Data for Development Challenge Senegal

Book of Abstracts: Scientific Papers

Orange uses big data for the benefit of the communities

Contact: Nicolas De Cordes, Orange, VP Marketing Anticipation, nicolasdecordes@orange.com
# Contents

## Session 1: Agriculture

1. [A01] Mobility profiles and calendars for food security and livelihoods analysis
   Pedro Zufrina  
   Damien Jacques  
3. [A03] Visual analysis on call data records for improving disaster resilience
   Brian Tomaszewski  
4. [A04] Unraveling correlations between agricultural events and phone traffic
   Rosa Benito

## Session 2: Energy

5. [E01] Using mobile phone data for rural electrification planning in developing countries
   Markus Schläpfer

## Session 3: Health

6. [H01] Quantifying the effect of movement associated with holidays on malaria prevalence using cell phone data
   Sveta Milusheva  
7. [H02] Uncovering the impact of human mobility on schistosomiasis via mobile phone data
   Marino Gatto  
8. [H03] Impact and scale of mobility on the spatial epidemiology of tuberculosis
   Bruno Lepri  
9. [H04] Human contact and the diffusion of Ebola in Senegal
   Rich Ling  
10. [H05] Detecting anomalies and supporting community to ensure healthy society
    Yutaka Hamaoka  
11. [H06] Weakening the incidence of transmittable diseases by taking advantage of mobile phone activity
    Manlio de Domenico
12. [H07] Human mobility and the spreading of waterborne diseases  
Enrico Bertuzzo

13. [H08] Progmosis: evaluating risky individual behavior during epidemics  
Antonio Lima

14. [H09] Developing an agent based migration model for Senegal for malaria transmission  
Adrian Tompkins

15. [H10] Forecasting Influenza in Senegal with call detail records  
Hao Wu

16. [H11] Mobile data as public health decision enabler: a case study of cardiac and neurological emergencies  
Edward Mutafungwa

Jonathan Leidig

**Session 4: National Statistics**

18. [N01] Towards connecting people, locations and real-world events in a cellular network  
Ramona Trestian

19. [N02] Spatial structure and efficiency of commuting in Senegalese cities  
Thomas Louail

20. [N03] Construction of socio-demographic indicators with digital breadcrumbs  
Timo Schmid

M. Saravananan

22. [N05] Detection of population mobility anomalies in Senegal from base station profiles  
Foued Melakessou

23. [N07] Virtual networks and poverty analysis in Senegal  
Neeti Pokhriyal, Wen Dong, Venu Govindaraju

24. [N09] Estimating population density and climate vulnerability using CDRs  
Emmanuel Letouze
# Session 5: Transports / Urbanism

<table>
<thead>
<tr>
<th>25. [T01] Roads+: estimation of travel times between cities in Senegal from CDR's</th>
<th>332</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jari Sarämaki</td>
<td></td>
</tr>
<tr>
<td>26. [T02] Deviations from the norm: detecting regularities and anomalies in human mobility patterns</td>
<td>338</td>
</tr>
<tr>
<td>Gijs Joost Brouwer</td>
<td></td>
</tr>
<tr>
<td>27. [T03] National and regional road network optimization for Senegal using mobile phone data</td>
<td>347</td>
</tr>
<tr>
<td>Erik De Romph</td>
<td></td>
</tr>
<tr>
<td>28. [T04] Building workers’ travel demand models based on mobile phone data</td>
<td>374</td>
</tr>
<tr>
<td>Liu Feng</td>
<td></td>
</tr>
<tr>
<td>29. [T05] Urban road construction and human mobility: evidence from Dakar</td>
<td>394</td>
</tr>
<tr>
<td>Amadou Sy</td>
<td></td>
</tr>
<tr>
<td>30. [T06] Travel demand analysis with differentially private releases</td>
<td>408</td>
</tr>
<tr>
<td>David Gundlegard</td>
<td></td>
</tr>
<tr>
<td>31. [T07] Cars and Calls: using CDR data to approximate official traffic counts</td>
<td>422</td>
</tr>
<tr>
<td>Vanessa Frias-Martinez</td>
<td></td>
</tr>
<tr>
<td>32. [T08] Towards mobile social crowd-sensing for transport information management</td>
<td>428</td>
</tr>
<tr>
<td>Georgios Bouloukakis</td>
<td></td>
</tr>
<tr>
<td>33. [T09] Operational planning of transport &amp; infrastructure using mobile data analysis</td>
<td>448</td>
</tr>
<tr>
<td>Ankit Singla</td>
<td></td>
</tr>
<tr>
<td>34. [T10] The municipality and the territory: two scales of understanding the city</td>
<td>452</td>
</tr>
<tr>
<td>Jerome Chenal</td>
<td></td>
</tr>
<tr>
<td>35. [T11] Sanitization of call detail records via differentially-private summaries</td>
<td>475</td>
</tr>
<tr>
<td>Mahammed Tuhin</td>
<td></td>
</tr>
<tr>
<td>36. [T12] A cell dwell time model for synthetic population generation from call detail records</td>
<td>486</td>
</tr>
<tr>
<td>Foued Melakessou</td>
<td></td>
</tr>
<tr>
<td>37. [T13] Improving mobility predictability based on user clustering on D4D Dataset</td>
<td>495</td>
</tr>
<tr>
<td>Jaeseong Jeong</td>
<td></td>
</tr>
<tr>
<td>38. [T14] From digital footprints to the dynamic population distribution and road network efficiency</td>
<td>501</td>
</tr>
<tr>
<td>Ruiqi Li, Dong Lei</td>
<td></td>
</tr>
<tr>
<td>39. [T15] Using mobile phone data for spatial planning simulation and optimization technologies</td>
<td>511</td>
</tr>
<tr>
<td>Serigne Gueye</td>
<td></td>
</tr>
</tbody>
</table>
40. [T16] Where do we develop? Discovering regions for urban investment in Senegal
   Derek Doran

41. [T17] Understanding traffic matrix for transportation planning with interregional connectivity
   Seungho Seungho Kim

42. [T18] High resolution traffic maps generation using cellular Big Data
   Ahmed El Mahdy

43. [T19] Structure and resilience of call networks in Senegal
   Anil Vullikant

44. [T20] Exploring relationships between human mobility motifs and rainfall
   Akihiro Fujihara

45. [T21] Text me - if you can: literacy, networks and mobility in Dakar
   Martina Kirchberger, Christopher Small

46. [T22] Cellphone’s users trajectories clustering from call detail records
   Paulin Melatagia Yonta

Session 6: Others

47. [O01] Generating anonymous sequence preserving datasets
   Sigal Shaked

48. [O02] On the anonymizability of mobile traffic datasets
   Marco Gramaglia

49. [O04] Human mobility during religious festivals in Senegal
   Christelle Scharff

50. [O05] Data for Development reloaded: visual matrix techniques for the exploration and analysis of massive mobile phone data
   Stef Van Den Elzen

51. [O06] Adaptive power load balancing in cellular networks
   Petko Bogdanov

52. [O07] Senegal Time-Space 2013
   Christopher Small

53. [O08a] High resolution mobility estimation from telecommunications data
   David Meyer

54. [O08b] Persistent Homology for Mobile Phone Data Analysis
   David Meyer

55. [O08c] Analyzing internal migration patterns using aggregated cell phone data
   David Meyer

56. [O08d] Glimpses of S’en’egal’s social network
   David Meyer
**Mobility profiles and calendars for food security and livelihoods analysis**


1 Universidad Politécnica de Madrid
2 United Nations World Food Program Senegal
3 Pulse Lab Kampala, United Nations Global Pulse

**1. INTRODUCTION**

Social vulnerability is defined as "the capacity of individuals and social groups to respond to any external stress placed on their livelihoods and well-being" [4]. Mobility and migrations are relevant when assessing vulnerability since the movements of a population reflect on their livelihoods, coping strategies and social safety nets. Although in general migration characterization is complex and open to controversy [6], changes in mobility patterns for vulnerable population groups are likely to indicate a change in livelihoods or coping strategies. These changes can also indicate that the population groups may be exposed to new shocks; hence, monitoring of changes in mobility patterns can be a powerful early warning mechanism.

Livelihoods in Senegal show a strong correlation with geographical location, and have been mapped out for analysis in different zones such as pastoralism, agriculture and fishing [2]. Within each of these zones, there are well-studied patterns of seasonal activities and population movements. However, such changes have until now been impossible to observe directly. Telecoms data therefore provide an important new opportunity to observe such changes in mobility patterns in real-time. For this purpose, we have developed statistical measures for profiling and calendarizing mobility in the context of livelihood zones and seasonal activity patterns in Senegal.

For each of the 13 mapped Livelihood Zones (LZ) of Senegal, we have characterized the profiles and calendars of the mobility flows from/to other LZs with different livelihood conditions. We have classified the population according to their mobility behaviors by clustering individual mobility trajectories into mobility classes. The timing of the displacements for each of the “mobility classes” has been aligned and compared with seasonal calendars and rainfall information. The calendar framework can be used to generate mobility baselines...
that combined with future real time data access could contribute food security early warning mechanisms.

2. MATERIALS AND METHODS

The proposed analysis is based on a model which gathers all the different types of variables considered to be relevant for characterizing any user mobility behavior. The model helps to integrate and analyse heterogeneous data with different time and space resolutions, by adjusting the domain of the variables from days to months or from antennas to livelihoods.

We start presenting the basic model variables based on the available data from the D4D datasets and external resources, newly defined variables and the developed analysis procedures.

2.1. Modelling variables

Here we present the different types of variables and their relationship with the available data.

2.1.1. Basic variables

The variables characterizing telephone users can be classified into:

1. User Behavior variables, $UB(t) = (l(t); c(t))$, gather both his/her geographical location $l = (l_a, l_o)$, and communication status $c$ along time.
2. Environment variables, $E(l_e, t)$, affect user behavior and depend on geographical location $l_e$ and time $t$ (e.g., rainfalls, holidays, etc.).
3. Indicators, $I(l_i, t)$, gather other relevant variables one may want to characterize (e.g., level of food insecurity of location $l_i$ at time $t$, etc.).

2.1.2. Derived secondary variables

Secondary variables can be derived from the basic variables. We can define two types:

1. User derived variables group the information (via time and/or space aggregation), keeping the (anonymized) user ID label. There are two types:
   a. Variables for which data are available (see Section 2.1.2.1).
   b. Variables defined for methodological purposes (see Section 2.1.2.2).
2. Environment or Indicator derived variables (see also Section 2.1.2.3).

2.1.2.1. Available data variables

The variables for which data are available in this challenge are:
• user bandicoot indicators \( b(t) \). Both Data-set 2 and 3, provide measurements of these variables in a monthly averaged basis.
• user Arrondissement location \( A(t) = A(l(t)) \) (derived from \( l(t) \)). Data-set 3 provides measurements of this variable for each user along the whole year.

2.1.2.2. Method related variables: Home or preferential location

These variables are required for the proposed methodology. They are relevant latent variables, since most environment variables and indicators depend both on space and time. Such variables can be estimated with different time and geographical resolutions, depending on the data employed. Based on Data-set 3, time aggregation procedures provide estimations of Daily-Home Arrondissement (DHA) and Monthly-Home Arrondissement (MHA) for each user. They can be complemented with the geographical location of the centroids corresponding to each Arrondissement. In addition, a geographical aggregation allows to consider the Monthly-Home Livelihood Zones (MHLZ) for each user. The D4D contextual data (shapefiles) have been used to aggregate the population from BTS to Arrondissements and from Arrondissements to regions or Livelihoods Zones (Figures 1a and 1b illustrate different levels of geographical resolution).

2.1.2.3. Daily Rainfall (DR) by Arrondissements or Livelihood Zones

They are obtained from a geographical aggregation of NASA’s TRMM sensed data [9], collected with a 0.25 resolution (longitude and latitude) in a daily basis.

2.2. Defining user feature vectors

MHA and MHLZ provide the location of users over time for the whole year with an Arrondissement and Livelihood Zone-month resolution respectively; this is complemented with the bandicoot information provided in Data-set 3, to define the feature vectors:

• Home Arrondissement User Vector (HAUV): A 13-dimensional vector comprising the user ID and his/her MHA for the 12 months.
• Home Livelihood Zone User Vector (HLZUV): A 13-dimensional vector comprising the user ID and his/her MHLZ for the 12 months (according to the map of Fig.1a/b)
• Bandicoot User Vectors (BUVs): for each bandicoot we have a 13-dimensional vector comprising the user ID and of his/her bandicoot value for the 12 months.

Our main objective is to unravel mobility patterns from the analysis of the HAUV, HLZUV and BUVs together with DR. Such analysis is aimed to classify the population mobility behavior into different groups depending on the period of the year and their geographical location. The results are complemented with the detection of general population movements associated with relevant events.
Fig. 1a: Livelihood zones map in Senegal. This map has been used to generate an Arrondissement to Livelihood assignment.

Fig. 1b: Levels of geographical resolution in Senegal based upon D4D datasets. The coverage regions of the antennas (red dots) correspond to Data-set 2 and can be aggregated to the Arrondissement level (line boundaries). In this work, the Arrondissements (Data-set 3) are grouped by Livelihood Zones (colors) rather than political boundaries.
2.3. Classification of feature vectors

2.3.1. Pre-filtering of mobility profiles

When considering mobility profiles, users can be filtered depending on different criteria. For instance, users whose HAUV or HLZUV components are all equal can be removed (considered as "non-moving" users). In addition, users for which the geographical distance corresponding to their Arrondissement change does not surpass a given ratio with respect to their radius of gyration (obtained from bandicoot data), can be also removed as “regular-travelers" and so forth. These filtering criteria were tested before the classification procedures (to be explained in the following Section 2.3.2).

2.3.2. Population clustering

Since a global country classification analysis does not seem feasible and useful for assessing livelihood related mobility patterns, the analysis has been performed by regions (at the Arrondissement and Livelihood Zone levels). At the Arrondissement level several alternatives have been evaluated for classification: if Arrondissement centroids or numbers are considered, the results, though very rich in terms of detailed information, are not easy to interpret. A binary representation indicating if the user is or not in the Arrondissement under consideration seems more tractable but the overall geographical interpretation remains complex. Therefore, the final analysis was addressed at a Livelihood Zone (LZ) level.

For each LZ, users who have visited it for some period of the year have been considered together, and their HLZUV analyzed. HLZUVs can be classified into different groups attending to mobility profiles. The binary representation indicating if the user is or not in the Livelihood Zone under consideration has been selected since it is easy to deal with and allows simple interpretations.

A new stage filtering procedure can be performed on the binarized HLZUV, depending upon several parametric temporal consistency constraints to remove noisy trajectories that may appear as a result of singular mobility profile or inaccuracy in the home location estimation:

- The user must have stayed at least $M_{\text{min}}$ consecutive months in the target LZ.
- The user must have not stayed more than $M_{\text{max}}$ months in the target LZ.
- The user must have stayed at least $M_{\text{out min}}$ months in some other LZ.
- The user must have stayed at a specific period of the year (when looking for specific types of mobility profiles, such as the ones related to rainfalls).

Then, the remaining binarized HLZUV are grouped into classes. Such clustering can be performed in different ways depending on:
• The type distance defined between vectors. Ultimately, the metric used should reflect
the specific perspective of users' behavior similarity in terms of mobility profile. So far
several distances (Euclidean, Manhattan, Cosine) have been used to generate
distance distribution between vector pairs, and they seem to provide similar results.
• The clustering procedure. Hierarchical clustering has been employed using a grouping
method relying on the "average" distance to build the tree nodes. The provided
dendrogram tree is cut by a maximum number of representative classes that may vary
between 4-5 classes for each LZ. Each of the cluster classes stands for a mobility
profile class within the population that has occupied the target LZ under the constraints
imposed.

The trajectories grouped together in different clusters, provide typical consistent mobility
profiles that can be used as seed information to understand migrations and social behaviours
to seasonal changes or large scale events.

2.4. Time gradients for event detection and period selection

The computation of time differences (or gradients) together with a threshold-based detection
scheme have been employed for global event detection and for relevant period selection.

2.4.1. Global event detection

Global event detection has been performed by analysing aggregated HAUV vectors; when
aggregating (by users) this information, general population movement behaviors can be
detected which are associated with relevant events in the country.

2.4.2. Time period selection for cluster analysis

Similar gradient computations on the profiles associated with user in a cluster provides the
relevant periods of time where most movements occur, allowing for a more specific analysis.

2.5. Complementary processing

2.5.1. Class characterization based on monthly locations and bandicoot data

Using the user IDs of each class, the corresponding bandicoot vectors have been classed up
together and statistically characterized with the mean and std, obtaining a behavioral
characterization of each class.

At the Arrondissement resolution level, a “Distance to Home Vector” (DHV) can also be built
using the HAUV and the estimated Arrondissements' centroids computed from the Senegal
map. The resulted averaged vector of the DHVs of each class shows a distribution of people
referred to the target Arrondissement weighted by the distance displaced. This information
has been expanded further by obtaining the “occupancy histogram” of each class. This histogram shows the number of people of the class that has occupied each Arrondissement along the time period comprised in the Data-set 3. This statistical characterization turns into a useful temporal characterization of people classes to be compared with other time series information, such as rainfall estimations, price changes, shocking events or seasonal cycles.

2.5.2. Validation of data processing

The resulting HAUV feature vectors have been compared to the vectors provided by a 1-step stationary Markov modelling of monthly displacements among Arrondissements. The correlation analysis between locations at different months derived from HAUV samples shows that they correspond to a non-stationary model: locations at summer months are less correlated with the rest of months locations; this validates seasonal (time dependent) population movements.

2.5.3. Rainfall estimations

Extracted from the TRMM-NASA project, they have been represented at different geographical resolution level in Senegal.

2.6. Visualization of mobility patterns and users’ characteristic mobility profiles

Real decision making tools must provide detailed temporal and geographical resolution of the population movements. Therefore, three different web tools have been designed and implemented to visualize the variables of the proposed model in an integrated way, adjustable to different data dimensionalities and resolutions:

1. **Viz1** [10]: Visualization of daily series of variables (primary ones and variations) at the Arrondissement level. This visualization includes a map based representation of the variables as well as an Arrondissement correspondence graph to complement such representation.

2. **Viz2** [11]: Visualization of the resulting distribution of HAUVs for different groups of people (datasets, filtered populations, clusterized classes,… ) by Arrondissements through time.

3. **Viz3** [12]: Visualization of the LZUVs also also for different types of population groups as a flow to understand the geographical distribution of the movements. It also embeds the visualization of the mobility profiles of the selected group and rain estimation diagrams.
3. RESULTS

The results obtained are:

1. Characterization of multi-scale mobility patterns for
   a. event detection;
   b. mobility profiling and calendarization of different communities.

2. We have characterized the relationship of mobility profiles with
   a. rainfall seasons;
   b. livelihood means;
   c. agricultural calendars.

3. For some regions, there are groups of inner population that show a yearly mobility profile in accordance to behaviors expected from other sources of information [3,7,8]. However, other groups display a profile which is not easily interpretable in such context and require further investigation. Even more, some regions seem to not have clear population groups following a specific pattern.

3.1. Multi-scale mobility patterns characterization

The different web tools developed allow for a characterization of mobility patterns at different aggregation levels with different applications.

3.1.1. Discovering events which drive strong mobility patterns abnormalities

**Viz1** [10] visualizes global movements among all Arrondissements: movements are coded via colors and arrows in a circle representing all the Senegal Arrondissements. For instance, daily user aggregated global movements can be represented, which is useful for general event detection.

Figures 2 to 4 show the potential of **Viz1** to understand and discover events or shock induced abnormalities in the mobility patterns. Figure 2 shows the behavior in a regular day: the color of each Arrondissement in the map based representation reflects the amount of mobility associated with it, whereas the Arrondissement correspondance graph provides a detailed origin-destination map of global movements (the color of each line represents the destination Arrondissement and its width is proportional to the amount of such movement; the size of each ring slice represents the total amount of people leaving such Arrondissement). Figures 3 and 4 illustrate the population movements corresponding to day numbers 355 and 357 of year 2013, when a national event occurred (Grand Magal at Touba). Therefore, the variations in the variables of the model may be exploited as a abnormality detection metric [5] to select candidates of significant events or shocks, when compared to the typical day characterization of movements in Senegal at a specific geographical level.
Figure 2. Regular distribution of movements in a normal day of year 2013.
Figure 3. Grand Magal detection on the 21st of December 2013. Both images show how people from almost all Arrondissements move to the city of Touba in the Arrondissement of Ndame.
3.1.2. Characterizing the destinations map of target populations through time.

Viz2 [11] visualizes movements from/to a selected Arrondissement: the processed HAUVs (e.g., those corresponding to a class) are shown on Senegal’s map: for each month the number of moving people to each destination Arrondissement is color-coded. This is useful for low resolution movements.

Figure 5 shows how Viz2 [11] helps to understand the mobility distribution of a population referred to one Arrondissement through the year.
Figure 5: Distribution along Senegal for the population that was completely localized in the Arrondissement of MATAM during July.

This type of visualization shows how well organized the mobility profiles of specific population groups are; also, it helps to discover the preferential destinations of any mobility pattern and how this pattern evolves during the whole year.

3.1.3. Understanding flows of moving population between different livelihood regions

**Viz3** [12] visualizes movements from/to Livelihood Zones: for a given a set of HAVUs (selected by some of the developed classification techniques), the number of moving people to and from each destination livelihood zone is coded by size and color as well as indicated by the corresponding arrows on a Senegal map. This is useful for livelihood related movements.

The module allows the visualization of incoming (blue) and outcoming (red) flows of people for specific target populations selected (via any specified criterion) from the whole Senegal, depicting the interactions between the different livelihood zones. Here we illustrate the visualization of population groups provided by the clustering scheme explained in Section 2.3.2. By embedding the temporal profile of the target population and other external signals at the same scale, we can add more dimensions to these livelihood zones interactions. Figure 6 shows the mobility profiles corresponding to the group in Livelihood 8 (agropastoral zone specialized in peanut culturing) that occupies this zone towards the second half of the year after the rain season, and Figure 7 shows another group mobility profile in the same livelihood area corresponding to people who leave this zone during the rainy season.
3.2. Relationship between onset of large mobility changes and the end of rainy season

Population movements in the north of Senegal are expected to start in October due to the end of the rainy season [3]. A major potential advantage of the use of CDR data is that the onset of population movements may be estimated with high accuracy by measuring the changes in the mobility patterns of the users. This way, the actual reaction of the population to the change of season can be quantitatively measured.
3.2.1. Estimation of rainy season

Rainfall estimations have been extracted from [9]¹. The estimation has been calculated for Jan’11 until Dec’13, in order to observe yearly variations for a better interpretation. Fig. 8 summarizes the averaged rainfall by Arrondissements for the period observed; 2013 did not have significant changes when compared to 2011 (a less rainy year) or 2012.

Figure 8. Top: Temporal evolution of precipitation levels by Arrondissement in 2013. Insets: Visualization of rainfalls over the map of Senegal at days: Feb 1st, Jun 1st and Aug 8th (from left to right). Bottom: Correlation between rainfall levels by Arrondissement in 2013 against 2011 and 2012.

3.2.2. Comparing the onset of mobility alterations with the end of rainy season

The results shown in Figures 6 and 7 feature the averaged rainfalls for the livelihood zone 8 (aggregated using the map in Fig. 1b) in a monthly time scale in order to simplify the analysis and visualization (as well as to compare them with agricultural calendars as explained below).

¹ http://erdos.mat.upm.es/d4d-senegal/rain.mp4
As observed, the selected mobility profiles are clearly influenced by the rainfall levels in the target candidate zone population that might be vulnerable to severe climate changes.

3.3. Discovering characteristic users’ mobility profiles depending on the livelihood means

We have used both HAUVs and HLZUVs to obtain different classes (profiles) of temporal patterns of mobility regarding both Arrondissements and Livelihood Zones (LZs). Focusing on the latter, the clustering method (see materials and methods) provides groups of people that show the same occupancy profile in the target LZ; this classification can be cross-checked with the tagging of each LZ of Senegal according to the data in [2]. This process has been repeated for each of the Livelihood Zones in Senegal: some LZs provide expected results where other LZs show a non-easily interpretable behavior, requiring further consideration.

3.4. Visualizing correlation between mobility profiles and agricultural calendars

The characteristic profiles displayed in Figures 6 and 7 can also be time correlated with agricultural calendars (see [2]) as shown in Figures 9 to 11. The existence of correlation could spotlight groups of people that migrate depending on the agricultural cycles of livelihood zones of Senegal. Potentially, with a long-term characterization through several years, this strategy could help to monitor in real-time population vulnerable to climate changes or production alterations. However, detailed local analysis should be carried out to confirm and validate this hypothesis.
For instance, in Zone 4 (Fig. 9), the calendar interval for planting and weeding -green [2]- of resources (millet, cassava, watermelon, peanuts, hibiscus) which implies the rise of employment due to local labor -dark blue-, seems to trigger changes in several migration classes, although there is not a specific profile with a strong temporal correlation with this interval (only the third profile seems to correlate with a significant delay).

A more clear correlation between a mobility profile (row 1) and a calendar interval is found in the Zone 6 (Fig. 10) during the milk sales, as this region is specialized in herding and transhumant livestocks. As a hypothesis, variations in this correlation and the significance of this mobility profile could provide information about the success of the milk production and sales which would impact the regions economy and vulnerability.
Figure 10. Zone 6 calendar of sources of income and activities against users’ mobility profiles.

A very important livelihood zone of Senegal is the one specialized in peanut production (zone 8 in Fig. 11) in the central area of Senegal. The first two mobility profiles seem to correlate with the beginning of the planting preparation and the planting, when there is an increase of population due to the peanut production cycle. However, this is a very complex region involving seasonal migrations since it is a major transit zone from west to east and north to south; hence, further and more precise analysis is demanded to understand seasonal mobility through this zone. Another complex zone of study is the zone 13 (Fig. 12), where only the fourth mobility profile seems to correlate with the planting and collection of the crops of this zone.

It is important to notice that these calendars represent a normal characterization of the country production cycles. However, migrations could greatly change depending on external factors such as the rainfall levels, the market prices or extreme conditions or shocks. Only a long-term observation of calendars considering external variables (such as the estimated rainfalls) would enable to distinguish when the profiles are really driven by agricultural calendars and when they may be modulated by external factors.
Figure 11. Zone 8 calendar of sources of income and activities against users’ mobility profiles.

Fig. 12. Zone 13 calendar of sources of income and activities against users’ mobility profiles.
4. DISCUSSION

This work has been motivated by the need of analyzing and quantifying the role of mobility patterns in the communities lifestyles and their access to basic resources, with the more precise and up-to-date information that the CDRs provide.

The developed processing and visualization prototype comprehensively integrates heterogeneous data for multiple use cases; here we have illustrated its potential by performing several off-line analyses (event detection, population mobility profiling and calendarization, etc.), with special focus on the possible interplay between mobility at the level of Livelihood Zones, accurate rainfall sensed data and agricultural calendars.

The time range limitation of the available Data-sets (an overall single year) does not allow for the robust design of on-line detection schemes, since seasonal reference baseline behaviors cannot be constructed. Nevertheless, the developed schemes can be easily extended to perform on-line detection provided larger time range data are available.

If the anonymization preserving additional limitations, either in time range or geographical resolution, of the Data-sets (fortnight time range for Data-set 2, Arrondissement resolution for Data-set 3) were removed, more robust and accurate results could be obtained. In general, this new approach to mobility patterns analysis could be very helpful to monitor vulnerable communities and to understand the impact of mobility patterns in the production means of Senegal.

ACKNOWLEDGMENTS

The authors want to thank the support of Ministerio de Ciencia e Innovación of Spain via project MTM2010-15102, and Cátedra Orange at the ETSI Telecomunicación in the Universidad Politécnica de Madrid (UPM), Spain. They also want to thank Juan Fernando Sánchez-Rada for his help in developing some of the web applications.

REFERENCES


Genesis of millet prices in Senegal: the role of production, markets and their failures

Damien Christophe Jacques 1, Raphael d’Andrimont 1, Julien Radoux 1, Francois Waldner 1 and Eduardo Marinho 2

1 Earth and Life Institute, Université Catholique de Louvain, Belgium
2 Rua Capistrano de Abreu, 33 - Rio de Janeiro, Brazil
31th December 2014

ABSTRACT
Staple prices are the main indicator of food access and a key determinant of the revenues of those living in agricultural zones. Differentials in prices between producing (low prices) and consuming (high prices) areas harm both groups and indicate the presence of market failures. In this study we model the millet prices formation process in Senegal in a spatially explicit model that accounts for both high transportation costs and information asymmetries. The model integrates a unique and diversified set of data in a framework that is coherent with the economic theory. The high ability of the model in reproducing the price differentials between 41 markets ($r^2 > 80\%$) opens a new avenue for the research on market integration which (i) integrates production data derived from remote sensing, (ii) simulates the demand and supply at the local level and (iii) the arbitrage process between imperfectly integrated markets.

Contact: damien.jacques@uclouvain.be

1 INTRODUCTION
In 2008, when international food prices reached their highest values in the last 30 years, a new global challenge has emerged as emphasized by FAO Director-General Jos Graziano da Silva: “If higher and volatile prices are here to stay, then we need to adapt to this new pattern.” Since then and at the request of Agricultural Ministers of the G20 2011, the top-leading group of international organizations working on food related matters have created the Agricultural Markets Information System in order to improve food market transparency and encourage policy coordination. Clearly enough, there is a true need for understanding how food markets operate, as it is a sine qua non condition for the implementation of appropriate food policies and ensure food security at the global scale.

This need is also true at the domestic scale, in particular in low-income and food-insecure countries where the rain-fed agricultural production relies on erratic rainfall patterns and where market failures result in the imperfect allocation of resources. Indeed, in such environments consumers suffer from high and volatile food prices that do not benefit producers since the price differentials between consumption and production markets can be substantially high. The Sahel is a a critical example of such situation. Indeed, in Senegal, average price differentials between markets can reach more than 50% as a consequence of market failures. Moreover, due to the high variability on the agricultural production, both yearly and spatially, market prices are very volatile at the local scale.

In this study we explore the formation process of food prices in Senegal, with a specific focus on millet. Our goal is to reproduce the whole dynamics behind the functioning of the Senegalese markets, from the production to the retail sale, by simulating profitable transfers of millet from surplus to deficit areas. We then assess what market failures are likely to generate the price differentials observed between markets in the country.

The spatially explicit model integrates a rich set of data coming from different sources. Local supply and demand are respectively derived from remote sensing and population density maps. The road network is used to establish the markets catchment areas and the distances between each couple of markets as a proxy of transportation costs. Finally, a unique dataset on mobile phone communications between the antennas within the country is used as a proxy for information circulation between the markets. This data is then put together in a model coherent with the economic theory. Actual millet prices are used for validation purposes.

2 MATERIAL AND METHOD
In order to model the market integration, the first step is to consider the extreme situation where all markets are independents i.e. when there is no exchange of merchandise between the markets. In this case, a pseudo-price can be defined for each market as a function of the demand (estimated by the population) and the supply (estimated by the production) found in each area covered by the market (Eq. 2).

$$P_i = f(D_i, S_i)$$

with

$$\frac{dP_i}{dS_i} > 0; \frac{dS_i}{dP_i} < 0; \frac{dP_i}{d^2S_i} > 0$$

where $P_i$, $D_i$, $S_i$ are the price, the demand and the supply for the market $i$.

Our approach (figure 1) is to model these pseudo-prices, that are expected to be proportional to the actual millet prices, using the population (see section 2.3) as a proxy of the demand and the local food production approximated using a vegetation index derived from satellite images (see section 2.4) as the supply input (Eq. 2). Both data are spatialized and the aggregation area of each
market (catchment areas) is the area that minimize the journey time (using road network) between each market (see section 2.2).

\[ P_i \propto P_{sPr_i} = \frac{\text{Pop}_i}{\text{Prod}_i + 1} \tag{2} \]

where \( P_{sPr_i} \), \( \text{Pop}_i \), \( \text{Prod}_i \) are the pseudo-price, the population and the production for the catchment area cover by the market \( i \).

The opposite situation is a completely open market where food flows freely from surplus markets to deficits markets or from areas of production (rural, agricultural areas) to areas of consumption (urban centers). In this particular case, transfers of merchandise occur until an equilibrium is reached with a unique price throughout the country.

The reality lies in between these two extreme situations. Transfers of merchandise between two markets occur if the transportation cost is less than the difference of prices between these two markets. In this study, the impact of the inefficiency of the circulation of the price information between the markets is also studied. In addition to the transportation cost, we therefore introduce an information cost that reflects the risk to move from one market to another market where the price is not well known. The higher the asymmetry of information, the higher information cost between two markets (Eq. 3):

\[ TC_{i,j} = \beta_{i,j} \times \frac{P_{sPr_i} - P_{sPr_j}}{d_{i,j} + 1} \tag{3} \]

where \( TC_{i,j} \), \( \beta_{i,j} \) and \( d_{i,j} \) are the transfer cost, the information cost (see section 2.5) and the distance (see section 2.2) between the market \( i \) and \( j \).

To mimic the dynamics behind the functioning of the Senegalese markets, we have built a model that simulates profitable transfers of millet from surplus to deficit areas. From a complete segregated market situation (Eq. 2), we have computed the pseudo-prices as observed at the equilibrium when several plausible combinations of information and transportation costs (i.e. several transfer costs) are carried out. The equilibrium is reached after all the profitable transfers of production units between the markets have been occurred i.e. when the difference of the new prices between two markets is not sufficient to justify a transfer of merchandise. The correlation between the pseudo-prices obtained at the different equilibrium and the actual prices are then computed. Situations leading to high correlation are assumed to be representative of the actual functioning of the market. The various contribution of the information and transportation cost can then be analyzed.

2.1 Market prices

Rice, millet and sorghum are the main subsistence food crops for Senegal’s rural population but millet is definitely the most vital. This crop beats out the other major staples as the most drought resistant and ne third of Senegal’s arable land (1 million hectares) is devoted to it. Most of the millet is grown in the regions of Kaolack, Kaffrine and Fattick where it is interchanged with peanuts. This crop rotation is crucial as peanuts fix nitrogen into the soil. Generally,
production of cereal food crops does not meet Senegal’s needs. Only in years of good rainfall, the country approaches self-sufficiency in the main staples in rural areas. In 2005 and 2006, for example, the total production of cereals was estimated at 1,177,782 MT, covering only 60% of the consumption needs. In years of poor rainfall, the shortfall in grains, especially millet, could be more difficult to cover because of low availability and trade of this grain in the region. Such constraints have been overcome with an increase in rice imports, leading to a shift from millet to rice consumption in households who can afford it [Dong, 2011].

Domestic price data are coming from the VAM Food and Commodity Prices Data Store of the UN World Food Program [VAM, 2014]. The data set consists in monthly retail prices (when available) from 41 markets (one market was discarded because its geolocation appeared uncertain) distributed in the 14 regions of Senegal for the years 2012.

2.2 Transportation modelling

Most of the food transport in Senegal relies on the road network. It was therefore assumed that the production transfers were driven by the proximity of producers to markets. The distance by road was used to approximate the transport costs and the catchment areas. A topological network has been built based on the Global Insight dataset and minimum travelling times computed using Dijkstra’s algorithm.

Transport cost is assumed to be directly proportional to the distance between the markets using the road network. The road network was therefore used to compute an origin-destination cost matrix for all markets. Figure 2 illustrates the transport cost for one of the markets.

The catchment areas of each market have been computed based on the best approximation of the most accessible markets. In absence of secondary travelling directions, the main underlying hypothesis is that the farmers will travel to the nearest main road and then go to the nearest market using the road network in order to sell their production. First, the closest road segments have been computed for each market. Each road segment is thus assigned to a single market based on the least travelling distance by road. Second, Euclidean (bird fly air) distance allocation of a raster grid to the nearest road segment yielded the catchment areas (figure 3).

In the absence of communication fluxes with other countries, foreign existing markets were not taken into account and the catchment areas were clipped to the boundaries of the Senegal. Therefore, border effects could occur but are likely to be very small due to the constraints of international trade for the small producers.

2.3 Demand and Population

To estimate the demand, population distribution maps from the Afripop project have been used [Linard, 2011]. Afripop maps present estimates of numbers of inhabitants per grid square with national totals adjusted to match UN population division estimates. As the population for the year of interest (2013) was not available, it was simulated with national population growth rates from the World Bank assuming that the growth is equally distributed over the Senegalese territory.

2.4 Supply and Production

Satellite remote sensing provides a suitable alternative for crop condition and yield estimation, as it gives a timely, accurate, synoptic, and objective estimation of various yield-directly related crop parameters such as net primary production [Ren et al., 2008]. Vegetation indices are widely used in crop growth monitoring and yield estimation based on remote sensing technology. Most of the vegetation indices are information-condensed which can reflect terrestrial vegetation cover and growth condition effectively and economically. Substantial research has shown that Normalized Difference Vegetation Index (NDVI) is a reliable index that can be related to crop yield [Manjunath et al., 2002] but also and more specifically to millet yield and production [Rasmussen, 1992]. NDVI is defined as the difference between near infrared and red reflection normalized by the sum of the two. The NDVI-values vary from 0.15 for bare soils to 0.80 for full green vegetation, with
all gradations in-between. A large number of metrics have been devised to relate NDVI with yield or production: maximum NDVI [Lewis et al., 1998], sum of NDVI between flowering and ripening [Genovese et al., 2001], Cumulative NDVI [Quarmby et al., 1993], or cumulative NDVI from the onset to the end of season, maximum NDVI throughout the crop season.

In order to link the NDVI metric to actual production, millet production statistics have been downloaded from the Senegal Ministry of Economy and Finance. However, as the granularity of these statistics is at the regional level (14 regions), it is needed to convert them at the market catchment level. To deal with this mismatch of spatial unit, a three-step procedure was followed to i) mask the agricultural areas, ii) define a spatially explicit proxy of the crop production and iii) redistribute subnational statistics at the catchment level.

First, the cropland areas have been masked using the Senegal Land Cover Map of 2005 at the 1:100,000 scale produced by the Global Land Cover Network [Leonardi, 2008]. Lacking reliable information on the spatial distribution of millet, it is here assumed that this crop is grown evenly within the cropland area.

Second, 10-day temporal synthesis of SPOT-VEGETATION NDVI at 1-km have been downloaded over the area of interest from 2012. In the multi-temporal image set, each pixel is thus characterised by a specific NDVI-time profile. However, since the raw profiles are still disturbed by cloudy measurements, the composites images are first submitted to a cleaning procedure by means of the Whittaker smoother [Eilers, 2003]. For each pixel within the cropland, NDVI values above 0.2 observed during the millet growing season were integrated, limiting thus the contribution of the soil to the signal. The actual millet production observed at the regional scale was then spatially distributed at the pixel level:

$$\text{Prod}_i = \frac{\text{Prod}_{\text{region}}}{\text{CUM}_{\text{NDVI}}^\text{region}}$$

where Prod$_i$ is the estimated millet production for a pixel $i$, CUM$_{\text{NDVI}}^i$ is the cumulated NDVI above 0.2 for the same pixel $i$, Prod$_{\text{region}}$ is the millet production of the region of pixel $i$ and CUM$_{\text{NDVI}}^\text{region}$ is the cumulated NDVI for the entire region.

Finally, using the market catchment area boundaries the pixel’s production values were aggregated to give millet production by catchment areas.

2.5 Information cost modelling

Mobile phone data have been provided by Sonatel Orange in the frame of the Data For Development (D4D) challenge. The D4D-Senegal challenge is an open innovation data challenge on anonymous call patterns of Oranges mobile phone users in Senegal [de Montjoye et al., 2014]. An original dataset of phone calls and text exchanges between more than 9 million of Oranges customers in Senegal between January 1, 2013 to December 31, 2013 have been sampled based on two criteria:

1. users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset)
2. users having an average of less than 1000 interactions per week.

The users with more than 1000 interactions per week were presumed to be machines or shared phones.

For commercial and privacy reasons, the exact location of the base transceiver stations (BTS), the mobile network antennas, has not been delivered. A new random geolocation has been associated to each site in its Voronoi cell i.e. in the region where all points are closer to that antenna than to any other. Among the three data sets delivered by Sonatel Orange, only the first one, antenna-to-antenna traffic for 1666 antennas on an hourly basis (number of sms, number of calls, duration of calls), has been explored. From the three variable proposed, the number of calls has been selected as it has been shown to be the more relevant variable for the purpose of the study.
Antennas in a buffer of 10 km around the market places have been aggregated and associated to each market. Due to their close proximity, the markets of Dakar; Diaobe and Sare Yoba; Ourossogui and Matam; have been merged. For each market the sum of the number calls from the associated antenna has been computed by month and averaged over the year. From this value, a contingency table (cross-tabulation) between all the markets has been defined giving the averaged number of calls over one year for each combination of origin-destination markets. From this, the parameter \( \beta \) used to estimate the cost of information is defined as:

\[
\beta_{i,j} = \begin{cases} 
1 & \text{if } \frac{1}{\text{IC}_{i,j}} \left( \frac{\text{log}(N_{\text{calls}_{i,j}}) - \min(\text{log}(N_{\text{calls}_{i,j}}))}{\text{IC}_{i,j} - \min(\text{log}(N_{\text{calls}_{i,j}}))} \right) > 1 \\
\text{else} & \end{cases}
\]  

(5)

where \( N_{\text{calls}_{i,j}}, \text{IC}_{i,j} \) are the number of calls and the between markets \( i \) and \( j \).

3 RESULTS AND DISCUSSION

As expected, before the transfers from surplus to deficit areas start, the correlation between actual millet prices and pseudo-prices is very low \( (r^2 = 0.26 \text{ for April, } r^2 = 0.23 \text{ for August}) \), which allow us to reject the perfect markets segregation in Senegal; while the perfect integration, or the law of one price, is directly rejected by the observed price differentials among markets. It leads us to explore the intermediary situation of imperfect arbitrage, i.e. the presence of information asymmetries and transportation costs. Figure 7 shows the correlation between pseudo-prices, under several regimes of transportation costs and the \( \beta \) parameter of information asymmetry, and millet prices in 4 selected months.

More pragmatically, the figure 8 shows an idea of the main markets where and when the circulation of the information is imperfect and could be improved. Using a model such as the one presented in this study could pave the way to address some market failures created by the asymmetry of the information.

4 CONCLUSION

This study aims at describing and simulating the formation process of millet market prices in Senegal. To the best of our knowledge, that is the first time that such an approach is implemented in a Sahelian country. The model shows a very good ability to reproduce the price differentials observed in the country with \( R^2 > 80\% \). This pioneer work opens a new avenue for (i) the already rich literature on market integration (ii) the integration of the two first pillars of food security, i.e. availability and access and (iii) the development of the food security early warning systems in the region. New findings are expected from the use of several years of mobile phone data and the expansion of the model to other Sahelian countries.
ACKNOWLEDGEMENT

Thanks to Orange Sonatel Senegal and the D4D team for providing the mobile phone data. Support from the Belgian National Fund for Scientific Research through a FRIA grant is acknowledged.

REFERENCES


**Abstract**

Natural disasters like floods have compounded effects on agriculture, which in turn, result in food insecurity and malnutrition. The Senegalese population is particularly vulnerable to floods due to existing poverty. We analyzed Data for Development (D4D) Call Detail Records (CDR) datasets to (1) find statistically significant spatial clusters of areas vulnerable to floods and (2) identify spatial interactions between call origins and destinations to understand calling behaviors during flooding events. In depth contextual inquiry was done to identify flood occurrences within the peak flood month of August and with a particular focus on Dakar and Saint Louis. Visual analysis was conducted on call records from August 27 and August 29. Our analysis revealed several interesting spatial interaction patterns. Specifically, we found that there were higher number of calls and sms during the peak flood dates. Visual analysis of call origins and destinations also potentially revealed existing social networks that are potentially utilized for economic and social relief during floods. Areas contacted during floods are located not only near large population centers like Dakar and Saint Louis but also in farther away regions like Kaolack and Ziguinchor. We argue that identifying spatial interactions during floods using CDRs can help build resilience to natural disasters and related events such as food insecurity and other flood-related health issues. We conclude the paper with specific recommendations for future work based on our findings.

**1. Introduction**

Senegal is one of the 16 countries that make up Western Africa. Its capital, Dakar, is located on the Cape Verde Peninsula and is the westernmost city in Africa. Senegal has a population of 13.7 million with 46.7% of the population living in poverty [2, 6]. Its average annual rainfall is 600mm which primarily occurs during the rainy season between June and October [1]. During the rainy season, Senegal often experiences extreme flooding events [5, 11]. Of the almost 14 million people, 11.4 million live in an area known as the Sahel region or the Sahel Belt. The Sahel region covers a major portion of Senegal, including Dakar. The combination of high poverty rates and seasonal flooding make many Senegalese living in the Sahel vulnerable to
natural disasters and the after effects of natural disasters [8]. People living in the Sahel often require immediate food assistance as a result of existing poverty conditions coupled with the effects of floods that in impact agricultural production causing food insecurity [3, 9, 10, 13]. In September 2013, the WFP (World Food Program) launched a mobile cash transfer program to the food insecure people in Senegal through mobile phones [3, 4].

The Orange Mobile Company provided a mobile Call Detail Records (CDR) dataset for Senegal from the year 2013. The data contains outgoing calls, incoming calls and text messages from 1666 antennas around Senegal. In this study, we examined calling patterns related to flooding events to understand how people and places in Senegal interact during flood events. Using the data provided by Orange, we found interesting patterns in the data that correspond to the flood dates in Senegal. We believe that these floods and the region’s food insecurity and malnutrition cases are correlated which could account for the 50% rise in malnutrition cases from 2012 to 2013 [5, 8]. In the following section, we first describe our methodology for processing, storing and retrieving the CDR datasets that were the basis for our analysis into calling behaviors, spatial interactions and flooding events.

2. Methodology

2.1 Database Creation

A database schema design was implemented that mirrored the structure of each file provided by Orange for the Data for Development (D4D) challenge [12]. Each .csv file in the dataset had a corresponding table created in the database. These files were then loaded into their appropriate table using SQL Server Business Intelligence Development Studio. After the tables were loaded, primary keys and foreign keys were created on the tables. Creating the keys after loading the data reduced the data import time. The primary keys also created indexes on the table which increased the speed of the querying the data. We then created views by joining voice, sms and site tables in order to find patterns for a given region which corresponds to multiple site_ids on the day of major events such as floods. We also used views to restrict number of rows for prototyping purposes. We leveraged the views to abstract table complexities and provide a cleaner virtual table containing all the information needed for further in-depth analysis. Views can perform as fast as the direct query on the tables, so we created indexes on tables to bolster view’s data retrieval performance [14].

2.2 Spatial Analysis

A geo-database was created with different reference datasets such health centers, cellular tower locations, settlement areas and nutrition. Using Geographic Information Systems (GIS), these reference datasets were then overlaid with CDRs extracted from our central database (discussed
previously in section 2.1) for specific flood dates identified for Dakar and Saint Louis (discussed next in section 3) to identify spatial patterns during the floods. We used the XY to Line Management tool\(^1\) of ArcGIS [7] to identify call tower origin and destination pairs for the flood dates using tower latitude and longitude coordinates combined with relevant call data (duration, call date, etc.). As seen in Figures 2, 3, and 4, the XY to Line tool provides a visual summary of tower origin to tower destination calls. The Getis-Ord-Gi* spatial statistic\(^2\) was then used to identify the statistically significant clusters or “hot spots” where the most calls were being made during flood dates using outputs from the XY to Line tool. The Getis-Ord-Gi* spatial statistic helped in identifying statistically significant spatial clusters of high values and low values by making use of aggregated data and dependence between the attributes – in this case, tower locations sending and receiving calls during a flood. In the following section, we describe our contextual inquiry into flooding in Senegal and results of our inquiry using the aforementioned methodology.

3. Contextual Inquiry: August 2013 Flooding in Senegal and Call Activity

Contextual inquiry on Senegal clearly illustrated that the major flood regions were Dakar and Saint Louis for specific dates in August 2013 (Figures 1.1 and 1.2).

![Figure 1.1: Dakar – August 2013 call patterns.](image1)

Note spike in calls on 27 August and then the gradual drop off after 27 August.

![Figure 1.2: St. Louis – August 2013 call patterns.](image2)

Note spike on 29 August and then the gradual drop off after 27 August.

We first conducted analysis on Dataset 1 (antenna to antenna traffic on an hourly basis) to identify specific calling patterns related to August 2013 flooding events. Visual exploration of

---

CDR data for August 2013 revealed several patterns which showed a drastic rise in number of calls and sms on particular days. Specifically, we identified 27 August 2013 for Dakar and 29 August 2013 for Saint Louis as significant flooding dates.

4. Results & Discussion

A significant pattern was found in analyzing Call Data Records (CDR) dataset 1 (Antenna to Antenna) with respect to specific flood dates in Dakar and Saint Louis. The results for the pre- and post- flood dates for Dakar and Saint Louis yielded a visual pattern which showed a drastic outreach to regions such as Kaolack, Thies and Louga identified through hot-spot analysis. Dakar also reached out to localities outside the aforementioned hot spots, irrespective of the proximity to the flood affected area such as Ziguinchor, Kolda, Tambacounda and Matam. We were not able to determine why these specifically areas were contacted. Existing research on flooding events and behaviors reflected in CDRs is still a very new field [16, 18, 19]. A recent study conducted by United Nations Global Pulse (2014) examining CDRs during floods in Mexico found that increases in mobile activity provide signals of flooding impact and call volumes increase in impacted areas [15]. However, the Global Pulse study did not discuss where people impacted by floods were calling to or why they called one region or another. From our study, we can infer that people being affected by floods (as reflected in high frequency of call volumes from certain cell tower locations) were calling other locations in Senegal more than normal. We wish to suggest that the reason for this behavior is that people affected by floods were likely contacting relatives, friends or other people more frequently than normal in order to provide social or financial support as this is a very common behavior during disasters [20]. This inference is illustrated in Figures 2.1 and 2.2. It is visually evident that normal calling interactions between Dakar and other areas increase dramatically during flood events.

![Figure 2.1: Dakar – 25 August (Pre-Flood Date).](image1)

![Figure 2.2: Dakar – 27 August (Flood Date).](image2)

Note increases in call activity as visually reflected in increased lines.
When mapping call flows from Saint Louis for the 29 August flood date, a similar pattern emerged like seen in Dakar. Increased call numbers were seen during the flood event, this time targeted towards the Kaolack, Thies and Dakar region (Figures 3.1 and 3.2). Calls originating from St. Louis also indicate that people living in St. Louis could have strong social and economic connections with Dshra, Linguere and Keberner.

![Figure 3.1: Saint Louis – 25 August (Pre-Flood Date).](image1)

![Figure 3.2: Saint Louis – 29 August (Flood Date).](image2)

A cumulative spatial analysis was done for both Dakar and Saint Louis based on the outgoing calls on an hourly basis from Dataset 1 which revealed an interesting pattern seen in Figures 4.1 and 4.2. This pattern shows a huge concentration of calls being made to Kaolack, Matam and Thies.
Figure 4.1: Outgoing Call Pattern for Dakar and Saint Louis on the 27 and 29 August flood dates.

Figure 4.2: Incoming and Outgoing calls on the 27 and 29 August flood dates overlaid on heat map
Finally, based on our analysis, we identified the top ten towers making calls during flood events in Dakar and St. Louis as a starting point for making recommendations based on our study (Figure 5.1 and 5.2).

**Figure 5.1:** Top ten call volumes by tower during August 2013 flooding events – Dakar.

**Figure 5.2:** Top ten call volumes by tower during August 2013 flooding events – St. Louis.
In both cases, call origins and destinations had the same tower ID. Thus, as per previous research on CDR during flooding events discussed previously in this section, it is likely that locations around these tower locations are particularly vulnerable to floods.

5. Conclusions and Recommendations

Analyzing the outgoing and incoming call data patterns for the most affected flood areas (i.e. Dakar and Saint Louis), we conclude that these places, during flood events, are likely contacting other locations throughout Senegal for social and economic support. We also observed call patterns outliers with calls being made to regions far away irrespective distance to flooded areas (as seen in Figures 2, 3, and 4).

Based on our analysis, we recommend the following:

1. Further analysis of outgoing calls in and around Koalack, Matam and Theis during the flood events. This could help determine (1) who is being called and why they are being called during floods for understanding food security and other health issues that arise during the floods and (2) gaining insight into existing social networks that can potentially be utilized before, during, and after natural disaster events for building societal resilience [17].

2. Conduct neighborhood-level disaster resilience studies on area around the tower locations shown in Figures 5.1 and 5.2. As previous research indicates, calling patterns can serve as surrogates for understanding location most vulnerable to floods [15]. Ideally, further understanding of the geographical context around these locations can also help to reduce risk and build disasters resilience.

3. Compare and investigate if correlations can be found between floods events, food security situations and mobile cash transfers like those done by Orange in 2013 [3]. Ideally, understanding this broader sequence of interrelated events can lead to better predictability and use of CDRs to identify and anticipate food security and other shock situations before they occur or mitigate their effects. Furthermore, understanding such event sequences can lead to improved resilience to disasters and better decision making for targeted relief efforts like those conducted by Orange [3].

References Cited


Unraveling correlations between agricultural events and phone traffic

Grupo de Sistemas Complejos, Universidad Politécnica de Madrid,
ETSI Agrónomos, 28040 Madrid, Spain

December 31, 2014

Abstract

In this work we have analyzed the impact that agricultural activities, such as the growing of groundnut, have on Senegal. To this end we have analyzed the NDVI time series of the whole of Senegal and spotted the regions where groundnut is grown. By analyzing phone calls at each region of the country we found that a significant fraction of antennas exhibit two well defined peaks of activity corresponding with the begging and end of the growing season. Antennas located on regions identified as growing regions present this pattern. However, other antennas, located in non growing regions, such as Dakar, also present the two peaks pattern. Hence, in a preliminary inspection of our results we think that there is a synchronization between growing regions and key points in cities that emerges from the agricultural activity of the country. Finally, we observed a significant difference in the community structure of the country during the pre-growing and growing seasons.

1 Introduction

Senegal is a small country with a total land area of about 196 722 square kilometres. It is part of a climatic zone characterized by a weak and random rainfall with severe and extended cyclical droughts, on one hand, and the predominance of arid and fragile lands susceptible to the erosion of natural resources, on the other. Soils are a key resource in Senegal. For over 60% of the population depend directly or indirectly on them for a living. Most of agricultural production of Senegal lies within the drought-prone Sahel region, with irregular rainfall and poor soils. Agriculture employs in Senegal over
70% of the labour force [1]. Senegal’s primary cash crops are groundnuts, cotton, and horticulture. Groundnuts are grown in the central region, particularly in the Peanut Basin (see figure 1), and horticulture is concentrated in coastal regions (ICRISAT, 2005). Sorghum and millet are grown in the northern and central regions, and rice is grown in the southern Casamance and Senegal River Valley. Maize and cowpeas are common subsistence crops. SUNEOR, Senegal’s groundnut refinery named before as SONACOS, counts among the largest industrial companies in the country. It controls the majority of domestic groundnut production. SUNEOR also processed groundnut oil and cakes, and supplies 70% of the domestic market for groundnut oil [2]. Due to all this, there are high moments of farming activities that should match important times for peanut crop and during their transport to the refineries. This entails a movement of people and communication among them. In order to contribute to the better understanding of the effects of this highly specialized economy on the population we have analyzed data from mobile phones calling activity relating to groundnut agricultural activity through the Normalized Difference of Vegetation Index (NDVI), an indicator used to assess whether an area contains photosynthetic active vegetation.

This article is structured as follows: first, we give a description of the experimental data. In the next section we explain our methodology. Next we show our main findings and finally we discuss them.

2 Description of the data

Call Detail Records are metadata produced by phone interactions (either calls or SMS). Orange and Sonatel Senegal have made available three mobile phone datasets extracted from CDRs in the framework of the D4D (Data for Development) challenge. The datasets are based on phone calls and SMS exchanged between 9 million users during the year 2013. All this data have been properly anonymized before being handled to researchers.

The first dataset contains the number and the duration of all calls between any pair of antennas on hourly basis. This database is composed of 1666 uniquely identified antennas whose geographical location is known. We have visualized all the antennas on figure 1. The second dataset contains fine-grained mobility data of 300000 individuals during lapses of two weeks, including bandicoot behavioral indicators. The information provided regards the random identifier, the time-stamp when the call was made and the antenna from which it was made. The third dataset consists of one year coarse-grain mobility data of about 150000 randomly sampled users, includ-
ing the random identifier of the user who made the call, the time-stamp and the arrondissement from where the call was made.

Figure 1: Map of Senegal with the Groundnut Basin highlighted in brown and the 1666 antennas plotted as dots.

In order to identify growing areas in Senegal, we have complemented the CDR datasets with georeferenced images of the country. We have obtained this images from the MODIS database available at The Land Processes Distributed Active Archive Center website. MODIS, which stands for Moderate-resolution Imaging Spectroradiometer, is a scientific instrument mounted on Terra and Aqua satellites that measures 36 spectral bands with high spatial resolution. The dataset contains images for the whole 2013 with a sampling frequency of 8 days and a spatial resolution of 500m.

3 Methods

The Normalized Difference Vegetation Index (NDVI) is a simple indicator that measures whether a region contains live green vegetation or not. It is defined by the following formula:

\begin{equation}
\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}
\end{equation}
Figure 2: Annual evolution of the NDVI in Senegal.

\[
\text{NDVI} = \frac{\text{NIR} - \text{VIS}}{\text{NIR} + \text{VIS}}
\]  

where VIS and NIR stand for the spectral reflectance measurements acquired in the visible (red) and near-infrared regions, respectively. For plants, the NDVI typically ranges between 0.0 and 1.0 although by design it can vary between -1.0 and +1.0.

Hence, to identify growing areas we have generated the NDVI time-series for the whole country. The evolution of the NDVI for the whole country is shown on figure 1. Next, we have associated a NDVI time series to each antenna by taking a three-by-three pixels lattice centered at the antenna coordinates and computing the average NDVI for those nine pixels. Each pixel corresponds to a 500x500 m square of land and the sampling frequency of the data is 8 days. In order to improve the signal-noise ratio, we have worked with the accumulated NDVI.
For growing regions the NDVI should be incremented during the growing season and decay after harvesting. In the accumulated NDVI time-series this pattern is reflected as an increase in the slope during the growing period (figure 3).

![Figure 3: Increase of the NDVI during the growing season, which has been highlighted with a light brown.](image)

Next, we have constructed the time-series of outgoing calls for each antenna. To build these time series we have used the first dataset described in section 2. On the top panel of figure 3 we show an example of an outgoing calls time series.

In order to understand the impact that the agricultural activity has the population of Senegal, we have compared the NDVI and calls time series to find possible correlations. We found that a significant fraction of antennas show peaks of activity corresponding to the beginning and at the end of the growing season.

Hence, we first classified the antennas according to the behavior of their calling activity and NDVI time-series. We took the following considerations into account. To filter the NDVI data associated to cities or forests, every
Figure 4: Activity and associated NDVI time-series of antenna 683. The growing season has been highlighted with a light brown background. Time-series with a yearly accumulated NDVI bigger than 25 or smaller than 10.5 are labeled respectively as forest type or city type. We then perform a linear regression for the pre-growing season of the accumulated NDVI time-series (from June the 1st to May 25th). If the $R^2 < 0.99$ we label that time-series as noisy. If the time-series is not noisy, we perform a second linear regression for the growing season period (from the beginning of August to the end of October). Next, we compute the ratio of the second obtained slope $m_2$ with respect to the first one $m_1$. If $m_2/m_1$ is bigger than 1.65 the time-series is tentatively labeled as crop field type. A final filtering is done by computing the first inflection point as the intersection of the first and second regressions. If it is located earlier than expected, the NDVI is classified as forest type. In figure 5 we present some examples of the different types of time-series.

In a second step we classify each antenna according to its calling pattern. In this case we have used a peak-searching algorithm to search for those that show two peaks during the growing season, one at the beginning and another
the end of the period. This algorithm assigns a score to each point of the

time-series. To do so, \( k \) values are taken from each side of the point (without

including the point itself) and the mean \( m \) and the standard deviation \( s \) of

these \( 2k \) values are computed. The score associated to the point \( x_i \) is

\[
\frac{x_i - m}{s}
\]

The points whose scores are beneath a threshold of 2.5 are filtered out.

Thus at the end of the process we classify each antenna as one of the

following three types (see Figure 6):

1) Antennas that present a two-peak structure in their calling activity
time-series and a crop field type NDVI (colored in red).

2) Antennas that show two activity peaks but no crop field NDVI (colored in blue).

3) Antennas presenting a crop field NDVI type but do not present a sig-
nificant peak during the growing season (colored in green).

4) Antennas that do not meet neither of the two conditions are not shown.

Figure 5: Examples of the different types of accumulated NDVI time-series and
their associated calling activity time-series. The highlighted period coincides with
the growing season. Figures (a) through (c) correspond to antennas located in cities
(Dakar, Touba and Kaolack); figures (d) through (f) belong to antennas placed on a
forest area and (g) through (i) are associated to antennas installed near crop fields.
4 Results

We have visualized the results of our classification overlaid on a map of Senegal on Figure 6. On this figure each antenna has been colored according to its classification. Over 7% of the antennas present both an increment of activity during the growing season and an NDVI associated to crop fields. Of this 7%, about 88% are located within the groundnut basin, where the majority of the antennas having a crop field-type NDVI are concentrated. These antennas form the 32% of the total antennas of the groundnut basin.

Figure 6: Map of Senegal showing the results of the classification of the antennas. The groundnut basin is highlighted with a light brown. Those antennas that present a two-peak structure in their calling activity time-series and a crop field type NDVI are colored red; those that show two activity peaks but no crop field NDVI associated are colored blue; the ones presenting crop field NDVI type but no special calling activity are green and antennas that do not meet neither of the two conditions are not shown.

In the light of our results there are some regions that present a calling activity that can be associated with agricultural activities in spite of not having a crop field-type NDVI associated. This can be explained by the
A large proportion of the labor force working in agriculture and, specially, in groundnuts: it is expected that not only rural but also urban population participate in one way or another in the agricultural process. An example of this behavior can be found on figure 5 (c), which corresponds to an antenna placed in the city of Kaolack, the main groundnut trading and processing center. Thus, the agricultural economic activity of the country is not limited to the growing regions, but required of a high degree of synchronization among many spatially distributed areas. The export of agricultural goods requires transportation, storage and importing fertilizers. In fact, Senegal imports fertilizers for value of 22 million dollars.

To further uncover the impact that the growing season has on Senegal's daily activity we have performed a community structure analysis to the antenna to antenna calling network. We used an algorithm based on modularity that was proposed by Blondel et al. [3]. On this network antennas are connected via directed links that quantify the flow of calls that go from one to another. We have performed the community structure analysis for three time periods. The first one corresponds to the pre-growing season. The second one to a 30 day time period centered around the first peak of activity associated to the beginning of the growing season. Finally, the third one corresponds to a 30 day time period centered around the second peak of activity corresponding to the end of the growing season. The community structure of the country has been illustrated on Figure 7 where each community is plotted in a different color. The partition of the country into communities exhibits some differences between the pre-growing season growing seasons, while showing a common structure for the two peaks of activity. The groundnut growing area is mainly formed by a single large

---

Figure 7: Community structure of Senegal for the pre-growing period (left) and during the growing period (right).
community in the pre-growing season. However, during the two peaks it brakes into three well defined communities showing an ordered structure. A similar phenomenon occurs in Dakar, where during the pre-growing period Dakar is divided into two communities, while a third community emerges at the beginning and end of the growing season (see Figure 8). Note that despite Dakar is a non crop growing area, it presents many antennas with the two peak pattern. This fact reflects that the economic activity around the growing and export of groundnut has an impact on the city. Thus, it is probable that the emergence of a third community on Dakar, corresponding to the previously discussed peaks of activity, results from the groundnut economic activity.

Figure 8: A, Community structure of Dakar for the pre-growing period. B, Community structure of Dakar during the growing period (right).

5 Conclusions and future work

In this work we have analyzed the impact that agricultural activities, such as the growing of groundnut, have on Senegal. To this end we have analyzed the NDVI time series of the whole of Senegal and spotted the regions where groundnut is grown. Next, we characterized the calling activity of each region of the country by analyzing the number of incoming and outgoing calls per day in each antenna. We found that a significant fraction of antennas exhibit two well defined peaks of activity corresponding with the beginning and end of the growing season. Almost all antennas located on regions identified as growing regions present this pattern. However, other antennas, located in non growing regions, such as Dakar, also present the two peaks pattern. Hence, in a preliminary inspection of our results we think
that there is a synchronization between growing regions and key points in cities that emerges from the agricultural activity of the country. Moreover, this idea is supported by the differences in the community structure of the country during the pre-growing and growing seasons. Thus, by further analyzing the patterns on the calling network among antennas associated to agricultural activity we can acquire knowledge of the economic activity the country. The next steps of the research involve further statistically testing that these patterns are caused by the agricultural activity. Moreover, it will be very meaningful to analyze the trajectories networks searching for patterns that emerge at the beginning and end of the growing process. This information would be very useful to identify the routes through which the groundnuts are transported for storage and export. Thus, this could highlight possible new highways connecting the rural and urban areas that would benefit the economy of the country.

References


Using Mobile Phone Data for Rural Electrification Planning in Developing Countries

Eduardo Alejandro Martinez-Cesena, Pierluigi Mancarella, Mamadou Ndiaye, and Markus Schläpfer

Abstract—Detailed knowledge of the energy needs at relatively high spatial and temporal resolution is crucial for the electricity infrastructure planning of a region. However, such information is typically limited by the scarcity of data on human activities, in particular in developing countries where electrification of rural areas is sought. The analysis of society-wide mobile phone records has recently proven to offer unprecedented insights into the spatio-temporal distribution of people, but this information has never been used to support electrification planning strategies anywhere and for rural areas in developing countries in particular. The aim of this project is the assessment of the contribution of mobile phone data for the development of bottom-up energy demand models, in order to enhance energy planning studies and existing electrification practices. More specifically, this work introduces a framework that combines mobile phone data analysis, socioeconomic and geo-referenced data analysis, and state-of-the-art energy infrastructure engineering techniques to assess the techno-economic feasibility of different centralized and decentralized electrification options for rural areas in a developing country. Specific electrification options considered include extensions of the existing medium voltage (MV) grid, diesel engine-based community-level Microgrids, and individual household-level solar photovoltaic (PV) systems. The framework and relevant methodology are demonstrated throughout the paper using the case of Senegal and the mobile phone data made available for the ‘4D-Senegal’ innovation challenge. The results are extremely encouraging and highlight the potential of mobile phone data to support more efficient and economically attractive electrification plans.

Index Terms — Electrification, human dynamics, mobile phone data, cellular networks, Microgrids, Photovoltaics

I. INTRODUCTION

Detailed knowledge of the energy needs at relatively high spatial and temporal resolution is crucial for the adequate energy infrastructure planning of a country. This is particularly relevant to the electrification of developing regions where new infrastructure needs to be built to foster socio-economic growth. However, such information is typically limited by the scarcity of comprehensive data on human activities. In this respect, during recent years the increasing availability of mobile phone data has proven to provide unprecedented insights into the mobility patterns of people and the distribution of the population in space and time [1]–[3]. Not surprisingly, this type of data has thus been hinted as promising for the design and operation of ‘smart’ infrastructures [4] and energy systems [5]. However, to the authors’ knowledge no quantitative study has so far investigated the potential applicability of mobile phone data for energy infrastructure planning, particularly in developing countries where cellular network data can actually be much more advanced than energy consumption data. For instance, taking Senegal as a typical case of a developing country, during the last decade its mobile phone usage has increased dramatically from 1.7 million subscribers in 2005 to 13.1 million in 2013, thus covering about 95% of the countries 14 million inhabitants [6]. In stark contrast to this upsurge in mobile communication, about half of the total population still has no access to electricity, and the electrification rate in rural areas is even as low as 28% [7]. Therefore, there is a clear potential to use mobile phone data for predicting a region’s energy demand and supporting its electrification process. In particular, compared to current approaches for electricity planning in developing countries that use, for instance, satellite imagery, mobile phone data can provide substantially more accurate information on the spatio-temporal activity centers [2], which could be combined with socioeconomic, geo-referenced or climate data for electrification planning purposes in both urban and rural areas.

On these premises, the aim of this work is the assessment of the potential use of mobile phone data to support rural electrification planning in developing countries. Specific objectives include the assessment of i) the suitability of mobile phone data as a proxy for current and future electricity needs and whether ii) this information can lead to more economical and more efficient electrification options. To that end, we develop a framework that brings together in an innovative way mobile phone data analysis, socio-economic and geo-referenced data analysis, and state-of-the-art energy infrastructure engineering techniques to quantify the techno-economic feasibility of different centralized and decentralized electrification options in developing countries. The electrification options considered here include extensions of the existing Medium Voltage (MV) grid (“centralized”
option), development of diesel engine-based community-level Microgrids, and installation of individual dwelling-level solar photovoltaic (PV) systems (“decentralized” options). The proposed methodology is clearly demonstrated throughout the report by taking the case of Senegal as representative of developing countries.

The report is organized as follows. The next section provides an overview of the different available ‘D4D-Senegal’ datasets, as well as the electricity context of Senegal, which provides the baseline for this work. Section III presents a high level description of the electrification planning methodology based on mobile phone data proposed in this work. The methodology involves the assessment of i) the energy requirements of Senegal, ii) the correlation between mobile phone data and electricity needs, iii) the population migration towards electrified areas and iv) the electrification potential. These steps are further detailed in Sections IV – VII. Section VIII describes possible follow-up studies that could be derived from this work and Section IX concludes.

II. OVERVIEW OF THE DATASETS

A. Mobile phone data

The anonymized mobile phone communication data used in this project was collected in Senegal between January 1, 2013 and December 31, 2013. These data were made available by the telecommunications provider Sonatel and the Orange Group within the framework of the D4D–Senegal challenge [8]. In 2013, Sonatel had 7.4 million mobile phone subscribers in Senegal, corresponding to a market share of about 60%. The data are organized into three sets:

- **Dataset 1** contains the hourly voice and text traffic between each pair of mobile phone towers (total call duration, number of calls and total number of text messages). The geographic location of the 1,666 mobile phone towers is depicted in Fig. 1, which also shows the topology of the electricity transmission and distribution networks. Note that a large number of towers lie outside the reach of the power grid; most mobile phone towers without grid access to electricity are in fact powered by diesel generators [9].

- **Dataset 2** contains the fine-grained mobilities of about 300,000 randomly sampled and anonymized users during each consecutive period of two weeks. For each time period, a new sample of about 300,000 users was selected and their trajectories recorded at the mobile phone tower level.

- **Dataset 3** contains the coarse-grained trajectories for about 150,000 randomly sampled and anonymized users during the entire year at the spatial level of Senegal’s 123 arrondissements (administrative subdivisions). This dataset is not considered in the present study due to its limitations in the spatial resolution.

A more detailed description of the three datasets is provided in [8].

B. Electricity consumption and infrastructure data

For the purpose of this project, the national electric utility in Senegal – the “Société Nationale d’Électricité” (Senelec) – kindly provided us with the hourly electricity consumption data for the entire year of 2013, aggregated at the national level (i.e., 8,760 data points) [10]. The overall yearly electricity consumption was 2,96 TWh. About 80% of the electricity was generated by diesel power plants and the remainder by gas-, steam- and hydro power plants, whereas most of these generators are owned and operated by Senelec [11]. The high-voltage (HV) transmission network consists of 90 kV national and 225 kV supranational lines totaling about 13,000 km in length (see Fig. 1). The 30 kV MV distribution network brings electricity from the transmission network to the consumption centers [13]. Both transmission and distribution networks are again managed by Senelec.

III. FRAMEWORK AND METHODOLOGY OVERVIEW

As mentioned above, the objective of this work is to build a framework and provide a quantitative assessment methodology for the use of mobile phone data to facilitate rural electrification planning in developing countries in general, and Senegal in particular. Mobile phone use and corresponding mobile phone charging requirements could, in principle, be extrapolated from the mobile phone data. This information could provide key insights into electrification planning of Senegal as mobile phone charging represents, along with lighting, a major energy demand in the country [6], [14]. In addition, mobile phone data could also be used as a proxy for current and future energy needs in a given area and even to estimate the spatio-temporal electricity profiles. This is due to the potential of mobile phone information (particularly if several years’ worth of information becomes available) to facilitate the mapping of human activity and migration within the country (e.g., people are more likely to migrate to areas with access to electricity, health and

![Fig. 1. Existing electricity infrastructure in Senegal and location of the mobile phone towers. The transmission and distribution network as well as the location of the power stations are adopted from [12].](image-url)
education, thus further increasing energy demands). Both data on human activity and migration can provide an accurate estimation of electricity needs and facilitate more sustainable electrification plans, particularly when combined with other data sources used in state-of-the-art electrification planning practices [15].

In the light of the above, the proposed framework and assessment methodology comprises the following four steps:

1) Assessment of the energy requirements and consumption characteristics of Senegal;
2) Evaluation of the use of mobile phone data as a proxy for current and future electricity needs via correlation analyses;
3) Estimation of potential future migration of population from non-electrified to electrified areas; and
4) Quantification of centralized and decentralized electrification options considering mobile phone data combined with socio-economic and geo-referenced information.

The assessment of energy requirements and consumption characteristics of Senegal is meant to provide context on the expected energy needs of the mobile phone users whose activity is recorded by the different mobile phone towers. This analysis is supported by socio-economic and geo-referenced information extracted from [15] detailing the population density and average distance between households in each area in Senegal, as well as the access to electricity, health, education, markets and so on. This information is used to further classify the mobile phone data compiled from the different mobile phone towers (i.e., Dataset 1 and Dataset 2), allowing the assessment of the correlation between the human activity and the aggregated electricity profile under different socio-economic conditions.

This study is expected to highlight the conditions that make the mobile phone datasets an accurate proxy for current and future electricity needs and profiles. Afterwards, potential migration trends towards electrified areas within the country are assessed based on the fine-grained mobility data (i.e., Dataset 2). Again, this information can provide insights into the future energy needs of an area after it is electrified, thus potentially improving electrification decisions. Finally, all this information derived from the mobile phone data is combined with geo-referenced information to build different state-of-the-art options for electrification, namely, MV grid extensions, development of diesel engine-based (community) Microgrids, and development of dwelling-level PV systems (see [15] for an example of the assessment of electrification options for Senegal based only on geo-referenced information). A detailed description of each of the methodological steps and relevant studies is provided in the next sections.

### IV. Energy Requirements and Consumption Characteristics of Senegal

The energy requirements and consumption characteristics currently available for Senegal are derived from the countrywide electricity demand profile, the solar radiation and temperatures in different areas, and the size and location of villages and their access to electricity, health, and educational services. The solar radiation and temperature profiles (8,760 hourly data points for 2013) were obtained from the SoDa solar energy services database [16]. A thorough description of the different types of villages in Senegal, their location, and their access to electricity, education and health services were obtained from a previous electrification study in Senegal prepared for the World Bank [15]. Table I lists typical services considered for villages of different sizes.

<table>
<thead>
<tr>
<th>Village size (population)</th>
<th>Hospitals</th>
<th>Schools</th>
<th>Markets</th>
<th>Public Lighting points</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1,000</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5,000</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>10,000</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>99</td>
</tr>
</tbody>
</table>

Fig. 2. Level of electrification in the Voronoi polygons defined by the location of the mobile phone towers. The locations of the settlements (cities, towns, villages) are adopted from OpenStreetMap.
V. CORRELATION STUDIES

In this section, the potential use of mobile phone data, specifically Dataset 1 and Dataset 2, as a proxy for electricity needs is assessed in terms of the correlation between the mobile phone activity at each mobile phone tower and the countrywide aggregated electricity load curve. To that end, we first measured for each mobile phone tower $i$ the linear correlation in terms of the Pearson coefficient $r_i$ with,

$$r_i = \frac{\sum_i (A_i(t) - \langle A_i \rangle) (D(t) - \langle D \rangle)}{\sqrt{\sum_i (A_i(t) - \langle A_i \rangle)^2 \sum_i (D(t) - \langle D \rangle)^2}}$$  \hspace{1cm} (1)$$

where $A_i(t)$ is the total mobile phone activity during hour $t$ (i.e., number or duration of calls, or number of text messages), $D(t)$ is the countrywide electricity consumption during the same time interval, and $\langle \rangle$ denotes here the average value over the entire year (8,760 hours). Fig. 3 shows the histogram of the Pearson coefficients, indicating a strong correlation between the hourly mobile phone activity from Dataset 1 and Dataset 2, and the electricity consumption for almost all mobile phone tower areas. The average values over all mobile phone towers for the call duration, number of calls and number of users are $\langle r^c \rangle = 0.76$, $\langle r^d \rangle = 0.8$ and $\langle r^n \rangle = 0.81$, respectively. This result demonstrates that mobile phone data are, in general, a reliable proxy for electricity consumption (and therefore infrastructure needs) in Senegal, even when considering all available data regardless of the characteristics of the villages in the mobile phone tower area (i.e., including non-electrified villages). The correlation study was repeated for areas with different amounts of mobile phone users, population and penetration levels of electricity, health and education. The penetration levels were calculated based on the expected percentage of people living in the mobile phone tower area with access to a given service (e.g., a 50% electrification penetration implies that only half of the individuals living within the mobile phone tower area have access to electricity). This further analysis indicates that low correlations can typically be attributed to the lack of mobile phone data or to areas with either very high or very low electrification levels. More specifically, in 63% of the areas with low correlation (i.e., $r_i < 0.4$), there were only 8 mobile phone users or less. This result is not surprising as mobile phone data cannot be used as a reliable proxy of the electricity needs whenever little or no data are available. Regarding the effects of different electrification levels, electrification was either “too low” (20% or less) or “too high” (80% or more) in 56% of the mobile phone tower areas with low correlation. This result could also be expected considering that the aggregated electricity profile may not be representative for small non-electrified villages, or for large and highly electrified villages. In the latter, for instance, energy-intensive industrial processes that cannot be inferred from human activity profiles may be more widespread.

Complementary to Fig. 3, Fig. 4 presents examples of the comparison of the mobile phone data and the electricity consumption data in terms of time series profiles for a mobile phone tower in a typical urban area (Dakar) and rural area (Fayil). More specifically, while mobile phone data are actual
information from the relevant cellular tower, the electricity profiles are scaled down from the national profile and in proportion to the amount of mobile phone users in the corresponding Voronoi polygon. The visual results confirm the adequacy of mobile phone data as a proxy for electricity needs, which in Senegal seems indeed mostly dictated by human activity (e.g., lighting, mobile phone charging). The figure also shows that good approximations of the electricity profile could be made with either the number of calls, call duration or number of users extracted from Dataset 1 and Dataset 2. Thus, good estimations of the electricity needs could still be made even if the information in the mobile phone datasets were limited.

Overall, the results of the correlation study suggest that, as long as sufficient mobile phone users are available, it is reasonable to use mobile phone data as a proxy for electricity needs under most conditions. Furthermore, this application of mobile phone data seems to be particularly accurate for the average electrified village. This is especially important for electrification planning as, after being electrified, villages are likely to resemble the average electrified village. Accordingly, the results of the correlation study highlight that the use of mobile phone data and a scaled version of the aggregated electricity profile are reasonable for electrification planning.

VI. MIGRATION STUDIES

In the long-term, the electricity needs of a village are dependent on the expected population growth and migration in the area. Traditionally, population growth and migration have been estimated via census data, which can be enhanced using satellite imagery. Nevertheless, emerging literature suggests that mobile phone data can be used to increase the accuracy of existing population mapping techniques [2].

Several years’ worth of mobile phone data beyond Dataset 2 would be needed to estimate population growth and migration in a given mobile phone tower area with a reasonable level of accuracy. However, considering that the main aim of this work is to illustrate the applicability of mobile phone data to enhance current electrification practices, it is assumed here that the available information in Dataset 2 suffices for a first estimation of migration to a mobile phone tower area (population growth is taken as 2.3% [15]). Future studies could improve the accuracy of the population mapping (including population growth estimations) should the required information become available.

In order to estimate the potential number of migrants attracted to different villages, the mobility patterns of mobile phone users were calculated based on Dataset 2. More precisely, for each mobile phone user we first determined the home location according to [17] and then identified all Voronoi polygons visited throughout the year. Subsequently, we aggregated the number of users that visited a given polygon, providing us with the total number of trips to that area. Finally, we binned the number of trips by the distance of the visitors’ home location and normalized it by the total number of trips. We applied the same procedure to determine the number of visits originating from a given area. Fig. 5 shows a sample of the results based on electrified and non-electrified areas. Our migration study shows that, as expected, the average amount of travels to an area decreases rapidly as distance increases. More importantly, people seem to travel longer distances to electrified areas than to those without access to electricity (qualitatively similar results apply to access to health and education). The relative difference between the number of people coming from and going to an area, averaged over all possible distances, is taken as the expected migration to an area. Accordingly, migration to electrified areas is assumed here to be between 7% and 13%. This range suggests that the attractiveness of a village is expected to increase if it offers access to electricity.

VII. ELECTRIFICATION STUDIES

In this Section, the energy needs and profiles derived from the mobile phone data, combined with the geo-referenced information extracted from [15], are used for the assessment of three possible electrification options, namely, MV grid extensions, Low Voltage (LV) community-level Microgrids powered by diesel generators, and dwelling-level PV generators. The different options were assessed based on their Net Present Costs (NPC), considering a planning horizon of 10 years and a discount rate of 10%, as recommended by the World Bank [15].

A. Grid extension

Traditional MV grid extensions involve installing additional MV lines that interconnect the consumption centers with the existing grid, as well as transformers and LV grids to supply the rural villages. This alternative can be particularly attractive to supply large villages near the existing grid, but it may become less economically attractive for smaller villages far from the grid.

The cost of the grid extension (GE) is calculated as the NPC denoted by (2), (3) and (4). The inputs for the relations are the specific characteristics of the village to be electrified (i.e., electricity consumption and peak) taken from the scaled electricity profile (adjusted for migration), the geo-referenced information extracted from [15] and the parameters given in Table II.
existing grid. The NPC of the Microgrid generation units (diesel generators in this case) and a LV village, it is possible to install a group of dwellings located far away from the existing grid. The NPC of the Microgrid (MG) electrification alternative \((NPC_{MG})\) is calculated with (5), (6), (7) and (8). Similarly to the previous case, the inputs for the equations come from the characteristics of the village to be electrified (i.e., electricity consumption and peak), the geo-referenced information from [15] and Table III.

\[
NPC_{GE} = I_{GE} + \sum_{i=0}^{T} \frac{C_{GE}(D_i + L_{MV,i})}{(1 + d)^i} + M_{GE}
\]

\[
M_{GE} = M_{MV} + M_{TR} + M_{LV}
\]

\[
I_{GE} = I_{MV} + I_{TR} + I_{LV}
\]

where \(I_{GE}\) is the total investment cost in CFA\(^3\), \(C_{GE}\) is the electricity generation cost in CFA/kWh, \(D_i\) is the annual demand of the village in kWh/year, \(L_{MV,i}\) represents annual power losses in kWh/year, \(d\) is the discount rate, \(i\) denotes a year, \(T\) represents the planning horizon (years), \(M_{GE}\), \(M_{MV}\), \(M_{TR}\), and \(M_{LV}\) denote the annual operation and maintenance (O&M) costs in CFA/year associated with the total investment cost, as well as with the MV line, transformer, and LV line, respectively. The parameters \(I_{MV}\), \(I_{TR}\) and \(I_{LV}\) denote the investment costs in CFA associated with the MV line, transformer, and LV lines, respectively. It is important to note that the investment costs are a function of the length of cable and capacity of the transformer to be installed.

The distance between the existing MV network and the villages was estimated based on the geo-referenced information from [15] and using an iterative procedure to find the minimum length between villages and existing network connection points. The length of the LV network was calculated by assuming that the mean distance between households varies between 8 m in villages with more than 5000 individuals to 30 m for villages with less than 500 people [15]. This distance is assumed to increase by up to 50% for dwellings located far from the center of the village.

**B. Diesel engine-based Microgrid**

Instead of extending the MV grid to the location of a village, it is possible to install a group of distributed generation units (diesel generators in this case) and a LV Microgrid to supply a village located far away from the existing grid. The NPC of the Microgrid (MG) electrification

\[
NPC_{MG} = I_{MG} - S_G + \sum_{i=0}^{T} \frac{C_{MG}(D_i + L_{MG,i}) + M_{MG}}{(1 + d)^i}
\]

\[
M_{MG} = M_G + M_{LV}
\]

\[
I_{MG} = I_{LV} + \sum_{i=1}^{\text{floor}(T/L_{G})} \frac{I_G}{(1 + d)^{iL_{G}}}
\]

\[
S_G = \frac{I_G}{(1 + d)^{\text{floor}(T/L_{G})L_{G}}} \cdot \frac{T - \text{floor}(T/L_{G})}{L_{G}}
\]

where \(I_{MG}\) is the total investment cost (in CFA), \(S_G\) is the salvage value of the generator, \(C_{MG}\) is the generation cost in CFA/kWh, \(L_{MG,i}\) represents the annual power losses (in CFA/year) and \(L_{G}\) represents the lifetime of the generator (in years). The parameters \(M_{MG}\) and \(M_G\) denote the annual O&M costs in CFA/year associated with the total investment and generator, respectively, and \(I_G\) represents investments in generators. It is important to highlight that several investments in generators may be needed throughout the planning horizon.

**C. PV systems**

Off-grid PV systems supplying individual households tend to be less economically attractive than the other technologies under normal conditions. However, due to their modularity, independent PV systems can be installed in each household without the need of an LV network. Thus, this option can become economically attractive for low population villages where dwellings are located far apart.

The NPC of a PV system \((NPC_{PV})\) is calculated with (9), (10), (11) and (12). The inputs for these equations were taken from the characteristics of the villages to be electrified, the geo-referenced information from [15] and Table IV.

\(^{3}\) 1 CFA = 0.0012 GBP = 0.002 US$
TABLE IV
OVERVIEW OF THE ECONOMIC AND TECHNICAL PARAMETERS
CONSIDERED FOR A PV SYSTEM ELECTRIFICATION ALTERNATIVE [15].

<table>
<thead>
<tr>
<th>PV Module</th>
<th>Capacity</th>
<th>Cost</th>
<th>O&amp;M costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>20W – 150W</td>
<td>88k CFA – 660k CFA</td>
<td>1%</td>
</tr>
<tr>
<td>Batteries</td>
<td>Capacity</td>
<td>14Ah – 38Ah</td>
<td>1%</td>
</tr>
<tr>
<td>Cost</td>
<td>40k CFA – 70k CFA</td>
<td>1%</td>
<td></td>
</tr>
<tr>
<td>Converter</td>
<td>Costs</td>
<td>28k CFA</td>
<td>1%</td>
</tr>
</tbody>
</table>

\[
NPC_{PV} = I_{PV} - S_B + \sum_{i=0}^{T} \frac{M_{PV}}{(1 + d)^i}
\]

\[
M_{PV} = M_{MO} + M_{CV} + M_B
\]

\[
I_{PV} = I_{MO} + I_{CV} + \sum_{i=0}^{\text{floor}(t/T_B)} \frac{I_B}{(1 + d)^{i + 1}}
\]

\[
S_B = \frac{I_B}{(1 + d)^{\text{floor}(t/T_B)}} \cdot \frac{T - \text{floor}(T / LT_B)}{LT_B}
\]

\[
i_{eff,i} = \frac{i_{rated}}{CAP(i)} \cdot \frac{N_{rated}}{N(DOD)} \cdot i_t
\]

where \( AhT \) is the expected throughput of a battery (in A), \( i_t \) is the discharge current (in A), and \( i_{eff,t} \) is the effective discharge current (in A). The parameters \( DOD_{rated} \) and \( DOD_t \) (in percent) are the depth of discharge under rated conditions and at a given time, respectively. \( CAP_{rated} \) and \( CAP(i_t) \) are the capacity of the battery subject to rated conditions and a given discharge current, respectively, \( N_{rated} \) and \( N(DOD_t) \) are the amount of operational cycles subject to rated conditions and a given depth of discharge, respectively, and the subscript \( t \) denotes a given hourly time step within a year.

D. Electrification option results and discussion

The three different options and relevant technologies discussed above were assessed for the electrification of the villages in Senegal, classified based on their population and access to services such as education, health and markets as extracted from [15]. Each electrification option was assessed based on parametric scenarios and taking the cheapest alternative as the recommended technology. In addition, the option to install small PV systems in small villages to supply only demand from mobile phones charging and lighting (which are the current main energy needs, as discussed in [6]) was considered too.

The parametric scenarios were formulated assuming different levels of electrification (from 20% to 80%) and population growth due to migration (from 0% to 13%), as suggested by our migration studies. The electrification levels represent potential electrification targets, such as the current target in Senegal to achieve 60% electrification of rural areas [15]. It is considered that electrification levels for each newly electrified village will be the same. For instance, if a 50% electrification level is considered, only 50% of the households in every village will be electrified, which would correspond to the dwellings in the densest area of each village and nearest to the center of the settlement. The results highlight that each of the electrification options can outperform the others under a specific set of conditions, as shown in Fig. 6.

Fig. 6(a) shows the costs per household associated with individual PV and diesel engine-based Microgrids, subject to different village population sizes and electrification levels. The PV systems are only deemed more attractive than the Microgrid for small villages where houses may be dispersed and the installation of a LV network would be too expensive, and for low electrification levels where the installation of a diesel engine may not be justifiable. The use of small PV systems is the only economically viable option for the electrification of small villages where mobile phone charging and lighting may be the main electrical load. This result is consistent with existing literature [15]. It is important to note that the results regarding the PV system present non-monotonic, but well defined increasing trends due to the nonlinearity and integer nature of the arrays of PV modules and batteries (i.e., it is not possible to install a fraction of a PV module or battery).
Fig. 6. (a) Costs of PV systems (PV) and diesel-based Microgrids (MG) for different electrification levels. The curves are averages over three migration levels (0%, 7%, 13%). (b) Maximum MV grid extensions that are economically competitive with PV and MG for different electrification levels. The curves are averages over different migration levels as in (a). (c) Costs of PV and MG for different migration levels. The curves are averages over different electrification levels (0%, 10%, 90%, 100%). (d) Maximum economically competitive MV grid extensions for different migration levels, averaged as in (c).

Fig. 6(b) shows the longest possible MV grid extension that would still be economically competitive with other alternatives (i.e., PV systems and Microgrids). This is calculated by first estimating the costs corresponding to a Microgrid and a PV system in the specific village, and then estimating the maximum potential MV grid extension (using (2) to (4)) that would be cheaper than the other technologies. The results suggest that it is more attractive to extend the grid when this enables high electrification levels for large villages, while it is attractive to use this option to electrify small villages when they are located in the proximity of the existing grid. It is important to note that investments in upstream generation and grid upgrades that might be needed to support MV line extensions are not considered in this study. Such costs would make line extensions – particularly when planning many line extension or long line extensions – for big villages less attractive, thus a decrease in the maximum recommended line extension (particularly after about 20 km) would be expected in the results.

Fig. 6(c) and Fig. 6(d) show the effects that migration could have on the preferred electrification option. It can be observed that the effects of migration on electrification planning are modest compared with those associated with different electrification levels. However, migration can still play a big role on the identification of small villages where PV systems are deemed the only feasible alternative. This is due to the potential of migration to increase (or decrease) the population of a village. It is worth noting that these results only highlight the potential impacts of migration in a parametric way and that additional research would be needed to better quantify the actual migration potential in Senegal.

E. Recommendations for the electrification of Senegal

Finally, a graphical representation of the results presented in Fig. 6 and applied to all the villages in the geo-referenced dataset extracted from [15] is presented in Fig. 7. As discussed above, for each relevant option the length of required MV extension was calculated based on the distance between the existing grid and the village, whereas the length of the LV networks were estimated based on the distance between households in each village, electrification level and growth due to migration. The electrification levels and population growth in combination with peak demand and total energy consumption, estimated from the mobile phone data, were used when sizing the PV arrays, generators and transformers.

Fig. 7 was calculated assuming 60% electrification and 13% migration and shows the zones where it would be more economically attractive either to extend the MV grid, or to install a diesel engine-based Microgrid or individual PV systems. In the green (dark) zones, there are two potential technologies to be deployed, namely Microgrids and PV systems. The latter are recommended for the smallest villages or for areas where only lighting and mobile charging infrastructure is to be electrified. Note that similar figures could be readily produced for alternative electrification and migration scenarios.

VIII. OPENINGS TO FURTHER WORK

If more detailed electricity and mobile phone data were available for longer observation periods, further work could be done to improve the analysis carried out here. For instance, a more detailed assessment of the upstream costs in the case of the MV extension option could be performed; also, a more in-depth analysis of possible changes in the mobile phone activity profiles due to the electrification of a settlement could...
be carried out, so as to improve the energy planning option assessment; finally, more electrification options could be considered, such as those based on wind turbines and/or fuel cells.

Further, given the novelty of the topic, there is significant research that could still be carried out for electrification planning and for other infrastructure-related applications based on mobile phone data. For instance, additional detailed mobile phone datasets covering several years would provide better proxies for electricity needs and population migration, particularly when combined with corresponding electricity consumption profiles for different areas. Moreover, the hourly activity curves derived from mobile phone data could be compared with environmental data such as hourly solar radiation or wind speeds [19] to quantify in more detail the potential for PV or wind power in a given area and estimate the need for energy storage.

More advanced studies could also be carried out in the context of developed countries. For instance, the dynamic population mapping derived from mobile phone data could be used for assessing the number of people that would be affected by a potential power blackout. Such ‘risk-maps’ could inform the extension and operation of existing power grids. Finally, as an example for future infrastructure scenarios, the derived people flows could provide valuable information on where to place charging stations for plug-in electric vehicles.

IX. CONCLUSION

This work has introduced an original framework and relevant assessment methodology to use mobile phone data for the enhancement of electrification practices in developing countries. This framework brings together in an innovative way mobile phone data analysis, socio-economic and geo-referenced data analysis, and state-of-the-art energy infrastructure engineering techniques. More specifically, mobile phone data have been used as a proxy for current and future electricity requirements in different areas. Subsequently, this information was used to quantify the techno-economic feasibility of different centralized and decentralized electrification options in Senegal.

The study shows that mobile phone data can be an accurate means to estimate the energy consumption, peak demand and even the electricity profile of different regions. This information, in turn, has proven to be able to facilitate detailed technical and economic assessments of the considered electrification options, namely, MV grid extension, diesel engine-based Microgrids, and individual household PV systems. The results clearly demonstrate how our framework and methodology can be adopted to quantify how the use of mobile phone data can effectively support electrification plans in developing countries with scarce information on local energy consumption and limited electrification in many areas. Several possible future extensions of the current work have also been discussed in detail, predicated on more extensive energy and mobile phone data.

ACKNOWLEDGMENT

The authors acknowledge Papa Alioune Ndiaye (Ecole supérieure polytechnique de Dakar) for providing the country-wide electricity consumption data for 2013. M. Schläfer thanks Luis Bettencourt, Paul Hines and Seth Blumsack for helpful discussions.

REFERENCES

Quantifying the Effect of Movement Associated with Holidays on Malaria Prevalence Using Cell Phone Data

Sveta Milusheva

April 2, 2015

Abstract

Studying the short term internal movement of populations is vital for developing targeted health policies to fight communicable diseases like malaria. Using mobile phone data, this study establishes short term patterns of movement around Senegal. This movement data is combined with weekly malaria data from ten health districts. Using a specification that accounts for time and place fixed effects, the paper shows a significant correlation between people returning from visits to high malaria places and malaria prevalence in the home district. No parallel effect on malaria prevalence is found in the places visited by people coming from high malaria places. The study finds that every person in the cell phone data who returns from a high malaria place is associated with a minimum of 1.135 new cases of malaria. In addition, the short term movement patterns are shown to be associated with the return of migrants to their place of origin, especially during the time of public holidays. This finding is important because it facilitates the ability to predict when a large number of people will be returning from short term visits and whether they visited high or low malaria places, which can help create more efficient and targeted policies to fight malaria and move towards its eradication.
1 Introduction

While migration has been shown to facilitate development through providing access to better labor markets and education opportunities, the movement of people back and forth between the place they are originally from and the place they have migrated to can also lead to negative consequences like the transmission of diseases. This is a difficult topic to study, due to limited data on internal migration, especially at a frequency level high enough to link it directly to increased prevalence of malaria. With the rise in the use of mobile phone data, it has become easier to trace internal migration and movement, which has allowed researchers to study the expected channels through which malaria travels from high prevalence to lower prevalence areas. Although these channels have been researched in some countries, there has yet to be a study that directly links a rise in malaria to movements within a country on a scale larger than one community or that measures the extent to which such movement impacts malaria prevalence. This paper provides evidence for such a link using mobile phone data for Senegal to measure movement within the country, and combines this with a unique dataset of weekly malaria cases for 20 health posts spread throughout the country.

Recent policies and interventions in Senegal have contributed to a large drop in malaria mortality and morbidity, with parasitaemia in children falling from 5.7% in 2008 down to 2.9% in 2010 -11 (Littrell et al. [2013]). Nevertheless, malaria is still a prevailing problem, especially in the southeast of Senegal. As long as there are still parts of the country where malaria prevalence is high, the whole country is at risk because people travel and can bring the disease to other areas. The study analyzes whether this link between movement and malaria exists and quantifies how big the effect is.

There are many ways to define movement within a country, from daily travel for work between villages and cities to permanent migration from one area to another. This paper focuses on short term movement that can be associated with longer term migration patterns. When individuals or families migrate from one area to another, they will often go back to visit family left behind. This is especially the case around important religious holidays that people want to spend with parents, children and extended families. It is also at those times that many work places are closed, providing individuals the opportunity to make these trips for longer periods of time. Yet when people go home for these longer periods of time, they could either contract malaria, if their home town has a higher rate of malaria, and bring it back to their new place of
residence, or else if their new place of residence has a high rate of malaria, they might bring the disease to their original home town when they visit. This paper studies the extent to which these visits home contribute to an increase in malaria prevalence.

There is extensive literature that looks at the impact of diseases like malaria on long term economic growth, showing a negative effect of these diseases. An increase in malaria prevalence from short term movements induced by migration could therefore have important negative repercussions for places of origin and destination. While this paper does not find increases in malaria in the places of origin, which are the places visited by migrants during holidays, it does find significant impacts on malaria prevalence in the destination locations, to which the migrants return after a visit home. Therefore, increases in migration could lead to increased short term movement that can have important health effects in the destination.

The paper begins with some background on the link between migration and the spread of disease, as well as the recent move towards the use of cell phone data for tracing movement within a country. Section 3 describes the mobile phone and malaria data used in the paper. That is followed by the empirical specification in section 4 and results in section 5. There is a discussion of the comparability of the findings to other data in section 6, extension of the data that looks at the link to long term migration in section 7, and the paper concludes with section 8.

2 Background

Migration and Malaria

The link between migration and the spread of communicable diseases is not a new topic (Prothero 1977). Many have focused on the spread of malaria and communicable disease internationally, finding, for example, that airline traffic is a factor in the spread of epidemics and drug-resistant parasite strains (Balcan et al. 2009, Huang, Tatem, et al. 2013, Tatem and D. L. Smith 2010). While malaria and other pathogens can spread internationally with increased airline travel, and also across borders between countries, movement within a country can also contribute to the spread of a disease and its reemergence in areas where it has been eliminated in the past. Internal migration has been looked at by some using census data or surveys to measure migration in order to describe migration routes and how these relate to the presence of malaria in different parts of a country (Lynch and Roper 2011, Stoddard et al. 2009). Yet the migration data available, especially for internal migration, is
often not sufficiently high resolution to establish a link between internal migration and the spread of a disease. In addition, the migration captured by surveys and the census often times misses short term movements, cyclical migration, or cases where migrants are in sensitive situations (Deshingkar and Grimm 2004).

A number of studies have been done looking at the link between malaria and travel in a specific location. Surveyors select a group of people that is diagnosed with malaria and a comparable group that is not and then conduct a survey that asks about travel history, along with other demographic characteristics that could contribute to malaria contraction. Siri et al. 2010 conduct this type of analysis in Kisumu, Kenya and find that the odds of a child contracting malaria were more than nine times greater for children who reported spending at least one night per month in a rural area. They also find that prior residence in a rural area with higher malaria is a risk factor for contracting malaria, which supports the idea of migrants (often moving to urban areas) visiting their places of origin and contracting the disease. In a similar study done in Quibdo, Colombia using cases that were not limited on age, Osorio, Todd, and Bradley 2004 find that staying outside of the town during the 8-14 days prior to disease onset was the strongest risk factor for contracting the disease. Finally, this is studied in Ethiopia by Yuki et al. 2013, who find that travel away from the home village in the last 30 days is a statistically significant factor for P. falciparum infection, although not for the other major malaria parasite that affects humans, P. vivax. These types of case-control studies have only been done on single locations, which makes it difficult to extrapolate and generalize the findings. Although the analyses that follow are also unable to incorporate all of Senegal, there is malaria data available from locations in different parts of Senegal with heterogeneous environmental characteristics and malaria endemicity levels, which allows the analyses to be more generalizable.

One of the reasons these studies have not been done on a larger scale is that to analyze movement and link it to malaria, it has been necessary to conduct surveys to collect travel history data, which can be time consuming and costly. The increased use of mobile phones, especially the exponential growth in their use in developing countries, and the proliferation of GPS units and GPS enabled smart phones, has opened the possibility to track internal movement and migration at a detailed level that can remain large scale and cover a whole country. Researchers have developed various methods for using the “traces” left when an individual uses a mobile phone in order to determine their location and how that location changes over time. This has been done using data for both developed (Gonzalez, Hidalgo, and Barabasi 2008).
Isaacman et al. (2011) and developing countries (Blumenstock 2012, Williams et al. 2014). Working with cell phone and GPS data has allowed researchers to develop new models of migration and daily movement of individuals and to link that to social and cultural characteristics.

The measurement of movement patterns within a country has implications for a variety of policies, including those surrounding transportation, infrastructure, conflict prevention and most importantly for this paper–disease control. Several papers have already explored the link between movement and disease using cell phone data. Vazquez-Prokopec et al. (2013) use GPS tracking to assess the impact of mobility on the epidemic propagation of a directly transmitted pathogen. Yet, GPS data is more difficult to collect and therefore is limited in scope so that it is not possible to examine temporal variations during and after holidays, which is a focus of this paper. Tatem, Qiu, et al. (2009) look specifically at the spread of malaria using cell phone data and movement between Zanzibar and the mainland, but they also look at only a short period of 3 months. Wesolowski et al. (2012) conduct a study of much wider scope, using cell phone records for over 14 million subscribers, allowing them to study mobility for all of Kenya and to combine it with malaria infection prevalence data to determine areas that act as sources of malaria and other areas that are “sinks”, which receive many migrants from malaria endemic regions. Enns and Amuasi (2013) do something similar in their paper that uses mobile phone data for Cote d’Ivoire. Although these papers are able to outline predictions for where malaria might come from and where it might spread to, they do not draw a direct link between the number of migrants from different areas and the number of cases of malaria in the places receiving migrants.

**Senegalese Context**

In 2013, 94 percent of the population in Senegal was considered in a high transmission area, with more than one case per 1000 people. The only malaria parasite is *P. falciparum* and it is carried mainly by the vectors *An. gambiae*, *An. arabiensis*, *An. funestus*, *An. pharoensis*, and *An. melas*. In total there were 345,889 cases of malaria reported in the country and 815 reported deaths. A number of interventions are used, including insecticide treated nets distributed for free and the use of intermittent preventive treatment (IPT) to prevent malaria during pregnancy. In addition, since the mid-2000s, Senegal has been much more active in detecting and testing for malaria (World Malaria Report 2014).
Past research on internal migration in Senegal has relied on household surveys and small scale surveys of a particular ethnic group or region of the country. A report from 2010 using data from the Senegalese Survey of Households (ESAM) finds that four of the eleven regions in Senegal are net receivers of internal migrants (Fall, Carretero, and Sarr [2010]). Yet, this information does not give an idea of circular movement within the country and the frequency of travel between different areas. Others looking at internal migration in Senegal have focused in on a specific group or tribe, which they are then able to interview extensively (Sahn and Catalina [2013], Benyoussef et al. [1974], Linares [2003], Pison et al. [1993]). A study done in Richard Toll, one of the districts in Senegal that is covered in the data used for this paper, tracked malaria cases over 12 weeks and used a questionnaire to learn more about how malaria was spreading (Littrell et al. [2013]). This is similar to the previous three studies discussed, though it does not involve a case-control methodology but instead is an investigation of confirmed malaria cases. The study found that one of the main risk factors for contracting malaria was travel that entailed an overnight stay. Yet, again, it is difficult to extrapolate from this study because it was conducted for a short period and in one particular area of the country that has few malaria cases. To understand the movement around a country of a pathogen such as malaria, it is necessary to have data from different regions, and it is important to have data on movement back and forth as people visit the families they have left behind.

3 Data

Cell Phone Records

Accurate data on internal migration is rarely available for developing countries, and this is especially true for high frequency data that allows researchers to look at daily, weekly, or even monthly movements within a country. To overcome this problem, researchers have turned to using phone record data collected by cell phone providers in order to measure movement. They track where calls are made from the same phone and follow the movement of the person if they make calls from different locations. This type of data comes with caveats as well. Depending on the percent of the population with access to a cell phone, the data might be capturing movement of only a certain portion of the population, such as those that are high income. In addition, if there is only data available from one mobile phone carrier and there are several major carriers in the nation, then the data would again only capture movement of a portion of the population, which could lead to different biases if the carrier type is associated with certain characteristics of the user. Nevertheless, in a context
with limited data on internal movement, cell phone data provides an opportunity to study short term effects of movement within a country, even if it is for a particular group of people.

The data used for this project comes from phone records made available by the mobile service provider Orange Telecom in the context of a call for projects with the objective to explore the potential of mobile call data to facilitate socio-economic development. The data used in this paper consists of calling data for Senegal at arrondissement level for 146,352 individuals between January 1, 2013 and December 31, 2013 (Montjoye et al. 2014). The current study only focuses on phone calls made between 7pm and 7am in order to determine the locations where individuals live and exclude as much as possible everyday travel for work. In 2013 there were 92.93 mobile phone subscriptions per 100 inhabitants in Senegal, which implies that a majority of the population was using cell phones, which would mitigate bias arising from heterogeneous phone ownership (Union 2013). Orange Telecom had between 56 and 62 percent of the cell phone market in 2013 (Régulation des Télécommunications et des Postes 2013). Although that is over half of the market, bias could arise if those with other cell phone providers differ from the people who use Orange. It is not possible to address this in the current paper, but is a caveat to consider when analyzing the results and their external validity.

Figure 1 demonstrates the total number of calls made each day by the 146,352 individuals in the dataset. At certain times during the year there are spikes in the number of phone calls. Vertical lines in the graph mark major holidays, which often correspond to the spikes in phone calls. Korité and Tabaski are the two biggest Muslim holidays, known also as Eid al-Fitr and Eid al-Adha respectively. The increased number of phone calls around these holidays is reassuring because it means that it will be possible to pick up individuals’ locations at that time and compare them to the locations where they normally reside.

Figures 2a and 2b demonstrate how the cell phone data can be used to pick up movement within Senegal. There are two holidays, the Prophet’s birthday and Magal de Touba, that involve large pilgrimages to a particular location. On the day of the Prophet’s Birthday, those of the Tijanimus Sufi brotherhood travel to the holy city of Tivaouane for the Maouloud festival in the Pambal arrondissement. Figure 2a shows how around this day there is a big jump in the number of people making calls from this arrondissement as a percent of the sample that is assigned to this
arrondissement as their home.

Similarly, for the Magal de Touba, those who follow the Mouride Sufi brotherhood take part in a pilgrimage to their religious center of Touba in the Ndame arrondissement. Again, during the two times when the holiday occurred in 2013, we see large jumps in the number of people making calls from Ndame as a percent of the sample who lives there. Both of these demonstrate how people visiting a different arrondissement make phone calls during their visit, allowing us to pick up when people have moved within the country for a short period of time.

Malaria Data

The data used to measure malaria prevalence comes from the Programme National de Lutte Contre le Paludisme (PNLP) (Bulletin de Surveillance Sentinelle du Paludisme No 1-46, 2013). This national program, which has the goal of controlling malaria in Senegal, has been collecting weekly data on number of malaria cases at twenty health posts around the country. This data has been collected starting in 2008 for some of the locations, but there is consistent data available online only from the middle of February 2013 until the end of November 2014. Figure 3 shows the location of the 20 sites where data has been collected. Since cell phone data is only available for 2013, only malaria data through the end of 2013 is used.

In addition, average malaria prevalence for 2013 from PNLP was used to characterize each arrondissement as having high or low malaria. A reproduction of the map used to assign malaria status to each arrondissement is shown in Figure 4 (Rapport Statistique PNLP Spécial: 2010-2013). Those arrondissements with under 15 cases of malaria were considered low malaria and those with over 15 cases are high malaria, and this is coded as a 0/1 dummy variable.

Scaling of the Data

Since weekly malaria data is available for only some health posts while the movement data is at the arrondissement level, it is necessary to adjust the data. There is data available on number of health posts in each health district in 2011 (Sanitaire 2011). This provides the ability to scale the malaria data up to the health district level. Since the twenty health posts are distributed so that there are two per health district, the weekly malaria numbers for each pair of health posts is averaged and

---

[1] A home arrondissement is assigned based on the most number of calls made from it. This is described in more detail in section 4.
then scaled by the number of health posts in the district. This provides an estimate for the average number of weekly malaria cases in the health district.

In order for the movement data to match the malaria data, the arrondissements are aggregated to the health district level. There are 123 arrondissements that were regrouped into the 74 health districts. The movement numbers are then calculated at the health district level rather than the arrondissement level in order to make the spatial units of the movement data and malaria data the same.

4 Empirical Specification

This paper focuses on measuring the effect of short term mobility on malaria prevalence. In analyzing the effect of short term movements on malaria, there are two possible channels through which malaria could be spread. People visiting from high malaria places could bring malaria to the places they visit and also individuals going to high malaria places could bring malaria back when they return home after a short visit. Both of these types of movement are analyzed to see whether they contribute to the spread of malaria. In order to conduct this type of directional analysis, both a “home” health district and a short term visit need to be defined.

Measuring Movement

A “home” district is assigned to each person based on where they have spent the most number of days out of the number of days in which they made phone calls in 2013. A location was assigned to each day based on the health district in which the person made the most number of phone calls between 7pm and 7am. In the case where the same number of phone calls were made from more than one district, one of the districts was randomly chosen as the main one. Similarly, when determining the “home” location for the year, if someone spent an equal number of days in two different districts, one was chosen at random. For over 90 percent of individuals, they spent more than half of their days in the district labeled as “home”. The remaining 10 percent move around quite a bit, making it difficult to accurately assign a “home” location, but the rule of picking the place where the person spends the most number of days seems reasonable. Robustness checks were conducted where the analysis was done excluding those people who spent less than half of their days in the data in a single district. The results found were even stronger than those including these people. Therefore, these ten percent of people are not driving the analysis and, if anything, are decreasing the effects seen by introducing measurement error.
A short term visit is defined as a visit between 3 and 14 days because it needs to be long enough that it does not capture daily movements for work between districts (especially in those cases of people living near the border of two districts), but is also not too long since this paper is focused on short term visits. Other studies that have looked at the link between movement and malaria using survey data have looked at the effect in as little as one night spent away in a high malaria place. The analysis here is conservative in measuring the effects of short term movements in looking at travel of at least three days spent away from home. The visit is identified in the data by comparing the individual’s “home” district to the district assigned to each day and picking out instances when the individual has 3 to 14 days in a row when he or she is away from the “home” district. The district visited is assigned based on the location where the person spent the most time on the visit, and it is randomly assigned in the case of a tie between multiple districts.

Figure 5 shows the number of people who return in a certain week to their “home” district after a visit away for 3-14 days. There are big jumps representing increased movement around the Prophet’s Birthday, Independence Day, Korité and Tabaski. This confirms the hypothesis that the majority of short term movements are focused around important holidays. It is important to note that the jumps in movement do not correspond strictly to number of phone calls seen in the first graph, since Independence day is not a day when there were many calls, but there is a large jump in movement. Therefore, the analysis is not being driven by number of phone calls.

Focusing on the ten health districts that are studied in this paper in which the twenty health posts are located, Figure 6 gives an idea of the movement patterns associated with these districts. The map for each district shows the percentage of trips to each of the other 73 districts out of all trips made in 2013. This is measured by the number of individuals who are assigned to that district as their home based on the phone data and who travelled to another district and then returned back to this district (all based on the earlier definitions given). As might be expected, the figure shows that individuals tend to travel most to the districts closest to their own district.

These analyses were also done measuring movement as 8 to 14 days, and they exhibit the exact same pattern; therefore, changing the definition of short term migration.
Malaria and Movement

Now we want to examine whether we might expect a link between the movement data and the malaria data. In order to relate malaria prevalence to movement around the country, it is important to measure not only short term movement but specifically movement from high malaria places, since we would not expect someone visiting from or returning from a low malaria place to influence malaria in the place where they visit or their home. The dummy variable for high malaria based on the 2013 annual numbers is multiplied times the number of people visiting or returning from that district. These products are summed across all the districts for each district of interest to get an aggregate measure for total number of people visiting from high malaria places or returning from high malaria places. These measures are aggregated at the weekly level in order to be comparable to the weekly malaria data.

Malaria in Senegal is seasonal, with almost all cases falling between July and December. This is largely driven by the rainy season which occurs during that time period, though has slight variations in timing depending on the region. Therefore, the analyses in this paper focus only on the weeks from July to December. Figure 7 shows for each of the ten health care districts with malaria data between July and December, malaria graphed on the left hand y-axis and number of people returning to the district after visiting high malaria places on the right hand y-axis. For many of the locations, we do not see a strong link because there are few cases of malaria. In the locations where there are more case of malaria, there does seem to be a pattern of bumps in movement followed by bumps in malaria.

In order to estimate the relationship seen in the graphs more robustly, the following equations are estimated:

\[
\begin{align*}
TotalMalaria_{it} &= \alpha + \beta R_{it} + \gamma_i + \lambda_t + \epsilon_{it} \\
TotalMalaria_{it} &= \alpha + \beta V_{it} + \gamma_i + \lambda_t + \epsilon_{it} \\
R_{it} &= \sum_{j=1}^{73} (R_{jt} \ast H_j) \\
V_{it} &= \sum_{j=1}^{73} (V_{jt} \ast H_j)
\end{align*}
\]

where \(TotalMalaria_{it}\) is the total number of malaria cases in location \(i\) in week \(t\), \(R_{it}\) is a measure of the number of people from district \(i\) who visited a high malaria district for 3 to 14 days and returned in week \(t\) to their home, \(V_{it}\) is the number of
people visiting district i and are originally from high malaria districts, $\gamma_i$ captures location fixed effects, $\lambda_t$ captures time fixed effects, $\epsilon_{it}$ is the error term, $R_{jt}$ and $V_{jt}$ are the number of returnees and visitors from district j to district i during week t, and $H_j$ is a dummy coded as 1 if district j is high malaria and 0 otherwise.

After an infected mosquito bites a non-infected human, *P. falciparum* incubates for 7 to 15 days (Doolan, Dobaño, and Baird 2009). In turn, once a non-infected mosquito has bitten an infected person, there is a temperature dependent extrinsic incubation period that was found to last between 9 and 14 days in 42 different area studies (Killeen, Ross, and T. Smith 2006). One of the studies specifically in Ndiop, Senegal found an incubation period of 9.3 days (Killeen, Ross, and T. Smith 2006). Therefore, migration is not expected to affect malaria prevalence in the same week. Instead, a period of at least one week is expected before a person who has travelled would show signs of malaria, and at least three to four weeks would be expected before other people infected due to the infection of that person can be diagnosed. To account for this, the variables $R_{it}$ and $V_{it}$ are lagged, with four weeks used in most analyses, although results are also shown for zero to eight lags.

In addition, it might be expected that malaria in previous weeks would affect malaria in the current week due to the way the disease is spread. Therefore, models are run that control for lagged number of malaria cases to account for this structure in the data. This could potentially lead to underestimates in the effect of migration though because malaria cases due to movement that lead to increased malaria in the following weeks would not be attributed to the movement. Finally, all models included corrected standard errors for the panel structure of the data.

5 Results

The main results are presented in Table 1. Column 1 shows the results from running equation 1 where number of malaria cases in the district is regressed on number of people returning from high malaria places with time fixed effects and district fixed effects included. Migration is lagged four weeks in order to take into account the incubation period within the infected person and the incubation period within the mosquito. We see a large and significant effect of 3.128. Column two shows a robustness check where the number of individuals returning from low malaria places is included in the regression and we see that there is no effect on either the size or magnitude of the coefficient of interest. Column 3 shows the results when number of malaria cases in the previous week is included. This variable is also extremely
correlated with malaria in the current week, as would be expected, and brings the R squared up to 0.905, which is common in autoregressive models. Nevertheless, the coefficient on number of returnees from high malaria places remains significant and larger than 1.

Additional robustness checks are shown in columns 4 and 5. In column 4, the health district with the highest number of malaria cases (Pikine) is dropped because it is an outlier in the sample in terms of number of cases. Once this location is dropped, the results are again significant at the .05 level for people returning from a high malaria place, though the coefficient is about half the magnitude. This implies that the extreme number of malaria cases in Pikine, which could be a function of other important factors aside from movement within the country, is potentially affecting the analysis more than the other districts. Nevertheless, though the results are affected by Pikine, they are not solely being driven by this one location. Finally, in the last column the last 4 weeks of the year are excluded because the data was only graphically available, and therefore it was necessary to infer the malaria numbers from the graph, which could introduce error. The coefficient continues to remain significant and the magnitude is comparable; therefore, these last four weeks are not leading to a bias in our findings. Although interpreting the size of the coefficients in these five regressions is difficult since we only have number of people returning based on a 10% sample of the population and the number of malaria cases only includes those that are actually reported; nevertheless, the results consistently show a significant and positive relationship between returnees and cases of malaria.

Table 2 shows the analogous results to table 1 except this time instead of looking at number of people returning from high malaria places it focused on number of people coming to visit from high malaria places. We see that visiting does not have an effect on malaria under any of the models. Although the results shown use four lags to be comparable to the results from people returning, the analysis was run using one, two and three lags and there is still no effect on number of malaria cases. Therefore, short term visitors do not seem to have an impact on malaria prevalence.

An additional robustness check is conducted by looking at only number of people returning from low malaria places. We would not expect this to have an effect on malaria since those individuals should not have contracted malaria while traveling and therefore should not get sick or pass on the disease to vectors in their home community. As expected, table 3 shows no effect of returnees from low malaria places under any of the models.
Finally, table 4 shows the results obtained from running nine different regressions each with a different lag in number of returnees from none to 8 weeks. As would be expected, there is no effect of number of returnees in the same week, the following week, or two weeks later. There is then a significant effect from three weeks post return up to five weeks post return and then again no effect. The magnitude also changes accordingly, with a smaller magnitude three weeks post return, at which point only those that have travelled are the ones that would show up as having malaria, but by four weeks after the magnitude is larger because others in the village might also have been affected. Although the individual returning might have been infected anytime during their visit, and not necessarily on the last day, there can also be a period of a few days even after symptoms begin before the individual actually seeks medical attention. The current malaria data only provides number of cases based on individuals who came to the health post, were tested, and were found positive for malaria. These regressions were run controlling for lagged malaria in order to be more conservative, but that means that the effect found in the regressions is probably smaller than the actual effect.

6 Discussion

After establishing a method of using the cell phone data to measure short term movement around the country and specifically looking at number of visitors to a district and number of residents returning to a district, the results indicate a positive correlation between returnees and cases of malaria, although no such correlation between visitors and malaria. These results remain significant after several robustness checks. These findings are important because they show how increase in short term movement could lead to increases in number of malaria cases. They also show the direction in which this relationship occurs, so that the effect is only concentrated on those returning from high malaria places rather than visitors from high malaria places. This allows policies to target specific locations that receive a large number of returnees from short term periods. It also points to specific strategies that can be used to reduce malaria such as providing travelers with bednets to use when visiting a high malaria place since there might not be extra ones for them to use.

These results can be compared to data available on malaria transmission and reproduction in Senegal. One measure of transmission that is a useful source of comparison when looking at the effect of movement on malaria is the reproduction number. The reproduction number for an infection $R_0$ is the average number of secondary cases a
typical single infected case will cause in a population (T. Smith and Schapira [2012]). This number takes into account the infectiousness of the human host, the density of vectors, their propensity to feed on humans and their survival. Gething et al. [2011] have created a global map of the P. falciparum reproduction number for each 5km x 5km pixel. This is done by aggregating data available from a variety of sources, and the data has been made available in raster form on the Malaria Atlas Project website (http://www.map.ox.ac.uk/). Looking at all the pixels that make up Senegal, the values for the reproduction number go from a low of 1.00433 to a high of 19.7154. Using this data, it is possible to obtain an average reproduction number for each of the ten health districts studied here. Five of them have the lowest possible value of 1.004 (Pikine, Mbao, Guediawaye, Podor and Richard Toll), Matam has an $R_0$ of 1.089, Linguere is 1.248, Ndoffane has an $R_0$ of 1.580, and the two highest are Baka and Kédougou with $R_0$s of 3.718 and 5.878 respectively. Taking the average of these, we get 1.853, suggesting that each case of malaria leads to 1.853 cases of malaria on average for these ten districts. This is very much in line with the coefficients in the analysis shown here.

7 Extension of the Data to Link Long Term Migration to Short Term Movement and Holidays

This paper focused on the effect of short term movement of 3 to 14 days on malaria in Senegal. This type of movement is often associated with a holiday, when individuals will visit their place of origin for several days both because of the importance of the holiday but also the time off from work, which makes it possible to visit for longer. This association between movement and holidays is of particular interest because of its connection to migration. As more people migrate permanently or semi-permanently within a country, we would expect an increase in this type of short-term movement home around holidays since there are more people living in a home that is different from their place of origin. Yet, as has been shown, short term movement can lead to the spread of disease. Therefore increased migration could also lead to increased prevalence of infectious diseases like malaria. In addition, if a link is established between short term movement and long term migration patterns, it would be possible to use migration patterns to predict short term movements around holidays, even without access to cell phone data to map out daily movement, which could help with the implementation of effective prevention policies. This section establishes this link between short and long term migration in the data. Since the malaria data is not used in this context, the analysis is done at the more detailed
arrondissement level rather than the health district level.

In order to link long term migration to short term movement, it is necessary to define longer term migration. Since the data only covers 12 months, it limits the definition of long term migration. In this context, long term migration is defined as a move from an arrondissement where the person has resided for at least 3 months to a different arrondissement where the person resides for the following three months. A person is defined as residing in an arrondissement based on the most number of days spent in an arrondissement in a month, where an arrondissement is assigned to a particular day based on the location of the most number of calls made on that day.

The use of this definition means no long term movement can be calculated for January-March or November-December since the data is censored on each end and it is not possible to know if an individual was in the same or different location in the months leading up to 2013 or after December 2013. Therefore, this is just an approximation of long term migration that is used to get a sense for how long and short term movements are related.

Once annual long term migration numbers were calculated, net long term migration for 2013 between pairs of arrondissements is then examined. Looking at pairs of arrondissements x and y, net long term movements can be characterized as going from x to y or from y to x. The opposite flows of movement should be seen around holidays. So if long term movements go from x to y, net movement before holidays should go from y to x and after the holiday it would again go from x to y as people return to the location where they have migrated long term. This is the correlation tested in order to see if short term movements around holidays are linked to long term migration.

In order to test this, short term movements are calculated by looking at net movement between pairs of arrondissements over seven day periods. This is done by taking each day d in 2013 starting with day 7 and looking at the 6 days prior and the day itself. The number of movements from arrondissement x to y and from y to x are then measured, where a movement is just a change from arrondissement x to arrondissement y on consecutive days, in order to come up with a net movement (so that if the majority of moves are from x to y, then the net movement is positive for y because people are entering and negative for x because people are exiting). Regressions are run of these 7 day movements between pairs for each day d, where the dependent

---

3Similarly to the district case, in the case where the same number of calls were made from two arrondissements, one was randomly chosen.
variable is the net movement over the 7 days and the independent variable is the net annual long term movement between the pair. These coefficients are graphed in figure 8, with the exception that insignificant coefficients are graphed as 0. In addition, all the public holidays are marked with vertical lines and described in the note below the graph. The graph shows that the times when the coefficient is significant, negative and large align perfectly with the days where there are public holidays. This implies that in the week prior to and including the holiday, movement is going in the opposite direction from long term migration. In addition, every one of these large drops below 0, is then followed by an equally large jump above 0, signifying that in the days after the holiday, people return to the places where they have migrated long term, and therefore movement goes in the same direction as long term migration.

Although the definition of long term migration is limited due to the timing of the data, nevertheless, the phone data is suggestive of significant movements of the population around holidays going in the opposite direction from regular long term movement patterns before the holiday and in the same direction after the holiday. Therefore, reverse migration patterns can predict short term visits, especially around holidays when the highest number of such visits occur, and these visits and returns home are correlated with increases in malaria cases. With the establishment of these links, it becomes possible for policy makers to create more targeted malaria policies even when only limited data is available. In the case that there is no cell phone data that can be used but there are surveys that measure long term migration, then those patterns can be used to predict short term movements around holidays. If cell phone data is available, then this can be used to look at annual short term migration patterns to determine the exact timing of movement, and it can also be used to determine long term migration patterns in order to predict the magnitude of short term movement that can be expected. Using this data, it becomes possible to determine which areas might see the most influx of returnees from high malaria places, which could help guide targeted interventions that effectively reduce the spread of malaria during high peaks of movement.

8 Conclusion

While previous literature has established how malaria might be expected to spread based on the movement of people, this paper tries to create a direct link between short term movements and increases in malaria cases. It also explores the timing of movement based on holidays and how it relates to long term migration patterns. This link and timing are important for policy in order to determine not
only to what extent movement around a country could lead to a rise in malaria, but also to find which are the spots that might be most at risk and at which times of year.

The analyses show an increase in malaria due to people returning home, but no effect on the places that people visit. The magnitude of the most conservative regression that controls for lagged malaria suggests that at the peak of four weeks after a returnee has come home after visiting a high malaria place, he or she contributes to 1.135 malaria cases. Without controlling for number of malaria cases lagged, the coefficient is around 3.1, suggesting that a returnee in the data is associated with 3.1 cases of malaria. The actual effect of migration should fall in between those two coefficients. Considering the fact that holidays cause the number of returnees to jump exponentially, specific times of year are especially susceptible to a rise in malaria cases and potentially the start of an epidemic.

An extension of this work would be to apply the current analysis to other communicable diseases such as influenza or meningitis. Preliminary analyses have been conducted to study the effect of movement on all patient visits to the 20 health posts (not just those related to malaria). These indicate an even stronger effect than the one seen for malaria, suggesting that such a link between these other diseases and movement could be established. Using the different incubation periods of diseases, it would be possible to separate out the migration effect since we would expect to see the effects of migration on the spread of the diseases at different times. This type of research would be extremely important for health policy as it contributes to an understanding of how communicable diseases could spread, allowing the implementation of directed policies tailored to each disease and hotspots with the highest risk.

The current research establishes a link between short term movements around the country and increased malaria. Many of these visits are driven by people who have migrated to another area of the country returning home to visit family. It is possible for policy makers to use the cell phone data to determine migration routes, which would help predict short term visits during holidays, and could then establish hotspots that are expected to see an increase in returnees from high malaria areas. They can then target these hotspots with malaria prevention campaigns before large holidays, such as providing travelers with bednets to use on their trip in case there are no extra ones in the place they are visiting. These types of policies could decrease the spread of malaria in Senegal and bring the country closer to eradication of the disease. In addition, this paper demonstrates how cell phone data to measure movement and high frequency malaria data can be used to test the accuracy of malaria
reproduction numbers. These numbers are devised using mathematical models and a number of assumptions because it is very difficult to collect some of the data necessary such as biting frequency. By looking at the effect of returnees from high malaria places, it gets at exactly the question of what is the rate of the spread of the disease when a new infected case comes in. Of course not all those who return from a high malaria place are infected, but it is still possible to come up with an estimate that can then be compared with the $R_0$ in order to check the accuracy of the models that are currently used to predict malaria transmission and are used by policy makers and academics in studying malaria and the possibility of its eradication.

References


Fall, Papa Déméba, María Hernández Carretero, and Mame Yassine Sarr (2010). “Country and Research Areas Report”. In:


Huang, Zhuojie, Andrew J Tate, et al. (2013). “Global malaria connectivity through air travel”. In: Malar J 12.1, p. 269.


Linares, Olga F (2003). “Going to the city...and coming back? Turnaround migration among the Jola of Senegal”. In: Africa 73.01, pp. 113–132.

Littrell, Megan et al. (2013). “Case investigation and reactive case detection for malaria elimination in northern Senegal”. In: Malar J 12, p. 331.


Montjoye, Yves-Alexandre de et al. (2014). “D4D-Senegal: The Second Mobile Phone Data for Development Challenge”. In:


Sahn, David Ezra and HERRERA Catalina (2013). Determinants of Internal Migration among Senegalese Youth. Tech. rep. CERDI.


Siri, Jose G et al. (2010). “Significance of travel to rural areas as a risk factor for malarial anemia in an urban setting”. In: The American journal of tropical medicine and hygiene 82.3, pp. 391–397.


Yukich, Joshua O et al. (2013). “Travel history and malaria infection risk in a low-transmission setting in Ethiopia: a case control study”. In: *Malar J* 12, p. 33.
Figure 1: Total Calls Per Day
Figure 2: People Calling per Day as Percent of Sample Living in the Particular Arrondissement
Figure 3: Map of Health Posts with Weekly Malaria Data Collection

Figure 4: Map of Malaria Prevalence in Senegal in 2013
Figure 5: Total Number of People Returning to “Home” Health District Each Week
Figure 6: Percent of People Returning from Each Health District to Each Health District Analyzed, 2013
Figure 7: Weekly Cases of Malaria and People Returning Home from High Malaria Places During the Malaria Season
<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Include Low Malaria Return</td>
<td>Include Lagged Malaria</td>
<td>No Pikine</td>
<td>25-48 Weeks</td>
<td></td>
</tr>
<tr>
<td>Return, High Malaria</td>
<td>3.128**</td>
<td>3.120**</td>
<td>1.136**</td>
<td>1.624**</td>
<td>3.708***</td>
</tr>
<tr>
<td></td>
<td>(1.231)</td>
<td>(1.224)</td>
<td>(0.487)</td>
<td>(0.698)</td>
<td>(1.269)</td>
</tr>
<tr>
<td>Return, Low Malaria</td>
<td>-0.150</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.145)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lagged Malaria Cases</td>
<td></td>
<td></td>
<td>0.865***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.0596)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>282.4***</td>
<td>284.8***</td>
<td>58.72</td>
<td>380.3***</td>
<td>371.4***</td>
</tr>
<tr>
<td></td>
<td>(73.24)</td>
<td>(73.43)</td>
<td>(39.51)</td>
<td>(55.38)</td>
<td>(71.01)</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>District Fixed Effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>230</td>
<td>230</td>
<td>230</td>
<td>207</td>
<td>190</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.542</td>
<td>0.545</td>
<td>0.905</td>
<td>0.524</td>
<td>0.560</td>
</tr>
<tr>
<td>Number of panels</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 1: Effect of Number of People Returning after Visiting a High Malaria Place for 3-14 days on Malaria Cases, Lagged 4 Weeks
<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Include Low</td>
<td>Include No</td>
<td>Lagged Malaria</td>
<td>Pikine</td>
<td>Weeks 25-48</td>
</tr>
<tr>
<td>Visit, High Malaria</td>
<td>-0.598</td>
<td>-0.660</td>
<td>0.262</td>
<td>0.410</td>
<td>-0.0790</td>
</tr>
<tr>
<td></td>
<td>(0.916)</td>
<td>(0.917)</td>
<td>(0.347)</td>
<td>(0.727)</td>
<td>(0.863)</td>
</tr>
<tr>
<td>Visit, Low Malaria</td>
<td>0.0893</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0841)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lagged Malaria Cases</td>
<td></td>
<td></td>
<td>0.887***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.0629)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>303.0***</td>
<td>298.1***</td>
<td>54.76</td>
<td>394.6***</td>
<td>379.0***</td>
</tr>
<tr>
<td></td>
<td>(77.07)</td>
<td>(76.59)</td>
<td>(41.39)</td>
<td>(56.12)</td>
<td>(79.38)</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>District Fixed Effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>230</td>
<td>230</td>
<td>230</td>
<td>207</td>
<td>190</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.506</td>
<td>0.508</td>
<td>0.901</td>
<td>0.500</td>
<td>0.498</td>
</tr>
<tr>
<td>Number of panels</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 2: Effect of Number of People Visiting from a High Malaria Place for 3-14 days on Malaria Cases, Lagged 4 Weeks
<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Include</td>
<td>No</td>
<td>Lagged Malaria</td>
<td>Pikine</td>
</tr>
<tr>
<td>L4.tot_lowreturn</td>
<td>-0.155</td>
<td>-0.0526</td>
<td>-0.0849</td>
<td>-0.142</td>
</tr>
<tr>
<td></td>
<td>(0.148)</td>
<td>(0.0733)</td>
<td>(0.0911)</td>
<td>(0.148)</td>
</tr>
<tr>
<td>L.district_cases</td>
<td>0.883***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0633)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>298.5***</td>
<td>59.67</td>
<td>400.1***</td>
<td>380.5***</td>
</tr>
<tr>
<td></td>
<td>(75.89)</td>
<td>(41.57)</td>
<td>(56.15)</td>
<td>(78.64)</td>
</tr>
<tr>
<td>Time Fixed Effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>District Fixed Effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>230</td>
<td>230</td>
<td>207</td>
<td>190</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.508</td>
<td>0.901</td>
<td>0.500</td>
<td>0.501</td>
</tr>
<tr>
<td>Number of panels</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3: Effect of Number of People Returning after Visiting a Low Malaria Place for 3-14 days on Malaria Cases, Lagged 4 Weeks
<table>
<thead>
<tr>
<th>Lags</th>
<th>Returnee Coef.</th>
<th>Lagged Malaria Cases</th>
<th>Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Lag</td>
<td>-0.0656</td>
<td>0.893***</td>
<td>33.90</td>
</tr>
<tr>
<td></td>
<td>(0.149)</td>
<td>(0.0575)</td>
<td>(31.50)</td>
</tr>
<tr>
<td>1 Lag</td>
<td>0.0224</td>
<td>0.891***</td>
<td>34.06</td>
</tr>
<tr>
<td></td>
<td>(0.491)</td>
<td>(0.0580)</td>
<td>(31.45)</td>
</tr>
<tr>
<td>2 Lags</td>
<td>0.617</td>
<td>0.886***</td>
<td>37.83</td>
</tr>
<tr>
<td></td>
<td>(0.494)</td>
<td>(0.0595)</td>
<td>(32.97)</td>
</tr>
<tr>
<td>3 Lags</td>
<td>1.112**</td>
<td>0.876***</td>
<td>85.04*</td>
</tr>
<tr>
<td></td>
<td>(0.478)</td>
<td>(0.0598)</td>
<td>(34.16)</td>
</tr>
<tr>
<td>4 Lags</td>
<td>1.136**</td>
<td>0.865***</td>
<td>57.14</td>
</tr>
<tr>
<td></td>
<td>(0.487)</td>
<td>(0.0596)</td>
<td>(35.39)</td>
</tr>
<tr>
<td>5 Lags</td>
<td>0.925*</td>
<td>0.850***</td>
<td>5.464</td>
</tr>
<tr>
<td></td>
<td>(0.503)</td>
<td>(0.0613)</td>
<td>(37.64)</td>
</tr>
<tr>
<td>6 Lags</td>
<td>0.0728</td>
<td>0.856***</td>
<td>26.69</td>
</tr>
<tr>
<td></td>
<td>(0.527)</td>
<td>(0.0682)</td>
<td>(37.21)</td>
</tr>
<tr>
<td>7 Lags</td>
<td>-0.217</td>
<td>0.861***</td>
<td>-2.094</td>
</tr>
<tr>
<td></td>
<td>(0.533)</td>
<td>(0.0728)</td>
<td>(38.40)</td>
</tr>
<tr>
<td>8 Lags</td>
<td>0.212</td>
<td>0.847***</td>
<td>49.19</td>
</tr>
<tr>
<td></td>
<td>(0.571)</td>
<td>(0.0783)</td>
<td>(39.60)</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4: Effect of Number of People Returning after Visiting a High Malaria Place for 3-14 days on Malaria Cases, 0 to 8 Weeks Lag
Figure 8: Coefficients from Regressions for Each Day in 2013 of Net Movement Between Arrondissements in Previous 7 days on Long Term Net Migration.

Uncovering the impact of human mobility on schistosomiasis via mobile phone data

Lorenzo Mari\textsuperscript{1}, Renato Casagrandi\textsuperscript{1}, Manuela Ciddio\textsuperscript{1},
Susanne H. Sokolow\textsuperscript{2}, Giulio De Leo\textsuperscript{2}, Marino Gatto\textsuperscript{1,*}

\textsuperscript{1} Dipartimento di Elettronica, Informazione e Bioingegneria,
Politecnico di Milano, Italy

\textsuperscript{2} Hopkins Marine Station,
Stanford University, United States of America

* Corresponding author: marino.gatto@polimi.it

December 30, 2014

Abstract

Schistosomiasis is a parasitic infection with chronic debilitating symptoms that is widespread in sub-Saharan Africa. In this work we study country-scale disease transmission dynamics in Senegal, where schistosomiasis represents a major health problem. The analysis is performed by means of a spatially-explicit model accounting for both local epidemiological dynamics and human mobility. Human hosts can in fact be exposed to contaminated water (and contribute to water contamination if infected) while traveling outside their home communities to carry out their activities. Mobility patterns are estimated from low-resolution movement routes extracted from anonymized mobile phone records made available by Orange and Sonatel in the context of the D4D-Senegal data challenge. The results of our analysis show that a relatively simple model can reproduce regional patterns of schistosomiasis prevalence quite reliably. Mobility is found to play a nontrivial role in disease spread, as it can either increase or decrease transmission risk – with the latter effect being predominant at large spatial scales. We also study the effectiveness of some intervention strategies aimed at reducing the burden of the disease and discuss how the model can be transformed into a decision-support tool to help eradicate schistosomiasis from Senegal.
1 Introduction

Schistosomiasis, also known as bilharziasis, is a major parasitic infection that affects about 250 million individuals in many areas of the developing world and that puts at risk about 700 million people in regions where the disease is endemic according to the World Health Organization (WHO 2014). Schistosomiasis is a major cause of mortality, being directly responsible for the death of about 12,000 people yearly (Lozano et al. 2012) and a co-factor in at least 200,000 deaths annually (Théiot-Laurent et al. 2013), and morbidity, with 20 million people suffering from severe consequences from the disease (Kheir et al. 1999) and an estimated disability-adjusted life years (i.e. the number of years lost due to ill-health, disability or early death) of 4.5 million (Penwick 2012). These figures likely make schistosomiasis one of the most common parasitic diseases (second after malaria) and the deadliest among neglected tropical diseases. The largest intensity of infection is usually observed in children, who are especially vulnerable and, if not treated, suffer chronic consequences into adulthood. Typically, parasites inside human tissues induce a response that causes local and systemic pathological effects ranging from anaemia, impaired growth and cognitive development, and decreased physical fitness, to organ-specific effects such as fibrosis of the liver, bladder cancer, and urogenital inflammation (Tzanetou et al. 2007; Colley et al. 2014). The burden of schistosomiasis is disproportionately concentrated in Africa, most notably in the sub-Saharan part of the continent (which accounts for at least 90% of cases worldwide, WHO 2014).

Schistosomiasis is caused by trematode parasites belonging to the genus Schistosoma (see again Colley et al. 2014). Most human infections are caused by three major species: S. mansoni, S. haematobium, or S. japonicum. These parasites need as intermediate hosts certain types of freshwater snails belonging to the genus Biomphalaria (for S. mansoni), Bulinus (for S. haematobium), or Oncomelania (for S. japonicum). The geographical distribution of schistosomes is linked to the specific range of the snail host habitat. As an example, only S. haematobium and S. mansoni are found in Africa. The infectious form of the parasite for humans is a freely swimming, short-lived larval stage, known as cercaria, that is shed by infected snails. Cercariae can infect humans penetrating their skin when they come into contact with contaminated freshwater. Within the human body, they develop into sexually mature adult parasites that can live for years, mating and producing hundreds to thousands of fertilized eggs daily. Eggs can leave the human host through faeces (S. mansoni or S. japonicum) or urine (S. haematobium). Eggs that reach freshwater hatch into so-called miracidia, a second short-lived larval form of the parasite that is infectious for snails. Miracidia undergo asexual replication in snail hosts, which can then shed tens of thousands of cercariae into water, thus completing the parasite’s life cycle (Fig. 1).

Spatial coupling mechanisms are very important in the spread, persistence and infection intensity of schistosomiasis (Gurarie et al. 2010). Parasites can in fact be carried in advective flows along canals
Figure 1: Schistosomiasis transmission cycle. Paired adult worms within human hosts (A) produce eggs (B; left to right: *S. mansoni*, *S. japonicum*, *S. haematobium*) that are shed through faeces or urine and hatch into miracidia (C). Miracidia infect intermediate snail hosts (D; left to right: genus *Biomphalaria*, *Bulinus*, *Oncomelania*), which then shed free-swimming cercariae (E) that can penetrate human skin and develop into reproductive schistosomes.

and streams as larvae, moved between aquatic and riparian habitats within intermediate snail hosts, and transported by human hosts as adult worms. While larval transport and snail movement may represent significant propagation pathways for the disease only over short spatial scales (in the order of hundreds of meters; see [Maszle et al., 1998](#) [Lowe et al., 2005](#) or long temporal windows (e.g. because of habitat expansion following water resources development; see [Steinmann et al., 2006](#) [Clemmon et al., 2007](#)), respectively, human mobility can play a significant role in disease propagation in and from endemic areas ([Bella et al., 1980](#) [Appleton et al., 1996](#) [Cetron et al., 1996](#) [Kloos et al., 2010](#)). Susceptible people can in fact become infected while traveling, and import the disease back in their home communities, while infected travelers coming from endemic regions can introduce schistosomiasis into villages that were previously disease-free. This type of medium-to-long range contamination is compatible with the successive focal transmission of the disease at the local level suggested by recent landscape genetic studies ([Criscone et al., 2010](#)).

Human mobility differs from the other ecohydrological pathways of schistosomiasis propagation in that human movement can occur over much larger spatial scales and much less predictably ([Remais, 2010](#)). As a matter of fact, despite recent advances in the modeling of human mobility (e.g. [González et al., 2008](#) [Song et al., 2010a](#) [Simini et al., 2012](#)), there still remain fundamental limits in our understanding of where, when, why and how people move ([Song et al., 2010b](#) [Lu et al., 2013](#)). Therefore, proxies of human
mobility that can be remotely acquired, properly anonymized and objectively manipulated represent an invaluable tool to inform epidemiological models. In this respect, the analysis of call detail records (CDRs) represents one of the most promising tools to infer human mobility patterns in a reliable (thanks to large sample size) and detailed (in both space and time) way (e.g. Palchykov et al., 2014). These characteristics are very important for epidemiological research, as shown by the increasing number of studies that make use of CDR analysis (see e.g. Wesolowski et al., 2012; Tatem et al., 2014; Tizzoni et al., 2014; Wesolowski et al., 2014).

In this work we aim at exploring country-wide patterns of schistosomiasis transmission in Senegal, where the urogenital form of the disease is widespread and represents a major health problem (Ndir, 2000; Schur et al., 2011), with average prevalence estimates as high as 25% (Briand et al., 2005). To that end, we apply a spatially explicit model of schistosomiasis dynamics proposed by Gurarie et al. (2010). Specifically, the model accounts for the dynamics of human hosts (characterized by different infection levels), intermediate snail hosts and larval parasite stages. Because of the large spatial scale of interest, human mobility is retained as the only mechanism being responsible for the spatial spread of the disease. Human movement patterns are extracted from anonymized CDRs made available for the D4D-Senegal challenge promoted by Orange and Sonatel. The model is parameterized with available epidemiological, demographic and socioeconomic data, and is calibrated against georeferenced records of urogenital schistosomiasis prevalence in the country. A specific aim of this work is thus to assess whether detailed information about human movement can alter the results that could be drawn from epidemiological studies in which mobility is simplified or neglected. The impact of some WASH (WAter, Sanitation and Hygiene) intervention strategies or IEC (Information, Education and Communication) campaigns is also evaluated. Finally, future research avenues are discussed, with the aim of showing how the framework applied within this project can be developed into a decision-support tool to help eradicate schistosomiasis from Senegal.

2 The model

2.1 A spatially explicit model for schistosomiasis dynamics

The human population is subdivided into $n$ communities (following e.g. administrative boundaries, health zones or geographical divides). Within each community $i$, the resident human population (of size $H_i$) is considered to be ‘stratified’, i.e. divided into different infection classes characterized by increasing parasite burden $p$ (from $p = 0$ to some maximum abundance $p = P$, Gurarie et al., 2010). Let $H^p_i$ be the abundance of individuals in community $i$ who host exactly $p$ parasites. Furthermore, let $S_i$ and $I_i$ be the densities of susceptible and infective snails in community $i$, and let $C_i$ and $M_i$ be the concentrations of cercariae and
miracidia in the freshwater resources accessible to community $i$. Schistosomiasis transmission dynamics can be described by the following set of coupled differential equations, which represents an extension of the classical Macdonald (1965) model (modified after Gurarie et al., 2010):

$$
\begin{align*}
\frac{dH_i^0}{dt} &= \mu_H (H_i - H_i^0) - F_i H_i^0 + \gamma H_i^1 \\
\frac{dH_i^p}{dt} &= F_i H_i^{p-1} - (\mu_H + \alpha^p_H + F_i + \gamma^p) H_i^p + \gamma^{p+1} H_i^{p+1} \quad [0 < p < P] \\
\frac{dH_i^P}{dt} &= F_i H_i^{P-1} - (\mu_H + \alpha^P_H + \gamma^P) H_i^P \\
\frac{dS_i}{dt} &= \mu_S (N_i - S_i) - bM_i S_i \\
\frac{dI_i}{dt} &= bM_i S_i - (\mu_S + \alpha_S) I_i \\
\frac{dC_i}{dt} &= \pi C_i - \mu_C C_i \\
\frac{dM_i}{dt} &= G_i - \mu_M M_i,
\end{align*}
$$

where $F_i = a \sum_{j=1}^n Q_{ij} \theta_j C_j$, $G_i = \pi_M \delta_i \sum_{j=1}^n Q_{ji} \frac{W_j}{\pi}$, and $W_j = \sum_{p=1}^P pH_j^p$. The first $n \cdot (P + 1)$ equations of model (1) describe the dynamics of human hosts, the following $2n$ the dynamics of intermediate snail hosts, the last $2n$ the dynamics of the larval stages of the parasite.

As for the dynamics of human hosts, $\mu_H$ is the baseline mortality rate, while $\mu_H H_i$ represents the total birth rate, here assumed to be constant (i.e. leading to a constant community size $H_i$ in the absence of disease-induced mortality). Human hosts progress from one infection class to the following because of exposure to water contaminated by cercariae. Specifically, $F_i$ is the force of infection for the inhabitants of community $i$: $Q = [Q_{ij}]$ is a row-stochastic matrix (that is a matrix in which rows sum to one) that describes the probability that a resident of community $i$ comes in contact with freshwater in community $j$ (possibly different from her/his home community as a result of human mobility), $\theta_j$ is the human exposure rate, i.e. the rate at which human hosts are exposed to contaminated freshwater in community $j$ (note that the exposure rate is assumed to be community-dependent to account for geographically heterogeneous access to safe water supplies), and $a$ is the probability that a cercaria successfully develops into a reproductive adult parasite following contact with a human host. The term $\gamma^p$ represents the parasite resolution rate, i.e. the transition rate from infection class $p$ to infection class $p - 1$ because of the death of one parasite, hence $\gamma^p = p\mu_P$, with $\mu_P$ being the per capita parasite mortality rate. Disease-related mortality in humans is accounted for by the term $\alpha^p_H$, which describes increasing mortality for increasing parasite burden ($\alpha^p_H = p\alpha_H$, where $\alpha_H$ is the additional mortality rate experienced by an infected host because of the presence of each adult parasite). As for the dynamics of snail hosts, $\mu_S$ is the baseline mortality rate, whereas $\mu_S N_i$ is the constant recruitment rate (local
population size in absence of disease is \(N_i\). The parameter \(b\) represents the exposure rate of susceptible snails to miracidia in the freshwater environment. We assume that exposure triggers a transition to the infective compartment (note, in fact, that possible delays between exposure and onset of infectivity are neglected here for the sake of simplicity). Infective snails suffer from an extra-mortality rate \(\alpha_S\). As for the dynamics of larval stages, cercariae are shed by infective snails at rate \(\pi_C\) and die at rate \(\mu_C\). Similarly, miracidia are shed by infected human hosts and die at rate \(\mu_M\); specifically, the total contamination rate for community \(i\) is \(G_i\), with \(\pi_M\) being the shedding rate of miracidia by infected humans, \(\delta_i\) the (possibly site-specific, because of local sanitation conditions) probability of contamination of accessible freshwater resources, and \(Q_{ji}\) the probability that an inhabitant of community \(j\) comes in contact with freshwater in community \(i\) (shedding is assumed to be proportional to the total number of adult parasite pairs \(W_j/2\) carried by the residents of human community \(j\)). Note that if \(Q\) is the identity matrix (no human mobility) system (1) reduces to a set of spatially disconnected local models.

It is useful to introduce some hypotheses that help simplify model analysis. By noting that the lifespan of the larval stages of the parasite is much shorter than those of the other biological agents involved in the transmission cycle of the disease (up to a few days vs. years; see e.g. [Colley et al., 2014], the concentrations of cercariae and miracidia can be considered at their equilibrium values (as obtained by setting \(dC_i/dt = 0\) and \(dM_i/dt = 0\)), thus considering the so-called slow-fast dynamics of the system. Also, as the dynamics of infection classes characterized by similar parasite burden are expected to be similar to each other, infection classes can be grouped into discrete infection levels \(k\) (Gurarie et al., 2010). Specifically, \(k = 0\) corresponds to parasite burden \(0 \leq p < p_1\), \(k = 1\) to \(p_1 \leq p < p_2\) and so on, up to the highest level \(k = K\) with parasite burden \(p_K \leq p \leq P\). For the sake of simplicity, it is convenient to assume regularly spaced infection levels of width \(\Delta\). If we also introduce the rescaled state variables

\[
\begin{align*}
    h_i^k &= \frac{H_i^k}{H_i}, & s_i &= \frac{S_i}{N_i}, & y_i &= \frac{I_i}{N_i},
\end{align*}
\]

which represent the prevalences of each human infection class and of susceptible/infected snails, system (1) introduced above can be written as

\[
\begin{align*}
    \frac{dh_0^i}{dt} &= \mu_H(1 - h_0^i) - F_i h_0^i + \gamma^1 h_1^i \\
    \frac{dh_k^i}{dt} &= F_i h_{k-1}^i - (\mu_H + \alpha_h^k + F_i + \gamma^k) h_k^i + \gamma^{k+1} h_{k+1}^i \quad [0 < k < K] \\
    \frac{dh_K^i}{dt} &= F_i h_{K-1}^i - (\mu_H + \alpha_h^K + \gamma^K) h_K^i \\
    \frac{ds_i}{dt} &= \mu_S(1 - s_i) - G_i s_i \\
    \frac{dy_i}{dt} &= G_i s_i - (\mu_S + \alpha_S) y_i,
\end{align*}
\]
with

\[ F_i = \sum_{j=1}^{n} Q_{ij} \beta_j N_j y_j, \quad \beta_j = \frac{\alpha}{\Delta \mu_C} \theta_j, \quad \gamma^k = \frac{p_k}{\Delta \mu_P}, \quad \alpha^k_H = p_k \alpha_H, \]

\[ G_i = \chi_i \sum_{j=1}^{n} Q_{ji} W_j, \quad \chi_i = \frac{b \pi M}{2 \mu_M} \delta_i, \quad W_j = H_j \sum_{k=1}^{K} p_k h^k_j. \]

2.2 Application to schistosomiasis dynamics in Senegal

Administrative boundaries and population distribution

The territory of Senegal is divided into 14 regions (first-level administrative units), each of which is further subdivided in departments (45 overall, second-level units). Finally, each department is divided into arrondissements (123 overall), which represent the third-level administrative units. In order to apply model (2) to study large-scale patterns of schistosomiasis dynamics in Senegal, human communities are identified with arrondissements. Note that arrondissements usually include several human settlements, yet we refrain from choosing smaller units for the sake of computational feasibility.

Population size for each arrondissement \((H_i, \text{Fig. 2A})\) is obtained from high-resolution population distribution maps available from the AfriPop project, which is part of the WorldPop project (data available online at [http://www.worldpop.org.uk/](http://www.worldpop.org.uk/)). Data include 2010 and 2014 estimates of population distribution with a spatial resolution of 30 arcsec (approx 100 m at the equator), and national totals adjusted to match United Nations estimates. The total number of people living in each arrondissement is thus computed by summing the 2014 population estimates of the grid squares that fall within the relevant administrative boundaries. Population-weighted centroids are also evaluated for each third-level administrative unit.

Human mobility

Human mobility in Senegal is estimated from the anonymized, low-resolution movement routes that have been released in the context of the D4D-Senegal challenge promoted by Orange and Sonatel (see [http://www.d4d.orange.com/en/home](http://www.d4d.orange.com/en/home)). Specifically, data consist of the trajectories at arrondissement level of about 150,000 randomly selected mobile phone users collected for one year, from January 1 to December 31, 2013. Each record in the dataset includes the user that made the call (anonymous identifier), and information about when (timestamp) and where (arrondissement) the call was initiated (de Montjoye et al., 2014). According to the definition given above, matrix \( Q = [Q_{ij}] \) represents the probability that people usually living in community (arrondissement) \( i \) come in contact with freshwater in community (arrondissement) \( j \) \((j = 1..123, \text{including } i)\). We assume that this probability is proportional to the time spent in arrondissement \( j \), and that the number of phone calls made by a user while being in
Figure 2: Data for model set-up. A) Population distribution at arrondissement level [number of inhabitants]. Black lines indicate regional borders. Regions are numbered as follows: 1–Dakar, 2–Thiès, 3–Diourbel, 4–Fatick, 5–Louga, 6–Kaolack, 7–Kaffrine, 8–Saint-Louis, 9–Kolda, 10–Sédhiou, 11–Ziguinchor, 12–Kédougou, 13–Tambacounda, 14–Matam. B) Human mobility fluxes [number of people]; the fluxes $\Phi_{ij}$ between any two arrondissements (say $i$ and $j$) are obtained as $\Phi_{ij} = H_i P_{ij}$. Only fluxes $\geq 100$ people are displayed as links between the relevant population centroids. C) Access to improved water sources [% of people with access]. D) Access to improved sanitation facilities [% of people with access]. E) People living in rural settings [%]. F) Prevalence of urogenital schistosomiasis according to the 1996 national survey (filled dots) or other georeferenced data sources (empty triangles; not used in this study) [% of infected people]. See text for technical details and data sources.
arrondissement $j$ is proportional to the time spent in that arrondissement. Therefore, the entries $Q_{ij}$ of matrix $Q$ are expected to be roughly proportional to the number of phone calls made by users usually living in arrondissement $i$ while being in arrondissement $j$.

To characterize human mobility patterns, first we use the data provided by Orange and Sonatel to identify the ‘home’ arrondissement for each anonymous user. Following a definition often used in the context of CDRs analysis (see e.g. Wesolowski et al. 2012, for a recent epidemiological application), we define home as the site (arrondissement) where most calls are made by a user during night hours (from 7pm to 7am) over the whole dataset (i.e. over a timespan of one year). If several arrondissements match this criterion home is randomly selected among the arrondissements that host most night calls for the user. Afterwards, for each arrondissement $i$, the number of calls made in arrondissement $j$ by users whose home site has been identified with $i$ is extracted from the dataset. This number, properly divided by the total number of calls made by users usually living in arrondissement $i$ (independently of location), represents an estimate of the entries of the mobility matrix (Fig. 2B).

Note that this definition of mobility matrix represents a time-averaged picture of the actual human movement patterns. However, the definitions of both home community and mobility matrix can be easily made time-varying. As an example, by looking at time horizons shorter than one year, it is possible to define a ‘monthly home’ (or even a ‘weekly home’) for each user. This in turn allows the analysis of migration patterns. Similarly, monthly/weekly/daily mobility patterns can be defined as well by looking at the spatial patterns of the calls made by each user over some defined time horizons (month/week/day). Migration and time-varying mobility matrices have been extracted from the data made available in the context of the D4D-Senegal challenge but not yet applied to the epidemiological model described above (but see Discussion).

**Water resources and sanitation conditions**

Use of model (2) requires the specification of the spatially heterogeneous parameters $\beta_i$ and $\chi_i$, which represent, respectively, the effective exposure and contamination rates for humans. Exposure and contamination are related to contact with environmental freshwaters. Communities that lack access to piped drinking water and/or improved sanitation, and that have to resort to unsafe water sources for their primary needs, are thus more prone (and more conducive) to schistosomiasis transmission (Rollinson et al. 2013; Ogden et al. 2014). Conversely, the availability of adequate water provisioning and sanitation infrastructures may represent an effective protection against schistosomiasis, as shown in a recent systematic review of available field data (Grimes et al. 2014). Georeferenced information about the use of improved water sources and sanitation facilities in Senegal (% of total population with access in 2012, Fig. 2CD) is available at department level through the Global Atlas of Helminth Infections (GAHI; data available
online at \url{http://www.thiswormyworld.org}, based on mapping and spatial analysis of cross-sectional survey data (Pullan et al., 2014). The available information can be spatially downscaled by assigning department data to the relevant arrondissements. Overall, 21% and 42% of the Senegalese people lack access to safe water supplies and improved sanitation, respectively.

Exposure and contamination rates can thus be expressed as

\[
\beta_i = \beta_0 (1 - \omega_i)^{\phi_{\beta}} \quad \text{and} \quad \chi_i = \chi_0 (1 - \sigma_i)^{\phi_{\chi}},
\]

respectively, where \(\beta_0\) and \(\chi_0\) are the maximum exposure/contamination rates, \(\omega_i\) and \(\sigma_i\) represent the fraction of individuals living in arrondissement \(i\) with access to water/sanitation, and \(\phi_{\beta}\) and \(\phi_{\chi}\) are two non-negative shape factors (Mari et al., 2012). As access to safe water and improved sanitation are highly correlated with each other (Pearson’s \(r = 0.71\)), and are also highly anti-correlated with the fraction \(\rho_i\) of the population living in rural areas (\(r = -0.58\) and \(r = -0.94\) for water and sanitation, respectively), which is also available from GAHI (Fig. 2E), an alternative formulation for the spatially heterogeneous parameters is

\[
\beta_i = \beta_0 \rho_i^{\phi_{\beta}} \quad \text{and} \quad \chi_i = \chi_0 \rho_i^{\phi_{\chi}}.
\]

We retain this latter definition for \(\beta_i\) and \(\chi_i\) and also set \(\phi_{\beta} = \phi_{\chi} = \phi\) for the sake of parameter parsimony.

Snail habitat

Model (2) also requires an estimate of the local densities of snail intermediate hosts (\(N_i\)). Country-scale malacological surveys (see e.g. Ndir, 2000) show that occurrence of snail species involved in schistosomiasis transmission is widespread in Senegal. However, the lack of quantitative data precludes the use of these descriptive surveys to inform the model about the spatial density of snail populations. The distribution of snail habitat can be linked to availability of environmental freshwater and suitable climatic conditions, and can be mapped via geo-statistical methods (see e.g. Stensgaard et al., 2013). As calibration and validation of such tools require considerable effort (and fall outside the scope of this work), we leave a more in-depth ecological characterization of snail habitat to future studies and assume \(N_i = N_0\) for all communities.

Epidemiological data and model calibration

Model calibration involves contrasting model outputs to available epidemiological data at a suitable spatial scale. Here we use urogenital schistosomiasis prevalence data collected during the national survey carried out in 1996 (Ndir et al., 1996; Ndir, 2000), which represents the most recent country-scale picture of the spatial distribution of the disease. Most survey data have been georeferenced by GAHI (Fig. 2F,
data available online at [http://www.thiswormyworld.org](http://www.thiswormyworld.org). Human prevalence records are assigned
to the arrondissement where they were obtained according to geographical coordinates. 61% of the
arrondissements have no prevalence records. For the others, prevalence is calculated as the mean value
of the available records. Prevalence data are then upscaled to departments/regions by assigning average
values (properly weighted according to the population size of each arrondissement) to second/first-level
administrative units. This aggregation procedure leads to 36% of the departments and 14% of the regions
having no prevalence data. We note that the representation of schistosomiasis prevalence provided by the
1996 national survey might not fully represent the current situation. Other data sources (also reported in
Fig. 2F, data available from GAHI; see [http://www.thiswormyworld.org/files/Senegal_references.pdf](http://www.thiswormyworld.org/files/Senegal_references.pdf) for data references) suggest in fact higher prevalence values in some regions, notably in Saint-Louis.

As schistosomiasis is endemic in Senegal, model outputs are evaluated by running system (2) up to
convergence to steady state \((O(100 \text{ years}))\) starting from an initial condition in which human communities
are set to be completely uninfected \((h_0^i = 1 \text{ and } h_k^i = 0 \text{ with } k > 0 \text{ in all arrondissements})\), while the
prevalence of infected snails is tentatively set to be 5% \((s_i = 0.95 \text{ and } y_i = 0.05 \text{ in all arrondissements})\).
Note that model (2) produces an estimate of the distribution of human hosts among infection classes
(and of the prevalence of susceptible/infected snails in each arrondissement). Comparing this output with
prevalence data requires the definition of an infection threshold in model (2). According to commonly
accepted biological evidence (reviewed in Gurarie et al., 2010), in fact, a minimum number of parasites
within a human host is required for pathogen reproduction to be effective and leading to a positive result
during epidemiological screenings. The infection threshold \((T)\) thus represents the minimum parasite
burden above which human hosts are considered to be infected. The prevalence \(u_i^M\) of infected human
hosts in each arrondissement can thus be evaluated as the sum of the prevalences of the infection classes
characterized by parasite burden larger than \(T\), i.e.

\[
    u_i^M = \sum_{k=k_T+1}^{K} h_i^k,
\]

where \(k_T\) is the lowest infection class with \(p_k \leq T\). These prevalence values can be easily upscaled to
departmental/regional scale via averaging (using arrondissement population sizes as weights).

Numerical simulations obviously also require the parameterization of model (2). Some of the param-
eters can be reliably estimated from the literature or from epidemiological/demographic records. Specifi-
cally, the baseline mortality rates of human hosts, snails and parasites can be evaluated as the inverse of
the average lifetimes of people in Senegal (61 years according to CIA, 2014 hence \(\mu_H = 4.5 \cdot 10^{-5} \text{ days}^{-1}\)),
snail intermediate hosts (about 1 year according to Feng et al., 2004 hence \(\mu_S = 2.7 \cdot 10^{-3} \text{ days}^{-1}\)) and
schistosomes (about 5 years according to Gryseels et al., 2006 hence \(\mu_P = 5.5 \cdot 10^{-4} \text{ days}^{-1}\)), respectively.
Parasite-induced mortality in human hosts is set to $\alpha_H = 1.1 \cdot 10^{-7}$ days$^{-1}$ parasite$^{-1}$ following a field study conducted in an endemic area of Sudan (Kheir et al., 1999), while the extra-mortality suffered by infected snails is set to $\alpha_S = 1.4 \cdot 10^{-2}$ days$^{-1}$ according to the observation that the lifespan of infected snails is about two months (Feng et al., 2004; Gryseels et al., 2006). As for parasite load in human hosts, we follow Gurarie et al. (2010) and consider a maximum burden of $P = 150$, discrete infection classes with a uniform width $\Delta = 10$ and a threshold for infection $T = 10$ (parasites). Therefore, the human population of each community is divided into 15 classes, with classes $k = 1..14$ being considered to be infected.

Conversely, numerical fitting is necessary to calibrate the remaining parameters, namely the baseline human exposure and contamination rates $\beta_0$ and $\chi_0$, the shape parameter $\phi$ and the snail density $N_0$. However, as $\beta_0$ and $N_0$ enter model (2) as a product, we actually calibrate the aggregate parameter $\beta'_0 = \beta_0 N_0$. To that end, we minimize the residual sum of squares of modeled vs. reported values of schistosomiasis prevalence in human communities at the spatial scale of interest, i.e.

$$RSS = \sum_x (u^D_x - u^M_x)^2,$$

in which $u^D_x$ is schistosomiasis prevalence in humans according to epidemiological data (1996 national survey only; other sources not considered because of heterogeneity in data collection) and $x$ is the index of the units within the administrative level being used for calibration. Parameter calibration is performed the the Nelder-Mead method (Lagarias et al., 1998).

2.3 Numerical results

Model (2) is able to reproduce the observed spatial patterns of urogenital schistosomiasis prevalence throughout Senegal (Fig. 3AB). Preliminary calibration runs have shown that the best calibration performances are obtained with coarse-grained prevalence data at the regional scale. All the results described in the reminder of this work thus refer to the model calibrated at this spatial scale. Model predictions are generally in good agreement with the available data at the regional scale (Fig. 3C, Pearson’s $r = 0.89$). Fit to data is obviously far from perfect, though, especially for the regions of Kédogou (12), Diourbel (3), Matam (14) and Fatick (4), where the model overestimates (Kédogou and Diourbel) or underestimates (Matam and Fatick) schistosomiasis prevalence by more than 7% (Fig. 3D). Note that the regions of Sédhiou (10) and Ziguinchor (11) are not included in this comparison, because the 1996 national survey does not include any georeferenced data pertaining to these regions.

The model can be used to evaluate the impact of human mobility on the spatial patterns of schistosomiasis prevalence. To that end, it is possible to contrast the best-fit model simulation shown in
Figure 3: Simulation of model (2) and comparison with epidemiological evidence. A) Regional prevalence of urogenital schistosomiasis [% of infected people] from the 1996 national survey. No data are available for the regions of Sédhiou (10) and Ziguinchor (11). B) Regional schistosomiasis prevalence [% of infected people] according to the best-fit model simulation. C) Quantitative agreement between recorded and simulated disease prevalence. D) Differences between simulated vs. recorded prevalence values (absolute differences sorted in decreasing order). Numbers in panels C and D refer to administrative regions as shown in panel A. Calibrated parameter values: $\beta_0' = 1.11 \cdot 10^{-3}$ [days$^{-1}$], $\chi_0 = 9.78 \cdot 10^{-4}$ [days$^{-1}$ parasites$^{-1}$], $\phi = 6.35$ [-]. See Methods for other parameter values and technical details.
Figure 4: Effects of human mobility on schistosomiasis dynamics. A) Differences in regional disease prevalence obtained without vs. with human mobility (absolute differences sorted in decreasing order). Numbers correspond to administrative regions as shown in panel B. B) Differences [%] in prevalence values obtained without vs. with human mobility at the scale of the arrondissements. Parameter values as in Fig. 3. To exclude the effects of human mobility Q has been set as the identity matrix.

The comparison shows that in the absence of human mobility schistosomiasis prevalence at the regional spatial scale is generally expected to increase (Fig. 4A). Differences in schistosomiasis prevalence that can be ascribed to human mobility are remarkable in several regions, e.g. exceeding 5% in Sédhiou (10) and Tambacounda (13). At a finer spatial scale mobility can either increase or decrease schistosomiasis prevalence. Specifically, mobility is expected to increase disease prevalence in the northwestern part of the country and in some arrondissements in the South, and decrease prevalence elsewhere (Fig. 4B).

3 The fight against schistosomiasis in Senegal: analysis of possible intervention strategies

Starting from the best-fit simulation shown in Fig. 3, epidemiological model (2) can be used to evaluate the effects of interventions strategies aimed at reducing schistosomiasis prevalence through improved access to safe water and sanitation (WASH projects; see Ogden et al., 2014, for some guidelines concerning the prevention of schistosomiasis and other neglected tropical diseases in Senegal), or promotion of hygiene and awareness (IEC campaigns, see e.g. Rollinson et al., 2013).

The effect of structural actions aimed at increasing access to safe drinking water and improved sanitation can be modeled as an increase of \( \omega_i \) and/or \( \sigma_i \) at the community level. In our simplified formulation of the spatially heterogeneous exposure and contamination rates (see again system (2)), this is equivalent to decreasing the fraction \( \rho_i \) of residents of community \( i \) that lives in rural conditions, at least as far as water and sanitation are concerned. These actions can be either targeted (i.e. implemented only in prevalently
rural communities) or untargeted (implemented in all communities). Let \( \tau \) and \( \eta \) be the planned extent of the interventions (evaluated as the number of potentially benefited people) and their supposed efficiency (i.e. the probability of success per effort unit). Untargeted actions can thus be described in the model as

\[
\rho'_i = \rho_i \left[ 1 - \eta \max \left( 1, \frac{\tau}{\sum_j H_j \rho_j} \right) \right],
\]

where \( \rho'_i \) represents the fraction of people in community \( i \) with no access to safe water supplies and improved sanitation after action implementation. Note that we assume that interventions are deployed in each community in a way that is proportional to the need of the community, evaluated as the fraction of people living in a rural context. Targeted interventions can be implemented in the model by sorting communities (according to some suitable criterion), selecting those for which \( \sum_i H_i \rho_i \leq \tau \) and setting \( \rho'_i = 1 - \eta \) therein. A natural sorting criterion is access to water/sanitation, but other options (like e.g. prioritizing communities with large inbound mobility fluxes) are obviously possible.

Concerning the implementation of WASH interventions, the model suggests that, all else being equal, widespread actions are generally more effective in decreasing country-wide average schistosomiasis prevalence, while actions targeted to rural communities where access to water and sanitation is lowest are more effective in decreasing maximum disease prevalence at the regional scale if the extent of the intervention is small. At least two million people should be involved in WASH improvement to eradicate schistosomiasis from Senegal – which would save about 1.5 million cases (Fig. 5ABC).

As for IEC actions aimed at promoting hygiene and increasing awareness about disease transmission pathways, their effect can be modeled as a decrease of the baseline exposure/contamination rates (from \( \beta_0 \) to \( \beta'_0 \) and from \( \chi_0 \) to \( \chi'_0 \)). Such interventions can be, again, targeted or untargeted. Untargeted interventions can be modeled as

\[
\beta'_0 = \beta_0 \left[ 1 - \eta \min \left( 1, \frac{\tau}{\sum_i H_i} \right) \right] \quad \text{and} \quad \chi'_0 = \chi_0 \left[ 1 - \eta \min \left( 1, \frac{\tau}{\sum_i H_i} \right) \right],
\]

whereas the implementation of targeted interventions requires sorting the communities, selecting those for which \( \sum_i H_i \leq \tau \), and setting \( \beta'_0 = 1 - \eta \) and \( \chi'_0 = 1 - \eta \) therein. Sorting criteria can prioritize, for instance, lack of access to safe water sources or sanitation facilities (possibly subsumed by the fraction of people living in rural settings) or high values of schistosomiasis prevalence.

According to the model, IEC campaigns targeted to rural communities are predicted to be more effective than untargeted ones. Prioritizing high-prevalence communities represents the best option to decrease average/maximum schistosomiasis prevalence only if the extent of the planned interventions is limited (Fig. 5DEF).
Figure 5: Evaluation of large-scale control strategies. A) Effects of WASH actions aimed at improving community access to safe water and sanitation on population-weighted average schistosomiasis prevalence. B) As in panel A, but for maximum regional schistosomiasis prevalence. C) As in panel A, but for the total number of avoided cases. DEF) As in ABC, but for IEC campaigns aimed at increasing hygiene and awareness. Targeted interventions prioritize communities where the fraction of resident population leaving in rural setting (blue) or schistosomiasis prevalence (green) is highest. Unspecified parameters as in Fig. 3.
4 Discussion

In this work we have proposed a model to study country-scale dynamics of schistosomiasis transmission in Senegal making use of the mobile phone data made available in the context of the D4D-Senegal challenge promoted by Orange and Sonatel. Despite its relatively simple structure, the model is able to reproduce large-scale patterns of schistosomiasis prevalence throughout the country quite reliably (Fig. 3) and can thus be used to study the effects of human mobility on disease dynamics, as well as to evaluate possible intervention strategies aimed at decreasing the burden of the disease.

Our results show that accounting for human mobility is crucial for an accurate reproduction of the observed spatial patterns of schistosomiasis prevalence (Fig. 4). At a regional spatial scale, the model surprisingly predicts that mobility may predominantly offer protection from exposure to the causative agent of the disease (Fig. 4). This finding can be explained by the fact that the largest mobility fluxes are attracted by the most populated and urbanized regions (Dakar, Thiès and Diourbel), where access to safe drinking water and improved sanitation are widespread – hence where schistosomiasis transmission is expected (and indeed found) to be low (Fig. 2). It is important to remark that field studies have often reported a positive relationship between human mobility and transmission emergence, parasite burden and disease spread (see Gurarie and Seto, 2009; Remais, 2010, and references therein). However, all those studies looked at rural, highly endemic sites, from where schistosomiasis could spill over to neighbouring areas because of human mobility and migration. Although this effect is accounted for in the model, at a regional spatial scale it is most likely clouded by the geometry of the mobility matrix, and by the heterogeneous spatial distribution of water supply and sanitation in Senegal. Model predictions at the spatial scale of the arrondissements may be less robust than regional projections, but show that mobility can also locally increase disease prevalence. We argue that high-resolution models targeted to specific regions of the country could elucidate the actual role of human mobility at different spatial scales, and help identify the focal hotspots of disease transmission.

High-resolution models would also require a more-in-depth look at the sources of complexity that are involved in the transmission cycle of the disease, and that have been neglected at present. As an example, human exposure and contamination are directly related to water contact patterns, which in turn are linked to demography and social structure. Including a simple, yet realistic, demographic model able to describe the age structure of the population at risk, and to track intra- and inter-annual changes in local population abundance would greatly improve the reliability of epidemiological projections (see e.g. Gurarie et al., 2010). While census microdata could be used to describe long-term mobility trends (Garcia et al., 2014), CDRs can be exploited to derive short-term migration patterns and/or time-varying mobility fluxes. A closer look at the connectivity matrices derived from CDRs (Fig. 6) shows in fact that human movement
is highly heterogeneous, not only in space but also in time. Overall mobility, evaluated as the fraction of people that leave their home community, displays clear weekly patterns (Dakar region), longer-term trends (possibly linked to seasonal economic activities, such as agriculture and fishing) and sudden peaks, with a space-time average of 27% of mobile people (Fig. 6A). Religious practice can produce remarkable mobility fluxes and the temporary displacement of hundreds of thousands of people, as in the case of the Grand Magal de Touba or of Kazu Rajab (also held in Touba, Fig. 6B). These religious gatherings attract pilgrims from all regions of Senegal, as shown in Fig. 6C, where daily mobility fluxes from the arrondissements included in the region of Saint-Louis are reported. The sustained mobility fluxes to the regions of Dakar (1), Louga (5) and Matam (14), all clearly visible in Fig. 6C, show the ‘gravitational’ nature of human mobility: the largest fluxes are directed towards the most ‘attractive’ region (Dakar, home to the homonymous capital city of Senegal) or to the closest ones (the neighboring Louga and Matam). Fig. 6C also remarks that temporal variability and seasonality are important components of human mobility fluxes. Therefore, these features need to be incorporated in future versions of the model that will resolve schistosomiasis dynamics at finer spatiotemporal resolution.

From a biological perspective, the ecology of the intermediate host has yet to be integrated in our modeling framework, specifically to account for the spatiotemporal variability of the environmental drivers that influence the distribution and abundance of snails, most notably water temperature and rainfall (Woolhouse and Chandiwana 1990a,b). In the absence of detailed country-scale malacological surveys, field evidence collected in other sub-Saharan countries (e.g. Poda et al. 1994) and geo-statistical modeling (Stensgaard et al. 2013) can be used to characterize Bulinus population dynamics. Particular attention should be devoted to studying the interplay between the seasonality of environmental signals and time-varying human mobility, which could induce non-trivial effects on disease transmission. Integrating the ecology of the intermediate snail host into the modeling framework described here is also crucial to planning and optimizing non-conventional intervention strategies based on biological control, e.g. as proposed in the projects “Aquaculture pour la santé: native prawn fisheries restoration for poverty alleviation and schistosomiasis control in the Senegal river basin” (Bill & Melinda Gates Foundation, principal investigator Giulio De Leo) and “Healthy ecosystems, healthy people: the coupled human health and environmental dynamics of schistosomiasis in sub-Saharan Africa” (US National Science Foundation, principal investigator Susanne Sokolow). One objective of these projects is to restore a native prawn species (namely Macrobrachium vollenhovenii) that has nowadays virtually disappeared from Senegal because of anthropogenic human alterations, namely the construction of the Diama dam in the 1980’s. M. vollenhovenii is a voracious snail predator, whose feeding activity can permanently interrupt disease transmission by suppressing the intermediate host population (Sokolow et al. 2014). Goal of these projects is also to achieve schistosomiasis eradication in a sustainable way, thanks to village-based prawn fishery (Alkalay).
Figure 6: A closer, time-explicit look at human mobility. A) Overall mobility, evaluated as \((1 - Q_{ii})\), \(i = 1..123\). Region labels (1..14) are shown for easier visual reference. B) Mobility fluxes to Touba (Ndame arrondissement, Diourbel region), evaluated as \(\sum_{i} \Phi_{i,Ndame}\), \(i = 1..123\). Peaks correspond to the most important religious gatherings held in Touba during 2013. C) Mobility fluxes from Saint-Louis region (8), evaluated as \(\sum_{i} \Phi_{i,j}\) with \(i \in \text{Saint-Louis}, j = 1..123\). Highlighted are religious gatherings that generate peak mobility fluxes (Grand Magal de Touba, GMdT; Gamou de Tivaouane; Kazu Rajab) and within-region mobility.
et al., 2014).

When building a fine-scale account of the ecological interactions that are relevant to schistosomiasis transmission, hydrological dispersal of the snail intermediate hosts, as well as of the larval stages of the parasite, has to be accounted for (see Gurarie and Seto 2009; Remais 2010 and references therein). This would require a detailed description of hydrological connectivity at a fine spatial scale. This analysis would also allow studying the effects of agricultural development, which requires the implementation of irrigation schemes and the construction of dam reservoirs. These interventions, in turn, can induce severe perturbations of the natural matrix that influences the population dynamics of snails and their natural enemies (see e.g. Poda et al. 2003, 2004; Steinmann et al. 2006; Li et al. 2007). As an example, the development of irrigation channels following the construction of the Diama dam resulted in increased transmission of *S. haematobium* and the introduction of *S. mansoni* in villages upriver of the dam, with a globally unprecedented velocity of transmission (Talla et al. 1990; Picquet et al. 1996). These observations highlight the need to explicitly address the inherent conflict between water resources development and schistosomiasis management (Steinmann et al. 2006).

On another side of water resources development, our results show that increasing access to safe water supplies and improved sanitation provides an effective way to reduce the burden of schistosomiasis (Fig. 5). While not surprising, this finding remarks the importance of a comprehensive approach to disease control, based not only on mass chemotherapy (Thériot-Laurent et al. 2013; Colley et al. 2014) but also on human development, specifically with programs for the improvement of life conditions (especially in rural communities), education campaigns aimed at promoting hygiene and awareness about the relevant risk factors, and transmission control (Rollinson et al. 2013).

Although preliminary, our study suggests that it is indeed possible to transform the modeling framework presented here into a support tool to help decision makers in the design of effective plans for schistosomiasis management, and in the optimization of sanitary and humanitarian efforts. Such a decision-support system should be able to accommodate real-time data assimilation (epidemiological reports, ecological surveys, demographic updates), as well as reliable projections of the relevant environmental drivers, such as temperature and rainfall. A well-established framework for real-time forecasting and decision making is represented by adaptive management (Allan and Stankey 2009). Adaptive management is an iterative process of robust decision making aimed at reducing uncertainty over time via system monitoring. Specifically, real-time information assimilation allows for the improvement of model forecasts and the evaluation of alternative interventions strategies, ideally in a multicriterial sense (Belton and Stewart 2002). In the adaptive management framework, predictions and decisions drawn from the model allow decision makers to gather additional information on the behavior of the system, which in turn further improves future forecasting and management practice, and helps identify possible knowledge gaps. To increase the robust-
ness of this iterative learning process, model simulations can be set in a Bayesian framework, in which not only optimal model trajectories but also their related uncertainties are estimated (see e.g. Gilks et al., 1995).

Achieving greater detail in the description of epidemiological dynamics, human mobility, ecological interactions, water resources development and interventions plans is likely to be unfeasible at the country scale, but becomes possible (and meaningful) when looking at smaller spatial scales (e.g. specific regions of Senegal), at which the underlying modeling hypotheses can be substantiated by knowledge gathered in situ, possibly with the help of local institutions. The lessons learned from local experiences could then be scaled up to define country-scale strategies to eradicate schistosomiasis from Senegal.

Acknowledgements

Anonymous mobile phone data have been made available by Orange and Sonatel within the framework of the D4D-Senegal challenge. The authors are also grateful to Andrea Rinaldo, Enrico Bertuzzo and Francisco Javier Perez-Saez (Laboratoire d’Ecohydrologie, École Polytechnique Fédérale de Lausanne) for stimulating discussions on schistosomiasis modeling.

References


24


1 Introduction

Tuberculosis (TB) is a global emergency responsible for 2 million deaths every year and it is one of the leading death causes among the curable infectious diseases. Specifically, Africa is facing the worst TB epidemic since the advent of antibiotic era. Driven by a generalized HIV epidemic and by weak health care systems and conditions that promote transmission of infection, this devastating situation has steadily worsened. Moreover, the World Bank estimates that loss of productivity attributable to TB is 4 to 7 percent of some countries’ GDP. Recently, several works have shown that the spatial spread of infectious diseases is affected by human mobility patterns [1], which can be quantified using data from mobile phones [2]. The importance of spatial aspects of TB transmission for optimizing resources dedicated to control has been recently recognized [3] but not sufficiently investigated. Here, we investigate the use of Call Detail Records (CDR) provided by Sonatel and Orange within the innovation challenge ”Data for Development Senegal” (D4D) to identify human mobility patterns in Senegal and evaluate their impact on the spatial epidemiology of TB through a mathematical model.

2 Methods

2.1 Demographic structure

The total surface of Senegal was divided in 9469 geolocated square cells (with a surface of about 20 km² each). Each cell was assigned to one of the 123 districts (“arrondissements”, in French) and to one of the 14 regions of Senegal, based on its coordinates. The population of each cell was initially assigned according to publicly available estimates on population density [4] and in such a way to obtain a total population and an age structure equal to that of Senegal in 1992. For the following years, the population is subject to birth and age-specific death rates from the Demographic and Health Survey program [5], so that the evolution of the total population and of the age structure reproduces the one observed by successive censuses. Spatial transmission from an origin district to a destination district occurs on the basis of a probability estimated from mobility data.

2.2 Mobility data

We exploit the dataset #3 made available by the D4D challenge organizers, encompassing one full year of Call Detail Records data produced by ~150000 users. As visible in the histogram depicted in Figure 1 minimum activity in the data under analysis falls, as could be expected, early in the morning, precisely between 04:00 and 05:00.

From CDRs we extract between-district mobility probabilities in the form of weighted adjacency matrices, using the following methodology. First, following the pattern visible in Figure 1, we define days as starting at 05:00 and ending at 04:59, under the assumption that the time range with minimum mobile network activity represents the time by which most of the subjects are sleeping at home. Then, for each day and each subject, we define a reference district as the one where the subject was at the time of his last daily communication. Finally, under the constraint that at least 5 daily communication events are present for the subject under analysis, we compute the percentage of time spent in the daily reference district
and in the others: the former compounds to the probability of staying within the reference district, while the latter compounds to the probability of moving between the reference district and each of the others. In the case of two consecutive CDRs happened in two different districts, hence representing a transition, we define time spent in each district as half the interval occurring between the two events.

We compare TB epidemiology predicted using this mobility matrix derived from CDRs with that predicted under two alternative hypotheses on spatial mixing: completely assortative mixing, i.e. infectious individuals can only transmit to members of the same district, thereby implying that mobility is irrelevant for transmission; and completely homogeneous mixing, which assumes that the probability of transmission is independent of the geographic location of infected individuals.

Figure 1. Histogram of the events per hour present in Dataset #3.

2.3 Epidemiological model

The epidemiological model is based on a previously published formulation [6], expanded to account for the increased TB risk in HIV infected individuals following recent approaches [7] [8]. All individuals are born TB-susceptible and upon infection enter a recent-infection state from which they can develop disease within 5 years (fast progression) with an age-dependent probability and a risk that decreases with time since infection [9]. Infected individuals can also progress to a chronic, latent TB infection (LTBI) state. Individuals with LTBI have a small, age-dependent risk of developing endogenous reactivation throughout their lifetime. LTBI also confers a partial protection from the risk of developing TB upon re-infection. Individuals with TB disease are assigned a different level of infectiousness, representing their smear status, at the moment of disease development. Individuals with TB disease are exposed to a TB-specific death rate and can be diagnosed and treated with a rate that increases over time, proportionally to the performance of the national TB programme (measured by the average case detection rate [10]). Treated individuals can relapse to TB disease at a given rate, and are assumed to maintain LTBI, thereby having the same risk of endogenous reactivation. HIV infection can occur in individuals between ages 15 and 49 according to the incidence rate provided by the UNAIDS program [11]. HIV-infected individuals progress with time over three stages of increasingly severe HIV infection, and have a stage-dependent increased
risk of progressing to TB upon recent infection, increased risk of reactivation, reduced protection from latent infection, increased all-cause death rates, increased TB-specific death rate, and decreased risk of positive smear status. HIV infected individuals have a stage-specific probability of being diagnosed and set on antiretroviral therapy (ART). ART prevents progression to further HIV stages and mitigates the increase of risks related to coinfection with TB and of all-cause mortality. Individuals diagnosed with TB can also be tested and diagnosed for HIV infection and set on ART according to probabilities given by national control programs [10].

2.4 Model calibration

The model is brought to epidemic equilibrium while keeping population size and age structure equal to that of Senegal in 1992. For the following years, the population is subject to demographic and epidemiological changes (e.g. changing HIV incidence and TB detection rates over time). The 10 free model parameters were calibrated using data on TB prevalence, incidence and mortality in the period 2008-2012 from the World Health Organization (WHO) for Senegal [12]. 10,000 combinations of parameter sets were sampled using Latin Hypercube Sampling [13] and model predictions for each parameter set were compared to corresponding data using the negative of the mean relative squared error, which allows comparison across quantities of different magnitude. The parameter set corresponding to the lowest error was used for simulations throughout the rest of this work.

3 Results

Figure 2 shows the mobility matrix calculated by the algorithm specified in section 2.2. Each pixel represents the fraction of time spent by an average individual belonging to an origin district in each destination district. The grid displayed in white separates districts belonging to different regions, for ease of interpretation. According to our data, individuals spend the overwhelming majority of their time within the district of belonging (range 73-98%), with basically all red pixels placed on the matrix diagonal. Most of the remaining time is spent in districts from the same region (orange pixels clustered in square blocks across the diagonal), with only some sparse exception between districts belonging to neighboring regions (e.g. between Thies and Diourbel, see map of Senegal on the right). Finally, the blue stripe at the bottom of the matrix indicates a certain amount of movements towards the region of the capital Dakar, especially from regions located at smaller geographical distance (Thies, Diourbel, Fatick, see map on the right).

Figure 3 reports the outcome of the fitting procedure. The model reproduces WHO estimates on the relative TB prevalence, incidence and mortality for the total Senegal population [12], with substantially all model predictions falling within the confidence intervals of the data.

Figure 4 compares regional TB notification rates in 2012, as provided by the Ministry of Health of Senegal [14], with corresponding model predictions, which account for region-specific case detection rates as estimated by the Senegal national TB programme [10]. Model predictions are quantitatively coherent for most regions, and at first sight this seems independent on the geographical mixing hypothesis. However, when we compare the Spearman rank correlation coefficient between model predictions and corresponding data (Table 1), the homogeneous mixing model appears uncapable of correctly identifying the relative importance of TB across regions. No significant difference (p-value=0.7133) can be found between the mobility and assortative model, coherently with the substantially assortative aspect of the mobility matrix with respect to interregional mobility. This implies that regional differences in TB notification rates are mainly a consequence of region-specific demography (population size, age structure) and access to healthcare systems (implicit in the different notification rates).

At a sub-regional scale, the mobility matrix differs significantly from the completely assortative model. We were not able to find appropriate epidemiological data at a district level to test whether the mobility
matrix reconstructs local differences in TB incidence better than the assortative mixing mode. However, we compared the difference in predictions between the two mixing models to show the possible impact of mobility at the district scale. Figure 5a shows a scatterplot of predictions by the two models (assortative on x, mobility on y) for each district, color-coded by region and with each point size proportional to the district population. While, as expected, the two predictions are strongly correlated, there are examples of significant variation, and we show that these are strongly correlated with mobility. For each district, we compute the rate of incoming individuals as the cumulative time spent in a given destination district by individuals belonging to all other origin districts, normalized by the population of the destination district;
Table 1. Spearman rank correlation between observed and predicted region-specific notification rates under different mobility hypotheses.

<table>
<thead>
<tr>
<th>Mixing model</th>
<th>Spearman’s $\rho$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobility</td>
<td>0.57</td>
<td>0.03663</td>
</tr>
<tr>
<td>Assortative</td>
<td>0.60</td>
<td>0.02617</td>
</tr>
<tr>
<td>Homogeneous</td>
<td>0.39</td>
<td>0.1782</td>
</tr>
</tbody>
</table>

Figure 4. Comparison between observed and predicted regional notification rates under different mobility hypotheses.

Similarly, the rate of outgoing individuals is calculated for each origin district as the cumulative time spent in all other destination districts by individuals belonging to the origin district, normalized to the district’s population. In figure 5a, the five districts with top rates of incoming individuals are highlighted with a diamond around the point, and the five districts with top rates of outgoing individuals are highlighted with a circle: the mobility model systematically overestimates the former five and underestimates the latter five with respect to an assortative mixing model. Panels b and c of Figure 5 confirm the same trend for all districts, by representing on a semi-log scale the percentage difference between district-specific incidences predicted by the two models against the rates of incoming and outgoing individuals respectively. Table 2 shows that the correlations between incoming and outgoing rates and the difference in model predictions are statistically significant.

Table 2. Spearman rank correlation between mobility rates and percent difference in model-predicted district-specific incidence (mobility vs. assortative).

<table>
<thead>
<tr>
<th>Mobility rate</th>
<th>Spearman’s $\rho$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incoming</td>
<td>-0.39</td>
<td>$10^{-5}$</td>
</tr>
<tr>
<td>Outgoing</td>
<td>0.49</td>
<td>$1.8 \cdot 10^{-8}$</td>
</tr>
</tbody>
</table>
4 Conclusions

Our elaborations on CDR data shows that individuals in Senegal spend the majority of their time in the district they belong to, and most of the remaining time in districts within the same region. For this reason, our TB transmission model was substantially equivalent to a model with completely assortative mixing (i.e., no mobility) when reproducing regional data. However, significant differences emerge between a completely assortative model and one that integrates mobility data in estimates of the district-specific incidence. In particular, in the model with mobility the amount of incoming individuals in a district tended to increase the local incidence, and the amount of outgoing individuals tended to decrease it. In summary, mobility in Senegal does not play a significant role in shaping regional differences in TB incidence. However, spatial transmission may have important effects at a sub-regional scale, especially for optimizing the spatial distribution of healthcare facilities, which can contribute to the reduction of TB burden through a higher accessibility to health services, with a consequent reduction in diagnostic times [15] and increase in adherence to treatment [16]. A more thorough, data-based assessment of spatial effects can also be important to identify possible hotspot areas, where the concentration of control efforts may have major benefits at a more global scale [3].

References


Figure 5. Differences in district-specific predicted incidence between mobility and assortative mixing.
Individual-based modelling of contact networks of epidemic diffusion using real-life data

Yi Yu, Gaoxi Xiao, Winson Peng and Rich Ling

Nanyang Technological University

(Dated: December 30th, 2014)

Epidemic outbreaks are obvious threats to human health. This is seen in historical examples such as the bubonic plague [1, 2], the 1918 flu pandemic [3-5], and more recently in the 2003 SARS outbreak [6-8] and the Avian Influenza [9-12]. Such episodes can affect the socio-political structure of a country. The case of Ebola in West Africa is a clear global health emergency [13-17]. This work will facilitate the disposition of resources to better understand the way that such a disease spreads.

Traditionally, studies of epidemic diffusion have used network based models [18-23]. In order to be accurate, these require in-depth insight into and accurate modeling of complex mobility and contact patterns. An alternative to this is agent-based modeling [24-30]. This alternative does not rely on sophisticated models of mobility/contact patterns. However it relies on large amounts of real-life data and it demands large-scale numerical simulations. The rise of so called big data and advanced analytic techniques, however means that examination of large scale data bases is becoming more realistic [31-35]. For example, the D4D mobile call data from Senegal gives us a rich data source, albeit in a country that has not experienced a serious outbreak of Ebola.

Developing accurate analytical models of human mobility/contact is theoretically possible but it is challenging, not least because of the correlation between mobility and contact patterns. For their part, the calculations of agent-based simulations for large populations, such as that of Senegal are highly time-consuming. In the work for the D4D challenge, we have developed a data-driven individual-based network modeling (DIN) approach. This has been constructed by using a contact network model relying on the D4D mobile communication call data records (CDRs). The DIN model requires no mathematical analysis or human mobility/contact modelling since this is empirically derived from the CDRs. We feel that the DIN approach provides for an expedient and reliable approach to modeling epidemic spreading since it opens for the possibility of using large empirical databases in the understanding of epidemiology.

The development of the DIN approach relies on the fact that people mostly stay within a small radius and that their travel patterns are repetitive [36]. This means that it is possible to map the spread of an epidemic using real-life data within a limited time frame. The two-week duration of the D4D data set (data set number 2) offers the option of working out such a contact network. Using the antenna-based data; we follow anonymized individuals as they share a particular location with others in ten-minute “time steps.” Their movement and co-location with others is
traced over two weeks. This is used to calculate the potential for contamination between the two persons. That is, we can derive the likelihood of disease transmission and its eventual progression. It should be noted that the model can be adjusted to describe the process with more complicated diffusion trajectories including different incubation periods. Further, if more fine-grained data becomes available, that can also be used to refine the model.

To validate the model we used material from the D4D dataset [37] to compare a purely data-driven agent-based simulation with an agent-based epidemic simulation in bi-week periods 1, 4, 7, 10, 13, 16, 19, 22 from the same dataset. In this case we used approximately 3% (about 10,000 individuals) of the population. Thus we used 40 samples to simulate the spread of infection.

To benchmark this, we assume that at every time step, the infected people have a probability $p$ of infecting others at the same location. One hundred users are randomly assigned to be the initial carriers in both the agent-based and contact network infection simulations. For a particular infected person in the sample, the average results from 100 rounds originating from him are considered the expected infection size and the average of the expected infection sizes of the 100 randomly selected carriers is the average infection size of this sample. Our analysis tested different infection rates ($p = 10^{-4}, 1.5 \times 10^{-4}, 2 \times 10^{-4}, 2.5 \times 10^{-4}$ and $3 \times 10^{-4}$ respectively).

Figure 1 shows that the results of the network-based simulation match those of the agent-based simulation. When considering early epidemic prediction, i.e. the most important window for addressing the spread of a disease, the proposed method gives accurate results of infection size. Applying the models to the actual CDRs for various regions shows the potential for contagion by district (See Figure 2). Clearly the granularity of this analysis can be enhanced through using data from smaller geographic areas.

We proposed an easily applied data-driven individual-based contact network approach with which to model the spread of contagious disease. We use the data from the D4D challenge to:

- evaluate the eventual effect of the disease when originating from different regions in Senegal, and,
- the speed with which the disease could spread to different regions.

The DIN approach can be quickly developed with no theoretical/statistical analysis of human mobility/contact patterns; rather it relies on empirical mobility patterns. Our analysis shows that the contact network based simulation matches the more computing intensive agent-based simulation results. Further the DIN approach has the flexibility with which to include various other real-life data as well as further refinement by using more fine-grained data. In the work we have also evaluated different scenarios of possible infection spreading in Senegal. We believe that this work can help control infection spreading in a variety of contexts.
Future work will obviously include the use of more parameters, such as latency periods and smaller geographical areas. The contact network model can also be verified if and when it is possible to follow actual infection sources, and also to include the effect of quarantine and also vaccination as they change the probability of contagion.

It is possible to conceptualize an application of the DIN approach in the development of easy to use mapping applications that will allow health care authorities to first map the actual points of infection and from there examine the likelihood that other locations will be infected. This will allow for better apportionment of health care resources so as to anticipate the trajectory of contamination.


[34] Mayer-Schönberger, V. Big data: a revolution that will transform how we live, work and think. (John Murray Publishers, 2013).


Figure 1 | **Mean infection size from random sources vs. infection rate $p$.** Error bars show the deviation for the 40 samples.
Figure 2 Mean infection size at different sites and mean time the first residents of a certain region get infected when the infection source is Ziguinchor. In the figure, besides each district name, $S=($,) includes the infection size of this region after agent-based simulation and contact network methods during the two weeks, respectively. $T=($,) includes the average time that the first case happens on the residents of that region.
Detecting Anomalies and Supporting Community 
To Ensure Healthy Society 

D4D 2014 Project Report 

December 2014 

Chiu, Tengchen, Zhenlie Han, Akamatsu Naoki, Ying Pu, Masanori Sueno, 
Ryosuke Takeuchi, Shumin Liu 
Graduate School of Business and Commerce, Keio University 

Yutaka Hamaoka 
Faculty of Business and Commerce, Keio University
Abstract

Putnam(2000) found benefits of social capital in many fields including education, economic growth, neighbor safety, democracy, and health. To realize decentralized-social capital based health care approach, two important problems must be solved. 1) How community should be organized. 2) How community should respond to anomalies, such as epidemiologic outbreak, natural disasters. This project tried to answer these questions using mobile phone call data.

For anomaly detection, anomaly was defined as deviation from regular call volume. For each antenna, VAR model was applied with lagged dependent variables and exogenous variables to estimate regular pattern. Although, explanatory variables are limited to holiday, month, and hour dummy that are easily defined, model fit was satisfactory. Regional weather could be affect call pattern, however, that is available only 11 locations. Analysis with minimum variables is effective strategy for less developed countries. Based on direction and regional distribution of anomalies, we classified them into four patterns. And we found "decreased" anomaly dominates that indicates natural disaster or technological problem damages infrastructure. Based on geographic distribution, we found anomalies at rural regions are frequent. Although, we cannot get enough information on cause of anomalies, support for rural area is urgent.

For organizing community, we classified antennas based on inter-antenna call volume. Detected community expands beyond administrative wards. As we mentioned earlier, inter-antenna call volume can be utilized as proxy of (inverse of) communication cost. Detected community can be utilized for the unit of region to community based decentralized health care system.

Keywords: Social capital, Anomaly detection, Community detection
1. Introduction


To ensure health of society, two approaches are possible: centralized and decentralized approach. Although centralized approach focuses large hospitals at the central location(s), large amount of resources are necessary to build and keep the system. For people at rural areas, access to the hospital is also problematic. On the contrary, decentralized approach that focus social capital: mutual help, support, and learning could achieve the same goal with low cost. A significant example in finance and community development is Graeme Bank (Yunus and Jolis 2003): microfinance banking system "based on mutual trust, accountability, participation and creativity "\(^1\).

To realize decentralized approach, two important problems must be solved. 1) How community should be organized. 2) How community should respond to anomalies, such as epidemiologic outbreak, natural disasters. Purposes of this project are to answer these questions using mobile phone call data.

2. Data

Call data

Call Detail Records (CDR) in Senegal from January 2013 to December 2013 are provided by Orange for the Data for Development (D4D) Challenge (de Montjoye et al. 2014)\(^2\). Among provided data sets, data set 1: the numbers of calls among 1666 towers are analyzed. Origin and destination of cell towers are identified by their longitude and latitude locations within Senegal; the Orange D4D Challenge also provides location information.

For 52 towers, no call was recorded and for 8 towers, data point was not enough, less than 48, to estimate lagged trend model. Thus our analysis includes 1614 towers. Data was processed using the R statistical software.

Event data

Information on holiday, natural disasters, political events were collected from Japanese embassy in Senegal\(^3\), that of US.

---


\(^2\) [Homepage of "Data for Development (D4D)" project](http://www.d4d.orange.com/en/home) (accessed 10 Oct. 2014)

\(^3\) "Monthly Report on Senegal" is available from Japan embassy in Senegal's home page in Japanese.

3. Anomaly Detection

Data Preparation

Data was given with format of outgoing tower x incoming tower x 24 hours x 365 days. Anomaly detection was conducted at each tower level, outgoing and incoming calls were aggregated to reduce data dimension. Thus, our data format was hourly incoming and outgoing calls (SMS) for each tower.

Basic Call pattern

Table 1 displays summary statistics of hourly number of call and SMS for 1614 towers. Although call and SMS have similar mean: 280 and 269 respectively, their maximum is quite different: some 38,000 vs. 444,000. SMS is more variable than mobile phone call. For call, the largest volume was observed on Jan. 23, 2013 20:00 a day before national holiday: Muhammad's birthday.

Figure 1 displays, average call pattern for 24 hours. Call volume is the lowest at 4:00 and highest at 20:00. In case of SMS, the highest slides to 22:00 and until midnight. In Figure 2, yearly call pattern is summarized. SMS increases during May, July, and August. Some spikes are also observed at 1 January and so on. Compared to SMS, mobile phone call volume is stable and it is lower than SMS. Based on this observation, focusing SMS is appropriate to detect anomalies.

<table>
<thead>
<tr>
<th>Table 1 Summary Statistics of Hourly Call and SMS (N=14,103,600)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Min.</td>
</tr>
<tr>
<td>1st-Qu.</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>3rd-Qu.</td>
</tr>
<tr>
<td>Date and Time when Maximum Call/SMS was observed (Tower no.)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Analysis

As we observed, call and SMS shows distinct pattern, we define anomaly as "deviation from regular pattern". To detect anomalies, call (SMS) volume were related to previous call volumes and exogenous variables. We have four time series variables: volume of incoming and outgoing volume for call and SMS, they are explained by lagged four variables and exogenous
variables. To analyze multiple time series, VAR model was applied. As exogenous variables, the following dummy variables were introduced. Although, weather related disaster could affect call volume, weather data was available only 11 locations⁴. Thus, they are not included in our analysis. Included variables are easily defined, it is beneficial for analysis, and especially data availability is restricted less developed countries.

List of exogenous variables
- Holiday dummy (1 on 14 national holidays, 0 for other date)
- Day of the week dummy (Base=Saturday)
- Month dummy (Base=February)
- Hour dummy (Base=4:00)

For 1,615 towers, VAR (Vector Auto Regression) model with 24 hours lagged variables are fitted. Although, included variables are limited, model fits very well. Means of $R^2$ were 0.898, 0.874, 0.716, and 0.730 for outgoing call, incoming call, outgoing SMS, and incoming SMS respectively.

Estimates for antenna 9 is displayed in Table 2. To reduce heterogeneity, log of volume was utilized for analysis. As the table shows, call volumes were significantly explained by some of listed variables. Based on the estimates, "expected call volume" were calculated as "regular call volume". Relative residual were calculated as follows. Relative residual more than 2 or less than -2 was defined as "anomalies".

Relative residual = \( \frac{\text{Observed volume} - \text{Expected volume}}{\text{Observed volume}} \)

---

⁴ We located weather data in Senegal for 11 points at the following sites.
Table 2 Results of Estimation (Antenna 9)

<table>
<thead>
<tr>
<th></th>
<th>Log (1^ incoming Call)</th>
<th>Log (1^ Outgoing Call)</th>
<th>Log (1^ incoming SMS)</th>
<th>Log (1^ Outgoing SMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>t-value</td>
<td>Estimate</td>
<td>t-value</td>
</tr>
<tr>
<td>Lagged variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln_call_o_H1</td>
<td>3.35E-01</td>
<td>11.10</td>
<td>1.98E-01</td>
<td>6.62</td>
</tr>
<tr>
<td>ln_call_L1</td>
<td>-4.59E-01</td>
<td>13.40</td>
<td>5.68E-01</td>
<td>17.76</td>
</tr>
<tr>
<td>ln_sms_o_H1</td>
<td>1.87E-02</td>
<td>1.29</td>
<td>5.48E-03</td>
<td>0.38</td>
</tr>
<tr>
<td>ln_sms_L1</td>
<td>9.58E-02</td>
<td>6.11</td>
<td>1.14E-01</td>
<td>7.15</td>
</tr>
<tr>
<td>const</td>
<td>7.35E-01</td>
<td>9.02</td>
<td>6.88E-01</td>
<td>8.51</td>
</tr>
<tr>
<td>trend</td>
<td>-2.51E-05</td>
<td>-1.33</td>
<td>-2.57E-05</td>
<td>-1.37</td>
</tr>
<tr>
<td>Holiday dummy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fg.holiday</td>
<td>-3.76E-03</td>
<td>-0.17</td>
<td>-4.78E-03</td>
<td>-0.22</td>
</tr>
<tr>
<td>Day of the week dummy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fg.dwk_Mo</td>
<td>-4.68E-02</td>
<td>-3.04</td>
<td>-4.22E-02</td>
<td>-2.78</td>
</tr>
<tr>
<td>fg.dwk_Tu</td>
<td>-5.03E-02</td>
<td>-3.99</td>
<td>-5.44E-02</td>
<td>-3.70</td>
</tr>
<tr>
<td>fg.dwk_We</td>
<td>-3.33E-02</td>
<td>-2.25</td>
<td>-3.04E-02</td>
<td>-2.08</td>
</tr>
<tr>
<td>fg.dwk_Th</td>
<td>-2.12E-02</td>
<td>-1.44</td>
<td>-1.92E-02</td>
<td>-1.31</td>
</tr>
<tr>
<td>fg.dwk_Fr</td>
<td>-5.67E-02</td>
<td>-3.85</td>
<td>-4.57E-02</td>
<td>-3.13</td>
</tr>
<tr>
<td>Month dummy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fg.mn_Jan</td>
<td>-1.87E-02</td>
<td>-0.71</td>
<td>-1.69E-02</td>
<td>-0.66</td>
</tr>
<tr>
<td>fg.mn_Feb</td>
<td>-1.42E-02</td>
<td>-0.61</td>
<td>-1.09E-02</td>
<td>-0.47</td>
</tr>
<tr>
<td>fg.mn_Mar</td>
<td>4.68E-02</td>
<td>1.37</td>
<td>4.44E-02</td>
<td>1.31</td>
</tr>
<tr>
<td>fg.mn_May</td>
<td>7.53E-02</td>
<td>1.64</td>
<td>7.17E-02</td>
<td>1.58</td>
</tr>
<tr>
<td>fg.mn_Jun</td>
<td>9.30E-02</td>
<td>1.59</td>
<td>8.72E-02</td>
<td>1.50</td>
</tr>
<tr>
<td>fg.mn_Jul</td>
<td>1.06E-01</td>
<td>1.48</td>
<td>1.05E-01</td>
<td>1.49</td>
</tr>
<tr>
<td>fg.mn_Aug</td>
<td>1.40E-01</td>
<td>1.64</td>
<td>1.47E-01</td>
<td>1.73</td>
</tr>
<tr>
<td>fg.mn_Sep</td>
<td>1.37E-01</td>
<td>1.39</td>
<td>1.39E-01</td>
<td>1.42</td>
</tr>
<tr>
<td>fg.mn_Oct</td>
<td>1.69E-01</td>
<td>1.49</td>
<td>1.72E-01</td>
<td>1.54</td>
</tr>
<tr>
<td>fg.mn_Nov</td>
<td>1.82E-01</td>
<td>1.43</td>
<td>1.79E-01</td>
<td>1.42</td>
</tr>
<tr>
<td>fg.mn_Dec</td>
<td>2.83E-01</td>
<td>1.82</td>
<td>2.27E-01</td>
<td>1.47</td>
</tr>
<tr>
<td>Hour dummy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fg.h.0</td>
<td>-2.52E-02</td>
<td>-0.59</td>
<td>-1.21E-02</td>
<td>-0.29</td>
</tr>
<tr>
<td>fg.h.1</td>
<td>8.93E-02</td>
<td>2.21</td>
<td>8.28E-02</td>
<td>2.36</td>
</tr>
<tr>
<td>fg.h.2</td>
<td>1.55E-01</td>
<td>4.20</td>
<td>1.73E-01</td>
<td>4.72</td>
</tr>
<tr>
<td>fg.h.3</td>
<td>1.59E-02</td>
<td>0.51</td>
<td>2.21E-02</td>
<td>0.71</td>
</tr>
<tr>
<td>fg.h.4</td>
<td>-2.89E-01</td>
<td>-9.30</td>
<td>-3.15E-01</td>
<td>-10.22</td>
</tr>
<tr>
<td>fg.h.5</td>
<td>-4.39E-01</td>
<td>-11.88</td>
<td>-4.25E-01</td>
<td>-11.60</td>
</tr>
<tr>
<td>fg.h.6</td>
<td>4.48E-01</td>
<td>11.10</td>
<td>5.21E-01</td>
<td>8.78</td>
</tr>
<tr>
<td>fg.h.7</td>
<td>8.89E-01</td>
<td>21.38</td>
<td>8.07E-01</td>
<td>19.45</td>
</tr>
<tr>
<td>fg.h.9</td>
<td>8.23E-01</td>
<td>18.84</td>
<td>7.52E-01</td>
<td>17.39</td>
</tr>
<tr>
<td>fg.h.10</td>
<td>5.41E-01</td>
<td>11.97</td>
<td>4.88E-01</td>
<td>10.87</td>
</tr>
<tr>
<td>fg.h.11</td>
<td>3.83E-01</td>
<td>8.28</td>
<td>3.50E-01</td>
<td>7.64</td>
</tr>
<tr>
<td>fg.h.12</td>
<td>3.52E-01</td>
<td>7.42</td>
<td>3.41E-01</td>
<td>7.23</td>
</tr>
<tr>
<td>fg.h.13</td>
<td>3.19E-01</td>
<td>6.51</td>
<td>2.96E-01</td>
<td>6.09</td>
</tr>
<tr>
<td>fg.h.14</td>
<td>2.74E-01</td>
<td>5.51</td>
<td>2.49E-01</td>
<td>4.97</td>
</tr>
<tr>
<td>fg.h.15</td>
<td>4.24E-01</td>
<td>8.52</td>
<td>4.09E-01</td>
<td>8.09</td>
</tr>
<tr>
<td>fg.h.16</td>
<td>3.02E-01</td>
<td>6.02</td>
<td>2.82E-01</td>
<td>5.68</td>
</tr>
<tr>
<td>fg.h.17</td>
<td>3.29E-01</td>
<td>6.41</td>
<td>2.94E-01</td>
<td>5.93</td>
</tr>
<tr>
<td>fg.h.18</td>
<td>3.40E-01</td>
<td>6.81</td>
<td>2.94E-01</td>
<td>5.94</td>
</tr>
<tr>
<td>fg.h.19</td>
<td>3.59E-01</td>
<td>7.29</td>
<td>3.11E-01</td>
<td>6.37</td>
</tr>
<tr>
<td>fg.h.20</td>
<td>3.77E-01</td>
<td>7.88</td>
<td>3.12E-01</td>
<td>6.59</td>
</tr>
<tr>
<td>fg.h.21</td>
<td>2.33E-01</td>
<td>5.01</td>
<td>2.15E-01</td>
<td>3.27</td>
</tr>
<tr>
<td>fg.h.22</td>
<td>1.34E-01</td>
<td>2.96</td>
<td>4.56E+02</td>
<td>1.11</td>
</tr>
<tr>
<td>fg.h.23</td>
<td>-2.55E-02</td>
<td>-0.58</td>
<td>-7.01E-02</td>
<td>-1.62</td>
</tr>
</tbody>
</table>

Note) Lagged variables were included up to 24 hours. Estimates for them are excluded due to space limitation.

Four types of anomalies

Figure 3 displays observed and expected SMS outgoing volume. Based on this chart two types of anomalies are exist: increased and decreased anomalies. Although clear relationship is not identified, we infer that increased more than regular volume tend to be caused by events that does not impact infrastructure. On the contrary, decreased anomaly caused by disasters that affect antennas.
The same analysis was conducted to other towers. Geographical and time trend distribution of anomalies is summarized in Figure 4. Geographical distribution exhibits relatively large residuals are observed at rural areas. From time-location distribution, two types of anomalies are also identified: national anomaly: towers all around Senegal indicates anomalies and local anomaly: antennas at limited region indicates anomalies.

Note) Size of x proportionate magnitudes of residual.

Figure 4 Distributions of Anomalies
Table 3 lists top 15 antennas with anomalies. In 2013/7/6 10:00, anomalies observed at 1499 antennas. We cannot identify event that caused this anomaly. Further examination is necessary to understand cause of anomaly.

Table 4 lists top 15 antennas with anomalies. As displayed in Figure 4, antennas in rural areas tend to have anomalies. Table 3 and Table 4, compares the number of "increased anomalies (Observation > Expected) and decreased anomaly. For top 15s, decreased anomaly is dominant that reflects problems in communication network caused by natural disasters or technical problems.

Table 3 Tops 15 Dates and Hour with Anomalies

<table>
<thead>
<tr>
<th>Rank</th>
<th>Date</th>
<th>Hour</th>
<th># Of Antenna with Anomalies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Obs&lt; Expected</td>
</tr>
<tr>
<td>1</td>
<td>2013/7/6</td>
<td>10:00:00</td>
<td>1499</td>
</tr>
<tr>
<td>2</td>
<td>2013/8/4</td>
<td>10:00:00</td>
<td>1301</td>
</tr>
<tr>
<td>3</td>
<td>2013/7/20</td>
<td>1:00:00</td>
<td>1033</td>
</tr>
<tr>
<td>4</td>
<td>2013/7/20</td>
<td>8:00:00</td>
<td>935</td>
</tr>
<tr>
<td>5</td>
<td>2013/7/19</td>
<td>18:00:00</td>
<td>896</td>
</tr>
<tr>
<td>6</td>
<td>2013/4/18</td>
<td>19:00:00</td>
<td>890</td>
</tr>
<tr>
<td>7</td>
<td>2013/3/7</td>
<td>8:00:00</td>
<td>875</td>
</tr>
<tr>
<td>8</td>
<td>2013/3/29</td>
<td>7:00:00</td>
<td>835</td>
</tr>
<tr>
<td>9</td>
<td>2013/3/29</td>
<td>8:00:00</td>
<td>547</td>
</tr>
<tr>
<td>10</td>
<td>2013/3/29</td>
<td>2:00:00</td>
<td>516</td>
</tr>
<tr>
<td>11</td>
<td>2013/4/30</td>
<td>23:00:00</td>
<td>486</td>
</tr>
<tr>
<td>12</td>
<td>2013/4/19</td>
<td>23:00:00</td>
<td>482</td>
</tr>
<tr>
<td>13</td>
<td>2013/5/2</td>
<td>23:00:00</td>
<td>480</td>
</tr>
<tr>
<td>14</td>
<td>2013/4/10</td>
<td>23:00:00</td>
<td>463</td>
</tr>
<tr>
<td>15</td>
<td>2013/5/6</td>
<td>3:00:00</td>
<td>421</td>
</tr>
</tbody>
</table>

Table 4 Top 15 Antennas with Anomalies

<table>
<thead>
<tr>
<th>Rank</th>
<th>Antenna No.</th>
<th># of Date-hour</th>
<th>Region</th>
<th>Department</th>
<th>Arr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Obs&lt; Expected</td>
<td>Obs&gt; Expected</td>
<td>Sum</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1174</td>
<td>1940</td>
<td>7</td>
<td>SEDHIOU</td>
<td>BOUNKILING BOGHAL</td>
</tr>
<tr>
<td>2</td>
<td>1177</td>
<td>1733</td>
<td>2</td>
<td>SEDHIOU</td>
<td>GOUDOMP DJIBANAR</td>
</tr>
<tr>
<td>3</td>
<td>1624</td>
<td>1651</td>
<td>4</td>
<td>KEDOUGOU</td>
<td>KEDOUGOU BANDAFASSI</td>
</tr>
<tr>
<td>4</td>
<td>1380</td>
<td>1384</td>
<td>0</td>
<td>TAMBCOUNDA</td>
<td>KOUPEOUUM KOUTHIABA</td>
</tr>
<tr>
<td>5</td>
<td>890</td>
<td>1381</td>
<td>1</td>
<td>KAOJACK</td>
<td>NIRO WACK</td>
</tr>
<tr>
<td>6</td>
<td>1296</td>
<td>1374</td>
<td>2</td>
<td>KOLDA</td>
<td>MEDINA YORO</td>
</tr>
<tr>
<td>7</td>
<td>1603</td>
<td>1269</td>
<td>69</td>
<td>KEDOUGOU</td>
<td>KEDOUGOU BANDAFASSI</td>
</tr>
<tr>
<td>8</td>
<td>1442</td>
<td>1318</td>
<td>1</td>
<td>KOLDA</td>
<td>VELINGARA BONCONTO</td>
</tr>
<tr>
<td>9</td>
<td>1361</td>
<td>1214</td>
<td>2</td>
<td>TAMBCOUNDA</td>
<td>KOUPEOUUM KOUTHIABA</td>
</tr>
<tr>
<td>10</td>
<td>1390</td>
<td>1141</td>
<td>11</td>
<td>KOLDA</td>
<td>VELINGARA SARE</td>
</tr>
<tr>
<td>11</td>
<td>1386</td>
<td>1133</td>
<td>4</td>
<td>KOLDA</td>
<td>VELINGARA PAKOUR</td>
</tr>
<tr>
<td>12</td>
<td>1456</td>
<td>1133</td>
<td>3</td>
<td>TAMBCOUNDA</td>
<td>TAMBCOUNDA MISSIRAH</td>
</tr>
<tr>
<td>13</td>
<td>289</td>
<td>1106</td>
<td>0</td>
<td>DAKar</td>
<td>GRAND</td>
</tr>
<tr>
<td>14</td>
<td>1500</td>
<td>1011</td>
<td>10</td>
<td>TAMBCOUNDA</td>
<td>TAMBCOUNDA MISSIRAH</td>
</tr>
<tr>
<td>15</td>
<td>1301</td>
<td>993</td>
<td>19</td>
<td>KAFFRINE</td>
<td>KOUNGHEUL IDA</td>
</tr>
</tbody>
</table>
Classification of anomaly

We identified two criteria to classify anomalies: direction (increase or decrease) and region (local or national). Table 5 tabulate anomalies. As we observed earlier, decreased-local anomalies dominates. Nationally increased anomalies category includes only 30 observations: Mar. 29 at 7:00 and July. 6, at 10:00 and so on.

Table 5 Classification of Anomalies (# of Date-hour -Tower with anomaly)

<table>
<thead>
<tr>
<th></th>
<th>Decreased (Observed &lt; Expected)</th>
<th>Increased (Observed &gt; Expected)</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td>82285</td>
<td>326</td>
<td>82611</td>
</tr>
<tr>
<td>National</td>
<td>13891</td>
<td>30 (Mar. 29 at 7:00, July. 6, at 10:00 etc.)</td>
<td>13921</td>
</tr>
<tr>
<td>Sum</td>
<td>96176</td>
<td>356</td>
<td>96532</td>
</tr>
</tbody>
</table>

4. Community Detection

Senegal consists of more than 40 peoples. Administrative wards are determined based on geographical, cultural and political reasons. To harness social capital, to classify similar regions into the same sub-areas is desirable to lower communication cost. In this section, we classify regions based on call volume among antennas.

Data and method

Same data with previous section, Data set 1: hourly call volume among antenna is utilized for analysis. This section aims to classify regions, thus volume was aggregated with time and our analysis is based on antenna x antenna call volume.

For most antennas, call and SMS with Dakar consist major part. For such pattern, hierarchical cluster method was unsuccessful, thus we employed Info MAP method (Rosvall et al. 2009). For this analysis, sum of outgoing and incoming call was utilized. And to avoid heterogeneity, log of them were utilized.

Figure 5 exhibits results of community detection. Dots in the same color indicates antenna belong to the same community based on inter-antenna call volumes. As we can see the same color expand beyond administrative wards.

5 Function "infomap.community" of library(igraph) of R was utilized for the analysis.
5. Summary and Discussion

Summary

To realize decentralized approach, two important problems must be solved. 1) How community should be organized. 2) How community should respond to anomalies, such as epidemiologic outbreak, natural disasters. This project tried to answer these questions using mobile phone call data.

For anomaly detection, anomaly was defined as deviation from regular call volume. For each antenna, VAR model was applied with lagged dependent variables and exogenous variables to estimate regular pattern. Although, explanatory variables are limited to holiday, month, and hour dummy that are easily defined, model fit was satisfactory. Regional weather could affect call pattern, however, that is available only 11 locations. Analysis with minimum variables is effective strategy for less developed countries.

Based on direction and regional distribution of anomalies, we classified them into four patterns. And we found "decreased" anomaly dominates that indicates natural disaster or technological problem damages infrastructure. Based on geographic distribution, we found anomalies at rural regions are frequent. Although, we cannot get enough information on cause of anomalies, support for rural area is urgent.

For organizing community, we classified antennas based on inter-antenna call volume. Detected community expands beyond administrative wards. As we mentioned earlier, inter-antenna call volume can be utilized as proxy of (inverse of) communication cost. Detected community can be utilized for the unit of region to community based decentralized health care system.
Limitation

We believe, our method contributes development of Senegal, however, we recognize the following problems that leads to further study.

Firstly, anomalies should be related to cause(s). Our analysis found anomalies, for example, Mar. 29 at 7:00, July. 6, at 10:00 and so on. Although, we tried to relate them with disasters or events, we failed to get information cause of this national level anomaly. Further information on disaster and event in Senegal is necessary to understand cause of anomaly.

Secondly, analysis of people's movement is more appropriate to classify areas. Although movement data was supplied for the D4D participants, limitation in computational power prohibited us to analyze it.

Thirdly, research on health care system in Senegal is necessary to implement decentralized health care system. Launch of Graeme Bank was based on research project on credit delivery system Yunus and Jolis(2003). Solution oriented research is effective to attack this problem.

Acknowledgement

We thank to Orange.com providing mobile phone call data. We also thank to steering team who enabled this competition.

Reference


Yunus, Muhammad and Alan Jolis (2003), Banker to the Poor: Micro-Lending and the Battle against World Poverty: PublicAffairs.
Weakening the Incidence of Transmittable Diseases by Taking Advantage of Mobile Phone Activity

J.T. Matamalas,¹ M. De Domenico,¹ and A. Arenas¹
¹Departament d’Enginyeria Informàtica i Matemàtiques, Universitat Rovira I Virgili, Tarragona, Spain
(Dated: December 31, 2014)

One of the primary goals in developing countries as Senegal consist in identifying outbreaks of endemic and communicable diseases, and preventing their spreading in the country and neighboring countries, without having a major health surveillance system. A possibility, exploited in the current work, is to analyze call records to infer mobility patterns and then the natural flow of epidemic contagion between individuals. The analysis is applicable in the full range of scales, from individuals to regions, providing a proxy for monitoring outbreaks and prevent their spreading. Our proposal consists in a memory-driven mobility model that captures the mobility patterns beyond the state of the art, by assigning high resolved probabilities to transitions between different arrondissements. Moreover, we complement the model with a game theoretical approach to the dilemma faced by individuals affording counter-measures to the epidemic spreading. Our results allow an accurate tracking of an epidemic outbreak, and provides enlightening control parameters to reduce the incidence of the spreading process based on cost reduction and social reinforcement.

I. INTRODUCTION

Nowadays, among the issues that low-income countries are facing, healthcare is probably the most urgent. In these countries, a significant fraction of the population lives in rural areas where infrastructures are often limited and physical information campaigns might be cost-effective. Additional concerns are related to the difficulty in accurately mapping important socio-economic indicators as poverty and health.

Recently, the private collaboration between researchers and mobile operators opened the door to a new way to gather information about the population through mobile phone data recorded by mobile carriers and indicating the closest mobile tower where an individual actively uses his or her mobile device. This possibility triggered a wide variety of studies showing, for instance, that mobile phones heterogeneously penetrated both rural and urban communities, regardless of richness, age or gender, providing evidences that mobile technologies can be used to build realistic demographics and socio-economics maps of low-income countries.¹ Mobile phone call records have been successfully used in a wide variety of cellphone applications as, for instance, to estimate population densities and their evolution at national scales, to confirm social theories of behavioral adaptation, where individuals change their patterns of communication to increase the similarity with their new social environment, and to capture anomalous behavioral patterns associated with a broad range of events, from religious and official holidays to earthquakes, floods, violence against civilians and protests. Even more recently, the public availability of mobile phone data sets thanks to data challenges further revolutionized the field by allowing ubiquitous sensing for mapping poverty, more efficient transportation planning, accurate monitoring of social segregation and information campaigns optimization to reduce epidemics spreading.

Although mobile phone data have biases and might not be representative of the whole population in some circumstances, they provide, at the same time, one the most powerful tools for reality mining, sensing complex social systems, studying human movement data and a valuable proxy for human mobility studies. The possibility to use individuals’ call detail records (CDR) allows to quantify mobility patterns that influence the spatial spread of infectious diseases, with important applications to containment and eradication programs, increasing the interest in understanding human dynamics its impact on the spreading of transmittable diseases.

In this study, we capitalize on the high-quality mobile phone data provided in occasion of the second edition of Data for Development (D4D) challenge to address healthcare issues in Senegal, with particular attention to diseases such as HIV AIDS and Tuberculosis, transmittable from human to human. Although the encouraging results achieved by Senegal in response, for instance, to HIV, certain social groups as sex workers still have very high rates close to 21%. Prostitution has been legal in Senegal since 1969, under severe restrictions, making Senegal a leading country in the fight against AIDS. However, many women and men are either not aware of being infectious or they do not want to make it evident to their family and the institutions, to avoid being marked as prostitutes, homosexuals or individuals having sexual relationships with prostitutes. In fact, this introduce a healthcare problem that prevent infectious people from doing medical tests and using prophylactics during their sexual activity. Remarkably, it has been recently shown that the role of clients in networks created by commercial sex workers is not marginal, because traveling clients shorten network distances between distant network neighborhoods and thus facilitate contagion among them more than sex workers.

For these reasons, we focus our attention on modeling the transmission of diseases in the population of Senegal while accounting, simultaneously, for mobility among
different arrondissements and the dilemma emerging in
the adoption of an innovation (e.g. condoms in the case
of AIDS) or a prophylaxis (e.g. treatments suggested by
World Health Organization in the case of Ebola). To
this aim, we introduce a memory-driven human mobility
model inferred from CDR and we extensively show that it
better captures existing correlations in mobility dynam-
ics, being more suitable than classical memoryless mod-
els [26, 28, 30, 32, 41, 42] for realistic simulations oriented
to policy-making based on short-term predictions. We
consider a susceptible-exposed-infected-recovered (SEIR)
transmission model within each arrondissement, modeled
as a meta-population with homogeneous mixing, and we
show that memory in human mobility is responsible for
slowing down the spreading of a disease. Finally, this
complex dynamics is coupled to the dilemma of adopt-
ing an innovation (or a vaccination), accounting for the
socio-economic costs associated to individual’s choice.

Our results can be used to propose an efficient dissem-
ination strategy, based on social reinforcement and with
roots on adoption dilemma, that makes individuals more
likely to adoption when they are made aware, through
mobile phone alerts or online social network platforms,
of the risks associated to the opposite behavior.

This work is organized as follows: In Sec. II we de-
scribe the available data sets and how we use them for
subsequent models and analysis. In Sec. III we introduce
our human mobility model based on adaptive geographi-
cal memory. In Sec. IV we describe the disease spreading
model coupled to mobility between arrondissements. In
Sec. V we introduce the human dilemma of innovation
(or vaccination) adoption, unifying such dynamics with
human movements through the country and the conta-
gion spread. We summarize and discuss the implications
of our findings in Sec. VI and, finally, we complement
this study by including appendices where we explain our
methods in more details and where we provide more de-
tailed supporting analysis.

II. DESCRIPTION OF THE DATA SETS

In this study, we make use of all data sets provided
by the D4D Challenge, some supplementary data sets
provided by partners and an additional ad hoc data set
built by us for the challenge.

The data provided by the D4D Challenge concern
communications 1666 communication towers distributed
across Senegal. In particular, three data sets are pro-
vided and described in the following. “SET1” reports
the number of calls, their duration and the number of
short texts exchanged by each pair of towers, aggregated
on an hourly basis. We exploit this information to map
communication patterns between different areas of the
country (i.e. the arrondissements). “SET2” provides the
source of calls from about 300,000 users in a two-week
period, giving information about the location (at tower
spatial resolution) of the user when she or he makes a call.

Different sub-sets are provided, each one containing infor-
mation about independent samples of users. We exploit
this data set to investigate the stochastic fluctuations in
human movements, to understand to which extent the as-
sumption of homogeneous mixing in a meta-population
model is satisfied (see App. D for details about this anal-
ysis). Finally, “SET3” provides 500 millions call records
of about 150,000 users along one year at a coarser spa-
tial resolution (i.e. arrondissement level) than SET2. We
use this data set, spanning a larger period of time than
SET2, to map individuals’ movements among different
arrondissements.

Demographics information has been obtained from the
Senegal data portal1 suggested as an additional official
resource for the Data for Development Challenge 2014.
It is worth noting that information has been manually
checked against inconsistencies and data about popu-
lation for the arrondissements of Bambilor, Thies Sud,
Thies Nord, Ndio and Ngothie were not available. We
reconstructed the missing information by combining mo-
 bile phone activity and available demographics data (see
Appendix and Fig. A.1 for further details). These data
have been used to infer more realistic contact rates to be
used in viral spreading simulations. The contacts among
individuals are generally quite difficult to track at coun-
try level. Their rate varies depending on several social
and demographical factors such as age, gender, location,
urban development, etc. [13, 47]. Nevertheless, there are
evidences from European and African countries that, on
average, the number of daily physical contacts among in-
dividuals range from 11 to 22 [13, 47]. In this work we
consider the same contact rate for all arrondissements be-
longing to the same region. Our choice is justified by the
lack of detailed data about the population density per ar-
 rondissement. Moreover, we assign 25 contacts per day
to the region with highest density (i.e. Dakar) and we
rescale the contact rate of all other regions proportion-
ally to the ratio between their density and the density of
Dakar, assuming that the minimum contact rate can not
be smaller than 10 contacts per day. We obtain a contact
rate between 10 and 11 for all regions, except Dakar.

Finally, we use an external data set to build an ad-
ditional layer of communication that can be exploited,
together with mobile phone calls, to target individuals in
different arrondissements for information dissemination.
We built this data set after monitoring the Twitter activ-
ity in the country and considering all geo-localized tweets
posted by individuals from 23 Nov 2014 to 25 Dec 2014
(33 days). The total number of collected tweets in our
data set is 67,914, covering the 88.9% of departments and
100% of regions, corresponding to a potential sensing of
75% of the whole population.

We show in Fig. 1 three maps of Senegal, color-coded
by different information used throughout this paper,

1 http://donnees.ansd.sn/en/
namely population, mobile phone and Twitter activity corresponding to each arrondissement.

III. MODELING HUMAN MOBILITY IN SENEGAL

We consider geographic areas, identified here at arrondissement level, as \( n \) nodes of a network embedded in space, where links encode mobility flux from one area to another one.

The movement of individuals between two areas is inferred from mobile phone activity records provided with the Data for Development Challenge 2014 (see Sec. II for details). Let \( \mathcal{A} \) be the set of all \( n = 123 \) physical arrondissements of Senegal. The physical mobility network, represented by the adjacency matrix \( M \), consists of the physical nodes connected by edges weighted by the fraction of people moving between pairs of arrondissements.

\[ M_{ij} = \frac{\sum_{\ell=1}^{L} m^{(\ell)}_{ij}}{\sum_{k=1}^{n} \sum_{\ell=1}^{L} m^{(\ell)}_{ik}}, \quad (1) \]

where \( m^{(\ell)}_{ij} \) is the number of times the individual \( \ell \) makes at least one call in arrondissement \( j \) after making at least one call in arrondissement \( i \). We do not impose a specific time window to calculate transitions, to avoid introducing biases and undesired effects due to the choice of the temporal range and it is worth remarking that other normalizations can be considered depending on data and metadata availability \[32\]. In the following, where not otherwise specified, we consider the mobility matrix obtained from the whole period of observation (see Sec. II for details). This model is known as “first-order” (or 1-memory) because the present state is the only information required to choose the next state. Although very useful, this has the fundamental disadvantage that it does not account for mobility memory. In fact, it is very likely that an individual moves to a neighboring area (by means of a car or public transportation) to work and after a few hours he or she will go back to the original position. This effect has been shown to be relevant also at much larger scale, e.g. at country level, where individuals fly from one city to another and often go back to their origin instead of moving towards a different city \[48\]. This memory is an intrinsic property of human mobility and must be taken into account for a realistic modeling of people movements between different geographic areas. When memory is taken into account, each physical node (e.g., \( i \in \mathcal{A} \)) is replaced by the corresponding state-nodes (e.g., \( i \diamond j \in \mathcal{\tilde{A}} \) if memory is of order 2) encoding the information that an individual is in arrondissement \( i \) when he or she comes from \( j \). While \( M \) encodes information about the network of \( n \) physical nodes, we need to introduce a new matrix \( \mathcal{\tilde{M}} \) to encode information about the network of \( n^2 \) state-nodes, accounting for all binary combinations (e.g. \( k \diamond j, j, k = 1, 2, ..., n \)) between physical nodes.

We clarify the difference between first-order and second-order model with one example, by considering mobile phone calls made by an individual during travels between three American cities, namely Chicago (C), San Antonio (S) and Baltimore (B). Let SSSCCCCSSSS-BBBBBBCCCB be the sequence of his or her calls during the period of observation. We show in Fig. 2 the transitions calculated using the two models (Fig. 2A and
Figure 2. Mobility models built from a representative sequence of mobile phone calls (SSSCCCSSSSBBBBBCCCCBB) made by an individual during travels between three American cities, namely Chicago (C), San Antonio (S) and Baltimore (B). The first-order and second-order models fail to capture evident returning patterns, whereas the adaptive memory approach, introduced in this work, correctly identify such patterns (see the text for further details).

2-State: \( XY \equiv X \rightarrow Y \)  

Mobility notation: \( X \rightarrow Y \rightarrow Z \equiv XY \rightarrow YZ \equiv Z \triangleleft Y \triangleleft X \)

Fig. 2B, respectively). While physical nodes correspond to the three cities, state-nodes involve information about the current physical node and its mobility relationship with another physical node. For instance, the fact that the individual makes a call in San Antonio conditioned to the fact that he or she made previously a call while in Chicago is encoded in the state-node labeled by CS, corresponding to \( S \triangleleft C \) following the notation introduced above. The second-order mobility model would better capture the probability of returning patterns as \( S \triangleleft C \triangleleft S \) (see Fig. 2), with respect to a first-order model, if the data used to build the corresponding transition probabilities were GPS traces. However, this is not the case of the present study, where calls are used to infer movements. In fact, when direct information about mobility is not available and indirect methods have to be used, a second-order Markov model can not be directly employed, because the memory in the sequence of calls can be quite different from the memory that should be accounted for in the sequence of movements. This is clear from the example shown in Fig. 2B where movement patterns like \( S \triangleleft C \triangleleft S \) and \( S \triangleleft C \triangleleft B \) are not captured from the bare sequence of calls that, conversely, favors patterns like \( C \triangleleft C \triangleleft S \) and \( S \triangleleft S \triangleleft C \) as expected from a simple second-order model.

### B. Adaptive memory-driven model

To overcome the issues of first-order and second-order models, here we propose a more general approach, that we name, in short, adaptive memory. As in the second-order model, we build the transition matrix between all possible 2-states, but instead of counting transitions in the sequence of calls, we count transitions in the sequence of unique movements. The main drawback of a second-order model is that the memory about the starting physical place, e.g., S, is lost after the second consecutive call in C (i.e., ...SCCC...), an evident undesirable effect. Here, we propose to preserve the information about the starting physical place by introducing the self-transition \( C \triangleleft S \leftarrow C \triangleleft S \): if the probability of this transition is larger than zero then the individual started his or her movements from state-node \( C \triangleleft S \) and he or she is assumed to stay in this state, corresponding to the call pattern \( C \triangleleft C \triangleleft ... \leftarrow C \triangleleft S \). The resulting mobility transitions between the three areas are quite different from the previous models, as shown in Fig. 2C. The adaptive memory approach correctly captures the existence of the returning pattern \( S \triangleleft C \triangleleft S \), at variance with the second-order model, and assigns a much higher probability to this transition if compared with the value obtained from the first-order model. Moreover, while a first-order model introduces spurious mobility patterns as \( S \triangleleft C \triangleleft B \), that are not observed in the sequence of movements, the adaptive memory approach successfully identify the absence of such transitions.

In general not all transitions between state-nodes are allowed. In our adaptive memory mobility model all second-order transitions like \( k \triangleleft j \leftarrow j \triangleleft i \) (including the case \( k \equiv j \equiv i \)) are allowed and encode the movements starting in node \( i \) and ending in node \( k \) while passing through node \( j \). However, our model allows (forbids) transitions that are forbidden (allowed) in a canonical second-order Markov model. The adaptive memory model does not allow transitions like \( j \triangleleft j \leftarrow j \triangleleft i \)
Figure 3. Mobility flow among a sub-set of Senegal’s arrondissements using two different models, first-order (A) and adaptive memory (B). This circular visualization is suitable for identifying mobility patterns (for instance, see [49]). The size of the flow is proportional to the width of the link and different colors are assigned to different links. The direction of the flow is encoded by the gap between link and circle segment at destination: links closer (farther) to the circle indicate origin (destination).

(with $j \neq i$), otherwise allowed in a second-order Markov model, where the memory about the origin physical node $i$ would be lost. In our model, this kind of dynamics corresponds to a self-loop in the state-node $j < i$, i.e. the transition $j < i \rightarrow j < i$. The adaptive memory model still defines a Markovian process between state-nodes, but it should not confused with a second-order Markov process because it can not be reduced to the first-order model represented in Fig. 2A. From the point of view of physical nodes, the dynamics is described by a non-Markovian process.

The impact of adaptive memory approach on modeling human mobility is significant and becomes evident when empirical data are used. In Fig. 3 is shown an emblematic case, where the modeled mobility flow among a sub-set of arrondissements is considered under the constraint that individuals must pass through Kael (arrondissement in the department of Mbacke, region of Diourbel) after departing from their origin and before reaching their destination. The figure shows the mobility modeled by means of first-order approach and adaptive memory, putting in evidence the different mobility patterns in the two cases. For instance, the adaptive memory model captures returning patterns (i.e. movements like $X \rightarrow$ Kael $\rightarrow X$) better than the first-order model. Moreover, the significant mobility pattern Ndame $\rightarrow$ Kael $\rightarrow$ Colobane exhibited at first-order turns to be fictitious when memory about origin is taken into account as in the adaptive memory approach. Such differences are fundamental because they dramatically affect the outcome of simulated dynamics devoted, for instance, to the prediction of traffic congestion or epidemics spreading.

C. Mobility equations

To build a consistent model of spreading dynamics accounting for human mobility patterns at country level, we first introduce a suitable framework that will allow us to write the equations of the dynamics in a compact and elegant form. In the following, to avoid confusion we use Greek indices for state-nodes whereas Latin indices will be used for physical nodes.

The dynamics of a first-order Markov model is governed by the mobility matrix $M$ whose elements $M_{ij} = p(j|i)$ corresponds to the probability that an individual would reach physical node $j$ if he or she was in physical node $i$ at the previous step. Similarly, the dynamics of the adaptive memory model is governed by $\tilde{M}$ where elements $\tilde{M}_{\alpha \beta} = p(\beta|\alpha)$ corresponds to the probability that an individual would reach state-node $\beta$ if he or she was in state-node $\alpha$ at the previous step. It is worth remarking that $\alpha, \beta = 1, 2, ..., n^2$ provide a convenient alternative representation of states $j < i$ and $j < k$, whereas $i, j = 1, 2, ..., n$. In general, mobility matrices $M$ and $\tilde{M}$ are not symmetric.

To quantify the impact of adaptive memory in modeling mobility, we simulated the movements of thousands
of individuals departing from each arrondissement independently and exploring the network for six months according to mobility matrices $M$ (FO) and $\tilde{M}$ (AM). We considered three widely adopted descriptors of such dynamics, namely the coverage -- defined as the fraction of nodes visited within the considered time window assuming to start from a specific origin, the mean first passage time -- defined as the expected time required to move between any pair of origin and destination nodes -- and the mean return time -- defined as the average time required to come back to the origin node. We show in Fig. 4A, B and C how these descriptors evolve over time when first-order (FO) and adaptive memory (AM) models are used. The difference between the two models in any mobility descriptor tends to increase over time and to be statistically significant already after one month. More in detail, we consider the coverage and the mean return time for each arrondissement separately, by averaging over hundreds of independent random realizations, and then we rank the nodes accordingly to answer questions like “from which arrondissement the diffusion runs faster through the country?” or “what is the average time to wait for a diffusive agent to go back to its origin?”. In Fig. 4D and Fig. 4E the ranking obtained by using the first-order model is shown against the one obtained by using the adaptive memory model. Although the two rankings are correlated, as expected, it is evident that arrondissements favoring diffusion in one case do not behave similarly when memory is accounted for, with the coverage estimated using AM systematically smaller than the coverage estimated using FO (the relative difference ranges from -25% to -40%). A similar result is found for the mean return time, with relative differences ranging from -80% to 20%. Our findings are robust even if the mobility matrix is built from monthly observations: we show in Fig. B.5 and Fig. B.6 the results obtained for the coverage and the mean return time by considering mobility in each month separately. Moreover, we show in Fig. B.4 an analysis of the similarity between monthly mobility that reveal interesting seasonal patterns. In Fig. 4F, for each origin-destination pair we show the relative difference between the mean first passage times ob-
tained using first-order and adaptive memory mobility models. Note that in the adaptive memory case, the mean first passage time is calculated for each arrondissement by averaging over times of the corresponding state-nodes. The difference range from -110% up to 360% in some cases. The first-order model favors a faster diffusion that leads to a smaller first passage time with respect to adaptive memory model (the underestimation is on average 50%), whereas the return time, whose values can be obtained from the diagonal entries of the mean first passage time matrix, is significantly overestimated by the first-order model.

These results have deep implications on applications where short-term or long-term predictions of dynamical processes (e.g. epidemics spreading) where human mobility is actively involved. In fact, a first-order model favors a faster diffusion whereas the adaptive memory model predicts a slower diffusion through the network, slowing down, for instance, the spreading of an epidemics as shown further in the text.

In the following, we use the mobility matrices corresponding to first-order and adaptive memory to build two different mobility models. Let \( N_i(t) \) indicate the population of the physical node \( i \in A \) at time \( t \), then the \( n \) mobility equations describing how the flux of people diffuse through the network are given by

\[
N_i(t + 1) = \sum_{j=1}^{n} M_{ji} N_j(t). \tag{2}
\]

In the case of adaptive memory, we indicate by \( \tilde{N}_\alpha(t) \) the population of the state-node \( \alpha \in \tilde{A} \) at time \( t \) and the \( n^2 \) mobility equations required to describe the same process are given by

\[
\tilde{N}_\alpha(t + 1) = \sum_{\rho=1}^{n^2} \tilde{M}_{\rho\alpha} \tilde{N}_\rho(t). \tag{3}
\]

The population in each physical node at time \( t \) is given by the sum of the population in the corresponding state-nodes. It is worth remarking that, in general, the matrix \( \tilde{M} \) can be a function of time as well and the equations would keep their structural form.

IV. MODELING INFECTIOUS DISEASE SPREADING IN SENEGAL

In general, it is not possible to model with a high accuracy level the spreading of any disease, because of the huge variability in infection profiles and dynamics. For instance, diseases can be transmitted from human to human by physical contact, by exchanging fluids and through air, or in many other cases they require carriers (e.g. insects or animals) for the transmission. Diseases like HIV or Ebola are quite complex and in general depend on many factors like sexual behavior (the basic reproduction ratio of HIV is 4 for the homosexual population in the United Kingdom and 11 for female prostitutes in Kenya) or resistance to policies. Nevertheless, such diseases, together with Measles, Chickenpox, Hepatitis, Cholera and Plague, can be still modeled by classical epidemiological dynamics for an appropriate choice of parameters.

In the following, we consider a static transition matrix for simplicity. A model of disease spreading coupled to mobility is easier to build when the dynamics is governed by a first-order mobility matrix, where each node is an arrondissement with an initial population that moves from one place to another according to modeled mobility patterns. In this scenario and in absence of information about the underlying contact network in each arrondissement, we consider each node as a meta-population where any individual can interact with a limited number of other individuals in the same node, although extensions to heterogeneous mixing would be possible depending on data availability. Conversely, in the case of higher-order models, as in the adaptive memory case, the meta-population is considered at state-node level and non-infected individuals in a state-node might interact with infected individuals in other state-nodes corresponding to the same physical node (see Fig. 5).

Some important transmittable diseases affecting Senegal in the past or at the present time, like AIDS, different strains of Hepatitis, Cholera and Plague, can

Figure 5. Interactions between susceptible and infected individuals in a meta-population. The SEIR transmission model involves four different compartments, with individuals being in one compartment or another depending at which stage of the disease they are (susceptible, exposed, infected, recovered or removed). In the first-order mobility model (A), susceptible individuals in the compartment \( S \) can be infected by individuals in the compartment \( I \) within the same physical node. In the adaptive memory mobility model (B) susceptible individuals in each state-node can be infected by interactions (here indicated by dotted arrows) with individual belonging to the compartment \( I \) of any other state-node within the same physical node.

not described by simple epidemiological models (e.g. susceptible-infected). In fact, pathogens responsible for these diseases reproduce rapidly within the host, who are infected but not yet infectious, and are not blocked by the immune system at the very beginning, although their abundance is too low to be transmitted to other susceptible individuals. This stage is modeled by adding an “exposed” compartment to the traditional susceptible-infected-recovered model, leading to a SEIR compartmental model [62–64]. In the following, we adopt a SEIR epidemiological dynamics coupled to mobility to model the spreading of diseases like the ones described above. In the case of diseases like AIDS it is important to account for the difference between homosexual and heterosexual contacts, as well as the possibility of vertical transmission, etc. Moreover, with a good level of approximation a SEIR model is still able to give insights about the spreading of the Ebola disease [65, 66], although more sophisticated approaches can be used [67, 68]. Nevertheless, to keep the description of the model as simple as possible, we will not consider such complex models, that can be easily integrated in the theory in more specialist studies. Here we do not focus on Ebola, because the recent outbreak in the West Africa did not affect Senegal thanks to the prompt response plan of its government to the case confirmed on 29 August 2014.

The discrete time step of the following models is \( \Delta t \approx 1 \) hour, approximately the median between two successive calls from the same individual. The parameters are demographics and epidemiological. Demographics parameters include the birth \( B \) and death \( \delta \) probability, whereas epidemiological parameters correspond to the latent period \( \tau_{E} \) of the infection, from which the probability \( \epsilon = \Delta t/\tau_{E} \) to pass to the infectious state is calculated, and the infectious period \( \tau_{I} \), from which the probability \( \gamma = \Delta t/\tau_{I} \) to recover from or die because of the infection is calculated. The last parameter is the effective transmission probability

\[
\beta(t) = 1 - \left( 1 - \tilde{\beta} \Delta t \frac{I_{i}(t)}{N_{i}(t)} \right)^{c_{i} \Delta t},
\]

an arrondissement-dependent parameter that depends on the average number of contacts per unit of time \( c_{i} \) had by an individual in node \( i \), the fraction of infected individuals in that node and the transmission risk \( \beta \Delta t \) in case of contact with an infectious individual. In fact, the definition of \( \beta(t) \) induces a type-II reaction-diffusion dynamics [21] computing for the fact that each individual does not interact with all other individuals in the meta-population, but only with a limited sample. If the number of infected agents is small (i.e. \( I_{i}(t) \approx 0 \)) the Taylor expansion of \( \beta(t) \) truncated at the first order gives the classical factor \( \beta \Delta t c_{i} \frac{I_{i}(t)}{N_{i}(t)} \) [64]. It follows that the equations describing the average spreading of a disease according to a SEIR model coupled to first-order mobility are given by

\[
\begin{align*}
S_{i}(t + 1) &= \sum_{j=1}^{n} M_{ji} \left[ (1 - \delta - \beta_{j}(t) I_{j}(t)) S_{j}(t) + B N_{j}(t) \right] \\
E_{i}(t + 1) &= \sum_{j=1}^{n} M_{ji} \left[ (1 - \epsilon - \delta) E_{j}(t) + \beta_{j}(t) S_{j}(t) \right] \\
I_{i}(t + 1) &= \sum_{j=1}^{n} M_{ji} \left[ (1 - \gamma - \delta) I_{j}(t) + \epsilon E_{j}(t) \right] \\
R_{i}(t + 1) &= \sum_{j=1}^{n} M_{ji} \left[ (1 - \delta) R_{j}(t) + \gamma I_{j}(t) \right]
\end{align*}
\]

whereas the coupling to the more realistic adaptive memory mobility model is given by

\[
\begin{align*}
\tilde{S}_{i}(t + 1) &= \sum_{\psi=1}^{n^{2}} M_{i\psi} \left[ (1 - \tilde{\beta}_{\psi}(t)) \tilde{S}_{\psi}(t) + B \tilde{N}_{\psi}(t) \right] \\
\tilde{E}_{i}(t + 1) &= \sum_{\psi=1}^{n^{2}} M_{i\psi} \left[ (1 - \epsilon - \delta) \tilde{E}_{\psi}(t) + \tilde{\beta}_{\psi}(t) \tilde{S}_{\psi}(t) \right] \\
\tilde{I}_{i}(t + 1) &= \sum_{\psi=1}^{n^{2}} M_{i\psi} \left[ (1 - \gamma - \delta) \tilde{I}_{\psi}(t) + \epsilon \tilde{E}_{\psi}(t) \right] \\
\tilde{R}_{i}(t + 1) &= \sum_{\psi=1}^{n^{2}} M_{i\psi} \left[ (1 - \delta) \tilde{R}_{\psi}(t) + \gamma \tilde{I}_{\psi}(t) \right]
\end{align*}
\]

\[
\tilde{\beta}_{\psi}(t) = 1 - \left( 1 - \tilde{\beta} \Delta t \frac{\tilde{I}_{\psi}(t)}{\tilde{N}_{\psi}(t)} \right)^{c_{i} \Delta t}
\]

where \( N(t) = \sum_{\psi=1}^{n^{2}} \tilde{N}_{\psi}(t) \) is the total population in the country at time \( t \), \( \lfloor \cdot \rfloor \) indicates the floor function and is used to identify the sub-set of state-nodes corresponding to the same physical node the population \( \tilde{S}_{\alpha} \) belongs to. We refer to App. C and Fig. C.7 and C.8 for a comparison between agent-based simulations and theoretical expectations in the case of a Meningitis outbreak.

V. MODELING THE ADOPTION DILEMMA IN SENEGAL

Preventing and containing severe diseases like AIDS or Ebola require the efforts of the community, although

4 Senegal faced an outbreak of this disease in 2012, see http://www.who.int/mediacentre/factsheets/fs161/en/
care must be taken in assuming that targeted individuals will adopt prevention or treatment policies. Emblematic cases are community-based strategies in the city for sex workers, shown to fail unless significant changes to the institutional arrangements are achieved [69]. In fact, theoretical models developed to reduce infection incidence generally assume that such super-spreaders will always agree to be vaccinated or adopt the innovation, completely neglecting the effect of policy resistance [70]. Voluntary adoption is often unlikely to be the best strategy, unless a huge disproportion between the perceived risk of disease and vaccination is in act and it has been shown that, in some cases, the perceived cost of the disease must be orders of magnitude higher than the perceived cost of vaccine-associated side effects for a successful adoption campaign [71], leading to very complicated dynamics [72]. Nevertheless, targeted information campaigns [10, 73, 74] have been shown to increase individual’s awareness about the disease, causing spontaneous behavioral changes in the population [75, 76] that can not be neglected to correctly inform public health decisions [78]. The recent outbreak of Ebola in West Africa [5] provided evidences that prevention and containment strategies based on information dissemination on specific geographic areas [10] can be very effective, as in Senegal [6] and Nigeria [7].

Here, we wonder if other strategies based on social mechanisms can be also effective in preventing the spreading of a severe communicable disease like Tuberculosis, Meningitis, Measles, AIDS or Ebola, to cite some ones that can be modeled by a SEIR dynamics.

We consider a population where a communicable disease is introduced, and infected individuals have a certain probability to transmit it to susceptible individuals. People getting sick suffer the effects derived from the illness symptoms, e.g. nausea, headache or joint pain. If we put such effects in terms of “fitness”, a sick person has smaller fitness than a healthy one, because he has to live with these symptoms. One way to model this scenario is by assuming that an infected individual has to pay a penalty $P$ in terms of fitness. Now imagine that some individuals get aware of a preventive measure, e.g. vaccination, usage of contraceptives, etc, that can reduce the probability of getting infected after a contact with an infectious individual. Such a measure has a cost $c$ that all people who adopted it $A$ have to pay, not just in economic terms but also in social terms – e.g. the social penalty for a woman found with condoms could be even higher than the cost of infection in some societies. We indicate people who do not adopt the measure by $\neg A$. Despite the reduction in the probability of being infected $\beta_\neg A >> \beta_A$, in the worst scenario some individuals who adopted the measure could still be infected ($\beta_A > 0$): they have paid the cost for the adoption and they are also infected [79]. This last possibility, together with the previous ones, raises in each individual an adoption dilemma where he or she must choose to adopt or not the countermeasure based on what it is going on in his or her neighborhood and on information provided by news broadcasters. In our model, we take into account all the possible outcomes of the prevention strategy and we summarize in Tab. I the different fitness for different infection/adoption scenarios.

![Infection Infection](#)

<table>
<thead>
<tr>
<th>$\neg$ Adoption</th>
<th>$c$</th>
<th>$c + P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\neg$ Infection</td>
<td>0</td>
<td>$P$</td>
</tr>
</tbody>
</table>

**Table I.** Possible fitness scenarios used in the adoption dilemma where $c$ is the cost associated to the adoption of a preventive measure and $P$ is the penalty that an individual “pays” when he or she gets infected. Individuals who adopt the suggested countermeasure and also get infected pay a higher penalty $c + P$. The values of $c$ and $P$ have to be negative in order to be considered a loss in terms of fitness.

In the following we consider diseases modeled by SEIR dynamics (see Sec. IV). In this case, the adoption of the countermeasure makes sense when susceptible (or exposed) individuals try to reduce their chances of getting infected. It is worth noting that exposed individuals, as susceptible ones, are targeted for the adoption process because they are not aware of having got the disease, and they also want to reduce their chances of getting sick. As we will show, this behavior has interesting implications on the evolution of the dynamics. Infected and recovered individuals do not need to use countermeasures because they have already got the infection.

The adoption process works as follows. An infectious agent is introduced in a healthy population; the infection follows its course until it reaches a predefined penetration level, after which some fraction of individuals adopts the preventive measure(s). This phase can be obtained by using an informative campaign about the benefits of adopting the countermeasure. At each time step each (susceptible or exposed) individual evaluates his or her neighborhood and employs a strategic choice based on his or her perception of how good is to adopt or not the countermeasure. This evaluation is performed in terms of the fitness, by computing the average fitness of the individuals who adopt or not the countermeasure. The equations that model how the fitness is calculated in the scenarios described in Tab. I are given by

\begin{align}
\text{Fitness} &= \text{Infection} + \text{Adoption} \\
\text{Infection} &= \beta \cdot S \\
\text{Adoption} &= \frac{A}{A + \neg A}
\end{align}
Figure 6. Comparison between the temporal evolution of a spreading process not accounting (A) and accounting (B) for social dilemma dynamics. The fraction of susceptible, exposed, infected and recovered individuals in the whole country is shown for each model versus time. Non-adopters get infected with probability \( \beta_{-A} = 0.07 \), whereas adopters get infected with probability \( \beta_A = 0.0007 \). The cost of the adoption \( (c) \) is considered to be 1% of the penalty \( P \) for being infected. The absolute difference between adoption and non-adoption dynamical processes is shown in panel C.

\[
\begin{align*}
    f_i^A(t) &= c \cdot \left( \frac{S_i^A(t) + E_i^A(t)}{N_i^A(t)} \right) + \\
    &\quad + (c + P) \cdot \left( \frac{I_i^A(t) + e^{-\frac{(t_R)}{\omega}} R_i^A(t)}{N_i^A(t)} \right), \\
    f_i^{-A}(t) &= P \cdot \left( \frac{I_i^{-A}(t) + e^{-\frac{(t_R)}{\omega}} R_i^{-A}(t)}{N_i^{-A}(t)} \right),
\end{align*}
\]

where \( X_i^Y \) indicates adopters \( (Y = A) \) and non-adopters \( (Y = -A) \) in a given compartment \( X = S, E, I, R \) and the whole population \( N \).

Equations (7) take into account the frequency of the different kinds of SEIR individuals in the populations, describing how the spreading dynamics and the adoption dynamics are coupled with each other. It is worth discussing here the role played by recovered (R) subjects in the calculation of the fitness. Recovered individuals have paid the cost of being infected at some point although, after recovering, they tend to forget that cost with time, reducing the impact on the average fitness. We model this behavior by introducing a factor \( e^{-\frac{(t_R)}{\omega}} \) that reduces the weight of recovered individuals in the calculation of the fitness, by accounting for the average time \( (t_R) \) that they have spent in this state and for a parameter \( \omega \) that specifies how fast the weight decays to zero.

After computing the fitness, individuals start imitating other individuals in their neighborhood. People tend to imitate the behavior of successful individuals, whereas individuals using the worst strategy in terms of fitness tend to change their strategy proportionally to the expected gain they will obtain after changing. This process is known as proportional imitation. If the fitness of the individuals who adopt the countermeasure is better than the fitness of the ones who do not adopt it, then the probability that an individual is a non-adopter is given
The fitness variables are given by

\[
\Pi_{A \rightarrow \rightarrow A}(t) = \frac{f^A_A(t) - f^\sim A_A(t)}{c - P}.
\]

Conversely, if the fitness of adoption is worse than the fitness of non-adoption then the probability that individuals who adopted the countermeasure will leave it to

\[
\Pi_{A \rightarrow \sim A}(t) = \frac{f^\sim A_A(t) - f^A_A(t)}{c + P}.
\]

Our model, accounting simultaneously for the human mobility between meta-populations (using our adaptive memory approach), the spreading dynamics and the adoption dilemma is given by the equations

\[
\begin{align*}
S^A_A(t+1) &= \sum_{\psi=1}^{n^2} M_{\psi A} \left[ (1 - \Pi^A_{\psi \rightarrow A}(t)) \left( (1 - \delta - \tilde{\beta}^{(\alpha)}_{\psi A}(t)) S^A_A(t) + B \tilde{N}^A_A(t) \right) + \\
&\quad (1 - \beta^{(\alpha)}_{\psi A}(t)) S^\sim A_A(t) + B \tilde{N}^\sim A_A(t) \right]
\end{align*}
\]

\[
\begin{align*}
\tilde{S}^A_A(t+1) &= \sum_{\psi=1}^{n^2} \tilde{M}_{\psi A} \left[ (1 - \Pi^\sim A_{\psi \rightarrow A}(t)) \left( (1 - \delta - \tilde{\beta}^{(\alpha)}_{\psi \sim A}(t)) \tilde{S}^A_A(t) + B \tilde{N}^A_A(t) \right) + \\
&\quad (1 - \beta^{(\alpha)}_{\psi \sim A}(t)) \tilde{S}^\sim A_A(t) + B \tilde{N}^\sim A_A(t) \right]
\end{align*}
\]

\[
\begin{align*}
E^A_A(t+1) &= \sum_{\psi=1}^{n^2} M_{\psi A} \left[ (1 - \Pi^A_{\psi \rightarrow A}(t)) \left( (1 - \delta - \epsilon) E^A_A(t) + \beta^{(\alpha)}_{\psi A} S^A_A(t) \right) + \\
&\quad (1 - \delta - \epsilon) E^\sim A_A(t) + \beta^{(\alpha)}_{\psi \sim A}(t) \tilde{S}^\sim A_A(t) \right]
\end{align*}
\]

\[
\begin{align*}
\tilde{E}^A_A(t+1) &= \sum_{\psi=1}^{n^2} \tilde{M}_{\psi A} \left[ (1 - \Pi^\sim A_{\psi \rightarrow A}(t)) \left( (1 - \delta - \epsilon) E^\sim A_A(t) + \beta^{(\alpha)}_{\psi \sim A}(t) \tilde{S}^\sim A_A(t) \right) + \\
&\quad (1 - \delta - \epsilon) \tilde{E}^A_A(t) + \beta^{(\alpha)}_{\psi A}(t) \tilde{S}^A_A(t) \right]
\end{align*}
\]

\[
\begin{align*}
I^A_A(t+1) &= \sum_{\psi=1}^{n^2} M_{\psi A} \left[ (1 - \delta - \gamma) I^A_A(t) + (1 - \Pi^A_{\psi \rightarrow A}(t)) \epsilon \tilde{E}^A_A(t) + \Pi^A_{\psi \sim A}(t) \epsilon \tilde{E}^\sim A_A(t) \right]
\end{align*}
\]

\[
\begin{align*}
I^\sim A_A(t+1) &= \sum_{\psi=1}^{n^2} \tilde{M}_{\psi A} \left[ (1 - \delta - \gamma) \tilde{I}^A_A(t) + (1 - \Pi^A_{\psi \sim A}(t)) \epsilon \tilde{E}^\sim A_A(t) + \Pi^\sim A_{\psi \sim A}(t) \epsilon \tilde{E}^\sim A_A(t) \right]
\end{align*}
\]

\[
\begin{align*}
R^A_A(t+1) &= \sum_{\psi=1}^{n^2} M_{\psi A} \left[ (1 - \delta) R^A_A(t) + \gamma \tilde{I}^A_A(t) \right]
\end{align*}
\]

\[
\begin{align*}
R^\sim A_A(t+1) &= \sum_{\psi=1}^{n^2} \tilde{M}_{\psi A} \left[ (1 - \delta) \tilde{R}^A_A(t) + \gamma \tilde{I}^A_A(t) \right],
\end{align*}
\]

where there is an equation for each SEIR compartment and the two classes of adopters (\(A\)) and non-adopters (\(\sim A\)). The fitness variables are given by

\[
\begin{align*}
\tilde{f}^A_A(t) &= \sum_{\rho=\lfloor \frac{n}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor + n + 1} \tilde{N}^A_A(t) + \sum_{\rho=\lfloor \frac{n}{2} \rfloor + n + 1}^{\lfloor \frac{n}{2} \rfloor + n + 1} \tilde{N}^A_A(t)
\end{align*}
\]

\[
\begin{align*}
\tilde{f}^\sim A_A(t) &= \sum_{\rho=\lfloor \frac{n}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor + n + 1} \tilde{N}^\sim A_A(t) + \sum_{\rho=\lfloor \frac{n}{2} \rfloor + n + 1}^{\lfloor \frac{n}{2} \rfloor + n + 1} \tilde{N}^\sim A_A(t)
\end{align*}
\]
\[ \tilde{f}^A(t) = P^\rho = \sum_{\rho = [\frac{1}{2}]n+1}^{\frac{n+n}{2}} \left( \tilde{I}_\rho^A(t) + e^{\tilde{R}_\rho^A(t)} \tilde{F}_\rho^A(t) \right) \]  

(12)

Influencing the value of the probabilities

\[ \tilde{\Pi}_{\psi}^{A \rightarrow A}(t) = \begin{cases} \frac{\tilde{f}_\psi^A(t) - \tilde{f}_\psi(t)}{|c + P|} & \tilde{f}_\psi^A(t) > \tilde{f}_\psi(t) \\ 0 & \text{otherwise} \end{cases} \]  

(13)

\[ \tilde{\Pi}_{\psi}^{A \rightarrow \bar{A}}(t) = \begin{cases} \frac{\tilde{f}_\psi^A(t) - \tilde{f}_\psi^\bar{A}(t)}{|c + P|} & \tilde{f}_\psi^A(t) > \tilde{f}_\psi^\bar{A}(t) \\ 0 & \text{otherwise} \end{cases} \]  

(14)

And effective transmission risks are given by

\[ \tilde{\beta}_{\psi^A}(t) = 1 - \left( 1 - \tilde{\beta}_{A \Delta t} \right)^{\rho = \frac{[\Delta t]}{\rho I_{\rho}^A(t) + L_{\rho}^{A}(t)}} \]  

(15)

\[ \tilde{\beta}_{\psi^A}(t) = 1 - \left( 1 - \tilde{\beta}_{\bar{A} \Delta t} \right)^{\rho = \frac{[\Delta t]}{\rho I_{\rho}^A(t) + L_{\rho}^{A}(t)}} \]  

(16)

VI. CONCLUSIONS

Summarizing, we have presented a model that improves significantly our prediction of the incidence of epidemics based on mobility patterns inferred from call detail records. The basic idea is to account for roundtrips by storing information about (memory) departure and arrival inferred from CDR and to take advantage of such information to infer mobility tracks. The adaptive memory model is complemented with a social dilemma that allows to explore initiatives based on social reinforcement, tax reduction, etc. to alleviate the cost (economic, cultural, social, etc.) of individuals who adopt countermea-
Figure 8. Simulated spreading of an infective disease in Senegal. The spatial incidence at arrondissement level after 1 day is shown for the first-order (A) and the adaptive memory plus adoption dilemma (B) models, where the color gradient is proportional to the number of infected individuals. The temporal evolution of the number of infected arrondissements obtained from both models is shown in panel C, whereas we show their relative difference in panel D. The percentage of infected individuals per arrondissement, with respect to the arrondissement’s population, is calculated and the median of the resulting distribution is shown in panel E, whereas the relative difference between the two models is shown in panel F.

Sures suggested by policy makers against epidemic processes. The results of our model are accurate enough to design targeted campaigns, using mobile phone communication channels, to widespread social action for weakening the incidence of an epidemics. In the case of Senegal, we were able to map with high resolution, at the level of arrondissements, the predicted incidence of the disease after a specific outbreak. These results could be of capital importance in helping policy-makers to simulate and forecast the outcome of their interventions in the health system. We devise that the current results can be enhanced by i) using transition matrices at tower level instead of district level; ii) considering the temporal evolution of the mobility matrices; and iii) including in the model additional layers of information awareness. All these facts makes the current work not a final method but an innovative starting point for new developments in social sensing and forecasting methods, where more high-quality social data from other sources (e.g. Twitter, Facebook) can be accounted for in a multilayer framework [80][82] (see Fig. 9) to ameliorate health policies devoted to predict, contain and eradicate contagion spreading.


using multilayer centrality [84]. to identify key arrondissements for information dissemination multivariate channels (e.g. CDR, online social networks, etc.) dilemma can be also coupled to communication patterns from a significant mobility or communication flow between them. Links between pairs of arrondissements indicate the existence of communication (calls and tweets) patterns in Senegal. Links Figure 9. Multilayer visualization [83] of mobility and com-

PNAS 111, 15888 (2014).
Appendix A: Inferring Missing Demographics Data from Mobile Phone Data

Available demographics data for men and women separately are scattered against available mobile phone activity data. Our analysis reveals that a second-order polynomial model in log-log scale is suitable to describe the observed correlation, as shown in Fig. A.1. Although more refined data could be used, we find this approach to be a reasonable trade-off between simplicity and accuracy, and we used the fitted polynomial model to infer the population in Bambilor, Thies Sud, Thies Nord, Ndiob and Ngothie, where mobile phone data are available while demographics data are missing in the database we used.

We investigated the possibility to obtain similar demographics information from another data set, i.e. the geo-localized activity on Twitter in Senegal. Although the statistics of tweets is not comparable to the one of mobile phone data used in this study, we found encouraging correlations between twitter activity, mobile phone activity and population in each arrondissement (see Fig. A.2A and B), suggesting that Twitter can be used as a communication layer complementary to mobile phone communications.

Appendix B: Monthly Mobility Flow

We have also studied the difference between the first-order and the adaptive memory mobility models for each month separately, instead of considering the whole year. The results are shown in Fig. B.3 for odd months. Remarkably, in the first-order model the patterns across months are stable, whereas thanks to the adaptive memory model it is possible to capture changing patterns over time.

To identify the emergence of different patterns over time, we first calculated from simulations the mean first passage time matrix $\tilde{T}_{ij}(\tau)$ between all origin-destination arrondissement pairs $(i, j = 1, 2, ..., n)$ for each month separately ($\tau = 1, 2, ..., 12$), using the adaptive memory mobility model. For each pair of months $\tau_1$ and $\tau_2$, we...
Figure B.3. Mobility flow among a sub-set of Senegal’s arrondissements using two different models, first-order (FO) and adaptive memory (AM). Monthly mobile phone data have been used in this study. The circular visualization is as in Fig. 3.

Figure B.4. Similarity matrix, based on Frobenius distance, between mean first passage time matrices calculated from monthly observations. The dendrograms are shown as a guide to the eye, to ease the identification of months with similar mobility patterns.

calculate the Frobenius norm of their difference

$$||F||_2(\tau_1, \tau_2) = \sqrt{\sum_{i,j=1}^{n} (\tilde{T}_{ij}(\tau_1) - \tilde{T}_{ij}(\tau_2))^2} \quad (B1)$$

to perform a similarity analysis among mobility in different months. The resulting similarity matrix is shown in Fig. B.4, putting in evidence interesting seasonal patterns. This study confirms that models involving long-term mobility should make use of time-varying mobility matrices to obtain more realistic results.

Finally, we show in Fig. B.5 and Fig. B.6 the results of analysis concerning the coverage and mean return time at arrondissement level (see Fig. 4D and E), obtained by using the first-order and the adaptive memory models, and mobility matrices obtained from monthly observations.

Appendix C: Agent-based Simulations versus Spreading Model

To understand the impact of mobility modeling on the spreading of transmittable diseases compatible with SEIR epidemiology, we consider a communicable disease like Meningitis, with transmission probability $\tilde{\beta} = 0.013$, an average contact rate $c_i = 14$ individuals per day (equal for all arrondissements), a latent period $\tau_E = 1$ day and a recovery period $\tau_I = 5$ days. As initial condition, we
consider an outbreak involving 0.01% of the population of Parcelles Assainies, in the region of Dakar. In Fig. C.7 and C.8 are shown, for completeness, the comparisons between results obtained from agent-based simulations and our analytical models (first-order and adaptive memory, respectively) of the spreading dynamics coupled to human mobility. In both cases, an excellent agreement between theoretical expectation and simulation is obtained.

Figure B.5. As in Fig. 4D, but considering mobility models from monthly observations. The coverage is calculated in each arrondissement separately and a scatter plot between their ranks obtained using FO and AM mobility models are shown. Color codes the relative difference between the values obtained from the two models, size is proportional to the population of the Department including the corresponding arrondissement as sub-administrative unit.
Figure B.6. As in Fig. 4E, but considering mobility models from monthly observations. The mean return time is calculated in each arrondissement separately and a scatter plot between their ranks obtained using FO and AM mobility models are shown. Color codes the relative difference between the values obtained from the two models, size is proportional to the population of the Department including the corresponding arrondissement as sub-administrative unit.

Appendix D: Stochastic Fluctuations in Human Movements Observed in Mobile Phone Data

It has been recently shown that human mobility is more predictable than expected [23, 26, 85], with locations visited rather regularly by the same individual. Although it is difficult to study human mobility from CDR only, because of the low resolution compared, for instance, to GPS records, it is possible to gain very interesting insights about human behavior. We consider the data provided in SET2 and for each individual in each sub-set we built the time-ordered list of towers where he or she made a call. To analyze high-quality data, we restrict our attention on individuals who visited at least
three different towers and who made at least 1000 calls within two weeks. Each time-ordered sequence is a symbolic sequence, where a symbol represents the tower ID. We transform each symbolic sequence into a numerical sequence obtaining a time series. It is worth remarking that the analysis performed in the following is robust with respect to the choice of the mapping from symbolic to numeric values, being a dynamical invariant the quantity we are interested into.

We use detrended fluctuation analysis (DFA) to analyze the resulting time series. The description of the DFA of non-stationary time series is out of the scope of the present work and we refer to [86] for details. In this context, it is worth remarking that DFA is a useful tool to investigate the scaling of fluctuations versus time scales in detrended time series. This scaling often follows a power-law where the exponent is known as Hurst ex-
Figure D.9. Detrended fluctuation analysis of human movements. Each analysis has been performed on an independent sub-set (from P01 to P25) of SET2, confirming the stability of the results (to ease the visualization, here only sub-sets from P01 to P06 are shown). The distribution of calculated Hurst exponent is shown in panel A, with a peak close to 1.5, corresponding to pure random walk dynamics. The goodness of the fit is calculated by Pearson’s $R^2$, whose distribution is shown in panel B for the considered trajectories: almost all values are larger than 0.85. In panels C and D the Hurst exponents, one for each mobility trace, are scattered against the number of unique towers visited and the number of calls made by the individual, respectively; solid lines represent the best linear regression models and suggest that Hurst exponents are weakly correlated with the number of unique towers and not correlated with the number of calls.

ponent and it is strictly related to the fractality of the underlying dynamical system. A Hurst exponent of 1.5 provides a clear evidence that fluctuations have the same dynamical properties of a random walk. The Hurst exponent is obtained by fitting the power-law scaling relationship and, in our case we observed that a reasonable scaling region was always present for time scales smaller than 60 hours. We show the result of our analysis in Fig. D.9.

Our results suggest that even if human movements are often regular and quite predictable on average, fluctuations in those movements are often described by a random walk dynamics. This result also suggests that the hypothesis of homogeneous mixing is not a bad approximation.
Human mobility and the spreading of waterborne diseases.

Javier Perez-Saez\textsuperscript{1}, Flavio Finger\textsuperscript{1}, Lorenzo Mari\textsuperscript{1,2}, Andrea Rinaldo\textsuperscript{1,3}, Enrico Bertuzzo\textsuperscript{1}

\textsuperscript{1}Laboratory of Ecohydrology ECHO/IIE/ENAC, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
\textsuperscript{2}Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133 Milan, Italy
\textsuperscript{3}Dipartimento di Ingegneria Civile, Edile ed Ambientale, Università di Padova, 35131 Padova, Italy

Abstract

Human mobility has long been recognized as a crucial determinant for the spreading of infectious diseases. Researchers have traditionally focused on diseases transmitted through direct contact between persons whereas the key role of mobility in the transmission of waterborne diseases (e.g. cholera and schistosomiasis) has been acknowledged only recently via spatially-explicit modeling. Here we analyze mobile phone records provided by the operator Orange for Senegal in order to estimate mobility measures that are crucial for the modeling of the spreading of waterborne diseases: the fraction of people moving outside their resident community and the distribution of trip destinations. Results show that these two quantities can be robustly estimated through the analysis of mobile phone records. A comparative analysis with alternative conceptual models of human mobility used in epidemiological contexts when mobility data are not available, namely gravity and radiation models, shows that mobile phone records are more informative and can be used to infer non-trivial mobility patterns that cannot be captured by simple conceptual models.
1 Introduction

Diarrheal waterborne diseases are caused by pathogenic microorganisms that are transmitted when contaminated water (or food contaminated by water) is consumed and are thus directly or indirectly hydrologically controlled. The pathogens include bacteria (e.g., *Vibrio cholerae*, protozoa (e.g., *Entamoeba histolytica* and *Shigella dysenteriae*), and viruses (e.g., *Rotavirus gastroenteritis*). These diseases are among the major causes of death in the world (World Health Organization 2008). Cholera is possibly the best-known lethal diarrheal disease, although other infections, such as rotavirus, cause more casualties (about 500,000 in 2004 alone, according to the World Health Organization). In Senegal the death rate of diarrheal diseases is about 60 per 100,000 individuals. In 2005 a severe cholera outbreak spread through the country exacting 31719 cases with 458 deaths (case fatality rate of 1.44%).

Although the spread of waterborne diseases is mediated primarily by the hydrological network, infections may be also spread by individuals moving between different places and acting as point sources of pathogens. While traveling or commuting, susceptible individuals can be exposed to pathogens and return as infected carriers to their settlement. Similarly, infected hosts can disseminate the disease away from their home community – in many cases infected individuals are asymptomatic and thus are not barred from their usual activities. In modelling terms, this requires the addition of a mobility network superposed to the hydrological one to describe the spatial dispersion of pathogens.

To understand, control, and predict waterborne disease epidemics, development of appropriate mechanistic models is fundamental, as they can play an important role in devising appropriate intervention measures. Recently, a spatially explicit framework was proposed to model the spread of pathogens through surface waters as a reactive transport process along networks, thus introducing the primary infection mechanism (Bertuzzo et al., 2008, 2010). Human mobility has been described by either diffusion-based (Righetto et al., 2011) or gravity-like models (Gurarie and Seto, 2009; Bertuzzo et al., 2011; Chao et al., 2011; Tuite et al., 2011; Mari et al., 2012b,a). The simultaneous inclusion of both the hydrological and the human mobility network has proved very effective in describing the Haiti cholera epidemic (Rinaldo et al., 2012).

The recent availability of mobile phone records offers an unique opportunity to analyze the regional travel patterns of large numbers of individuals (i.e. of the order of millions as in Bengtsson and al. (2011) and Wesolowski et al. (2012)) over a significant temporal framework in selected spatial settings pertinent to
waterborne diseases. The parallel assessment of human mobility patterns can sensibly reduce the uncertainty associated to the prediction of the spreading of an epidemic outbreak. In this section we first present a general model for the spreading of waterborne diseases in order to highlight what are the components of human mobility that are crucial for the spreading of the disease. In the following sections we will propose a road-map to derive such quantities of interest through the analysis of mobile phone data.

1.1 A spatially-explicit model of waterborne disease dynamics

The model subdivides the total population into \( n \) human communities spatially distributed within a domain that embeds both the human mobility and the hydrological networks. Let \( S_i(t), I_i(t) \) and \( R_i(t) \) be the local abundances of susceptible, infected and recovered individuals at time \( t \) in each node \( i \) of the network, and let \( B_i(t) \) be the environmental concentration of pathogens at site \( i \) (figure 1). Epidemiological dynamics and pathogen transport over the hydrological and human-mobility networks are described by the following set of ordinary differential equations:

\[
\frac{dS_i}{dt} = \mu(H_i - S_i) - \beta_i \left[ (1 - m_i) f(B_i) + m_i \sum_{j=1}^{n} Q_{ij} f(B_j) \right] S_i + \rho R_i
\]

\[
\frac{dI_i}{dt} = \beta_i \left[ (1 - m_i) f(B_i) + m_i \sum_{j=1}^{n} Q_{ij} f(B_j) \right] S_i - (\gamma + \mu + \alpha) I_i
\]

\[
\frac{dR_i}{dt} = \gamma I_i - (\rho + \mu) R_i
\]

\[
\frac{dB_i}{dt} = -\mu_B B_i - l_i \left[ B_i - \sum_{j=1}^{n} \frac{P_{ji} W_j}{W_i} B_j \right] + \frac{P_i}{W_i} \left[ (1 - m_i) I_i + \sum_{j=1}^{n} m_i Q_{j,i} I_j \right]. \tag{1}
\]

The evolution of the susceptible compartment (first equation of model 1) is a balance between population demography and infections due to contact with the pathogen. The host population, if uninfected, is assumed to be at demographic equilibrium \( H_i \) (the size of the \( i^{th} \) local community), with \( \mu \) being the human mortality rate. The parameter \( \beta_i \) represents the site-dependent rate of exposure to contaminated water, and \( f(B) \) is the probability of becoming infected because
Figure 1: Schematic representation of the general model for waterborne diseases of the exposure to a concentration $B$ of pathogens. Because of human mobility, a susceptible individual residing at node $i$ can, while travelling, be exposed to pathogens in the destination community $j$. This is modeled assuming that the force of infection in a given node depends on the local concentration $B_i$ for a fraction $(1 - m_i)$ of the susceptible hosts and on the concentration $B_j$ of the surrounding communities for the remaining fraction $m_i$. The parameter $m_i$ represents the community-level probability that individuals travel outside their node. The concentrations $B_j$ are weighted according to the probabilities $Q_{ij}$ that an individual living in node $i$ would reach $j$ as a destination. The dynamics of the infected compartment (second equation of model 1) is a balance between newly infected individuals and losses due to recovery or natural/pathogen- induced mortality, with $\gamma$ and $\alpha$ being the rates of recovery and mortality due to the disease, respectively. Recovered individuals (third equation of model 1) lose their immunity and return to the compartment of susceptibles at a rate $\rho$ or die for natural mortality at a rate $\mu$. The evolution of the local concentration of pathogens that live free in the aquatic environment (fourth equation of model 1) assumes that bacteria, viruses, or protozoa are released in water (e.g., excreted) by infected individuals.
and immediately diluted in a wellmixed local water reservoir of volume $W_i$ at a site-specific rate $p_i$, depending on local sanitation conditions. Similarly to exposure, the total pool of infected people possibly shedding pathogens accounts for human mobility and in particular for infected people who travel from community $j$ to the focal community $i$. Freeliving pathogens are also assumed to die at a constant, site-dependent rate $\mu_{B,i}$, which can widely vary according to inhomogeneous climatic and ecological conditions. As regards the hydrological transport, the spread of pathogens over the river network is described as a biased random-walk process on an oriented graph (Bertuzzo et al. 2007). Specifically, we assume that pathogens can move, at a rate $l_{ij}$ from node $i$ to node $j$ of the hydrological network with a probability $P_{ij}$. The rate $l_i$ depends on both downstream advection and other possible pathogen transport pathways along the hydrological network, for example, short-range distribution of water for consumption or irrigation or pathogen attachment to phyto- and zooplankton.

Two crucial quantities are therefore directly related to human mobility patterns, namely $m_i$, the fraction of people living in community $i$ that travel outside their resident node, and the matrix $Q_{ij}$ which represents the probability that an individual living in node $i$ would reach $j$ as a destination. In previous epidemiological applications, these quantities were inferred using different mathematical modeling schemes, say based on gravity-like approaches (Gurarie and Seto, 2009; Bertuzzo et al., 2011; Chao et al., 2011; Tuite et al., 2011; Mari et al., 2012b,a) or on the so-called radiation models (Simini et al., 2012), whose parameters were calibrated contrasting model results against epidemiological records.

2 Methods

2.1 Extraction of mobility patterns from mobile phone data

Mobile phone records are analyzed to extract mobility patterns both in terms of trips and distribution of time spent by each user in distinct locations (here termed pause times). In the perspective of linking mobility to distinct geographical areas we choose to use the coarse grained dataset provided by Orange (SET 3) which consists of mobile phone records of 150’000 users for a period of 12 months. Records consist of a timestamp, a user ID and a location for whenever a given user engages in a communication. The objective being to quantify the origin/destination fluxes for any location, the first step of the analysis consists in determining the home location of each user, in order to then determine trips and pause time distributions. Both measures are relevant for epidemiological
modeling. Indeed long pause times can be associated with high risk of exposure to human-to-human transmission of infectious diseases or increased probability of consumption of contaminated food or water. On the other hand the number of inbound trips can be linked to exogenous contamination rates of a location by travelers which is of great importance for waterborne disease transmission. A user’s home is defined to be the location where the user spends the most time at night, here considered to be the timespan between 10 PM and 6 AM (computation (a) in figure 2). Having the home location id of a given user, outbound trips to other locations were defined as the realized presence of the user in any other location for any pause time in all user trip loops corresponding to records between two consecutive mobile phone usages in the user’s home location (computation (b) in figure 2). Here pause time is defined as the time between two consecutive communications. Pause time distribution for any given user is computed as the cumulative time spent in location different from the user’s home (computation (c) in figure 2).

![Figure 2: Methodology flowchart](image)

2.2 Using geographical and demographical information to model mobility patterns in epidemiological models

The gravity model (Erlander and Stewart, 1990) is commonly used within spatially explicit epidemiological models to describe fluxes of susceptible, infected and recovered individuals within a network of geographical locations. Here we
compare the mobility fluxes obtained through different implementations of the gravity model as well as the radiation model proposed by Simini et al. (2012) with those extracted from mobile phone data.

2.2.1 Geographical and demographical data

Inputs to mobility models are the population size of all network nodes as well as a measure of distance between the nodes. In this project we use two kinds of distance measures:

- Euclidean distance between centroids of the arrondissements;
- Travel time between the arrondissements.

Whereas the computation of euclidean distances between nodes is straightforward, its use might be subject to several disadvantages. It does not take into account the presence or absence of roads and detours that must be taken to travel from one point to another. This might be particularly important in Senegal because of Gambia, which can only be crossed in a few places, which can significantly prolong journeys between the northern and the southern parts of the country. The euclidean distance does not take into account the presence of topographical obstacles and mountain ridges, which also might prolong a journey, although this effect might be of lesser importance in Senegal due to the relative flatness of the country. As an alternative one can try to estimate the travel time between two network nodes and use this information as distance measure in the gravity and radiation models. An overview of the data and methods we use for this purpose is given in table 1 and in figure 3. A cost surface raster is built from the road network and the digital terrain model of Senegal, where the cost (units of time) of crossing a raster cell depends on the presence or absence of a road.

<table>
<thead>
<tr>
<th>Data</th>
<th>Source</th>
<th>Accessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTM</td>
<td>SRTM (NASA)</td>
<td>23.11.2014</td>
</tr>
<tr>
<td>Administrative boundaries</td>
<td>d4d</td>
<td>–</td>
</tr>
</tbody>
</table>
as well as on the terrain slope. The shortest path between every pair of pixels is then computed using the Dijkstra algorithm (Dijkstra, 1959). To get the travel time between two nodes (e.g. *arrondissements*), the population weighted mean of the travel time from pixels belonging to the two nodes is computed.

![Flow chart of the creation of the travel time matrix](image)

**Figure 3:** Flow chart of the creation of the travel time matrix

### 2.2.2 Gravity and radiation models

We compare the performances of the following implementations of the gravity model as well as the radiation model with respect to the mobility patterns extracted from mobile phone data:
\[ Q'_{ij} = \frac{H_j e^{-d_{ij}/D}}{\sum_{k \neq i}^n H_k e^{-d_{ik}/D}} \]  
\[ Q''_{ij} = \frac{H_j d_{ij}^\delta}{\sum_{k \neq i}^n H_k d_{ik}^{-\alpha}} \]  
\[ Q'''_{ij} = \frac{1}{1 - H_i \left( \sum_{k=1}^n H_k \right)^{-1} \left( H_i + s_{ij} \right) \left( H_i + H_j + s_{ij} \right)} \]  

Equation 2 is a gravity model characterized by an exponential kernel, as used in Rinaldo et al. (2012), where \( H_j \) is the population of node \( j \), \( d_{ij} \) is a distance measure between nodes \( i \) and \( j \) and \( D \) is a parameter standing for the deterrence distance. Equation 3 is a gravity model with a power function as a kernel, the exponent \( \delta \) being a parameter. Equation 4 is the radiation model as described by Masucci et al. (2012), where \( s_{ij} \) stands for the population residing at a distance smaller than the one to the destination \( j \) from home node \( i \), with home population not included. Gravity models rely on one or several parameters which have to be estimated, whereas the radiation model is parameter free. As discussed in the Introduction, an additional parameter \( m_i \) is needed to describe the fraction of population taking travels outside of the resident node \( i \).

3 Results

3.1 Mobility patterns extracted from mobile phone data

To elicit mobility patterns at the national level, the 10 administrative regions composing the agglomeration of Dakar (arrondissement IDs 1-10) where combined into one location in the following plots. Trips and pause times from and to these arrondissements where summed to obtain total fluxes to the city. In the following, yearly or monthly fluxes refer to statistics obtained using the whole year or a single month of data, respectively.

Figure 4 illustrates the preponderant role of Dakar in yearly mobility fluxes in terms of trips with most of the mobility occurring between the arrondissements of the capital and the rest of the country. The other large cities of the country act as local basins of attraction for mobility fluxes such as Ziguinchor and Kao-lack. Regional fluxes between in-land districts are also visible, in particular fluxes connecting the Northern arrondissements along the Senegal river.
Patterns highlighted in the number of trips connecting each region can also be found in fluxes expressed in terms of time spent in other arrondissements (figure 5). A trend that could be observed is that pause times connecting inland arrondissements to Dakar tend to be relatively larger than in the case of the number of trips. This trend is confirmed when comparing number of trips to pause time as illustrated in figure 6.

When comparing trips to pause times no linear relation can be found (figure 6). Indeed a greater variability of pause times for low numbers of monthly trips indicate a strong heterogeneity in the type of trips users took during the survey period. Nevertheless increased number of trips for a given user can be associated to longer pause times which could correspond to a certain socio/economic activities. Interestingly no visible difference exists between the dry and wet periods.

Figure 4: Yearly trips. Color code refers to the log of flux intensities. Fluxes smaller than 10,000 trips per month are not shown.
Figure 5: Yearly pause times. Color code refers to the log of pause times. Total monthly pause time shorter than $10^7$ minutes are not shown.

in terms of trip and pause time distributions.

The distribution of the number of monthly trips follow an exponential-like distribution (figure 7), with a larger amount of trips originating from Dakar in comparison with the rest of the country as observed in the geographical distribution of fluxes in figure 4. Furthermore no clear difference in the number of trips during the wet (May-November) and dry season (December-April) is observed. Nevertheless a seasonal trend can be observed in the outliers of the distribution of trips both for Dakar and the rest of the country, with maximal outlier values in the middle of the wet period (figure 8). On the other hand pause time distribution is bell-shaped in both seasons and for both trips from Dakar and the rest of the country with an average around 12 hours. Pause time distribution does not have strong seasonal fluctuations at the national scale (figure 8).
Figure 6: Trips vs pause time (minutes) distribution (log values).
(a) Trips histogram estimate

(b) Pause times (minutes) histogram estimate

Figure 7
(a) Box plot of monthly trips

(b) Box plot of monthly pause times (minutes)

Figure 8
3.2 Modelling mobility patterns based on travel time

3.2.1 Distances and travel time

Figure 9 shows the relation between the travel time computed according to the algorithm presented in figure 3 and the euclidean distance. Points are close to a straight line indicating that the euclidean distance is a good measure of distance for most pairs of points, even if they go from the northern to the southern part of the country or the contrary. This pattern is probably due to the dense road network, in particular the two roads crossing Gambia, as well as the absence of major topographical obstacles in the country.

![Figure 9](image_url)

Figure 9: Estimate travel time between two nodes vs. euclidean distance. Red: trajectories from *arrondissements* south of Gambia (Regions Kolda, Sedhion and Ziguinchor) to other arrondissements, and vice versa. Blue: all other trajectories.
3.2.2 Mobility patterns estimated using gravity and radiation models

For further analyses we consider the number of yearly trips between nodes and the travel time as mobility estimates and distance measure, respectively. The 10 arrondissements of Dakar are kept aggregated. The elements of the matrix $Q_{ij}$ are estimated from the mobile phone data by dividing the number of trips between node $i$ and $j$ by the total number of trips originating from node $i$ during a certain time span. The parameters of the gravity models 2 and 3 are then calibrated such that the output of the models would fit $Q_{ij}$ best, using the sum of square errors as optimization objective. The output of the radiation model 4 was computed without calibration, being parameter free.

From figure 10 it can be seen that the two gravity models represent the elements of $Q_{ij}$ from and to Dakar better than the radiation model. It seems however that the gravity models overestimate the importance of other major cities such as Diourbel, Mbour and Kaolack. For better visibility all elements of $Q_{ij}$ from and to Dakar are removed in figure 11. It is now possible to see that the radiation model does a good job predicting local movements between nearby communes, whereas it predicts almost no long distance movements.

Figure 12a shows the row-wise mean square error between the gravity (equation 2) and radiation models and the observed $Q_{ij}$. A row of the matrix $Q_{ij}$ represent all the fluxes originating from location $i$. It can be seen that both models do a good job predicting movements originating from Dakar (first bar), whereas the gravity model clearly outperforms the radiation model for the rest of the origins. Figure 12b shows the elements of the observed $Q_{ij}$ for trajectories originating from Dakar, compared to their equivalents according to the gravity (equation 2) and to the radiation models.

From the analysis of these figures it seems that the fluxes from and to Dakar are extremely important, such that they dominate the other elements of $Q_{ij}$ during parameter estimation for the gravity models. The gravity models are able to represent this effect, but only by putting too much emphasis on the fluxes to other highly populated places. The radiation model is unable to represent long distance movements within the country, but represents localized movements correctly.
Figure 10: Elements of matrix $Q_{ij}$ as estimated from the mobile phone data (a), the radiation model (b), and the two gravity models (c and d). For better readability only the three biggest outgoing values of each node are shown. The color represents the value of the element, going from 0 (green) to 1 (red).
Figure 11: Identical to figure 10, except that fluxes from and to Dakar are excluded to allow for better readability.
(a) Row wise root mean square errors of the gravity model (3) and the radiation model compared to the matrix $Q_{ij}$ based on the mobile phone data. Each row corresponds to a different origin of trajectories. Note the low RMS for Dakar (first arrondissement on the left) compared to the other origins.

(b) Row of the matrix $Q_{ij}$ which corresponds to outgoing trajectories from Dakar, as estimated using mobile phone data and modelled using the gravity and the radiation model.

Figure 12
3.2.3 Estimation of the fraction of population travelling

As described in the Introduction, an important parameter to estimate for all mobility models is the fraction $m_i$ of population living at a certain node $i$ travelling to other nodes during each timestep. An estimate of the value of this parameter can be derived from the mobile phone data by dividing the number of trips originating from a node $i$ by the number of cell phone users in the dataset whose home node is $i$.

Figure 13a shows the histogram of the population travelling per day in different *arrondissements*. The seasonal variability of this distribution is low in mean and in variance, as can be seen from figure 13b.
(a) Histogram of the yearly average fraction of population travelling per day by origin node.

(b) Boxplots of the monthly average fraction of population travelling per day by origin node. Note the low seasonal variability in terms of mean and variance.

Figure 13
4 Conclusions

Mobile phone records offer the opportunity to shed light on human mobility patterns at different spatial and temporal scales. The analysis presented herein is guided by the possible application to a spatially explicit model of transmission of waterborne diseases. To that end we have extracted mobility patterns from the coarse grained dataset (set 3, arrondissement level). At this spatial scale every pair of origin/destination has enough data to reconstruct robust statistics on the underlying mobility patterns. Moreover this dataset allows to follow users for an entire year and therefore to possibly detect seasonal trends of the mobility behaviours.

The general model of waterborne disease transmission illustrated in section 1.1 highlights how the two human-mobility related quantities that are deemed crucial for the application of the model at large scale are the fraction of people moving outside their resident community and the distribution of trip destinations. The results show how these two quantities can robustly be estimated from mobile phone records. We have analyzed the distributions of trip destinations given a certain origin in terms of both number of trips and times spent at destination (pause time). As discussed in section 2.1, depending on the type of disease and the related transmission dynamics, both measures can be relevant for pathogen disseminations. Results also show that there are not strong seasonal variations on the mobility patterns at these spatial and temporal scales of observation.

We have also performed a comparative analysis to assess whether, in absence of mobile phone records, one can resort to conceptual models of mobility which rely only on demographic and geographical information. To this end we have compared mobility patterns extracted form phone records with those obtained trough the application of a gravity (Erlander and Stewart, 1990) and a radiation models (Simini et al., 2012). Results show that mobile phone records are more informative and can be used to inferred non trivial mobility patterns that cannot be captured by simple conceptual models.

Overall this preliminary analysis suggests that mobile phone records are a necessary ingredient for the future development of mechanistic models to understand, control, and predict waterborne disease epidemics, as well as to devise appropriate intervention measures.
References


**Prognosis:**

Evaluating Risky Individual Behavior During Epidemics Using Mobile Network Data

**Antonio Lima**  
Massachusetts Institute of Technology (MIT)  
University of Birmingham, UK

**Veljko Pejovic**  
University of Birmingham, UK

**Luca Rossi**  
University of Birmingham, UK

**Mirco Musolesi**  
University of Birmingham, UK

**Marta Gonzalez**  
Massachusetts Institute of Technology (MIT)

**Abstract**

The possibility to analyze, quantify and forecast epidemic outbreaks is fundamental when devising effective disease containment strategies. Policy makers are faced with an intricate task of drafting realistically implementable policies that strike a balance between risk management and cost. Two major techniques policy makers have at their disposal are: epidemic modelling and contact tracing. Models are used to forecast the evolution of the epidemic both globally and regionally. While contact tracing is used to reconstruct the chain of the people that have been potentially infected, so that they can be tested, isolated and treated immediately. However, both techniques might provide limited information, especially during an already advanced crisis when the need for action is urgent.

In this paper we propose an alternative that goes beyond epidemic modelling and contact tracing, and leverages behavioral data generated by mobile carrier networks to evaluate contagion risk on a per-user basis. The individual risk represents the loss incurred by not isolating or treating a specific person, both in terms of how likely it is for this person to spread the disease as well as how many secondary infections it will cause. To this aim, we develop a model, named Prognosis, which quantifies this risk based on movement and regional aggregated statistics about infection rates. We develop and release an open-source tool that calculates this risk based on cellular network events. We simulate a realistic epidemic scenario, based on an Ebola virus outbreak; we find that gradually restricting the mobility of a subset of individuals reduces the number of infected people after 30 days by 24%.

1 Introduction

The world is now facing many severe healthcare challenges and, indeed, the recent Ebola outbreak seems one of the most worrisome and urgent. Mr David Nabarro, Special Envoy of the UN Secretary-General said at an informal UN meeting that he had never encountered a challenge like Ebola in 35 years of his professional life: “This outbreak has moved out of rural areas and it’s com-
ing to towns and cities. It’s no longer just affecting a very well-defined location, it’s affecting a whole region and it’s now impacting the whole world”

Nowadays transportation systems make it possible for people to travel easily across a country and across the globe, but, unfortunately, they make that possible for diseases too. The spread of diseases is facilitated by today’s rich transportation networks that enable human disease carriers to quickly move across distant regions Merler and Ajelli (2010). In this context, drastic measures like banning transportation to disease-affected areas are difficult to implement, have a high cost and are actually believed to worsen the outbreak Chamary (2014) Meloni et al. (2011). The need for smaller, targeted interventions matches the increasing availability of large-scale data, especially coming from mobile networks. The benefit of mobile-phone records to combat quickly-spreading diseases like Ebola is unquestionable Economist (2014).

When an outbreak becomes global, an infected person can be found anywhere, in cities as well as rural areas, and regardless of country boundaries; this might suggest that no place is really safe. However, we argue that some people and places are more exposed to risk than others.

We propose to use such heterogeneity to our advantage and to use mobile networks to unveil such heterogeneity. We envision a system that utilizes the data coming from mobile carriers and, where available, social networks and smartphones, to construct individual-based risk models. The system can assess the risk associated with a person, primarily based on that person’s mobility patterns and, optionally, on other demographic or behavioral indicators that can be inferred from the data. We would like to characterising features of the proposed solution: first, it can use data that is readily available (such as cellphone carrier data), and second, it is be able to operate under uncertainty (it does not require the knowledge of the identity of the infected).

The risk model can be used in several real-world scenarios, especially when urgent response is required. Thus, the model can be used to answer the following questions. Who should be tested early for signs of the disease, and possibly put into quarantine if positive, given that vaccinations can be produced and performed with a certain rate? Who should get vaccinated first? Who should receive information about prevention, for example by means of text messages? All these scenarios describe individual-based interventions that are very hard to administer quickly over large populations. This model can prioritize the people to be targeted with the intervention sooner rather than later.

2 Motivation

People behaviour is highly heterogeneous. Existing epidemic models are based on analyses conducted at population level to assess how infectious a disease is, based on the basic reproductive ratio $r_0$, i.e., the average number of secondary cases generated by a single infected person. However, several studies have investigated that spreading processes are usually highly heterogeneous and some individualism account for a large proportion of the spreading. The presence of these influential spreaders has been investigated for generic networks Kitsak et al. (2010), as well as in epidemics processes. Superspreading seems to be a common featufire of the spread of diseases and targeted individual-based control measures are much more effective than population-wide measures, as reported by Lloyd-Smith et al. Lloyd-Smith et al. (2005). For this reason, identifying superspreaders is extremely important in order to contain epidemics.

Existing techniques, such as contact tracing, are not sufficient. Moreover, efforts in fighting disease outbreaks mainly focus on contact tracing techniques, as it is happening for Ebola Murphy (2014). Contact tracing works by finding all the people who have been in contact with an infected person, and then interviewing, monitoring, isolating them when necessary. The process is repeated for everyone who is found to be infected. While contact tracing can be effective, it has some drawbacks. First of all, information provided by people might be subject to errors, due to fear, shame, faulty memory or other reasons. Secondly, contact tracing needs time: contact tracing only starts when a person is diagnosed with the disease already, or at least shows symptoms. Tracing the contacts also takes time: if the disease has an asymptomatic phase or highly infective, the contacts might be likely to have infected others before they are traced.

\(^1\)http://webtv.un.org/watch/david-nabarro-ebola-virus-outbreak-general-
assembly-informal-meeting-69th-session-10-october-2014/3832613824001
Localization techniques have already been used successfully during critical scenarios. Recently, Nigeria also resorted to GPS technology to improve, scale up and speed up contact tracing, repurposing GPS devices used for polio vaccinations Fasina et al. (2014); Gates (2014). The huge effort of the country resulted in eradication of Ebola and Nigeria was declared “Ebola-free” by the WHO (cite one of the many news that report this). While this success story demonstrates how location tracking can be very useful during similar scenarios, the very same strategy could have not been used if the epidemic was in a more advanced state, i.e., if many more people had already been infected. For this reason, we think it is very important to investigate the use of alternative systems that can provide coarser location tracking but for a very large number of individuals.

Medical treatment is scarce and costly. For example, in the case of Ebola, while it is seen as a serious challenge by the whole world, vaccinations have to face serious technical and financial issues\(^2\). When a commodity such as vaccinations is scarce, who should be given priority to receive vaccination?

3 Risk Model

In this section, we propose a method to quantify the risk associated with each person during an outbreak, depending on their mobility behavior, inferred from their phone-activity. Here we refer it as the risk model. Our goal is not the estimation of the individual cost (i.e., the chance of getting infected), but the cost that an entire community faces by not treating a specific person. Early testing, medical treatment, vaccination, quarantine of specific individuals might reduce cost sustained by the community at later time.

A general estimate of the total risk \( R \) associated to a set of events \( E \) is defined by:

\[
R = \sum_{E} P_{E} \times L_{E}
\]

(1)

where \( P_{E} \) and \( L_{E} \) are the probability and the expected loss for each event, respectively Vapnik (1998).

We bring this definition to the epidemiology domain by considering a scenario in which several geographic areas are associated different values of contagion risk that change in time. The risk measures how likely it is for an individual to get infected in a region. As in common models of infectious diseases, we assume it is directly proportional to the fraction of infected people in the region and we also assume homogeneous mixing within the region. Similarly, we assume that the risk to infect a healthy individual is directly proportional to the fraction of susceptible people in the region.

By staying in a geographic area with non-zero risk, a person will have some chances to get infected; the same person will also have a chance to infect someone else, increasing the risk of the geographic area. When moving between two or more areas, the person will affect the risk of these areas. We will not determine whether each person is in a susceptible, infective or recovered state. Instead, we will consider them in all the states and we will assess how risky their mobility behavior is.

In general, the way people transmit disease across geographic areas has been extensively studied in literature Bajardi et al. (2011); Balcan et al. (2009a); Merler and Ajelli (2010). Most of the studies dealing with the effects of mobility on epidemic spreading usually make the assumption that the mobility patterns of individuals in a subpopulation are homogeneous Colizza and Vespignani (2007), while they are indeed highly heterogeneous Dalziel, Pourbohloul, and Ellner (2013); Merler and Ajelli (2010). This is particularly true for developing countries, where highly irregular and temporally unstructured contact patterns have been observed Vazquez-Prokopec et al. (2013).

We consider a disease that has contagion rate per contact \( \beta \) (i.e., given a friendship between an infected and a susceptible person, a contagion will happen with rate \( \beta \)). Assuming the user \( u \) spends \( T_{u,i} \) fraction of his time in each location \( l \in \mathcal{L} \) (hence, \( \sum_{i} T_{u,i} = 1 \)) we define a time-dependant contagion risk:

\[
C_{u}(t) = \beta \sum_{l,m \in \mathcal{L}} T_{u,l} T_{u,m} [i_{l}(t)s_{m}(t) + i_{m}(t)s_{l}(t)].
\]

(2)

\(^2\)http://news.sciencemag.org/health/2014/10/leaked-documents-reveal-behind-scenes-ebola-vaccine-issues
where \( i_l(t) \) and \( s_l(t) \) refer to the fraction of infected and susceptible population in location \( l \) at time \( t \), respectively. Note that now the probability of the event occurring, in this case, is the probability that a person becomes infected in a region, according to the time fraction spent there, while the expected loss is the number of people expected to be infected in another region, according to the time fraction spent there. As we do not know where the person might be infected, this formula accounts for all the combinations, which are assumed as equally likely. The maximum risk value, for a specific state of the network, is reached by an individual who equally spends his time in the region with the highest infected fraction of individuals and in the region with the highest susceptible fraction. We might calculate this normalized value but, for ranking purposes, it is not necessary, as it is a common factor; we can also ignore the rate \( \beta \) for the same reason.

Our proposed model could also be generalized by defining different risk classes depending on demographic indicators, which can be inferred from mobile data Zhong et al. (2013) or other behavioral indicators, such as those provided with the D4D-Dataset de Montjoye et al. (2014). It is important to emphasize that our model uses only information that is either already collected when outbreaks occur (e.g., estimated number of infected people in various geographic regions) or that can be obtained from telecommunication companies, provided that such use is compatible with existing legislation in the country.

4 Evaluation

Next, we evaluate the effectiveness of the risk identification and containment model proposed above. We set up a realistic epidemic scenario and perform stochastic simulations, following an approach similar to that implemented in GLEaM Balcan et al. (2009a), while keeping track of the movement of individuals following the real traces found in the dataset. We use the SEIR model, where each individual can be in one of the following discrete states at any given time instant: susceptible (S), exposed (E), infected (I), permanently recovered or deceased (R). This model has been used for the 2002 seasonal influenza outbreak Balcan et al. (2009a) and the 2014 Ebola outbreak Althaus (2014), among other outbreaks. It is described by the following set of equations:

\[
\begin{align*}
\frac{dS}{dt} &= -\beta S(t)I(t)/N \quad (3) \\
\frac{dE}{dt} &= \beta S(t)I(t)/N - kE(t) \quad (4) \\
\frac{dI}{dt} &= kE(t) - \gamma I(t) \quad (5) \\
\frac{dR}{dt} &= \gamma I(t) \quad (6)
\end{align*}
\]

We inform a spreading model with the realistic parameters taken from estimates of the 2014 Ebola outbreak in Sierra Leone Althaus (2014), as reported in Tab. 4. Where \( \sigma^{-1} \) and \( \gamma^{-1} \) are the average durations of incubation and infectiousness, respectively. The transmission rate per day in absence of control interventions is \( \beta \), and \( r_0 = \beta/\gamma \) is the basic reproduction number.

| \( \sigma^{-1} \) | 5.3 [days] |
| \( \gamma^{-1} \) | 5.61 [days] |
| \( r_0 \) | 2.53 |
| \( \beta \) | 0.45 |

Table 1: Parameters assumed for the simulation.

We simulate the epidemics in the following different contexts:

- in total absence of any treatment;
- when treatment is given with rate \( \xi \) per day and people given treatment are chosen randomly;
- when treatment is given with rate \( \xi \) per day to highest ranked people, according to the risk measure \( C_u \).
For simplicity, in this paper we focus only on treatment that takes the form of travel restrictions, not allowing high-risk individuals to travel outside the metapopulation they are found when the treatment is applied. This is an extreme scenario, realistic only for diseases for which specific treatments or vaccinations are not available (e.g., Ebola virus). Without loss of generality, we can investigate the effects of vaccination and/or early treatment of people with higher-risk movement patterns. Since we use the same parameters for each metapopulation, and the treatment does not directly affect the epidemic process (i.e., it is not a vaccination or a cure) but only the movement of individuals, the local epidemic profiles will be similar and will be more or less shifted in time, depending on the travel fluxes. We will first show how much we can reduce synchronization by restricting the travel of high-risk individuals in a simple example.

As an illustrative case, we simulate a synthetic model. In Fig. 1 we show the total number of infections since the beginning for two metapopulations, in two specific contexts. Individuals are equally assigned to either metapopulation and they belong to two classes: a fraction of people \((1 - f)\) who do not travel out of their metapopulation, and a fraction of people \(f\) who spend an equal amount of time, on average, in both. We use SEIR with the parameters mentioned before and we initialize the epidemics with a single infected case in one of the two metapopulations, chosen randomly. The top plot \((f = 0.1)\) shows a high level of synchronization, while the bottom plot \((f = 0.01)\) displays a clear delay in the growth of the epidemic size.

We then test our approach initializing simulations with real-data, so that a single randomly chosen region is the unique source of infection with 100 cases. We use the first six months, from January to June 2013, to learn the movement habits of individuals. Then we perform simulations under the three scenarios mentioned above: no countermeasures, people quarantined randomly and people quarantined according to their risk rank. We set an adaptive quarantine rate of \(\xi = \beta i(t)\) to match the countermeasure efforts with the speed of growth of the outbreak. Fig. 2 shows results for the month of July 2013, in terms of how the global prevalence of the disease changes in time in the three cases. Despite the number of randomly quarantined people is pretty high at the end of the month (10% of the population), it does not delay the spreading. Targeted quarantine based on risk, instead, manages to delay the spreading; at the end of the month there are 24% fewer infected individuals than in the baseline cases.

This effect is obtained by restricting individuals who are in the areas with higher risk, specifically those who travel to low risk areas. This determines an increased number of infection cases in high-risk areas, as shown in Fig. 3 and a decreased number of infection cases in low risk areas, as shown in Fig. 4.
Figure 2: The top plot shows how the total number of infected people changes in time when no countermeasures are taken (none), when people are quarantined randomly (random) and according to the highest risk rank (risk). The bottom plot shows how the number of people who have been put into quarantine grows in time. The proposed identification method reduces the number of infected individuals with fewer people in quarantine, using only aggregated information of the number of infected and mobility patterns from mobile phone data providers.

Figure 3: Number of infected (top plot) and quarantined (bottom plot) in the region where the first cases were initialized (hence, a region with higher risk than the others). Our proposed approach determines an increased number of infections in this region, while reducing the total aggregated number of infections.

5 Discussion and Limitations

This model assesses risk using data collected from mobile phones, hence it excludes people who do not use the mobile phones or share them with others. Since mobile penetration rates is already high and increasing in the vast majority of countries, including developing countries, we believe this problem will fade out as time goes. Another potential problem when dealing with network-data is their sparseness of the call activity, but recent studies try to overcome this limitation Leontiadis et al. (2014) by interpolating information in space and time. Furthermore, we would like to remark that the goal of this method is not to find every high-risk individual, but a large proportion of them,
Figure 4: Number of infected (top plot) and quarantined (bottom plot) in a low-risk region. Our proposed approach determines a decreased number of infections in regions that have been less affected by the epidemic, such as the one shown here; this determines a delay in how the global number of infected grows in time.

given the data available. Moreover, it is worth noting that this method might be also be combined with other existing disease prevention and containment techniques already in use, such as contact tracing.

The model described in this paper requires access to sensitive data about individual call and mobility patterns. It is very important to take into account the important ethical and legislative issues arising from the use of these highly personal data. However, solutions based on the analysis of mobile data, such as that presented in this work, can play a critical role during emergencies. For this reason, we believe it is acceptable to use such system when the benefits exceed the risks. We envision the use of such a system only in well-defined circumstances, within specific time intervals and geographic boundaries, within the limits defined by the law and under user informed consent. The model could also be used to design a system that informs users only the users themselves about their own behavior, evaluating their the risk level and, potentially, suggesting them appropriate actions tailored to their risk profile (e.g. get tested, seek help, change lifestyle habits, etc.).

Finally, it is worth noting that we evaluate the model on traces that correspond to an epidemics-free case. People might change their mobility behaviour once they are aware of the epidemic Meloni et al. (2011). Future adaptations of the model might estimate this change by analysing mobility data in real-time.

6 Related Work

Human behaviour can have a significant impact on infective disease dynamics. In turn, a complex interplay of disease spread, awareness of the disease, and population beliefs affect human behaviour Funk, Salathé, and Jansen (2010). The mobility of a person, whether that person is infected or not, is a particularly important factor of disease spread Rizzo, Frasca, and Porfiri (2014). Awareness-induced changes in movement patterns, such as a decision to avoid unsafe infected areas, often have a detrimental effect and might lead to even higher disease spreading, since they result in bringing the infection into previously isolated communities Meloni et al. (2011); Wang et al. (2012). At the same time, international travel restrictions have been shown to have a limited impact on disease spreading, due to the high heterogeneity of human mobility patterns Bajardi et al. (2011). In fact, it is this heterogeneity, both in terms of population behaviour and a-priori infections, that drives disease development. In her discussion of HIV and other STDs transmission Aral argues that bridge groups, such as truckers, the police and the military personnel, transmit infections from highly infected groups, e.g., sex workers, to previously uninfected populations Aral (2000). Our work is
founded on the above observation, and we propose a model that explicitly takes the transmission of risk into account. While previous models consider artificial simulations Buscarino et al. (2014) and long-distance Merler and Ajelli (2010) or multiscale Balcan et al. (2009a) mobility networks in order to quantify possible outcomes of different metapopulations movement patterns on disease spread, we build our model upon individual mobility and interactions, as recorded by fine-grain cellular network traces.

Our work relies on mobile phone call records for estimating risk transfer in a population. The suitability of CDRs for tracking population movements and identification of spatial events in populations has been shown by Bengtsson et al. (2011) and Candida et al. Candida et al. (2008). Furthermore, when it comes to infectivity modelling, in Eames, Read, and Edmunds (2009) Eames et al. show that simple interaction potential measures, such as the total number of a user’s connections (total degree), perform almost as well as more complex measures of interaction, such as individually weighted links. In further work the total node degree might be used to approximate a user’s potential for contact. Finally, in this work we do not modify the interaction network over time. Such modifications, orthogonal to our approach, are discussed in Kamp (2010), and can be accounted for by having a time-dependent contact network.

7 Conclusions

In this paper we have propose Progmosys, a model that goes beyond traditional epidemic modelling and contact tracing, and leverages behavioral data generated by mobile carrier networks to evaluate contagion risk on a per-user basis. The individual risk represents the loss incurred by not isolating or treating a specific person, both in terms of how likely it is for this person to spread the disease as well as how many secondary infections it will cause. We have developed and released an open-source tool that calculates this risk based on cellular network events. We have also simulated a realistic epidemic scenario, based on an Ebola virus outbreak. We have found that gradually restricting the mobility of a subset of individuals greatly reduces the number of infected people.

References


Developing an agent based migration model for Senegal for malaria transmission

ADRIAN M. TOMPKINS

Nicky McCreesh

ABSTRACT

This report presents a preliminary analysis of the D4D dataset to determine the characteristics of journeys that result in an overnight stay. These are considered relevant for malaria transmission. The report then outlines a new, highly memory and computationally efficient agent-based migration model called WISDOM that sets some of its parameters using the D4D analysis. First simulations of the agent based model are shown. These show the agent based model WISDOM reproduces the zero-order patterns of migration involving overnight stays but further improvements are required before the next step of coupling WISDOM to the spatially explicit malaria model VECTRI can be undertaken.

1. Introduction

Population movements drive the transmission patterns and intensity of many communicable diseases. In locations where endemicity is spatially heterogeneous population mobility can also act to transport malaria parasites and may lead to outbreaks in epidemic zones (Martens and Hall 2000; Wesolowski et al. 2012). The strong gradient in malaria transmission intensity in Senegal implies that human mobility may have a significant role in transporting malaria parasites to the northern epidemic-prone districts. Previously such a role of human mobility has been studied using statistical techniques such as diffusion models (Wesolowski et al. 2012). Here the aim is to develop an agent based model for population movements, and to couple this to spatial dynamical model for malaria in order to explicitly account for population movements. To construct the migration model, information about population movements is required. Mobile phone data is one key source of information that can aid this (Tatem 2014), and has been used to assess population cyclic and international mobility in a number of African countries (Blumenstock et al. 2012; Buckee et al. 2013). Here we first make a preliminary assessment of the D4D dataset in order to assess journeys that involve overnight stays and subsequently introduce the new agent based mobility model. Finally the report outlines the next steps to be taken to produce the coupled agent-based transmission model.

2. D4D data analysis

a. method

The D4D dataset was analyzed, using the first 18384 entries in this preliminary analysis. The dataset, which identifies the randomized phone users at the arrondissement level for a period of one year, is studied to identify statistics relating to journeys that result in an overnight stay. We focus on such journeys since the key vectors bite predominately during the evening and night hours (e.g. Braack et al. 1994; Githeko et al. 1996). Thus we assume that shorter journeys where the agent returns home during the same day do not result in transmission of malaria parasites and can be neglected.

The first step of the analysis was the definition of each individual’s home, defined as the place from which a call was made on the greatest number of days. People who lived in or visited locations on the border between arrondissements presented problems for the analysis, as it was difficult to determine where these people lived, and where they spent each night. Locations were therefore recoded, with all observations in locations adjacent to an individual’s home location considered to occur in their home location. Each individuals second most commonly visited arrondissement was then identified, and locations adjacent to that recoded. This was then repeated for their third, fourth, etc most commonly visiting locations.

We attempted to determine where each individual spent each night. Four rules were used. In order of priority, they were:
i. If the last call by an individual on day \( n \), and any call on day \( n+1 \), were from location \( x \), then the individual spent night \( n \) in location \( x \).

ii. If the last call by an individual on any day, and the next call on any day, were from the same location, and there were not more than 60 hours between the two calls, then the individual was considered to have spent any nights between the calls in that location.

iii. If the last call on a day occurred after 7pm, then the individual spent the night in the location from which that call was made.

iv. If the first call on a day occurred before 7am, then the individual spent the preceding night in the location from which that call was made.

The number and destinations of all trips made by each individual were then identified. Each trip was considered to start when an individual made a call from a location other than their home location, and to end when they next made a call from their home location. The destination of the trip was considered to be the location from which a call was made that was located the furthest distance from the individuals home location. Distances were measured as the distance between the centre points of two locations, measured as the crow flies. The proportion of trips that involved a night away were then calculated. A trip was considered to involve a night away if either the night before or the night of the first day of a trip occurred at a location other than the individuals home location. If the location in which an individual spent the night before or the night of the first day of a trip was not known, then that trip was excluded from the analysis.

A number of assumptions were made for the analysis. The ten arrondissements covering the Dakar area (labelled 1-10) were considered to be one location for the purposes of the analysis. In addition, arrondissements 19 and 20 were considered to be one location. The dataset contained a considerable number of erroneous observations that required removing to avoid biasing the statistics. These included events such as two calls by the same individual from different, non-adjacent arrondissements a short time apart. Implausible calls or groups of calls were removed if they would have required a travel speed of greater than 50km h\(^{-1}\), with distance measured as the crow flies between the closet points of the arrondissements. This may have resulted in some genuine calls being removed if individuals traveled by plane, however the low volume of domestic air travel in Senegal means that the effects of this are likely to have been negligible.

b. results

The probability density function of the number of regular migration destinations that individuals visit within the year of data is given in Fig.1. It shows that the majority of individuals either only travel to random locations (i.e. the province is only visited once in the year) or at most have one regular location. Less than 10% of individuals have 3 or more regular locations. As with all the analysis presented in this report, the caveat of the non-representative nature of phone ownership and financial means to make calls is reiterated, which implies that the assessment of population mobility is biased high. This point is adequately described in literature (e.g. Tatem 2014) and is not discussed in detail here.

We next analyze the probability of a journey occurring per day between two points as a function of the distance separating them, considering all journeys made irrespective of whether they involve an overnight stay. We also then analyze the proportion of these journeys that involve an overnight stay. We applied a best fit exponential function to both of these data (Fig. 2) which shows that the overnight proportion increases with an e-folding distance of \( \tau_{\text{overnight}} = 62 \) km. The resulting curve shows that the separation at which overnight stays peaks is at approximately 50km. This relationship is likely to be spatially highly heterogeneous depending on the population density (Simini et al. 2012), and much further analysis is required.

Finally, we show the number of journeys made from each province in Senegal that involve overnight stays (Fig. 3). This map shows that the journey number in general follows the population density (not shown), with the exception of the northern-most provinces within Saint-Louis that border Mauritania. For these the journey number considerable exceed those made from other provinces of a similar population density and distance from the capital Dakar. Possible reasons for this are outlined in the model discussion below.
Fig. 2. Graph of best fit relationship to overnight stays (right axis) and journey probability as a function of distance (left axis).

Fig. 3. Map of total journeys from province marked to other provinces.

3. WISDOM model

a. model description

A new agent based model for cyclic and permanent migration within developing countries has been constructed. The model intends to incorporate personal wealth and welfare into a decision based process to govern decisions regarding both cyclic and permanent movements and is referred to as the Welfare Indexed Societal Demographic Migration model (WISDOM). This first beta release of the code only considers cyclic migration. The model is still under active development and the beta version used only incorporates limited information from the D4D datasets; thus the results presented must be considered very preliminary.

The model presently divides a square domain that includes all Senegal into a regular grid-mesh using a 5km resolution. The population density in each cell is given by the AfriPop dataset (Linard et al. 2012). The model then allocates a weighting between each grid point that governs the likelihood of a cyclic migration journey occurring between the two points. The D4D analysis was carried out in terms of the distance $r_{ij}$ between two locations $i$ and $j$, as the first intention was to use a version of the gravity law to govern this flux, whereby the probability $p_{ij}$ of a journey from points $i$ to point $j$ was a function of the population density in each location $m_i$ and $m_j$:

$$p_{ij} = \frac{m_i m_j}{f(r_{ij})}$$  \hspace{1cm} (1)

In the analysis of the previous section it was seen that an exponential functional form gave a reasonable fit for $f$. However, recently, Simini et al. (2012) have suggested an alternative model referred to as the radiation model that addresses some of the shortcomings of the gravity model, such as nonphysical behaviour in the limits $m_j \gg m_i$ for example. In this model, the distance functionality is substituted by the radial total population within distance $r_{ij}$, denoted $\hat{m}_{ij}$. Simini et al. (2012) then give the journey probability per individual as

$$p_{ij} = K \frac{m_i m_j}{(m_i + \hat{m}_{ij})(m_i + m_j + \hat{m}_{ij})},$$ \hspace{1cm} (2)

where in this context $K$ is the mean migration rate per individual per unit time. From the data we set $K = 0.017$. Simini et al. (2012) show how the radiation model addresses the key short-comings of the gravity model, and we therefore initially tested the model using the radiation model. However it was found that the standard radiation model gave a poor fit since the $\hat{m}^2$ term in the denominator gives a very strong $O\left(r^{-4}\right)$ dependency of the migration probability. In fact, Simini et al. (2012) themselves point out that for the special case of uniform population density, the radiation model reduces to a gravity model with $r^{-4}$ dependency. We note that for the comparison of the radiation and gravity models in Simini et al. (2012), the authors fail to state how the gravity model was fitted or which power dependency was used for the distance function. In order to produce reasonable results we found the need to introduce an empirical modification using a factor $\gamma$ to the radiation model to define the probability weights between two locations as

$$p_{ij} = K(1 - e^{-r_{ij}/\gamma}) \frac{m_i m_j}{(m_i + \hat{m}_{ij})(m_i + m_j + \hat{m}_{ij})}.$$ \hspace{1cm} (3)

We used $\gamma = 0.5$ in a preliminary simulation, but the results below show that a lower value will likely result in an improved fit to the data. By retaining the radiation model framework the advantages of improved analytical behaviour is kept but further work is required to improve the fit of the radiation model to the data. The additional
exponential term in brackets $1 - e^{-\frac{t}{\tau_{overnight}}}$ accounts for the fact that short journeys do not result in an overnight stay.

In this preliminary version of the model, the distance between points is simply calculated as the point to point direct distance and does not account for the road network or national boundaries. For example, overland travel between key locations in Casamance such as Ziguinchor and Dakar involves transit through The Gambia, although this is obviously not the case for flights or ferry travel. Also, presently the model does not distinguish The Gambia from Senegal and allows journeys between the two countries, although these are not presented in the analysis. Future developments of the model will involve an implementation of Dijkstra’s algorithm to include increased weightings for minor roads and transnational routes for example.

The model is initialized using 1000 agents in each 5km$^2$ cell, which is also assigned as the home location for the agent. Thus the initial conditions are somewhat artificial as each agent starts in its home location. Discounting ocean cells in the simulation domain, the simulation includes a total of 2856000 agents. Each agent thus represents a different number of individuals according to the location. For example, in an urban location with a population density of 2000 km$^{-2}$, each agent represents 50 people, while in a sparsely populated rural location of 20km$^{-2}$, each agent represents half an individual. This highlights the fact that individuals per agent is a member of the real set and can take non-integer values. A future development will introduce an adaptive grid mesh to allow a finer spatial resolution to be used in urban locations. Birth and death of individuals is not accounted for in these runs which conduct a one year present day simulation.

Each agent is assigned a probability to migrate per day, which is set to give the same mean number of journeys per individual in a year (approximately 6) as observed in the D4D dataset. In each timestep, a random number is chosen for each agent and used to decide if an agent makes a journey that will involve an overnight stay. The full decision tree of the model is shown schematically in Fig. 4. The destination choice is divided between a regular location or a location chosen at random within Senegal.

As the D4D dataset indicated that the vast majority of individuals had less or equal than two regular migration destinations, at the simulation outset each agent is initialized by selecting three regular locations for each agent. A different probability weighting is assigned to each regular

1As a technical footnote, the model is written using memory-efficient pointer techniques in the Fortran 2003 language and this year-long simulation with approximately 3 million agents can run on a single processor of an average laptop in under a minute wall-clock time. This was a key issue in the model design since it will allow future simulations to be conducted with numerous couplings to malaria, dengue, and HIV transmission modules using a high number of agents to resolve age, wealth and other socio-economic characteristics of society.

Fig. 4. Key decision tree for the v1 beta release version of the agent based migration model WISDOM

location, such that a migration event is far more likely to result in a journey taken to regular location 1, with location 2 and 3 increasingly less likely. In the first model release, all agents are initialized equally, thus the number of agents that make $n$ trips to regular locations, and the number of identifiable regular locations per agent per year of simulation will be Poisson distributed, and underdispersive relative to the data. Again, this will be an aspect that will be improved once socio-economic status is added to the agents’ characteristic matrix to allow for differential probability of migration between agents. If an agent is in a location away from home, it may decide to undergo a further migration to a third location, or alternatively the agent may decide to return to the home location. The probability of returning home $p_{home}$ is greatly exceeds that of making a journey ($p_{home} = 0.38$) to ensure that journey have a duration that similar to observed. Thus the probability of a journey lasting $n$ nights or more is equivalent to $(1-p_{home})^n$. Thus only approximately 4% of journeys involving an overnight stay have a duration of a week or more.

b. preliminary simulations

We show the very preliminary results of the agent based model in terms of the number of individuals arriving in a location per square km per day. Thus the maps are not directly comparable to the D4D data shown earlier and much further analysis is required, but the maps give a first indication of the model simulation (Fig. 5). The model predicts the highest journey flux to/from Dakar as expected, but also identifies the high flux from the western, more highly populated provinces in the vicinity of Dakar, as well as parts of Casamance. One key difference between the observations and model is the lack of higher population movements predicted in the northern counties, which have
Fig. 5. Preliminary results of the WISDOM model simulations. Units are individuals per square km arriving per day.

a low rural population density. Although geographically far from Dakar, they are connected by recently improved N2 highway that runs right through this region. In addition to much of the western counties, the region close to the northern border of Senegal is also predominately Wolof, the major ethnic group politically and numerically within Senegal. Thus one may expect enhanced migration to/from these regions relative to other Eastern counties as many Wolof based in the capital or nearby provinces may have family ties in the region. This highlights the need to incorporate ethnic background into the agent based characteristic matrix. The advantage of the agent-based approach is that the addition of such characteristics is easily accomplished and is memory efficient.

4. Conclusions and next steps

This short report has shown a brief analysis of the D4D Sonatel telephone data in terms of attempting to determine journeys away from home that involve at least one overnight stop, since it is assumed that most transmission is occurring by vectors with evening or night time biting preferences. A new and efficient agent-based migration model, WISDOM, was then introduced. While still very preliminary and under active development, the model was shown to be able to represent the zero-order distribution of journey’s involving overnight stays from the western provinces and Casamance, but lacked the enhanced migration from/to the northern border provinces.

There are several further developments that are required in the model, including but not limited to:

- An improved formula for the journey probability matrix, based on the updated radiation model.
- Accounting for the transport network in the distance weightings, such as differentiating major and minor highway routes, and accounting for air-travel possibilities as well as coastal ferry connections between Dakar and Casamance.
- Accounting for ethnicity in the agent characteristics.
- Accounting for the agent’s socio-economic status to ensure that the distribution of journeys is less under-dispersive.

Once these improvements have been instigated, the next goal is to couple the WISDOM model with the spatially-explicit, climate driven malaria model VECTRI (Tompkins and Ermert 2013), which has been used for forecasting and climate change applications (Caminade et al. 2014; Tompkins and Di Giuseppe 2015). In this way, the impact of cyclic population migration on malaria transmission can be assessed in a fully dynamical framework for the first time.

REFERENCES


Forecasting Influenza in Senegal with Call Detail Records

Hao Wu*, Prithwish Chakraborty†‡, Saurav Ghosh†‡ and Naren Ramakrishnan†‡

*Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA
†Department of Computer Science, Virginia Tech, Arlington, VA 22203, USA
‡Discovery Analytics Center, Virginia Tech, Arlington, VA 22203, USA

Abstract—As part of the D4D Senegal Challenge we describe the use of call detail records (CDRs) in seeding parameters for an epidemiological model around metapopulations. We apply this model to the study of influenza-like illnesses and validate model results against epidemiological surveillance data.

I. INTRODUCTION

Epidemiological surveillance and forecasting has become an established discipline with the availability of myriad direct surveillance and surrogate data sources. Perhaps the most mature methods are available for forecasting influenza-like illnesses (ILI). Researchers have explored the integration of social and physical indicators [1]; while physical indicators (e.g., humidity, temperature, season) contribute the most to performance quality, social indicators (e.g., activity on Twitter) do provide a measurable improvement especially in identifying non-traditional progression of disease. Most recently, researchers have explored the possibility of non-traditional data sources such as restaurant reservations [2] and hospital parking lot imagery [3].

In this paper, we describe the use of mobile call detail records (CDRs) from the D4D Senegal Challenge in seeding parameters for an epidemiological model around metapopulations. We apply this model to the study of influenza-like illnesses and validate model results against epidemiological surveillance data. Related research in this space is discussed in greater detail toward the end of the paper.

II. PRELIMINARIES AND PROBLEM FORMULATION

In this section, we will first introduce some notation and preliminaries to be used in the rest of the paper, followed by the formal definition of the problem studied here.

A. Preliminaries

a) Call Detail Records (CDR): A mobile call detail record (CDR) is a three element tuple \((u, t, \text{loc})\), where \(u\) specifies a mobile phone user (e.g., a numerical ID), \(t\) represents the time stamp that this call event happens, and \(\text{loc}\) denotes the location where this call is originated. The \(\text{loc}\) element can possibly be at different granularity levels, e.g., the arrondissement level or the mobile phone antenna tower level. When the \(\text{loc}\) is at the antenna tower level, further CDR metadata is also available in the format of a four element tuple: \((\text{site}, \text{arr}, \text{lon}, \text{lat})\), where \(\text{site}\) and \(\text{arr}\) represent the antenna tower ID and the arrondissement ID the tower located in, respectively, and \(\text{lon}\) and \(\text{lat}\) denote the longitude and latitude of the antenna tower.

Given a CDR tuple \(r\), we will use \(r[u], r[t]\) and \(r[\text{loc}]\) to represent its mobile user ID, time stamp and location values, respectively. \(\text{date}(r[t])\) is used to represent the date associated with the time stamp \(r[t]\) of the CDR tuple \(r\). When \(\text{loc}\) is at the antenna tower level, \(r[\text{loc}][\text{arr}], r[\text{loc}][\text{lon}],\) and \(r[\text{loc}][\text{lat}]\) are used to represent the corresponding arrondissement area, tower longitude, and latitude values, respectively.

b) Network Notations: A weighted directed network \(G\) is defined as a ternary tuple \(G = (V, E, W)\) where \(V\) is the vertex set, \(E\) is the edge set and \(W\) is the weight matrix for the edges in \(E\). Given any two vertices \(v_i, v_j \in V\), \(e_{ij} = (v_i, v_j)\) represents the directed edge from vertex \(v_i\) to \(v_j\), and \(W(e_{ij})\) denotes the corresponding weight for edge \(e_{ij}\).

c) Disease Spread Model: To simulate a disease spread model we used a modified SIR model with interacting metapopulations in this work. Given any node (arrondisment or tower) \(i\) and time \(t\), we denote the number of susceptible people in the node by \(S_i(t)\), number of infected people by \(I_i(t)\) and the number of recovered people by \(R_i(t)\). At any time point \(t\), the summation over all \(I_i\) is used as the total predicted ILI case count in the country.

B. Problem Formulation

Given a CDR dataset \(R = \{r_i\}\), the problem is to infer a weighted directed network \(G = (V, E, W)\) which captures the mobile phone user mobility information in \(R\), and forecast the spread of influenza by imposing a disease spread model over \(G\).

III. INFLUENZA FORECAST MODEL

In this section, we describe our approach to forecasting influenza from CDR datasets. We will begin by estimating the active mobile phone population at the arrondissement and mobile phone antenna tower levels, followed by a method to instantiate a disease propagation network.

A. Active Mobile Phone Population Estimation

Given CDR data \(R = \{r_i\}\), we first classify \(R\) into two groups according to its location granularity, e.g. \(R_{\text{arr}}\) and
we find all the sampled mobile phone users that are involved
in the CDR set. We estimate the active mobile phone population at the antenna tower level, which could be more specific, the estimated influenza propagation network should be ordered by the corresponding CDR time stamps in increasing order (e.g. time stamp \( t_1 \) being smaller than time stamp \( t_2 \) means the CDR related to \( t_1 \) happens earlier than that related to \( t_2 \)). To be more specific, a trajectory \( T_{d_j} \) for mobile phone user \( u_i \) at date \( d_j \) is defined as:

\[
T_{d_j} = (r_1[\text{loc}], r_2[\text{loc}], ..., r_l[\text{loc}]),
\]

where \( r_k \in R_{\text{tower}} \), \( r_k[u] = u_i, \text{date}(r_k[t]) = d_j, k = 1, ..., l \)

If a location \( r_k[\text{loc}] \) belongs to a trajectory \( T_{d_j} \), we use \( r_k[\text{loc}] \in T_{u_i} \) to represent this aspect. To compute the trajectories from the CDR set \( R_{\text{tower}} \), we just follow the trajectory definition described above.

With calculated mobile phone user trajectories \( T_{d_j} = \{ T_{u_1}^{d_j}, T_{u_2}^{d_j}, ..., T_{u_n}^{d_j} \} \) at a particular date \( d_j \), we will continue to construct the propagation network \( G_{d_j} = (V_{d_j}, E_{d_j}, W_{d_j}) \) for date \( d_j \). The basic idea is to create the network \( G_{d_j} \) in such a way that it captures the mobile phone user flow information between any two of the antenna towers. Thus, we will use each antenna tower as a vertex in the network \( G_{d_j} \), that is \( V_{d_j} = ID_{\text{tower}} \). To construct the edge set \( E_{d_j} \), for each trajectory \( T_{u_i}^{d_j} \in T_{d_j} \), if any pair of consecutive locations in \( T_{u_i}^{d_j} \) is different from each other, e.g. \( r_k[\text{loc}] \neq r_{k+1}[\text{loc}] \) for \( r_k[\text{loc}], r_{k+1}[\text{loc}] \in T_{u_i}^{d_j} \), we add a directed edge \((r_k[\text{loc}], r_{k+1}[\text{loc}])\) into edge set \( E_{d_j} \), and increase its weight by 1. Algorithm 1 illustrates this network construction procedure, where \( \text{len}(T_{u_i}^{d_j}) \) represents the number of locations in the trajectory \( T_{u_i}^{d_j} \).

Finally, we use the average propagation network during a period with the edge weights normalized between 0 and 1 as our estimation of the influenza propagation network. To be more specific, the estimated influenza propagation network
during the period from date $d_1$ to $d_m$ would be:

$$
\hat{G}_{d_m}^{d_1} = (V_{d_m}^{d_1}, E_{d_m}^{d_1}, W_{d_m}^{d_1})
$$

where $V_{d_m}^{d_1} = ID_{\text{tower}}$, $E_{d_m}^{d_1} = \bigcup_{j=1}^{m} E_{d_j}$

$$
W_{d_m}^{d_1} = \frac{W_{d_m}^{d_1}}{\max \left( W_{d_m}^{d_1} \right)}, \quad W_{d_m}^{d_1} = \frac{1}{m} \sum_{j=1}^{m} W_{d_j}
$$

where $\max(\cdot)$ denotes the largest element of the given edge weight matrix.

C. SIR Influenza Spread Model

We used a discrete SIR meta-population model to capture the spread of influenza in the network. For each epidemiological week we consider each node (arrondissements level or tower level) to be a sub-population experiencing a single SIR dynamic process for influenza.

Under discrete approximations, for the node $i$ the SIR evolution equation can be given as:

$$
I_i(t+1) \sim \text{NegBin} \left( \lambda_i(t+1), I_i(t) \right)
$$

where $\text{NegBin}$ signifies the negative binomial distribution and $\lambda_i(t+1)$ denotes the expected number of new infections in unit time in node $i$.

Following similar steps as in [5], we model this expected count as follows:

$$
\lambda_i(t+1) = \frac{\beta(t) \times S_i(t) \times (I_i(t) + \bar{I}_i(t))^\alpha}{N_i(t)}
$$

where $\beta(t)$ indicates the transmissibility of the disease which we model to be independent of spatial characteristics. $\bar{I}_i(t)$ captures the spatial spread from neighboring nodes and $\alpha$ is a factor used to correct for discrete assumptions (see [5]).

We use a gravity model to capture the spatial force $I$. Since influenza data is reported weekly, we assumed the same population estimate $N_j(t)$ for node $j$ for the full epidemiological week. Under this assumption, we model the spatial force as a Gamma process, i.e.,

$$
\bar{I}_i(t) \sim \text{Gamma}(m_k(t), 1)
$$

Here $m_k$ signifies the spatial coupling. Under the generalized gravity model the coupling induced from node $k$ to $j$ can be given as:

$$
m_{k \rightarrow j}(t) \propto I_k(t) \cdot N_j(t)/d_{kj}
$$

d_{kj}$ here signifies the directed edge distance (inverse of edge weight) from node $k$ to $j$. The overall spatial coupling for node $j$ can then be given as:

$$
m_j(t) = \theta N_j(t) \sum_{j \neq k} \frac{I_k(t)}{d_{kj}}
$$

Here $\theta$ is a constant signifying the strength of spatial interactions.

### Table I

<table>
<thead>
<tr>
<th>Method</th>
<th>Percentage Relative Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMA</td>
<td>43.25</td>
</tr>
<tr>
<td>National Level</td>
<td>67.50</td>
</tr>
<tr>
<td>Arrondisment Level</td>
<td>80.25</td>
</tr>
<tr>
<td>Tower Level</td>
<td>70.30</td>
</tr>
</tbody>
</table>

Fig. 1. Average Curves generated by different simulations

IV. Simulation

We used the discrete SIR meta-population model as described in section III and ran multiple simulations over the network to compare the efficacy of using mobility as surrogates for influenza network spread. We randomly initialized a single node with an infected individual and ran the stochastic model described earlier. We used a static $\beta(t)$ for this work. Parameters of this process such as $\theta$, $\alpha$ and $\beta$ play a crucial part in setting up the model. We used a Latin hyper-cube sampler to create a grid for the said parameters and found the best parameters through cross-validation. The cross-validation procedure used the accuracy measure defined as:

$$
\text{accuracy} = \left( 1 - \frac{|\text{actual} - \text{predicted}|}{\max(\text{actual}, \text{predicted}, 10)} \right) \times 100
$$

We ran the simulation with the tower graphs (i.e. where the nodes and hence the sub-populations are at the tower level) and at the arrondisement level. We also ran a discrete SIR process without any spatial force using the national level data (i.e. no network structure) for comparison. Finally, we also implemented a simple ARMA model to compare against the epidemic models. All the influenza data used in the simulations are downloaded from WHO FluNet [6]. In Table I, we present the accuracy results (percentages) while predicting two weeks ahead and assuming full knowledge of the mobile network. As can be seen, the arrondisement level meta-population gives the best accuracy. Also, from ARMA to the arrondisement level we can see a 100% increase in accuracy. For better visual comparison, Figure 1 presents the average plots for each of the different methods.
V. Related Work

In this section, we provide a brief survey of related research. In particular, we discuss work related to estimating human mobility from CDR data and other type of mobile data, and modeling epidemics over networked metapopulations.

A. Human Mobility Modeling

Learning human mobility patterns provides insight into understanding and solving key problems in epidemiology and social science. Thus, concomitant with the development of advanced mobile technology, much active research has been conducted towards understanding human mobility with mobile data. In [7], the authors analyzed billions of anonymous CDR data to characterize the mobility patterns of thousands of people, and explored different aspects of human mobility, e.g., daily travel range and traffic volumes. The authors of [8] studied the travel patterns of 500,000 individuals in Cote d’Ivoire using mobile phone CDR datasets. Through considering both the uncertainty of movements and temporal correlations of individual trajectories, the authors performed a theoretical analysis of the limits of predictability in human mobility.

Beyond mobile phone CDR datamany, research has also explored other forms of mobile data in studying human mobility. In [9], the authors studied the trajectory data of 100,000 mobile phone users whose positions are tracked for a six-month period through their cellphones. They found that individual travel patterns could be collapsed into a single spatial probability distribution, indicating the inherent similarity of human travel patterns. In [10], the authors collected close proximity interactions (CPIs) data from 788 individuals in an American high school using wireless sensor network technology, and thus, inferred the human contact network for estimation of infectious disease transmissions. However, collecting such non-CDR data requires additional mobile devices or software, which may be inconvenient to apply to large scale populations.

Brennan et. al. [11] study the relationship between the human mobility and the spread of infectious disease at a global level. They aim to solve the task of predicting the prevalence of flu-like illness in a given city. The flows of individuals between cities are inferred with geo-tagged twitter status of travelers. The authors of [12] studied the global spread of smallpox after an intentional release event through the simulation over a large-scale structured metapopulation model considering human mobility.

B. Epidemic Modeling over Metapopulations

Balcan et. al. developed and presented the Global Epidemic and Mobility (GLEaM) model in [13], which integrates sociodemographic and population mobility data into a spatially structured stochastic disease approach. The flexible structure of GLEaM makes it suitable for computational modeling of global epidemic spread while considering population mobility at the same time. In [14], by considering three European countries and the corresponding commuting networks at different resolution scales, the authors explored the approach of using proxies for individual mobility to describe the commuting flows and predict the diffusion of an influenza-like illness epidemic. Goufo et. al. presented a fractional SEIR model over metapopulation system in [15] to study the spread of measles between four distinct cities. The condition for the stability of the disease-free equilibrium was discussed, and the numerical simulation showed that infection was proportional to the size of population in each city. Wang et. al. [16] provide a survey on the latest progresses on spatial epidemiology on networked metapopulation, in which empirical and theoretical findings that verify the validity of networked metapopulation modeling are discussed.

VI. Conclusion

Our initial exploration of CDRs shows promise in creating a synthetic model upon which we can impose and study different epidemiological scenarios. Future work will be aimed at capturing behavioral interventions as well as detecting significant shifts of population-level activity and studying their effects on (or influences by) disease progression.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

Mobile Data as Public Health Decision Enabler: A Case Study of Cardiac and Neurological Emergencies

Edward Mutafungwa\textsuperscript{d}, Frantz Thiessard\textsuperscript{b}, M. Pathé Diallo\textsuperscript{b}, Ross Gore\textsuperscript{c}, Vianney Jouhet\textsuperscript{b}, Chiheb Karray\textsuperscript{a}, Nouha Kheder\textsuperscript{a}, Rym Saddem\textsuperscript{a}, Jyri Hämäläinen\textsuperscript{d}, Gayo Diallo\textsuperscript{b}

\textsuperscript{a} Faculté des Sciences de Tunis University of Tunis, Tunisa
\textsuperscript{b} ERIAS INSERM U897, ISPED, University of Bordeaux, F-33000, France
\textsuperscript{c} Virginia Modeling Analysis and Simulation, Old Dominion University, VA, USA
\textsuperscript{d} Department of Communications and Networks, Aalto University School of Electrical Engineering, Espoo, Finland

Abstract

The establishment of hospitals in an area depends on many parameters taken into account by health authorities. We would like to investigate whether data from the use of mobile phones could feed this reflection. In order to do this, we chose two diseases that require rapid hospitalization for their care: myocardial infarction and stroke. The objective of the study is to show the areas in which the absence of a nearest hospital can result in death or serious sequelae in Senegal.

In the approach that we propose, the antenna coverage was estimated by the use of Voronoi diagrams. The real population density in each antenna area was estimated with the mobile population density. A total of 40 hospitals located across the 14 regions of Senegal where considered for the study. The maximum distance around each hospital was estimated to be reached in 90 minutes or three hours (corresponding to the time limit for the two diseases considered). The numbers of expected cases for the two diseases were estimated with the incidence rates of stroke and myocardial infarction in the population, and the number of people in each antenna area. As a result, from the expected 13,508 strokes each year, only 462 (3.42%) will occur too far from a hospital to be able to have the thrombolysis treatment, because 96% of the population can reach a hospital in less than 3 hours. By cons, from the expected 24,315 Myocardial infarctions, 4,241 (17.4%) will occur too far from a hospital to be able to have the balloon treatment because they cannot reach the hospital in less than 90 minutes.

Keywords: Mobile data, Public Health, Stroke, Myocardial Infarction, D4D Challenge Senegal

I. Background

Some medical emergencies require rapid hospitalization for their care. Myocardial Infarction and Stroke are two diseases that fall within this framework and with a known maximum time limit for the treatment.

Myocardial Infarction (MA) is an absolute cardiological emergency, the incidence remains high with 120,000 cases per year in France. According to WHO data, ischemic heart disease is the leading cause of death with 7.2 million of coronary heart disease death over the 50 million annual deaths worldwide. The prognosis remains serious since the MA is still responsible for 10 to 12% of total annual adult mortality. In case of MA, it is possible to perform a mechanical unblocking by the expansion of a balloon in a coronary to be carried out within 90 minutes after the first signs of the crisis.

Stroke is the leading cause in Western countries of acquired disability in adults, the second cause of dementia after Alzheimer's disease (30% of dementias are wholly or partly due to stroke), and the third leading cause of mortality. In Europe, the annual incidence of stroke is between 101 and 239 per 100 000 for men and between 63 and 159 per 100 000 for women. In developing countries, such as Senegal, the burden of stroke and other non-communicable diseases has risen sharply. In Dakar, stroke is the most frequent neurological disease with the highest mortality.

Taking health related issues, thanks to the recent advancements in data analysis over huge amounts of data sources, almost all the determinants of our health – from our individual genetic coding to our particular habits – is becoming knowable. In that context, Big Data may be the future for healthcare. Besides the possibility of achieving personalized medicine (FDA, 2013), cross-linking and analyzing various heterogeneous data sources can help early identification of factors that influence peoples health.

In this paper, we propose an approach based on the use of huge amount of recorded and anonymized mobile data to identify and estimate population at risk of major Public Health issues, in particular stroke and MA. To that end, we rely on data provided in the context of the Data For Development (D4D) challenge launched in 2014 by Orange France Telecom\textsuperscript{1}. This year is the second edition and Senegal is the country concerned.

Senegal is a West African country bordered by 5 countries (Gambia, Guinea-Bissao, Guinea, Mauritania and Mali) and totaling 196,712 km\textsuperscript{2}. The country is subdivided in 14 regions. It is further subdivided by 45 Départements, 123 Arrondissements and by a set of Collectivités Locales. The total population is estimated to 13,508,715 people according to the last General Census of the Population. The capital city is Dakar. It concentrates the main business activity of the country. We provide in Table 1 main figures about the population in Senegal (RGPHAE, 2013).

II. Materials

Dataset used

Orange Senegal Mobile Data

\footnote{1\url{http://d4d.orange.com/en/home}}
The dataset provided by Orange Senegal are based on fully anonymized Call Detail Records (CDR) of mobile phone calls and SMS between the company clients in Senegal between January 1st 2013 and December 31st 2014 (Montjoye et al. 2014). The collected CDR which initially comprises 9 million unique aliased mobile phone numbers have been reduced following two criteria (Montjoye et al. 2014):

- users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset)
- users having had an average of less than 1,000 interactions per week. The users with more than 1,000 interactions per week were presumed to be machines or shared phones.

**Dataset 1**: it contains metadata about the traffic between each antenna for 2013. It includes both voice and text traffic. Table 2 gives examples for voice traffic between sites.

**Dataset 2**: this dataset contains two weeks basis fine-grained mobility data. It is constituted of the trajectories at site (antenna) level of about 300,000 randomly selected users meeting the two previously mentioned criteria. Table 3: example of dataset 2 presents an example of the dataset for a given user. This dataset comprises 25 different files.

Orange provides also a coarse-grained mobility dataset (referred to as dataset 3) which contains trajectories at arrondissement level. We have not used it in the current study.

**Table 1: Main figures about population in Senegal**

<table>
<thead>
<tr>
<th>Region on the map</th>
<th>Name of the region</th>
<th>Number of males</th>
<th>Number of females</th>
<th>Global Population</th>
<th>Area (km²)</th>
<th>Density (/km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dakar</td>
<td>1 579 020</td>
<td>1 558 176</td>
<td>3 137 196</td>
<td>547</td>
<td>5735.3</td>
</tr>
<tr>
<td>2</td>
<td>Thies</td>
<td>896 572</td>
<td>892 292</td>
<td>1 788 864</td>
<td>6670</td>
<td>268.2</td>
</tr>
<tr>
<td>3</td>
<td>Diourbel</td>
<td>716 460</td>
<td>780 995</td>
<td>1 497 455</td>
<td>4824</td>
<td>310.4</td>
</tr>
<tr>
<td>4</td>
<td>Kaolack</td>
<td>474 404</td>
<td>486 471</td>
<td>960 875</td>
<td>5357</td>
<td>179.4</td>
</tr>
<tr>
<td>5</td>
<td>Saint-Louis</td>
<td>453 315</td>
<td>455 627</td>
<td>908 942</td>
<td>193241</td>
<td>47.2</td>
</tr>
<tr>
<td>6</td>
<td>Louga</td>
<td>433 715</td>
<td>440 478</td>
<td>874 193</td>
<td>24889</td>
<td>35.1</td>
</tr>
<tr>
<td>7</td>
<td>Fatick</td>
<td>353 716</td>
<td>360 676</td>
<td>714 392</td>
<td>6849</td>
<td>104.3</td>
</tr>
<tr>
<td>8</td>
<td>Tambacounda</td>
<td>344 475</td>
<td>336 835</td>
<td>681 310</td>
<td>42364</td>
<td>16.1</td>
</tr>
<tr>
<td>9</td>
<td>Kolda</td>
<td>335 018</td>
<td>327 437</td>
<td>662 455</td>
<td>13771</td>
<td>48.1</td>
</tr>
<tr>
<td>10</td>
<td>Kaffrine</td>
<td>282 093</td>
<td>284 899</td>
<td>566 992</td>
<td>11262</td>
<td>50.3</td>
</tr>
<tr>
<td>11</td>
<td>Matam</td>
<td>276 481</td>
<td>286 058</td>
<td>562 539</td>
<td>29445</td>
<td>19.1</td>
</tr>
<tr>
<td>12</td>
<td>Ziguinchor</td>
<td>281 813</td>
<td>267 338</td>
<td>549 151</td>
<td>7352</td>
<td>74.7</td>
</tr>
<tr>
<td>13</td>
<td>Sedhiou</td>
<td>229 468</td>
<td>223 526</td>
<td>452 994</td>
<td>7341</td>
<td>61.7</td>
</tr>
<tr>
<td>14</td>
<td>Kedougou</td>
<td>78 867</td>
<td>72 490</td>
<td>151 357</td>
<td>16800</td>
<td>9.0</td>
</tr>
</tbody>
</table>

Total: 6 735 417 male, 6 773 298 female

<table>
<thead>
<tr>
<th>Number of calls</th>
<th>Area (km²)</th>
<th>Density (/km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>138</td>
<td>547</td>
<td>5735.3</td>
</tr>
<tr>
<td>136</td>
<td>6670</td>
<td>268.2</td>
</tr>
<tr>
<td>121</td>
<td>874 193</td>
<td>35.1</td>
</tr>
<tr>
<td>272</td>
<td>714 392</td>
<td>104.3</td>
</tr>
</tbody>
</table>

**Table 2: Example of voice traffic between antennas**

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Outgoing site</th>
<th>Incoming site</th>
<th>Number of calls</th>
<th>Total call duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-04-01 00</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>138</td>
</tr>
<tr>
<td>2013-04-01 00</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>136</td>
</tr>
<tr>
<td>2013-04-01 00</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>121</td>
</tr>
<tr>
<td>2013-04-01 00</td>
<td>2</td>
<td>5</td>
<td>13</td>
<td>272</td>
</tr>
<tr>
<td>2013-04-30 23</td>
<td>1651</td>
<td>1632</td>
<td>1</td>
<td>3601</td>
</tr>
<tr>
<td>2013-04-30 23</td>
<td>1653</td>
<td>575</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2013-04-30 23</td>
<td>1653</td>
<td>1653</td>
<td>2</td>
<td>385</td>
</tr>
<tr>
<td>2013-04-30 23</td>
<td>1659</td>
<td>608</td>
<td>1</td>
<td>3601</td>
</tr>
</tbody>
</table>

Contextual Data

In addition to CDR related data, the administrative organization of Senegal is provided as well as the antenna GPS coordinates and the arrondissement to which they belong to.

We have also used data from the National Agency of Statistics and Demographics, in particular the last available Senegal General Population and Housing Census (RGPHAE 2013).
The list of hospital is compiled from the online Senegal Medical Directory (SMD, 2014) and the SenDoctor web site (SD, 2014).

Eventually, in addition to the shape files for Senegal provided by the D4D challenge organizers, we used data from OpenStreetMap.org (OSM, 2014).

III. Description of the approach

The overall followed methodology in our study is conducted in regards with the following hypothesis:

- The average incidence rate of stroke in Senegal is estimated to 100 per 100,000 inhabitants. As there is no recent documented study which gives us figures that could be used, we based our hypothesis on the fact that the incidence of stroke in Europe is between 63 and 159 for 100,000 women, and between 101 and 239 for 100,000 men. For Senegal it is supposed to be less.

- The average incidence rate of MA is estimated to 180 per 100,000 inhabitants in France. Based on that, we assume that the incidence rate in Senegal is about 150 per 100,000 inhabitants.

- We base our estimation distance from home to the nearest hospital in case of emergency of stroke on the recommendations of the French High Authority for Health (HAS, 2014). The recommendations estimate that for a severe stroke, the patient needs to be taken in charge no more than 3 hours.

- For MA the maxim time for an efficient management is estimated to 1h30 (90 minutes).

Estimating antenna coverage

The geographical coverage areas of mobile (cellular) networks have often been described using equal-sized hexagonal coverage areas around each cell site. These hexagonal grid models have routinely been used for system performance studies for instance in Third Generation Partnership Project (3GPP) standardization studies (Holma and Toskala 2009). However, in reality the cellular layout is highly irregular due to constraints on the where the cell site could be located, spatio-temporal variations in mobile penetration and population density, the surrounding topography, presence of buildings, and so on (Holma and Toskala 2009).

The use of Voronoi diagrams have been proposed as tessellation that overcomes the inaccurate hexagonal grid cellular representation when compared to real world cellular network layouts (Baert and Seme 2004). In the Voronoi diagram approach, cellular network for an area covered by N cells sites area is subdivided into convex polygonal regions around N points that correspond to the locations of the N sites. The irregular shape of the polygons allows providing a relatively better approximation of cell size by taking into account irregular site locations and promity of neighbouring sites. The Voronoi tesselation generated for the provided 1,666 cell sites in Senegal is shown in Figure 1.

Computing mobile population density

In order to have an overall view of the distribution of mobile population at regional level, we used data from dataset 2. The idea is to aggregate a daily average mobile phones identified in a particular antenna. We make the hypothesis that census corresponds to the place where people are globally. We use a first correction factor α for adjusting identified unique users to the 300,000 two weeks based Orange users.

We then used a second correction factor β to adjust the number of people phoning to the expected number of people in the area according to the census provided by (RGPHAE 2013) and the Orange market share in Senegal.

$$\alpha = U_S * 1.2 * \frac{1}{\text{Oms}}$$

$$\beta = \frac{U_R}{\text{Or}}$$

Where

- $U_S$ is the number of unique users per day and per antenna computed from the dataset 2
- Oms is the estimated orange market share in 2013 according to GSMA survey2.
- The factor 1.2 is obtained by adjusting the total numbers of unique users to 300,000 as of dataset 2.

The α and β adjustment coefficients are important to make the corrections of all counted number of people calling during a given day. This help having an idea of the real number of people at each place (antenna location) knowing that the total number of people should be 13 508 715 according to Senegal population census. For instance, the β correction factor for the Dakar region is 9.05 while it is 160.11 for the Sédhiou region and even 509.43 for the Kédougou region. Table 4: Average number of unique users per site daily gives an example of estimated population by antenna site.

Table 4: Average number of unique users per site daily

<table>
<thead>
<tr>
<th>Site</th>
<th>Region</th>
<th>Unique User/Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dakar</td>
<td>160,24</td>
</tr>
<tr>
<td>1583</td>
<td>Tambacounda</td>
<td>170,05</td>
</tr>
<tr>
<td>1405</td>
<td>Saint-Louis</td>
<td>181,04</td>
</tr>
</tbody>
</table>

2 http://www.gsma.com/
Figure 2 represents distribution of antenna as well as the computed density of the population across the country. Let’s assume that N represents number of people in a given antenna coverage. We have used 5 different colors respectively grey for \( N < 100 \), yellow for \( 101 < N < 1000 \), red for \( 10001 < N < 100,000 \), brown for \( 10001 < N < 100,000 \) and black for \( N > 100,001 \).

**Figure 2: Distribution of Senegal population according to the antennas**

**Estimating risk zones**

Many health geographers use distance as a simple measure of accessibility, risk, or disparity in terms availability of health services in different locations (Dummer 2008). Time-critical medical emergencies like heart attacks and strokes considered in this study require that hospitals that provide emergency care are within a certain distance of the victims requiring immediate care. For the case of medical emergencies due to heart attacks and strokes, locations from which victims are unable to reach the hospitals within a given time are considered to be high risk areas.

In this study we utilize the mobile datasets provided to evaluate the areas that are considered at high risk based on given time criteria. This involves mapping the hospital and populated distribution on to the cellular network layout. A total of 40 hospitals located across the 14 regions of Senegal were considered for the study (see Figure 3 with the capital city Dakar highlighted). These hospitals have been retrieved from the online Senegal Directory and the SenDoctor web site. The geographical location of the hospitals was approximated by representing them using site IDs of the nearest antenna site. Using this approximation it was noted that 85% of the 40 hospitals considered were within 2 km of their real geographical locations (see Figure 4). The impact of this error is minimal when evaluating the travel time to the hospitals.

The segmentation of locations according to their proximity to hospitals is also done based on cell areas. To that end, a common travel time is assumed to reach a hospital from a particular antenna for all people located in the same cell area. Furthermore, the antenna site location is assumed to the centroid of the cell area and the distance to the hospital for cell is the distance between the cell antenna site and the antenna site to which the hospital is associated.

**Figure 3: Location of the 40 hospitals considered in the study. They are represented in red symbols while the road network is shown in black lines. The Dakar region is shown inset.**

**Figure 4 Distance of hospitals to closest antenna sites**

A significant number past health planning studies have used the straight-line (“as the crow flies”) to measure the distance between two points on the map (Boscoe et al 2013). The approach uses either the spherical distance for geo coordinates (latitude and longitude) or Euclidean distance for projected coordinates. However, the real drive distances (and hence travel times) between the two points tend to be longer due to the fact that roads are built around natural obstacles, such as, mountains, boulders and so on. This is clearly visible in the road network of Figure 3. To that end, a correction factor known as *detour index* representing the ratio of the drive-distance to the straight-line distance has been introduced to obtain more accurate distance estimates with less computation or measurement effort (Boscoe et al 2013). The detour index approaches the lower bound of 1 the denser the road network. In developed economies detour indices in the range of 1.2 to 1.6 have been noted.

For this study we calculate the straight-line distance from each hospital to all cell sites and we then use a detour index of 2 to evaluate the drive distances between the points. The detour index assumption is rather conservative to take into account the relative low road network density and the less than ideal road conditions. Furthermore, for simplicity an average driving speed of 60 km/hr is assumed for all areas. An estimate of the travel time ranges from each cell area to the nearest hospital is illustrated in *Erreur ! Source du renvoi introuvable.*. The maximum of 90 and 180 minutes are considered based to the treatment time windows for the cardiovascular conditions considered in this study. The cell
areas beyond the 180 minutes catchment area are considered high risk areas for all conditions.

Figure 5 Estimated travel times from different areas to the nearest hospital. (Green: less than 90 minutes, Yellow: 90 to 180 minutes, Red: over 180 minutes).

From the expected 13,508 strokes each year, only 462 (3.42%) will occur too far from an hospital to be able to have the thrombolysis treatment, because 96% of the population can reach an hospital in less than 3 hours (assuming that all the hospital are able to do the treatment) according to Figure 6.

Figure 6. Estimated incident cases of strokes from different areas too far from the nearest hospital to be treated by balloon. (Grey < 5 victims, Yellow 5 to 25).

By cons, from the expected 24,315 MA, 4,241 (17.4%) will occur too far from an hospital to be able to have the balloon treatment because they cannot reach the hospital in less than 90 minutes (Figure 7).

Figure 7. Estimated incident cases of Myocardial Infarction from different areas too far from the nearest hospital to be treated by fibrinolysis. (Grey < 5 victims, Yellow 5 to 25, Orange 26 to 50, Red 51 to 100).

We have shown that by using a considerable amount of recorded mobile data combined with census data it is possible to perform location based estimation of the people in risk. This will help Public Health decision makers in their early stage decision making.

Limitations of the study

The current study presents some limitations due to bias introduced by the data used and some choices for designing the study. The first bias is related to the extrapolation of population at a given antenna coverage area as it is not possible to estimate precisely the Orange share market in Senegal during the period covered by the provided data. In addition, a filtering on data is performed by Orange as indicated in section 2. Another bias which may affect the results is related the computation of unique users per day for a given antenna site. We did not performed the estimation based solely on the users during night, which is likely to be more accurate. Eventually, we based the study on an estimated incidence rate of the considered medical emergency as there is no official figures.

Future work

There is a room for improvement of the current study. First, we envision to investigate a more fine-grained estimation of the population density. Indeed, currently we perform our estimation by using CDR mobility data at antenna level. Even if we adjust our figures with the census information, we do not take into account difference between night and day, neither more fine grained time frames according for instance to off pick hours.

For the medical emergencies considered in this study (stroke and myocardial infarctions) it has been assumed that all the considered hospitals possess the requisite capabilities for treatment of the emergencies. A more precise studies would take into account the individual treatment capabilities of each hospital to obtain more precise mapping of the high risk zones for the considered medical emergencies.

In order to generalize our approach to other emergency cases, we also plan to base our approach on a domain knowledge model represented by an ontology. The idea is to represent formally and semantically the different emergency case which
may lead to death or irreversible sequelae (cardiovascular diseases, stroke, etc.) and describe the different factors to take into account for an efficient management. To do so, we will rely on semantic web technologies (Shadbolt et al., 2006).

Another issue that needs to be addressed is taking into account the ever growing available data through the Open Data initiative, which make available Linked Open Data (weather conditions, traffic jams, traffic networks, etc.). Coupled with the available Big mobile data, it should be possible to build on-demand or stream-based applications for tackling major Public Health issues.

Conclusion

In this paper, we have described our approach for the identification of risk zones for helping Public Health decision makers to take the required action on the earlier. Two major concerns in Public Health have been considered: myocardial infarction and stroke in the context of Senegal. Thanks to the use of anonymized mobile data provided in the context of the 2014 D4D challenge, we have been able to estimate population in risk.

References

- Boscoe et al. 2013: A Nationwide Comparison of Driving Distance Versus Straight Line Distance to Hospitals, Professional Geographer, 64(2), 2013.
- RGPHAE, 2013: Recensement Général de la Population et de l’Habitat
- SMD, 2014 : The online Senegal Medical Directory Web Site http://www.annuairemedical-senegal.com, consulted on December 30, 2014

Acknowledgments

We would like to thanks the organizers of the D4D Challenge for providing us the data necessary to perform the current study. We also thanks medical doctors from Senegal who helped in getting information about health in this country.

Address for correspondence

Gayo.Diallo@u-bordeaux.fr and edward.mutafungwa@aalto.fi
Modeling Ebola Virus Diffusion in Senegal using Mobile Phone Datasets and Agent-based Simulation

Jonathan P. Leidig*, Christopher Theisen*, Nicholas Vogel*, Doug H. Graham†, Jerry Scripps*, Greg Wolffe*

*School of Computing and Information Systems, Grand Valley State University
†Department of Biomedical Sciences, Grand Valley State University

Abstract—Mobile call detail records assist in capturing human behavioral trends that are not otherwise possible to ascertain. They are especially useful when applied to developing areas due to widespread adoption of cellular devices, general lack of governmental resources, and difficulty of activity modeling. This project mined the Data for Development (D4D) datasets to build latent population and mobility models of the underlying population of Senegal. Combining this information with existing census data, simulation software was developed to model Ebola virus diffusion and transmission routes in Senegal in light of an existing epidemic in neighboring countries. Experiments were performed to study the effect of disease outbreak mitigation strategies and governmental policies for optimization of resources and efforts, e.g., quarantine and border closures. The modeling and simulation software may be used (by conducting simulation studies) to inform Senegalese government policy on disaster prediction, preparation, public health response, and recovery, in addition to recommending best practices.

Index Terms—Ebola, planning, modeling and simulation

I. INTRODUCTION

The cost of infectious diseases has a significant negative impact on the economy, health, and well being of countries. The cost of succumbing to and recovering from infectious diseases often places the highest burden on poor and disadvantaged citizens. These individuals are unable to acquire sufficient preventative, diagnostic, and treatment services. In addition, the effects of an infection are generally more severe in high-risk groups, e.g., pregnant, HIV-positive, and the elderly. Thus, it is especially important to maximize the use of public resources and optimize policies related to health. In West Africa, communicable diseases (e.g., Ebola and meningitis), vector-borne diseases (e.g., malaria and yellow fever), and parasitic diseases (e.g., Schistosomiasis) have a significant impact on health and the economy. Senegal in particular is at high-risk for infectious diseases and epidemics. Medical resources (such as physicians and pharmaceutical treatments) and infrastructure (hospitals and clinics) are limited in many developing areas despite the prevalence of infectious diseases. Thus, it is not straightforward to prevent and respond to epidemics or eradicate diseases. Proper planning is required to best allocate and use available resources, especially if an epidemic threatens to exhaust resource availability. Governmental policies are required for closing borders, closing schools, stopping commerce, surveillance, mass media outreach, distributing pharmaceutical treatments, restricting travel, quarantine, and isolation. However, it is not known a priori which optimal combination of mitigation strategies will best prevent or end an epidemic. Simulation results provide a computational basis for predicting how a given disease will spread in a given population and scenario. Conducting a simulation study allows governmental officials to set policy based on predicted future events, costs, and attack rates.

Several simulation software tools have been developed to simulation the spread of diseases [3], [4], [7], [18]. These tools have been used extensively in setting public health policies for several national governments and aid organizations. The tools handle a range of diseases including avian flu, pertussis, smallpox, and malaria. However, the current generation of software tools must be modified to accurately simulate the spread of Ebola. These tools have been used to set policies in several developed countries with population models built on government census records and individual activity questionnaires. The tools require models for populations, social networks, individual behavior, movement, and diseases. However, these models have not been developed or parameterized to work well with developing countries and the Ebola virus. Many assumptions and population models produced for developed countries break down when applied to West Africa. There are differences in family and household sizes, age structures, school sizes and attendees, lifestyles, social networks, movement models, migration, seasonal population shifts, transportation infrastructure, and the locations and availability of healthcare resources. Limited governmental resources and the difficulty of travel in certain areas have prevented some countries for producing censuses and activity models. With the scarcity of information regarding remote areas, it has not historically been possible to conduct accurate simulations regarding these regions.

Mobile phones are ubiquitous in developing countries. Anonymous call detail records (CDR) provide metadata regarding the time and location a person sends or receives a call and/or SMS message. With anonymized CDR datasets, researchers are able to track relative population levels in each area of the country, individual movements, seasonal locations, population shifts, and migration. With these datasets, data mining as applied to the frequency and timing of calls and/or texts make it possible to identify population trends. With recently made available CDR datasets, population and movement models can now be produced that enable the simulation of epidemics in these areas. Population models of Senegal now contain high-resolution detail on population travel, daily movement, and interactions.

To set public policy in Senegal, public health officials may utilize the newly developed models and modified simulation software. This approach has been benchmarked and calibrated against predictions by CDC models.

II. EBOLA BACKGROUND

Ebola virus causes a severe, often-fatal disease in humans. It was first identified in 1976 during two simultaneous outbreaks in Sudan and the Democratic Republic of Congo (DRC, then Zaire), the latter outbreak occurring near the Ebola river for which the virus was named. Since then, there have been 22 documented outbreaks of Ebola virus disease (EVD; formerly Ebola hemorrhagic fever) in sub-Saharan Africa, with case-fatality rates (the proportion of those infected who die) ranging from 25% to 89%, and total cases ranging from a few dozen to the unprecedented case count of over 18,000 in the current outbreak in West Africa [5].

Ebola virus is an RNA virus in the genus Filoviridae. To date, five different Filovirus strains have been
identified, four of which are known to cause disease in humans: Zaire ebolavirus (EBOV), Sudan ebolavirus (SUDV), Tai Forest ebolavirus (TAFV), and Bundibugyo ebolavirus (BDBV). The reservoir of Ebola virus is thought to be several species of bat, which have been shown to harbor the virus asymptomatically [14]. The virus is also capable of causing disease in a number of wild animal species, many of which, including fruit bats, are part of the local diet in many parts of Africa [17]. Outbreaks are sparked by a spillover event during which a human (i.e., the index case) becomes infected through contact with an infected animal or contaminated ‘bush meat,’ and subsequently propagated via person-to-person transmission.

Ebola virus targets and replicates within cells of the immune system, where it then disseminates to the lymph nodes, liver, and spleen. It subverts proper immune system function by disabling several key anti-viral mechanisms, and induces excessive release of pro-inflammatory cytokines (chemical mediators), which in turn leads to systemic inflammation, dysfunction of the clotting cascade, destruction of blood vessels, and multiple organ failure [2]. Widespread internal bleeding leads to shock and ultimately death. Toward the later stages of infection, viral load in patient tissues, especially liver, spleen, and blood, can exceed 10^6 pfu/ml [19]. It is also abundant in body fluids (urine, feces, saliva, sputum, sweat, vomit, mucus, tears, breast milk, and semen), which facilitates transmission during close contact.

Following an incubation period of 2 - 21 days, infected individuals experience an acute onset of fever, headache, weakness, vomiting, and diarrhea. Patients are considered infectious at the onset of symptoms, at which point virus can be transmitted via direct contact (through broken skin or mucous membranes), through contact with the aforementioned fluids, as well as via contaminated objects (e.g., syringes) and surfaces. Since Ebola virus is transmitted in a direct, person-to-person manner, individuals at greatest risk of infection are clinicians in healthcare settings, and those caring for sick family members at home. Contact tracing often reveals a chain of transmission sequentially going through every member of a family.

The rate at which Ebola virus spreads through a susceptible population is described by a parameter called the basic reproductive number, or R_0. It is the average number of secondary infections generated by an infected index case. When R_0 drops below unity, an epidemic eventually stops. As of September, estimates of R_0 for the 2014 outbreak in West Africa were 1.51 in Guinea, 1.59 in Liberia, and 2.53 in Sierra Leone [1]. These are consistent with R_0 estimates from two previous EVD outbreaks: 1.3 in the DRC and 2.7 in Uganda [8]. R_0 is an important summary measure of the ‘strength’ of an epidemic and plays a key role in determining the scale and extent of required control measures such as patient isolation, school closures, and cancelation of social and economic gatherings [9].

There are many challenges to controlling EVD outbreaks in Africa. Their timely identification is often hampered by the non-specific nature of the early symptoms, which are often misdiagnosed as malaria or any number of other endemic diseases. This, combined with a lack of epidemiological surveillance and diagnostic capability, leads to critical delays in detecting outbreaks, and allows the virus to spread from remote settings into areas of higher population density. Based on published estimates of a number of epidemiological parameters from past EVD outbreaks (e.g., incubation period, infectious period, time from illness onset to death, R_0, case fatality rate), Chowell and Nishiura [9] developed a simulation model to illustrate the relationship between the timeliness of control interventions, and the likelihood of an Ebola outbreak occurring. Their results indicate that a delay in outbreak detection on the order of one week still affords a roughly 80% chance of preventing a full-blown epidemic, whereas with a delay of 30 days, this probability fell to below 20%. To put this in context, the current West African outbreak was not recognized by local authorities until it had been under way for close to three months [12].

Additionally, the region suffers from an extremely low ratio of health care workers to general population, is under-resourced in the way of essential personal protective equipment (gloves, gowns, masks), and lacks the public health infrastructure necessary to effectively trace contacts and isolate infected individuals. Epidemic spread is further amplified by traditional funeral practices, which involve communal touching of the diseased. In Sierra Leone, one funeral was linked to 365 EBV deaths, and in Guinea an estimated 60% of all cases are linked to traditional burials [21].

Prior to 2014, EVD outbreaks occurred in more or less remote and isolated settings, a fact that facilitated their eventual control. The current outbreak in West Africa is the first time that Ebola virus has moved into urban environments. The comparatively high population density of the three capital cities involved has exacerbated the spread of the disease and severely hampered efforts to bring the epidemic under control. Given the high rate of population growth projected for this region in the coming decades [11], the inexorable increase in connectivity and travel between its hinterlands and urban centers, and the (thus far) lack of an effective vaccine or antiviral drugs, future outbreaks of EVD on a large scale are likely.

III. Public Health and Healthcare Policy

The field of computational epidemiology makes use of computing to improve the health of a population. Modeling and simulation provides healthcare planners with the ability to predict the results of scenarios. These scenarios consist of a hypothetical situation consisting of hundreds of variables. The population for the simulation may consist of a village, arrondissement, department, nation, or set of countries. Research groups have produced software applications for computational experimentation related to the spread of diseases. These tools typically require population models, activity models, and disease models. Population models include individuals’ demographics, population age structures, and population densities. Activity models describe travel patterns, daily schedules, and interactions between individuals. Disease models required for these simulation applications have been developed during research concerning the spread of diseases in other countries. However, disease models for Ebola have not gained widespread adoption and use in the computational epidemiology community due to the slow emergence of concrete facts of the current outbreak. With these models and datasets, a software application is needed to predict the diffusion of a disease throughout the population. Stochastic software models are used to predict the probabilistic spread between individual hosts.

Public health actions and mitigation strategies play a large role in preventing the emergence of an epidemic and facilitating its eradication. Experimental studies in the field of computational epidemiology are conducted by executing the simulation software using the underlying models and datasets in order to compare the results of thousands of scenarios. These results then lead to the identification of best practices.

IV. Simulation Details

The software produced in this research was modified from an open-source software package, FluTE. FluTE is a stochastic, agent-based simulation system for modeling the spread of influenza, based upon United States census data [7]. The program was modified and rewritten to model the direct-contact transmission of the Ebola virus within the Senegalese population. The process for producing a simulation engine for Ebola in Senegal required three steps.
A population model of Senegal was produced using content provided in the D4D Challenge.

A mobility model for travel and movement within Senegal was produced using content provided in the D4D Challenge.

FuTe’s source code was modified to provide a platform for Ebola transmission and Ebola related parameters.

The simulation system requires four datasets as input to produce a prediction: geo-political models, worker movement data, employment data, and scenario configuration files. This section presents a summary of and justification for the modifications of the existing computational epidemiology platform, as well as a discussion of the limitations of the simulations.

A. D4D-Informed Synthetic Populations

Access to CDR data provided by D4D and the Senegalese government allowed the development of a population model previously unavailable to stochastic, agent-based simulation platforms. Active user ids, defined as having at least one CDR record, were used to provide details of human mobility and locations based on arrondisements centers and antennas. The antenna-based dataset consists of 25 two-week duration records each with 300,000 individuals and overlap of some individuals between records. When synthesizing Senegal’s population, the ratio of the percentage of mobile users per geographic antenna location, derived from CDR dataset, was used to scale the spatial population density for over 1,600 antenna locations. By multiplying Senegalese census regional population data by this ratio, the population was distributed with fine-resolution. The population’s age distribution was applied to each area and antenna range. Using the cumulative ratios of age distribution, ages were assigned to each individual. Household size distributions were constructed as an array using Senegalese household size data [15]. Using the cumulative ratios of household sizes, the size of each household was randomly assigned. In constructing the set of individuals occupying a household, an adult of working age was first added to the household (either aged 15-24 or 25-64), with associated probabilities based on population age distribution. If the household size was larger than one, the other individuals were randomly assigned according to age distribution probabilities of the population. Figure 1 provides an overview of antenna-based population modeling. Figure 2 displays several geospatial factors in the spread of Ebola in Senegal, e.g., roadways, border crossings, and sampling locations (i.e., antenna locations).

B. D4D-Informed Mobility and Activity Modeling

The population, worker-flow, and employment models classify where individuals live, work, and travel based on observed traces of an individual’s movement. Each individual’s home and work locations were based on the most frequently used antenna by a person when making or receiving calls or texts between the times of 7:00pm-7:00am (home) and 7:00am-7:00pm (work). The working population was based upon the population derived from CDR and employment data. The employment data, such as working age population and percentage of employment, is used to model the percent of employed working age individuals who travel to non-home locations for work. The antenna-based work population (7:00am-7:00pm) was multiplied by the percent of working age individuals and the percent of employment in each area in order to randomly assign employment to individuals at each antenna location. This brings workers into contact (through the daytime movement of individuals) who are located at work within the same antenna range. This mobility model improves the accuracy of social connections and mixing between agents. This is done through the datasets of mobility traces for hundreds of thousands of random individuals located throughout the country. Previously, simulations of remote or developing areas relied on coarse-grain, fully mixing models that were developed from the same assumptions made for developed countries with different social patterns or else extrapolated from limited, small scale surveys conducted by workers on the ground. Through the D4D datasets, it is now possible to assign mobility and activities in remote locations based on actual observed behavior.

C. Disease Modeling

Along with the construction of D4D-informed population models, the simulation software was modified to support prediction of Ebola. Table I details several modifications required to properly describe Ebola. In contrast with previously studied diseases, the incubation period is many times longer the standard duration period of a few days. The incubation time period was changed to 2-22 days, depending upon each individual [5]. This also delays the peak of the disease and leads to a slow growing, multi-year epidemic. While influenza outbreaks are expected to spread in terms of seasons, Ebola outbreaks require more computation due to the longer days required to simulate each scenario. Burial rituals were added to the simulation platform based on cultural practices and may be modified through intervention policies.
Some features required to simulate Ebola already existed as parameters in FluTE. However, minor software updates and scenario file parameterizations were required. The ascertainment delay in Table II is based on the current time required to diagnosis Ebola in a patient using experimental lab tests. This is the time delay before viral detection in an individual is possible. This range represents the earliest length of time in which Ebola is detectable with current lab procedures [6]. The viral load trajectory (which determines the extent to which the virus replicates within a host) was also modified. Two viral load trajectories are possible in our version, one resulting in death. This was based upon historical and current case fatality rates of Ebola cases, ranging from 40-70% [19]. As FluTE does not simulate death, this aspect was added in order to investigate the effects of burial practices. In terms of the simulation, after the person is deceased, they are isolated at home to simulate potential burial practices in which a family member(s) prepare the deceased person before burying them. A range of 1-3 days after death was used to represent that a body would be present in the household [13]. During this time, the deceased individual has their viral load set to the highest level within their trajectory to simulate the high viral loads seen in patient just before and after death was used to represent that a body would be present in the household [13]. During this time, the deceased individual has their viral load set to the highest level within their trajectory to simulate the high viral loads seen in patient just before and after dying from Ebola [10]. After the person’s funeral, they are set to recovered and removed from the population. Multiple parameters are used to implement governmental policies, e.g., closing schools and borders. The remaining parameters are based on natural biological factors of the host and disease. The software allows seeding infected persons in major ingresses, such as airports. Thus, geospatial datasets regarding the location of airports, ports, and border crossings were added to the software’s underlying assumptions. In addition, the software includes features such as mitigation-limiting measures to reduce the extent of an outbreak. In order to transition from influenza models to Ebola, the basic reproduction, the number of secondary infections generated from an infectious person, denoted $R_0$, was changed to the range 1.51-2.53 [1].

There were aspects of FluTE that are not applicable to the Ebola disease model. One of these aspects is antiviral kits and vaccinations. Although research continues toward this goal, neither pharmaceutical intervention is available for mitigation efforts. Another inactivated aspect is temporal seasonality, as historical outbreaks of Ebola do not correlate to specific times of the year [16].

There are several limitations to our simulation. There is a general lack of knowledge regarding workforce, workplace locations, and school structures. Additionally, the model does not deal with the locations and resources available at entities such as hospitals, health clinics, and caregivers. Another limitation is smoothing the number of places that a person travels to a few locations, e.g., home and work. This aspect could be enhanced to demonstrate the richness of CDR data available.

## V. Study Results

The simulation software may be used to answer public health questions regarding the spread of Ebola in Senegal. Multiple simulation runs were conducted as a demonstration of the software. Additional studies are required to answer specific questions of interest. As examples, studies might attempt to answer “what is the affect of closing a specific border” or “what is the affect of various compliance rates regarding a potential nationwide call to modify traditional burial practices.”

### A. Alignment with CDC Models

Producing the simulation model required extensive modifications of the FluTE simulation platform. To verify the results of the system, worse-case scenarios were compared with output from a publicly available model provided by the Centers for Disease Control and Prevention [20]. The CDC model provides a prediction of daily Ebola infections for a generic population of a user defined size. However, the model does not take the population structure, dynamics, or topology into consideration or allow for predicting the effects of mitigation strategies. However, the model does provide a baseline expectation for the spread of Ebola in the absence of any public policy or basic actions taken by individuals (e.g., staying home when sick).

Figures 3 and 4 display the results of the two models. For these figures, simulations were run with population of 13,401,076, one initial index case, and parameters encoding current assumptions regarding the characteristics of Ebola. The D4D-informed model produces results in alignment with CDC predictions in addition to providing finer-resolution information regarding each infected individual. As shown in Table III, the CDC model predicts between 6.9 and 11.9 million infections depending on the average number of days a person is infectious. This is in alignment with the stochastic model’s prediction of 8,971,606 infections.

![Fig. 3. Simulation results from the D4D-informed model detailing the count of individuals that have been infected by the end of that day (cumulative symptomatic) and currently infected on that day (symptomatic).](image-url)
The datasets provided through the D4D challenge were utilized in each stage of the software building process. The data provided by dataset 2 and 3 were used to provide a higher-resolution population models and densities than available in a recent census. These datasets were also used to determine the locations visited by individuals starting from their home location. Travel patterns were then developed based on the set of individual routines and applied to the entire population. These datasets provided the ability to determine mobility patterns within short distances in large cities and between large areas in remote locations.

Further work is needed to develop population models of additional countries in West Africa. Of particular importance are the countries between Senegal and Ivory Coast, where Ebola cases have been concentrated. Additional models will allow for the simulations to predict the impact of mitigation efforts in light of the entire region.

**ACKNOWLEDGEMENTS**

We thank France Telecom-Orange and the Data 4 Development Challenge for providing access to mobile datasets regarding Cote d’Ivoire and Senegal.

**REFERENCES**


Fig. 5. Cumulative simulation results from the D4D-informed model detailing the count of infected individuals by day.

Fig. 6. Daily (current) simulation results from the CDC model detailing the count of infected individuals by day.


Towards Connecting People, Locations and Real-World Events in a Cellular Network

R. Trestian, P. Shah, H. Nguyen, Q-T. Vien, O. Gemikonakli, B. Barn
School of Science and Technology, Middlesex University, London, UK
{r.trestian, p.shah, h.nguyen, q.vien, o.gemikonakli, b.barn}@mdx.ac.uk

ABSTRACT
Being able to react fast to exceptional events such as riots, protests or disaster predictions is of paramount importance, especially when trying to ensure peoples’ safety and security, or even save lives. In this paper we study the use of fully anonymized and highly aggregate cellular network data, like Call Detail Records (CDRs) in order to analyze the telecommunication traffic and connect people, locations and events. The goal of this study is to see if the CDR data can be used to detect exceptional spatio-temporal patterns of the collective human mobile data usage and correlate these ‘anomalies’ with real-world events (e.g., religious festivals, public concerts, traffic congestion, riots, protests etc.). These observations could be further used to develop an intelligent system that detects exceptional events in real-time from CDRs data monitoring. Such system could be used in intelligent transportation management, urban planning, emergency situations, network resource allocation and performance optimization, etc.

Keywords
Cellular Networks, Human Mobility, Call Detail Records

1. INTRODUCTION
In the ever-evolving telecommunication industry, smart mobile computing devices have become increasingly affordable and powerful, leading to a significant growth in the number of advanced mobile users and their bandwidth demands. This, together with the improved next generation telecommunications infrastructure, motivates the continuing uptake of the mobility around the world. People can now connect to the Internet from anywhere at any time, while on the move (e.g. on foot, in the car, on the bus, stuck in traffic etc.) or stationary (e.g., at home/office/airport/coffee bars, etc.). The number of mobile users increases continuously as the penetration of both fixed and mobile broadband solutions becomes more affordable for the masses and more accessible around the globe. The connection to the Internet is possible and can be done via wireline or wireless solutions. Depending on the user location, wireless connectivity is enabled by different Radio Access Technologies (RATs) such as: Global System for Mobile Communications (GSM), Enhanced Data Rates for GSM Evolution (EDGE), Universal Mobile Telecommunications System (UMTS), High Speed Packet Access (HSPA), Long Term Evolution (LTE), Worldwide Interoperability for Microwave Access (WiMAX), Wireless Local Area Networks (WLAN), Wireless Personal Area Network (WPAN), etc. Use of all these RATs is rapidly spreading, covering various geographical locations in an overlapping manner.

Additionally, this increasing expansion of the telecommunication infrastructure could bring economic, social and technological benefits especially to the far reaching regions. For example, it can bring education to the remote regions; it can contribute to enabling innovations in healthcare (e.g., remote monitoring and diagnostics), smart grid solutions, social networking sites, economy, etc.

One of the key characteristics of these mobile networks and the mobile computing devices is that every time they are used a digital signature is recorded. Voluntarily or not, whenever people interact with the telecommunications networks or any type of social media platform, they leave behind digital traces. All these traces have become a powerful tool to analyze human behavior patterns. For example, the data collected by the cellular telecommunications systems referred to as Call Details Records (CDRs) is done on a regular basis for billing and troubleshooting purposes. Moreover, these CDRs contain the information details about every call carried within the cellular network, including information about the location, call duration, call time, and both parties involved in the conversation. Thus there is an increased interest on making use of the information provided by the CDRs in order to analyze the human mobility cheaply, frequently and especially at a very large scale. In general, understanding the human mobility patterns could have broad applicability on a wide range of areas, such as: network resource optimization, mobile computing, transportation systems, urban environment planning, events management, epidemiology, etc.

In this work we explore the use of anonymized CDRs containing both voice-calls and SMS activities, from a cellular network in Senegal in order to connect people, locations and events. The goal of this study is to identify the exceptional spatio-temporal patterns of the collective human activity from fully anonymized and highly aggregate cellular network data, like CDRs, and correlate these ‘anomalies’ with real-world events (e.g., religious festivals, parades, public concerts, riots, protests, conflicts, etc.). These observations could be further used to develop an intelligent system that detects exceptional events in real-time from CDRs monitoring. The benefits of such systems could be threefold: (1) the network operators could benefit by detecting congested cells and optimize their network resources in advance of an exceptional event, e.g., make use of the wi-fi offloading solutions, enabling adaptive bandwidth allocation to their radio cells, etc.; (2) the society could benefit from intelligent transportation and urban planning and management; (3) the individual could benefit from traffic information and prediction, emergency management. For example, a real-time event detection system could be used in case of emergency situations, such as riots protests which could be more efficiently handled if detected and handled on time.

Within this context, our research questions are: can the CDR data be used to detect exceptional spatio-temporal patterns of the
collective human mobile data usage? Can we correlate these exceptional usage patterns to real-world events?

2. RELATED WORKS

Recently, there has been extensive academic research related to the use of user-generated traffic in mobile communications networks as a powerful tool to analyze human behavior. As billions of people around the world own at least one mobile device, the data from the mobile communications networks could help study different aspects of human mobility and their interactions on a large scale. This section provides a survey of the current research on the use of user-generated traffic within the mobile communications networks to understand different aspects of human movement and their interaction. The existing approaches in this area are divided in three main categories: (1) definition of the universal law for human mobility; (2) urban planning and real-time traffic forecast; and (3) human localization and mobility patterns.

2.1 Definition of the Universal Law for Human Mobility

When looking into understanding the human mobility patterns, several works tried to define some basic laws governing the human motion. Gonzalez et al. in [1] showed that individuals follow simple reproducible patterns despite the diversity of their travel history. The study was conducted by tracking the position of 100,000 anonymized mobile phone users over a six month period. The authors observed that individuals present significant regularity in their trajectories as they often return to several of their highly frequented locations (e.g., home, work). Song et al. in [2] used empirical data on human mobility to show that the predictions provided by the continuous time random walk (CTRW) models conflict with the empirical results. The authors made use of two data sets: (1) the first dataset contains the CDRs of 3 million anonymized mobile phone users over a one year period; (2) the second dataset contains the hourly location record of 1,000 anonymized users who signed up for a location-based service, over a two week period. The authors explore the limitations of the traditional random walk models and propose a new individual mobility model that takes into consideration the exploration and the preferential return of the individual. The exploration refers to the probability of the individual moving to a new location. As the authors state, observing the trajectory of an individual long enough it becomes harder to find unexplored locations within the vicinity of their home or workplace. The preferential return refers to the probability of the individual returning to one of the previously visited locations. It has been shown that human tend to return to the highly visited locations, like home or workplace, which is not considered by the random walking models. Another study on the limitations of predictability in human mobility is provided in [3]. The study is conducted using the CDRs of 50,000 anonymized mobile phone users over a three months period. The authors observe that most of the individuals can be well localized within a specific neighborhood with only few users traveling widely. The study shows that there is a probability of 93% that the human location could be predicted regardless of how far the person travels within the preferred locations. Palchikhov et al. in [4] showed that human mobility can be predicted by using a simple model based on the frequency of the mobile phone calls between two locations and their geographical distance. The authors use the data provided by Orange for Ivory Coast, consisting of CDRs from 50,000 anonymized mobile phone users collected over a 150 days period. Using only the aggregated call data and the geospatial information as inputs, three different models were tested: the gravity model, the communication model based on the number of calls between two locations, and a modified version of the radiation model. The results showed that out of the three models the communication model is the most accurate in this setting.

Table 1 presents a summary of the solutions focused on defining the universal law for human mobility.

Table 1. Definition of the Universal Law for Human Mobility - Solutions Summary

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Application</th>
<th>Number of Users</th>
<th>Period Covered</th>
<th>Location</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>defining individual human mobility</td>
<td>-100,000 anonymized mobile phone users</td>
<td>- 6 months for the 100,000 users and one week for the 206 users</td>
<td>not mentioned</td>
<td>- CDRs – identity of the closest communication tower when the user initiates or receives a call or a text message</td>
</tr>
<tr>
<td>[2]</td>
<td>defining individual human mobility</td>
<td>-3 million anonymized mobile phone users</td>
<td>- one year for the 3 million users</td>
<td>not mentioned</td>
<td>- CDRs – identity of the closest communication tower when the user initiates or receives a call or a text message</td>
</tr>
<tr>
<td>[3]</td>
<td>limitations of predictability in human mobility</td>
<td>-50,000 anonymized mobile phone users</td>
<td>- 3 months</td>
<td>not mentioned</td>
<td>- CDRs – identity of the closest communication tower when the user initiates or receives a call or a text message</td>
</tr>
<tr>
<td>[4]</td>
<td>human mobility</td>
<td>-50,000 anonymized mobile phone users</td>
<td>-150 days from December 1, 2011 to April 28, 2012</td>
<td>Ivory Coast</td>
<td>- CDRs – identity of the closest communication tower when the user initiates or receives a call or a text message</td>
</tr>
</tbody>
</table>

2.2 Urban Planning and Real-Time Traffic Forecast

Several works explored the use of cellular network data for urban planning and real-time traffic forecast. Isaacman et al. [5] look into the mobility patterns of two cities, such as Los Angeles and New York. This study could further help at investigating the environmental impact of daily commutes. Yuan et al. [6] propose a framework referred to as DRoF, Discovers Regions of different Functions (e.g., educational areas, entertainment areas, historic oriented areas, etc.) in a city using a combination of both human mobility data among different regions and points of interests (POI). The proposed framework is evaluated using large-scale and real-world datasets consisting of two POI dataset of Beijing collected in 2010 and 2011, and two 3 month GPS data used to represent human mobility, generated by 12,000 taxi cabs in Beijing in 2010 and 2011. The authors state that their proposed solution outperforms other two baseline methods that base their findings of functional regions solely on POIs or mobility data. Calabrese et al. [7] present a real-time urban monitoring platform that makes use of a broad range of datasets in order to provide a visualization map of the vehicular traffic status and the pedestrians’ movement. The platform was tested for the city of Rome in Italy. The authors aim to provide a visualization tool that gives a qualitative understanding of how the mobile phone data and vehicle real-time location data could be used to provide valuable services in the context of urban planning and tourist management.
Di Lorenzo et al. [8] combine the use of people trajectories and geographical preferences in order to propose a method for evaluating human spatio-temporal activity patterns. The authors used two datasets covering the state of Massachusetts consisting of the individual human trajectories extracted from anonymous mobile phone traces and the geographical features of places in the area, such as land use. The data was collected over a 4 months period from one million unique devices. The authors define a measure of land use distribution used to characterize the human activity on a specific day, and they have identified 4 distinct patterns that can be mapped to a specific number of kilometers traveled in the day. The authors state that these patterns can be further integrated into activity-based transportation models.

A summary of the above solutions is presented in Table 2.

Table 2. Urban Planning and Real-Time Traffic Forecast – Solutions Summary

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Application</th>
<th>Number of Users</th>
<th>Period Covered</th>
<th>Location</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>[6]</td>
<td>-urban planning</td>
<td>-12,000 taxi calls for GPS data</td>
<td>year 2010 and 2011 for the POI data sets</td>
<td>Beijing, China</td>
<td>-POI - coordinates and category like restaurants and shopping malls -GPS trajectory datasets representing human mobility</td>
</tr>
<tr>
<td>[7]</td>
<td>-real-time urban monitoring</td>
<td>-30,000 calls</td>
<td>-weekdays</td>
<td>Rome, Italy</td>
<td>-mobile phone position information, call in progress, SMS sending, handover, etc. -GPS data -real-time traffic noise from sensor networks</td>
</tr>
<tr>
<td>[8]</td>
<td>-human spatio-temporal activity patterns</td>
<td>one million unique devices</td>
<td>4 months</td>
<td>Massachusetts</td>
<td>-individual human trajectories extracted from anonymous mobile phone traces -geographical features of places</td>
</tr>
</tbody>
</table>

2.3 Human Localization and Mobility Patterns

One of the first studies that provided evidence of geographic correlation between users’ interests within a cellular network was conducted by Trestian et al. [9]. The authors categorized the user interests into six groups based on the type of the application they are accessing, such as: mail, social networking, trading, music, news, and dating. The main focus of the study was to correlate these users’ interests with their location, e.g., home or work. Their results showed that in general the users tend to spend a significant fraction of their time in their top three locations only. Additionally, the authors showed that the location affects the applications the users are accessing. For example, at certain locations users might be interested in accessing one particular type of application regardless the time of day.

Another study that focuses on identifying important locations in humans’ live from mobile data traces was conducted by Isaacmanc et al. [10]. The authors propose several algorithms based on logistic regression that are used to identify important places from CDRs and then apply semantic meaning to these important locations, namely Home and Work. The authors state that their algorithms identify the key locations with media errors under one mile.

Zang et al. [11] propose a dynamic profile-based paging/location management technique in order to increase the efficiency of the location management process within a cellular network. In order to propose their solutions, the authors examine the data from a network operator providing CDMA2000 with support for voice, data and SMS services to their customers. The data was collected from hundreds of thousands of users over a one month period in three locations: Manhattan, Philadelphia, and Brisbane. The authors state that the proposed solution increases the average paging success rate across voice/data/SMS calls above 90% in Brisbane and 85% in Manhattan. Additionally such a mechanism could reduce the signaling overhead by up to 90% at a cost of a small increase in paging delay.

Motahari et al. [12] investigated the impact of temporal factors on the randomness and the size of mobility and the spatial distribution. The study was conducted on CDRs collected from several thousands of users in San Francisco area. Two temporal factors were considered in the analysis: the day of the week and the time of the day. The authors studied how these temporal factors impact four characteristics of human mobility, such as: location entropy, radius of gyration, step size, and spatial probability distribution of user locations. The results show that the spatial distribution is most concentrated during work hours and most scattered on the weekends. However, a different pattern is observed during the non-working hours of weekdays where the spatial distribution is concentrated around home, work and the commute path. The authors state that by considering the temporal factors, the place predictions mechanisms can improve their accuracy by 15% compared to the case without the temporal factors. Table 3 presents a summary of the above solutions.

Table 3. Human Localization and Mobility Patterns – Solutions Summary

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Application</th>
<th>Number of Users</th>
<th>Period Covered</th>
<th>Location</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>[9]</td>
<td>-correlation between people’s application interests and mobility properties</td>
<td>-281,394 users</td>
<td>-seven days</td>
<td>large metropolitan area of 1,900 square miles (approx. 5,000 square kilometers)</td>
<td>-CDR containing information about voice calls and text messages</td>
</tr>
<tr>
<td>[10]</td>
<td>-identifying important places in people’s lives</td>
<td>-hundreds of thousands of phones</td>
<td>-78 consecutive days from November 15, 2009 to January 31, 2010</td>
<td>Los Angeles New York</td>
<td>-CDR containing information about voice calls and text messages</td>
</tr>
<tr>
<td>[11]</td>
<td>-location management</td>
<td>-1,061,000 users in Manhattan 543,000 users in Philadelphia 404,000 users in Brisbane</td>
<td>-one month, from February 2 to February 28, 2006</td>
<td>Manhattan, Philadelphia, Brisbane</td>
<td>-per call measurement data, containing: call starting time, call duration, initial cell, final cell, service type, call direction, number of pages, etc.</td>
</tr>
<tr>
<td>[12]</td>
<td>-mobility characteristics and location prediction</td>
<td>-several thousands of users</td>
<td>-4 days, 16 days, and 3 months</td>
<td>San Francisco</td>
<td>-CDRs including voice, SMS, and data sessions</td>
</tr>
</tbody>
</table>

All these studies have shown that understanding the humans’ mobility patterns could be a crucial component in several areas, such as: network optimization opportunities for cellular network operators in handling the explosive growth in traffic observed from CDRs; transportation planning and management, modeling...
commuting flows, content delivery services and context-aware applications, etc.

3. DATA COLLECTION METHODOLOGY AND CHARACTERISTICS

3.1 Data Collection and Preprocessing

In this paper we use the anonymous CDR data provided by the Orange Group within the Orange Data for Development (D4D) challenge. The CDRs are anonymized phone calls and SMS exchanges between more than nine million Orange customers in Senegal. The anonymized CDRs were collected from a random set of cellular phones over one year, between January 1, 2013 and December 31, 2013. The territorial expanse of the dataset on Senegal is illustrated in Figure 1. Senegal is located in West Africa having an area of 197,000 square kilometers and an estimated population of 13.5 million inhabitants. The capital is Dakar, however the country is subdivided in 14 administrative regions, each region having a regional capital. The country telecommunications sector is dominated by mobile telephony with Orange, owned by Sonatel, being one of the leaders in the market, recording two thirds of the cellular market. For the purpose of this study, there are three sets of data provided by Orange Group and described in the following sections.

![Figure 1. Territorial expanse of the dataset – Senegal.](https://maps.google.com/)

3.2 Dataset 1: Antenna-to-Antenna Communication

The first dataset contains the aggregated number of calls as well as the calls durations within one hour, between any antennas pair. The dataset was stored in 12 files each corresponding to a one month interval. All the datasets are provided in Comma Separated Values (CSV) file format. For the Antenna-to-Antenna dataset each line stores information about the date, time, originating antenna, terminating antenna, number of voice calls, and the duration of the voice calls in minutes for a given hour. A second set of 12 files is provided which contain the monthly aggregated number of SMS exchanged between any antenna pair within one hour.

3.3 Dataset 2: Individual Trajectories – High Spatial Resolution Data

The second dataset provides high resolution individual movement trajectories of 300,000 randomly sampled customers split into consecutive two-week periods. The data was stored in 24 CSV files, and in order to protect the customers’ privacy new random identifiers for each customer are chosen in every two-week time period. Each line in the file contains information about the customer identification number, the connection date and time and the antenna identification number they are connected to.

3.4 Dataset 3: Individual Trajectories – Long Term Data

The third dataset contains the long term, low spatial resolution trajectories of the 146,352 randomly selected customers over one month period. The low spatial resolution is obtained by replacing the antennas identifiers with the arrondissement identifier of the antenna the customer is connected to. Senegal has a total number of 123 arrondissement administrative regions. Figure 2 illustrates the arrondissements and their identifiers along with the Orange antennas locations as provided in the datasets. There are a total number of 1666 antennas. Figure 2 also highlights specific regions like Dakar, Pikine, Kaolack, Saint-Louis etc., where some identifiers are highlighted later on in the paper with respect to connecting people and real world events.

![Figure 2. Senegal Arrondissement and Orange antennas location.](https://maps.google.com/)

Each file in this dataset contains information on customer identification number, the connection date and time and the arrondissement identifier that contains the antenna the user is connected to.

3.5 Limitations of the Datasets

Even though the Call Detail Records represent a good source of location information they have several significant limitations:

- they are generated only when the mobile device is engaged in a voice call or exchanges text messages, thus no information about application usage type (voice/text/data) is available.
- the location granularity is at cell tower level or sub-prefectures, no information about the exact user location is provided.
- no information about the individual call duration is provided.

4. TELECOMMUNICATION FLOWS

This section aims at analyzing the characteristics of the telecommunication traffic flows from Dataset 1, in terms of number of voice calls, voice calls duration and SMS exchange.

---

1. https://maps.google.com/
4.1 Antennas Density vs. Population Density
Apart from the 14 regions, the country is further subdivided by 45 departments and 123 Arrondissements. We made use of the information provided by GeoHive\(^2\) about the population census estimates for 2013. Based on this information, we projected the map of Senegal so that the 45 regions of the country are represented by a color proportional to its population as illustrated in Figure 3. Table 3 lists the main cities with the higher population sizes.

The placement of the antennas within a telecommunication network represents an important decision for any mobile service provider. This will determine how many people will be able to access the network and the quality of the calls. A crucial factor is given by the population density. Wherever there are more people, the density of the antennas should be higher. However the rural population is very important as well, as the access to the mobile communication network could impact their development, e.g., access to real-time agricultural services, access to education, etc.

Table 4. Senegal Main Cities and Population

<table>
<thead>
<tr>
<th>City</th>
<th>Population (2013-11-19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pikine</td>
<td>1,101,859</td>
</tr>
<tr>
<td>Dakar</td>
<td>1,081,222</td>
</tr>
<tr>
<td>Mbacké</td>
<td>879,506</td>
</tr>
<tr>
<td>M'Bour</td>
<td>641,068</td>
</tr>
<tr>
<td>Thies</td>
<td>636,088</td>
</tr>
<tr>
<td>Kaolack</td>
<td>466,421</td>
</tr>
<tr>
<td>Rufisque</td>
<td>462,741</td>
</tr>
<tr>
<td>Tivaouane</td>
<td>431,956</td>
</tr>
<tr>
<td>Podor</td>
<td>356,408</td>
</tr>
<tr>
<td>Louga</td>
<td>354,989</td>
</tr>
</tbody>
</table>

As illustrated in Figure 2, we mapped the 1666 antennas coordinates on a standard latitude-longitude projection, represented by the black dots on the map. It can be noticed that the antennas are spatially very unevenly distributed with a dense distribution around Dakar region, where the capital is located. However towards the center of the country the number of antennas is reduced. It can be seen that the antennas distribution correlates with the population density within the country. Thus in the most populous regions, there is an increased number of base station.

4.2 Population Density vs. Telecommunication Traffic
The first dataset provided by Orange contains information about the traffic exchanged between every antenna pair. It provides information about the number of voice calls, voice calls duration, and number of SMS exchanged within one hour. In order to study if there is any correlation between the population size and the amount of traffic exchanged within the communication network, we aggregated the antenna level communication at a department level, and differentiated between incoming and outgoing traffic. Figure 4, 5 and 6 illustrate the population per department vs. the telecommunications traffic exchanged, in and out for each department along with the corresponding smoothed fit curve with confidence region. The information is aggregated over the full year of data.

---

\(^2\) http://www.geohive.com/cntry/senegal.aspx
It can be noticed that along with the increase in the population size, the intensity of the traffic in that location is also increasing.

### 4.3 Inter-City Telecommunication Flows

After analyzing how the telecommunication flowing into and out of the 45 departments from Senegal, scale with the population size, we are interested in looking into the inter-antennas/inter-arrondissement telecommunication flows. Figure 7 illustrates the total traffic intensity for each antenna, in terms of the total number of in and out voice calls, with their size and color representation reported to the load intensity. Thus, heavily loaded antennas are represented by wider points and higher intensity color. It can be seen that the traffic load is correlated with the population size represented in Figure 3 and is mainly located in the main city areas as identified in Figure 2. In this context, an adaptive power allocation for the antennas could be employed so that, in the conditions where some antennas presents high traffic load and dense user activity, their transmission power could be increased to improve the resource allocation capacity and the calls Quality of Service (QoS). Whereas for the antennas where the demand is low and there is light traffic load the transmission power could be reduced. In this way the network operator could also save on energy consumption.

When analyzing the inter-arrondissement telecommunication flows we compute the traffic matrix by associating the antennas to their corresponding arrondissement and we compute the total number of calls, total duration of calls, and total SMS exchanged between any two pairs of arrondissements. The logarithmic representation of the matrices are listed in Figure 8, 9 10. It can be noticed that for all three matrices the highest values are mostly concentrated along the diagonal and the most intensity in the origin. If we look at the previous analysis about the population and the main cities presented within the Arrondissements, as illustrated in Figure 2, we can notice that the Arrondissement IDs near the origin are the ones with the region of the capital and where most of the antennas are deployed.
Thus, from the Figures above we can notice that most of the communication is done with the Arrondissements near the capital, and within the same Arrondissements. Communication between arrondissements that are both far from the capital is highly uncommon. We can also notice that people are more likely to call than send an SMS, if we compare the voice matrices with the SMS traffic exchanged illustrated in Figure 10.

Looking into the inter-antenna traffic exchange for a more fine-grained view, Figure 11, 12, and 13 present the voice calls exchanged, the voice calls duration and SMS exchanged matrices. It is noticeable also at a fine-grained level that the communication is mainly happening within the same antenna, as given by the high values along the diagonal. It is also noticeable that the communication near the origin is more intense. This is because the antennas near the origin are within or very close to the Dakar region. It can also be noticed in terms of voice calls duration, the communication is more intense when calling customers located near the capital. In terms of SMS, it is visible that SMS exchange between antennas far away from the capital region is highly uncommon. This shows that calls are more common than SMS, as the calls matrices are visible denser than the SMS matrix.

As null values are indicated using the white color, the white lines illustrated on the above matrices indicate that some antennas are not generating any kind of traffic over the full year. These findings are of importance for network operators when considering the deployment or development of their network for cost saving and efficient use of resources.

4.4 Distance vs. Telecommunication Traffic
Several studies have stated that another parameter that influences the communication intensity between cities is the distance. In this section we look at the impact of the distance between antenna pairs and the telecommunication traffic exchange between antennas. To this extent, we have aggregated the antenna to antenna communication over the full year. Using the information provided about the antennas location, we make use of the latitude and longitude coordinates of each antenna and we calculate the distance in kilometers, between them using the haversine formula. Approximatively, a number of 2.4 million antenna-to-antenna interactions were found and the results are plotted in Figure 14, 15, and 16 indicating the distance vs. total number of voice calls, distance vs. total duration of voice calls, and distance vs. SMS.
exchanged, respectively. On each graph, the corresponding smoothed fit curve is plotted. It can be seen that the telecommunication exchanged traffic is mainly happening between antenna pairs close located, and as the distance increases the communication intensity is decreasing.

4.5 Symmetry of the Telecommunication Traffic

In this section we analyze the symmetry of the telecommunication traffic at the arrondissement and antenna level. We look into how the communication flows into and out of the arrondissement/antenna. In Section 4.2 we have analyzed how the in and out traffic intensity scales with the population size. We have seen that the traffic is more intense in areas where the population is higher. Here we aggregated the number of voice calls, the total voice calls duration and the SMS exchanged in and out for each arrondissement and each antenna. The results are plotted in Figure 17, 18, and 19 for the arrondissements and Figure 20, 21, and 22 for the antennas. It can be visible that the incoming and outgoing communications are highly symmetric for both situations the arrondissements and the antennas. Thus the calls or SMS in one direction always find a match in the opposite direction.
Figure 23 represents the voice in and out traffic flows between the 14 regions of Senegal. The source and destination of the voice traffic flows are represented by the circle’s segments, where nearby regions are positioned close to each other. The width of the link also indicates the size of the traffic flow. It can be noticed again that most of the traffic is happening between the same region, and Dakar region occupies almost half of the traffic. Moreover, the traffic is higher between the regions that are close to each other, as previously observed.

5. IDENTIFYING EVENTS IN DATA TRAFFIC

By analyzing the traffic exchange from dataset 1, we are able to identify several anomalies in the traffic.

5.1 Dead Antennas

The dataset should contain information from 1666 antennas spread around the country. However, when analyzing the
aggregated traffic exchanged over the full year in order to create the traffic matrices introduce in Section 4.3, we have noticed that several antennas do not generate any kind of traffic, indicated by the white lines in Figures 11-13. The reason could be that the information about them is missing from the datasets or they are dead antennas, meaning that they are not working. The dead antennas identified from dataset 1 are illustrated in Figure 24. An interesting fact is that 52 antennas (red points on Figure 24) out of 1666 are not recorded in the dataset to generate any kind of in or out voice nor SMS traffic, whereas 2 antennas (blue points on Figure 24) generate a very small amount of voice traffic, 4 and 64 voice calls, but no SMS traffic.

5.2 Anomalies in the Telecommunication Traffic

In order to identify anomalies in the telecommunication traffic, we first plot the aggregated total number of voice calls by month along with the median, the quartiles, and the maximum and minimum values, as illustrated in Figure 25. It can be seen that there is a very low value in the voice calls in March, several drops in traffic are registered in June, July and September, whereas peaks in traffic are registered in August and October.

To have a more fine grained level we plot the aggregated voice calls, duration and SMS exchanged per day of the year along with the corresponding smoothed fit curve with confidence region as illustrated in Figures 26, 27 and 28, respectively. It can be seen that for all three types of traffic there is a minimum traffic exchange in 29th of March. Several peaks are recorded on 15-16th of October and 9th of August for the voice calls exchanged, and 9th of August, 16th of October, and 7th of August for the voice calls duration, whereas for SMS traffic exchange the peaks are 16, 15, 17th of August.
5.3 Low Activity on 29\textsuperscript{th} of March

As seen previously there is a very low activity recorded on 29\textsuperscript{th} of March which might be due to a gap in the Dataset 1 or it can be associated to an electric failure. Also it happens that the date is associated to Good Friday.

The traffic matrix for this particular day is listed in Figure 29. It can be noticed that the traffic is mainly local. The activity recorded is only between 12 to 1am, with a total number of voice calls of 153,119 and a total voice calls duration of 1,626,8224 minutes.

5.4 Ramadan Period

In Senegal, Islam is the predominant religion and is practiced by approximately 94\% of the country’s population, whereas Christian community is represented by almost 5\% of the population. Ramadan represents the month the Quran was revealed and is celebrated by Muslims through fasting during the daylight hours from dawn to sunset. In 2013, the Ramadan started on 9\textsuperscript{th} of July and continued for 30 days until the 7\textsuperscript{th} of August.

The marking of the end of the Ramadan period is celebrated through the Feast of Breaking the Fast, referred to as Eid al-Fitr. The celebration will start on the last day of Ramadan, on 7\textsuperscript{th} of August and continues until the next day evening, 8\textsuperscript{th} of August.

As during Ramadan, people fast during the daylight hours we want to analyze if this will impact their calling habits within the network operator. To this extent, we aggregated the number of voice calls and voice calls duration per month and we plot the aggregated values per hours of the day, including the median, the quartiles, and the maximum and minimum values, as illustrated in Figure 30 and 31, respectively. We can notice that there are several peak changes during the night period. The number of voice calls doubles starting from 11pm to 5am during the months of July and August. Thus, this means that during this month people are highly active at night and have a slow start during the morning and the day. This is reflected in the voice calls duration as well. People tend to speak more during July and August starting from 10pm until 6am, when the voice call duration doubles.
starting with 15th of October. This explains the high peak in the number of voice calls and voice calls duration over 15th to 16th of October.

6. CONNECTING PEOPLE, LOCATIONS AND REAL-WORLD EVENTS

From the spatio-temporal patterns of the collective customers’ activity within the mobile network traffic datasets introduced previously, the correlation between people, locations and events is analyzed. Specifically the interest is on studying the correlation between exceptional patterns detected in the mobile usage within a cellular network and real-world events such as public concerts, parades, religious festivals, riots protests, etc.

Understanding the exceptional data usage patterns could significantly improve the spatial and temporal awareness when taking decisions. An example would be in the case of event management, when organizing parades/carnivals/concerts, etc.

For this analysis the data provided in Dataset 1 and Dataset 2 were used.

6.1 Connecting People and Locations

Considering the data from Dataset 2, we computed the overall antennas activity in terms of how many users are connected to it, over the full monitoring period. Taking the location coordinates of these antennas, it was possible to identify their position within a certain city. The results are illustrated in Figure 32. The dots represent the antennas locations whereas their size and color representation is reported to the load intensity over the 365 days period. Thus, heavily loaded antennas are represented by larger points and higher intensity color. The results show that the highly loaded antennas are spread across the main cities of Senegal as identified in Figure 2. Comparing the results with the indicative population map of Senegal as illustrated in Figure 3, it can be noticed that data usage activity is mostly registered in densely populated areas, as expected.

These findings have significant impact and they can be correlated to the important cities of the country. These observations led to the correlation between antennas activity within a cellular network and their geographical location. Thus by analyzing the user activity and their mobility patterns within a cellular network only, it is possible to identify the major cities/locations within a country/city.

Understanding the people-location interaction could represent a potential for location-based services. For example time-independent interactions refer to overlapping trajectories between distinct people irrespective of the actual time of overlap. This information is very useful in social recommender systems which are based on location-based tagging services.

The total number of active users over each period of Dataset 2 is represented in Figure 33. It is visible that there are some particular dates when the users are highly mobile. Several such events are identified in the following sections.

6.2 Le Grand Magal de Touba

By analyzing the antennas activity and user diversity (e.g., the number of distinct users connected to an antenna over the monitoring period), it was possible to identify particular religious festivals. The data showed that there was intensive users activity and users diversity in the area of Touba, Mbacké region, towards the end of December. Figure 34, illustrates the total number of active users over the December period for three antennas located in Touba area, specifically Antenna IDs: 1019, 1024 and 1025. It can be noticed that the user activity increases more than six times by 22nd of December. Taking these observations and looking at the real-time events happening in that specific location during exactly that period, we come to know about the Magal Festival taking place on 22nd of December. Consequently, these pattern exceptions in the antenna usage are perfectly correlated with the real-world event, such as: Magal Festival.

Figure 32 Total Users Activity per Antenna Over the Full Year

Figure 33 Total Active Users per Period

Figure 34

During the Magal Festival, more than a million pilgrims, members of the Mouride brotherhood flocked to ‘Africa’s Mecca’ from all over the world in the holy town of Touba. On 22nd of December 2013 the 119th edition of the Magal festival was celebrated.

### 6.3 Tivaouane Maouloud Festival

Another exceptional event was registered in January near Tivaouane, Thies region. Figure 35 illustrates the total number of active users over the January period for five antennas located in the Thies area, specifically Antenna IDs: 599, 604, 606, 608, and 609. It can be noticed that the user activity increases more than ten times by 23rd of January. Taking these observations and looking at the real-time events happening in that specific location during exactly that period, we come to know about the Maouloud Festival\(^4\). Each year millions of visitors are celebrating the birth of the prophet Muhammad through the Maouloud festival, also known as Gamou.

High user activity during the Maouloud festival was registered in Kaoulack as well, as illustrated in Figure 36 for Antenna ID 944.

### 6.4 Casamance Conflict

Apart from the religious festivals that can be detected from the CDR Datasets, we were able to detect conflict events as well. An exceptional event in the CDR datasets 1 and 2 was detected in the area of Kafoutine during February. The total number of voice calls along with the active users during the February period for Antenna ID 622 is listed in Figure 37. It can be seen that there is an anomaly in the traffic that starts increasing from 1st of February with a peak in the communication on the 2nd of February.

The antenna is located in near the Kafoutine city in the Casamance region. On 1st of February 2013 there was reported an attack of the rebels from the Casamance Movement for Democratic Forces over the Credit Mutuel bank in Kafountine\(^5\). After the clashes between the rebels and the government soldiers, four dead including a Frenchman were reported. The effect of this attack can be noticed in the CDRs. After the news about the attack came out on 2nd of February there was a peak in the number of voice calls made and a significant increase in the voice calls duration.

### 7. CONCLUSIONS

In this work we explore the use of anonymized Call Detail Records (CDRs) containing both voice-calls and SMS activities, from a cellular network in Senegal in order to study the telecommunication traffic exchanged and connect people, locations and real-world events. The study presented in this paper, shows that CDR data can be used to detect exceptional spatio-temporal patterns of the collective human mobile data usage and that these ‘anomalies’ in the usage patterns can be correlated to real-world events (e.g., religious festivals, riots, conflicts, etc.). Understanding the exceptional data usage patterns could significantly improve the spatial and temporal awareness when taking decisions and this knowledge could be further used to develop an intelligent system that detects exceptional events in real-time from CDRs monitoring. For example, a real-time event detection system could be of crucial importance to ensure people’s safety in case of emergency situations, such as riots protests which could be more efficiently handled if detected on time, e.g., the Casamance Conflict.

---

\(^4\) Senegal Maouloud Festival [http://en.wikipedia.org/wiki/Tivaouane]

8. ACKNOWLEDGMENTS

We would like to thank the D4D Challenge committee and the Orange Group for Development and Initiative for the datasets provided in this study. We would also like to express our gratitude to Dr. Franco Raimondi for his useful expertise and insights he provided to this project.

REFERENCES


Spatial structure and efficiency of commuting in Senegalese cities

Rémi Louf\textsuperscript{1}, Giulia Carra\textsuperscript{1}, Hadrien Commenges\textsuperscript{2}, Jean-Marie Dembele\textsuperscript{3}, Riccardo Gallotti\textsuperscript{1}, Maxime Lenormand\textsuperscript{4}, Thomas Louail\textsuperscript{1}, Marc Barthelemy\textsuperscript{1,5}

\textsuperscript{1} Institut de Physique Théorique, CEA-CNRS (URA 2306), F-91191, Gif-sur-Yvette, France
\textsuperscript{2} Laboratoire Eau, Environnement et Systèmes Urbains, Ecole des Ponts, F-77455, Marne-la-Vallée, France
\textsuperscript{3} Université Gaston Berger, UFR SAT, Saint-Louis, BP 234, Senegal
\textsuperscript{4} IFISC, Instituto de Física Interdisciplinar y Sistemas Complejos (CSIC-UIB), Campus Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain and
\textsuperscript{5} Centre d’Analyse et de Mathématique Sociales, EHESS-CNRS (UMR 8557), 190-198 avenue de France, FR-75013 Paris, France

Senegal is experiencing an unprecedented urban revolution: according to the latest UN projections, its urban population will be multiplied by 3 in the forthcoming decades, to reach 18 million people by 2050. While cities are often lauded as the solution to mankind’s socio-economical and environmental issues, when badly managed, they can also be recipes for disasters. In order to propose well-informed transport and planning policies in Senegal, it is crucial to first measure and understand the key spatial dynamics that shape its cities. In this report, we uncover commuting patterns in the 12 largest urban areas of Senegal and their coarse-grained spatial properties using mobile phone data, allowing us to characterize the efficiency of commuting and to compare cities with each other. At the inter-urban level, we show that for most cities, the vast majority of commuters live and work in the same urban area, meaning that Senegalese cities are well-integrated employment markets. We then compute the geographical area of influence of each city. This confirms the importance of large cities such as Dakar, but also highlights smaller cities which play an important economical role such as Tambacounda or Touba. At the intra-urban level, we quantity the spatial mismatch between residence and workplace locations and we propose a measure of the ‘optimality’ of the commuting structure. We find that Dakar, with a high optimality index, has a coherent spatial structure with nested residential and employment areas which is reflected in the fact that 80% of the residential and activity hostspots overlap. Smaller cities however—such as Louga, Kolda, Mbour—are far from the optimal commuting solution. The methods proposed in this study could help urban planners in identifying locations and areas which are the most penalized by inefficient commuting, a source of economic loss and stress on people’s life and the environment.

Introduction

Urban transition in Senegal

Senegal, along with a large part of African countries, experiences a rapid economic growth with an average of 3% GDP growth per year \cite{1}. This economic growth goes hand-in-hand with a dramatic increase of its urban population: while Senegal is estimated to currently host 6 millions people in its cities, the UN estimates that this number should be multiplied by 3 by 2050. It is therefore an understatement to say that Senegalese urban areas are changing fast. The current transition carries a lot of promises in terms of development; yet, badly managed, it could as well lead to a socio-economic and environmental disaster. In order to make useful investments in transportation infrastructures and propose well-informed planning policies, policy-maker need to understand which areas are going to develop, at what rate, and quantify the ties between different neighbourhoods (at the intra-urban level) and different urban areas (at the inter-urban level).

In general, rapid urbanization goes along with unplanned settlements, health and sanitary problems, as well as congestion problems. In particular, a key ingredient of sustainable cities is an efficient organization of commuting. Non-pedestrian commuters in Senegal travel almost exclusively by cars, and the longer the commuting, the larger the pollution and CO\textsubscript{2} emissions are. Dakar urban and industrial areas expanded in direction of Thies, and zones of informal settlements appeared. Poor people are pushed toward the periphery which is usually under-equipped and with much lower level of transport infrastructure. It is therefore important to identify areas and trips with large volumes and strong demand for a better, targeted planning. Such information is however usually difficult to obtain and it has been almost a decade now that scientists have realized that geolocated traces passively generated by individuals’ ICT devices could revolutionize the quantitative and theoretical understanding of human spatial dynamics, and urban dynamics \cite{5}. To give a few examples of urban phenomena whose understanding has been enhanced in the recent years through renewed quantitative approaches applied to new sources of ICT data, we can mention the statistical and spatial properties of individuals’ mobility in cities \cite{6–10}; the universal structure of subway networks and streets networks \cite{11, 12}; the number and the spatial organisation of centers in urban areas \cite{13, 14};
the spatial properties of social networks in countries and cities [15, 16]; and the scaling of diverse quantities with the population size of the city [17–19].

For a few years now, we can investigate these urban questions in much more detail than during previous decades, by analyzing vast amounts of data available at different spatial and temporal scales. Problems that in the past were addressed through surveys by geographers and transport scientists are nowadays addressed by interdisciplinary teams, with many different data sources, data that are more precise both spatially and temporally. Sixty years after its emergence as an academic research field, spatial analysis and quantitative geography may be living their second quantitative and theoretical revolution.

In this paper, we will present an example of such studies, and by using mobile phone data recorded in Senegal, we will show by proposing new measures how we can extract useful information about the spatial structure of urban areas and commuting.

Mobile phone data and urban mobility

Limitations of classic data sources Traditionally, urban planning has relied on travel surveys and censuses. However these data sources have several limitations:

- They require an important logistic, time, and are expensive.
- For these reasons, they are performed over long time intervals, typically every decade or so. With such data it is then impossible to estimate changes of the urban structure over short time scales.
- Transport surveys are often based on samples of a view thousands of individuals only.

In contrast, individual mobile phone data provide anonymized location information about a large fraction of the population with a temporal resolution below 24h, and with a spatial resolution which depends on the density of antennas. For most cities this resolution is of the order of a few hundred meters in the centers, and 1 to 2 kms in the peripheral neighbourhoods. In Senegal the mobile phone penetration rate was estimated to be about 85% in mid-2013 and projections estimated a 110% figure at the end of 2014 [40]. It thus allows to monitor population displacements, such as the daily journey to work, important events implying many travels such as Touba’s Magal, or long term residential migrations (mostly rural to urban migrations).

Interestingly, a recent study in Madrid and Barcelona – the two largest Spanish urban areas – demonstrated that new sources of mobility data (mobile phone data in the first place, but also geolocated tweets) provide at the city scale a very comparable picture of the commuting structure, when compared to the information obtained with transport surveys [4]. This result opens the door to a more systematic use of new sources of ICT data to work on mobility issues in cities.

Previous studies of urban mobility in developing countries with ICT data. For obvious structural and economical reasons, most of the ICT-related studies of the last decade have focused on cities located in rich and developed countries. Few results have focused on cities of other regions of the world and other continents, notably Eastern Europe, South-America and Africa. Several important papers have claimed to uncover universal mobility and urban patterns and propose general models, but are so far limited to a small number of geographical areas [6, 18, 24]. Continents and countries have however different urbanization histories, exhibit different spatial properties such as densities and spatial organization [21, 22], and the universality of urban patterns is not yet proved to be correct for all regions of the world. To this day, Africa is the less urbanized continent, and is currently experiencing a very fast urban transition [23]. Most large African cities, including Dakar, have no subway network, bike-sharing or car-sharing systems, and their highways systems are much less developed than in US, European or Asian cities of the same size. Apart from the informal collective transport (‘bus rapides’) and the numerous taxis, public collective transport in Dakar is provided by the municipality bus service (‘Dakar Dem Dikk’). It is currently an important issue to obtain a better spatial knowledge of trips in order to develop this service. For these reasons, measuring and comparing the spatial properties and the structure of intra-urban mobility in Senegal is particularly important both for Senegal and Dakar urban planning questions, and also on the scientific side for the elaboration of the emerging science of cities. Such a quantitative knowledge could help to guide planning policies of rapidly urbanizing areas.

The previous edition of Orange’s D4D challenge provided communication datasets in Ivory Coast, and allowed for quantitative studies of mobility patterns in Abidjan and Ivory Coast. Numerous interesting studies related to transport and mobility were performed. For example Kung et al. [25] used the data to test the long-lasting hypothesis of a universal, fixed time-budget for daily mobility (often referred as ‘Zahavi’s law’). They obtained surprisingly large commuting times, and their results questioned the possibility of using individuals mobile phone data to infer travel duration [25]. Berlingerio et al. [26] developed an interactive and modular application “to optimize the public transport network, with the goal to improve ridership and user satisfaction”. They used individuals’ travel and activity patterns detected in the data to extract origin-destination (OD) matrices and individuals’ travel preferences, to determine optimal design of potential new transit services. A number of studies proposed methods to automatically extract OD matrices from individual data, notably [27]. Wakita et al. analysed the temporal patterns of communication data and could infer the dominant type of land-use of the geographical area covered by each antenna, and proposed
maps of land-use at various scales in the country [29]. Andris and Bettencourt [28] applied network analysis on the communication network at three different scales (individuals, cities, and the whole urban system), drew the communication networks upon natural resource layers and discussed the resulting maps. While providing advice about possible future developments, they didn’t characterize the attractiveness of cities, nor their spatial organization and its interplay with mobility patterns of individuals. In addition, none of these 2013 projects proposed a comparison of cities based on the spatial properties of commuting, especially from an efficiency perspective. Such a comparison could help urban planners in identifying cities which are the most penalized by inefficient commuting, a source of economic loss and stress on people’s life and the environment.

Objectives of the study and organization of the report.

The objective of our project is to extract from mobile phone data the intra-urban and inter-urban commuting patterns of individuals, discuss these patterns, and to provide a set of quantitative characterizations of the organisation of journey-to-work mobility in Senegalese cities. Another important motivation is to develop coarse-grained indicators summarizing these large amounts of data, able to provide synthetic and large scale pictures of the structure of individual mobility in cities. Such meso-scale information is also particularly useful for validating synthetic results of urban mobility models (such as [30] for example), for comparing different cities and also for comparing different models. An accurate modeling of mobility is indeed crucial in a large number of applications, including the important case of epidemic spreading which needs to be better understood, especially at the intra-urban level [31, 32]. The recent case of the Ebola virus, that didn’t spread in Senegal but probably had serious impact on international migrations and touristic activities in the country, provides a contemporary illustration of the societal usefulness of such an understanding.

In this report, we focus on commuting (journey-to-work trips) in Senegalese cities, which represents everywhere the largest part of the daily mobility. We have extracted the 12 largest cities from the map of Orange’s antennas in Senegal (see Table I), and we have computed the origin-destination (OD) matrices from the twenty-five 2-weeks individual activity datasets (see Methods). We focused successively on two different spatial scales. We first studied the inter-urban case and characterized the resulting maps of land-use at various scales in the country [29]. Andris and Bettencourt [28] applied network analysis on the communication network at three different scales (individuals, cities, and the whole urban system), drew the communication networks upon natural resource layers and discussed the resulting maps. While providing advice about possible future developments, they didn’t characterize the attractiveness of cities, nor their spatial organization and its interplay with mobility patterns of individuals. In addition, none of these 2013 projects proposed a comparison of cities based on the spatial properties of commuting, especially from an efficiency perspective. Such a comparison could help urban planners in identifying cities which are the most penalized by inefficient commuting, a source of economic loss and stress on people’s life and the environment.

\[ C = N^o + N^{-} \] (1)

These three types of commuting are represented schematically on Fig. 1.
FIG. 1: Schematic representation of the different types of commuters: internal, convergent and divergent commuters.

The values measured for the 12 Senegalese cities we identified are given in Table I, ranked in decreasing order of commuting population size.

In order to estimate the relative importance of the different flows, we define the two following ratios

\[ f_0 = \frac{N^\rightarrow + N^\leftarrow}{N^\circ} \]
\[ f_1 = \frac{N^\leftarrow}{N^\rightarrow} \]

whose values are also given in the two last columns of Table I. The quantity \( f_0 \) characterizes the importance of internal commuters, and \( f_1 \) simply compares the number of convergent and divergent commuters.

For most cities studied here, convergent and divergent commuters represent roughly 10% of the internal commute, signifying that Senegalese cities are well-integrated employment markets. The smallest fraction of out-commuting is found in Dakar (3%) and the largest in Tivaoune (31%), which is an important religious center in Senegal, thereby explaining this large value.

The ratio \( f_1 \) of the number of divergent and convergent commuters allows us to divide Senegalese cities in three classes:

- Cities for which \( f_1 > 1 \) - Diourbel, Louga, Touba, Tambacounda - where the number of ingoing commuters is larger than the number of outgoing commuters. This implies that there are more people present in the city at daytime when compared to nighttime.

- Cities with \( f_1 \approx 1 \) - Dakar, Kaolack, Kolda - where the number of outgoing and ingoing commuters is roughly the same, meaning that the total population present in the city is basically the same during day- or nighttime.

- Cities with \( f_1 < 1 \) - Tivaoune, Thies, Ziguinchor, Saint-Louis - where the number of outgoing commuters is larger than the number of ingoing commuters, also meaning that there are more individuals present in the city during the night than during the day.

This simple indicator \( f_1 \) allows to illustrate the balance between jobs and residences. An ‘attractive’ city (\( f_1 > 1 \)) for example means that the job/activity market is larger than the residential offer. In contrast, a city with \( f_1 < 1 \) is in majority residential. We note however that, surprisingly, there are very few differences between cities, and that the ratios are almost all of order 1, suggesting an equilibrium between the number of people that live outside a city and spend the day in it, and the number of people who live in a city and spend the day outside.

We can further characterize the convergent commuters by measuring the average distance they are traveling. We call this distance the attraction radius \( r \) of the city, computed as

\[ r^\rightarrow = \frac{1}{N^\rightarrow} \sum_n \ell_n \]  \( (2) \)

where \( \ell_n \) is the distance traveled by the \( n \) convergent commuters. This quantity characterizes the regional influence of cities in terms of commuting and job/activity market. The radius for Senegalese cities are shown on Fig. 2, and their values listed in the Table II.

Dakar has a large zone of influence – as expected – however there are some surprises such as Tambacounda which is a small city but displays a large attraction radius. It is interesting to note that this large value probably reflects the fact that Tambacounda is an important commercial stop on the road to eastern countries such as Mali and to Casamance, and for stock trading.

We believe that this first study of convergent and divergent commuting should be interesting for planning purposes, but we note that the available data here are too sparse to reach very reliable conclusions. Indeed, when extracting journey-to-work OD matrices (see Methods), we realized that on a 2-weeks period, there is an important proportion of individuals whose mobility is not reg-

<table>
<thead>
<tr>
<th></th>
<th>( C )</th>
<th>( N^\circ )</th>
<th>( N^\leftarrow )</th>
<th>( N^\rightarrow )</th>
<th>( f_0 )</th>
<th>( f_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakar</td>
<td>3,099,116</td>
<td>4,049,579</td>
<td>47,540</td>
<td>49,537</td>
<td>0.03</td>
<td>0.95</td>
</tr>
<tr>
<td>Thies</td>
<td>261,050</td>
<td>248,950</td>
<td>10,718</td>
<td>12,100</td>
<td>0.09</td>
<td>0.88</td>
</tr>
<tr>
<td>Touba</td>
<td>230,498</td>
<td>218,987</td>
<td>13,980</td>
<td>11,511</td>
<td>0.11</td>
<td>1.21</td>
</tr>
<tr>
<td>Mbour</td>
<td>216,418</td>
<td>204,312</td>
<td>11,111</td>
<td>12,106</td>
<td>0.11</td>
<td>0.91</td>
</tr>
<tr>
<td>Kaolack</td>
<td>178,365</td>
<td>170,821</td>
<td>8,097</td>
<td>7,544</td>
<td>0.09</td>
<td>1.07</td>
</tr>
<tr>
<td>Saint-Louis</td>
<td>162,875</td>
<td>154,910</td>
<td>5,843</td>
<td>7,965</td>
<td>0.08</td>
<td>0.73</td>
</tr>
<tr>
<td>Ziguinchor</td>
<td>135,562</td>
<td>130,264</td>
<td>4,255</td>
<td>5,298</td>
<td>0.07</td>
<td>0.81</td>
</tr>
<tr>
<td>Tambacounda</td>
<td>65,577</td>
<td>62,900</td>
<td>3,049</td>
<td>2,677</td>
<td>0.09</td>
<td>1.13</td>
</tr>
<tr>
<td>Tivaoune</td>
<td>63,198</td>
<td>53,815</td>
<td>7,532</td>
<td>9,383</td>
<td>0.31</td>
<td>0.80</td>
</tr>
<tr>
<td>Diourbel</td>
<td>62,345</td>
<td>59,357</td>
<td>3,585</td>
<td>2,988</td>
<td>0.11</td>
<td>1.19</td>
</tr>
<tr>
<td>Louga</td>
<td>59,883</td>
<td>56,964</td>
<td>3,649</td>
<td>2,919</td>
<td>0.11</td>
<td>1.25</td>
</tr>
<tr>
<td>Kolda</td>
<td>57,895</td>
<td>55,676</td>
<td>2,304</td>
<td>2,219</td>
<td>0.08</td>
<td>1.03</td>
</tr>
</tbody>
</table>
FIG. 2: Attraction radius for the major Senegalese cities. We represent the radius of attraction computed using the OD matrices that we extracted from the data. The radius for different cities are shown on three different panels for a matter of clarity.

<table>
<thead>
<tr>
<th>City</th>
<th>$r^\text{attr}$ (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakar</td>
<td>137</td>
</tr>
<tr>
<td>Thies</td>
<td>60</td>
</tr>
<tr>
<td>Touba</td>
<td>109</td>
</tr>
<tr>
<td>Mbour</td>
<td>53</td>
</tr>
<tr>
<td>Kaolack</td>
<td>77</td>
</tr>
<tr>
<td>Saint-Louis</td>
<td>91</td>
</tr>
<tr>
<td>Ziguinchor</td>
<td>91</td>
</tr>
<tr>
<td>Tambacounda</td>
<td>141</td>
</tr>
<tr>
<td>Tivaouane</td>
<td>72</td>
</tr>
<tr>
<td>Diourbel</td>
<td>69</td>
</tr>
<tr>
<td>Louga</td>
<td>69</td>
</tr>
<tr>
<td>Kolda</td>
<td>98</td>
</tr>
</tbody>
</table>

TABLE II: Attraction radius $r^\text{attr}$. Cities are ordered by total number of commuters, showing that there is no clear correlation between the attraction radius and the total number of commuters.

A first interesting look on a city’s structure is provided by the locations of hotspots, which are local maxima of the density of individuals. Using the extracted OD matrices, we can identify the most important residential and daily activity areas. In order to determine hotspots we used the ‘LouBar’ method based on the Lorenz curve of the distribution of densities [14]. For most cities the hotspot spatial structure is simple and monocentric. Only for the city of Dakar we have an interesting organization shown on Fig. 3, where we see a polycentric structure as it is observed for other large cities in the world [13].

We show in this figure 3, the map of residential and activity hotspots in Dakar, determined with the method described in [14].

Intra-urban organisation of Senegalese cities

Spatial structure of hotspots A first interesting look on a city’s structure is provided by the locations of hotspots, which are local maxima of the density of individuals. Using the extracted OD matrices, we can identify the most important residential and daily activity areas. In order to determine hotspots we used the ‘LouBar’ method based on the Lorenz curve of the distribution of densities [14]. For most cities the hotspot spatial structure is simple and monocentric. Only for the city of Dakar we have an interesting organization shown on Fig. 3, where we see a polycentric structure as it is observed for other large cities in the world [13].

We show in this figure 3, the map of residential hotspots (i.e. with large night activity) and job/activity hotspots (i.e. daily activity). The overlap between these...
maps is large which means that important residential areas in Dakar are also the important daily activity centers. Our spatial delimitation of the urban area of Dakar (see Methods) includes major surrounding municipalities (Pikine, Rufisque), and some parts of these municipalities are indeed the most important crowded parts of the urban area. The hotspots of Dakar are not all along the coast but are located in internal crowded neighbourhoods (Liberté, Derklé, Sacré-Coeur, Medina, etc.).

The ‘work’ hotspots identified in the urban area of Dakar (26% of all the antennas in the city) contain 37% of the total daily population (as identified by our OD matrices). The ‘home’ hotspots (24% of all the antennas in the city) on the other hand contain 39% of the total population identified by our OD matrices (over the whole year). We note that rich residential neighbourhoods (Les Almadies, Le Plateau), which are also the areas where most of the administration and buildings of the major companies are located, are not residential hotspots. For example one can see on the work hotspots map that some areas of ‘Le Plateau’ are indeed tagged as employment/daily activity hotspots, but not as residential hotspots.

Surprisingly, home and work hotspots overlap then quite well in the city of Dakar: 80% of the work hotspots are also home hotspots. It gives us a first intuition that in Dakar, the commuting distances could be quite short in average, and that the city seems coherently organized, in the sense that many individuals don’t have to travel long distances with cars everyday to travel from their home to their workplace.

Mobility of individuals within cities: hotspots and organization level Once we have extracted the commuting network of individuals we can investigate the spatial properties of this network. We first measure the distance daily traveled by individuals to go from their home to their main activity place. These distances for Dakar and Thies are mapped on Figure 4.

These two figures reveals two typical forms of organization of cities. In the case of Dakar, we observe a polycentric structure with individuals traveling longer distances as they live further from the main activity hotspots (see figure 3), but also with secondary activity centers appearing in the suburbs (Rufisque, Pikine), where many people live and work, resulting in shorter commuting distance on average. Still in Dakar, while several secondary activity centers have developed as the city expanded to Rufisque and Pikine, the historical center remains the most influential and attracts more commuters from these secondary centers than the opposite (see Figure 5). In contrast to this polycentric structure, we observe for the smaller city of Thies a clear monocentric structure with a unique central zone that attracts most daily activity. Consequently, individuals that live in the center commute over short distances, while the further people live from the center, the longer their daily commuting distance is.

An intriguing question is if we can characterize the level of organization of cities from the spatial structure of commuting flows. The first metric that we have is the total commuting distance $L$. The value of $L$ however does not have a clear interpretation by itself, but we can compare it to reference values. We will then compare $L$ with the two extreme situations:

- A totally disorganized mobility structure, where commuting patterns occur at random, while conserving the static structure of the city given by the numbers of inhabitants and employees attached to each antenna. In this case, we obtain a total commuting distance denoted by $L_R$.

- An optimally organized mobility structure, for which mobility patterns are such that the total commuting distance in the city is minimum (while conserving the number of inhabitants and employees attached to each antenna). Given the constraints of numbers of inhabitants and employees, this leads to a minimum total commuting distance denoted by $L_O$.

In order to compute the values of $L_R$ and $L_O$ for each city, we calculate the corresponding OD matrices of the two extreme cases, optimal and random. The OD matrix for the disorganized state is calculated by adding flows at random, making sure to respect the in- and out-degree of each node in the network (see Methods). The distance $L_R$ is averaged over 100 random realisations. The OD matrix for the optimal state is obtained by simulated annealing, a local search optimisation method [35]. Once we have calculated the two quantities $L_R$ and $L_O$ we can then define an organization index of the city

$$O = \frac{L - L_R}{L_O - L_R}$$

This index $O$ is equal to 0 when the mobility patterns are completely disorganised ($L = L_R$) and equal to 1 when the mobility patterns are completely organised ($L = L_O$). This measure is indeed a measure of the spatial mismatch in the city: the larger the value, the more organized the city is, and individuals in this case live very close to their activity location. We give in Table III the values of $L/L_O$, $L/L_R$ and $O$ for the twelve Senegalese cities identified by our urban areas detection method.

We observe in this table that for most cities the ratio $L/L_R$ is small and approximately constant (on average equal to 0.25), the ratio $L/L_O$ displays larger variation: on average we obtain 3.8 and extremes such as 8.98 (Tivaoune) and 1.92 (Kaolack). Although we would need the corresponding values of cities in other parts of the world, these numbers suggest that there is probably some room for improving the commuting figures in many cities in Senegal. In particular, in the case of Dakar, we obtain the values indicated in Table IV (as a reference we also give the typical distance of the city, taken as the square root of its surface $\sqrt{A}$).

It is interesting to note that in the largest city of the country, the individuals journey-to-work mobility is
FIG. 4: **Average commuting distance at the antenna level** (Left) Dakar: Average outreach per antenna, i.e. average distance commuted by people living in the area covered by this antenna. When calculating the average we do not take into account the people who are flagged as living and working at the same place. (Right) Same measure for the urban area of Thies. These results are obtained by averaging the OD matrices over the whole year.

FIG. 5: **Main directed flows of commuters at the municipality scale, in the urban area of Dakar.** As the urban area expanded in direction of the surrounding municipalities, several secondary activity and employment hotspots appeared in the urban area of Dakar. Still the historical activity centers, located in the municipality of Dakar, attract more commuters that live in the surrounding municipalities, than the opposite.

rather short, and closer to the optimal situation than to the random, disorganized situation. This result suggests that there is indeed a good match between locations where people live and the ones where they perform their daily activity. It would be interesting to calculate these values for cities in European and US countries for which mobile phone datasets have been used in many papers in the recent years (France, Spain, Portugal, etc.).

**Discussion**

We extracted the OD matrices for the 12 largest Senegalese urban areas from mobile phone data and proposed several measures that can help in characterizing the spatial structure of commuting and its efficiency, and to compare different cities with each other.

At the interurban level, we could show that Senegalese cities display a good integration of labor and housing markets. In addition, the attraction radius of cities allowed us to identify important large cities and also important economical nodes. Cities with the largest attraction radius – such as Dakar, Tambacounda, Touba, Saint-Louis and Ziguinchor – would then naturally benefit from transport infrastructure improvements at the country scale, linking cities.

At the intra-urban level, we characterized the efficiency...
of the spatial organization of residential and working areas in terms of commuting. In particular, we showed that for Dakar, commuting distances can be very short and that the city seems to be coherently organized in this respect. However for other cities such as Louga, Kolda, Mbour, it seems that we are far from the optimal commuting solution. One could have naively expected that the larger the city, the more ‘anarchic’ it becomes, but our results prove that this naive representation is wrong. These preliminary conclusions would benefit from further investigations, in order to understand the origin of the spatial mismatch for these cities.

Our results show that mobile phone data can effectively be used to characterize how well an urban area is organized. In this respect, they can help in identifying the more fragile urban areas that deserve a particular attention for future intra-urban transport planning.

**Material and Methods**

**Delimitation and selection of cities.**

When one wants to compare cities, an important issue is to rely on a common, reasonable spatial definition/delimitation applied to all cities [20, 33]. For example, Dakar as a geographical object cannot be restricted to the municipality of Dakar, both in terms of morphology and function. The spatial layers provided in the project include larger spatial delimitations corresponding to administrative entities. Since we didn’t find any documentation explaining the territorial criterion chosen - if any - to construct these spatial objects, we cannot assume that such spatial delimitations are suitable to properly define Senegalese cities.

For this reason, we conceived a method to delimitate cities using the spatial points pattern of Orange’s mobile phone antennas. We proposed a simple density-based clustering method based on the hypothesis that the density of antennas reflects the density of population. For each antenna we count the number of its neighbors within a growing distance threshold ranging from 0 up to 20 km, with a fixed step of 0.5 km. We then classify those quasi-linear distributions (see Figure 6) and distinguish between Dakar’s antennas with a steep slope, Touba’s and other cities with a less steep slope, and non-urban antennas with a slight slope. Our delimitation includes urban cores but also peripheric neighborhood (see Figure 6).

We compare the criterion used to satellite pictures of the corresponding cities [41] and to the the OpenStreetMap boundaries of each city, and find that the definition captures well the built areas of each city (see Figure 7).

**FIG. 6:** Illustration of the method used to delimitate cities. For each antenna of the dataset, we count its number of neighbors in a circle of increasing radius (x-axis of the top figure). Each curve represent an antenna. We then apply a hierarchical clustering method on the resulting set of vectors, and represent the average profile of each class on the bottom figure.

**FIG. 7:** Comparison of the antennas chosen to delimitate four of the twelve cities selected, following our criterion based on the density of antennas (red), and the corresponding OpenStreetMap layers. We checked each city individually (others not shown here).
Extraction of the Origin-Destination matrices

In the following, we call commuters the people identified by our algorithm, and commutes the trips as they appear in the resulting Origin-Destination matrix.

The reference files to calculate the OD matrices are SET2_PXX.CSV of Dataset 2, where XX varies from 01 to 25. The output of the method is a \( m \times n \) matrix where \( C_{ij} \) is the number of commuters that live in place \( i \) and whose main daily activity is located in \( j \). In the following we call 'Home' the residence cell of the user and 'Work' the cell of its main daily activity place.

For each user, the extraction procedure is the following: for each hour of the two weeks period - weekends excepted - during which the user used her phone at least once, we identify the most visited cell/antenna during this hour. This cell/antenna is the one from which the user has given/received the most calls/sms during this particular hour. Hours are partitionned in two groups: (1) the daily hours that are spent at work/school for most people during weekdays (hours between \( \min_W \) and \( \max_W \)); (2) the late evening, night and early morning hours, spent at home for most people (hours between \( \min_H \) and \( \max_H \)). For both groups of hours, we identify the cell to which the user has been 'attached' the greatest number of hours. We then calculate the proportion of time spent in the cell (number of hours / total number of hours during which the user called). Finally, if in both cases these proportions are greater than a parameter \( \text{prop} \), then the two cells are tagged as the user's work/home cell and the user's home cell. Otherwise the user is not selected because her locations don't show enough regularity to assume than the two most frequent antennas are resp. her workplace and home.

Once we have applied the extraction procedure to all users we end up with an OD matrix of commuting flows for the whole country, for each two weeks period of the dataset 2.

Sensitivity analysis We analyzed the influence of the value of \( \text{prop} \) on the number of users selected, and also on the proportions of intra-cell flows (i.e. the proportion of individuals who have the same Home and Work cell) for the file SET2_P01.CSV, by using the following parameter values:

- \( \min_W = 8 \)
- \( \max_W = 17 \)
- \( \min_H = 19 \)
- \( \max_H = 7 \)

In the figure 8 we can see that the number of users decreases when \( \text{prop} \) increases, which is an expected effect. However the proportion of intra-cell flows tends to sharply increase when \( \text{prop} \) increases. In order not to remove users and try to keep the network structure we chose to fix \( \text{prop} \) to 1/3.

Computation of the mobility networks to random and optimum networks

Random Matrices and optimal matrices

In the following, we detail the methods used to obtain the random origin-destination matrices—corresponding to a situation where people would choose their working location at random from the existing possibilities—and the optimal origin-destination matrices—corresponding to a situation where people would be assigned to a working location so that the total distance commuted at the city level is optimal. What we are interested in here is the optimality of the commuting patterns given the existing spatial distribution of homes and jobs. We therefore constrain the random and optimal OD matrices in such a way that the number of homes and jobs at each antenna are identical to those observed in data. If we interpret the OD matrix as representing a network between homes and jobs in different locations, then both our random and optimal null model preserve the in- and out- degree sequences.

To generate a random graph that conserves the in- and out-degree of each node of the reference graph, we use the Molloy-Reed algorithm [34] which complexity is in \( O(n) \), where \( n \) is the sum of the weights of the edges (i.e. the number of individuals in the OD case).

In order to generate the optimal network, we use simulated annealing, a probabilistic method for global optimization problems [35]. At each step, we invert 2 OD pairs \((a,b)\) and \((c,d)\) to \((a,d)\) and \((c,b)\) (so that the conservation of in- and out-degrees is guaranteed). The move is accepted with probability 1 if the proposed solution is better than the previous one, i.e. if the total commuting distance is smaller than that of the previous situation. If the new situation is worse, however, the move can still be accepted with a probability that depends on temperature \( T \) as
and the temperature is decreased as the search progresses. This trick allows to avoid getting stuck in local minima.

Using calls data to estimate population movements

Framework

We define the activity $\tilde{A}_i$ of an antenna $i$ as the total number of calls and text messages sent and received from $i$ over a time window $\tau$. We call $\tau$ the resolution – the minimal resolution is 1 hour, imposed by the dataset $(D1)$. In other words, we are theoretically able to follow the change of activity in the city at the one-hour level.

$$\tilde{A}_\tau(i) = \int_\tau A(i, x) \, dx$$ (5)

where $A(i, x)$ is the measure of activity (in-call, out-calls, number of users) for the antenna $i$ at the hour interval $x$. Then we define the mass of the antenna as

$$M_\tau(i) = \frac{\tilde{A}_\tau(i)}{\sum_{i \in \mathcal{N}} A_\tau(j)}$$ (6)

Depending on the area of study $\mathcal{N}$ and the time-window $\tau$ we might be able to catch different phenomena. For instance, setting $\mathcal{N}$ to be a city, $\tau$ to be of the order of an hour, one can identify the patterns of daily activity in cities. Setting $\mathcal{N}$ to be an entire country and $\tau$ to be of the order of a week, a month... one can possibly identify internal migrations.

Difference between day and night activity

We first start with comparison between day and night activities in the city. Census traditionally give information on the residential population in cities, sometimes also on the working population. Mobile phone data allow us to get information with better time resolution and to follow the locations of people within the city during the day. Such information is important to have, for instance for emergency evacuation plans, for understanding epidemic spreading over short time-scales, etc. For each antenna $i$ we plot the quantity

$$DN(i) = \frac{M_{\text{day}}(i) - M_{\text{night}}(i)}{M_{\text{night}}(i)}$$ (7)

which represents the relative difference of activity of the antenna between night and day. With other data to calibrate the relation between mass of antennas and populations, one should be able to estimate the relative differences in population from these differences in antenna mass. Even without calibration, the differences in the mass of antennas give a good idea of the changes in the spatial location of populations over time.

Authors contributions

TL, RL, GC, RG, HC, ML, J-MB and MB designed the study; RL, TL and MB coordinated the study. RL, ML, HC, GC, TL and RG processed and analysed the data. RL, GC, HC and ML made the figures. TL, RL, and MB wrote the paper. All authors read, commented and validated the final version of the manuscript.

Acknowledgements

MB is supported by the European Commission though the FET-Proactive project PLEXMATH (Grant No. 317614), and the project EUOIA (FP7-DG.Connect-318367). TL is supported by the EU commission through project EUOIA (FP7-DG.Connect-318367). ML acknowledges funding from the PD/004/2013 project, from the Conselleria de Educació, Cultura y Universidades of the Government of the Balearic Islands and from the European Social Fund through the Balearic Islands ESF operational program for 2013-2017. RG is supported by the European Commission FET-Proactive project PLEXMATH (Grant No. 317614)

FIG. 9: Relative difference of antenna mass between night and day in Dakar. Each Voronoi cell is colored with respect to the value \( \frac{M_{\text{day}}(i) - M_{\text{night}}(i)}{M_{\text{day}}(i)} \) calculated for the corresponding antenna \( i \).


[34] Molloy, M. and Reed, B. A critical point for random graphs with a given degree sequence. Random structures and algorithms, 6, 161–180 (1995).


[40] NB: a penetration rate value superior to 100 means that there are more active phone numbers than individuals living in the country, which happens if for example some people have two numbers, plus professional lines, etc.

[41] not shown for copyright reasons
Cookbook for a socio-demographic basket

Constructing key performance indicators with digital breadcrumbs

Fabian Bruckschen
Humboldt Universität Berlin
fabian.bruckschen@cms.hu-berlin.de

Timo Schmid
Freie Universität Berlin
timo.schmid@fu-berlin.de

Till Zbiranski
Humboldt Universität Berlin
till.zbiranski@cms.hu-berlin.de

ABSTRACT
While developed countries are confronted with increasing public opposition against excessive data collection efforts from government agencies, developing countries often still lack basic knowledge about the people who live within their boundaries. Traditional data collection methods such as censuses or household surveys impose great financial and organizational burdens upon chronically underfunded, ill-equipped national statistical systems. The rise of new information and communication technologies offers promising sources to mitigate these shortcomings. In this paper we show that socio-demographic indicators from official statistics can be rebuilt in a uniform approach using mobile phone data from the Senegal. Therefore, geocoded survey data is used as benchmarks, the results are tested for robustness across different spatio-temporal aggregates and finally smoothed to create high-resolution thematic maps of the Senegal. The results yield three insights: First, high-resolution data can help to uncover so far hidden local heterogeneity such as pockets of poverty or illiteracy. Second, the ad-hoc availability of information can reduce the time to prepare countermeasures. Third, key performance indicators based on big data can help to reduce the scope thus costs of surveys, since key variables can be modelled and re-calibration accepts longer time intervals. Smaller, cheaper surveys could then be conducted more frequently, thereby increasing the value of the non-modelled data and making the national statistical system more efficient.

1. INTRODUCTION
“If you can’t measure it, you can’t manage it.”, Michael Bloomberg, former Mayor of New York City. A state’s budget can hardly be allocated efficiently, if the state does not know where the money is needed the most. Citizens cannot hold the government accountable for reforms, if they do not know which effects they had. Thus, knowledge about the dynamics of a society is the foundation of evidence-based policymaking. Traditionally, this knowledge is collected via censuses and household surveys and is provided by institutions of the national statistical system as a public good, the ‘official statistics’. The value of this knowledge is determined by its relevance. However, censuses are conducted about every 10 years, major household surveys every 3-5 years and both require a well-functioning infrastructure, starting from cars for the interviewers to computers and well-trained personnel for the analysis. With national statistical systems in developing countries often being subject to unstable funding and a lack of human resources, the collection and processing of relevant data imposes a great challenge, which too often cannot be met [5].

The underlying problem of this challenge is twofold: First, long cycles between censuses decrease the value of its information over time, shorter cycles are hardly financeable. Second, household surveys can partly fill the gap, however, they are based on samples and thus might not capture important local heterogeneity such as poverty pockets. High frequency data on individual level, often referred to as big data, covers a significant share of the population and might help to mitigate these shortcomings.

Mobile phone data, often referred to as digital breadcrumbs [9], contains the advantage of being collected as a side product. While it has been successfully used in e.g. constructing early warning systems for influenza [7], humanitarian responses after crises [2], disease modelling [10], transport optimization [3] and population estimation [4], it can also be misused for surveillance and thus is rightfully subject to privacy concerns.

To the best of our knowledge, this paper explores a so far untouched area in research by presenting an easily applicable approach to model a basket of socio-demographic indicators using mobile phone data only. Therefore, the analysis sets out to test the usability of mobile phone data, in this case antenna-to-antenna traffic in the Senegal from 2013, in official statistics by constructing fine granular key performance
indicators (KPI) for major socio-demographic variables. It is based on the assumption that sub-populations exhibit a distinct communication behaviour and can thus be identified. We show that mobile phone data is a promising tool to improve inter-censal estimations for a variety of variables while at the same time respecting the privacy of the individual. These findings have multiple implications: First, fine-grained data can capture important local heterogeneity and thus helps to discover pockets of misery. Second, estimates from big data might help to tailor the scope of surveys, since data that can be modelled does not have to be collected in every survey round, but for re-calibration of the model longer time intervals are sufficient. This can decrease the response burden and thus, the costs of surveys. Third, lower survey costs and smaller scopes can facilitate shorter survey cycles, thus improving also the relevance of data that cannot be sufficiently modelled using mobile phone data. Together with fine granular, almost real-time estimations, providers of official statistics can therefore become more efficient. Finally, more relevant data can contribute to improved decision-making.

The remaining paper is structured as follows: In Section 2, we briefly describe the datasets used in the analysis before explaining the methodological approach of this paper in Section 3. Section 4 presents the results of our analysis. After concluding in Section 5, we use Section 6 to point to caveats in the analysis and derive fields of possible further research.

2. DATA DESCRIPTION

The Republic of Senegal, short: The Senegal, is located in West Africa at the Atlantic Ocean between Mauritania to the North and Guinea-Bissau to the South. At the most Western tip lies Dakar, the country’s capital and also largest city. The set-up of administrative areas in the Senegal is complex, but can be divided into three different levels that are of interest in this paper: 14 regions, 45 départements and 123 arrondissements. The total population is estimated at about 13.5 million (2013) and consists of several ethnic groups, e.g. the Wolof or the Serer [14].

The mobile phone data used in this analysis consist of call detail records (CDR), covering the year 2013, from the Sene-
galese telecommunication company Sonatel and is provided in the context of the Orange/Sonatel Data for Development (D4D) Challenge 2014. To ensure privacy, the call detail records are anonymized and the locations of the towers slightly modified (+/- 2 kilometres). Further, only users with inter-
actions on more than 75% of the days, but less than 1000 interactions a week are included. While three different da-

tasets are provided, antenna-to-antenna traffic of all 1666 antennae, fine-grained mobility patterns of frequently chang-
ing sub-populations and coarse-grained mobility patterns for a selected sub-population, we concentrate on the first dataset in our analysis. Mobility patterns require sophisti-
cated privacy protection mechanisms and are thus expected to reduce the implementability of an approach [4]. Antenna-
to-antenna traffic is defined as the incoming and outgoing number of calls and SMS as well as the duration of each call for every tower on an hourly basis. For details on this and the other two datasets, see the official D4D documents [11].

We use geocoded data from the Demographic and Health Survey (DHS) 2011 as a benchmark. The DHS 2011 is a re-
presentative sample of 7,902 households covering 77,269 in-
dividuals in 391 geographic clusters in the Senegal [1]. While most variables such as age, sex or educational attainment are available for all household members, the variables literacy, religion and ethnicity are imputed in this paper from DHS sub-samples of 15,688 women and 4,929 men, respectively.

Figure 1 shows the GPS locations of the DHS cluster centres (orange crosses) next to the D4D antennae (blue points).

Using the DHS 2011 instead of using data from the recent 2013 general population census has several disadvantages: First, the power of our analysis heavily relies on the as-
sumption of representativeness of the survey. Second, sam-
ple variation and a possible sampling bias is not explicitly modelled in our analysis. These simplifications are assumed to be valid, as the DHS is a renowned global survey program with sample sizes in the survey clusters of the Senegal which mostly feature more 100 individuals (four clusters with less than 100 individuals). Table 1 shows the distribution parameters for the cluster sizes. Hence, the direct estimator is considered to be a fairly accurate proxy for the true population value.

Table 1: Distribution of DHS survey cluster sizes

<table>
<thead>
<tr>
<th>Min</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>163</td>
<td>197</td>
<td>197.6</td>
<td>230.5</td>
<td>409</td>
</tr>
</tbody>
</table>

Third, the mobile phone data used in this analysis describes network activities in 2013. Target variables, however, were measured in 2011. In order to derive valid relations from these two datasets, time invariance of the target variables has to be assumed. While, to a certain degree, this seems to be a reasonable assumption, especially given the relatively short time difference, it still limits the re-usability of the estimated parameters. However, the re-usability of the approach itself remains unaffected.

Figure 2 shows the ten arrondissements of the region Dakar. It gives a first impression of the population size and the antenna traffic on the administrative area level. The blue bar represents the SMS counts which range from 68 million
to over 500 million per arrondissement over the year in this region. The calls (orange bar) range from about 77 million to over 370 million.

**Figure 2: Population and mobile traffic (Dakar)**

Population numbers on the administrative area level are retrieved from the recent 2013 general population census [14]. The allocation of towers to the administrative area level and map data are extracted from the support files of the D4D dataset [11].

Landlines and the use of internet-based mobile communication services such as Skype, WhatsApp or Viber may cause distortions in communication patterns. A stagnating landline penetration rate of 2.8 % [12] and estimates of app downloads from Priori Data make the case for assuming a negligible selection effect. The all-time downloads of messaging applications compared to other countries are extremely low (e.g. WhatsApp 124,818 and Viber 95,891 on iOS as of 12/18/2014). Nevertheless, given the data preparation by Orange/Sonatel, some types of users may systematically be excluded. Explicitly modelling them, however, exceeds the scope of this paper.

### 3. METHODOLOGY

**Key performance indicators (KPIs) ought to mirror the structural patterns of the variables they rebuild. In order to create reliable KPIs, we benchmark our indicators on survey data, cross-validate the results and test their robustness in both the spatial and the time dimension. In time, we move from yearly to monthly aggregates; spatially, we downscale to the tower level. This section describes aggregation, allocation, model fitting and testing mechanisms to construct KPIs for a variety of socio-demographic indicators in the Senegal.**

#### 3.1 Cleaning and Preparation

Tower locations are provided as spatial points, i.e. single pairs of GPS coordinates (longitude - latitude). Also provided as spatial points are the geocoded DHS survey clusters. While the mapping of towers to administrative areas (arrondissements) is extracted from the D4D dataset, the DHS survey clusters are allocated by verifying in which boundaries the spatial points fall. Six of the 391 survey clusters do not provide GPS information and are therefore excluded from the analysis. The treatment of spatial polygons as spatial points shows a caveat of the analysis which should be addressed here: Since the actual coverage areas of clusters and towers are unknown, a distinct allocation of tower traffic to administrative areas is not possible. Hence, additional variation due to overlaps cannot feasibly be modelled.

Mobile phone data variables are aggregated to the year and growth rates are based on monthly aggregates. Survey variables are, if necessary, imputed, weighted and grouped into binary variables on the individual level and then calculated as percentages, i.e. shares on the aggregated level. Distances are measured in kilometres using the great-circle distance between spatial points. A list of all initial covariates and details on the transformation of the survey variables can be found in the appendix. Regarding towers that do not record any calls/SMS during the whole year, we cannot differentiate whether this is due to no events taking place, technical issues at these towers or anonymization as mentioned in the data description section.

#### 3.2 Modelling

The intuition behind modelling sub-populations (e.g. illiterates) in a spatial area with mobile phone data stems from the underlying assumption that sub-populations exhibit a distinct call and SMS behaviour. For example, one could assume that illiterates prefer to call than to write SMS, poor people prefer to be called than to call and people without steady electricity supply can communicate only irregularly. This paper aims to rebuild important socio-demographic indicators such as literacy and poverty from mobile phone data in a uniform approach as accurately as possible and does not intend to provide insights into possible causal relationships. Therefore, the model choice focuses substantially on fit optimization under robustness aspects and less on the interpretability of the standard errors. Here, we draw from the impressive work done in other papers [13] on creating covariates based only on antenna-to-antenna traffic to unlock as much information as possible from this dataset.

Since our dependent variables are population shares, we use simple linear regression models with dummy variables for the regions to control for heterogeneity between them. We use backward elimination on the set of initial variables and, if necessary, stepwise forward selection for interaction terms. In mathematical notation, our key performance indicators adhere to the following general structure:

$$E(Y|X) = \tilde{\beta}_0 + X_1^i\tilde{\beta}_1 + \ldots + X_n^i\tilde{\beta}_n + \ldots + X_{m}^i\tilde{\beta}_{(n+j)r}$$

$$\forall r \in R \land \forall i \in 1...n \land \forall j \in 1...m$$

$R$ consists of the 14 regions of the Senegal and $\tilde{\beta}_0$ contains an intercept for each of these regions. The term $X_i^j\tilde{\beta}_{(n+j)r}$ represents an interaction term, where $i$ is one of the covariates (maximum $n$ covariates) multiplied with a region-specific value $r$. The subscript $j$ reaches from 1 to $m$, i.e. the maximum number of interaction terms.

To avoid overfitting, we cross-validate our results across spatial dimensions. Instead of cross-validating predictions for
3.5 The Recipe
data only is summarized in the recipe below.

mely estimates of socio-demographic variables from antenna
The methodological approach to acquire high resolution, ti-
used.
with its 1174 tower, a grid with around 2000 hexagons is
grid with around 1000 hexagons, for the remaining Senegal
values. For the Dakar region with its 492 towers, we use a
xagons based on the distance to observed, here predicted,
fall into the same hexagon and then interpolate empty he-
grid over Senegal, average the tower estimates in case they
administrative areas in its distributions. To paint a more
this paper may exclusively adhere to the borders of the
Neither literacy, nor poverty, nor other variables modelled
indicators such as literacy or educational attainment. Fre-
ent food security or population estimates, however, espe-
ially in times of crises, can be of immense value for disaster
response. Thus, the validity of estimations based on monthly
aggregates is tested. Therefore, the procedure described
above is repeated for all twelve monthly aggregates indivi-
dually, the estimated coefficients are then compared over the
months and against the yearly aggregate. This approach has
another advantage. By following the parameter changes over
time, time patterns can be extracted and taken into account
for modelling future developments more accurately.

3.3 Robustness
So far, we have used annual aggregates for estimation. Cal-
culations based on such aggregates suffer, however, from two
disadvantages. First, data preparation involves high comput-
ing capacities, which might not always be available. Second,
sub-anual variability of the target variables is not captured.
Sub-anual availability of data appears to be less relevant for
indicators such as literacy or educational attainment. Fre-
frequent food security or population estimates, however, espe-
cially in times of crises, can be of immense value for disaster
response. Thus, the validity of estimations based on monthly
aggregates is tested. Therefore, the procedure described
above is repeated for all twelve monthly aggregates indivi-
dually, the estimated coefficients are then compared over the
months and against the yearly aggregate. This approach has
another advantage. By following the parameter changes over
time, time patterns can be extracted and taken into account
for modelling future developments more accurately.

3.4 Smoothing
Neither literacy, nor poverty, nor other variables modelled
in this paper may exclusively adhere to the borders of the
administrative areas in its distributions. To paint a more
data-driven picture, we use inverse distance weights for in-
terpolation on the tower level. Therefore, we lay a hexagonal
grid over Senegal, average the tower estimates in case they
fall into the same hexagon and then interpolate empty he-
xagons based on the distance to observed, here predicted,
values. For the Dakar region with its 492 towers, we use a
grid with around 1000 hexagons, for the remaining Senegal
with its 1174 tower, a grid with around 2000 hexagons is
used.

The methodological approach to acquire high resolution, ti-
emly estimates of socio-demographic variables from antenna
data only is summarized in the recipe below.

3.5 The Recipe
1. Match survey/census data and cell phone aggregates
   on sub-national (administrative area) level
2. Set up a basic linear model to regress socio-demographic
   indicators on cell phone data variables
3. Use model parameters for predictions on the tower le-
   vel and to calculate the IS-RMSE
4. Inflate basic model by adding interaction terms to find
   minimum of the sum of RMSE and IS-RMSE
5. Eliminate variables to further minimize the sum of
   RMSE and IS-RMSE
6. Use spatial smoothing on the tower predictions for high
   granular heat maps to identify e.g. pockets of misery.

4. RESULTS
Results of our analysis are presented using the literacy rate
 SHARE OF LITERATES PER AREA AS AN ONGOING EXAMPLE. DETAILED
results for other key variables are available upon request
from the authors. Figure 3 shows benchmark values and pre-
dictions with the 95% confidence intervals (CI) under nor-
mality. The blue line represents the share of literates calcula-
ted from the DHS survey which is used as the benchmark for
the model predictions based on mobile data (orange line).

Figure 3: Benchmark values and predictions with CI

Literacy varies strongly across arrondissements as depicted
in Figure 3. The model succeeds to capture most of the va-
riation with the confidence interval (grey band) staying close
to the estimated values, signifying a robust fit.

Table 2 presents different performance measures for selected
socio-demographic indicators on the arrondissement level.
The adjusted R² shows the share of total variation captured
from the respective model, thereby penalizing for the num-
ber of variables used for explanation. The root mean squa-
red error (RMSE) describes the precision of the estimates in
terms of bias and variance. Both measures indicate how well
given socio-demographic indicators can be re-built using mo-
bile phone data only. While this already demonstrates the
value of mobile phone data for national statistics, this va-

cue can further be leveraged by exploring the robustness of
predictions across different (spatial) dimensions, using the
IS-RMSE. The difference of RMSE and IS-RMSE for each
model shows how robust the model performs inter-spatially.
For example, the value for Literacy rate rises from 0.0502 to
0.0841 (about 67.5%) while Poverty rate rises from 0.1334 to
0.1587 (about 19%). Hence, the latter model does not per-
form as well in the first place but shows more inter-spatial
robustness.
Table 2: Basket precision

<table>
<thead>
<tr>
<th>Variable of interest</th>
<th>Adj. R²</th>
<th>RMSE</th>
<th>IS-RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literacy rate</td>
<td>0.9647</td>
<td>0.0502</td>
<td>0.0841</td>
</tr>
<tr>
<td>Poverty rate</td>
<td>0.8697</td>
<td>0.1334</td>
<td>0.1587</td>
</tr>
<tr>
<td>Primary completion rate</td>
<td>0.9177</td>
<td>0.0338</td>
<td>0.0456</td>
</tr>
<tr>
<td>Share of Minors</td>
<td>0.9966</td>
<td>0.0255</td>
<td>0.0343</td>
</tr>
<tr>
<td>Share of Elders</td>
<td>0.9625</td>
<td>0.0174</td>
<td>0.0177</td>
</tr>
<tr>
<td>Share of Women</td>
<td>0.9984</td>
<td>0.0162</td>
<td>0.0398</td>
</tr>
<tr>
<td>Share of Wolofs</td>
<td>0.9326</td>
<td>0.0805</td>
<td>0.1024</td>
</tr>
<tr>
<td>Share of Poular</td>
<td>0.9722</td>
<td>0.0421</td>
<td>0.0518</td>
</tr>
<tr>
<td>Share of Serer</td>
<td>0.9256</td>
<td>0.0248</td>
<td>0.0377</td>
</tr>
<tr>
<td>Safe water access rate</td>
<td>0.9244</td>
<td>0.1524</td>
<td>0.3092</td>
</tr>
<tr>
<td>Electricity access rate</td>
<td>0.8677</td>
<td>0.1483</td>
<td>0.1800</td>
</tr>
<tr>
<td>Share of cell phone owners</td>
<td>0.9910</td>
<td>0.0629</td>
<td>0.0801</td>
</tr>
</tbody>
</table>

One can see on the values of the adjusted $R^2$ that model performance fluctuates across indicators. This can have multiple reasons: First, linear regression models on shares only perform well if shares are approximately normally distributed between 0 and 1. In the case of excessive zeros and ones, other models may perform better. Second, the analysis is based on the assumption that sub-populations exhibit a distinct communication behaviour. While this assumption seems reasonable for some variables, it may not hold for others. Figure 4 provides insights into the in-sample fit of estimations on the arrondissement level. It shows the differences (in %) between benchmarks and model estimates for five basket variables to identify areas of model inaccuracy. Areas without benchmark information are greyed out.

Dark areas point to larger differences between survey values and estimates. For the variables displayed, these differences never rise above 15%. Accurate in-sample fits, however, do not prove the out-of-sample performance of the models.

Below, the key performance indicator for literacy is listed as an example. The parameter vector `region` is of length 14 and represents the region-specific intercepts/slopes. Formulas for other variables of the socio-demographic basket such as ethnicity can be found in the appendix.

\[
\text{literacy_rate} = \text{region} - \\
0.0002 \times \frac{\text{call_volume}}{\text{outgoing_sms}} - 0.4019 \times \text{sms_ratio} + \\
0.0833 \times \text{call_ratio} - 0.1096 \times \text{sms_entropy} + \\
0.1055 \times \text{call_entropy} - 0.0216 \times \text{mean_sms_distance} + \\
0.1240 \times \frac{\text{calls_to_Dakar}}{\text{outgoing_calls}} + \text{region} \times \frac{\text{calls_to_Dakar}}{\text{outgoing_calls}}
\]
The following plot shows the survey literacy rate (blue) next to the aggregated tower-level predictions (orange) on arrondissement level. The aggregated predictions are calculated as unweighted averages, since population values on the tower level are not available. Population weights could, however, improve the interpretive power of the averaged tower predictions.

Figure 5: Comparing benchmark values of literacy to its aggregated tower predictions

Compared to Figure 3 and as described in Table 2, the aggregated tower predictions perform only slightly less well than the arrondissement estimations. These results underpin the overall robustness of the approach across spatial dimensions. Robustness across time is verified by aggregating antenna data not to an annual, but to a monthly basis. While results are not displayed here, monthly aggregates of rather time-invariant variables such as literacy exhibit insignificant variation across time. Thus, a reduction to monthly antenna data aggregates seems feasible, thereby reducing the computing burden and facilitating sub-annual variation detection for disaster response.

The robustness of an approach is essential for its practical implementability and scalability. However, without an appropriate presentation of the results, an analysis can lose much of its potential impact. The presentation of results based on administrative areas is common, but carries two fundamental weaknesses: First, a granularity of 123 entities might be too coarse to uncover local heterogeneity such as pockets of poverty or illiteracy. Second, socio-demographic phenomena may not exclusively adhere to static geographical structures such as administrative districts. The visually implied homogeneity within administrative boundaries might thus be misleading. To reduce the possibility of misperception, we present in Figures 6 to 9 a more data-driven picture. Dakar and the rest of the Senegal are therefore split up to account for different tower densities.

Figure 6 shows the expected image. Literacy peaks in central Dakar and fades away with increasing distance to the city centre. In Figure 8, pockets of illiteracy are clearly visible. Overall, rural Senegal exhibits higher illiteracy. The coastal Casamance, Touba and other major cities in Senegal can be identified by brighter areas, signifying lower illiteracy. As mentioned in Section 3, this paper is not intended to provide insights into and therefore does not further discuss possible causal relationships.
Figure 8: Ethnicities and Poverty in Senegal (without Dakar)

Figure 9 shows the distribution of the three major ethnic groups in the Senegal: Wolof, Poular and Serer. The fourth heat map indicates the population share of the poorest fifth of the country. Poor areas of the Senegal are dominated by the Poular. This also holds true for areas in otherwise Wolof dominated regions.

Figure 9: Age, Education, Gender and Cell Phone Ownership in Senegal (without Dakar)

Figure 10 presents the shares of under 18 years olds, of people without completed primary education, of women and of people owning a cell phone. The share of women varies stronger than one would assume. Areas with a high share of people without school education often exhibit low cell phone penetration rates.

5. CONCLUSION

In our analysis we showed that key performance indicators can be estimated in unprecedented detail, virtually ad-hoc and, compared to traditional means of data collection such as surveys and censuses, at potentially little costs. This paper offers a uniform approach to modelling socio-demographic indicators from mobile phone data that can be extended to other variables without putting privacy at risk, since it is based on aggregated antenna-to-antenna traffic data only - data that is less prone to privacy concerns than e.g. mobility patterns. Hence, it is expected to improve the accessibility of mobile phone data.

However, mobile phone data is not omnipotent. Sub-populations such as illiterates can only be modelled reliably if they exhibit distinct communication patterns. Thus, it cannot replace traditional data collection methods. Mobile phone data key performance indicators still have to be ground-truthed, meaning, calibrated by the reality on the ground. Furthermore, even though treasures from new data sources are just started to be tapped in the context of national statistics and mobile phone data is just one part of it, it remains to be seen if today’s mobile phone data can capture the complexity of livelihoods to the degree modern-day surveys or censuses can. Nevertheless, mobile phone data, especially when combined with traditional data collection methods, can help to increase the relevance of data by adding detail and increasing timeliness.

Not only big data analytics develop, traditional data collection methods improve, too. Computer-assisted personal interviewing (CAPI) devices such as tablets are becoming global standard, standardized questionnaires and intelligent survey designs reduce the sample size on the one hand and increase the re-usability on the other hand and therefore, the value of the data collected. It is not unrealistic to assume convergence of big data and traditional data collection methods in the near future. For some parts of national statistics such as price data [8] or agricultural statistics [6] this is already in use.

6. FUTURE RESEARCH

This paper has shown that socio-demographic data can be estimated from mobile phone data, however, it has also pointed to caveats and weaknesses of the approach. First, the model presented in this paper is based on a couple of simplifications. While the assumptions seem to be appropriate in the respective situation, the approach would overall benefit from explicitly modelling so far neglected uncertainty. Sources of non-modelled uncertainty in this paper are: the use of surveys for benchmarking - here, a verification of the results using recent census data would increase the power of the analysis; the variation resulting from the imputation of missing values in the survey and possible overlaps and allocation uncertainty of coverage areas and administrative borders. Second, the time discrepancy between survey data collection (2011) and mobile phone data collection (2013) reduces the re-usability of our parameters, however, not of our approach. Nevertheless, in order to create KPIs that can further be used in practice, the time frame for data collection of both mobile phone and survey data should be the same. This is also expected to improve the model performance for fast-changing indicators. Third, the majority of the paper is
a cross-sectional analysis at a certain point in time. While in Section 4 we analyse the development of monthly aggregates, a more thorough analysis is needed to verify the prediction power of our models and to spot time patterns in the coefficients. Furthermore, the approach described in this paper has been tested for a basket of socio-demographic variables. It still has to be shown whether the approach can be extended to other variables, e.g. from health, economics or trade. Fourth, the analysis is based on the assumption that sub-populations exhibit distinct communication patterns. Additional variables generated from antenna-to-antenna data such as hourly antenna traffic or tower density can help to capture the diversity of call and SMS patterns in greater detail. This might further improve overall model performance. Fifth, we use basic linear models with dummy variables for the regions which includes the assumption of homogeneity of unobserved characteristics inside each region. More sophisticated models, e.g. random intercepts with random slopes would not need that assumption and also lose fewer degrees of freedom.

Another question yet to be answered for mobile phone data in the field of national statistics is whether it will be possible to institutionalize the cooperation of data providers and statistical agencies, while respecting the privacy of an individual and protecting citizens from governmental surveillance at the same time. In this paper, we used antenna-to-antenna traffic data only, since monthly or even hourly aggregates of network activity per antenna need little anonymization effort and are thus less prone to privacy concerns. As sampling theory revolutionized census practice, big data could revolutionize survey practice. The D4D challenge and this analysis are a first step towards it.

References


APPENDIX

The following covariates were calculated for calls and SMS separately. For simplification they will both be referred to as events in the following explanations.

A couple of variables are simple aggregations over the year such as **outgoing sms**, **outgoing calls**, **outgoing call volume**. Outgoing in this sense means that the tower is seen as the source of the event. The opposite is called incoming in our analysis. For each event the **ratios** are also calculated by **# of outgoing events / # of incoming events**.

**Mean distance** is defined as the average distance for an event per tower. It is calculated on the tower level by taking the distance of the outgoing tower to the incoming tower for each event and dividing it by the amount of events. The distance itself is calculated via the great-circle distance, which takes the earth’s curvature into account.

The **distance-to-dakar** covariate uses the distance from each tower to a centroid which is calculated by the centred location of the four arrondissements of the department Dakar. On the arrondissement level this distance is a weighted mean over the towers allocated to the arrondissement **Isolation** mirrors the variation of connections a tower maintains during the year. The idea of constructing this variable leads back to last year’s D4D challenge where it was already successfully implemented. [13] Activities between two towers adds a one to the isolation counter of the outgoing which ranges theoretically from 0 to 1666 (total number of towers available).

\[ I(t_i) = \sum_j E(t_i, t_j) \forall j \in 1, ..., 1666 \]

The indicator function \( I \) is 1 if the condition \( E(t_i, t_j) \) is true, i.e. an event happened between the two towers, and 0 otherwise. The intuition behind this variable is to quantify the diversity of interactions by users of a tower.

**Entropy** measures the average amount of information an event contains. The intuition behind this variable is that the more unlikely an event is to happen, the more information it contains. On a tower level we calculated the entropy with the following formula:

\[ E(t_i) = -\sum_{j \neq i} p(t_i, t_j) \log(p(t_i, t_j)) \]

The entropy of a tower is then defined as the sum of the probability of an event \( p(t_i, t_j) \) between this tower and all other towers times the logarithm of this probability. For each event and also call volume we calculated the **monthly growth** as well as the **variation** (i.e. variance) of monthly aggregates. **Calls-to-dakar** and **sms-to-dakar** reflects the amount of calls or sms for each tower that were directed at towers located in the Dakar department. **Interaction terms** are used in our analysis as random slopes to model area-specific effects of covariates. Both, the concept of random intercepts and random slopes are used for models with multiple hierarchical levels in the data. In the case of this paper, the analysed geographical units ‘arrondissement’ and ‘tower’ are both nested in the higher level hierarchical structure of a region.

The indicators from the DHS 2011, originally categorical variables, have been grouped to binary variables for this analysis. For ‘literacy’, the categories “able to read only parts of sentence” and “able to read whole sentence” have been grouped to ‘1’, ‘0’ otherwise. For ‘poverty’, the category ‘poorest’ was set to ‘1’ and ‘0’ otherwise. Every religion and every ethnicity listed in the DHS was transformed into a binary indicator, respectively, ‘young’ is classified as being 18 years of age or below, ‘old’ as 55 years of age and above at the time of the data collection. For ‘education’, an educational attainment of ‘completed primary’ or higher is grouped to ‘1’, ‘0’ otherwise. For ‘protected water’, ‘piped into dwelling’, ‘piped to yard/plot’, ‘public tap/standpipe’, ‘protected well’ and ‘protected spring’ are grouped to ‘1’, ‘0’ otherwise.

The indicators literacy, ethnicity and religion in the DHS survey had to be imputed. The Demographic and Health Survey provides additionally to the household survey male and female specific surveys on sub-populations of the household. The survey data is multiple imputed (mi5), grouped to binary variables and in this analysis, for simplicity, averaged. While the simplification ignores uncertainty induced by the imputation the resulting bias is assumed to be negligible, since the selection to the subsamples is random, thus the resulting unit non-response for subsample specific questions appears to be missing completely at random (MCAR).

**KPI formulae for remaining basket variables:**

\[
\text{poverty_rate} = \text{region} + 0.0002 \times \frac{\text{call_volume}}{\text{outgoing_sms}} + 1.2633 \times \text{sms_ratio} + 0.5342 \times \text{call_ratio} + 0.2964 \times \text{sms_entropy} - 0.2931 \times \text{call_entropy} + 0.0037 \times \text{mean_call_distance} + 0.0010 \times \text{distance_to_dakar} + \text{region} \times \text{call_ratio}
\]

\[
\text{share_of_minors} = \text{region} - 0.0042 \times \frac{\text{outgoing_sms}}{\text{outgoing_calls}} + 0.0001 \times \frac{\text{call_volume}}{\text{outgoing_sms}} - 0.0200 \times \text{call_entropy} - 0.2273 \times \frac{\text{calls_to_dakar}}{\text{outgoing_calls}} + \text{region} \times \frac{\text{calls_to_dakar}}{\text{outgoing_calls}} + \text{region} \times \frac{\text{outgoing_sms}}{\text{outgoing_calls}}
\]

\[
\text{share_of_elders} = \text{region} + 0.0095 \times \frac{\text{outgoing_sms}}{\text{outgoing_calls}} + 0.0702 \times \text{sms_ratio} - 0.0364 \times \text{call_ratio} - 0.0122 \times \text{sms_entropy} + 0.0117 \times \text{call_entropy} + 0.0001 \times \text{mean_sms_distance} + 0.0074 \times \frac{\text{calls_to_dakar}}{\text{outgoing_calls}}
\]

\[
\text{share_of_females} = \text{region} + 0.0231 \times \frac{\text{outgoing_sms}}{\text{outgoing_calls}} + 0.0001 \times \frac{\text{call_volume}}{\text{outgoing_sms}} - 0.0954 \times \text{sms_ratio} + 0.0002 \times \text{sms_isolation} - 0.0011 \times \text{call_isolation} - 0.0335 \times \text{sms_entropy} + 0.0194 \times \text{call_entropy} - 0.0005 \times \frac{\text{distance_to_dakar}}{\text{region}} + \text{region} \times \frac{\text{distance_to_dakar}}{\text{region}} + \text{region} \times \text{call_entropy}
\]

The entropy is defined as the average distance for an event per tower. It is calculated on the tower level by taking the distance of the outgoing tower to the incoming tower for each event and dividing it by the amount of events. The distance itself is calculated via the great-circle distance, which takes the earth’s curvature into account. The distance-to-dakar covariate uses the distance from each tower to a centroid which is calculated by the centred location of the four arrondissements of the department Dakar. On the arrondissement level this distance is a weighted mean over the towers allocated to the arrondissement. Isolation mirrors the variation of connections a tower maintains during the year. The idea of constructing this variable leads back to last year’s D4D challenge where it was already successfully implemented. Activities between two towers adds a one to the isolation counter of the outgoing which ranges theoretically from 0 to 1666 (total number of towers available).

\[ I(t_i) = \sum_j E(t_i, t_j) \forall j \in 1, ..., 1666 \]

The indicator function \( I \) is 1 if the condition \( E(t_i, t_j) \) is true, i.e. an event happened between the two towers, and 0 otherwise. The intuition behind this variable is to quantify the diversity of interactions by users of a tower.

Entropy measures the average amount of information an event contains. The intuition behind this variable is that the more unlikely an event is to happen, the more information it contains. On a tower level we calculated the entropy with the following formula:

\[ E(t_i) = -\sum_{j \neq i} p(t_i, t_j) \log(p(t_i, t_j)) \]

The entropy of a tower is then defined as the sum of the probability of an event \( p(t_i, t_j) \) between this tower and all other towers times the logarithm of this probability. For each event and also call volume we calculated the monthly growth as well as the variation (i.e. variance) of monthly aggregates. Calls-to-dakar and sms-to-dakar reflects the amount of calls or sms for each tower that were directed at towers located in the Dakar department.

Interaction terms are used in our analysis as random slopes to model area-specific effects of covariates. Both, the concept of random intercepts and random slopes are used for models with multiple hierarchical levels in the data. In the case of this paper, the analysed geographical units ‘arrondissement’ and ‘tower’ are both nested in the higher level hierarchical structure of a region.

The indicators from the DHS 2011, originally categorical variables, have been grouped to binary variables for this analysis. For ‘literacy’, the categories ‘able to read only parts of sentence’ and ‘able to read whole sentence’ have been grouped to ‘1’, ‘0’ otherwise. For ‘poverty’, the category ‘poorest’ was set to ‘1’ and ‘0’ otherwise. Every religion and every ethnicity listed in the DHS was transformed into a binary indicator, respectively, ‘young’ is classified as being 18 years of age or below, ‘old’ as 55 years of age and above at the time of the data collection. For ‘education’, an educational attainment of ‘completed primary’ or higher is grouped to ‘1’, ‘0’ otherwise. For ‘protected water’, ‘piped into dwelling’, ‘piped to yard/plot’, ‘public tap/standpipe’, ‘protected well’ and ‘protected spring’ are grouped to ‘1’, ‘0’ otherwise.

The indicators literacy, ethnicity and religion in the DHS survey had to be imputed. The Demographic and Health Survey provides additionally to the household survey male and female specific surveys on sub-populations of the household. The survey data is multiple imputed (mi5), grouped to binary variables and in this analysis, for simplicity, averaged. While the simplification ignores uncertainty induced by the imputation the resulting bias is assumed to be negligible, since the selection to the subsamples is random, thus the resulting unit non-response for subsample specific questions appears to be missing completely at random (MCAR).

KPI formulae for remaining basket variables:
\[
\text{share_of_wolof} = \text{region} - \\
0.0004 \ast \frac{\text{call_volume}}{\text{outgoing_sms}} - 0.2679 \ast \text{sms_ratio} + \\
0.3978 \ast \text{call_ratio} - 0.6220 \ast \text{call_entropy} - \\
0.0004 \ast \text{mean_sms_distance} + 0.0012 \ast \text{mean_call_distance} - \\
0.2462 \ast \frac{\text{calls_to_dakar}}{\text{outgoing_calls}} - 0.0264 \ast \text{distance_to_dakar} + \\
\text{region} \ast \text{distance_to_dakar} + \text{region} \ast \text{call_entropy}
\]

\[
\text{safe_water_access_rate} = \text{region} - \\
0.0001 \ast \frac{\text{call_volume}}{\text{outgoing_sms}} - 1.5330 \ast \text{sms_ratio} + \\
0.5433 \ast \text{call_ratio} - 0.0015 \ast \text{sms_isolation} + \\
0.0082 \ast \text{call_isolation} - 0.1034 \ast \text{sms_entropy} - \\
0.2727 \ast \text{call_entropy} - 0.0005 \ast \text{mean_sms_distance} - \\
0.2738 \ast \frac{\text{calls_to_dakar}}{\text{outgoing_calls}} - 0.0119 \ast \text{distance_to_dakar} + \\
\text{region} \ast \text{distance_to_dakar}
\]

\[
\text{share_of_poular} = \text{region} - \\
0.1686 \ast \text{call_ratio} + 0.1885 \ast \text{sms_entropy} + \\
0.0026 \ast \text{mean_call_distance} + 0.1125 \ast \frac{\text{calls_to_dakar}}{\text{outgoing_calls}} + \\
0.0143 \ast \text{distance_to_dakar} + \text{region} \ast \text{distance_to_dakar} + \\
\text{region} \ast \frac{\text{calls_to_dakar}}{\text{outgoing_calls}} + \text{region} \ast \text{sms_entropy}
\]

\[
\text{share_of_serer} = \text{region} - \\
0.0220 \ast \frac{\text{outgoing_sms}}{\text{outgoing_calls}} - 0.00002 \ast \frac{\text{call_volume}}{\text{outgoing_sms}} - \\
0.0379 \ast \text{sms_ratio} + 0.0002 \ast \text{sms_isolation} - \\
0.0003 \ast \text{call_isolation} - 0.0125 \ast \text{sms_entropy} - \\
0.0168 \ast \text{mean_sms_distance} + 0.0093 \ast \text{mean_call_distance} + \\
0.1657 \ast \frac{\text{calls_to_dakar}}{\text{outgoing_calls}} + 0.0050 \ast \text{distance_to_dakar} + \\
\text{region} \ast \text{distance_to_dakar} + \text{region} \ast \text{mean_sms_distance} + \\
\text{region} \ast \text{mean_call_distance}
\]

\[
\text{primary_completion_rate} = \text{region} + \\
0.0797 \ast \frac{\text{outgoing_sms}}{\text{outgoing_calls}} - 0.4642 \ast \text{sms_ratio} + \\
6.6716 \ast \text{call_ratio} - 0.0793 \ast \text{sms_entropy} + 0.0914 \ast \text{call_entropy} - \\
0.0015 \ast \text{mean_call_distance} - 0.2498 \ast \frac{\text{calls_to_dakar}}{\text{outgoing_calls}} + \\
0.0125 \ast \text{distance_to_dakar} + \text{region} \ast \text{call_ratio} + \\
\text{region} \ast \text{distance_to_dakar}
\]

\[
\text{electricity_access_rate} = \text{region} - \\
2.0363 \ast \text{sms_ratio} + 10.3692 \ast \text{call_ratio} - \\
0.4880 \ast \text{sms_entropy} + 0.4278 \ast \text{call_entropy} - \\
0.0044 \ast \text{mean_call_distance} - 0.9215 \ast \frac{\text{calls_to_dakar}}{\text{outgoing_calls}} + \\
\text{region} \ast \text{call_ratio}
\]
Mapping and Measuring Social Disparities in Senegal Regions using Mobile Phone Subscribers Data

ABSTRACT

The identification and measuring of social disparities among communities is an important function for the betterment of any country. These disparities, typically falling along lines defined by the indicators like gender, race/ethnicity, urbanization, infant mortality and social class, are quite often trivially visible in developing countries like Senegal. The advancement in Information and Communication Technologies has leads to the development of wide range of applications for use in community development. This has driven us to explore the usability of the mobile phone data for fact finding of disparities in communities. The proposed approach is all about the use of mobile data for finding the users’ location and tracking their mobility pattern. Using this information, we have made an effort to discuss few important hypotheses related to mapping and measuring of social disparities in Senegal region. We recommend that our findings and suggestions be incorporated into routine analyses of social development activities for the improvement of overall facilities in Senegal.

KEYWORDS

Social disparities; Urbanization; Mobility Patterns; Hypothesis Testing; Sustainable Development

1 Introduction

Ethnic diversity is one of the major forms of social complexity found in most contemporary societies [2]. We look into the facts of Senegal, an African country that has home for more than 11 ethnic groups with different backgrounds and distinct characteristics. Being predominantly rural in most of the regions, with limited amount of natural resources, there is a high need to stabilize the economy and boost the social status of the nation. The distribution of people belonging to various ethnic groups among the 14 major administrative divisions is not in line with the available opportunities to improve living conditions. Each of the groups have defined and historical ways of living, which are neither, at large, substantiated by the existing conditions, nor are there any measures to improve the quality of life.

Figure 1. Ethnic group distribution in Senegal
Figure 2. Regions in Senegal and their area

Given the differences in characteristics of the existing ethnic groups, there also exists a plethora of disparities in various social domains, including education, employment, health care, gender and so on, between the rural and urban populace of the nation. Rural regions are mostly devoid of any or all of the basic facilities, all of it being concentrated in the urban centres. According to the UN Human Development Report, 2014, Senegal’s HDI value for 2013 is 0.485—in the low human development category—positioning the country at 163 out of 187 countries and territories [1]. The really concerned disparities are chiefly to be addressed in order to improve the standard and accelerate the development of the nation.

‘Traditional’ causes of inequality such as land concentration, urban bias, the dominance of a highly concentrated mining sector and inequality in education do explain most of the variation in cross-country inequality. Decomposition analysis of inequality per region showed that, in 2011, about two-thirds of inequality is formed by within-region inequality and one third by between-region inequality. ‘New’ causes, as identified, are strongly linked to the neoliberal policy reforms that have been increasingly adopted. The causes include-

- **Land ownership** – It explains the high levels of income inequality, in mainly rural areas, and also the high income concentration in the urban areas by depressing minimum urban wages.
- **Education levels** - There is a strong negative linkage between average years of education and measured income inequality. A recorded low average level of education leads to increased inequality. People with better education levels are preferred more when it comes of job offers.
- **Consumption** - In Senegal, the richest fifth of the population is responsible for about half of total consumption. The share of the poorest fifth has stayed roughly constant at just seven percent of the total. The ratio of the proportion of consumption taken by the richest fifth of the population to that consumed by the poorest fifth has not really changed between 2001 and 2011. This means that richest fifth of the population consumes seven times more than the poorest fifth, leading to a high percentage of inequality.
- **Technology** - New technologies generate a demand for skills, which favours higher-skilled workers over lower-skilled ones and leads to increasing wage differentials between skilled and unskilled workers. Technical advances might have a greater impact on inequality in the future by turning formerly non-tradable services into international tradables
- **Globalization** - Globalization, in general, and trade liberalization, in particular, are perceived to have had a negative impact on income inequality. The import of world class technology—or the shift to high tech exports requiring highly educated labour—is one of the chief causes.
- **Liberalization** - Liberalization of the domestic and international financial system has caused an increase in income inequality much greater than that caused by other policy changes such as trade and labour market liberalization and privatisation.

- **Regressive taxation** - Through the combined effects of taxes and expenditure, the government can have a significant impact on levels of income inequality. Progressive tax and pro-poor expenditure policies will reduce inequality. However, over the last two decades, the main policy trend has broadly been in the opposite direction. Tax systems appear to have evolved towards greater use of indirect taxes (above all Value Added Tax-VAT) further increasing the inequality. Tax and pricing policies as well as public expenditure policies can have significant impacts on the distribution of opportunities and income between rural and urban areas.

ICT has been used as the basic approach to solve some of the existing problems. The reason for the choice is that they have been found to improve economic prosperity, employment creation and substitution, as well as social welfare. From the recent study, Senegal is an emerging leader in the field of ICT development. Mobile phone is one of the emerging ICT tool for the betterment of country’s infrastructure and development.

![Four clusters emerge based on each country’s current Internet economy and its foundations for future performance](source)

**Figure 3.** Senegal among its peers in Internet economy

![Senegal’s Internet penetration in 2012](source)

**Figure 4.** Senegal’s Internet penetration in 2012
Mobile voice services represent a mature market that has affected the evolution of the Senegalese economy during 2004-2011 and is now used by the majority of the population. Apart from the basic communication services, mobile networks offer the necessary platform to launch innovative services. In certain cases, these services increase the capacity of existing business channels (retail stores, geo-location, transport services) while in others, they tackle the lack of access to traditional services. Mobile infrastructures have significant effects on the creation of new markets and services, hence reducing unemployment and openness of the economy.

The importance of digital connectivity in the country may improve the sectors like education, health and others. Coverage is a key issue, as the relatively rural country would greatly benefit from the widespread availability of data services. Senegal is a fertile technological Greenfield already ripe for an abundance of applications and services that could drastically reduce poverty, increase life quality, sustain growth and promise a brighter future. A plethora of issues dealing with various communities residing in a country such as urban planning, transportational efficiency, spread of viruses and pathogens have long been difficult to understand and counteract owing to the lack of tools and manpower. However, since the mobile phones exploded into a ubiquitous presence in everyone's lives, plenty of research works have been emphasizing the fact of realizing these issues using patterns obtained from mobile phone records. From increasing transportational efficiency to determining the spread of malaria in rural lands, mobiles have been vital in the estimation and thus the prevention of various issues that the communities deal with.

Immense research has been done, with respect to the mobile subscribers, which aims at increasing the standard of living and addressing various concerns of the society. It is important to clearly distinguish the mobility patterns of rural and urban communities, in order to lay out clear cut conclusions of the ethnicity-based research. Mobility pattern analysis has been done in order to formulate the characteristics that make up the mobile users and to provide an elaborate understanding of the difference among the rural and urban subscribers. A comparative analysis, of the existing living standards between rural and urban societies, is done using mobile phone data.

**Mobility patterns**- a term generally used to characterize user movement, i.e. the different trajectories of subscribers. Human mobility patterns are influenced, on a large scale, by the regional socio-economic factors, such as population density, income and unemployment rate. Understanding communication patterns from mobile phone records is important to characterize human movements. Mobile phone records contain detailed information of the spatio-temporal localization of users, which can be used as proxies for human mobility. Individuals live and travel to different areas, yet each user can be assigned to a well-defined location where he/she is present most of the times.

Listed below, are the hypotheses about the disparities in the living conditions of rural and urban customers that will be examined and proved in detail by this paper:

**Hypotheses:**

- Gender disparity in employment is a serious concern
- Urbanization affects the growth of agricultural sector
- Infant and child mortality is high in rural areas compared to urban regions due to social disparity
- Social disparity is directly connected with lack of quality education, inadequate teaching materials and teachers
- Disease Spread and Socioeconomic disparity are highly connected
- Some ethnic groups, in spite of the location, involve in same profession
- Government policies do not directly address employment, gender, social protection and sustainable development
2 Related Work

Cell phones, in recent years, have become a quintessential part of human lives. The ever growing capabilities and power that the mobile phones bestow upon the subscribers has an incredible potential to influence even the miniature works that one undertakes and also helps to analyze and understand the subscribers’ characteristics along various fronts. The research and subsequent development of technologies targeting mobiles such as Mobile Banking, Mobile Health, Personal Assistant Applications, etc. underline their huge potential. Hence, data from simple mobile phones could enable us to understand how each person in a nation behaves and moves about, and also helps in medical, transportation, socio-economic and demographic implications. Beginning with the seminal work of Robert Park (1950), there have been numerous efforts to develop sociological theories to classify and explain the variations in ethnic inequality. Yet there is no single paradigm that dominates the field. The basis of most of the theories on ethnic inequality is that the type and organization of society shapes the structure of ethnic relations. Other factors such as the relative size of majority and minority groups are additional contingencies that affect the relations.

"Models of ethnic stratification" are basically measuring and interpreting interethnic socioeconomic inequality, typically in terms of income, occupation, and education. The interpretation is founded on causal models, estimated with multivariate regression techniques that measure how social background characteristics affect differential socioeconomic achievement. Another pioneering study along these lines was Lieberson and Fuguitt’s (1966) application of Markov chain techniques to estimate the separate effects of social origins and social mobility on black-white occupational differences.

The recent development of models within the field of social stratification can be dated from 1967 with the publication of Blau and Duncan’s The American Occupational Structure. The "model" uses the statistical techniques to test the causal or theoretical framework. Duncan’s clear exposition of how to "model" and interpret ethnic differences in stratification has greatly influenced contemporary studies of this topic.

One of the major frontiers of research on ethnic stratification is comparative analysis. Through, comparative analysis, one can specify certain attributes of the social structure of geographical areas or cities that may affect the levels and processes of ethnic stratification. Such studies are increasingly possible with the availability of very large data files to measure geographic areas and ecological characteristics. Hence, we can clearly perceive from the mentioned works that mobile phone usage and mobility information could aid us answer questions about the ethnic and socio-economic disparities in various corners of Senegal regions.

3 Mobility Patterns

3.1 What are mobility patterns?

Mobility patterns illustrate the movement of a mobile user, and how their location, velocity and acceleration changes over time. Mobility patterns can be depicted with the help of models. Mobility models characterize user movement patterns and the behavior of subscribers. Such models can provide performance parameters and can derive valuable solutions for complex cases.

3.2 Need for mobility patterns

Senegal is home to people belonging to a wide variety of ethnic groups. These people move around the cities to earn a living. Mobility patterns can lay out such movement trends among the people of different cities that can then be used to identify what kind of occupation is carried out in each city. Also, with basic knowledge of the main occupation of each region, it is easy to identify the different occupations each ethnic group is associated with. Furthermore, with the available traffic patterns, it is possible to determine the relative population density of the regions. Conclusions can be arrived at on details like economic status and relative literacy rates of each region, given they have direct relation with the mobile phone usage patterns.
3.3 Detection of mobility and usage patterns

The complete skeleton of the system is depicted in the figure below. The dataset is mined to derive region wise statistics on different grounds. This division can also be envisioned by the hypothesized view of the system. The results, obtained through the process, for all regions, are combined to obtain the inferences. This helps in providing the appropriate recommendations for the categorized rural and urban regions independently.

![Figure 5. Steps involved in analysis of data](image)

**Steps Involved**

1. **Preprocessing of data**
   
   Data is available from two different sources - statistical data and the mobile phone data. Data from mobile phones is available in abundance in the form of event-logs. The datasets are based on Call Detail Records (CDRs) having fields that give the complete information pertaining to a transaction involving a mobile subscriber. For the work presented in this paper, more than 9 Million transactions that took place over a particular period of time, from a reputed mobile service provider of Senegal has been used. These transactions include the all users belonging to the entire nation. Therefore, there was a need to segregate these data into segments of regional data. This was done with the help of the Site ID of a CDR, which identifies the Base Transceiver Station (BTS) or a sector of a BTS within a Location Area Code (LAC). This was used to group the users into the different region in Senegal. The dataset was further preprocessed to remove noise.

   Statistical data, on the other hand, is derived from a variety of sources available on the Internet, and
from reports published by the United Nations, Central intelligence Agency and World Bank and various other sources. These statistics are segregated based on different regions and are analyzed along various socio-economic parameters.

2. Attribute Identification
A call detail record can contain a variety of attributes and most of these attributes may not be relevant for the derivation of characteristic inferences. It was therefore necessary to identify only those set of attributes, which would be the ideally useful in recognizing the patterns. Based on previous research that has been done in this field, a set of attributes was selected. These include duration, time of day, frequency of calls made, the incoming and outgoing sites and regions to which the sites belong. These in turn were used in the arrival at the derived attributes like total movement, workforce, etc.
Along statistical lines, attributes such as social status, gender, education, occupation and health are filtered from the collected data and are used for analysis.

3. Mobility analysis
Having selected the necessary attributes, analyses on the respective sets of regional users were done in a comprehensive manner. The complete analysis is averaged out for different regions individually, and the comparison is presented in the following sections of this paper. The considered hypotheses have either been proven or disproven based on the findings of analysis, in order to identify the disparities among the communities.

4 Methods used

A. Chi-Square Goodness of Fit Test
Generally, this test is applied when the data has one categorical variable from a single population. It is used to determine whether the sample data are consistent with a hypothesized distribution. The chi-square goodness of fit test [23] is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- The population is at least 10 times as large as the sample.
- The variable under study is categorical.
- The expected value of the number of sample observations in each level of the variable is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

B. Hypothesis Test for Regression Slope
This section describes how to conduct a hypothesis test to determine whether there is a significant linear relationship between an independent variable \( X \) and a dependent variable \( Y \). The test focuses on the slope of the regression line

\[
Y = B_0 + B_1X
\]

where \( B_0 \) is a constant, \( B_1 \) is the slope (also called the regression coefficient), \( X \) is the value of the independent variable, and \( Y \) is the value of the dependent variable.

If we find that the slope of the regression line is significantly different from zero, we will conclude that there is a significant relationship between the independent and dependent variables.

C. Inferential Statistics
The methods of inferential statistics are (1) the estimation of parameter(s) and (2) testing of statistical hypotheses.

5 Mobile Data Set

The Call Detail Records (CDR) have been collected for a year’s period, from January 1, 2013 to December 31, 2013, with the customer identifiers are anonymized. The original dataset contained more than 9 million unique aliased mobile phone numbers. The datasets retained only users meeting the given criteria-

1. Users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset)

2. Users having had an average of less than 1000 interactions per week. The users with more than 1000 interactions per week were presumed to be machines or shared phones.

Three sampled and aggregated datasets were used that contains details as follows-

- **Dataset 1**: One year of site-to-site traffic for 1666 sites on an hourly basis
- **Dataset 2**: Fine-grained mobility data (site level) on a rolling 2-week basis with behavioral indicators at individual level for about 300,000 randomly sampled users, meeting the two criteria mentioned before, for each 2 week period
- **Dataset 3**: One year of coarse-grained (arrondissement level) mobility data with behavioral indicators at individual level for about 150,000 randomly sampled users, meeting the two criteria mentioned before, for a year.

6 Evaluation of the mentioned hypotheses

The hypotheses that were listed at the beginning of this paper were evaluated using the statistical analysis and prediction methodology. This section will look at the results of the evaluation.

- **Hypothesis 1: Gender disparity in employment is a serious concern for the development**

The total population of Senegal amounts to 13.6 million as of 2014. The working age group comprises of people ranging from 15-64 years of age. The percentage of total workforce is 54.65%, out of which the male and female workforce constitutes 47.06% and 52.94% respectively.

**Testing and Results**

The employment and unemployment percentages of both genders were obtained and participation ratios were calculated. Although the women eligible labor force is greater than that of men, the actual participation of women is found to be much less, compared to the male participation in certain communities. Results derived from the data set analysis also were in favor of the alternative hypothesis.

<table>
<thead>
<tr>
<th>Table 1. Workforce Statistics from Senegal Population</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Employed Workforce</strong></td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

*Percentages given in brackets

From the above table, we infer that the present employment ratio of male (0.8822) significantly high compare to that of female (0.6590). The above statistical data is further substantiated with the mobile phone data from the local Telecommunication operators. The following table consists of data gathered from a sample of the entire mobile phone usage among the population in the two selected distinct
regions for understanding gender disparity. Dakar is highly populated and more urbanized and Ziguinchor is less populated and rural.

**Table 2. Workforce Statistics based on mobile phone dataset**

<table>
<thead>
<tr>
<th>Region</th>
<th>Total Workforce</th>
<th>Eligible Workforce</th>
<th>Total Population</th>
<th>Employment Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakar</td>
<td>4459</td>
<td>2298</td>
<td>51.5%</td>
<td></td>
</tr>
<tr>
<td>Ziguinchor</td>
<td>58</td>
<td>27</td>
<td>46.55%</td>
<td></td>
</tr>
<tr>
<td>Entire Senegal</td>
<td>5506</td>
<td>2721</td>
<td>49.4%</td>
<td></td>
</tr>
</tbody>
</table>

This can be reasoned with analysis on the presence of the various ethnic groups and their characteristic day-to-day life hailing in the corresponding regions.

- Wolof, the most predominant ethnic group in Senegal, is influenced largely by Western culture. The Wolof women are not restricted and appear freely in public and employ themselves in a variety of activities. Hence, this could attribute to the larger employment in Wolof occupied region like Dakar (urban). On the other hand, Ziguinchor hosts only a handful of Wolof women and hence a small proportion of women employment is observed.

- Ziguinchor (rural) with low population consists mainly of ethnic groups such as Mandinka, Diola, Bandial, Mancagne and Manjack, where all of them exhibit similar characteristics close to that of Mandinka. The Mandinka women are highly orthodox and mostly spend their entire lifespan as housewives, taking care of their house and children. Hence, this could signify a meagre contribution of woman occupants to the employment in Ziguinchor.

These results can be generalized to the entirety of Senegal to prove the hypothesis of gender disparity and discussed some relevant reasons for the inequality. Thus, it is seen that there is need for women empowerment among the communities on a larger scale, given that currently only few of them allow women participation in labour. This awareness can be brought about with the help of education, so that people realize the need and implement the same in their everyday life.

- **Hypothesis 2: Urbanization affects the growth of agricultural sector**

The annual urban population growth rate has been 3.3% in the recent years. The total urban population was 41% as of 2013 and expected to grow up to 56.41 in 2021 [1]. People from other regions move to more urbanized regions like Dakar, Diourbel and Thies. With the increase in rural migration towards urban zones, it can only go in the direction of exponential increase in the urban percentage in the horizons of 2020. Although the area of coverage of Dakar is less than 0.3% of the total area of Senegal, it shelters more than 25% of the total population of the country that is more than 50% of the total urban population, which is 3.1 million as of 2013.

**Testing and Results**

The trends show a steady increase in urban population over the last four decades in conjunction with decrease in rural population, agreeing with the hypothesis. The data is analyzed to decipher results, which prove that urbanization has had a negative impact on the agricultural sector in Senegal.

**Table 3. Urban population growth trends in Dakar**

<table>
<thead>
<tr>
<th>Year</th>
<th>Percentage of Urban Population compared to the Total Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1961</td>
<td>22%</td>
</tr>
<tr>
<td>1976</td>
<td>32%</td>
</tr>
<tr>
<td>1988</td>
<td>39%</td>
</tr>
<tr>
<td>1994</td>
<td>45%</td>
</tr>
<tr>
<td>2015</td>
<td>56%</td>
</tr>
</tbody>
</table>
The continuation of this tendency will involve a rate of urbanization equal to 56.4% in 2021, and the major part of this population would be in Dakar, which accommodates each year 1,20,000 new comers.

Table 4. Results from secondary data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban Population (%)</td>
<td>23</td>
<td>35.8</td>
<td>38.9</td>
<td>40.3</td>
<td>42.4</td>
</tr>
<tr>
<td>Rural Population (%)</td>
<td>77</td>
<td>64.2</td>
<td>61.1</td>
<td>59.7</td>
<td>57.6</td>
</tr>
</tbody>
</table>

As more and more people move towards urban regions, there is need for increased residential area. Also, due to the recent developments in ICT and other fronts, more of the available land area is taken up for such activities, thus leaving behind lesser percentage for agricultural purposes.

The above data was available till the year 2012, which was then extended till 2014 with the help of trend projections using linear regression techniques.

Figure 6. Agricultural land in Senegal

Figure 7. Arable land in Senegal
The above data was available till the year 2012, which was then extended till 2014 with the help of trend projections using regression techniques.

![Forest Area](image)

**Figure 8.** Forest area in Senegal

The steady decrease in forestland area can be attributed to both natural factors and man-made ones like deforestation and new projects that require land for building their development centers and factories.

Data gathered from the mobile phone data set lay out the same trends, depicting the low workforce in rural areas and higher proportions in the urban. This is analyzed using two cities representing each characteristic, in a case study.

**Table 5.** Results from mobile phone dataset

<table>
<thead>
<tr>
<th>Regions</th>
<th>Total Workforce</th>
<th>Workforce within same locality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakar</td>
<td>2298</td>
<td>1304</td>
</tr>
<tr>
<td>Ziguinchor</td>
<td>27</td>
<td>14</td>
</tr>
<tr>
<td>Entire Senegal</td>
<td>2721</td>
<td>1524</td>
</tr>
</tbody>
</table>

Dakar houses most of the markets and trading centres, whereas Ziguinchor hosts primarily agricultural activities. It is clear that workforce engaged in agriculture is far less compared to that of the urban counterparts, thus showing the drastic effects of urbanization on the agricultural sector.

Urbanization can be attributed to better opportunities in the urban cities and lesser support for agriculture in the rural regions. Although effects of poor rainfall and extreme drought cannot be avoided, measures and policies can be introduced to revive the laydown and improve the revenue and GDP value contributed by agriculture, providing better future and thus persuading people to stay in their own residential regions.

- **Hypothesis 3:** Infant and child mortality is high in rural areas compared to urban regions due to social disparity

Child Mortality rate is one of the key factors that Senegal ranks highest among its Sub-Sahara African counterparts. According to the Senegal DHS 2010-11 study, for every 1000 children,

- 102 of the under-five children end up in the grave in the countryside compared to 62 in cities
Infant mortality stands almost same with 49 and 55 deaths under 1 year of life in cities and villages respectively.

However, the deviations were predominant in children after 1 year, where 46 of provincial children lose their lives in comparison with 19 downtown.

Hence, this deviation is one of the serious causes of concern in Senegal that is further driven by the existence of Social disparity per se.

**Testing and Results**

Analysis and results show that, though there has been a steady decline in the overall infant and child mortality rates over the years, there is still a great deal of deviation in the rates with respect to the social status of the people. The economically poor regions are devoid of necessary health-care facilities, leading to greater mortality, thus confirming the positivity of the hypothesis.

**Table 6. Health status indicators in Senegal**

<table>
<thead>
<tr>
<th>Stillbirth rate per 1,000 total births</th>
<th>34</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neonatal mortality rate per 1,000 live birth</td>
<td>24.4</td>
</tr>
<tr>
<td>Neonatal deaths</td>
<td>12,481</td>
</tr>
<tr>
<td>Infant mortality rate per 1,000 live births</td>
<td>45.2</td>
</tr>
<tr>
<td>Number of infant deaths</td>
<td>23,052</td>
</tr>
<tr>
<td>Under-five mortality rate per 1,000 births</td>
<td>59.6</td>
</tr>
<tr>
<td>Number of under-five deaths</td>
<td>29,975</td>
</tr>
</tbody>
</table>


**Figure 9.** Distribution of under-five deaths by age group
Figure 10. Geographic variations in under-five deaths

Source: Senegal DHS 2010-11

Figure 11. Distribution of causes of neonatal and under-five deaths

Figure 12. Inequities in mortality rates and health care

The chief reason for inequities among the urban and rural regions is that rural regions are devoid of basic facilities at large. Also, due to the lack of education, people don't have the knowledge of the necessary childcare routines.

Table 7. Results of Infant Mortality Rate in Rural and Urban regions

<table>
<thead>
<tr>
<th>Region</th>
<th>Infant Mortality Rate (per 1000 live births)</th>
<th>Rural/Urban</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ziguinchor</td>
<td>73</td>
<td>Rural</td>
</tr>
<tr>
<td>Fatick</td>
<td>88</td>
<td>Rural</td>
</tr>
<tr>
<td>Matam</td>
<td>89</td>
<td>Rural</td>
</tr>
<tr>
<td>Kaffrine</td>
<td>93</td>
<td>Rural</td>
</tr>
<tr>
<td>Sedhiou</td>
<td>142</td>
<td>Rural</td>
</tr>
<tr>
<td>Kolda</td>
<td>145</td>
<td>Rural</td>
</tr>
<tr>
<td>Kedougou</td>
<td>154</td>
<td>Rural</td>
</tr>
<tr>
<td>Thies</td>
<td>53</td>
<td>Urban</td>
</tr>
<tr>
<td>Dakar</td>
<td>59</td>
<td>Urban</td>
</tr>
<tr>
<td>Louga</td>
<td>80</td>
<td>Urban</td>
</tr>
<tr>
<td>St. Louis</td>
<td>91</td>
<td>Urban</td>
</tr>
<tr>
<td>Kaolack</td>
<td>98</td>
<td>Urban</td>
</tr>
<tr>
<td>Tambacounda</td>
<td>100</td>
<td>Urban</td>
</tr>
<tr>
<td>Diourbel</td>
<td>104</td>
<td>Urban</td>
</tr>
</tbody>
</table>

Greater mortality rates can be attributed to the fact that rural regions are served by lesser number of hospitals and specialists, than the urban regions. Providing all regions with better facilities on an equal basis can help reduce the disparities and mortality rates.
Hypothesis 4: Social disparity is directly connected with lack of quality education, inadequate teaching materials and teachers

Though education is compulsory and free up to the age of 16, due to inadequate facilities, many school-age children seek education and training through more informal means. Also, there are children involved in the workforce, although the legal age for such apprenticeships is supposed to be 15. Children who live in rural parts of the country are also at a disadvantage, and usually work in agriculture instead of attending school. Despite the increase in enrollment rates, education quality is severely constrained by a lack of trained teachers, a shortage of instructional resources and a challenging school environment. As a result, many Senegalese children have insufficient skills for their grade level.

Testing and Results

The results from the available statistics prove that urban children have more technical courses, schools and teachers at their disposal whereas countryside children don’t have such luxury, in affirmation with the hypothesis.

Table 8. School attainment levels- by region, age group

<table>
<thead>
<tr>
<th>Age group</th>
<th>Highest education level attained</th>
<th>Urban</th>
<th>Rural</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No schooling</td>
<td>155628</td>
<td>484483</td>
<td>640111</td>
</tr>
<tr>
<td>10-14</td>
<td>Primary or less</td>
<td>335442</td>
<td>277338</td>
<td>612780</td>
</tr>
<tr>
<td></td>
<td>Not completed lower secondary</td>
<td>17164</td>
<td>4008</td>
<td>21172</td>
</tr>
<tr>
<td></td>
<td>Completed lower secondary</td>
<td>1305</td>
<td>1687</td>
<td>2992</td>
</tr>
<tr>
<td></td>
<td>Not completed higher secondary</td>
<td>904</td>
<td>0</td>
<td>904</td>
</tr>
<tr>
<td>15-24</td>
<td>No schooling</td>
<td>300653</td>
<td>742406</td>
<td>1043061</td>
</tr>
<tr>
<td></td>
<td>Primary or less</td>
<td>348872</td>
<td>196126</td>
<td>544997</td>
</tr>
<tr>
<td></td>
<td>Not completed lower secondary</td>
<td>96283</td>
<td>31705</td>
<td>127988</td>
</tr>
<tr>
<td></td>
<td>Completed lower secondary</td>
<td>64236</td>
<td>17803</td>
<td>82040</td>
</tr>
<tr>
<td></td>
<td>Not completed higher secondary</td>
<td>93158</td>
<td>15461</td>
<td>108618</td>
</tr>
<tr>
<td></td>
<td>Completed higher secondary</td>
<td>15110</td>
<td>3123</td>
<td>18233</td>
</tr>
<tr>
<td></td>
<td>Technical &amp; professional</td>
<td>12248</td>
<td>193</td>
<td>12441</td>
</tr>
<tr>
<td></td>
<td>Higher education</td>
<td>9779</td>
<td>888</td>
<td>10667</td>
</tr>
</tbody>
</table>

Figure 13. Employment/Unemployment vs. Education Levels [20-24-age group]
Although there are more people moving towards the urban cities in search of better opportunities, not enough highly paid jobs are available. Hence, for want of more income, even children are sent for jobs that pay less or in unpaid activities. Job opportunities available for highly educated people are not in abundance.

![Figure 14. Wage Employment/Unpaid Employment vs. Education Levels](image)

The above figure confirms the stated, that more and more people with less education levels engage in unpaid or less wage activities as they do not possess the necessary skills for the technical jobs. Only people with better educational skills are paid more, though not much of such opportunities are available.

Table 9. Education relevant data in Urban and Rural Regions

<table>
<thead>
<tr>
<th>Category</th>
<th>Urban (%)</th>
<th>Rural (%)</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Schooling</td>
<td>16</td>
<td>36.85</td>
<td>52.85</td>
</tr>
<tr>
<td>Primary or less</td>
<td>18.55</td>
<td>9.75</td>
<td>28.3</td>
</tr>
<tr>
<td>Less than Lower secondary</td>
<td>5.1</td>
<td>1.6</td>
<td>6.7</td>
</tr>
<tr>
<td>Completed Lower secondary</td>
<td>3.4</td>
<td>0.9</td>
<td>4.3</td>
</tr>
<tr>
<td>Less than Higher secondary</td>
<td>4.95</td>
<td>0.75</td>
<td>5.7</td>
</tr>
<tr>
<td>Completed Higher secondary</td>
<td>0.8</td>
<td>0.15</td>
<td>0.95</td>
</tr>
<tr>
<td>Technical &amp; Professional</td>
<td>0.65</td>
<td>0</td>
<td>0.65</td>
</tr>
<tr>
<td>Higher Education</td>
<td>0.5</td>
<td>0.05</td>
<td>0.55</td>
</tr>
<tr>
<td><strong>Total (%)</strong></td>
<td><strong>49.95</strong></td>
<td><strong>50.05</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

Mobile phone usage is directly related to the level of education. For example, Dakar has the highest recorded literacy rate whereas Kedougou, one of the poorest regions, has a very less literacy rate.

Table 10. Relevant Results from mobile phone dataset

<table>
<thead>
<tr>
<th>Regions</th>
<th>Literacy Rate</th>
<th>Mobile Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakar</td>
<td>~100%</td>
<td>28852</td>
</tr>
<tr>
<td>Kedougou</td>
<td>7.9%</td>
<td>303</td>
</tr>
</tbody>
</table>

Rural regions are deprived of quality teachers and schools. Due to this, parents prefer sending their children to work as apprentices. But, their low educational background also doesn't earn them a better income. Urban regions on the other hand have better education facilities, thereby allowing them to pursue higher education and earn a better living, leading to the disparities. Thus, providing rural regions with such necessary facilities can help significantly reduce the disparities.
Hypothesis 5: Disease spread and socioeconomic disparity are highly connected

Some of the most common medical problems in Senegal include child mortality, maternal death, malaria, and sexual diseases including HIV/AIDS. There is a high disparity in both the quality and extent of health services between urban and rural areas. The greatest problems in public health are in the East and South (Louga, Kaolack, and Tambacounda) and the region of Cassamance. Some of the greatest barriers to health care utilization include lack of information, lack of communication, low number of health care workers, and social and religious barriers.

Major disparities exist in health care access for those living in urban versus rural areas-
1. Approximately 70% of doctors and 80% of pharmacists and dentists are located in Dakar, the capital city. However, only 42% of the Senegalese population lives in urban areas, such as Dakar, which means that few doctors are available to rural residents.

2. Of every 10,000 women who give birth, 24 will die in urban areas, but nearly 100 will die in rural areas. Additionally, there are major disparities in children's nutrition in urban versus rural areas, with those in rural areas being more heavily disadvantaged.

3. Also, distance from health care facilities, rough roads, and improper means of transportation limit healthcare access in Senegal. For 80.5% of households, the poorly equipped health post is the only accessible health facility in an average distance of 4.3 kilometers. The closest hospital is located, on average, 20km away from the village of the household. Therefore, only 32% of rural households have regular access to a health center and thereby access to a medical doctor.

4. According to data from 2005, 14.5% percent of Senegalese children under the age of 5 are underweight. Only 42% of children between 12 and 23 months receive all necessary vaccinations

Testing and Results

The results state that disease spread is directly attributed to the disparity among the societal and economic statuses of the people. Hence, this means that unavailability of standard hospital facilities and inadequate skilled personnel attending to patients are the main causes of the spread of the diseases. People in economically weaker regions succumb more to ill-fated epidemic diseases, thus proving our hypothesis.

<table>
<thead>
<tr>
<th>Background characteristics</th>
<th>Percentage of all facilities offering malaria diagnosis and treatment services</th>
<th>Total number of facilities</th>
<th>Guidelines for diagnosis and treatment of malaria</th>
<th>Diagnostics</th>
<th>Number of facilities offering malaria diagnosis and treatment services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facility type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital</td>
<td>93</td>
<td>17</td>
<td>69</td>
<td>51</td>
<td>66</td>
</tr>
<tr>
<td>Health center</td>
<td>99</td>
<td>30</td>
<td>80</td>
<td>53</td>
<td>84</td>
</tr>
<tr>
<td>Health post</td>
<td>99</td>
<td>317</td>
<td>97</td>
<td>60</td>
<td>78</td>
</tr>
<tr>
<td>Managing authority</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>100</td>
<td>302</td>
<td>90</td>
<td>57</td>
<td>86</td>
</tr>
<tr>
<td>Private</td>
<td>93</td>
<td>62</td>
<td>59</td>
<td>20</td>
<td>55</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dakar</td>
<td>94</td>
<td>71</td>
<td>69</td>
<td>36</td>
<td>68</td>
</tr>
<tr>
<td>Diourbel</td>
<td>100</td>
<td>25</td>
<td>89</td>
<td>43</td>
<td>88</td>
</tr>
<tr>
<td>Fathe</td>
<td>100</td>
<td>22</td>
<td>94</td>
<td>31</td>
<td>100</td>
</tr>
<tr>
<td>Kaffrine</td>
<td>100</td>
<td>14</td>
<td>100</td>
<td>91</td>
<td>62</td>
</tr>
<tr>
<td>Kaolack</td>
<td>100</td>
<td>23</td>
<td>80</td>
<td>61</td>
<td>94</td>
</tr>
<tr>
<td>Koldougou</td>
<td>96</td>
<td>9</td>
<td>80</td>
<td>43</td>
<td>72</td>
</tr>
<tr>
<td>Kolda</td>
<td>100</td>
<td>17</td>
<td>81</td>
<td>57</td>
<td>72</td>
</tr>
<tr>
<td>Louga</td>
<td>100</td>
<td>23</td>
<td>96</td>
<td>75</td>
<td>77</td>
</tr>
<tr>
<td>Matam</td>
<td>100</td>
<td>17</td>
<td>100</td>
<td>75</td>
<td>83</td>
</tr>
<tr>
<td>Saint Louis</td>
<td>100</td>
<td>29</td>
<td>95</td>
<td>85</td>
<td>91</td>
</tr>
<tr>
<td>Sédhiou</td>
<td>100</td>
<td>12</td>
<td>97</td>
<td>65</td>
<td>86</td>
</tr>
<tr>
<td>Tambacounda</td>
<td>100</td>
<td>24</td>
<td>83</td>
<td>61</td>
<td>80</td>
</tr>
<tr>
<td>Thiesio</td>
<td>100</td>
<td>48</td>
<td>82</td>
<td>49</td>
<td>84</td>
</tr>
<tr>
<td>Ziguinchor</td>
<td>100</td>
<td>29</td>
<td>86</td>
<td>80</td>
<td>77</td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>364</td>
<td>85</td>
<td>58</td>
<td>78</td>
</tr>
<tr>
<td>Health hut</td>
<td>89</td>
<td>24</td>
<td>61</td>
<td>34</td>
<td>60</td>
</tr>
</tbody>
</table>
Figure 15. Malaria- healthcare facilities

<table>
<thead>
<tr>
<th>Background characteristics</th>
<th>Percentage of facilities that have TB smear microscopy</th>
<th>System for diagnosing HIV among TB clients</th>
<th>First line treatment for TB</th>
<th>Injectable streptomycin</th>
<th>Number of facilities offering any TB diagnostic, treatment, and/or follow-up services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facility type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital</td>
<td>72</td>
<td>86</td>
<td>57</td>
<td>49</td>
<td>31</td>
</tr>
<tr>
<td>Health center</td>
<td>71</td>
<td>98</td>
<td>71</td>
<td>81</td>
<td>70</td>
</tr>
<tr>
<td>Health post</td>
<td>3</td>
<td>87</td>
<td>31</td>
<td>58</td>
<td>15</td>
</tr>
<tr>
<td>Managing authority</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>15</td>
<td>94</td>
<td>42</td>
<td>69</td>
<td>29</td>
</tr>
<tr>
<td>Privé</td>
<td>21</td>
<td>56</td>
<td>17</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dakar</td>
<td>21</td>
<td>73</td>
<td>49</td>
<td>74</td>
<td>43</td>
</tr>
<tr>
<td>Doualba</td>
<td>37</td>
<td>100</td>
<td>68</td>
<td>68</td>
<td>28</td>
</tr>
<tr>
<td>Fatim</td>
<td>6</td>
<td>88</td>
<td>30</td>
<td>30</td>
<td>6</td>
</tr>
<tr>
<td>Kaffrine</td>
<td>9</td>
<td>100</td>
<td>16</td>
<td>32</td>
<td>11</td>
</tr>
<tr>
<td>Kedougou</td>
<td>23</td>
<td>100</td>
<td>32</td>
<td>65</td>
<td>66</td>
</tr>
<tr>
<td>Kolda</td>
<td>10</td>
<td>94</td>
<td>4</td>
<td>03</td>
<td>10</td>
</tr>
<tr>
<td>Lomé</td>
<td>10</td>
<td>94</td>
<td>4</td>
<td>03</td>
<td>10</td>
</tr>
<tr>
<td>Louga</td>
<td>39</td>
<td>100</td>
<td>39</td>
<td>65</td>
<td>23</td>
</tr>
<tr>
<td>Matam</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Saint Louis</td>
<td>100</td>
<td>100</td>
<td>57</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Sédhiou</td>
<td>13</td>
<td>100</td>
<td>63</td>
<td>76</td>
<td>6</td>
</tr>
<tr>
<td>Tambacounda</td>
<td>17</td>
<td>87</td>
<td>9</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>Thies</td>
<td>11</td>
<td>84</td>
<td>39</td>
<td>79</td>
<td>24</td>
</tr>
<tr>
<td>Ziguinchor</td>
<td>9</td>
<td>100</td>
<td>46</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>98</td>
<td>39</td>
<td>81</td>
<td>24</td>
</tr>
<tr>
<td>Health hut</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

NA = Not applicable

Figure 16. Tuberculosis- healthcare facilities

It is mostly localized in Dakar and Thies, and affects more men than women, with approximately 9500 cases of tuberculosis per year in Senegal and a 2.4% mortality rate.

<table>
<thead>
<tr>
<th>Background characteristics</th>
<th>Percentage of all facilities with HIV testing system</th>
<th>Number of facilities</th>
<th>Percentage of facilities with HIV testing system that have the following features</th>
<th>Number of facilities having HIV testing system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facility type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital</td>
<td>61</td>
<td>17</td>
<td>100</td>
<td>46</td>
</tr>
<tr>
<td>Health center</td>
<td>79</td>
<td>30</td>
<td>100</td>
<td>64</td>
</tr>
<tr>
<td>Health post</td>
<td>83</td>
<td>317</td>
<td>100</td>
<td>68</td>
</tr>
<tr>
<td>Managing authority</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>91</td>
<td>302</td>
<td>100</td>
<td>46</td>
</tr>
<tr>
<td>Privé</td>
<td>39</td>
<td>62</td>
<td>100</td>
<td>56</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dakar</td>
<td>62</td>
<td>71</td>
<td>100</td>
<td>65</td>
</tr>
<tr>
<td>Doualba</td>
<td>93</td>
<td>25</td>
<td>100</td>
<td>66</td>
</tr>
<tr>
<td>Fatim</td>
<td>88</td>
<td>22</td>
<td>100</td>
<td>93</td>
</tr>
<tr>
<td>Kaffrine</td>
<td>100</td>
<td>14</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Kedougou</td>
<td>83</td>
<td>23</td>
<td>100</td>
<td>93</td>
</tr>
<tr>
<td>Kolda</td>
<td>80</td>
<td>9</td>
<td>100</td>
<td>34</td>
</tr>
<tr>
<td>Louga</td>
<td>68</td>
<td>17</td>
<td>100</td>
<td>15</td>
</tr>
<tr>
<td>Matam</td>
<td>80</td>
<td>23</td>
<td>100</td>
<td>75</td>
</tr>
<tr>
<td>Saint Louis</td>
<td>100</td>
<td>17</td>
<td>100</td>
<td>13</td>
</tr>
<tr>
<td>Sédhiou</td>
<td>91</td>
<td>12</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>Tambacounda</td>
<td>83</td>
<td>24</td>
<td>100</td>
<td>35</td>
</tr>
<tr>
<td>Thies</td>
<td>84</td>
<td>48</td>
<td>100</td>
<td>32</td>
</tr>
<tr>
<td>Ziguinchor</td>
<td>91</td>
<td>29</td>
<td>100</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>82</td>
<td>364</td>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td>Health hut</td>
<td>NA</td>
<td>74</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

NA = Not applicable

Figure 17. HIV/AIDS- healthcare facilities

The Cassamance region has the highest prevalence of HIV/AIDS at 2.0%, which can be attributed in part to the Cassamance conflict. There are about 59,000 people in Senegal living with HIV/AIDS, according to a 2009 estimate.
Cardiovascular diseases - healthcare facilities

It is the second largest cause of mortality in Dakar. Risk factors include:
1. High BP [50%]
2. Smoking [47%]
3. Obesity [23%]
4. Cholesterol [12.5%]
5. Diabetes [11.6%]

<table>
<thead>
<tr>
<th>Disease</th>
<th>Urban Centres</th>
<th>Rural Centres</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malaria</td>
<td>239</td>
<td>119</td>
</tr>
<tr>
<td>TB</td>
<td>95</td>
<td>59</td>
</tr>
<tr>
<td>HIV / AIDS</td>
<td>190</td>
<td>108</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>233</td>
<td>118</td>
</tr>
</tbody>
</table>

Figure 18. Cardiovascular diseases - healthcare facilities

Figure 19. Distribution of basic healthcare facilities in Senegal

Table 11. Disease Spread data
Table 12. Disease Spread in measurable level from secondary data

<table>
<thead>
<tr>
<th>Disease</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malaria</td>
<td>44.4</td>
<td>48</td>
<td>22.6</td>
<td>21.1</td>
<td>27.2</td>
<td>163.3</td>
</tr>
<tr>
<td>TB</td>
<td>12.4</td>
<td>7.2</td>
<td>9.7</td>
<td>6.7</td>
<td>6</td>
<td>42</td>
</tr>
<tr>
<td>Angina</td>
<td>13</td>
<td>5.2</td>
<td>8.7</td>
<td>7.7</td>
<td>8.6</td>
<td>43.2</td>
</tr>
<tr>
<td>Arthritis</td>
<td>21.2</td>
<td>17.1</td>
<td>21.5</td>
<td>19.3</td>
<td>18.6</td>
<td>97.7</td>
</tr>
<tr>
<td>HIV</td>
<td>1</td>
<td>1.2</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>3.8</td>
</tr>
<tr>
<td>Total</td>
<td>92</td>
<td>78.7</td>
<td>63.1</td>
<td>55.3</td>
<td>60.9</td>
<td>350</td>
</tr>
</tbody>
</table>

[Q1 – Poorest Quintal and Q5 – Richest Quintal]

Mobile data suggests similar conditions in Senegal. Considering as a case study, Dakar and Tambacounda, the number of hospitals per arrondissement give a clear picture of the scenario.

Table 13. Hospital – Arrondissement Ratio

<table>
<thead>
<tr>
<th>Regions</th>
<th>Number of Hospitals</th>
<th>Arrondissement count</th>
<th>Hospital/arrondissement ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakar</td>
<td>50</td>
<td>10</td>
<td>5.000</td>
</tr>
<tr>
<td>Tambacounda</td>
<td>4</td>
<td>13</td>
<td>0.308</td>
</tr>
</tbody>
</table>

As is evident from the above data, urban centers are equipped with better health care facilities than their rural counterparts. People from the backward regions have to travel miles to reach the nearest facility. Thus, they neither made aware of any new diseases nor are they provided necessary treatments. This makes it mandatory to equip the rural regions with better health care facilities so that they are deprived of the pain endured.

Hypothesis 6: Some ethnic groups, in spite of the location, involve in same profession

Senegal houses more than 11 ethnic groups among its 13 million population. The concentration of various ethnic groups with respect to the total population residing in Senegal is –

- Wolof – 43.3%
- Pulaar – 23.8%
- Serer – 14.7%
- Diola – 3.7%
- Mandinka – 3%
- Soninke – 1.1%
- European and Lebanese – 1%
- Others – 9.4%
Testing and Results

The results suggest the importance and credibility of our hypothesis. It is fairly evident that since most of Senegal’s occupation is agro-based, most of the people indulge themselves in agriculture and related work. It also tells that despite their diverse localities, the various ethnic groups are involved in same kind of activities for a living.

**Table 14.** Ethnic groups and their main occupation

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Presence in Regions</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wolof</td>
<td>Dakar, Kaolack, St. Louis, Theis, Fatick, Louga, Diourbel and Kaffrine</td>
<td>Sedentary farmers and trade</td>
</tr>
<tr>
<td>Lebou</td>
<td>Dakar and St. Louis</td>
<td>Fishermen, farmers, construction and supply business</td>
</tr>
<tr>
<td>Pulaar</td>
<td>Dakar, St. Louis, Kaolack, Tambacounda, Matam and Kolda</td>
<td>Commercial, breeding, peanut farming</td>
</tr>
<tr>
<td>Serer</td>
<td>Dakar, Thies, Diourbel, Kaolack, Tambacounda, Fatick and Kaffrine</td>
<td>Peanut cultivation</td>
</tr>
<tr>
<td>Diola</td>
<td>Dakar, Ziguinchor and Sedhiou</td>
<td>Rice cultivation and fishing</td>
</tr>
<tr>
<td>Mandinka</td>
<td>Dakar, Tambacounda, Kolda, Sedhiou and Kedougou</td>
<td>Peanuts, rice, millets and husbandry</td>
</tr>
<tr>
<td>Fula</td>
<td>Dakar and Ziguinchor</td>
<td>Livestock, goat and sheep herding</td>
</tr>
<tr>
<td>Bassari</td>
<td>Tambacounda and Kedougou</td>
<td>Rice, millets and corn cultivation</td>
</tr>
<tr>
<td>Mancagne</td>
<td>Ziguinchor</td>
<td>Rice cultivation and fishing</td>
</tr>
</tbody>
</table>
Results from mobile phone data set agree with those from secondary data. Assuming regions with similar main occupation have similar workforce per area ratio, it is seen that regions with agriculture as chief occupation agree with one another, as the other regions with trading and markets, and other activities. Thus the ethnic groups residing in a region decide the chief occupation of a region, and regions bearing the same ethnic group population host the same chief occupation.

Table 16. Relevant Workforce Results from sample mobile phone dataset

<table>
<thead>
<tr>
<th>Region</th>
<th>Workforce</th>
<th>Area(sq.m)</th>
<th>Workforce/Area(sq.m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAKAR</td>
<td>2298</td>
<td>547</td>
<td>4.20138</td>
</tr>
<tr>
<td>THIES</td>
<td>168</td>
<td>6670</td>
<td>0.02533</td>
</tr>
<tr>
<td>DIOURBEL</td>
<td>36</td>
<td>4824</td>
<td>0.00753</td>
</tr>
<tr>
<td>KAOLACK</td>
<td>25</td>
<td>5357</td>
<td>0.00475</td>
</tr>
<tr>
<td>SAINT-LOUIS</td>
<td>82</td>
<td>19241</td>
<td>0.00430</td>
</tr>
<tr>
<td>ZIGUINCHOR</td>
<td>27</td>
<td>7352</td>
<td>0.00371</td>
</tr>
<tr>
<td>FATICK</td>
<td>9</td>
<td>6849</td>
<td>0.00134</td>
</tr>
<tr>
<td>SEDHIOU</td>
<td>5</td>
<td>7341</td>
<td>0.00078</td>
</tr>
</tbody>
</table>
Considering case studies of different ethnic groups:

- Ethnic groups like Serer, Diola, Pulaar, Fula and Mandinka are basically farmers and resort to cultivation of crops for a living. Dakar hosts people belonging to most of these groups, but fails to provide them agricultural opportunities. Hence these people tend to move outside of Dakar to earn a living. This is substantiated by the higher movement values pertaining to Dakar.

- Ethnic groups like Lebou and Wolof tend to concentrate in, or move to urban cities, given their main way of earning a living is trading and markets. Also, these people are less concentrated in other cities, like St. Louis. The other groups resort to farming in their residential area. This explains the lower movement values in those regions, as suggested by the mobile phone data set.

**Table 17. Movement rates derived from mobile phone dataset**

<table>
<thead>
<tr>
<th>Regions</th>
<th>Movement rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakar</td>
<td>4458.66675</td>
</tr>
<tr>
<td>St. Louis</td>
<td>91.51252088</td>
</tr>
<tr>
<td>Kedougou</td>
<td>5.974995727</td>
</tr>
</tbody>
</table>

Ethnic groups have certain historical characteristics and are skilled in one particular area of work that they do for a living. Their region of residence doesn't affect their occupation. They make sure that they exercise their best skills, even if they have to travel around for the opportunities that their residential region lacks.

- **Hypothesis 7:** Government policies do not directly address employment, gender, social protection and sustainable development

The economic and social policies implemented should be beneficial to the populations. The quest for better health coverage, a quality education system, and access to water, sanitation and a decent living environment are necessary at the moment. Also, there is need to ensure social protection, eradicate exploitation of children and improve education services.

**Testing and Results**

**Employment related policies:**

- The new National Employment Policy is based on the relevant public policies on the promotion of employment.
- The average annual number of new potential job seekers is 202000.
- To ensure full employment, the economy should generate close to 150,000 jobs every year.
- For the next five years, the strategy targets the following major objectives-
  
  1. Promote massive job creation- through the promotion of public investments in highly labour intensive activities.
  2. Improve monitoring and management of labor market

- There are plans to create at least 500,000 jobs during the 2013-2017 period.
Although measures have been said to be taken to decrease unemployment rates over the years, no significant change has been observed.

**Figure 20.** Unemployment rates in Senegal over the years

As revenue from agricultural activities decrease, more and more people migrate to urban cities in search of jobs. This demand for jobs is not compensated by equal job creation. Thus, large numbers of people are left unemployed or are made to work for no wages.

**Social Protection related policies:**

- Social protection comprises of series of measures to protect populations against the occurrence of social risks.
- It includes both public and private, or community based security schemes and is driven by three principles-
  1. Assistance
  2. Insurance coverage for various services
  3. Empowerment of social groups

- Senegal has taken the Social Protection Floor(SPF) initiative which aims to provide an enhanced access to essential social services and transfers for the poorest and most vulnerable.
- The strategic objectives inclue-
  1. Enhance social security for workers and retired persons
  2. Improve the socio-economic conditions of vulnerable groups
  3. Extend social protection to the informal sector and vulnerable groups

**Crime in Senegal**

- Minor street crime is very common in Senegal, particularly in urban areas.
- Dakar is considered to be a high-crime city, according to the number and frequency of crime incidents.
- Corruption and immunity remain a serious issue in the police force in Senegal.
- About 77% of households in Senegal consider the police to be corrupt, according to a recent survey.
Table 18. Different crimes and their quantifiers

<table>
<thead>
<tr>
<th>Crime</th>
<th>Quantifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Murder rate per million</td>
<td>83.91</td>
</tr>
<tr>
<td>Rape rate per million</td>
<td>5.6</td>
</tr>
<tr>
<td>Number of people possessing guns</td>
<td>2 in 100</td>
</tr>
<tr>
<td>Drug users</td>
<td>0.03%</td>
</tr>
<tr>
<td>Number of prisoners</td>
<td>~5500</td>
</tr>
</tbody>
</table>

**Gender parity in elective offices:**

There has been significant rise in the participation of women in elective offices.

![Figure 21. Percentage of seats held by women in Elective Offices](image)

Currently there is 43.3% women participation in the national parliaments. The sudden increase in the year 2012 and hence constant participation of women can be attributed to the fact that a new governmental policy came into effect that year, stating increased reservation of seats for women.

**Basic old-age and disability pension:**

- In Senegal, old people above 60 years of age and persons of working age (15-59) who are disabled are eligible for basic pension
- About one-percent of the working age population is eligible for a disability pension.
- Although only 6% of total Senegal population is eligible for this pension, about 45% of households benefit indirectly from it.

Table 19. Beneficiary rates of old-age and disability pension

<table>
<thead>
<tr>
<th>Senegal</th>
<th>Dakar</th>
<th>Other urban</th>
<th>Senegal</th>
<th>Dakar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligible persons</td>
<td>4.5</td>
<td>6.0</td>
<td>6.7</td>
<td>6.1</td>
</tr>
<tr>
<td>Individuals living together with beneficiary</td>
<td>40.8</td>
<td>50.6</td>
<td>58.7</td>
<td>53.1</td>
</tr>
<tr>
<td>Households with beneficiary</td>
<td>29.6</td>
<td>43.8</td>
<td>53.6</td>
<td>45.5</td>
</tr>
</tbody>
</table>

Table 20. Targeted cash transfer for vulnerable communities

<table>
<thead>
<tr>
<th>Senegal</th>
<th>Dakar</th>
<th>Other urban</th>
<th>Rural</th>
<th>Senegal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligible households</td>
<td>9.2</td>
<td>13.8</td>
<td>14.0</td>
<td>12.7</td>
</tr>
<tr>
<td>Individuals living in eligible household</td>
<td>4.5</td>
<td>7.9</td>
<td>9.0</td>
<td>7.8</td>
</tr>
</tbody>
</table>
Sustainable development related policies:

- 60% of the population depend on natural resources-related sectors such as agriculture, forestry, fisheries and tourism.
- The excessive and increasing exploitation of natural resources in a context of environmental degradation have created new constraints on economic growth and on the prospects of job creation.
- Proper management of biodiversity and increased awareness of its vitality can generate significant economic benefits and contribute to poverty reduction.
- Following are the strategic objectives:
  1. Mitigate effects of climate change on ecosystem
  2. Strengthen capacities in environmental and natural resource management
  3. Promote green economy
  4. Make rural ecosystems less vulnerable to the effects of climate change

Table 21. Decline in total forest area

<table>
<thead>
<tr>
<th>Forest Year</th>
<th>Total Forest Area (ha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>9,348,000</td>
</tr>
<tr>
<td>2000</td>
<td>8,898,000</td>
</tr>
<tr>
<td>2005</td>
<td>8,673,000</td>
</tr>
<tr>
<td>Annual Change 1990-2000 (ha</td>
<td>%)</td>
</tr>
<tr>
<td>Annual Change 2000-2005 (ha</td>
<td>%)</td>
</tr>
<tr>
<td>Total Change 1990-2005 (ha</td>
<td>%)</td>
</tr>
<tr>
<td>Change in rate (%)</td>
<td>5.06%</td>
</tr>
</tbody>
</table>

Figure 22. Forest area in Senegal

The steady decrease in forestland area can be attributed to both natural factors and man-made ones like deforestation and new projects that require land for building their development centres and factories.
Table 2. Environment statistics

<table>
<thead>
<tr>
<th>Threatened species</th>
<th>2013</th>
<th>103</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forested area (% of land area)</td>
<td>2011</td>
<td>43.8</td>
</tr>
<tr>
<td>CO2 emission estimates (000 metric tons and metric tons per capita)</td>
<td>2010</td>
<td>7053/0.6</td>
</tr>
<tr>
<td>Energy consumption per capita (kg oil equivalent)</td>
<td>2010</td>
<td>141.0</td>
</tr>
<tr>
<td>Precipitation in the capital city, total mean (mm)</td>
<td></td>
<td>514</td>
</tr>
<tr>
<td>Temperature in the capital city, mean °C (minimum and maximum)</td>
<td></td>
<td>21.7/27.6</td>
</tr>
</tbody>
</table>

Though there has been a grain of seriousness with which various troubling issues such as employment, environmental protection and gender disparity are addressed by governmental and national policies, it can be observed that the effects of them are not very satisfactory and not meeting the expectations.

- Employment, despite garnering serious attention, refuses to show much improvement and still unemployment is seen increasing
- Though more women participation in Governmental offices are observed, overall there is still the concerns of gender equality among the other organizations
- Social protection policies too aren’t much beneficial for Senegal households
- The steady and alarming rate with which forests are disappearing contradicts the various policies enacted to prohibit them from happening

Although there are objectives that have to be met, there has not been any significant effort to exercise them, especially in areas like crime and environment. The trends over the years suggest the same, showing no improvement, and in the worst case, even degradation. It is therefore necessary that relevant action be taken in order to meet the specified goals.

5 Discussion and Suggestions

We now discuss about the reasons for the presence of the various issues among the Senegalese communities. We also enlist the rectification measures that could curtail the spread of the issues with respect to the inferences from the above hypotheses and also suggest a few much sought after features that could have a tremendous impact.

Table 23. Summary of results and recommendations

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Issues</th>
<th>Rectification Measures</th>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender disparity in employment is a serious concern for development</td>
<td>Women of many ethnicity remain orthodox, which leads to no contribution to the workforce</td>
<td>Women empowerment can be effected with the help of educating the womanhood among the regions inhabited by orthodox ethnic groups</td>
<td>Setting up of educational institutions or schools to educate the women of lesser groups</td>
</tr>
<tr>
<td>Urbanization affects the growth of agricultural sector</td>
<td>Better employment in urban cities with little support for agriculture</td>
<td>Projects and policies to better support agriculture and avoid the effects of severe climatic conditions</td>
<td>Instructing the rural households about the significance of agriculture and persuading them to stay</td>
</tr>
<tr>
<td>Infant and child</td>
<td>Less number of hospitals</td>
<td>Setting up and providing</td>
<td>Inspiring and motivating</td>
</tr>
</tbody>
</table>
mortality is high in rural areas compared to urban regions due to social disparity

Social disparity is directly connected with lack of quality education and teachers

Disease spread and socioeconomic disparity are highly connected

Some ethnic groups, in spite of the location, involve in same profession

Government policies do not directly address employment, gender, social protection and sustainable development

and healthcare facilities countryside

Disparities due to provision for higher education for cities and presence of apprentices in villages

Sophisticated hospitals in cities to almost not even first-aid facilities in villages

Some professions aren’t much supported in certain regions leading to disparities in economy

Insignificant efforts to achieve the objectives in various categories

healthcare facilities on an equal basis among rural regions

Setting up of education facilities with competent teachers for professional and technical studies in rural areas

Gradual establishment of preliminary health facilities with growing sophistications over the years

Avoiding the migration of rural communities towards urban cities and instructing about their importance for the productivity and ingenious commodities

Supporting potent personnel at various cadres governmental bodies to help in bypassing or meeting the expectations and projections

the youth to take on nursing and doctoral careers supported by scholarships

Introducing scholarships and other opportunities motivating youth in urban areas to aid their countrymen

Spreading awareness among the low-level people about diseases and instructing counter measures

Training of unfortunate immigrants to hone their skills according to the prevalent profession

Clearly laid out policies to better the quality of employment, crime and other departments by skilled bureaucrats

6 Conclusion

The work presented in this paper emphasizes the need for identifying the socio-economic disparities of a nation, in order to accelerate the upliftment of its status on multiple fronts. It is therefore imperative to understand the subtle differences in the characteristics of people in order to provide them with adequate facilities for the improvement of their socio-economic status. This paper is just a starting step in such an analysis and a lot of other hypotheses can be regarded and proven to get a better understanding of the communities residing in a nation. This information, if well utilized, can guarantee expected growth and achievement of development goals of a nation and its resident communities.

References

3. Yan Ji. Understanding Human Mobility Patterns Through Mobile Phone Records: A cross-cultural study (June 2011)
5. The World Factbook, Central Intelligence Agency
7. Senegal, World Development Indicators, World Bank
10. Diallo Ngagne. A Diachronic Study on the Urban Growth of Dakar
12. Senegal Demographic and Health Survey - Multiple Indicator Cluster Survey (2010-11)
14. UNICEF. Understanding youth employment outcomes in Senegal
15. Education In Senegal, Wikipedia  
   http://en.wikipedia.org/wiki/Education_in_Senegal
20. Senegal, Worldmark Encyclopedia of Nations
Detection of Population Mobility Anomalies in Senegal from Base Station Profiles

Foued Melakessou, Thierry Derrmann, Raphael Frank, German Castignani, and Thomas Engel,
SnT, University of Luxembourg, Luxembourg,
Email: {foued.melakessou, thierry.derrmann, raphael.frank, german.castignani, thomas.engel}@uni.lu

Abstract—The analysis of Call Detail Records has captured the attention of traffic and transportation researchers to optimize people’s mobility. In our work, we analyze Call Detail Records in order to extract realistic human mobility models adapted to the Senegal use case. In this paper, we describe our analysis of the available D4D datasets. The first contribution is the modeling of the daily traffic demand profile of each antenna, by considering voice and messaging activities. The evaluation of mobility models will help to better design and develop future infrastructures in order to better support the actual demand. A classification has been performed into urban, suburban and rural modes. An algorithm has been developed to detect traffic anomalies in 2013, based on the daily profiles. The second contribution corresponds to the generation of inter-antenna and inter-district mobility graphs for each month of 2013. The rest of this paper is organized as follows. In Section I, we present the problem statement. Then, in Section II, we overview general information of the Senegal use case. We present demographics, economy and statistics that enable a better understanding of cause of population mobility in Senegal. Next in Section III we describe the datasets provided for the D4D challenge by Sonatel/Orange. In Section IV we present the pre-processing operations that have been performed in order to optimize computation time and memory space, our model analysis and traffic characterization. In Section V we describe the daily traffic (voice and text) profile for each antenna. In fact each model characterizes the normal traffic behavior supported in each antenna on a daily basis. Thus, in Section VI we propose an algorithm to detect traffic anomalies from the computed profiles. Followed in Section VII by the analysis of user mobility in Senegal according to the different datasets. Finally, Section VIII provides related works and Section IX concludes the paper.

Keywords: D4D Challenge, Big Data, CDR, Mobility, Graph

I. INTRODUCTION

Collecting, analyzing and modeling the distribution of the mobility flows is an important factor that enables the design or modification of the current transportation infrastructure (bus routes, train schedules, river transport, ferries routes, etc.) in order to efficiently support the user demand. We plan to evaluate the impact of the transportation infrastructure modifications according to simulation studies. To this end, we suggest to use Call Detail Record (CDR) datasets, provided by Orange Senegal, combined with demographic data and maps (base stations geolocalization), in order to compute, characterize and identify the mobility flows of Senegalese population. In this work, we propose to build a multimodal demand model (origin-destination matrix) for the Senegal use case, based on the MATSim simulation framework. This facilitates the impact evaluation of infrastructure modifications, e.g. addition of new roads and railway tracks as well as optimum reachability placements for new public facilities such as hospitals, defined by several simulation scenarios on the efficiency of the transportation services. Mobility flows can be generated between Points of Interest such as commercial facilities (shopping, petrol stations, sport centers, etc.), public facilities (hospitals, schools, place of worship, etc.) and tourist attractions. Our goal is to analyze historical CDRs data in order to model behaviors of traffic volume exchange between antennas. Our analysis is based on CDRs extracted from a set of 1666 antennas distributed all over Senegal. In fact, each mobile operator manages a population of customers that are geographically distributed all around the country. We propose to build mobility maps from simple statistical analysis of CDRs, provided by Sonatel/Orange. CDRs can be used to retrieve mobility patterns of the population under study. The precise location and mobility patterns of each customer is an unknown parameter of our problem. The only way to compute this information relies on the use of discrete aggregated CDRs. As soon as a user started a mobile phone activity, a log entry is added into the global CDR and provides information such as the timestamp when the activity began, the base station where the caller and callee are connected and the duration of the call. In our contribution, we propose to analyze the behavior of mobile phone activities. We show how mobility patterns can be highlighted. This work explains the transportation dynamics in Senegal. As goods and people transportation are highly correlated with economic activities, we present in the next section, demographics and general statistics that enable a better understanding of the characteristics of population mobility in Senegal.

Our team is currently working on a research project named MAMBA that intends to propose and validate a multimodal mobility platform that relies on new Internet technologies to interconnect different mobile services to provide relevant travel advice based on the context of the users, so as to optimize overall system performance. Taking into account real time traffic conditions, the status of existing public transport services (e.g., buses, trains) and user preferences, a personalized travel assistant will proactively suggest the best transportation possibility to reach a desired destination while also balancing the load over the different transportation modes in the multimodal system. It is planned to apply the methodology presented in this paper to the Luxembourg use case as defined in the MAMBA project.

1http://mamba-project.lu
II. SENEGAL: DEMOGRAPHICS AND ECONOMY

Senegal is a country located on the west part of the African continent, between latitudes 12.35° and 16.65°N, and longitudes 11.37° and 17.52°W. Rural population represents 55% of the total population. Senegal is composed of 14 regions and 123 districts [1]. We present in TABLE I. the general statistics of all Senegal regions in 2013.

The population reaches 13.5M citizens and its distribution is mainly concentrated in the west part of the country, in regions such as Dakar, Thies, Djourbel, Kaolack, Loubâ and Saint-Louis. The highest population density is located in the area of Dakar, the capital of Senegal. The second most populated city is Touba, located in the region of Djourbel. The fertility rate ranged 5 to 5.3 between 2005 and 2013, with 4.1 in urban areas and 6.3 in rural areas. Illiteracy is high, particularly among women. Life expectancy is estimated to 57.5 years. According to [2], the distribution of the population (Age) is:

- 0-14 years: 42.5%
- 15-24 years: 20.5%
- 25-54 years: 30.4%
- 55-64 years: 3.8%
- 65 years and over: 2.9%

The median age is 18.4 years. Approximately 60% of the population are less than 24 years old. This factor needs to be taken into account in order to build realistic models that will generate mobility behaviors according to the users’ activities. People’s mobility can be explained by many reasons. We propose to distinguish at least two classes, i.e., professional trips over the country and personal trips. The professional trips are correlated with the national economy, mainly based on food processing, mining, cement, artificial fertilizer, chemicals, textiles, refining imported petroleum and tourism. The economy is focused on Europe (mainly France and Italy), and India [3]. Senegal faces socioeconomic disparities, as 5% of the richest population has 47% of the revenues and 80% of poorest people have less than 28% of the revenues. Unemployment reached 48% and 40% among young city-dwellers in 2010. Also, 45.7% of the population lives under the national poverty threshold. Primary sector employs 70% of the active population, which includes:

- Agriculture (cereals such as mil, sorgho, rice in Casamance/Kolda, peanuts, sugar, fruit and vegetables such as tomato and greens in Niayes, Cotton in Kâhonne, Tambacounda, Casamance and Kedougou)
- Fishing
- Phosphate (Calcium Phosphate in Taiba, Aluminium Phosphate in Thies)
- Gold in oriental Senegal (Faléme river)

Informal services (fishing, business, and handicraft), remain with the State, the first employer and revenue resource of citizens.

The mobility of citizens is facilitated with a road network that interconnects the different regions of Senegal. There are mainly 7 primary roads [4]:

- N1: Diourbel Kaolack Tambacounda Kidira (Mali)
- N2: Kaolack This St-Louis Richard Toll Ouro Soguí Kidira
- N3: This Diourbel Toubâ Lingure Ouro Soguí
- N4: Kaolack (Trans-Gambia Highway) Bignona Ziguinchor (Guinea-Bissau)
- N5: Bignona Diouloulou (Gambia)
- N6: Tambacounda Kolda Ziguinchor
- N7: Tambacounda Dar-Salâm Niokolo-Koba Kedougou Sgou (Guinea)

In large cities, the transport network is well developed with taxis, taxi-brousse (Peugeot 505, 7 places), bus and "speed collective cars" (yellow and blue van, 20 places), "N’Diaye”, white van (Mercedes, 30 places). The majority of vehicles are trucks and people are often doing hitch-hiking accordingly. In August 2013, a new toll highway between Dakar and Diamniadio has been inaugurated. CDRs have been used in Section VI in order to detect the impact of this event on the behavior of citizens. There was a unique train line connecting Dakar (west Senegal)-Kidira (East Senegal)-Bamako (Mali) (1230km in 72h, Transrail) offering a single round-trip link per week, with a maximum speed of 65 km/h. Unfortunately, this line has been stopped in May 2009 and only merchandise trains circulates now. PTB “Petit train de banlieue” Little Blue Train (respectively Autorail Dakar-Thies) also connects Rufisque, Mbao, Thiaroye, Hann and Dakar (respectively Colobane, Hann, Thiaroye, Rufisque, Bargny, Sebkhotane and Pout).

In the following section, we describe the datasets provided for the D4D challenge.

III. D4D DATASETS

Sonatel/Orange (Société nationale de télécommunications) is the main mobile telecommunication operator (55.52% of the population) in Senegal, followed by Tigo (24.23%) and Expresso (20.25%) [5]. Anonymous call patterns of Orange’s mobile phone users in Senegal have been released for the 2015 D4D challenge [6]. We present in this section the available datasets, based on CDRs of phone calls and text exchanges between 9 million of Orange customers during 2013.

A. Context Data

The file SITE_ARR_LATLON.CSV provides the coordinates of all antennas:

- Site_id is the antenna ID,
- Reg_id corresponds to the district ID where the antenna is located,
- Lon is the antenna longitude,
- Lat provides the antenna latitude.

An extract of this file is presented below:

<table>
<thead>
<tr>
<th>Site_id</th>
<th>Arr_id</th>
<th>Lon</th>
<th>Lat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,-17.525142,14.746832</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The geographic border information of each district has been aggregated traffic at the district and region levels. The second and fifth columns have been used to evaluate the antenna density per region has been computed in TABLE II. Dakar, Thies and Djourbel present the highest values. They are in fact the more urbanized region of Senegal.

Our visualization tool is based on the Scilab software [7] and the NARVAL toolbox [8]. As a consequence, a matrix format for the data has been used. We have generated the file ant_gps_senegal_region.data where we added, for each antenna, in the last column, the region to which the base station belongs. For instance, here are its first 5 entries:

<table>
<thead>
<tr>
<th>Antenna</th>
<th>Region Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,2,-17.524360,14.747434</td>
</tr>
<tr>
<td>2</td>
<td>3,2,-17.522576,14.745198</td>
</tr>
<tr>
<td>3</td>
<td>4,2,-17.516398,14.746730</td>
</tr>
<tr>
<td>4</td>
<td>5,2,-17.512870,14.740658</td>
</tr>
</tbody>
</table>

The second and fifth columns have been used to evaluate the aggregated traffic at the district and region levels.

The file Shapefile_senegal.zip contains the shapefile document that enables the visualization of Senegal districts. The geographic border information of each district has been extracted from the file senegal_arr_2014_wgs.shp according to the free and Open Source Geographic Information System (QGIS) [9] (geojson), and a Perl script into the file senegal_arr_2014_wgs.data. The country boundary and road network have been extracted from the OpenStreetMap database [10] and the software Osmosis.

The antenna density per region has been computed in TABLE II. Dakar, Thies and Djourbel present the highest values. They are in fact the more urbanized region of Senegal.

![context_data_visualization.png](attachment:context_data_visualization.png)

Fig. 1. Context data visualization.

**TABLE I. General statistics of Senegal in 2013 (Agence Nationale de la Statistique et de la Demographie, ANSD).**

<table>
<thead>
<tr>
<th>Region</th>
<th>Area (km²)</th>
<th>Population</th>
<th>% Poverty Index</th>
<th>% Literacy</th>
<th>% Education</th>
<th>% child registered (civil status)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakar</td>
<td>547</td>
<td>3,137,196</td>
<td>26.1</td>
<td>68.6</td>
<td>64.3</td>
<td>91.80</td>
</tr>
<tr>
<td>Thies</td>
<td>6670</td>
<td>17,888,64</td>
<td>41.3</td>
<td>53.8</td>
<td>53.4</td>
<td>86.4</td>
</tr>
<tr>
<td>Djourbel</td>
<td>4824</td>
<td>14,974,55</td>
<td>47.8</td>
<td>35.1</td>
<td>28.9</td>
<td>69.4</td>
</tr>
<tr>
<td>Fatick</td>
<td>6849</td>
<td>71,439,2</td>
<td>67.8</td>
<td>45.8</td>
<td>61.1</td>
<td>79.5</td>
</tr>
<tr>
<td>Longa</td>
<td>24,889</td>
<td>87,419,3</td>
<td>26.8</td>
<td>36.4</td>
<td>30.3</td>
<td>65.8</td>
</tr>
<tr>
<td>Kaolack</td>
<td>5,357</td>
<td>9,608,75</td>
<td>61.7</td>
<td>50.2</td>
<td>46.4</td>
<td>72.5</td>
</tr>
<tr>
<td>Kaffrine</td>
<td>11,262</td>
<td>5,669,92</td>
<td>63.8</td>
<td>42.0</td>
<td>26.7</td>
<td>64.1</td>
</tr>
<tr>
<td>Saint-Louis</td>
<td>19,241</td>
<td>9,089,42</td>
<td>39.7</td>
<td>53.2</td>
<td>57.4</td>
<td>71.5</td>
</tr>
<tr>
<td>Kolda</td>
<td>13,771</td>
<td>66,245,5</td>
<td>76.6</td>
<td>43.7</td>
<td>51.9</td>
<td>56.6</td>
</tr>
<tr>
<td>Sedhiou</td>
<td>7341</td>
<td>4,529,94</td>
<td>68.3</td>
<td>47.3</td>
<td>62.7</td>
<td>56.5</td>
</tr>
<tr>
<td>Ziguinchor</td>
<td>7352</td>
<td>5,491,51</td>
<td>66.8</td>
<td>65.0</td>
<td>87.3</td>
<td>82.2</td>
</tr>
<tr>
<td>Kedougou</td>
<td>16,000</td>
<td>15,135,7</td>
<td>71.3</td>
<td>35.0</td>
<td>63.2</td>
<td>75.8</td>
</tr>
<tr>
<td>Tambacounda</td>
<td>42,364</td>
<td>68,131,0</td>
<td>62.5</td>
<td>35.0</td>
<td>44.4</td>
<td>55.2</td>
</tr>
<tr>
<td>Matam</td>
<td>29,445</td>
<td>56,253,9</td>
<td>45.2</td>
<td>28.4</td>
<td>42.1</td>
<td>67.8</td>
</tr>
<tr>
<td>Senegal</td>
<td>196,712</td>
<td>13,087,15</td>
<td>54.7</td>
<td>45.7</td>
<td>51.4</td>
<td>71.1</td>
</tr>
</tbody>
</table>

The first dataset is composed by 24 traces:
<table>
<thead>
<tr>
<th>ID</th>
<th>Region</th>
<th>#Antennas</th>
<th>Antenna density (#/km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dakar</td>
<td>489</td>
<td>0.893</td>
</tr>
<tr>
<td>2</td>
<td>Thies</td>
<td>158</td>
<td>0.024</td>
</tr>
<tr>
<td>3</td>
<td>Djourbel</td>
<td>120</td>
<td>0.025</td>
</tr>
<tr>
<td>4</td>
<td>Fatick</td>
<td>68</td>
<td>0.010</td>
</tr>
<tr>
<td>5</td>
<td>Louga</td>
<td>94</td>
<td>0.004</td>
</tr>
<tr>
<td>6</td>
<td>Kao lucrack</td>
<td>76</td>
<td>0.014</td>
</tr>
<tr>
<td>7</td>
<td>Kaffrine</td>
<td>64</td>
<td>0.006</td>
</tr>
<tr>
<td>8</td>
<td>Saint-Louis</td>
<td>125</td>
<td>0.006</td>
</tr>
<tr>
<td>9</td>
<td>Kolda</td>
<td>73</td>
<td>0.005</td>
</tr>
<tr>
<td>10</td>
<td>Sedhiou</td>
<td>61</td>
<td>0.008</td>
</tr>
<tr>
<td>11</td>
<td>Ziguinchor</td>
<td>78</td>
<td>0.011</td>
</tr>
<tr>
<td>12</td>
<td>Kedougou</td>
<td>59</td>
<td>0.004</td>
</tr>
<tr>
<td>13</td>
<td>Tambacounda</td>
<td>129</td>
<td>0.003</td>
</tr>
<tr>
<td>14</td>
<td>Matam</td>
<td>72</td>
<td>0.002</td>
</tr>
</tbody>
</table>

**TABLE II. ANTELLA DISTRIBUTION.**

- SET\_i.S\_i.CSV with $i \in [01:12]$ provide the text traffic for each month of the year (total size: compressed 2.2GB, uncompressed 15.0GB).
- SET\_i.V\_i.CSV with $i \in [01:12]$ provide the voice traffic for each month of the year (total size: compressed 7.2GB, uncompressed 38.8GB).

SET 1 provides antenna-to-antenna traffic for 1666 antennas on an hourly basis.

For the voice service, each trace line follows the format $T, I_S, I_D, N, D$:
- $T$ is the 1H timestamp where calls are collected (YYYY-MM-DD HH),
- $I_S$ corresponds to the ID of the antenna where the callers are connected,
- $I_D$ gives the ID of the antenna where the callees are connected,
- $N$ provides the total numbers of calls between $I_S$ and $I_D$ in $T$.
- $D$ is the total duration of all calls between $I_S$ and $I_D$ in $T$.

For instance, here are the 5 first entries of SET\_i.V\_01.CSV:

<table>
<thead>
<tr>
<th>Year-Month-Day</th>
<th>Hour-Minute-Seconds</th>
<th>CallsID1</th>
<th>CallsID2</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-01-01</td>
<td>00,1,1,1,54</td>
<td>1</td>
<td>1</td>
<td>1,54</td>
</tr>
<tr>
<td>2013-01-01</td>
<td>00,1,2,1,39</td>
<td>1</td>
<td>2</td>
<td>1,39</td>
</tr>
<tr>
<td>2013-01-01</td>
<td>00,1,24,1,2957</td>
<td>2</td>
<td>24</td>
<td>1,2957</td>
</tr>
<tr>
<td>2013-01-01</td>
<td>00,1,186,1,56</td>
<td>1</td>
<td>186</td>
<td>1,56</td>
</tr>
<tr>
<td>2013-01-01</td>
<td>00,2,2,22,418</td>
<td>2</td>
<td>22</td>
<td>2,418</td>
</tr>
</tbody>
</table>

Thus, in respect with the fifth line, on the first January 2013, 22 calls have been started between 00AM and 01AM, 1 of the callers connected to the antenna 1, 24 to the antenna 1, 186 to the antenna 1, 2 to the antenna 2, and their callees connected to the antenna 418.

- $I_D$ gives the ID of the antenna where the callees are connected,
- $N$ provides the total numbers of messages between $I_S$ and $I_D$ in $T$.

For instance, the 5 first entries of SET\_i.S\_01.CSV are listed below:

<table>
<thead>
<tr>
<th>Year-Month-Day</th>
<th>Hour-Minute-Seconds</th>
<th>MessagesID1</th>
<th>MessagesID2</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-01-01</td>
<td>00,1,61,1</td>
<td>1</td>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>2013-01-01</td>
<td>00,1,340,1</td>
<td>1</td>
<td>340</td>
<td>1</td>
</tr>
<tr>
<td>2013-01-01</td>
<td>00,1,419,1</td>
<td>1</td>
<td>419</td>
<td>1</td>
</tr>
<tr>
<td>2013-01-01</td>
<td>00,1,420,1</td>
<td>1</td>
<td>420</td>
<td>1</td>
</tr>
<tr>
<td>2013-01-01</td>
<td>00,1,447,2</td>
<td>1</td>
<td>447</td>
<td>2</td>
</tr>
</tbody>
</table>

Thus, in respect with the fifth line, on the first January 2013, 2 messages have been sent between 00AM and 01AM, for callers connected to the antenna 1, and their callees connected to the antenna 447.

**C. Dataset: SET 2**

The second dataset is composed by 50 traces. It provides fine-grained mobility data on a rolling 2-week basis for a year, which includes roughly 300,000 randomly sampled users:

- SET2\_Pi.CSV with $i \in [01:25]$ (total size: compressed 3.3GB, uncompressed 37.0GB),
- INDICATORS\_SET2\_Pi.CSV with $i \in [01:25]$ (total size: compressed 524.2MB, uncompressed 1.4GB).

For a voice traffic trace, each line follows the format $I, T, A$:
- $I$ is the user ID,
- $T$ provides the timestamp when the user started his call (YYYY-MM-DD HH:MM:SS), the time resolution is in consequence 10 min,
- $A$ corresponds to the ID of the antenna where the caller is connected.

For instance, the 5 first entries of SET\_2P\_01.CSV are listed below:

<table>
<thead>
<tr>
<th>Year-Month-Day</th>
<th>Hour-Minute-Seconds</th>
<th>CallsID1</th>
<th>CallsID2</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-01-07</td>
<td>13:10:00,461</td>
<td>1</td>
<td>2</td>
<td>1,461</td>
</tr>
<tr>
<td>2013-01-07</td>
<td>17:20:00,454</td>
<td>1</td>
<td>2</td>
<td>1,454</td>
</tr>
<tr>
<td>2013-01-07</td>
<td>17:30:00,454</td>
<td>1</td>
<td>2</td>
<td>1,454</td>
</tr>
</tbody>
</table>
This file provides the trajectories of about 300,000 users during the first 2 weeks of January. In that case, we can observe the trajectory of the user 1 who was connected to the antenna 461 at 13h10, and thereafter moved to the antenna 454, started a call at 17h20, and another one at 17h30. Thus for each user, the real trajectory can be extracted for each 2 weeks period.

Behavioral indicators computed by the bandicoot toolbox [11] have also been provided for each user.

**D. Dataset: SET 3**

The third dataset is composed by 24 traces. It provides one year of coarse-grained mobility data, month by month, at district level for about 150,000 randomly sampled users:

- **SET3_Mi.CSV** with $i \in [01:12]$ (total size: compressed 2.0GB, uncompressed 16.2GB),
- **INDICATORS_SET3_Mi.CSV** with $i \in [01:12]$ (total size: compressed 124.3MB, uncompressed 312.5MB).

For a traffic trace, each line follows the format $I, T, A_r$:

- $I$ is the user ID,
- $T$ provides the timestamp when the user started his call (YYYY-MM-DD HH:MM:SS), the time resolution is in consequence 10 min,
- $A_r$ corresponds to the ID of the district where the caller was located.

For instance, here are the 5 first entries of **SET_3_M01.CSV**:

<table>
<thead>
<tr>
<th>ID</th>
<th>Date</th>
<th>Time (HH:MM:SS)</th>
<th>District ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>37509</td>
<td>2013-01-29</td>
<td>15:00:00</td>
<td>3</td>
</tr>
<tr>
<td>84009</td>
<td>2013-01-14</td>
<td>07:00:00</td>
<td>3</td>
</tr>
<tr>
<td>84009</td>
<td>2013-01-14</td>
<td>07:00:00</td>
<td>3</td>
</tr>
<tr>
<td>84009</td>
<td>2013-01-14</td>
<td>07:00:00</td>
<td>3</td>
</tr>
<tr>
<td>80150</td>
<td>2013-01-27</td>
<td>16:50:00</td>
<td>3</td>
</tr>
</tbody>
</table>

Thus, the user 37509 was located on the 2013-01-29 at 15:00:00 in the district 3, namely Grand Dakar.

In order to optimize our mobility analysis and traffic characterization in terms of time computation and memory space, we have performed few pre-processing operations on the original raw datasets.

**IV. DATA PRE-PROCESSING**

Entries in each file of the dataset SET 1 have been chronologically sorted. Thus we have generated new files t_SET1V_i.data (Voice) and t_SET1S_i.data (Text) with $i \in [01:12]$ (total size: 400KB). They provide for each month the index within each file where the traffic for a certain period of time, can be directly extracted. We have used Perl regular expressions and Linux SED commands [12]. This operation allows avoiding parsing each dataset. In consequence, we drastically reduce the computation time of extraction and search operations. In fact, a simple search can become extremely long, as we are processing very large datasets. As an example, here are the 5 first entries of t_SET1V_01.data (Voice traffic in January).

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Caller ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-01-01</td>
<td>00 01</td>
<td>1</td>
</tr>
<tr>
<td>2013-01-01</td>
<td>01 215630</td>
<td></td>
</tr>
<tr>
<td>2013-01-01</td>
<td>02 327611</td>
<td></td>
</tr>
<tr>
<td>2013-01-01</td>
<td>03 389658</td>
<td></td>
</tr>
<tr>
<td>2013-01-01</td>
<td>04 427875</td>
<td></td>
</tr>
</tbody>
</table>

In this case, all the voice traffic of the date 2013-01-01 for the timeslot 00 (between 00h and 01h) is found between the lines $l = 1$ and $l = 215630 - 1$ of the file SET1V_01.data. In the same way, all the voice traffic of the timeslot 01 (between 01h and 02h) is located between the lines $l = 327611$ and $l = 327611 - 1$. For instance, here are the 5 first entries of t_SET1S_01.data (Text traffic in January).

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Caller ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-01-01</td>
<td>00 1</td>
<td></td>
</tr>
<tr>
<td>2013-01-01</td>
<td>01 396521</td>
<td></td>
</tr>
<tr>
<td>2013-01-01</td>
<td>02 601149</td>
<td></td>
</tr>
<tr>
<td>2013-01-01</td>
<td>03 697731</td>
<td></td>
</tr>
<tr>
<td>2013-01-01</td>
<td>04 754552</td>
<td></td>
</tr>
</tbody>
</table>

In this case, all the text traffic of the date 2013-01-01 for the timeslot 00 (between 00h and 01h) is found between the lines $l = 1$ and $l = 396521 - 1$ of the file SET1S_01.CSV. In the same way, all the voice traffic of the timeslot 01 (between 01h and 02h) is located between the lines $l = 601149$ and $l = 601149 - 1$. We have also generated the files MAX_SET1V_i.data (Voice) and MAX_SET1S_i.data (Text) with $i \in [01:12]$, that provide the maximum number of calls, the maximum duration of calls and the maximum number of text messages, for each antenna, and for each month (total size: 400KB). These values are used during the visualization process. For instance, here are the 5 first entries of MAX_SET1V_01.data (Voice traffic in January).

<table>
<thead>
<tr>
<th>Time</th>
<th>Caller ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>00 1</td>
<td>1</td>
</tr>
<tr>
<td>01 396521</td>
<td></td>
</tr>
<tr>
<td>02 601149</td>
<td></td>
</tr>
<tr>
<td>03 697731</td>
<td></td>
</tr>
<tr>
<td>04 754552</td>
<td></td>
</tr>
</tbody>
</table>

In this case, in January (month 01), the maximum number of calls generated by callers connected in the antenna 1 during any 1h timeslot rates 19. The maximum duration of all calls inside any 1h timeslot of all callers connected to the antenna 1 is 3600 min. The relation between the quantity and the duration of calls is unfortunately not known. In fact, less callers can initiate longer calls. For instance, here are the 5 first entries of MAX_SET1S_01.data (Text traffic in January).

<table>
<thead>
<tr>
<th>Time</th>
<th>Caller ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>00 1</td>
<td>1</td>
</tr>
<tr>
<td>01 396521</td>
<td></td>
</tr>
<tr>
<td>02 601149</td>
<td></td>
</tr>
<tr>
<td>03 697731</td>
<td></td>
</tr>
<tr>
<td>04 754552</td>
<td></td>
</tr>
</tbody>
</table>

In this case, in January (month 01), the maximum number of calls generated by callers connected in the antenna 1 during any 1h timeslot rates 19. The maximum duration of all calls inside any 1h timeslot of all callers connected to the antenna 1 is 3600 min. The relation between the quantity and the duration of calls is unfortunately not known. In fact, less callers can initiate longer calls. For instance, here are the 5 first entries of MAX_SET1S_01.data (Text traffic in January).

<table>
<thead>
<tr>
<th>Time</th>
<th>Caller ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>00 1</td>
<td>1</td>
</tr>
<tr>
<td>01 396521</td>
<td></td>
</tr>
<tr>
<td>02 601149</td>
<td></td>
</tr>
<tr>
<td>03 697731</td>
<td></td>
</tr>
<tr>
<td>04 754552</td>
<td></td>
</tr>
</tbody>
</table>
In this case, in January (month 01), the maximum number of text messages generated by any caller connected to the antenna 1 during any 1h timeslot rates 65. We also generated the following files that give additional information about the quantity of calls initiated by callers (total size: 70MB):

- 2013_V_NS_year_[mean/std].data: mean/standard deviation of the number of calls emitted by callers located in each antenna (level of aggregation=1 year), 1666 lines * 1 column,
- 2013_V_NS_month_[mean/std].data: mean/standard deviation of the number of calls emitted by callers located in each antenna (level of aggregation=1 month), 1666 lines * 12 columns,
- 2013_V_NS_week_[mean/std].data: mean/standard deviation of the number of calls emitted by callers located in each antenna (level of aggregation=1 week), 1666 lines * 52 columns,
- 2013_V_NS_day_[mean/std].data: mean/standard deviation of the number of calls emitted by callers located in each antenna (level of aggregation=1 day), 1666 lines * 365 columns,
- 2013_V_NS_hour.data: annual traffic for callers located in each antenna (level of aggregation=1h), 8760 (365*24) lines * 1666 columns,
- 2013_V_NS_stat.data: general statistics [minimum, first quartile, median, mean, third quartile, maximum, standard deviation], 1 line * 7 columns.

The same files have been generated for the number of calls received by callees (total size: 71MB):

- 2013_V_ND_year_[mean/std].data,
- 2013_V_ND_month_[mean/std].data,
- 2013_V_ND_week_[mean/std].data,
- 2013_V_ND_day_[mean/std].data,
- 2013_V_ND.data,
- 2013_V_ND_stat.data.

These files allow analyzing the geographical correlation between callers and callees (detection of short/long distance calls). The same files have also been generated for the duration of the calls for callers 2013_V_DS* (total size: 93MB) and callees 2013_V_DD* (total size: 94MB), and also for the number of messages for callers 2013_S_S* (total size: 66MB) and callees 2013_S_D* (total size: 66MB). These files have been used to study the correlation between each traffic variable (number of calls, duration of calls or number of message).

In the next section, we propose to compute the daily traffic profile on each antenna. This information characterizes the normal traffic behavior supported on each antenna on a daily basis. These profiles have been used to detect traffic anomalies and predict unusual behaviors that are caused by specific events such as cultural or sport events.

V. DAILY PROFILE MODEL

Human activities heavily impact the behavior of calls and messages emitted and received during each period of the day (00h to 24h). We propose in this section to model the daily traffic profile of each antenna. We used the files 2013_V_NS.data, 2013_V_ND.data, 2013_V_AD.data, 2013_S_S.data and 2013_S_D.data, already described in the previous section. The traffic of each antenna is composed by 8760 values (365*24). For instance, we have shown in Fig. 2, and Fig. 3, the 2013 year traffic of callers and callees respectively connected to the Antenna 2.

![Fig. 2. Antenna 2 traffic for callers.](image)

Each traffic variable $X$ (number of calls #Calls, duration of calls TCall, and number of messages #Text) presents a high variability along each day of the year. In order to compute the daily profile for each traffic variable $X$, we have computed average values for each timeslot, after an outliers removal process at $k_x = 2 \sigma$ (standard deviation). The value of $k_x$ has been empirically selected. It permits to take the largest set of valid points into consideration in order to generate the best average profile that is the closest to the normal traffic conditions. The generic profile of the antenna 271 has been depicted in Fig. 4, in order to explain our modeling. This antenna supports the highest traffic peak. Let $P^a_X = \{X_1, X_2, \ldots, X_{24}\}$ be the profile of the antenna $a$ for the variable $X$. $X_i$ with $i \in [1:24]$ provides the expected value of the traffic variable $X$ collected in the antenna $a$ during the $i^{th}$ one hour timeslot.

Let $E^n_a = \{E_1E_2\cdots E_{24}\}$ be the list of extrema of the antenna $a$. $E^n_a$ is computed according to a gradient algorithm that solves $\frac{dP^a}{dt} = 0$. It is composed by a succession of local
minimums and maximums. In order to classify the set of all antennas, we apply a k-means algorithm [13] in order to find a set of maximally disjoint clusters based on the following parameters:

- $T_{max_f}$ is the time when the first local maximum $N_{max_f}$ occurs (beginning of diurnal activities),
- $T_{max_l}$ corresponds to the time when the last local maximum $N_{max_l}$ occurs (end of diurnal activities),
- $M$ (respectively $S$) is the average traffic (respectively the standard deviation) of the diurnal activities,
- $N_e$ provides the number of extrema,
- $T_{min}$ gives the time when the global minimum $N_{min}$ occurs.

We apply the k-means function of NaN, a statistics and machine learning toolbox for Scilab focusing on data with and
without missing values encoded as NaN’s [14]. \textit{nangkmeans} has for input parameters $[T_{max}, T_{max}, S, M, N_{max}, N_{c}, T_{min}, N_{min}]$. $k$ is the number of classes. We tested the set of values $k \in [2:4]$. 1602 antennas provide significant results and have been used during the classification process. The remaining antennas have been defined as the class C0. TABLE III is a summary of the classification process. It provides for each variable $X$ (#Call, Tcall and #Text for caller and callee) and for each value $k$, the average traffic value of each class. The same behavior has been discovered for the traffic metrics Voice #Call, Voice TCall and Text #Text (for caller and callee) We observe that $k$ has a minor impact on the parameters $T_{max}$, $T_{max}$, $T_{min}$ and $N_{c}$ between each class. A significant impact can be highlighted on the parameters $S, M, N_{max}, N_{max}$ and $N_{min}$. Let us consider the case $k = 4$. For #Call, each class can be easily separated with (for $i > j$ with $i, j \in [1:4]$):

- its average traffic value $M_{Ci} > M_{Cj}$,
- its variability $S_{Ci} > S_{Cj}$.
Thus this classification permits to separate three distinct behaviors for antenna location:

- urban (class 1),
- suburban (class 2 and 3),
- rural (class 4).

Let us consider the case of a Caller. Urban sites have the profiles with the highest #Call, TCall and #Text. The first and largest peak of traffic occurs around 12AM. The diurnal activities start at 08AM in the morning. This corresponds to the well known behavior of traffic demand in any large cities, even if in the case of Senegal, these activities finishes at 11PM. We observe also a strange behavior for several antennas belonging to class 2 and 3, where their daily profiles present 3 high peaks at 10AM, 01PM and 03PM for #Call and TCall. The location of these anomalies have been displayed in Fig. 7. The position is located around Touba.

For text messages, we observe a high activity during the night (08PM to 02AM) with a peak at 11PM. A possible reason for this behavior is the average age of citizens in Senegal (less than 20 years). In fact, this mode of communication is the most used by young people, especially at night. We now propose to study the geographic distribution of each class of antennas. Fig. 8, provides the spatial distribution of all classes of antennas. These maps allow finding a high correlation between the class of each antenna and the demographics of Senegal (see Fig. 1). There are minor differences in the localization of classes resulting from the k-means classification based on each traffic metric (#Call, TCall and #Text) for caller and callee. This means that a single traffic metric such as #Call is sufficient during the classification. However, it can be interesting to see the minor differences, that are probably due to different user behaviors related to the mode of communication. Thus, antennas of class 1 are located in urban areas.
areas. As a consequence, they intrinsically support more traffic than antennas that belongs to class 4, corresponding to rural areas. We also observe that Class 1, 2 and 3 are in general close to the road network of Senegal. Fig. 9, provides the geographical distribution of the caller and callee activities (M and S parameters for #Call, TCall and #Text). The value of S on each antenna for caller (respectively callee) is displayed with a blue (respectively red) angle in the top (respectively bottom) direction. Its side length is proportional to its value. For instance, for the voice traffic, the maximum value reaches 3025 calls per hour. This visualization allows detecting the locality of calls, and in consequence the partial population connected to each antenna. Let us consider the antenna i in Fig. 10. The remaining antennas are represented by the node j ≠ i. The traffic $S_{ii}$ is considered as local if callers connected to the antenna i initiate calls towards callees that are also connected to the antenna i. The remaining traffic $S_{ij}$ corresponds to the case when the callees are connected to an antenna j different from i. The population of callers (voice) is defined as $MS = S_{ii} + S_{ij}$. This analysis is also valid for TCall and #Text. The population of callees rates $MD = S_{ii} + S_{ji}$ where $S_{ji}$ is the aggregated traffic emitted.
from $j$ towards $i$. We have $MS - MD = S_{ij} - S_{ji}$. If $MS - MD = 0$ then $S_{ij} = S_{ji}$. Thus if the two angles in an antenna have the same size $MS = MD$, we can estimate that there is no privileged direction between callers connected to the antenna $i$ and callees connected to the antenna $j$. In other words, the probability to initiate a call from the antenna $i$ in the direction $i \rightarrow j$ is the same than the one in the opposite direction ($j \rightarrow i$). Unfortunately, we can not separate the contribution of callers from other antennas that also start call to this antenna. If $MS - MD > 0$, then $S_{ij} > S_{ji}$. Thus for any random call between a caller and a callee that can be located in $i$ or $j$, we can conclude that it is more probable that the caller is located in $i$ and the callee in $j$. The direction $i \rightarrow j$ is more probable. If $MS - MD < 0$, then $S_{ij} < S_{ji}$. In that case, the direction $j \rightarrow i$ is more probable. In order to distinguish local call from distant call ($S_{ii} = (S_{ii} + S_{ij})$), we also need to extract $S_{ii}$ from $MS$. Due to time constraints, this analysis has not been treated in this paper, but it will be considered in future works. We observe that the level of number of calls emitted by callers located in each antenna, reaches approximately the same values that the number of calls received by callees in the same antenna. Thus we conclude that
there is no privileged direction. However, if we also consider the standard deviation (Std(#Call) in Fig. 9.), some locations present asymmetry between callers and callees. For instance, in Touba, the value for callers is much larger than the ones for callees. This implies that callers have much more activity there, and they generally call people that are located in a different antenna site. We conclude that the behavior is similar between #Call, TCall and #Text. For instance, an antenna that supports a high value of Mean(#Call) generally handles also a high value of Mean(TCall). There is in fact a high correlation between the traffic collected on each antenna and the local population located close-by (see Fig 1). In the next section, we propose to analyze spatio-temporal traffic anomalies from these computed profiles.

VI. Traffic Anomalies Detection

The modeling of daily profiles is the basis of detection of traffic anomalies. We propose in this section to detect unusual traffic from the dataset SET 1. For instance, we propose to analyze the traffic of callers connected to the antenna 1 and detect each day of the year that presents an anomaly. Fig. 11. provides the aggregated traffic (level of aggregation=1 day) of antenna 1 during the complete year 2013 for the variable #Call, TCall and #Text. For each day, the average and standard deviation of each variable have been computed. Fig. 12. is a dispersion plot of the previous results. Each black cross represents a single day of traffic (365 elements). Each blue marker corresponds to a single week of traffic (52 elements). Each green disc is a single month of traffic (12 elements). We have plotted for each level of time aggregation (day, week and month) two discs:

- straight line disc centered in the average mean value $\text{Mean}$ and average standard deviation $\text{Std}$ of all
elements, with a radius of $\sigma$ (standard deviation for all elements),

- dashed line disc centered in $[\text{Mean}, \text{Std}]$, with a radius of $2\sigma$.

If an element belongs to a disc, it presents a normal traffic. Else it is considered as an anomaly. We have represented in Fig. 11. the time position of all anomalies ($\sigma$ in green and $2\sigma$ in red) detected with the previous scheme. The set of anomalies at $\sigma$ are of course included inside the set of anomalies at $2\sigma$. We obtained a high correlation between traffic anomalies for the different variables (#Call, TCall and #Text). We have preformed the previous detection scheme for all antennas. Then, we have computed in Fig. 13. the quantity of antennas that present an anomaly for each day of the year. We assume that global/national anomalies are related to events such as holidays, etc. Here is the list of global anomalies for #Call:

$[1:2,23:24]/1$
$29/3$
$15/6$
$19/7$
$[7:10]/8$
$[2,12,14:26]/10$
$[13,28:30]/11$
$[1:2,5,7,13:14,16:25,27:31]/12$

Here is the list of global anomalies for TCall:

$[1,24]/1$
$[29]/3$
$[15:18]/6$
$[9]/7$
$[4,6:10]/8$
$[16:25]/10$
$[28:30]/11$
$[1:9,12:18,21:22,27:28]/12$

Here is the list of global anomalies for #Text:

$[1]/1$
$[14]/2$
$[21:22,24:29]/7$
$[3:4,7:9,14:31]/8$
$[1:8]/9$
$[16]/10$
$[31]/12$

We provided in TABLE IV. a non-exhaustive list of events that occurred in Senegal in 2013. These events can explain the global anomalies represented in Fig. 13. The majority of anomalies are correlated with religious events. In fact, Islam is the predominant religion in the country (94%). Then during religious events, we have directly observed a radical change in the majority of antennas. We can observe a radical change in the number of anomalies related to #Text during summer holidays. In fact, Ramadan occurred during this period. Besides, the population in Senegal is young. Then during this period, the majority of the population is on vacations. This also explains the traffic growth. Thus CDRs have been used to highlight specific events related to the cultural background of the country. We propose to apply our anomaly detection scheme in order to discover the largest movement of citizens in Senegal, during the Magal of Touba event. One of the largest religious orders in Senegal, the Murid organized each year a pilgrimage, located in Touba. Between the end of December

Fig. 13. Detection of global/national anomalies.
(15) and the beginning of January occurred the Magal of Touba (3 millions of participants). Fig. 14. represents the traffic of the antenna 1047 located in Touba. The largest peak for each traffic variable confirms this event. Further analysis is needed to understand unusual behaviors such as the growth #Call and TCall between mid-March and mid-June.

Another important event related to transportation occurred on the 1st August 2013 In fact, the new highway between Dakar and Diamniadio has been inaugurated. The list of antennas $L$ near the new highway is:


We have illustrated in Fig. 15. and Fig. 16., the impact of new highway on the traffic of two antennas 237 and 397 of $L$.

![Fig. 14. Detection of traffic anomalies in Touba.](image1)

![Fig. 15. Impact of the new highway on the traffic of the antenna 237.](image2)

![Fig. 16. Impact of the new highway on the traffic of the antenna 397.](image3)
the use of the new highway that has attracted new users. This growth has been verified for the majority of antennas of L in Fig. 17, where we plotted for each antenna the traffic growth after the opening of the new highway.

![Traffic growth graph](image)

Fig. 17. Traffic growth after the new highway opening.

The average growth is 15% for #Call, 23% for TCall and 29% for #Text for equally sized time slots. Unfortunately, it is not possible to know if local events attracting citizens have also participated to this growth. Further investigations need to be considered. We propose now to analyze the mobility of users according to the datasets SET 2 and SET 3.

VII. MOBILITY DEMAND

The fine-grained (SET 2) and coarse (SET 3) mobility data have been used to identify by region and month of the year, the mobility of users in Senegal. This information will help national authorities to prioritize the deployment of new infrastructure such as road building, public transportation, civil equipment or tourist attractions. This analysis enables the visualization of the primary flows of population. We have performed the maps of primary flows for each month of the year. We have developed a visualization tool [8] to display the primary flows of population in the Scilab environment. Results for SET 3 have been summarized in Fig. 18., Fig. 19., Fig. 20., Fig. 21.). Results for SET 2 have been represented in Fig. 22., Fig. 23., Fig. 24., Fig. 25.). We displayed in each figure the mobility graph, composed by all interconnections between antennas (respectively districts) where users moved during the duration of the traffic collection. We also displayed the congestion map \( C \) of the mobility graph. \( C \) provides, for each antenna \( a \), the number of users that have crossed \( a \) during the duration of the traffic measurement. Thus this metric can be used to detect the most used antennas during each month of 2013.

We observed that each congestion map generally presents high values in antennas near to urban areas. The mobility graph is highly correlated with the road network. In fact, movements of citizens in Senegal are mainly done via the roads. The majority of population flows are concentrated around the region of Dakar. We observed that there are many long links between far antennas. The reason of these links is the intrinsic characteristic of CDRs. There is a lack of information about intermediate locations where the users traveled, if the inter-arrival time between two mobile phone activities is too large.
VIII. RELATED WORK

In [15] Morales et al. showed the correlation between the main roads of Ivory Coast, and the mobility patterns they extracted from 2013 D4D mobile phone datasets. They showed that a large part of the trajectories follow the primary road network in the country. We confirm this conclusion from our analysis of the Senegal use case. They highlighted the main flows of population between the different regions, especially in the axis between the north and the south of the country.
They underlined that the network of trajectories is in fact a reflection of the country infrastructure and demography, but also economic activities. Bengtsson et al. found in [16] a high degree of alignment between the mobile phone activity based population counts and the population extracted from official surveys. In [17], Glass et al. showed that the analysis of movements based on mobile network activity data presents challenges, because locations are not observed at regular time intervals but rather only when user activities occur. CDR entries are in fact generated only when a user is engaged.
in a voice call or a text message activity. In [18], Dusi et al. analyzed the peak traffic hours for outgoings calls. They discovered two modes at 10h in the morning and 20h in the evening. In the case of Senegal, we also highlighted the daily profile of each antenna, that generally starts earlier (08h) and finishes later (23h). The authors studied the temporal profile of calls and also the distance between callers and callees. They also showed that users in highly populated areas with population larger than 100K citizens, generally initiate short distances calls. In [19], Lu et al. used specific metrics,
In this paper we have presented a modeling of the daily traffic demand profile of each Orange base station in Senegal for the year 2013, by considering voice and messaging activities. A classification has been performed into urban, suburban and rural modes. We showed that the number of calls (and duration) and text messages made from an antenna is proportional to the population of its covering area. An algorithm based on an outlier identification scheme has been developed to detect traffic anomalies, e.g., unusual situations of mass population movements, based on the daily profiles. We illustrated this traffic anomalies detection with different examples. We have extracted mobility maps from CDR datasets, where we showed population flows inside the country in 2013. CDRs provide a good opportunity to retrieve mobility flows from mobile phone datasets. Thus it becomes feasible to monitor the impact of important events in urban areas.

ACKNOWLEDGMENT

The authors would like to thank the National Research Fund of Luxembourg (FNR) for providing financial support through the CORE 2013 MAMBA project (C13/IS/5825301).

REFERENCES

[14] https://atoms.scilab.org/toolboxes/nan

Fig. 25. fourth quarter mobility analysis (SET 2).

e.g. average travel distance and radius of gyration in order to measure the mobility patterns of individuals. They also extracted the entropy of day-to-day movement of individuals. We will consider those metrics in a future work.
Virtual Networks and Poverty Analysis in Senegal

Neeti Pokhriyal  
Computer Science and Engineering  
State University of New York at Buffalo  
neetipok@buffalo.edu

Wen Dong  
Computer Science and Engineering  
State University of New York at Buffalo  
wendong@buffalo.edu

Venugopal Govindaraju  
Computer Science and Engineering  
State University of New York at Buffalo  
venu@cubs.buffalo.edu

Abstract
Do today’s communication technologies hold potential to alleviate poverty? The mobile phone’s accessibility and use allows us with an unprecedented volume of data on social interactions, mobility and more. Can this data help us better understand, characterize and alleviate poverty in one of the poorest nations in the world. Our study is an attempt in this direction. We discuss two concepts, which are both interconnected and immensely useful for securing the important link between mobile accessibility and poverty.

First, we use the cellular-communications data to construct virtual connectivity maps for Senegal, which are then correlated with the poverty indicators to learn a model. Our model predicts poverty index at any spatial resolution. Thus, we generate Poverty Maps for Senegal at an unprecedented finer resolution. Such maps are essential for understanding what characterizes poverty in a certain region, and how it differentiates from other regions, for targeted responses for the demographic of the population that is most needy. An interesting fact, that is empirically proved by our methodology, is that a large portion of all communication, and economic activity in Senegal is concentrated in Dakar, leaving many other regions marginalized.

Second, we study how user behavioral statistics, gathered from cellular-communications, correlate with the poverty indicators. Can this relationship be learnt as a model to generate poverty maps at a finer resolution? Surprisingly, this relationship can give us an alternate poverty map, that is solely based on the user behavior. Since poverty is a complex phenomenon, poverty maps showcasing multiple perspectives, such as ours, provide policymakers with better insights for effective responses for poverty eradication.

1 Introduction and Motivation
According to the United Nations Development Program’s 2014 Human Development Index (HDI), Senegal is ranked 163 out of 187 countries with an HDI index of 0.485. HDI measures achievement in three basic dimensions of human development: health, knowledge, and standard of living. Senegal has a population of 14.1 million, with 43.1% urban population, and the median age of 18.2 years. It is one of the poorest country in the world, with over 9.2 million people living in multidimensional poverty. Wealth distribution in Senegal is very unequal.

Poverty incidence remains high, affecting about 47% of the population. There are wide disparities between poverty in rural areas (at 57%) and urban areas, where the poverty rate is 33%. More than 42% of the population lives in rural areas, with a population density that varies from 77 people per square kilometer to 2 people per square kilometer in the dry regions of the country.

On the other hand, the growth in mobile-cellular technology has been very impressive in recent decades. It is estimated that there are 95 mobile-cellular telephone subscriptions per 100 inhabitants worldwide [1]. In Senegal, there are 93 mobile phone subscriptions per 100 people, according to the latest world-bank report [2].

The power of growth of mobile technology poses a question: Can their accessibility be used to identify, characterize, and, in turn, alleviate poverty? Ours is a case study towards answering this question. An expected outcome is a high resolution poverty map of Senegal, and its poverty analysis, with some recommendations for effective policies for an inclusive growth. We believe that such poverty analysis with the growth of virtual mobility will be beneficial to a developing econ-
omy like Senegal.

1.1 Poverty Maps Currently the poverty maps are created using nationally representative household surveys, which requires a lot of man-power, and time, and continues to lag for Sub-Saharan Africa compared to the world [3]. The data is updated yearly, and assessed for poverty progress in 3 years.

Poverty has traditionally been measured in one dimension, usually income or consumption, called income poverty. Another internationally comparable poverty measure is the World Bank's $1.25 per day, which identifies people who do not reach the minimum income poverty line.

In 2010, the Oxford Poverty and Human Development Initiative (OPHI) launched a Multidimensional Poverty Index (MPI). It is a composite of 10 indicators across three areas - education (years of schooling, school enrollment), health (malnutrition, child mortality), and living conditions.

We use MPI for our poverty analysis, since it closely aligns with the Human Development Index, and is widely accepted to study poverty. MPI is robust to decomposition within relevant sub-groups of populations, like urban vs rural, geographic regions (districts/provinces/states), religion and ethnicity, gender; so that targeted policies can be planned for specific demographics.

The MPI data is available at region level for each of the 14 regions in Senegal. Figure 1 depicts the latest (2011) poverty map of Senegal.

2 Contributions

The main contributions of our work are two-fold. First we construct a virtual network for Senegal from cellular-communication data (Dataset 1), identify network-theoretic measures that correlate well with poverty indicators, learn a model that predicts poverty at a finer resolution, and finally build a poverty map for Senegal at an arrondissement level. Second, we learn a model solely based on the relationship of aggregated user behavior (Dataset 3) with poverty indicators, and generate a poverty map at a finer resolution.

Here are detailed technical contributions of our work:

- We construct a virtual network for Senegal, which is defined as a who-calls-whom network from the mobile communication data. Intuitively, a virtual network quantifies the mobile connectivity, and accessibility to the population. It signifies the macro-level view of connections or social ties between people, dissemination of information or knowledge, or dispersal of services.

- We study Senegal’s virtual network to empirically get the most important spatial regions. We assign each region a unique score based on its importance in the virtual network. We find that a network theoretic based measure, called centrality, provides a strong correlation with the poverty index. The more important the region, the less poor it is on the poverty index.

- As MPI is a composite index, we find how well each of the component of MPI correlate with the importance of the regions.

- We apply linear regression to learn a relationship between centrality and the components of the poverty indicators. Our model is, then, used to estimate the poverty at a finer spatial resolution of arrondissements. We show our predicted region-level poverty map, and validate that it correlates well with the the true poverty map.

- We provide an in-depth region-level analysis of the correlation and poverty, and explain the bias caused by the Dakar region as it has very high centrality and very low poverty.

- We also attempt to understand, and characterize poverty. Since regions may be poor because of various reasons - information poor, but resource rich; or resource poor, but information rich. To study this, we use the behavior indicators of the users provided in Dataset 3. Some of these indicators are: entropy of contacts, percentage of calls from home, radius of gyration, etc. These indicators characterize individual users based on how they call, move, and interact within the cellular network. We studied their correlation with the poverty index and interestingly, found that the correlation was not biased by Dakar. Using one of such indicators, which has a very strong negative correlation with the poverty index, we learn a model for predicting MPI and construct an arrondissement level poverty map for Senegal.

3 Related Work

Call data records (CDR) allow a view of the communication and mobility patterns of people at an unprecedented scale. In the past, several researchers have used CDR data to understand human mobility [6, 5]. However, there has been limited work in understanding relation between CDR data and poverty [7, 12, 8, 11]. Eagle et al. [11] correlated the diversity in communication with socioeconomic deprivation and found a strong positive
correlation between the diversity in calling patterns and socioeconomic deprivation in England. The closest work to that presented here is by Smith et al. [11] who have calculated a number of features like introversion, diversity, residuals, and activity to find their correlations between poverty index (MPI) of Côte d’Ivoire, and further build a finer granularity poverty index based on the feature that gave the best correlations. Our work is different in the following aspects: a). we study the virtual communication network using a rigorous network science approach and find that centrality based measures which focus on the importance or influence of nodes provide a better correlation with the poverty index, b). we treat MPI as a composite index and provide estimate models that predict the individual components of the index, c). we provide an in-depth region wise analysis of the correlation and discover the bias caused by the Dakar region because of its unique nature in Senegal, and d). we also compare the finer level poverty maps generated using the network centric approach with a human behavior based method. Our work on relating human behavior with MPI is motivated from work done in the past in which behavior indicators are extracted from CDR data and used to predict the socioeconomic indicators of a region [12, 8]. Specifically, Soto et al. [12] have proposed a Support Vector Machine model which uses 279 features (calling behavioral, mobility, and social) extracted from an individual users CDR to predict the socioeconomic levels at a census region. However, the predictive model requires knowledge of finer granularity poverty data and partial knowledge of a user’s home information. Instead, we use the 33 indicators provided in the D4D challenge data and provide a methodology to correlate the indicators at region and arrondissement level without requiring additional information about the users.

4 Senegal’s Virtual Network
We define the physical network of a country, as composed of transportation landscape like road, railways, ports, which are the nation’s arteries that fuel its economic growth. With the burgeoning growth of mobile communication, we define a virtual network of a country, that describes who-calls-whom network. Calls are placed for a variety of reasons including request of resources, information dissemination, personal. The call data records (CDR), provided by the Orange, provides an interesting way to characterize and understand the virtual network of Senegal.

While the physical network determines how people move, and goods are transported, virtual network determines how information or knowledge flows. Currently, a good portion of the information and services are dispersed virtually. While the physical network is limited by the inherent capacity of the roads, and railway network, the virtual network is dynamic. Due to people’s mobility across spatial regions, and the ubiquitous nature of cellular technology, both physical and virtual
networks interact creating complex dynamism. For a holistic understanding of any complex phenomenon, we need to understand both the networks.

Static maps are easy to get, but how to get the virtual network. Existing gravity models can provide an estimate of the flow (with the knowledge of a constant), however such estimates are static and over-reliant on the spatial proximity between sites. The CDR data, however, provides the actual measure of the information flow at a finer spatial and temporal resolution. We construct a virtual network of Senegal from the CDR data. Such a network is generic, and can be used for understanding multiple phenomenon involving dynamic interactions with the physical network, like e-health (while the physical network determines where disease spreads next, virtual network determines how it can be contained by proper dissemination of preventive knowledge), e-education, and e-commerce.

4.1 How to construct the Virtual Network To construct the virtual network, we need two entities: spatial regions, where calls are originated from or are received in; and virtual paths that signify communication among them.

In virtual network for Senegal, the spatial regions correspond to administrative areas (that can be arrondissement, departments or regions), and the virtual path between each pair of nodes corresponds to the volume of mobile communication (number of calls and texts) between them during the whole year of 2013.

For this study, we used the hourly antenna-to-antenna traffic available for 1666 cell phone towers (sites) to measure communication between sites for 2013 (Dataset 1), as follows:

- Create an information flow matrix at site-level \( M^s \) with 1666 rows and 1666 columns, such that the entry \( M^s_{ij} \) denotes the number of calls and texts exchanged between site \( i \) and \( j \) during the whole year. Each entry \( M_{ij} \) represents the calls and texts originated at site \( i \) and received at site \( j \).

- To get the arrondissement level information flow, we “coarsen” the site to site matrix into a 124 \( \times \) 124 matrix \( M^a \), such that the entry \( M^a_{ij} \) denotes the total number calls and texts originated at all sites in arrondissement \( i \) and received at all sites in arrondissement \( j \).

- To get the region level information flow, we “coarsen” the arrondissement to arrondissement matrix into a 14 \( \times \) 14 matrix \( M^r \), such that the entry \( M^r_{ij} \) denotes the total number calls and texts originated from all sites in region \( i \) and received at all sites in region \( j \).

The resulting virtual network is shown in Figure 2. On the map, each region is depicted by the latitude and longitude of its geographical centroid. In the graph, the size of each node denotes the total number of incoming and outgoing calls and texts for the region for the entire year. The thickness of the link indicates the volume of calls and texts exchanged between the corresponding pair of regions. Looking at the map, we see that regions, e.g., Dakar, Thies and Ziguinchor, which have low MPI are important nodes in the virtual communication network. On the other hand, regions with high MPI, e.g., Kaffrine and Kolda are not well connected with other regions. However, there are regions that are well-connected but are poor (e.g., Tambacounda) and regions which are poorly connected but are relatively less poor (e.g., Kedougou). This indicates that poverty is a complex phenomenon and needs to be understood from multiple perspectives like relationships with bordering countries, unique geographical settings, etc.

4.2 Virtual Network and Poverty Analysis Figure 2 shows a relationship between the importance of a region in the virtual network with the poverty index. This motivates us to find a quantitative measure of the importance of the region.

As a standard nomenclature, in a network the spatial regions are called nodes, and virtual paths are called edges. Network analysis involves extracting some quantitative measures or properties, associated with the structure of the network, nodes and/or edges. A popular measure is the relative importance of the nodes in the network. This measure is referred to as centrality [10]. It identifies the most important nodes in a network, and assigns a quantitative score to each node. Since importance can have many definitions ranging from nodes being central or cohesive, there are several centrality measures for networks. To calculate the measure we normalize the raw communication matrix \( M^r \) as discussed below.

\[
\tilde{M}_{ij}^r = \frac{M^r_{ij} \ast d_{ij}}{n_i n_j}
\]

where \( n_i \) (and \( n_j \)) is the number of cell-phone towers in region \( i \) (and \( j \)) and \( d_{ij} \) is the as-the-crow-flies distance...
Figure 2: Virtual network for Senegal at Region-level with MPI (Multi-dimensional Poverty) as an overlay. Thickness of links indicate the volume of calls and texts exchanged between a pair of regions. Size of the circle at each region indicates the total number of incoming and outgoing calls and texts for the region. Note that regions with plenty of strong links have lower poverty, while poor regions look isolated.

between the centroids of regions $i$ and $j$. We normalize the matrix $M^a$ in a similar way.

We use the number of sites ($n_i$) as an indicator of the size of the calling population in the given region. Thus, the value $\frac{n_i n_j}{d_{ij}}$ in (4.1) is proportional to the expected number of calls between two regions, which is very similar to the well-known Gravity model, which has been used in the past to predict the intensity of mobile phone calls between cities [9] ($numcalls \propto p_i p_j d_{ij}$). However, instead of using the population of the two regions ($p_i$ and $p_j$), we use the number of sites to get an estimate of the population with mobile phones. We set the exponent $\alpha$ for distance as 1 as it gave the best correlation with poverty level. The normalization procedure removes the impact of regional population and spatial distance on the information flow and measures the residual flow, assuming that all regions have equal population and are equidistant from each other.

4.2.2 Measures of Importance and their Correlation with Poverty Besides graph-theoretic measures, we also investigated direct features that can be calculated from the raw communication matrix ($M^r$) or the normalized communication matrix ($\hat{M}^r$) discussed as follows.

- **Activity**: This feature is a simple aggregate of outgoing flows from a region ($= \sum_{i \neq j} M^r_{ij}$ for region $i$). We also used a similar feature derived from the normalized matrix ($= \sum_{i \neq j} \hat{M}^r_{ij}$ for region $i$). Additionally, we investigated other variants such as the count of incoming flows, within flows, and total flows and found similar relationships with the poverty indicators.

- **Eigen Vector and Page Rank Centrality**: This is derived from the normalized matrix $\hat{M}^r$, and is a measure of the influence of a node in a graph. Eigen vector centrality of a node, $v_i$, is a weighted sum of centralities of all of its outgoing connections:

  \[
  x_i = \frac{1}{\lambda} \sum_j \hat{M}^r_{ij} x_j
  \]

  where $\lambda$ is some constant. In matrix notation, this can be written as $\lambda \mathbf{x} = \hat{M}^r \mathbf{x}$, such that $\mathbf{x}$ is the eigenvector of the matrix $\hat{M}^r$ corresponding to the leading eigenvalue.

**Page Rank** is a variant of eigen vector centrality and is widely used for ranking websites by search engines such as Google. However, the actual role of Page Rank is to rank nodes in a network based on their importance. It has been noted that the
Table 1: Pearson’s r Correlation of region-wise poverty indicators with communication graph features. \( H \) – Incidence of Poverty, \( A \) – Average Intensity Across the Poor, \( MPI \) – Multidimensional Poverty Index.

<table>
<thead>
<tr>
<th>Measure</th>
<th>( H ) corr</th>
<th>p-value</th>
<th>( A ) corr</th>
<th>p-value</th>
<th>( MPI ) corr</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>-0.87</td>
<td>( 6 \times 10^{-5} )</td>
<td>-0.81</td>
<td>0.0004</td>
<td>-0.82</td>
<td>0.0003</td>
</tr>
<tr>
<td>Eigenvalue Centrality</td>
<td>-0.83</td>
<td>0.0002</td>
<td>-0.80</td>
<td>0.0005</td>
<td>-0.79</td>
<td>0.0007</td>
</tr>
<tr>
<td>Gravity Residual</td>
<td>-0.81</td>
<td>0.0003</td>
<td>-0.76</td>
<td>0.0015</td>
<td>-0.79</td>
<td>0.0007</td>
</tr>
<tr>
<td>Introversion</td>
<td>0.82</td>
<td>0.0002</td>
<td>0.70</td>
<td>0.0040</td>
<td>0.79</td>
<td>0.0006</td>
</tr>
<tr>
<td>Activity (Normalized)</td>
<td>-0.81</td>
<td>0.0008</td>
<td>-0.76</td>
<td>0.0003</td>
<td>-0.79</td>
<td>0.0015</td>
</tr>
<tr>
<td>Activity (Raw)</td>
<td>-0.80</td>
<td>0.0006</td>
<td>-0.68</td>
<td>0.0075</td>
<td>-0.71</td>
<td>0.0040</td>
</tr>
</tbody>
</table>

classic eigen vector centrality (See (4.2)) performs poorly for directed networks while the Page Rank measure can handle directed networks better.

- **Gravity Residual**: As shown in (4.1), each entry of the normalized matrix \( \hat{M}^r \) measures the “residual” from node \( i \) to \( j \) after normalizing for population and spatial distance. We compute the total outgoing residual flow from each node as:

\[
Residual_i = \sum_j \hat{M}^r_{ij}
\]  

(4.3)

In the past [11], similar measures have been shown to correlate negatively with MPI, indicating that regions that communicate more are less poor.

- **Introversion**: This measures the tendency of the population within a region to communicate within the region instead of outside. The introversion measure can be calculated as:

\[
Introversion_i = \frac{M^r_{ii}}{\sum_j M^r_{ij}}
\]  

(4.4)

All the above measures give a score for each region in Senegal, based on its relative importance. Further, we study how these measures correlate with the poverty index of the regions. The MPI reflects both the incidence of poverty, i.e., the proportion of the population that is multidimensionally poor, and the average intensity of their poverty, i.e., the average proportion of indicators in which poor people are deprived. The MPI is calculated by multiplying the incidence of poverty by the average intensity across the poor (\( H \times A \)). Hence, we study the correlation of the network features with \( H, A, \) and \( MPI \). Table 1 shows the Pearson’s \( r \) correlation and the corresponding p-values. We observe that the various metrics have a strong negative correlation with the \( H \) value, which is the headcount ratio of poverty. Similarly, the metrics have a marked negative correlation with \( A \), which is the incidence of poor, and also with MPI of the regions.

4.2.3 Strong influence of Dakar region Although pagerank exhibits strong correlation with the indicators \( H, A, \) and \( MPI \), when we plot the correlation (see Figure 3) we see that Dakar has a very unique characteristic of very high centrality, and very low MPI, and occupies a corner in the scatter plot, whereas all other regions are spread at the other corner, with mid-to-high MPI and low-to-mid pagerank. We, then, remove Dakar, to see its effect on the correlation of pagerank with the poverty indicators. Surprisingly, we lose the high correlation between pagerank and poverty indicators when Dakar is removed. This is evident from Table 2. The correlation drops significantly with high p-values.

We attribute this to its geopolitical heritage and past history as a port during the colonial times. It is the largest city with 2.47 million people, followed closely by Grand Dakar at 2.35 million. There has been excessive economic activity in Dakar, which makes up more than half of the Senegalese economy in less than 1 percent of the national territory. But sustained economic development, there needs to be de-centralized development focusing on marginalized areas. This fact is also validated by the International Monetary Fund’s 2013 report on Senegal [4].

4.2.4 Generating Finer Resolution Poverty Maps To illustrate how to derive finer resolution poverty maps (at department or arrondissement levels), we use pagerank from Table 1 to predict \( H \) and \( A \). We learn two linear models using ordinary least squares regression to predict the Incidence of Poverty (\( H \)) and Average Intensity across Poor (\( A \)). The learnt models are:

\[
\tilde{H}_i = -708.32 \times \text{PageRank}_i + 131.94
\]  

(4.5)

\[
\tilde{A}_i = -346.66 \times \text{PageRank}_i + 84.58
\]  

(4.6)

Finally, we combine the two estimates to predict the MPI as:
The estimated model for predicted MPI is shown in Figure 4. Using this model, we can estimate the MPI at a finer spatial resolution, as long as we can compute the pagerank for the target spatial areas.

\[
\tilde{MPI}_i = \frac{\tilde{H}_i}{100} \times \frac{\tilde{A}_i}{100}
\]

The region level predicted MPI map is shown in Figure 5. Note its similarity with the true MPI Map of Senegal in Figure 1.

Figure 6 motives the need for finer granularity poverty maps than regions. We can observe a significant variability in the centrality measure across arrondissements within the same region. This indicates that a region has varying levels of poverty. For targeted distribution of economic resources, we need finer level poverty maps than regions.

To generate the arrondissement level poverty map, we first compute its Page Rank using the normalized matrix \( \tilde{M} \).

Then we use the models in (4.5)–(4.7) to predict the MPI for each arrondissement. The predicted poverty map is shown in Figure 7. It is interesting to see that regions are composed of arrondissements with varying poverty index.

5 Correlating Behavioral Indicators of users and Poverty

In this section, we study how user behavioral statistics, gathered from cellular-communications, correlate with the poverty indicators. Can this relationship be learnt as a model to generate poverty maps at a finer resolution?

As previous researchers have shown [12, 8], human behavioral information extracted from CDR data can be used to measure the socio-economic development of a region. We study the relationship of several human behavior indicators extracted from CDR data with MPI with the goal of identifying key indicators which can then be used to predict MPI at a finer spatial resolution.

5.1 Data

For this study we use the one year of coarse-grained mobility data available at arrondissement level for 146,352 users (referred to as Set3 data). For each user, the data records the location (at arrondissement level) and time (at hourly level) at which the user makes a call or sends a text. Additionally, the data also contains a monthly set of 33 behavioral indicators which capture calling/texting patterns (14), mobility patterns (6), and social behavior (13) of each user.

5.2 Aggregating User Behavior

For each user, we compute the median of the 12 monthly values, for each of the 33 indicators. To relate these individual level indicators with region level MPI data, we need to assign a “home region” to each user. This information is not provided in the data set. We employ the following localization procedure to assign an arrondissement (and a region) to each user in the sample of 146,352.

5.2.1 Localization of Users

For each user we consider the calls made between 8 PM and 12 PM on each of the 365 days of the year. We measure the following quantities:

1. \( d_i \): Fraction of days (out of 365) the user \( i \) made at least one call between 8 PM and 12 PM in the whole year.
2. \( a_i \): The integer id (between 1 and 123) of the arrondissement that the user \( i \) called most frequently from during those hours.
3. \( c_i \): Fraction of total calls made by the user \( i \) between 8 PM and 12 PM from the arrondissement \( a_i \).

The arrondissement \( a_i \) is assigned as the “home arrondissement” for user \( i \) and the corresponding region is the “home region”. We filter out individuals with insufficient (low \( d_i \) ) or ambiguous (low \( a_i \) ) information by ignoring all users for whom \( d_i \leq 0.5 \) and \( c_i \leq 0.95 \), i.e., we only considered those users who made a call in the night at least half of the days in the year and who called from a single arrondissement 95% of the time. After filtering the sample contained 33,323 individuals (23% of the original sample).
Figure 3: Illustrating influence of Dakar on the relationship between MPI and Page Rank for regions.

Table 2: Pearson’s r Correlation of $H$, $A$ and $MPI$ with Pagerank of the regions considering Dakar and NOT considering Dakar.

<table>
<thead>
<tr>
<th>Measure</th>
<th>$H$ corr</th>
<th>$p$-value</th>
<th>$A$ corr</th>
<th>$p$-value</th>
<th>$MPI$ corr</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank with Dakar</td>
<td>-0.87</td>
<td>$6 \times 10^{-5}$</td>
<td>-0.81</td>
<td>0.0004</td>
<td>-0.82</td>
<td>0.0003</td>
</tr>
<tr>
<td>PageRank without Dakar</td>
<td>-0.68</td>
<td>0.01</td>
<td>-0.65</td>
<td>0.016</td>
<td>-0.64</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Figure 5: Predicted region level Poverty map using the virtual network.
Figure 6: Visual depiction of what happens when MPI is calculated at a region level. All arrondissements within each region are assigned the same poverty, but they have varying centrality measures, signifying importance! Thus, need to generate finer poverty maps for targeted eradication of poverty. The dashed box in the top panel is expanded in the bottom panel.

Figure 7: Predicted Arrondissement level Poverty map using the virtual network.
To verify that the filtered sample represents the entire country we compare the region wise distribution of the individuals with true 2011 population share and the region wise share of the number of cell phone antenna sites in Figure 8. We observe that while our filtered set of users oversamples from some of better developed regions of Senegal, the distribution approximates the population as well as number of sites (which in turn is an indicator of the number of mobile users) across regions.

Figure 8: Comparing the region-level population share for the indicator sample and the true population.

5.2.2 Region-Level Behavioral Indicators

To compute the indicators for each region, we consider all users assigned to that region. For each indicator, we compute the median value for each indicator. Thus we obtain 33 median indicators for each region.

5.3 Aggregated Behavioral Indicators and Poverty Analysis

For each indicator we compute the Pearson’s $r$ correlation between the region level median value for that indicator and MPI. Out of 33 indicators, 11 had an absolute correlation of 0.90 or greater with $p$-value $< 0.00001$. We chose one of these indicator variables with the strongest correlation with MPI – *Percentage Initiated Conversation (PIC)*. PIC had a negative correlation of -0.93 ($p$-value $= 2 \times 10^{-06}$) with MPI. Additionally, this indicator (as well as all other indicators) were not significantly influenced by Dakar (correlation without Dakar = -0.89, $p$-value $= 4 \times 10^{-05}$).

Result for PIC indicate that in regions with low MPI users tend to initiate more call/texts than the users belonging to regions with higher MPI. Similar to previous approach, we found the linear regression models to predict incidence of poverty ($H$), average intensity across poor ($A$) and eventually, the MPI for a given geographical region. The parameters of the linear models are:

$\hat{H}_i = -302.65 \times \text{PIC}_i + 119.35$ (5.8)

$\hat{A}_i = -151.53 \times \text{PIC}_i + 78.84$ (5.9)

The MPI is calculated by multiplying the two estimates

(See (4.7)). The estimated model for MPI using (5.8) and (5.9) is shown in Figure 9. Using a similar procedure as discussed in previous section, we generate an arrondissement level poverty map for Senegal as shown in Figure 10.

6 Conclusions

We analyze the virtual network for Senegal, constructed from call data records (CDR) in the context of understanding poverty. We propose a novel methodology to construct such networks at varying spatial resolutions, such as regions or arrondissements. We apply network centric methods, such as centrality, to measure the importance of each node in the virtual network, where the node either corresponds to one of the 14 regions or 123 arrondissements in Senegal. We show strong correlation of centrality and other measures with the poverty index of the region level nodes.

Since Multi-dimensional Poverty Index (MPI) as a composite of two individual indices, we learn a model that correlates poverty with each of the indicators. This allows us to learn a better relationship between the network centric measures and MPI. We provide an in-depth region-level analysis of the correlation between
centrality and MPI and discover a bias induced by the Dakar region and further analyze the cause of such bias. We provide an approach to utilize the user behavioral indicator data to understand their relationship with the MPI. This is the first time such analysis has been done to understand MPI. Through our analysis we discover indicators which are not only strongly correlated with MPI at region level (0.92 Pearson’s r correlation) but also are not biased by any particular region, as was observed for the centrality measures.

Since poverty is a complex phenomenon, poverty maps showcasing multiple perspectives, such as ours, provide policymakers with better insights for effective responses for poverty eradication. Poverty maps at arrondissement and department levels, or at any spatial levels, will enable targeted policies for inclusive growth of all the regions in Senegal. The poverty maps generated using the behavioral indicators can be used to focus policies for certain demographics of the society that are specially vulnerable to poverty, such as women and specific ethnic groups.

Acknowledgment

We gratefully thank Professor H. C. Pokhriyal, at University of Delhi, India and Dr. Varun Chandola, SUNY Buffalo for many thoughtful discussions.

References


N09

**COULD AND SHOULD BIG DATA COMPLEMENT OFFICIAL STATISTICS?**

**PERSPECTIVES FROM CELL PHONE DATA, POPULATION DENSITY AND CLIMATE VULNERABILITY IN SENEGAL**

Co-authors*:
Ana Areias¹, Emmanuel Letouzé², Gabriel Pestre³,
Bessie Schwarz⁴, Beth Tellman⁵, Emilio Zagheni⁶

¹Program Manager, Data-Pop Alliance (Harvard Humanitarian Initiative, MIT Media Lab, Overseas Development Institute), USA ²Director, Data-Pop Alliance; Visiting Scholar, MIT Media Lab; Fellow, Harvard Humanitarian Initiative; Senior Research Associate; Overseas Development Institute; PhD Candidate, UC Berkeley, USA, ³Research Assistant, Data-Pop Alliance, USA, ⁴Director of the Communications and Outreach Director, Yale Project for Climate Change Communication, Research Affiliate, Data-Pop Alliance, USA ⁵PhD Student, Department of Geography, Arizona State University, Research Affiliate, Data-Pop Alliance, USA, ⁶Assistant Professor, Department of Sociology, University of Washington; Research Affiliate, Data-Pop Alliance, USA.

December 2014

* This draft has received extensive inputs and feedback from Espen Beer Prydz, Economist at the World Bank Research Group and Research Affiliate with Data-Pop Alliance, USA, who cannot be formally listed as a co-author until its final version is reviewed and cleared by the World Bank. It also benefitted from discussions with Dr Ibrahima Hathe, Research Director for the “Initiative Prospective Agricole et Rurale” (IPAR), Senegal, an institute with which will collaborate on the final version and extensions of this paper.
1. Background and Objective

Basic reliable recent socio-economic and demographic data are notoriously lacking in poor countries, especially in Africa, which has been described as facing a “statistical tragedy”\(^1\) constraining development and humanitarian policy and programming.\(^2\) Much attention is paid to the potential of Big Data in general and cell-phone data analytics in particular, to fill data gaps and perhaps help ‘leapfrog’ statistical systems in developing countries\(^3\), but relatively few empirical analyses have so far addressed this theme.

Our paper examines specific aspects of this question in the case of Senegal, using cell-phone data from Orange along with other datasets, notably census data, Demographic and Health Survey (DHS) data, and data derived from satellite imagery. Our central question is simple, but we think potentially highly consequential. We investigate whether cell-phone activity can be a good proxy for population density at different geographical levels at different points in time, which may prove useful in responding to a crisis or shock that may affect vulnerable populations—such as flooding in Senegal.

Our starting hypothesis is that the correlation between cell-phone activity and ‘official’\(^4\) population figures is strong for large areas on average (over a year), but weak for small areas at any given point in time, which may point to a lack of reliable information on population distribution at low levels of temporal and spatial granularities. If so, perhaps cell-phone activity could help improve knowledge of where people are at a given point in time. Specifically, our paper aims to:

1. Analyze correlations between estimates of population density\(^5\) derived from cell-phone activity at different geographical levels and official statistics, after correcting for differential cell-phone penetration ownership across areas;
2. Analyze whether and how much using more granular population densities estimated from cell-phone activity instead of official static figures may yield a different picture of population vulnerability at different point in time and across regions.

Although preliminary and subject to several caveats—this being still a working paper in its early stages—our initial findings seem to confirm this hypothesis. If true upon further investigation, that conclusion could be read in (at least) two different ways.

(a) Dismiss these CDR-based estimates as just bad and useless, because of inherent biases etc.;
(b) Interrogate the accuracy of ‘official’ population densities once spatial and temporal granularity increase, and investigate whether and how CDR-based measures of population densities could be developed, taking into account sample biases.

2. Big Picture: Official Statistics and Cell-Phone Data Analytics

Demand side: Africa’s “statistical tragedy”

Despite improvements in the availability of human development statistics—notably following the introduction of the DHS programs in the mid-1980s and the systems set up to monitor the Millennium Development Goals (MDG) since 2000, overall a good indicator of a region’s underdevelopment continues to be the absence of recent, reliable development indicators\(^6\)—at the risk of placing them “off the map”.\(^7\)

A basic statistic often missing is population size. Only 12 of 49 countries in sub-Saharan Africa have held a census since 2004, and several African countries haven’t had a census in three decades. And so their population structures and distributions are essentially educated guesses (small differences in estimated vs. actual demographic growth rates rapidly make a big difference).

These realizations have notably presided over the 2013 call of the High Level Panel on Post-2015, for a “Data Revolution”\(^8\) that culminated in the recent publication of a report by a UN-appointed group of independent experts\(^9\). Frustration over the current state of the data landscape is also fueled by the ‘supply side’—i.e. the growing availability of new kinds of data, new tools and methods that may be able to alleviate in part the ‘statistical tragedy’.
Supply side: Big Data for development and CDR analytics

The field of Big Data for development is only a few years old, but it has already spurred a great deal of both excitement and skepticism since at least a much cited 2009 paper that found that light emissions picked up by satellites could track GDP growth and suggested they could supplement national accounting in data-poor places. Ten.

These debates go beyond the scope of this paper—spanning conceptual, definitional, theoretical, technological, political, ethical and legal dimensions—but it is useful and fascinating to be mindful of them, and summarize a few key parameters.

We define Big Data as the analysis of “traces of human actions picked up by digital devices”¹¹, generated about and by people as the by-product of their use of digital devices and web-supported tools and platform “whose primary purpose is not statistical inference” ᵃ². Since 2012, enough data have been produced to fill 80 billion 16GB iPhones that would circle the earth more than 100 times.¹³

Cell-phone activity is captured in CDRs, recorded by a telecom operator each time a call is made using a landline or a cell phone, which belong to the subset of “digital breadcrumbs” ᵃ⁴; structured data that may not constitute the bulk of Big Data size wise (videos are very large files) but probably hold the greatest potential for research and policy given the information about collective human behavior (rather than beliefs) they contain, though one of Big Data’s three ‘functions: descriptive (e.g., a map); predictive—either for proxying one variable based on another, or forecasting a future event; or prescriptive (i.e. causal inference) functions. ᵃ⁵ Our analysis sits squarely in the ‘proxying’ category.

The rapid spread of cell phones has spawned both examples of how they can be used to improve policy and programming¹⁶, but equally, controversies over technical issues (such as sample bias) and broader ethical issues. A handful of cases are widely cited there is no need for replication and duplication.¹⁷ Of particular interest are studies that use data aggregated to the level of cell towers or above rather than individual records, which somewhat reduce privacy concerns. (experiments using individual level cell phone credit purchases and calling patterns inevitably raise even more serious ethical issues.¹⁸)

Two papers have explored aggregated CDRs in relation to poverty in Ivory Coast, by Smith-Clarke et al. (2014)¹⁹ and Smith-Clarke et al. (2013) who undertook a similar analysis using 2005 DHS data for ground-truthing. A weakness is the multi-year gap between both datasets, leading the authors to conclude that, in the case of Ivory Coast but with broader applicability, “a valuable extension to our work would … be to obtain more up to date socioeconomic data” (p. 6).

In this paper, we are able to use recent census and DHS data for the same period as that covered by CDRs.

Using Cell-Phone Activity to Estimate Population Density

In particular, our paper builds on that of Deville et al. (2014). These authors use data on call density aggregated to the level of cell phone towers to estimate local population density and its variation across time in France and Portugal on the basis of CDRs.²⁰ The model was able to reflect population patterns caused by work-day versus weekend patterns of activity, as well as holiday-related resettlement. They validated their model against WorldPop data, derived from satellite imaging, and found that the CDR-derived estimates had the same degree of accuracy overall, but the advantage of providing much more timely information since estimates could be updated daily. Such high temporal resolution is vital in the case of natural disasters, disease and other emergency situations.

A well-identified challenge identify is to account for diverse cell phone usage patterns where mobile penetration is low; in their words: “applying the method to low-income countries where penetration rates are increasing rapidly but still exclude an important fraction of the population would require further sensitivity analyses of the impact of phone use inequalities, especially as marginalized populations also are the
most vulnerable to disasters, outbreaks, and conflicts” (n.p.). They explicitly mention the DHS as one potential source of data on mobile penetration—which we use.

A nascent literature has discussed variations of cell phone penetrations and their resulting biases, and whether and how it can be ‘corrected’ for—which conditions the appropriateness of inferential statistics. In Rwanda, for instance, phone owners were found to be ‘disproportionately owned and used by the privileged strata of Rwandan society’, especially men.21 Similar findings were made in Kenya.22 Despite such bias, in Kenya, the same authors show that “mobility estimates are surprisingly robust to the substantial biases in phone ownership across different geographical and socioeconomic groups”.23 The extent to which population count estimates are affected by such biases requires further investigation.

3. Population Density Estimations

Data and Strategy

Our analysis uses anonymized CDR of phone calls and SMS exchanges between more than 9 million of Orange’s customers in Senegal between January 1, 2013 and December 31, 2013, released as part of Orange’s 2014-15 Data for Development Challenge; we use one of the 4 datasets that were available,24 providing individual trajectories for 50,000 customers.

Our analysis is able to test the relationship between call activity and ‘official’ population count and density observed therefore “ground-truths” the estimates derived from call records. This is made possible by the fact that the 2013 Demographic and Health Survey (DHS) was conducted over nearly the exact same time period as from September 2012 through June 2013, a period similar to that of the CDRs.

For this paper, Voronoi polygons are formed by partitioning the map of Senegal into 1,666 cells with each cell containing an antenna and all the points that are closest to that antenna.25 For each user ID, the antennas to which they connected the most times over the period under analysis were identified as their “home tower”. An estimate of active cellphone users for each tower is the number of users that tower has as home tower. We then sum up voronoi data at the level of their corresponding arrondissements (103) and départements (45) to build and test the nature and extent of changes in the predictive power of CDR-derived estimates between these different levels.

Three sources of imprecision may skew our results. Our model assumes that all areas are covered by cellphone patterns. Second, we only have CDRs from one operator, whose penetration may be skewed across the country—a potential source of bias we are not able to control for lack of sufficiently granular data. Third, we use cell phone ownership at the département level, which is not necessarily representative of the rate of cell phone ownership at lower levels of granularity.

Population size as a function of users

We developed a model to evaluate the relationship between population size for geographic regions within the Senegal and Call Detail Records (CDRs) for Senegal made available by Orange as part of the D4D challenge 2014-15.26 To calibrate and validate our model, we used census data at the level of 45 départements and and estimates produced by the WorldPop project at the 103 arrondissements level.27

We tested the relationship between the number of cellphone users in a region and population size for the same region:

\[
P_i = U_i^k
\]  

(1)

Where \( P_i \) is the population size in the geographic region \( i \), \( (\text{voronoi, arrondissement and départements}) \) \( U_i \) is the number of active cellphone users in the Orange sample in the geographic region \( i \). \( k \) is the exponent of \( U \) and a parameter that has to be estimated.

The parameter \( k \) in the model that we described above can be estimated using a linear regression approach, after a logarithmic transformation of the variables:

\[
log(P_i) = k \log(U_i) + e_i
\]  

(2)
For our empirical analysis we estimated the number of cellphone users $U$ in each region $i$, by evaluating the cellphone tower to which each user connected the most. In other words, we evaluated the modal tower for each user and assigned the residence of each user to the region where the modal cellphone tower is located.

At the level of the département, we observe a very strong superlinear relationship between the logarithm of cellphone users and the logarithm of population size is quite linear and predictable (see Figure 1).

The $R^2$ of the regression model is equal to 0.92, meaning that most of the variability in population size across départements is accounted for by the geographic distribution of mobile phone users $U$. The relationship described in equation 2 is not new. It has been shown that it holds in other contexts. For instance, Linard et al. (2014) used the same type of model to show that population distribution can be estimated from the spatial distribution of mobile phone users in countries like Portugal and France.

What is new, however, is the application of this model to a setting where mobile phone penetration is not complete, and the use of DHS data on actual cell-phone penetration levels to correct for the variation across regions in penetration rates.

Figure 1: Relationship between the logarithm of census population and the logarithm of Orange mobile phone users at the department level in the Senegal.

**Attempting to correct for sample bias**

Mobile phone penetration rates are expected to be quite uniform in developed countries, but there may be substantial heterogeneity within and across developing countries. Below we describe our modeling approach.

We assume that the model described in equation 2 does not provide an unbiased representation of the underlying relationship. More specifically, variations in mobile phone penetration rates across the departments generate some systematic bias in the model and we expect the following underlying relationship to hold:

$$\log(P_i) = k \log(U_i) + \text{bias}_i + \epsilon_i$$

where:

$$\text{bias}_i = f(\text{mobile phone penetration}_i)$$

Equation 3
Estimates of mobile phone penetration are obtained from the DHS for the Senegal, conducted in 2013. We used the information collected about the share of individuals in the household with a cellphone. Estimates from the DHS survey are representative of macro-regions within the country. Table 3 summarizes the variation in mobile phone ownership across regions in the Senegal. For example, we observe that virtually everybody has a mobile phone in the city of Dakar while in other parts of the country mobile phone penetration is lower, at around 70%.

Table 3: Fraction of individuals in the household with a cellphone, by geographic region of the Senegal (see annex)

The expression presented in equation 4 is supported by empirical evidence. Thus we ran a linear regression model, where the dependent variable $Y$ is the vector of the residuals from equation 3 (it is the sum of bias and random noise) and the independent variable is mobile phone penetration rates across geographic regions:

$$Y_i = \beta_0 + \beta_1 \text{ mobile phone penetration}_i + \epsilon_i \quad (5)$$

We found a strongly significant positive relationship. This means that the higher the mobile phone penetration in a region, the lower the bias. In other words, the higher the mobile phone penetration, the smaller the underestimation of the population size. In those regions where cellphone use is less widespread, we expect that using the number of cellphone subscribers to estimate population size would lead to more sizeable underestimates.

The coefficients of equation 5 are strongly statistically significant. The relatively low variance explained by mobile phone penetration is related to the fact that estimates of mobile phone penetration from DHS are for macro-regions of the Senegal. At the cluster level, estimates from DHS are statistically representative. We assumed that all the areas within a cluster have the same mobile phone penetration. In the current setting we capture only a small portion of the variability across regions. That explains the low $R^2$ for the model in equation 5.

The main idea is that using mobile phone penetration rates, we can partially evaluate the extent of the bias at the department level using the model described in Equation 5. Plugging in the expected values of the bias into equation 3 would then allow us to improve the fit for our model that predicts population size based on number of mobile phone users.

In order to test the predictive power of our approach, we split our data set into two portions:

1. A first one, which contains about 2/3 of the observations, has been used as a training data set.
2. The remaining 1/3 of the observations have been used to test the ability of the model to make predictions “out-of-sample”. In other words, we calibrate our model using data on population size, number of mobile phone users and penetration rates for 2/3 of each the 3 types of areas considered.

We then assume that we only had data about number of mobile phone users (from CDRs) and mobile phone penetration rates (from DHS) for the remaining départements, and we test the ability of our model to predict population size for these départements. More specifically, we estimate the coefficients of the regression models in equations 2 and 5 using the training data set. We then plugged the estimated parameter $k$ from equation 2 and the estimated bias, conditional on penetration rates, from equation 5, into equation 3 for the test data set and evaluate the mean absolute percentage error (MAPE).

Our preliminary results indicate that the MAPE calculated on the test data set when no bias is accounted for, and thus the model reduces to the one described in equation 1, is equal to 0.0195 (for the model that predicts the logarithm of population size). When we include a bias correction based on equation 5, the MAPE reduces to 0.0195. The improvement in predictive power when we account for bias is modest, but it is robust to the choice of random samples for the training and test data sets.
At the arrondissement level, the correlation is statistically significant but lower, at .71, with a MAPE of 3.12. The relationship between cellphone ownership and the residuals of the prior regression has an $R^2$ of 0.06 and significant $p(0.018)$; the population estimate on the test data, correcting for bias, has a slightly better MAPE of 2.97.

At the voronoi level, results are very weak/non existent: the $R^2$ between WorldPop estimates and subscriber counts at the antenna level is significant but very low at 0.02. This model predicts population with a MAPE of 7.373.

Of key interest to policy makers is the extent to which such granular data can be used to predict the spatial distribution of poverty.

**Figures 2: Deviation in population densities between ‘official’ and predicted from cellphone activity: (i) Départements (right); (ii) Arrondissements (left); and (iii) Voronoi (bottom)**

These maps show how the correlation is much stronger at the department than at the arrondissement and voronoi—as conveyed by the % or deviation from the ‘official’ statistics, which never exceeds 1.2% in the test departments, vs. 4% in the test arrondissements vs. up to 100% at the voronoi level.

To state the obvious, this is most likely simply a reflection of the fact that people rarely move further away from the ‘official’ home to be ‘assigned’ a different department than where they reside, while this is probabilistically much likely to be the case at the level of a cell-tower.
4. Relevance and Application: the Case of Climate Vulnerability

**Rationale**

One reason why getting reliable—i.e. as accurate as possible—population density estimations is Senegal is because the country has faced major flood events in recent years. Although flood recession farming techniques have historically been used in various parts of the country, notably the River Valley, catastrophic floods have taken a high human toll as well as destroyed crops in recent years.28

In a subsequent version of this paper we refine our initial assessment of whether how much, when and where, using different population density estimations affects vulnerability monitoring and mapping. The rationale is that population density is one of the factors determining vulnerability in most models—and specifically in the one we intend to test against different population distribution estimations.

**Model**

To do so, we will use a flood vulnerability prediction model that analyze social and physical risk to riverine flooding based on multiple data inputs and runs basic statistics on the populations most vulnerable—developed and owned by two co-authors of this paper.29

The model is built in the new cloud geotechnology platform Google Earth Engine API and uses publicly available physical satellite imagery and demographic data extracted from censuses, surveys and/or reputable sources—such as WorldPop.

The model first assesses the hydrology of Senegal, second determines the social vulnerability within the flood zone, and finally predicts the characteristics of the high-risk areas. In this version of the paper, we were only able to run the baseline model using WorldPop data at a highly granular level (census data not being available at that level yet).

For this paper, we customized a qualitative index for likelihood of flooding within each pixel using several biophysical indicators: slope, elevation, watersheds with a high percentage of impervious surface, and locations with a high topographic index (ratio of upstream contributing area to slope). Elevation is based on 90m SRTM elevation dataset. Slope is calculated in degrees using the elevation layer. Impervious Surface is calculated from vegetation data using an NDVI index from a 2012 global composite of Landsat 7.

Finally Topographic Index describes the spatial distribution of the soil moisture and related landscape processes30, calculated with continental scale flow accumulation rasters. The top 10% of Senegal most likely to be flood is the high-risk zone or floodplain. We estimated the population within this zone from the WorldPop population rater for Senegal.

To assess the overall social vulnerability of Senegal we use a method similar to the physical vulnerability methodology: first creating a qualitative risk surface using three variables for age structure and population density. Indicators for social vulnerability use 100m resolution data from WorldPop a global high-resolution database and are selected based in part on vulnerability scholarship primarily from the Hazard and Vulnerability Institute31.

Young populations make areas more vulnerable because young children require assistance during evacuation and recovery. This model considers children to be people ages 0-5 years. Conversely, elderly populations require physical and informational assistance, and people ages 65 and up are considered elderly. Last and critically for our investigation as mentioned above, high densely populated areas tend to be more vulnerable, and we calculate density from the raw WorldPop population grid.

However, one can understand and monitor population density statically or dynamically. Our argument is that CDR-based estimates of population distribution may yield more accurate vulnerability maps at low spatial and temporal levels than one based on official statistics—by definition largely time invariant as of now, even if granular.
In this version, indicators are each divided by overall population per 100 meters, normalized, and combined into one surface. The top 15% of the country most socially vulnerable to disaster is considered the high social vulnerability area.

**Results:**

The initial model predicted 782 square Kilometers of Senegal to be at high risk from flooding based social and physical indicators. 242,739 Senegalese total are exposed to flooding and 91,732 of these Senegalese are highly socially vulnerable.

Using the CDR derived population density estimates, the area exposed to flooding is only 154 kilometers squared and number of people exposed to flooding is only 48,151 people, as this is a smaller area of analysis. The number of highly vulnerable people is 5,550 people. This is because we only have population density estimated for a very limited area (see below).

If we run our original model on this limited area (same spatial area as CDR estimates), the total number of people exposed to flooding is evidently the same, but the number of highly vulnerable people exposed to flooding increases to 22,000 people. It seems the WorldPop has a higher population density than the CDR estimates in these areas. In other words people are predicted as vulnerable with CDR data—perhaps because they have adapted their behaviors in the face of risk.

**Figures 3: Vulnerability based on WorldPop data.**

**Figures 4: Vulnerability based on CDR-derived population density estimates**

Pink areas below are vulnerable people WITHOUT CDR data. Blue areas are vulnerable to flooding.

Notice that fewer people appear socially vulnerable.
5. Concluding remarks

Our paper points to several preliminary results as well as interesting avenues that need further investigation.

We feel that main question of interest is indeed an important one, and that our approach could in time yield important findings that may inform policy. To reiterate, the main hypothesis, which our initial results seem to confirm, is that time-invariant and average (over a year, typically, or at-mid-year) official statistics on population distribution and density may not be the best way to capture underlying vulnerabilities to shocks at high level of spatial and temporal granularities. We have attempted to test this, trying to control for differential ownership of cell phones found in DHS data. This, we believe, is a novel and potentially fruitful direction.

If this result is indeed robust to additional research, it means that in the case of flooding, for example, official statistics may not be the best indicator of where to focus response. In particular, it may pick up the fact that communities adapt their movements to risks in ways that static observations do no pick up well. This may open the possibly of developing more timely and targeted response systems.

The paper also has obvious limitations, which are material for opportunities. As stated in the introduction, this paper is work in progress, and needs further elaboration. We hope to be able to conduct more-in-depth and broader analysis in collaboration with IPAR, the African School of Economics and CODESRIA to discuss our results and get local perspectives on how they can be improved.

We also do not discuss in this paper criticisms made to others of its kind—that these data were aggregated and used with little if any consent of the population that generated them, and that the results may not be used for practical purposes nor used to build local capacities.

We intend to use this paper’s subsequent version and its results to engage in the debates around ethics and empowerment of at-risk communities. Broadly, a risk that has not received the attention it merits is Big Data’s potential to create a ‘new digital divide’ that may widen rather than close existing gaps in income and power worldwide. One of the ‘three paradoxes’ of Big Data is that because it requires analytical capacities and access to data that only a fraction of institutions, corporations and individuals have, a ‘data revolution’ may serve to disempower the very communities and countries it promises to serve. People with the most data and capacities will be in the best position to exploit Big Data for economic advantage, even as they claim to use them to benefit others.

A last basic challenge is that of putting the data to use. Most discussions about the ‘data revolution’ assume that ‘data matter’; that poor data are to blame for poor policies. But lack of data has historically played only a marginal role in the decisions leading to bad policies and poor outcomes. And a blind ‘algorithmic’ future may undercut the very processes that are meant to ensure that the way data are turned into decisions is subject to democratic oversight.

But it is of course is not always clear whether and how much having better, finer data—even the best, most accurate, near-real time data possible—would make any difference to the rather derelict state of the world; in other words to what extent and through which mechanisms if any the pervasiveness of poverty, rise of income inequality, and perpetuation of social oppression and environmental degradation, among other bad outcomes, are consequences rather than mere correlates or even causes of bad data.

However, putting these data in the hands of at-risk communities would certainly help—allowing them to monitor population densities and vulnerability in case of a shock may go a long way towards building more resilient communities and saving lives.

We very much hope to be able to work towards this objective in the next few months.
Selected References


Letouzé, E. Big Data for Development Opportunities and Challenges (2012), UN Global Pulse, New York


**Statistical annex**

1. **Department level**

**Table 1.1. Population and subscriber counts at the Department level**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>log_censuspop13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log_sub</td>
<td>0.623***</td>
<td>(12.31)</td>
</tr>
<tr>
<td>_cons</td>
<td>7.216***</td>
<td>(16.79)</td>
</tr>
</tbody>
</table>

N 30  
$r^2$ 0.826

* $p < 0.05$,  ** $p < 0.01$,  *** $p < 0.001$

**Table 1.2. Cell phone ownership and residuals from Table 1.1.**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>res</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cell_own</td>
<td>0.250</td>
<td>(0.36)</td>
</tr>
<tr>
<td>_cons</td>
<td>-0.218</td>
<td>(-0.36)</td>
</tr>
</tbody>
</table>

N 30  
$r^2$ 0.826

* $p < 0.05$,  ** $p < 0.01$,  *** $p < 0.001$

2. **Arrondisement level**

**Table 2.1. Population and subscriber counts at the Arrondisement level**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>log_worldpop</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log_sub</td>
<td>0.517***</td>
<td>(17.28)</td>
</tr>
<tr>
<td>_cons</td>
<td>6.609***</td>
<td>(22.78)</td>
</tr>
</tbody>
</table>

N 92  
$r^2$ 0.717

* $p < 0.05$,  ** $p < 0.01$,  *** $p < 0.001$

**Table 2.2. Cell phone ownership and residuals from Table 1.2.**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>res</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cell_own</td>
<td>1.266*</td>
<td>(2.41)</td>
</tr>
<tr>
<td>_cons</td>
<td>-1.105*</td>
<td>(-2.39)</td>
</tr>
</tbody>
</table>

N 92  
$r^2$ 0.0643

* $p < 0.05$,  ** $p < 0.01$,  *** $p < 0.001$

3. **Voronoi(Antenna) cell level**

**Table 3.1. Population and subscriber counts at the antenna level**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>log_worldpop</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log_sub</td>
<td>0.133***</td>
<td>(5.05)</td>
</tr>
<tr>
<td>_cons</td>
<td>7.754***</td>
<td>(40.03)</td>
</tr>
</tbody>
</table>

N 1144  
$r^2$ 0.0245

* $p < 0.05$,  ** $p < 0.01$,  *** $p < 0.001$
Table 3.2. Cell phone ownership and residuals from Table 3.1.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>res</td>
<td></td>
</tr>
<tr>
<td>cell_own</td>
<td>0.00808</td>
<td>(0.02)</td>
</tr>
<tr>
<td>_cons</td>
<td>-0.00731</td>
<td>(-0.02)</td>
</tr>
<tr>
<td>N</td>
<td>1144</td>
<td></td>
</tr>
<tr>
<td>r2</td>
<td>0.000000494</td>
<td></td>
</tr>
</tbody>
</table>

* \( p < 0.05, \)  \( ** \) \( p < 0.01, \)  \( *** \) \( p < 0.001 \)
functions, to make sure that no group is being left behind (HLP 2013, p. 23).
3. We use population count and density almost interchangeably since the former is the former per some area.
4. In this paper we use official to refer to figures that are either produced by an official statistical system or meant to be accurate and widely used. We will refine this research focusing exclusively on census data when it is fully made public.
5. Population count and density are equivalent concept, the latter being a normalized version of the latter.
6. Claire Melamed, lead author of the Independent Expert Advisory Group on the Data Revolution"most of what we think of as facts in development, are actually estimates. We have actual numbers on maternal mortality for just 16% of all births, and on malaria for about 15% of all deaths. For six countries in Africa, there is basically no information at all. "http://www.theguardian.com/global-development/poverty-matters/2014/jun/31/data-development-reliable-figures-numbers
9. According to the Report of the High Level Panel, “data gathered will need to be disaggregated by gender, geography, income, disability, and other categories, to make sure that no group is being left behind (HLP 2013, p. 23).
11. Letouzé, Vinck and Meier, 2013
13. Computations by Letouzé and Preste for Scidev.net
14. Alex ‘Sandy’ Pentland—the Edge interview, 2012
15. Letouzé, Vinck and Meier, 2013
17. For example how CDR analysis has been used to study malaria spread in Africa (Buckee et al. 2014), or socioeconomic levels in a Latin American city [Soto et al. 2011], and many others For a review, see UN Global Pulse 2013, Letouzé various years, World Bank 2014, etc.
18. For example, Soto et al. (2011) use CDRs to predict poverty at the level of cell tower areas in a Latin American city with about 500,000 citizens, and compare their findings with official estimates – using information about the aggregated behavioural, social network and mobility of users, this approach predicted the socio-economic status of 80% of areas correctly. Gutiérrez et al. (2013) derived a proxy-wealth indicator for Senegal on the basis of information on phone credit top-ups (they hypothesized that poorer people would be likely to top up their phone credit in smaller amounts and with greater frequency) – but their results have not yet been validated against any established wealth indicator (cited in Smith-Clarke et al. 2014).
19. In this paper they report on two experiments they conducted in Ivory Coast and a “Region B” to preserve confidentiality, Region B is described simply as “another developing region”. For the former, they use CDRs on total traffic between cell phone towers for over 5 million phone users to construct geographically detailed income poverty maps, and to ‘ground-truth’ these data using 2008 IMF poverty estimates that are available nationally and for 11 subnational regions. For the latter, they use call records of around 928,000 mobile phone subscribers from early 2012. They follow a similar method in both cases.
23. Wesolowski et al. 2013
http://en.wikipedia.org/wiki/Voronoi_diagram
27. Details about the methods used to generate estimates of population size by the WorldPop team are described in Linard et al. (2012) and at http://www.worldpop.org.uk. WorldPop population estimates for regions within the Senegal can be downloaded from: http://www.worldpop.org.uk/data/summary/?conte selec=t_Africa&countselect=Senegal&typeselect=Population
28. For example see http://reliefweb.int/map/senegal/flooding-dakar-senegal-august-2013
29. Bessie Schwarz and Beth Tellman.
Estimation of travel times between cities in Senegal from mobile phone call records

Rainer Kujala, Talayeh Aledavood, and Jari Saramäki
Department of Biomedical Engineering and Computational Science,
Aalto University School of Science, P.O. Box 12200, FI-00076 AALTO, Finland
(Dated: December 23, 2014)

Whenever someone makes or receives a call on a mobile telephone, a Call Detail Record (CDR) is automatically generated by the operator for billing purposes. Recently, it has been realised that CDR’s have a wide range of applications beyond billing, from social science to data-driven development purposes. We have developed a method for estimating travel times between cities from CDR’s, and applied it to data from Senegal, released by Orange for the 2014 Data for Development Challenge. As the outcome, we present tables of typical travel times and speeds between Senegalese cities. This information fills an existing gap for Senegal (accurate travel time estimates are not available); however, the method is readily applicable anywhere where CDR information is available. It is also suitable for near-real-time operation for early detection of problems in the road network, or longer-term monitoring of e.g. road deterioration or the effects of improvements.

I. INTRODUCTION

Call Detailed Records (CDR’s) are secondary products of mobile phone communication that contain detailed information on each call event (caller, callee, time, duration, tower ID). Since CDR’s are automatically produced for billing purposes by mobile telephone operators, any other usage of them comes with no extra cost of data collection, neither to the company nor mobile phone users [2]. Because of their ubiquity, CDR’s have been proven to be useful for many purposes beyond billing, especially where traditional methods of data collection are too-labor intensive and costly, from population mapping to studying human mobility [3][4]. Further, because CDR’s are continuously being collected, they may be used for real-time monitoring and for providing crucial information for deciding on immediate actions after an incident (e.g. flows of displaced people after an earthquake [5]). As a downside, there are privacy issues that make CDR data sensitive and difficult to share. However, initiatives such as the Data for Development Challenge by Orange have made CDR-based data accessible to researchers by means of aggregation and anonymization. This facilitates studies that are application-oriented, and therefore results can more or less directly be applied to practical issues for development and governance.

For many developing countries, there is a lack of consistent, up-to-date information on roads, their condition, and travel times between locations. This is also the case for Senegal, the location of the 2014 Data for Development challenge, where anonymised call and location information has been released for development-related projects. Online travel time estimates between Senegalese cities vary greatly; e.g. estimates for travel times between Dakar and Tambacounda are typically up to 2× higher on travel guides and web forums than what Google Maps yields. This can result from several reasons including temporary variation (e.g. delays from checkpoints) to changing road conditions. At the same time, Dakar, the capital city of Senegal, is the second largest port in West Africa and therefore a major hub for transfer of goods to other countries in the region by land. Thus there is an immediate need for more accurate travel time estimates, as well as a monitoring system that keeps track of travel times on a continuous basis.

This paper introduces a method that provides travel time estimates from CDR data. The method is applied to D4D Challenge data on Senegal. As the end result, we present tables of estimated actual travel times between all largest cities in Senegal. From these, it is clear which roads are in worse-than-average condition or congested, and may need improvement. The method has been especially designed with a situation in mind where travel time information is not readily available from other sources or where it is too costly to collect; it can be readily applied to CDR data from any country, and modified to near-real-time use.

II. METHODS AND RESULTS

A. Data

In this study, we have used the 25 anonymized mobility data sets provided by Orange for the D4D Challenge. Each set contains ∼300,000 mobility traces for a two-week time span; a mobility trace contains the cell tower ID’s and time stamps of calls of one anonymised customer. In the provided data, users whose traces span less than 75%
of the days of a given two-week period, and users who have more than 1000 weekly events have been filtered out. Because of privacy and commercial reasons, only approximate coordinates are available for the cell towers instead of their exact locations. The provided data has a time resolution of 10 minutes. For a full description of the data set, see Ref. [7].

As our focus was on cities, not individual towers, we first obtained a list of cities and their geo-coordinates from www.tageo.com [8]. With the help of these and the tower locations, we assigned a set of cell towers to each city, such that each cell tower is mapped to its closest city whenever their distance is at most 10 km. The locations of cities and associated cell towers are displayed in Fig. 1A.

For an approximative proxy of Senegal’s road network and its current condition, we use the shapefiles provided by African Development Bank Group [9]. This data contains only the largest roads, and since this often leaves ‘holes’ at cities, we connected all loose ends of the road network to each other within a 7 km radius. For an illustration of the road network, see Fig. 4A. As the nodes of the road network denote intersections and turns of major roads, associating cities with their nearest road network nodes was a straightforward task. We determined the travel distance between two cities from the road network, using the shortest-path distance between the nodes corresponding to the two cities, and added the distances between the original positions of each city and the associated road network nodes.

B. Determination of travel times

Given two cities \(i\) and \(j\) and the corresponding sets of cell towers \(I\) and \(J\), we say that user \(u\) has made a travel from city \(i\) to city \(j\) whenever the mobility trajectory of \(u\) first contains one of the cell towers corresponding to city \(i\), and at a later point, one of the cell towers corresponding to city \(j\). The trajectory may contain other locations between \(i\) and \(j\). Thus a travel from \(i\) to \(j\) consists of a series of time-ordered observations (time, user, tower ID), where the first tower ID corresponds to city \(i\) and the last tower ID to city \(j\). The travel time is then defined simply as the time taken between the first and last observation.

To estimate the typical travel time from city \(i\) to city \(j\), we pool all travel times and investigate their distribution. In theory, the shortest observed travel time should be considered indicative of how fast one can travel between the two cities. However, there are some known errors in the source data [10]. Additionally, travel by air yields much shorter travel times than travel by land. As our interest is in estimating the typical time it takes to reach city \(j\) from city \(i\) along the road network, we thus need a more sophisticated method.

We have observed that the travel time distributions typically have a clear peak, and the position of this peak can be viewed as the typical travel time. However, some smoothing of the distribution is required. In more detail, our algorithm for estimating the travel time peak corresponding to travel by land is as follows:

1. Smooth the data-driven travel time distributions using a Gaussian kernel whose width corresponds to 30 minutes, and compute the smoothed density estimates with 5 min interval spacing. Smoothing is necessary for obtaining more reliable results when the number of observations between two cities is low.

2. Find all local maxima of the smoothed probability density functions, for which the corresponding travel speed does not exceed 120 km/h. (To filter out air traffic and errors in the data.)

3. Select the peak with smallest travel time such that the height of the peak is at least 0.5 times the height of the largest valid peak. Typically this condition results in picking simply the highest peak of the distribution. However, with a low number of observations this condition was found to be more meaningful, as it was able to provide more robust results.

The code implementing the above algorithm for extracting typical travel times from CDR data is freely available at https://github.com/rmkujala/d4dttimes

C. Validation of the travel time estimates

Since there is no accurate ground truth against which to compare our estimates, we have investigated two different consistency requirements in order to validate our results. First, we have focused on the relationship between the on-road distance and the estimated travel time, and found out that the relationship is approximately linear, which agrees with common sense. This linear dependence is emphasized if we exclude estimates that are based on small samples of data (\(n_{obs} < 1000\) data points) as is shown in Fig. 2A.

As a second sanity check of our algorithm, we have investigated whether our travel time estimates are symmetric: the travel time from \(i\) to \(j\) should approximately equal the time taken to travel from \(j\) to \(i\). This result is shown in Fig. 2B. As before, by discarding estimates computed with \(n_{obs} < 1000\) data points, the results become significantly less prone to noise.

When the above diagnostics were performed by varying the number of required estimates, \(n_{obs} = 1000\) turned out to be a reasonable threshold for reliable results (not shown). Thus in the following, this value is used for distinguishing between reliable and prone-to-error travel time estimates.
FIG. 1: Validation of estimates. Extraction of typical travel times between cities. Panel A: Each cluster of cell towers corresponding to a city are colored with same color. Black dots indicate cell towers that have not been assigned to any city. The thick green line in combination with the road network defines our split of Senegal into two parts which are handled separately in analyzes requiring accurate on-road distances between locations. Panels B & C: Two examples of travel time distributions (B: from Kaolack to Tambacounda; C: from Dakar to Ziguinchor). The empirical distribution of travel times is shown with black dots, and the red line denotes the kernel density estimate of probability density of travel times. The blue vertical line denotes the estimated typical travel time corresponding to the peak. In Panel B, we see a typical pattern with a single clear peak that stands out, at 265 minutes (∼ 4.4 hrs). This gives us an estimate of the typical travel time from Kaolack to Tambacounda. In Panel C, however, there are two peaks. The first peak is located at ∼100 minutes and is presumably from air traffic between Dakar and Ziguinchor. The second peak at 870 minutes (∼ 14.5 hours) can then be attributed to the typical travel time by land. This peak is wide, which implies variation in the travel times between Dakar and Ziguinchor on land.

D. Travel times and travel speeds between Senegalese cities

The end results of our method are displayed in Fig. 3. There are three tables that show the computed on-road distance, the estimated travel times, and the travel speeds between the 20 largest cities in Senegal (according to the 2000 estimate [8]).

To illustrate the practical applicability of our results, we consider the travel time between Dakar and Saint-Louis. According to Lonely Planet [11], the time it takes by frequent sept-place taxis to travel between the two cities is roughly five hours. As can seen in the travel time table of Fig. 3, our estimates for the typical travel time estimates (5h and 5h 5mins) exactly coincide with this duration. To the contrary, Google Maps and an Open-Street-Map-based routing service [12] provide travel time estimates of approximately 3.5 hrs. These estimates can well be attainable travel times, but it is likely that they do not reflect the typical travel duration.

The obtained travel speed estimates between Senegalese cities can also be visualized on a map and compared with the quality of the road network (earth or asphalt). This comparison is presented in Fig. 4. In the visualization, only city pairs that have more than 1000 trips in both directions and are located in the same part of the Senegal are considered. For each city pair, the speed estimate is the average of the two travel time estimates between the cities divided by the on-road distance between the cities.

By comparing the visualizations of the road network and the travel speed estimates, we can make a number of observations. First, looking at travels between the outskirts of Dakar region to cities like Tambacounda and Kedougou, we can see that the high estimated travel speeds are in line with the fact that the roads are qualified as asphalt roads instead of earth roads. Our speed estimates are also in line with the road network data when e.g. the travels between Tambacounda and Ouru Sogui are investigated. Now the travel speed estimate using the shortest path route is slow, and the roads corresponding to the shortest paths are earth roads.

However, there are also cases where the travel time estimates do not reflect the quality (asphalt/earth) of the road network. For example, travel speeds between Dakar and other cities are typically low, even though most roads around Dakar are asphalt roads. This reflects the known

FIG. 2: Validation of estimates. Panel A shows that the travel times are approximately linear with respect to distance. Panel B displays a scatter plot of travel times from j to i versus travel times from i to j; in the ideal case the travel times should be symmetric. Here it is seen that symmetricity is improved when the number of observations from which the travel times are computed is large enough. In these results, only city pairs that reside on the same part of Senegal are considered (see Fig. 1).
## On-road distance (km)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>65.0</td>
<td>204.9</td>
<td>814.7</td>
<td>76.4</td>
<td>247.7</td>
<td>148.3</td>
<td>192.3</td>
<td>313.3</td>
<td>480.9</td>
<td>179.8</td>
<td>648.8</td>
<td>874.5</td>
<td>140.3</td>
<td>818.0</td>
<td>110.2</td>
<td>247.0</td>
<td>116.4</td>
<td>49.3</td>
<td>109.1</td>
<td>56.6</td>
<td>434.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65.0</td>
<td>204.9</td>
<td>814.7</td>
<td>76.4</td>
<td>247.7</td>
<td>148.3</td>
<td>192.3</td>
<td>313.3</td>
<td>480.9</td>
<td>179.8</td>
<td>648.8</td>
<td>874.5</td>
<td>140.3</td>
<td>818.0</td>
<td>110.2</td>
<td>247.0</td>
<td>116.4</td>
<td>49.3</td>
<td>109.1</td>
<td>56.6</td>
<td>434.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>101.1</td>
<td>313.3</td>
<td>353.7</td>
<td>270.3</td>
<td>152.6</td>
<td>23.1</td>
<td>462.4</td>
<td>260.2</td>
<td>661.6</td>
<td>251.0</td>
<td>314.0</td>
<td>822.7</td>
<td>348.9</td>
<td>353.7</td>
<td>273.6</td>
<td>287.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Travel time (min)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>246.2</td>
<td>351.0</td>
<td>950.9</td>
<td>65.3</td>
<td>212.8</td>
<td>120.5</td>
<td>131.3</td>
<td>407.2</td>
<td>113.7</td>
<td>598.0</td>
<td>988.8</td>
<td>227.9</td>
<td>759.2</td>
<td>196.7</td>
<td>217.9</td>
<td>121.8</td>
<td>150.5</td>
<td>152.6</td>
<td>23.1</td>
<td>65.3</td>
<td>212.8</td>
<td></td>
</tr>
</tbody>
</table>

## Speed on shortest path (km/h)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>31.2</td>
<td>43.9</td>
<td>56.2</td>
<td>35.9</td>
<td>49.4</td>
<td>35.1</td>
<td>44.6</td>
<td>37.2</td>
<td>41.1</td>
<td>54.9</td>
<td>32.8</td>
<td>38.3</td>
<td>43.8</td>
<td>33.1</td>
<td>37.5</td>
<td>31.2</td>
<td>31.5</td>
<td>39.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## FIG. 3: Distance, travel time, and travel speed for the 2000 population estimate

In all tables, values are color-coded such that values on different parts of the map are colored red and big values are colored white. Non-diagonal gray entries denote city pairs which reside that travel to another city, and based of the view from (the view) and should be interpreted with caution. Star indicates results that are based on less than 1000 observed travels between cities.
congestion problems in and around Dakar.

E. Case study: opening of the Dakar–Diamniadio toll highway

As our results above point out, Dakar has suffered from congestion problems. To mitigate these problems, a new highway between Dakar and Diamniadio was opened on August 1st, 2013. We have evaluated the impact of this new highway by obtaining the approximate coordinates of its endpoints, and computed the travel time distributions between those before and after the opening of the highway. The computations were done in a similar manner as for the whole country, except that the radii that were used for choosing the cell towers associated with the end points were suitably adjusted.

The resulting distributions are presented in Fig. 5 and show how travel times have become shorter after the introduction of the new high way. Especially, the share of 20-minute travel times is greatly increased, which is in line with the anticipated 15 min travel time between Dakar and Diamniadio under normal traffic conditions [13].

Due to the limited temporal accuracy of the data provided by Orange (10 min), our peak detection method is not able to detect this improvement. However, with the original 1 second accuracy of the billing system, a peak with shorter travel times would likely be detected. Moreover, knowing the exact locations of the cell towers would greatly improve these estimates.

III. DISCUSSION

In this paper, we have estimated average travel times between cities in Senegal from CDR’s with the help of a method developed for the task. Especially for Senegal, such information has not been publicly available – while e.g. Google maps provides some estimates, their accuracy varies a lot, and while online forums may contain first-hand accounts by travelers, this information is scattered. The strength of our method lies in its ability to provide a picture of times actually taken for travel between any two cities, provided that the statistics in the source data are good enough.

It is worth stressing that our results have been obtained from the trajectory samples released for the D4D Senegal competition, and their accuracy would be greatly improved were the calculations based on the whole set of CDR data and no artificial accuracy restrictions. In some cases, the improved accuracy might allow for travel time estimates for different modes of transport [2]. Moreover, were an operator to apply the travel time estimator on a continuous basis, it would be possible to monitor road conditions in almost real time, e.g. on a daily or weekly basis. This would be very helpful for spotting problems early on, from increased congestion at certain hours of the day to deterioration of road infrastructure [10] or to delays because of road harassment [17]. We would also like to point out that since the method in the end works with aggregate distributions, it could be operated in a privacy-preserving way because individual mobility traces do not need to be stored, just the travel times. Finally, the usefulness of the method is by no means limited to developing countries - in general, with the help of their CDR data and the introduced method, mobile telephone operators can play a key role in the governance and planning of transport systems.

[10] Some of the positions of the base transceiver stations (BTS) that correspond to the cell tower IDs in the text are known to have changed locations during the time span of the CDR data. (Personal e-mail communication with the D4D-Senegal organizers, September 2014).
[14] Mali/Senegal: Road Development and Transport Facili-
FIG. 4: Comparing our results with the road network. In Panel A, we present the road network as obtained from Ref. [9], where blue color denotes asphalt roads, red denotes earth, and yellow patches close to cities denote roads that we added manually to bind together major roads that intersect at cities. In Panel B, we show the estimated travel speeds between various Senegalese cities. When the road network is compared to the estimated travel speeds computed using shortest paths on the road network, the travel speed and road type seem to correlate. However, travel speeds on asphalt roads around Dakar are in general lower compared to other parts of the country, which is indicative of the congestion problems Dakar is suffering from.

FIG. 5: Travel time distributions from Dakar to Diamniadio before and after the opening of the new toll highway. These distributions clearly demonstrate how travel times have been shortened due to the new highway. Please note that the 10-minute accuracy comes from the limited time resolution of the released data. Nevertheless, even with this limitation, the shift of the distribution towards shorter travel times is clearly visible. Were the same analysis performed with more accurate data, we expect that the lower peak around 20 minutes would be more emphasized and detectable using our algorithm.
Remote sensing has become an extremely useful tool to study, model and predict human mobility patterns during normal daily life, but also when these regular routines are disrupted by major events such as sociopolitical upheaval or natural catastrophes. Here, we defined a set of metrics that captured both regularities and anomalies in human mobility patterns, using the Orange Data for Development cellphone detail records. For each cellular event of each user, we extracted the distance to the most common cell phone tower of that user, the 'surprise' of that user visiting a particular cell phone tower, and the probability of seeing transitions between two specific cell phone towers. Averaging across all users that registered a cellular event at a specific tower at a specific time, we derived an average distance, surprise and probability of transitioning to that tower at that moment in time. Using dimensionality reduction, we were able to group sets of towers that shared one or more specific patterns. These patterns revealed both regularities and anomalies during the year the data were collected. The most obvious regularities were the daily commuting patterns of users in urbanized environments, while the anomalies represented major displacements as a result of religious observances. By detecting regularities, our metrics can be applied to infrastructural planning (resource allocation). By detecting anomalies (other than those anticipated) our metrics become an early warning system for events that change or disrupt daily life. This allows for a swift and informed response to such events.

Being ubiquitous and ever near us, the cell phone has become an extremely useful tool for remote sensing. When and where we decide to use our cellphones serves as a proxy for certain behaviors that are otherwise too time-consuming or expensive to measure. Records of cellular events reveal regular patterns of mobility and activity, and deviations from them. Such deviations can be indicators of major catastrophic events, such as earthquakes, socio-political instabilities like mass protests or even smaller, more local infrastructural problems (loss of electric power, water shortages). Observing and modeling people's collective actions (reflecting their collective needs and beliefs) prior, during and after such events provides useful insights on how we should respond to future crises.

Developing nations are most vulnerable to crises, lacking the necessary resources and infrastructure to cope with disruptive events (e.g. crop failures and epidemics). In addition, poverty and political, social, and economic inequalities within developing countries predisposes them to conflict. Described as quasi-democratic, Senegal is stable compared to other countries within its region. However, like many other African nations, it faces difficult developmental challenges. Senegal relies heavily on seasonal and non-irrigated agriculture on poor soil, making it very susceptible to natural vectors, such as droughts, floods and pests. It also experiences heavy urbanization, leading to extreme poverty in these urban areas. In addition, the literacy rate in 2011 was estimated to be 52.1%, making Senegal the 9th most illiterate country in the world. However, in a short period of time, mobile technology has become an integral part of everyday life for the people of Senegal: as of early 2014, 81% of all Senegals owned a cellphone, 13% of which were internet-enabled smartphones. This justifies using Senegal cell phone records as a proxy for user behaviors (e.g. mobility) and characteristics (e.g. demographics).

The current work used data provided by Orange as part of the Data for Development Senegal Challenge. It represents a subsampling of all records collected during 2013, and is large enough to be representative of all cell phone communications (at least in densely populated areas). This is the second time Orange has issued such a challenge; in 2013 they released cell phone

Acknowledgements. We would like to thank Dan Hill, Soren Larson, Kiril Tsemekhman, Evgeny Shmelkov, Katia Eliseeva, Igor Zabukovec, John Kittrell, Mansi Parikh and Nesha Burghardt for their useful comments and suggestions.
records collected in Cote d'Ivoire. That challenge yielded useful proposals and methodologies for using these cell phone records to aid development. This included recommendations on how to improve infrastructure and public transport\textsuperscript{11-13}, prevent epidemics\textsuperscript{14-16} as well as insights into the effects of weather\textsuperscript{17} and conflict\textsuperscript{18} on human mobility patterns.

Here, we present a set of intuitive metrics that can be used to predict and monitor both regular and anomalous human mobility patterns from cell phone records. More specifically, for each user we computed the probability of the user transitioning between two towers, the distance between each visited tower and the tower most commonly visited by the user, and the surprise of a user visiting a specific tower, regardless of the origin. We found these metrics to be correlated with each other, but more importantly, to be capable of detecting major events. We feel these metrics can be instrumental in optimally allocating resources (e.g. electricity) when patterns are regular and predictable. In addition, the detection of anomalies can be used as an early warning system of local or global events that have the potential to severely disrupt normalcy and put people at risk.

\section*{Methods}

\textbf{Raw call detail records (CDR)}

The raw call detail records (CDR) were released by telecommunications provider Orange. In terms of market share, Orange Senegal was the dominant carrier with a market share of 58.34\% at the end of September 2013, equivalent to around \textasciitilde7.4 million users\textsuperscript{19}. In the current study, we used one out of the three data sets provided by Orange. This fine-grained data set contained a random sampling of users. For each user, the data contained all of their cellular events (voice call or text message), when each event occurred and what cell phone tower registered the event. Since calls (and texts) were assigned to the closest cell phone tower within range, the cell phone tower's physical location becomes a useful proxy for the user's physical location. A new set of users was selected every two weeks, thus mobility patterns of a single user could only be observed within this two-week period. For privacy reasons, users were assigned a random identifier, rather than their cell phone number. In addition, the timestamp of each cellular event was rounded to the nearest 5-minute bin. The data set did not cover the full year 2013; records for the beginning of January and the end of December were not present.

\textbf{Anomaly detection in human mobility patterns: transition probabilities.} Transition probabilities were computed as follows. For each user, we removed all subsequent cellular events (voice calls and texts) at the same tower, as such events did not provide any information about the user's mobility. Based on the users' sequence of cellular events we derived a transition matrix $T$. Within this $n \times n$ square matrix ($n =$ number of unique towers visited by the user), each $T_{ij}$ entry counts the number of times a cellular event at tower $i$ was followed by a cellular event at tower $j$. Normalizing this matrix (by dividing all elements within each column by the sum across the column) creates a Markov model\textsuperscript{20} of cellular events: the probability that each cellular event at tower $i$ will be followed by a cellular event at tower $j$. It has been shown that these probabilities are typically not uniform. In most cases, the transition matrix is dominated by the two towers that represent the home and work location\textsuperscript{21}.

After the transition matrix had been established for each user, we replayed the entire sequence of cellular events of the user and extracted the probability of each event from the transition matrix. We associated this probability with the destination tower of the transition and recorded its timestamp. We then computed, for each tower, the hourly average transition probability of users reaching this tower as their destination (Fig. 1). This computation resulted in an $nxm$ matrix, where $n$ is the number of time points (24 hours x 365 days) and $m$ is the number of cell phone towers. In this matrix, if the average probability was high at a specific tower at a specific moment in time, the users entering this destination had a high probability of arriving there. Most likely, these probabilities were dominated by users commuting to this destination on a daily basis (either to work, or returning home). If this probability was low, then the destination tower was being visited by users that normally would not be in the vicinity of the tower. Alternatively, it could mean that the route taken to this tower was less probable, even though the tower visited was a common destination (e.g. detours). Note that it was also possible that this behavior was not caused by actual mobility, but rather by unavailability of the most probable destination tower.

\textbf{Anomaly detection in human mobility patterns: distance and surprise.} Two additional measures were computed for each cellular event. First, for each unique user, we computed the haversine distance between the tower at each transition destination and the tower representing the mode for a particular user. The mode was the most visited tower of a user, and represented either their home or work location. Thus an increased distance from the mode indicated that the user was further from home. The distribution of these distances has revealed important characteristics of human mobility patterns\textsuperscript{22,23}. 

\section*{References}

\begin{itemize}
  \item \textsuperscript{1}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{2}"\textit{Anomalous behavior in human mobility patterns.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{3}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{4}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{5}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{6}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{7}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{8}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{9}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{10}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{11}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{12}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{13}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{14}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{15}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{16}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{17}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{18}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{19}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{20}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{21}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{22}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
  \item \textsuperscript{23}"\textit{Anomalous human mobility patterns from cell phone records.} Anomaly detection in human mobility patterns.}
\end{itemize}
Figure 1. Steps in the computation of transition probability, distance and surprise. (A) For each user, we collected the number of times they visited each available cell phone tower. From this we computed the mode tower (the tower most frequently visited) for this user. In addition, we computed the measure of surprise of seeing the user at a specific tower, which we defined as \((1 - \text{frequency of visiting the tower}) / \text{maximum frequency across all towers}\). (B) For the transition probability, we computed the probabilities of each transition between available towers occurring in each user. (C) Replaying the sequence of observed cellular events for each user, we obtained a time stamped list of observed distances, surprises and transition probabilities. We assigned these metrics to the destination tower. Finally, we aggregated these metrics on an hourly basis, averaging the metrics associated with all observed cellular events within that hour.

Second, we computed a metric of surprise, which we defined as \((1 - \text{frequency of a cellular event occurring at this tower}) / \text{maximum frequency across all towers}\). The high numerator captured that towers visited very infrequently were very surprising. However, visiting a large number of distinct towers suppressed the maximum frequency any one tower could be visited. Therefore, the denominator (maximum frequency across all towers) normalized the metric to account for this. As with the transition probabilities, both the distance and surprise measures gave us an \(nxm\) matrix, where \(n\) is the number of time points and \(m\) is the number of cell phone towers.

**Between-metric correlations.**
After computing the metrics, we observed a highly non-linear dependency of the metrics on each other. Because of this, we captured their statistical dependency using mutual information, rather than a linear correlation. We first binned the transition probabilities, surprise and distance. We then calculated the joint probability distribution \(p(x,y)\) and the two marginal distributions \(p(x)\) and \(p(y)\). Mutual information was then computed using the following equation:

\[
I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \left( \frac{p(x,y)}{p(x)p(y)} \right)
\]

To determine whether the amount of information between each pair of metrics was statistically different from chance, we used a permutation test in which we randomly permuted the indices of probabilities and recalculated mutual information. Doing this a large number of times (1000x) generates a baseline distribution of mutual information. The actual observed mutual information was deemed statistically significant if it was greater than the 95th percentile of this distribution.

**Detecting regularities and anomalies**
To detect both regularities and anomalies in the patterns of human mobility, we used principal component analysis (PCA). Specifically, we computed the principal components of the matrices representing our metrics: 1) average probability of users arriving at a specific tower, 2) distance from the mode for users arriving at a specific tower and 3) average surprise of observing a user arrive at a specific tower. Projecting the original data on these principal components provides a matrix of component scores, sorted by their eigenvalue (amount of variance explained by each component). Components with sufficient amounts of explained variance represent mobility patterns that are shared across a sufficient number of towers.

**Grouping behaviors into memberships and communities**
The patterns observed at each tower are a linear combination of the full set of principal components. We estimated the weights \(W\) on these components for each tower using the original \(nxm\) matrix \(X\) (where \(X\) can represent any of the 3 measures computed). More specifically, we retained only the first 20 component scores, giving us an \(nx20\) matrix \(C\), where \(n\) is the number of time points. Estimating the matrix of weights \(W\) on each of these 20 components was done using ordinary least squares:

\[
W = (C^T C)^{-1} C^T X
\]

These weights allowed us to determine the extent to which each tower contributes to each behavioral component. In addition, we detected communities within the matrix of weights using a community detection algorithm. Specifically, the \(W\) was
The main source of variance in the mobility data was explained by the regular daily patterns of commuting back and forth between the home location and work. Inset shows the regular patterns observed over a few days. The second source of variation revealed four major religious observances: the Grand Magal of Touba, the end of Ramadan, Kazhu Rajab and the birth of the prophet Mohammed. Cell phone tower membership to the regular daily commuting pattern was high for urbanized regions, low for more rural regions. Cell phone tower membership to the four major religious observances was positive in Touba and the surrounding area, the destination of major pilgrimages during these events. In (C) and (D) each dot represents a cell phone tower.

weight matrix transformed into an adjacency matrix $D$, where each element $D_{ij}$ represented the similarity between the weights vectors of towers $i$ and $j$ on the components. While many different algorithms exist to detect communities within adjacency matrices, here we computed them using the algorithm described in detail by Arenas et. al. We found that the similar algorithms yielded very similar results.

Results

Component patterns of human mobility. Using the fine-grained data set, we computed three metrics that aim to capture both regularities and anomalies in human mobility patterns. Specifically,
Figure 3: Mobility patterns, part 2. (A) The third source of variance in the mobility data was explained by the religious holiday of Eid El Kebir (feast of sacrifice) and a reduction in the average distance metric within the month of Ramadan. (B) The fourth source of variation revealed a singular deviation from the norm around September 5th. (C) Cell phone tower membership to third source of variance was high for urbanized regions, low for more rural regions. (D) Cell phone tower membership during the Magal of two Rakkas of Ndar was high in St. Louis and the surrounding area where the event took place. In (C) and (D) each dot represents a cell phone tower.

(although ordered differently in terms of explained variance). Therefore, we outlined our results using a single and perhaps most intuitive metric: distance.

**Mobility component 1: commuting patterns.** The first component, explaining more than 36% of the observed variance, captured the most basic pattern of human mobility: traveling back and forth between home and work. Average distances from home were lowest in the middle of the night, increased in the early morning as people left their home location, and decreased as people returned home in the evening. Assigning each cell phone tower a level of membership to this component revealed that this pattern was most prominent in Senegal’s urban areas. Co-localized cell phone towers with a high positive weighting on the component revealed the location of the major cities of Senegal: Dakar, Thies, Tambacounda, Kolda, Ziguinchor, St. Louis and Kaolack (Fig. 2).

**Mobility component 2: pilgrimages.** The second component, explaining 8% of all variance, captured a few anomalies occurring on different days during the year. The deviations from the normally observed patterns of distance coincided perfectly with a number of important religious events unique to Senegal (Fig. 2). The biggest excursion was seen on December 22nd, the day of the Grand Magal of Touba. Every year, millions of Muslims from Senegal and around the world undertake a pilgrimage to Touba, honoring the memory of Sheikh Amadou Bamba, founder of the Mouride brotherhood. The second peak, observed around
Figure 4. Component-based communities. Community analysis of the weights placed on the 20 first components revealed a total of 13 distinct communities. These communities were geographically separated, even though no direct information about distance was used to detect the communities. Each dot represents a cell phone tower.

June 6th coincides with the Kazhu Rajab, a celebration in Touba to commemorate the birth of Serigne Fallou Mbacke, the second Caliph Mouride General. Consequently, the cell towers in Touba and its surrounding area received high positive weights on this component. A further number of smaller peaks line up with more traditional Muslim holidays, including the birth of the prophet Mohammed (January 24th) and the end of Ramadan (August 8th).

Mobility component 3: Ramadan. The third component captured some of the same events already captured by component 2, with slightly different dynamics. In particular, we observed a slight depression during Ramadan, in addition to the peaks at the start and end of this month. Ramadan is a very important social, cultural and religious event for Muslims, and changes people's daily mobility habits substantially. We analyzed the month of Ramadan below. In addition to Ramadan, this third component also captured another religious event on October the 15th: Eid El Kebir (feast of sacrifice), see Fig. 3.

Mobility component 4: Magal of the two Rakkas of Ndar. The final component, explaining 1% of the variance was very well localized geographically in the northwestern city of St. Louis. A slight deviation from the norm for this component was observed around September the 5th, which coincided with the religious commemoration of the Magal of two Rakkas of Ndar taking place in St. Louis (see Fig. 3).

Communities in regular and anomalous behavior.
So far, we showed the degree of membership of each cell phone tower to each discovered component. Next, we combined the relative weights of the towers on each component, in order to detect communities within the mobility patterns of Senegal. Formulated in this way, a community is formed by a set of cell towers that have a similar weighting on the components. Using the algorithm outline by Arenas et. al.27, we detected a total of 13 communities. These communities were geographically clustered, even though no actual distances were used to compute the adjacency matrix (Fig. 4). We could only detect with confidence, anomalous events that gave rise to the first few mobility components. These events were associated with major events that were reported on in the news. However, we argue that the later components contain more subtle anomalies that change and/or disrupt regular daily mobility patterns locally. These could include loss of electric power, loss of cell phone connectivity, poor weather conditions and disruptive events, such as violence, protests, fires and accidents.

Daily patterns during and outside Ramadan.
Averaging across all towers and days of the year revealed the daily patterns of our metrics. Starting around midnight, the distance metric drops, indicating that people stay home or at least closer to home at night. The distance metric rises again in the early morning, as people rise and get to work. A similar behavior was seen for the surprise metric. This, too, is explained by considering that people tend to stay at home at night, which is the least surprising place for most individuals to be. Finally, the same trend was observed for the probability metric, which was less intuitive. Why would we observe less probable transitions at night? One explanation is that not many transitions occur at night, since humans tend to travel or commute during the day. These daily transitions are highly predictive, but most nighttime transitions are not: they represent events that are less ordinary. This perhaps include a range of behaviors such as a medical or other kind of emergency, attendance at a party or other social gatherings, or sports, music and art events that typically occur at night.

How is the month of Ramadan different in this respect? During Ramadan participating Muslims do not eat from sunrise to sunset. However, immediately after sunset (~7.30PM during the 2013 Ramadan in Senegal), families have the fast-breaking meal known as Iftar. Social gatherings are frequent at Iftar. It is a time of being close to families, relatives and surrounding communities.
Figure 5. Daily activity patterns prior (black), after (gray) and during the month of Ramadan (red). Each line represents a circular trace of a full day of measurements, starting at midnight, and moving clockwise. (A) During Ramadan, nightly transition probabilities were significantly higher than observed outside of Ramadan, whereas the average surprise (B) and average distance (C) did not show such a significant deviation (although a trend is visible). See text for details.

With this in mind, it is not surprising to see that during Ramadan nights, people do not differ significantly in the distance metrics, although small but insignificant increases can be seen (Fig. 5). The same was not observed for the probability and surprise metrics. During Ramadan transitions observed at night were significantly more predictable, but also significantly more surprising. Although somewhat counterintuitive, this might perhaps be explained by the nightly prayers that accompany Ramadan. If those observing these nightly prayers did so by traveling back and forth between their home location and a location of prayer (e.g. mosque), we would expect to observe increased transition probabilities at night during Ramadan. At the same time, such locations are still slightly more surprising than observing, for example, the work or home cell phone tower, explaining why average surprise was larger during the nights of Ramadan.

Between-metric correlations
Intuitively, we expected our metrics to be dependent on each other. Breaking daily routines decreases the probability of observing a particular transition, and increases the surprise of visiting particular towers. Similarly, less probable transitions typically occur further away from home (during travel). However, no such simple relationships were immediately evident when the joint probability distributions between pairs of metrics were computed. First, the resulting distributions were multimodal, as a result of differences in distances between cell phone towers in Senegal. To account for such variations in tower density, we clustered the cell phone towers by the interdistance between the tower and its ten nearest neighbors. This revealed two well-separated clusters: one for dense urban areas in which cell phone towers were close and abundant, and one for rural areas where interspacing was much larger. Separately for each of these clusters, we calculated the amount of mutual information between each pair of metrics. This yielded significant dependencies between all metrics: in each case, the joint probability distribution was significantly different from the outer product of the two marginal distributions (statistical independence). We suspect that the metrics show a complex interaction that is conditional on other factors in the data that are not accounted for at this time.

Discussion
The human mobility patterns showed a mixture of regular and anomalous behaviors, each of which can aid development in the short and long-term.

Regularities
The regular patterns we observed clearly captured daily life in an urban setting. The transition matrices of people in these urban areas were highly non-uniform, with high probabilities of transitioning between two specific towers in particular: the one located nearest home, and the one located nearest work. Similar observations have been made for data collected elsewhere. The degree to which these patterns occur at a particular location can be used to remotely sense
the degree of urbanization, as well as unemployment levels. Furthermore, local deviations from these daily patterns can be used to detect local infrastructural issues (such as traffic jams, accidents). In addition, these daily patterns can be used for optimal allocation of energy resources. Available capacity can be shifted in accordance with where we expect the majority of people to be over the course of a day, and during different seasons. A larger sampling of cell phones could reveal similar or different regular patterns for remote areas. Seasonal migration is common in rural areas, but the sampling of cell phone records in the D4D data set was too low, and users were tracked too briefly (2 weeks), to really pick up on these migratory patterns.

Here, we assigned the metrics to the towers that users visited. An alternative approach is to assign them directly to the users and infer communities within the set of users (rather than the towers). Although this would be a powerful monitoring tool, this does come with risk of compromising people's privacy. Proper anonymization becomes key in that scenario.

Anomalies
The anomalies we detected were easily identified as major religious events, such as the end of Ramadan, and the pilgrimage to the holy city of Touba. The events are obviously known beforehand, but some useful information can still be extracted. For example, the patterns observed during a pilgrimage do not only reveal where the pilgrimage takes people (which is known), but also where they originate from (which might not be known). Our ability to detect these events suggests that we would also detect events of a more catastrophic nature: events that severely disrupt people lives. This includes natural disasters, such as earthquakes, tsunamis and droughts as well as sociopolitical events, such as protests and armed conflicts. They can also include infrastructural failures such as power outages and water shortages. Cell phone records have been used to study the effect of such major events on human mobility patterns, allowing us to model the response of humans to such events. This in turn allows us to predict the pattern for any future event and respond optimally to it.

Although Senegal is in definite need of humanitarian aid and development, fortunately it has not seen the level of violence experienced in nearby countries such as Mali and Côte d'Ivoire. For now, it has also been spared the 2014 outbreak of Ebola that is hitting Sierra Leone, Liberia and Guinea. However, it is prone to floods, droughts and pests that threaten its fragile agricultural system. Although we did not find any records of these events occurring in 2013, we feel optimistic about being able to detect future events that could severely disrupt and threaten human lives. Again, a higher sampling of cell phone records, especially in rural areas, would greatly aid in that ability.

On a local level, detected anomalies should correlate with local power outages and other infrastructural problems. Senegal's energy infrastructure is known to be unstable, and frequent power outages did occur in 2013. Unfortunately, we lacked a full record of these power outages and when and where they occurred. Such a list could be compared to the cell phone traffic and mobility within the affected areas during, after and even before the outage. It is possible that an influx of people into a region in a short or longer time window can be predictive of power outages occurring, as the infrastructure becomes overused. Cell phone towers do come equipped with backup power generators, so detecting a power outage by looking at cell phone data is not trivial. Having a ground truth of actual power outages, their location and duration would be extremely helpful in future work on this issue.

In summary, we computed a set of intuitive measures that capture and detect both the regularities and anomalies in human mobility patterns in Senegal. Such patterns provide important insights that would otherwise be too time consuming, expensive or even impossible to collect. It allows policy makers to optimally allocate infrastructural resources. Finally, it provides an early warning system for disruptive and catastrophic events allowing us to minimize the loss of human life and resources by a fast and informed response.

References

1. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. Plos Med 8:e10001083.
16. Wesolowski A, and Buckee CO (2013) Are gravity models appropriate for estimating the spatial spread of malaria?
National and Regional Road Network Optimization for Senegal Using Mobile Phone Data

Yihong Wang¹, Gonçalo Homem de Almeida Correia¹, and Erik de Romph∗¹,²

¹Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands
²DAT.mobility BV, P.O. Box 161, 7400AD Deventer, The Netherlands

Abstract

Due to the scarcity of mobility data in Senegal, mobile phone data provided by the D4D Challenge is used for optimization of the national and regional road network in Senegal.

We first applied a filtering algorithm to estimate inter-departmental origin-destination trip matrices (OD matrices) of sampled users in 2013. We name these matrices relative OD matrices, since we believe that they can reflect the mobility patterns in Senegal in a relative way.

Secondly, based on a literature study on the relations between travel and telecommunication, we explored such relations empirically by comparing the estimated relative OD matrices and the cell phone interaction matrices. The cell phone interaction matrices contain the number of calls and text messages of all Orange’s users, between pairs of departments. We found that the number of trips made by sampled users between each two departments is almost proportional to the number of cell phone interactions and inversely proportional to the travel cost between departments.

Thirdly, based on this observation, we constructed a new type of gravity model, based on the number of cell phone interactions instead of population where the traditional gravity model is usually based on. We estimated the parameters of this new model which gave us a model to predict elastic travel demand pattern for potential road network changes.

In the final step, we used this model to optimize the national and regional network for Senegal. We used an optimization model with the objectives of efficiency and equity. In the model two kinds of action can be performed: the construction of a new road of a given level; and the upgrading of an existing road to a higher level. A local search algorithm is used to find the solutions to this road network design problem.

We found that the created tool gained good insight into where and how to expand the Senegal network.

1 Introduction

The D4D Challenge provides anonymous data of Orange’s mobile phone users in Senegal for the study on several priority subject matters. Based on the advice from local authorities, some possible thematic issues regarding transport and infrastructure are listed on the website of the D4D Challenge. From those, we are inspired to select the specific topic of road network design in the interior regions.

When most people explore a country on Google maps, one of the components they would notice at first sight is the road network, which connects different parts of a country to satisfy travel demand. For long-term development, the government makes decisions on whether they should add new roads or upgrade the existing
ones to improve the level of service provided of roads. This task is especially important and urgent for the
government of Senegal, since it has been found that population growth is outstripping road development
there (World Bank 2004). In a less developed country like Senegal, it is particularly important to consider the
cost efficiency of road network planning due to the strong trade-off between increasing demand and budget
limitations. This goal can be achieved by road network optimization, which is regarded as one of the most
challenging transport topics (Yang, Bell, and G 1998), and in this case there is added complexity due to the
scarcity of mobility data in Senegal. Since travel demand is regarded as one of the most important issues
of an optimization-based road network design model (Santos, Antunes, and Miller 2009), before designing a
national and regional road network in Senegal, travel demand in this country should be investigated. The
D4D Challenge gives an opportunity to solve this kind of problem using mobile phone data.

In this context, the objectives of this research are to provide insights into how the mobility information
of a country can be derived from the mobile phone data and to advise decisions on national and regional
road network planning for a country based on the derived mobility information.

1.1 Original Mobile Phone Datasets

The datasets provided by the D4D Challenge are based on Call Detail Records (CDR) of phone calls and
text exchanges between more than 9 million of Orange’s customers in Senegal between January 1, 2013 to
December 31, 2013. Antenna-to-antenna traffic of calls and text messages of more than 9 million users for
1666 antennas on an hourly basis is provided as Dataset 1, which we name cell phone interaction data since
these data contain the information of interaction intensity (both number and duration of calls and number of
text messages) between two zones in the country. Dataset 2 provides fine-grained mobility data on a rolling
2-week basis for a year at individual level for about 300,000 randomly sampled users having more than 75% of
the days with interactions in one year. Once a user made a phone call or had a text message with others,
the location of the antenna to which this user connected at that time was recorded. Thus, his trajectory can
be captured over two weeks. Dataset 3 provides one year of coarse-grained inter-arrondissemental mobility
data at individual level for about 150,000 randomly sampled users having more than 75% of the days with
interactions in one year. Once a user made a phone call or had a text message with others, the arrondissement
where he connected to the antenna at that time was recorded. Thus, his trajectory can be captured over one
year. We call Dataset 2 and Dataset 3 as mobile phone traces. It should be noticed that the users presumed
to be machines or shared phone users are excluded in these datasets.

1.2 Spatial Data and Road Network Information

The D4D challenge provides the geographic information system (GIS) shapefile of Senegal, which contains
the information of administrative divisions of Senegal (arrondissement, department and region, ordered by
size from small to big). Among all kinds of administrative divisions, we focus on department, as spatial unit
used in this study. The population and area data of Senegal, collected per department, are found on the
website of the National Agency of Statistics and Demography (ANSD 2013).

The information of the road network in Senegal is found on the website of Digital Logistic Capacity
Assessment (2013). The roads in Senegal can be classified into five levels: national roads (N), regional roads
(R), department roads (D), urban way (VU) and classified tracks (P). National roads provide long distance
connections between several administrative regions and with neighboring states. Regional roads provide
connections between different departments of the same region. The other three levels of roads provide the
connections within the departments. The focus of this study is on the network of national and regional roads
which connect the different departments in the country.

A GIS layer of Senegal road network in 2002 is found on the website of ArcGIS provided by the D4D
challenge, including 1139 roads of different levels. The source of road network information is the Autonomous
Agency of Road Work of Senegal (AATR).

In addition, OpenStreetMap (OSM) is used as a layer in GIS software to show more details in the country,
and the latest road network information can be complementary to the layer of Senegal road network in 2002.
One of the most significant changes occurred in the meantime is the construction of a highway in Dakar, which was opened to traffic in two phases: the Patte d’Oie-to-Pikine section was opened first in 2011, followed by the Pikine-to-Diamniadio section on August 1st, 2013 (Eiffage 2013).

2 Literature Overview and Problem Description

The main problem addressed in this study is how to advise decisions on national and regional road network planning for Senegal with the best use of the cellphone data that the D4D challenge has provided us. To this end, we collect and review the related literature.

In this section, the literature about deriving mobility information from mobile phone information is reviewed at first, which suggests ways of exploring mobile phone traces (Dataset 3) for mining mobility information in Senegal. Some limitations of this method and their possible solutions are discussed. Secondly, a paper by Santos, Antunes, and Miller (2009) is reviewed to find the importance of elastic travel demand prediction for network optimization purposes using an unconstrained gravity model. It is discussed how to estimate such a gravity model in our study, and functional forms of gravity model are listed. In addition, the limitation of this kind of model is mentioned. Based on the literature review about the relations between telecommunication and travel, we discuss the possibility of using the cell phone interaction data (Dataset 1) as a proxy for current travel demand pattern or to predict elastic travel demand pattern. At the end of this section, a brief literature review of road network design problem is presented.

2.1 Origin-Destination Estimation Using Mobile Phone Traces

In Senegal, it was reported that the mobile use penetration passed 88% of the population in 2012 (Eto 2012). Along with the development of technology, it becomes possible for mobile phone to play a major role as a wearable sensor to collect data, especially the data representing the geographic locations of individual users (Ratti et al. 2006). To that extent, their trajectories can be traced over a period (Demissie, Correia, and Bento 2013), like what is included in Dataset 3. Due to both the fast expansion of market penetration and the availability of technology, many researchers have found the possibilities to derive travel demand for transport studies using mobile phone data, which are regarded as a game changer to build origin-destination trip matrices (OD matrices) (Caceres, Wideberg, and Benitez 2007; Nanni et al. 2014; White and Wells 2002; Calabrese et al. 2011a). This method could be quite efficient, compared with traditional methods like mobility surveys which are too costly, time-consuming and static.

However, there have always been limitations regarding this kind of estimated OD matrix. The first limitation is that the estimated OD matrix only includes the trips by sampled mobile phone users. To check whether the sample is biased, some researchers compared the density of mobile phone users’ homes and the density of population (Calabrese et al. 2011a). The second limitation is that some trips might be missed in this estimated OD matrix since the user may not use mobile phone during his trip. In most studies (Calabrese et al. 2011a; Hoteit et al. 2014), the sampling rate was examined to analyze the existence of this limitation. The third limitation is that a long trip could be divided into many partial trips in this estimated OD matrix. A possible solution which can be found in many studies is to focus only on the commuting trips (Nanni et al. 2014; Csáji et al. 2013).

Moreover, the OD matrix derived by mobile phone traces is often questioned about its validity, especially when the quality of data (e.g. sampling rate, penetration rate, etc.) is not good. There are two ways to validate the accuracy of the estimated OD matrix. One way is to check if the estimated OD matrix can fit well to a gravity model (Calabrese et al. 2011a; Csáji et al. 2013). No knowledge about the parameter values of the model are required. A high adjusted R-squared would mean the higher validity of the estimated OD matrix. The other way is to compare the estimated OD matrix with the available Census data from existing mobility surveys (Calabrese et al. 2011a). However, as mentioned before, this is not available for Senegal.
2.2 Traditional Gravity Model based on Population

The estimation of the current mobility patterns is not adequate, and a prediction of future travel demand is needed for planning. In the paper by Santos, Antunes, and Miller (2009), it was argued that in many cases of road network optimization, travel demand is assumed to be known in advance. However, this is a poor assumption since the addition of new arcs and the improvement of existing arcs will influence travel costs and thus change the distribution of existing trips and even create the new trips. Santos, Antunes, and Miller (2009) solved this problem by applying an unconstrained gravity model iteratively to predict the elastic travel demand for all possible solutions, responding well to different possible travel costs between each pair of two zones as they change with different networks. In such a case, it is simply assumed that population of different zones would not change in the future.

As known, in a gravity model concerning trip distribution, the number of trips between two zones should be proportional to a trip generation indicator (e.g. population) and inversely proportional to the travel cost between the zones (Dios Ortuzar, Willumsen, et al. 1994). A simplest version of the gravity model concerning trip distribution has the following functional form:

\[ T_{ij} = K_0 \frac{P_i P_j}{d_{ij}^2} \]  \[1\]

where, the scaling constant \( K_0 \) is the gravity constant for trip distribution, and \( T_{ij} \) is the number of undirected trips between two zones, and \( P_i \) and \( P_j \) are respectively the population of zone \( i \) and zone \( j \), and \( d_{ij} \) is the Euclidean distance between zone \( i \) and zone \( j \).

The model was further generalized by assuming that the effect of distance or ‘separation’ could be modeled more precisely by a cost function, which can be a function of distance or travel time or generalized cost between the zones (Dios Ortuzar, Willumsen, et al. 1994; McNally 2008). Also, the improvements included the use of total trip ends \( (O_i \text{ and } D_j) \) instead of total population. Due to the lack of information regarding trip ends, sometimes \( O_i \) and \( D_j \) can be replaced by a power function of the population (Csáji et al. 2013). The improved model can be written as (Dios Ortuzar, Willumsen, et al. 1994; Csáji et al. 2013):

\[ T'_{ij} = K_1 P_i^a P_j^b f(c_{ij}) \]  \[2\]

where, \( T'_{ij} \) is the number of directed trips between two zones, and \( a \) and \( b \) are the parameters for populations. \( f(c_{ij}) \) is the cost function. \( c_{ij} \) is the travel cost between \( i \) and \( j \). The travel cost \( c_{ij} \) can be distance or travel time or generalized cost. The popular versions for cost function can be classified into exponential function, power function and combined function, which are formulated respectively as follow (Dios Ortuzar, Willumsen, et al. 1994):

\[ f(c_{ij}) = e^{-\beta c_{ij}} \]  \[3\]

\[ f(c_{ij}) = c_{ij}^{-n} \]  \[4\]

\[ f(c_{ij}) = c_{ij}^n e^{-\beta c_{ij}} \]  \[5\]

where, \( \beta \) and \( n \) are the exponential parameter for cost function and the power parameter for cost function respectively.

One problem is that this kind of gravity model may not perform well all the time. It was found that if the distance between two zones is larger than 150 kilometers, the number of trips no longer depended on the actual distance (Csáji et al. 2013). Also, population census is always questioned regarding its accuracy, which might lead to the inaccuracy of the gravity model. Moreover, in this kind of gravity model, the social interaction between two zones is not taken into consideration. For example, imagine that two densely populated areas are close to each other, while the people in these two areas use different languages. It can be assumed that there would not be that many number of trips as the gravity model predicts.
2.3 The Relation between Telecommunication and Travel

We should not forget that communication is the basic function of a mobile phone. A mobile phone is able to record not only the trajectories of its own user, but also the interactions he makes, either by calls or by text messages, with others. Some researchers tried to explore the relations between telecommunication and travel. It was found again and again that there is a complementarity effect between telecommunication and travel (Mokhtarian 2002; Calabrese et al. 2011b; Kamargianni and Polydoropoulou 2013), especially at aggregate level (Plaut 1997; Calabrese et al. 2011c; Hsiao 2007). Our question is whether this kind of relation can help us understand more regarding the mobility pattern in Senegal.

It was found in a case study in Belgium that the total call duration between two zones was proportional to the product of population of two zones and that an inverse-square law decrease was found between the call duration and the distance, and a gravity model concerning the intensity of telecommunication was then estimated (Krings et al. 2009):

\[ I_{ij} = K_2 \frac{P_i P_j}{d_{ij}^2} \] \[ 6 \]

where, the scaling constant \( K_2 \) is the gravity constant for a timespan of 6 months of calling activity, and \( I_{ij} \) is the undirected communication intensity (total call duration) between two zones, and \( P_i \) and \( P_j \) are respectively the population of zone \( i \) and zone \( j \), and \( d_{ij} \) is the Euclidean distance between zone \( i \) and zone \( j \). It should be noted that this gravity model of communication intensity was fit to the reality in Belgium, where the distance between each two zones is not large.

This equation indicates that the intensity of telecommunication would not change in the future if we simply assume population would not change.

If we combine the gravity model regarding intensity of telecommunication and the simplest gravity model regarding trip distribution mentioned in the previous subsection (Eq. [1]), it results a linear relationship between the intensity of telecommunication and the number of trips between two zones:

\[ \frac{T_{ij}}{I_{ij}} = \frac{K_0}{K_2} \] \[ 7 \]

If this linear relationship holds true, it indicates that intensity of telecommunication between two zones could play a role as a proxy for current travel demand between two zones. However, it does not have any power to predict future travel demand, since we cannot predict changes in the cell phone interaction data.

However, it is obvious that if we combine this gravity model regarding intensity of telecommunication and an improved gravity model regarding trip distribution (e.g. Eq. [2]), the result would not be a constant. The ratio might be dependent on travel cost or population as well. Especially, the impedance of travel cost is much likely to be different for intensity of telecommunication and travel demand. It can be reasonably hypothesized that the relationship between intensity of telecommunication and travel demand is dependent on travel cost. If this is true, it would be possible to make a model based on intensity of telecommunication and travel cost to predict elastic travel demand, hence allowing applying our network design model for Senegal.

2.4 Road Network Design Problem

The network design problem is usually formulated as a bi-level problem, like a Stackelberg game, in which the network designer is the leader and the travelers are the followers (Snelder et al. 2007). The higher-level problem addresses the question of where new arcs should be constructed or which existing arcs should be upgraded. The lower-level problem concerns the estimation of demand in the network (Yang, Bell, and G 1998).

Regarding the lower-level problem, as mentioned in Section 2.2, elastic travel demand should be considered not only for trip distribution but also for traffic induction. Traffic assignment is usually made according to the user-equilibrium principle or 'all-or-nothing' principle.
Regarding the higher-level problem, the objective of network design problem is to optimize a given system performance measure. Some system performance measures can be: efficiency (to maximize the weighted average accessibility), robustness (to maximize the weighted reserve capacity of the network), equity (to limit the computation of accessibility to the zones with the lowest accessibilities) (Santos, Antunes, and Miller 2009) and environmental objectives (to minimize carbon monoxide emissions) (Cantarella and Vitetta 2006). In some studies, the total costs of road investments can also be the objective of the network design problem (Snelder et al. 2007). However, this is most considered as a constraint of network design problem (Yang, Bell, and G 1998).

Historically, the network design problems have two kinds of solution: a discrete form dealing with the additions of new arcs or roadway segments to an existing road network, and a continuous form dealing with the optimal service improvement of existing arcs (Yang, Bell, and G 1998). However, this classification has been challenged by a number of recent studies. Firstly, these two forms can be combined. To that extent, the existing arcs can be upgraded and the new arcs can be added at the same time (Santos, Antunes, and Miller 2009). In addition, it was argued that an important issue of the real-world road network planning is the multilevel discrete nature of service improvement. A discrete form dealing with the optimal service improvement of existing arcs or potential new arcs was suggested (Santos, Antunes, and Miller 2009). However, solving the problem of such discrete form is rather difficult, requiring heuristic methods (Yang, Bell, and G 1998).

3 Research Questions

The following research questions result from the objectives of this work and the literature review.

- Can cell phone interaction data be used as a proxy not only for current travel demand pattern but also to predict elastic travel demand pattern which is required for solving the lower-level network design problem?

  To answer this research question, some subquestions should be answered as well: What is the statistical relation between number of cell phone interactions and the estimated number of trips by sampled users empirically found in this study? Is this relation dependent or independent on travel costs?

- Which is the better model to predict elastic travel demand in this study, a predictive model based on cell phone interaction data (if any) or the traditional gravity model based on population?

  To answer this research question, some subquestions should be answered as well: What are the model performances of the respective models? Why do they perform different?

- What will be the optimal design of the national and regional road network for Senegal with regard to different objectives?

  To answer this research question, some subquestions should be answered as well: What will be the differences between the solutions towards different objectives? What will be the sensitivity of the solutions to a budget change?

4 Methodology

The methodology to answer the research questions in this study is illustrated in Figure 1. The main steps are listed as follow.

Firstly, we start from the four external arrows in Figure 1. We explore the census data, the GIS data and the original cell phone datasets respectively. The population of departments can be collected from the census data. Based on the current national and regional road network, the fastest path network analyses
are done under two scenarios, without and with the newly-opening Pikine-Diamniadio highway section, in Section 5.1. The skim matrices of shortest travel times between each two departments can be generated under these two scenarios. Then the cell phone interaction matrices of all users derived from Dataset 1 are aggregated at department scale for twelve months in 2013 in Section 5.2, and the estimated inter-departmental OD matrices of sampled users for twelve months in 2013, named as the relative OD matrices, can be derived from Dataset 3. Moreover, we examine the monthly fluctuations of the estimated mobility data. If there are no obvious seasonal fluctuations during the whole year, in order to apply the cross-validation technique for model validation afterwards, we classify the relative OD matrices into two groups as the training set and the test set, which are respectively the matrices under the first scenario (without the newly-opening Pikine-Diamniadio highway section, before August 1st, 2013) and the matrices under the second scenario (with the newly-opening Pikine-Diamniadio highway section, after August 1st, 2013).

We estimate a traditional gravity model based on population using the training set of the relative OD matrices, the shortest travel time calculated under the first scenario and population of each department. This part of the work is presented in Section 6.1. We explore the relations between the cell phone interaction matrices and the estimated relative OD matrices in Section 6.2. In Section 6.3, we determine if the cell phone interaction data can help us to predict elastic travel demand. Otherwise, we can only use the traditional gravity model based on population to solve the road network design problem. If it is proved that the cell phone interaction data can be used to predict elastic travel demand, we can furthermore build a new predictive model based on cell phone interaction and compare this model with the traditional gravity model based on population regarding their model performance of predicting the test set of the relative OD matrices.

After we have determined whether we use the traditional gravity model based on population or the new predictive model based on cell phone interaction to predict elastic travel demand in order to solve the road network design problem, we can start the work of road network planning. A detailed flowchart...
of methodology regarding road network planning is shown in Section 7.

5 The Exploration of Spatial Information and Mobile Phone Datasets

5.1 Network Analysis

In Figure 2, we can see that Senegal is divided into 45 different departments. The depth of color indicates the population of each department in 2013. The names of these departments and their codes are listed.

We clean the GIS layer of Senegal road network, including road shape and road information. First of all, only the national and regional roads in the network are kept and the lower levels are removed. We define department center as the traffic generation centroid of each department. In most of them we choose the capital of the department for generating the traffic. The nodes of the network are thus the centroids of departments or the intersections of roads. The separated links are merged if they are just part of the same road, and some low-level roads such as departmental roads are added only if they are necessary for inter-departmental connections. It should be noticed that roads are not extended to the foreign countries except the country of Gambia which is an enclave of Senegal. The newly-constructed Dakar-Diamniadio toll highway and the ferry service at the Banjul-Barra crossing point and at the Trans-Gambia crossing point are complementary to this network. Note that the Pikine-to-Diamniadio section was open on August 1st 2013.

In order to calculate shortest travel time, we should know average travel speed, which is influenced by speed limits, capacity and traffic volume to a large extent. Since we have limited knowledge about speed limits, capacity and daily traffic on these roads, 60 km/h, 45 km/h and 30 km/h are simply assumed as the average service speeds on national, regional and departmental roads respectively. We assume 80 km/h as the average service speed on Dakar-Diamniadio toll highway though in the reality this value is even higher. We do this to take in consideration the effects of the road toll. In addition, it is assumed that it would take people around 4.5 hours and 3.5 hours (including travel time, waiting time and effects of ferry tariff) to take...
the ferry services at the Banjul-Barra crossing point and the Trans-Gambia crossing point respectively.

The simplified road network in the country is shown in Figure 2. Based on this network, the Dijkstra’s Algorithm is applied to calculate the shortest travel time between each two departments. Due to the opening of the highway section in August, the calculation of shortest travel time is made without and with this section. The skim matrix of shortest travel time calculated without the section is used to estimate predictive models using the training set of the relative OD matrices, and the one calculated with the section is used to test how accurately the estimated models can predict the test set of the relative OD matrices.

5.2 Aggregation of the Cell Phone Interaction Data

As it was explained previously, the number and the total duration of calls as well as the number of text messages between each two antennas of all the Orange’s mobile phone users in 2013 are provided per hour. According to the document provided by the D4D Challenge (Montjoye et al. 2014), the number of Orange’s mobile phone users has reached 9 million, and the population of Senegal is about 13 million, yielding a penetration rate of nearly 70% of the population. We have no available information regarding how these 9 million users are distributed in the different departments, however, we can somehow regard this dataset as a persuasive sample of the interaction pattern in Senegal because of the considerable penetration rate.

We combine both the number of calls and the number of text messages as the intensity of telecommunication used in this study. Compared with the intensity of telecommunication used in Krings et al.’s study (2009), which was defined as the total call duration between two zones, our intensity of telecommunication is more comprehensive since it includes not only the interaction by calls but also by text messages. Moreover, we can simply assume that the duration per call is constant. To that extent, it is sufficient to use the number instead of the total duration to indicate intensity of interaction.

We aggregate the data and build the cell phone interaction matrix in which every cell presents the number of one-year calls and text messages from one department to another department. It can be observed that this directed matrix is rather symmetric. This means that the number of cell phone interactions aggregated in one year from department A to department B is almost equal to the one from department B to department A.

In the same way, twelve cell phone interaction matrices can be built respectively for each month in 2013. We firstly examine the total number of monthly cell phone interactions. It can be observed that the number of cell phone interactions increased in the second half of the year, mainly in August. However, the correlation coefficient between cell phone interaction matrices of each two months is higher than 0.99, which seems to indicate that the cell phone interaction pattern between departments in Senegal keeps almost the same.

5.3 Relative Origin-Destination Matrix Estimation by Tracing the Trajectories of Sampled Mobile Phone Users

5.3.1 Primitive Estimation

In Dataset 3, the traces are recorded at arrondissement scale. First of all, we can aggregate the traces at department scale. Then the consecutive traces at the same department of each user can be fused together. To that extent, every inter-departmental move of an individual user can be observed if it was detected that he used his cell phone in one department, and later he used his cell phone in another department. The number of inter-departmental moves can be recorded into an OD matrix, which reflects the inter-departmental movements of the sampled users. However, this primitive estimated OD matrix cannot reflect the mobility pattern of the whole population in Senegal, and it cannot even reflect the real mobility pattern of sampled users because of some limitations we have mentioned in Section 2. The solutions to those limitations in this case are given in the following sections.
5.3.2 Sampling Rate

We have estimated a primitive OD matrix by tracing the trajectories of each user. Once a user made a call or had a text message in a different department, it was indicated that he had already made a trip. This is the basic idea of a very simple algorithm to derive the movements of sampled users.

However, it could be argued that some trips would have been missed in this primitive estimated OD matrix since the user might not use mobile phone during those trips. In most of studies (Calabrese et al. 2011a; Hoteit et al. 2014), sampling rate was examined at first. If sampling rate is high enough, we can say there is enough evidence to only focus on the recorded trajectories. In addition, the acceptable value of sampling rate should be related with the area of the spatial unit. For example, the frequency of inter-departmental trips made by one person should be much lower than the frequency of his intra-urban trips. It can be said that the sampling rate in this study is high enough since we only focus on the inter-departmental trips and the sampled users are active enough (having more that 75% days with interactions in one year).

5.3.3 A Filtering Algorithm

Another problem is that in our primitive OD matrix, a long trip could possibly be divided into many partial trips. If a user passed by a department and used his cell phone there, this department should not be a real origin or a real destination. A possible solution which can be found in many studies is to only focus on the commuting trips (Nanni et al. 2014; Csáji et al. 2013). It was assumed in these studies that the place where a user was most frequently traced to stay and the place where he was second most frequently traced to stay should respectively be the location of his home and the location of his work. However, this method cannot capture non-work trips and the patterns of weekday and weekend as well as seasonal variations (Csáji et al. 2013). There is another problem in our study if we only focus on the commuting trips. Except the departments in Dakar region of which the area is relatively small, a department in Senegal has an area of larger than 1000 square kilometers. Since the typical size of a department is very large, it is quite possible that most people live and work in the same department. To examine if this problem exists in this case, we explore the data to find the locations of each user’s home and work. We find that the possibility of using mobile phones in the 'home departments' for about 75% sampled users is higher than 80%, which seems to indicate that most of inter-departmental trips in Senegal belong to irregular trips instead of commuting trips.

In this study, we make an attempt to filter the traces by using a threshold of least duration at one department plus travel time passing this department. If a user only stayed in one department for a very short time, it can be derived that this department should not be an origin or a destination and should be where he passed by. We can simply assume the least duration of one user at one department should be at least two hours. Based on these ideas, we apply an algorithm, of which the approach can be understood as follows:

At first, the consecutive traces at the same department of each user should be fused together, as mentioned previously. For every sampled user \( u \), the \( r \)th trace that he had is at department \( D_{ur} \), and the time he made the first interaction at \( D_{ur} \) is at \( FT_{ur} \), and the time he made the last interaction at \( D_{ur} \) is at \( LT_{ur} \). As assumed, the least duration of \( u \) at \( D_{ur} \) should be 2 hours. The shortest travel time between \( D_{u(r-1)} \) and \( D_{ur} \) is \( t(D_{u(r-1)}, D_{ur}) \), and the shortest travel time between \( D_{ur} \) and \( D_{u(r+1)} \) is \( t(D_{ur}, D_{u(r+1)}) \). It is simply assumed that user \( u \) always made phone calls and had text messages at department centers. Then, the interval \( FT_{u(r+1)} - LT_{u(r-1)} \) should be larger than the sum of least duration at department \( D_{ur} \) and \( t(D_{u(r-1)}, D_{ur}) \) and \( t(D_{ur}, D_{u(r+1)}) \).

Therefore, if \( FT_{u(r+1)} - LT_{u(r-1)} < t(D_{ur}, D_{u(r+1)}) + t(D_{u(r-1)}, D_{ur}) + 2 \) (unit: hour), the \( r \)th trace that the user \( u \) had should be removed from the traces since department \( D_{ur} \) is computed as where \( u \) passed by.

In this algorithm, \( u \in \{1, 2, ..., 160000\} \), \( r \in \{2, 3, ..., \} \), \( D_{ur} \in \{1, 2, ..., 45\} \), \( FT_{ur} \) and \( LT_{ur} \) are in the form of yyyy-mm-dd hh:mm:ss, the function of \( t(D_{u(r-1)}, D_{ur}) \) or \( t(D_{ur}, D_{u(r+1)}) \) is supported by the shortest travel time matrices calculated in Section 5.1.
It can be observed that this algorithm can solve the aforementioned problem efficiently. Moreover, we find that the algorithm is able to solve two more problems which are specific in this case to some extent. One problem is regarding sensing errors (there are some impossible traces in the original dataset, e.g. moving too fast). The other one is that some very short trips around the boundary between departments might have been sensitively recorded as inter-departmental trips.

After applying this algorithm, about 30% records are eliminated to improve the primitive estimated OD matrix.

5.3.4 Sample and Population

In Dataset 3, to guarantee the sampling rate of every user, only 0.16 million Orange’s users, who are the most active users, are sampled. The question is whether the movements of them can reflect the travel demand pattern of the whole population in Senegal. To answer this question, we follow the idea of Calabrese et al. (2011a) to compare the population distribution and the home location distribution of sampled users. It can be assumed that people often stay at home from 18 in the evening until 7 in the morning. The department where one user is traced most frequently during that night interval in the whole year is detected as the location of that user’s home. Therefore, the number of the sampled users’ homes of each department are known. A linear regression in log 10 scale is made to find the relationship between population and the number of sampled users’ homes of each department. In Figure 3, a power law increase with an exponent close to 0.5 can be observed. It is indicated that in the densely populated departments, there are relatively more active Orange’s users who are sampled, which prove that this sample is biased to some degree.

It was recommended by Nanni et al. (2014) that a factor which depends on the marketing penetration, cell phone ownership and cell phone usage can be weighed to make an estimate of actual traffic flow. However, it has been found that people who have more cell phone interaction with others will generate more trips (Kamargianni and Polydoropoulou 2013; Nobis and Lenz 2009). This conclusion makes us believe that simply multiplying by a factor would even make the calibrated traffic flow more biased. Moreover, based on this conclusion, we can somehow assume the sampled users are the most active travellers in each department since they are the most active cell phone users there. To that extent, the movements of sampled users are representative to some degree.

Since this sample is the best mobility data of Senegal which can be obtained, we can estimate the mobility
in Senegal based on nothing but this sample. However, it should be kept in mind that the estimated OD matrix records the movements of sampled users, which can somehow reflect the mobility pattern of the population in a relative way, or in other words, this estimated OD matrix cannot provide the actual traffic flow.

5.3.5 Modal Split

A specific problem in our case is that the estimated OD matrix may show the number of trips between departments by all possible modes, while we would like to only focus on the trips by road transport in this study. In a report by the World Bank (2004), it was said that road passenger share in Senegal was above 99%, and road freight share was above 95%. It can be confirmed that most of the inter-departmental trips are made by road transport.

5.3.6 Results and Analyses

As a result, a one-year relative OD matrix can be estimated, of which the symmetry is observed. In the same way, twelve relative OD matrices can be estimated for each month as well. The total number of estimated trips made by sampled users per month are calculated, and it can be observed that the fluctuation seems smooth. The correlation coefficient between relative OD matrices of each two months in 2013 is higher than 0.99, which indicates that the relative mobility pattern keeps almost the same during the whole year. As we know, the Pikine-to-Diambadié highway section was open on August 1st, 2013. It does not lead to any significant changes of overall mobility pattern in Senegal because this section is short compared to the total distance of road network and it is parallel to an existing national road. However, the travel demand induced by the newly-opened highway can be observed if we only focus on the subregion around this new section. As shown in Table 1, we focus on Dakar (1), Guediawaye (2), Pikine (3), Rufisque (4), Thies (5) and Mbour (6) and examine the "before-after" impact of the opening of new highway section on the average estimated number of trips made by sampled users per month between these departments. It can be observed that most of these numbers are increased, and especially, a sharp increase can be observed between Guediawaye and Mbour.

<table>
<thead>
<tr>
<th>OD Pair</th>
<th>Average Estimated Number of Undirected Trips between OD Pairs Made by Sampled Users per Month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before the opening of new highway section (From January to July)</td>
</tr>
<tr>
<td>1-4</td>
<td>54563</td>
</tr>
<tr>
<td>1-5</td>
<td>15023</td>
</tr>
<tr>
<td>1-6</td>
<td>14287</td>
</tr>
<tr>
<td>2-4</td>
<td>6219</td>
</tr>
<tr>
<td>2-5</td>
<td>1937</td>
</tr>
<tr>
<td>2-6</td>
<td>1444</td>
</tr>
<tr>
<td>3-4</td>
<td>87416</td>
</tr>
<tr>
<td>3-5</td>
<td>10806</td>
</tr>
<tr>
<td>3-6</td>
<td>8981</td>
</tr>
</tbody>
</table>

Table 1: The "Before-After" Comparison of Average Travel Demand

For the purpose of cross-validation, two relative OD matrices can be estimated respectively for the period before the opening of the Pikine-to-Diambadié highway section and the period after the opening of the highway section. In Section 6, the first matrix is used as a training set to fit models that can be used to predict travel demand pattern, while the second one is used as a test set to assess the predictive power of models.
6 Modelling

In the previous section, the relative OD matrices reflecting mobility pattern in Senegal are estimated. It is suggested that one of the best ways to validate them is to fit them to a gravity model and then to examine the fitness. Also, a gravity model can provide insights into the effect of the travel costs in the impedance to travel in the study area. To that extent, the gravity model can be used to predict the changes of future mobility pattern with the potential changes of travel cost.

6.1 Traditional Gravity Model Based on Population

A traditional gravity model based on population, which indicates that the mobility between two zones is almost proportional to the product of population of two zones and inversely proportional to the travel cost between two zones, is used to fit the training set of the relative OD matrices (for the period before the opening of new highway section), and the parameters of this model are estimated.

The Eq. [2] described in Section 2.2 is used as the functional form for traditional gravity model. Regarding the cost function, we follow the idea of Csáji et al. (2013): fitting both the power law decay and the exponential decay, to find the one that provides a better fit. The functional form can be either Eq. [8] or Eq. [9].

\[ T_{ij} = K_1 P_i^a P_j^b t_{ij}^{-n} \]  

\[ T_{ij} = K_1 P_i^a P_j^b e^{-\beta t_{ij}} \]

Where, \( i \) and \( j \) represent the department of origin and the department of destination. \( i \in \{1, 2, ..., 45\} \), and \( j \in \{1, 2, ..., 45\} \), and \( i \neq j \). \( P_i \) and \( P_j \) are population of origin and destination, and \( a \) and \( b \) are the parameters for population. The shortest travel time \( t_{ij} \) between \( i \) and \( j \), as calculated without new highway section, is used as the component of cost function. \( \beta \) and \( n \) are the exponential parameter and the power parameter for cost function respectively. \( K_1 \) is a scaling constant for a timespan of one month. It should be noticed that we fit the gravity model using our training set, which is the relative OD matrix for the period before the opening of new highway section. \( T_{ij} \) is not an exact number of trips between two departments in this case. Technically speaking, it is the average estimated number of directed trips made by the sampled users per month before August 1st, 2013, and it can reflect in a relative way the directed mobility between zones during that period.

After fitting the data to models, it is observed that the gravity model with the power parameter for cost function fits better to the training set than the one with the exponential parameter. The fitness of the better one is shown in Figure 4, and the estimated values of parameters and adjusted R-squared are listed in Table 2. The similar values of \( a \) and \( b \) indicate that the trips made by the sampled users are symmetric. The value of \( n \), 2.53015, is a reasonable one which can reflect the impedance of travel costs. The value of adjusted R-squared indicates that our training set somehow fits well to the gravity model. Moreover, it can be observed in Figure 4 that when the travel time between two departments is shorter, the model fits better. The observation in Csáji et al.’s research (2013) is reproduced. These results validates our estimation of relative OD matrix to a certain degree. The functional form of this estimated traditional gravity model based on population is given as follows:

\[ T_{ij} = (1.27e - 09) \times P_i^{1.07067} \times P_j^{1.08714} \times t_{ij}^{-2.53015} \]  

6.2 Exploring the Relation between the Cell Phone Interaction Data and the Estimated Mobility Data

To explore the relation between the cell phone interaction data and the estimated mobility data, we plot them in Figure 5, where y-axis represents the average estimated number of undirected trips between each two
Figure 4: The Fitness of Traditional Gravity Model Based on Population with the Power Parameter for Cost Function

Table 2: The Estimated Values of Parameters and Adjusted R-Squared

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log_{10} K_1$</td>
<td>-8.89705</td>
<td>0.34927</td>
<td>-25.47</td>
</tr>
<tr>
<td>$a$</td>
<td>1.07067</td>
<td>0.04305</td>
<td>24.87</td>
</tr>
<tr>
<td>$b$</td>
<td>1.08714</td>
<td>0.04305</td>
<td>25.25</td>
</tr>
<tr>
<td>$n$</td>
<td>2.53015</td>
<td>0.05034</td>
<td>50.26</td>
</tr>
<tr>
<td>Adjusted R-Squared</td>
<td></td>
<td></td>
<td>0.7299</td>
</tr>
</tbody>
</table>

departments made by sampled users per month before August 1st, 2013, and x-axis represents the average number of undirected cell phone interaction between each two departments made by all Orange’s users before August 1st, 2013, and after August 1st, 2013 in Figure 6. It should be noticed that the undirected estimated number of trips and the undirected cell phone interaction are used here instead of directed ones. This is because we assume the direction of cell phone interaction would not indicate anything regarding the direction of trips. However, in this case, using directed and undirected cell phone interaction or travel does not make any different since both the cell phone interaction matrix and the relative OD matrix are nearly symmetric. It can be observed again in these figures that both mobility pattern and interaction pattern stay unaltered after the opening of the new highway section. In addition, a power law increase with an exponent close to 1 can be observed, which indicates that there exists a close-to-linear relationship between mobility data and interaction data.

Figure 5: The Relation between Cell Phone Interaction Data and the Estimated Mobility Data before August 1st, 2013

Figure 6: The Relation between Cell Phone Interaction Data and the Estimated Mobility Data after August 1st, 2013
To test the hypothesis whether the relation between the cell phone interaction data and the estimated mobility data is dependent on travel cost, we indicate the travel time by color in Figure 5 and Figure 6. It can be observed that when the travel time between departments is shorter, the ratio between mobility data and interaction data is mostly higher. This observation proves that the hypothesis is true, and the mobility between two departments is almost proportional to the number of cell phone interactions and inversely proportional to the travel cost. To that extent, we can build a new form of gravity model by replacing the product of population of two zones with the number of cell phone interactions between these two zones.

### 6.3 New Gravity Model Based on the number of cell phone interactions

The functional forms of gravity model based on the number of cell phone interactions (with the power parameter and with the exponential parameter for cost function) are given as follow:

\[
T_{ij} = K_3 I_{ij}^\alpha t_{ij}^{-m} \tag{11}
\]

\[
T_{ij} = K_3 I_{ij}^\alpha e^{-\theta t_{ij}} \tag{12}
\]

Where, \(i\) and \(j\) represent the department of origin and the department of destination. \(i \in \{1, 2, ..., 45\}\), and \(j \in \{1, 2, ..., 45\}\), and \(i < j\). \(I_{ij}\) is the average number of undirected cell phone interaction per month. \(\alpha\) is the parameter for \(I_{ij}\). Since we have observed a close-to-linear relationship between mobility data and interaction data, we assume \(\alpha\) would be estimated as about 1. \(t_{ij}\) between \(i\) and \(j\), as calculated without the new highway section, is used as the component of cost function. \(\theta\) and \(m\) are the exponential parameter and the power parameter for cost function respectively. \(K_3\) is a scaling constant for a timespan of one month. \(T_{ij}\) is not an exact number of trips between two departments in this case. Technically speaking, it is the average estimated number of undirected trips made by sampled users per month before August 1st, 2013, and it can reflect in a relative way the directed mobility between zones during that period.

In the same way as we did in Section 6.1, we fit the training set to the new forms of gravity model based on the number of cell phone interactions. The fitness of two models are shown in Figure 7 and Figure 8, and the estimated values of parameters and the adjusted R-squared are illustrated in Table 3 and Table 4.

![Figure 7: The Fitness of New Gravity Model Based on Cell Phone Interaction with the Power Parameter for Cost Function](image1)

![Figure 8: The Fitness of New Gravity Model Based on Cell Phone Interaction with the Exponential Parameter for Cost Function](image2)
The fitness of both the new gravity models looks better than the fitness of the traditional gravity models, and the values of the adjusted R-squared are higher. Especially, as observed in Figure 4, the traditional gravity model is not fit well when the travel time between two departments is large. By contrast, it can be observed that this problem does not exist when fitting the new gravity models.

Despite the higher adjusted R-squared value of fitting new gravity model with power parameter, it can be observed in Figure 7 that the model overestimates the mobility between departments with the highest mobility. Therefore, we choose the new gravity model with the exponential parameter as the one to be compared with the traditional gravity model regarding their model performance. The functional form of this estimated new gravity model based on the number of cell phone interactions is illustrated as follows:

$$T_{ij} = 0.00493 \times I_{ij}^{1.00108} \times e^{-0.3497 \times t_{ij}}$$  \[13\]

The value of $\alpha$, 1.00108, indicates that when travel time is same, mobility between departments is proportional to the number of cell phone interactions. It should be noticed that since the new gravity model based on the number of cell phone interactions is trained using estimated relative OD matrix, $T_{ij}$, hence what this model can predict is a relative value, and actual traffic flow cannot be predicted. Since $T_{ij}$ represents a relative value, the constant, 0.00493, is not important in this functional form.

### 6.4 The Comparison between Two Gravity Models

Two different estimated gravity models are used to predict the mobility pattern after August 1st, 2013. This test set is used to assess which model has a greater predictive power. The comparison of their model performance of predicting the test set is shown in Figure 9 and Figure 10. It can be observed that the traditional gravity model based on population does not perform well especially when the travel time is higher. Root Mean Square Error (RMSE) is used as the indicator to test model performance by comparing observed values and predicted values. The undirected mobility data in the test set are used as observed values, and we transfer the directed travel demand predicted by the traditional gravity model to the undirected one in order to be compared with the undirected travel demand predicted by the new gravity model. As calculated, RMSE of using the traditional gravity model based on population is 157229.3, while RMSE of using the new gravity model based on the number of cell phone interactions is only 5590.4. As a result, we choose the new gravity model based on the number of cell phone interactions, which performs much better, in order to support the decisions on road network design.

From our point of view, there are some possible reasons why the new gravity model based on the number of cell phone interactions performs better:

- The cell phone interaction data are more reliable, more precise and more updatable than the population census data.
- The cell phone interaction data can reflect the social interaction between zones, which population cannot reflect.
7 Road Network Planning

In the previous section, we have estimated a new gravity model based on the number of cell phone interactions, which can be used to solve the lower-level problem of road network planning regarding travel demand distribution. In this section, based on the estimated new gravity model and the cell phone interaction data we have, national and regional road network planning is made for Senegal.

7.1 Planning Approach

The approach to road network planning in this study follows the main principles listed below:

- Planning decisions include adding new links of given levels or upgrading existing links to higher levels.
- Efficiency is the main objective, and equity is taken into consideration as well.
- Construction costs of adding and upgrading links should not exceed the budget.
- Travel demand is elastic with road network design.

The flowchart of planning approach is illustrated in Figure 11.

Firstly, we start from the three external arrows. We use the road network with newly-opened Pikine-Dianniaudio highway section as the current road network, as shown in Figure 2, including five types of road which are: toll highway, national road, regional road, departmental road and ferry connection. We have assumed different levels of average service speed on these different types of road, and shortest travel time can be calculated based on these assumed speeds. To upgrade existing links, we improve road levels by improving the corresponding speed levels since speed is the only design characteristic of road levels to be considered in this case. In fact, besides speed, the capacity of roads should also be considered as an important design characteristic. However, in this case, as stated in Section 5.3.4, actual traffic flow cannot be estimated, and we can only estimate a relative mobility pattern in Senegal, not to mention the scarcity of the information regarding capacity of roads in Senegal. Therefore, we have no knowledge about the ratio
between traffic flow and capacity on each road, and thus we only use assumed average speeds, instead of capacity, as the design characteristic to indicate different service levels of different road types.

Apart from upgrading existing links, adding new links is also considered. The potential links should be determined. If neighboring departments are not well connected, a potential link is made straightly between them unless there are physical barriers (e.g. mountains, forests, etc.). All potential links are shown in Figure 12. For solving road network design problems, the average service speeds of these potential links are assumed as zero. In this case, highway is considered as the supreme level of all road types since in recent years there are more projects regarding construction of new highway in Senegal. We assume that the average speed in a highway is 80 km/h, same as the assumed average speed on toll highway in Dakar. All road types can be upgraded to highway level. In addition, regional roads can be upgraded to national roads, and departmental roads can be upgraded to regional or national roads. Potential links can be added as regional or national road or highway. It is noteworthy that we consider in this planning whether the two ferry services should be replaced by bridges. Because of long waiting time and limited capacity of ferries, it can be supposed that those ferry services are mobility bottlenecks. Thus we assume that ferry connections can be upgraded to bridges, and we assume that the average service speed on bridges is the same as the one on national road, 60 km/h. The average service speeds of each road level and the relative unit costs for road construction and upgrading are shown in Table 5. We take the relative unit costs used in the study by Santos, Antunes, and Miller (2009) as a reference for determining the ones used in our study.

The best assignment of 2171 monetary units (which represents 10% of the total budget required to construct all potential links as highway and to upgrade all existing links to the highest level, assumed as the available budget in this case) is determined to improve the existing road network. Under the budget constraint, there are still millions of solutions regarding how to add and upgrade links. Different solutions would lead to different new networks, which result in new shortest travel times between departments. With these potential network changes, we apply our estimated new gravity model based on the number of cell

Figure 11: The Flowchart of Methodology Regarding Road Network Planning
phone interactions to predict elastic travel demand pattern, in terms of a relative OD matrix which is meant to reflect as best as possible predicted mobility pattern. Travellers are assumed to follow the fastest paths, travelling at the average service speeds consistent with the road levels of the links included in their routes.

We assess the solutions with regard to efficiency and equity objectives. Regarding efficiency, we use the maximization of the accessibility of centers in the country as the measure. According to Santos, Antunes, and Miller’s study (2009), accessibility was defined as (proportional to) the spatial interaction between the center and all other centers, and the typical expression used to calculate weighted average accessibility based on traditional gravity model is given as follows:

$$Z = \sum_{i \in N} A_i \times \frac{P_i}{P} \quad \text{and} \quad A_i = \sum_{j \in N \setminus i} P_j \times f(c_{ij})$$  \hspace{1cm} [14]$$

Where, $Z$ is the measure of weighted average accessibility. $N$ is the set of traffic generation centers. $P$, $P_i$ and $P_j$ are the total population of country, the population of center $i$ and the population of center $j$. $A_i$ is the accessibility of center $i$. $c_{ij}$ is the travel cost between center $i$ and center $j$, such as distance or travel time. $f(c_{ij})$ is the cost function which reflects the impedance of travel cost. The cost function estimated in the gravity model can be used here directly.

Since the new gravity model based on the number of cell phone interactions is applied in this case, and in the estimated model, mobility between departments is proportional to the number of cell phone interactions and inversely proportional to travel cost, a new expression used to calculated weighted average accessibility

<table>
<thead>
<tr>
<th>Existing Level</th>
<th>Upgraded Level</th>
<th>Potential</th>
<th>Departmental</th>
<th>Regional</th>
<th>National</th>
<th>Highway</th>
<th>Bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential</td>
<td>0 km/h</td>
<td>–</td>
<td>–</td>
<td>1.2</td>
<td>1.6</td>
<td>4</td>
<td>–</td>
</tr>
<tr>
<td>Departmental</td>
<td>30 km/h</td>
<td>–</td>
<td>–</td>
<td>0.2</td>
<td>0.6</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>Regional</td>
<td>45 km/h</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.4</td>
<td>2.8</td>
<td>–</td>
</tr>
<tr>
<td>National</td>
<td>60 km/h</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>2.4</td>
<td>–</td>
</tr>
<tr>
<td>Ferry</td>
<td>0-2 km/h</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 5: Design Characteristic of Different Road Levels and Relative Unit Costs for Road Construction and Upgrading
based on the new gravity model is adapted as follows:

\[ Z = \sum_{i \in N} A_i \times \frac{1}{2I} \text{ and } A_i = \sum_{j \in N \setminus i} I_{ij} \times f(c_{ij}) \]  \[15\]

Where, \( I \) is the total number of undirected cell phone interaction between all pairs of departments per month. \( I_{ij} \) is the undirected number of cell phone interactions between department \( i \) and department \( j \) per month. \( f(c_{ij}) \) is the cost function estimated as a component of the new gravity model based on the number of cell phone interactions. In this case, the expression of this cost function can be derived from Eq. [13]:

\[ f(c_{ij}) = e^{-0.3497c_{ij}} \text{ and } c_{ij} = t_{ij} \]  \[16\]

Where, \( t_{ij} \) is travel time between department \( i \) and department \( j \).

In this case, the accessibility measure of the efficiency objective is rather compatible with the new gravity model which can only predict relative mobility pattern, since the functional form of the accessibility measure based on the new gravity model does not necessarily incorporate a scaling factor, or in other words, the actual travel flow between each OD pairs is not required for calculating this measure.

Regarding equity, we use the maximization of accessibility for the centers with the lowest accessibility in the country as the measure. The expression is given as follows (Santos, Antunes, and Miller 2009):

\[ E = \sum_{i \in N_{low}} P_i \times A_i \text{ and } A_i = \sum_{j \in N \setminus i} P_j \times f(c_{ij}) \]  \[17\]

Where, \( E \) is the measure of equity. \( N_{low} \) is the set of centers with lowest accessibility. In this case, we focus on the 20% of department centers with the lowest accessibility.

Also, we can adapt this equation based on our new gravity model based on cell phone interaction in this case:

\[ E = \sum_{i \in N_{low}} A_i \text{ and } A_i = \sum_{j \in N \setminus i} I_{ij} \times f(c_{ij}) = \sum_{j \in N \setminus i} I_{ij} \times e^{-0.3497t_{ij}} \]  \[18\]

We choose efficiency as the unique objective at first, and a best solution to achieve this objective can be found. Then we take equity objective into consideration, giving different weights to accessibility and equity, leading to different solutions. Afterwards, all the solutions to achieve different objectives can be compared.

Since this road network design problem is non-linear, the optimal solutions are difficult to be found without using heuristic methods. In this study, a local search algorithm is applied to help us find the best solutions efficiently. In every iteration, a new solution is generated through the local search algorithm. We compare the new solution assessed in each iteration with the best existing solution obtained in previous iterations. Once the new solution is better than the existing best solution, it becomes the existing best solution, and if it is found that the existing best solution cannot be improved any more, the iteration will stop.

### 7.2 Optimization Model

To accomplish the approach explained previously, an optimization model should be solved in each iteration. This model is illustrated as below:

\[ \max V = w_Z \times \frac{Z(y) - Z_0}{Z_B - Z_0} + w_E \times \frac{E(y) - E_0}{E_B - E_0} \]  \[19\]

subject to:

\[ Z(y) = \sum_{i \in N} \sum_{j \in N \setminus i} I_{ij} \times e^{-0.3497 \times t_{ij}(y)} \times \frac{1}{2I}, \quad \forall i, j \in N \quad (i \neq j) \]  \[20\]
\[ E(y) = \sum_{i \in N_{\text{low}}} \sum_{j \in N \setminus i} I_{ij} \times e^{-0.3497 \times t_{ij}(y)}, \quad \forall i, j \in N \quad (i \neq j) \tag{21} \]

\[ \sum_{m \in M_l} y_{lm} = 1, \quad \forall l \in L \tag{22} \]

\[ \sum_{m \in M_l} e_{lm} \times y_{lm} \leq b \tag{23} \]

\[ T_{ij} \geq 0, \quad \forall i, j \in N, \quad l \in L, \quad y_{lm} \in \{0,1\}, \quad l \in L, \quad m \in M \tag{24} \]

Where \( V \) = normalized value of a solution; \( w_Z \) and \( w_E \) = weights attached to efficiency and equity objectives; \( Z \) and \( E \) = values of a solution in terms of each objective (which are not scalable); \( Z_B \) and \( E_B \) = best values obtained for each objective in previous iterations; \( Z_0 \) and \( E_0 \) = worst values obtained for each objective in previous iterations; \( I_{ij} \) = the number of cell phone interactions between department \( i \) and department \( j \); \( t_{ij} \) = the shortest travel time between department \( i \) and department \( j \), which is dependent on \( y_{\{l\}} \); \( y_{\{l\}} \) = matrix of binary variables equal to one if link \( l \) is set at road level \( m \) and equal to zero otherwise; \( N \) = set of departments; \( L \) = set of links; \( M_l \) = set of possible road levels for link \( l \); \( e_{lm} \) = cost of setting link \( l \) at road level \( m \); and \( b \) = budget.

The objective function (Eq. [19]) of this optimization model is set to maximize the normalized value of the road network planning solution. The weights \( w_Z \) and \( w_E \), which can reflect the relative importance of accessibility and equity objectives, are given to the normalized values of the solutions. The values of the solutions are normalized using the range of variation of solutions. The values of the solutions \( Z \) and \( E \) are essentially dependent on the decisions made regarding road levels which are expressed as \( y \). The constraints Eq. [20] and Eq. [21] are the expressions of accessibility and equity based on new gravity model, which have been explained in the previous subsection. The constraint Eq. [22] is used to guarantee that each link should be set at only one level. The constraint Eq. [23] is used to guarantee that the cost should not exceed the available budget. Expressions Eq. [24] gives the domain for each decision variable.

### 7.3 Solution Algorithm

In this study, a local search algorithm is used to find the best solution. For solving non-linear problems, a local search algorithm generates a new solution based on the current solution by applying a transformation to the current solution in every iteration. This method can prevent from exhaustively searching the entire space of possible solutions (Michalewicz and Fogel 2004). The key to apply a local search algorithm is to find how a transformation can be applied to the current solution in a specific case. Santos (2009) introduced a specific local search algorithm for road network design problem. This algorithm includes three procedures: add, interchange and drop, which are three ways to transform the solutions. The add procedure starts with the initial network and selects the one-level upgrade link change that improves the objective measure most in successive iterations. The interchange procedure starts with the add solution and selects the combination of one-level upgrade and downgrade link changes that improves the objective most. The drop procedure starts when no further accessibility increase is possible, and it downgrades the links which are previously upgraded by one level.

According to Santos’s evaluation (2009), this local search algorithm performs decently, and the computation time is rather short compared with other algorithms. Especially when the number of links in the network is more than 100, the solution quality becomes better, and the computation time is much shorter than the computation time of other algorithms. In this case, the number of links in the network is 107, which is one of important reasons for us to choose this local search algorithm to solve the problem.
7.4 Results for a Single Efficiency Objective

Firstly, we only consider efficiency objective by setting $w_Z$ as 1 and setting $w_E$ as 0. The new network of the best solution is shown in Figure 13, and all the links which are upgraded in this solution are highlighted in Figure 14. It can be observed that three main lines of national roads originated from Dakar are suggested to be upgraded to highway in this planning solution. Along these three lines, the existing Dakar-Diamniadio highway could be extended to Dagana (24), Mbacke (8) and Bignona (34) respectively. Especially, the line extended to Bignona passes through Gambia, where Trans-Gambia ferry service is on the way. In this planning solution, a bridge is suggested to be built to replace the ferry service. The connection between Tivaouane (7) and Bambey (10) and the connection between Fatick (12) and Foundiougne (13) are found as the most important regional roads for the accessibility in the country. Thus, they are suggested to be upgraded to national roads. Moreover, a link is suggested to be added between Thies (5) and Mbour (6), and the national road between Pikine (3) and Rufisque (4), which is parallel to the newly-opened Pikind-Diennniadio highway section, is suggested to be upgraded to highway. The departmental connection between Guediawaye (2) and Pikine (3) is suggested to be upgraded to highway as well. All the links suggested to be upgraded are in the western part of Senegal, where the departments are more densely populated.

In this planning solution, the value of efficiency measure $Z$ increases by 6.548% from the value of the current network.

7.5 Impact of Adding an Equity Objective

If efficiency is the only objective considered for road network planning, this would lead to the improvement of roads next to the centers where travel demand is higher. To that extent, the dissimilarities between large and small centers’ welfare will be potentially increased. For sustainable development, Santos, Antunes, and Miller (2008) takes equity issue into account in road network planning. We believe that this is also an important issue in Senegal.

As mentioned in planning approach, we choose the accessibility to low-accessibility centers as our equity measure. Firstly, we give the full weight to equity objective by setting $w_Z$ as 0 and setting $w_E$ as 1. The best solution is depicted in Figure 15 and Figure 16. It can be observed that three main lines upgraded to highway radiate from Tambacounda (39). Two potential links are suggested to be added as national roads between Medina Yoro Foulah (27) and Bounkiling (32) and between Kedougou (36) and Salemata (38). The existing departmental link between Kedougou (36) and Saraya (37) is suggested to be upgraded to national roads.
roads. A bridge is suggested again to be developed to replace the Trans-Gambia ferry service. This is the only same link change in the two different planning solutions for different objectives. Most links suggested to be upgraded to achieve the equity objective are in the southeastern part of Senegal, where the departments are less populated.

Figure 15: New Network of the Optimal Solution to Achieve the Single Equity Objective

Figure 16: All Upgraded Links of the Optimal Solution to Achieve the Single Equity Objective

In this planning solution, the value of equity measure $E$ increases by 21.758% from the value of the current network.

However, it is not possible for government to only consider the equity objective since the solution to achieve the equity objective is not a good one for the efficiency objective. Therefore, to make a trade-off between the different objectives, the different weights are usually given to them. In this case, we include the efficiency objective and the equity objective, assigning equal weights (0.5) to them. The best solution obtained is depicted in Figure 17 and Figure 18.

Figure 17: New Network of the Optimal Solution to Achieve both Efficiency and Equity Objective

Figure 18: All Upgraded Links of the Optimal Solution to Achieve Both Efficiency and Equity Objective

It can be observed that this planning solution includes the improvement of roads both in the eastern
part of Senegal, where the departments are not populated, and in the western part, where the departments are populated. It is noteworthy that the Trans-Gambia ferry service is suggested again to be replaced by a bridge.

From the values of assessment measures of the current network, the value of equity measure $E$ increases by 18.341%, and the value of efficiency measure $Z$ increases by 3.537%.

### 7.6 Sensitivity Analysis

To test the sensitivity of the solutions to a budget reduction, the budget level is considered as 50% of the initial budget for the single efficiency objective and for the objective of 50% efficiency and 50% equity.

Under budget constraint of 1086 monetary units, the best solution for the single efficiency objective is depicted in Figure 19 and Figure 20. From the values of assessment measures of the current network, the value of equity measure $Z$ increases by 4.644%, and the value of efficiency measure $E$ increases by 1.608%.

Under the budget constraint of 1086 monetary units, the best solution for both efficiency and equity objectives is depicted in Figure 21 and Figure 22. From the values of assessment measures of the current network, the value of equity measure $E$ increases by 8.988%, and the value of efficiency measure $Z$ increases by 2.158%.

In Table 6, the increase of the assessment measure values from the current measure values are presented under different scenarios (for different objectives and under different budget constraints). It can be observed that the reduction of budget has less impact on the efficiency measure than on the equity measure. In other words, the increase of efficiency measure slows down with the increase of budget, and on the other hand, there is still much room for improvement of the equity of road network in Senegal, which explains why the use of budget is sensitive to the increase of equity measure.

The Trans-Gambia ferry service is suggested to be replaced by a bridge in all the planning solutions not only for the efficiency objective but also for the equity objective under different budget constraints. Thus, we are not surprising to find that the construction of a bridge has been planned for a long time, though the plan has not come to fruition (Wikipedia 2013). In addition, the Dakar-Diamniadio highway is suggested to be extended to Thies (5) and Mbour (6) in most of the planning solutions, which is exactly similar with what the government of Senegal is planning as the phase 2 of the Dakar Toll Road Project (ADBG 2014). The consistency between the model results and the reality validates the model to a certain degree.
Figure 21: New Network of the Optimal Solution to Achieve Both Efficiency and Equity Objective Given 1/2 Budget

Figure 22: All Upgraded Links of the Optimal Solution to Achieve Both Efficiency and Equity Objective Given 1/2 Budget

<table>
<thead>
<tr>
<th>Solution</th>
<th>Measure</th>
<th>Budget 50%</th>
<th>Budget 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>For the Single Efficiency Objective</td>
<td>Z (Efficiency)</td>
<td>4.644%</td>
<td>6.548%</td>
</tr>
<tr>
<td>For Both Efficiency and Equity Objective</td>
<td>Z (Efficiency)</td>
<td>2.158%</td>
<td>3.537%</td>
</tr>
<tr>
<td></td>
<td>E (Equity)</td>
<td>8.988%</td>
<td>18.341%</td>
</tr>
</tbody>
</table>

Table 6: The Increase of Assessment Measure Values From the Current Measure Values Under Different Scenarios

8 Conclusions

In this study, based on the cell phone interaction data and the mobile phone traces that the D4D Challenge has provided us, we find that the mobility between departments is proportional to the aggregated number of cell phone interactions between departments and inversely proportional to the travel costs between departments in Senegal. To that extent, using the filtered mobile phone traces, we estimate a new gravity model based on the number of cell phone interactions, and compare it with the traditional gravity model based on population regarding the model fitness and the predictive accuracy. Because of the better model fitness and the stronger predictive power, the estimated new gravity model based on the number of cell phone interactions is used to solve the lower-level problem of the national and regional road network planning in Senegal. Under the assumed budget constraints, we select the efficiency and the equity as the objectives of solving this network design problem by giving them different weights, and we adapt the functional forms of the efficiency measure and the equity measure, which are originally based on traditional gravity model, to the version based on the new gravity model. The model results show a consistency with some potential plans for the roads and the bridges in the near future which have been announced by the government.

We believe that the methodology presented in this study have possible uses for development in the following aspects:

- The filtering algorithm introduced in this project can be used to filter the mobile phone traces and thus to improve the OD estimation.
- The empirically found relation between telecommunication and travel, and the new gravity model based
on cell phone interactions, allow the government to better understand and predict mobility patterns in Senegal.

- The optimization model based on the new gravity model can help the government to make better decisions on national and regional road network planning using mobile phone data. Based on the actual planning goal, the government can determine the weights of different objectives and the actual available budget in the model by themselves, in order to obtain the best solution under a certain scenario.

In this study, we use the mobile phone traces to derive the mobility information in Senegal as best as possible, and furthermore regard them as the ground truth to find the relationship between telecommunication and travel and to estimate the gravity models. Even though we apply a filtering algorithm to improve the OD estimation, it may still be questioned whether the filtered traces can represent the real mobility of people, and some people might furthermore argue that the strong relationship between telecommunication and travel that we find is a result of the fact that we estimate the mobility information using mobile phone data. Nevertheless, these questions cannot be answered without additional traffic information. Therefore, we recommend that the government can use additional traffic information, such as road counts and mobility survey data, to validate the estimated relative OD matrices and the estimated gravity models.

References


Building workers’ travel demand models based on mobile phone data

Feng Liu\textsuperscript{a}, Davy Janssens\textsuperscript{b}, JianXun Cui\textsuperscript{c}, Geert Wets\textsuperscript{b}

\textsuperscript{a,b}Transportation Research Institute (IMOB), Hasselt University, Wetenschapspark 5, bus 6, B-3590, Diepenbeek, Belgium
\textsuperscript{c}Department of transport engineering, Harbin Institute of Technology (HIT), 1500, Harbin, China

\textsuperscript{a}Corresponding author: Tel: +32 0 11269125 fax: +32 0 11269199

E-mail addresses: feng.liu@uhasselt.be (F. Liu), davy.janssens@uhasselt.be (D. Janssens), cuijianxun@hit.edu.cn (J.X. Cui), geert.wets@uhasselt.be (G. Wets)

Abstract

Daily activity-travel sequences of individuals have been estimated by activity-based transportation models. The sequences serve as a key input for travel demand analysis and forecasting in the region. However, the high cost along with other limitations inherent to traditional travel data collecting methods has hampered the models’ further advancement and application, particularly in developing countries. With the wide deployment of mobile phone devices today, we explore the possibility of using mobile phone data to build such a travel demand model.

Our exploration consists of four major steps. First, home, work and other stop locations for each user are identified, based on their mobile phone records. All the obtained locations along with their particular orders on a day are then formed into stop-location-trajectories and classified into clusters. In each cluster, a Hidden Markov Model (HMM) is subsequently constructed, which characterizes the probabilistic distribution of activities and their related travel of the sequences. Finally, the derived models are used to simulate travel sequences across the entire employed population.

Using data collected from natural mobile phone usage of around 9 million users in Senegal over a period of one year, we evaluated our approach via a set of experiments. The average length of daily sequences drawn from the stop-location-trajectories and the simulated results is 4.55 and 4.72, respectively. Among all the 677 types of the stop-location-trajectories, 520 (e.g. 76.8\%) are observed from the simulated sequences, and the correlation of sequence frequency distribution over all the types between these two sequence sets is 0.93. The experimental results demonstrate the potential and effectiveness of the proposed method in capturing the probabilistic distribution of activity locations and their sequential orders revealed by the mobile phone data, contributing towards the development of new, up-to-date and cost-effective travel demand modelling approaches.

Keywords activity-travel sequences, Hidden Markov Model, activity-based transportation models, travel surveys, mobile phone data.
1. Introduction

1.1. Activity-based transportation models
The main premise of activity-based transportation models is the treatment of travel behavior as a derived demand of activity participation. In this modeling paradigm, travel is analyzed through daily patterns of activity behavior related to and derived from the context of land-use and transportation network as well as personal characteristics such as social-economic background, lifestyles and needs of individuals (e.g. Bhat & Koppelman, 1999; Davidson et al., 2007; Wegener, 2013).

All the above information, complemented with a training set of household travel surveys which record the full daily activity-travel sequences of a small sample of individuals during one or a few days, is analyzed and translated into heuristic decision making rules, using machine learning techniques, e.g. decision trees (e.g. Arentze & Timmermans, 2004; Bellemans et al., 2010). These rules represent the scheduling process of activities and travel by the individuals. Once established, the activity-based models can be used as the probabilistic basis for a micro-simulation process using Monte Carlo methods, in which complete daily activity-travel sequences for each individual in the whole region are synthesized. The synthesized sequences are then aggregated into travel measures, e.g. the average number of trips or travel distances per day, or an origin-destination (OD) matrix. The OD matrix represents the number of trips between each pair of locations of the region, and it can be assigned to a road network through traffic assignment algorithms. The derived travel measures as well as the amount of travel assigned to specific roads can subsequently serve as essential input for travel analysis in the region, such as travel demand forecasting, emission estimates, and the evaluation of emerging effects caused by different transport policy scenarios. Fig. 1 illustrates the entire process of an activity-based transportation model.

![Fig. 1. The entire process of an activity-based transportation model](image)

1.2. Problem statement
Despite comprehension and advancement of activity-based transportation models, e.g. Albtross (Arentze & Timmermans, 2004), TASHA (Roorda et al., 2008), Feathers (Bellemans et al., 2010), the availability of household travel surveys has been a prerequisite condition for the model building, regardless of the following drawbacks of the data collection method (e.g. Asakura & Hato, 2006; Cools et al., 2009). (i) The entire survey is a lengthy process; from the initial data gathering to data cleaning and the exploitation of the first results, it could take months even years, causing a time lag between the data initially obtained and the results that are required for objective and up-to-date activity-travel behavior analysis. (ii) It imposes a significant burden on respondents, resulting in low response rates and under-reporting of short trips. (iii) Despite the above disadvantages, the data is very expensive to collect, leading to
only a limited number of respondents and a (or a few) day(s) being involved in the surveys. Consequently, this tends to obfuscate the less frequent activities, such as sports or telecommuting activities which are often carried out once a week or once a month. Questions are also raised about the capability of such limited sample size in representing activity-travel behavior of a whole population.

Apart from travel surveys, travel information has also been gathered from sensors, e.g. loop detectors and video cameras, which are installed in a road network to monitor traffic flow. However, the sensors are usually set up on highways, as it is expensive to instrument a whole region with such static devices. Consequently, the collected data is only limited to the high-capacity roads, and sheds little light on the traffic flow in the rest of the area (e.g. Gühnemann et al., 2004).

Due to the data constraints, the existing methods on travel behavior analysis and travel demand modeling are restricted to only a (or a few) statistical average day(s) and a relatively small region as well as to a subset of the population, because of the lack of a large dataset that is spatially and temporally extensive as well as involves more individuals. Consequently, the results are difficult to be generalized to evaluate travel demand in various types of days (e.g. weekdays, weekend and holidays) and at a higher geographical scale (e.g. an entire city or a whole country). For a long time, data problems have been one of the essential challenges of the current research on travel demand modelling. The problems have seriously hampered further development and application of the existing techniques (e.g. Hartgen, 2013; Janssens et al., 2012). Having accurate, reliable, while affordable travel data for the estimation of travel demand and the subsequent analysis on transport network systems has thus been a major concern, particularly in developing countries.

1.3. Mobile phone data: a new data source for travel demand modelling

The wide deployment of mobile phones has created the opportunity to use the devices as a new data collection method to overcome the lack of reliable travel data (Jiang et al., 2013). Location data recorded from mobile phone devices reflects up-to-date travel patterns on a significantly large sample of a population, making the data a natural candidate for the analysis of mobility phenomena in the region (e.g. Do & Gatica-Pereza, 2013; Schneider et al., 2013). In addition, the data collection is a by-product of mobile phone companies for billing and operational purposes that generates neither extra expenses nor respondent burden.

The importance and added value of mobile phone data in the field of transportation research have been manifested by a variety of studies, ranging from the investigation of key dimensions of human travel, such as travel distances and time expenditure at different locations (e.g. González et al., 2008; Schneider et al., 2013; Song et al., 2010), to the discovery of typical mobility patterns (e.g. Bayir et al., 2009; Berlingerio et al., 2013; Calabrese et al., 2011), and to the examination of the status and efficiency of current transport network systems (e.g. Angelakis et al., 2013; Steenbruggen et al., 2013). Particularly, mobile phone data has been employed to explore the possibilities of building travel demand models, e.g. OD matrices (e.g. Becker et al., 2011; Calabrese et al., 2011; Shan et al., 2011). The research by (Shan et al., 2011) can represent the typical process of such exploration. The study utilizes mobile phone data of more than 0.3 million users collected in the metropolitan area of Lisbon, Portugal for an entire month. In this process, the two most frequent call cell towers for each of the users are first identified as the residential and employment locations, respectively. Using the two obtained locations, an OD matrix depicting home-to-work commuting trips in the morning is then built. Based on a census survey, this derived OD matrix is subsequently scaled up to account for the total employed population of 1.3 million in the study area. The adjusted matrix is ultimately used to compare against the travel demand during the same morning period forecasted by an integrated land use and transportation model.
developed in this region. The results show comparative performance of this OD matrix in estimating the morning travel demand in this region.

However, despite its advancement by incorporating mobile phone data into the modeling process, the OD-based method does not consider the sequential information which is imbedded in activity-travel patterns. A detailed analysis of the sequential dependencies of the daily activities from activity-travel behavior is thus ignored in the modeling process. It has been widely acknowledged that the choice of activities is dependent on the preceding activity engagement (e.g. Joh et al., 2008; Wilson, 2008), exemplified by the fact that, during one particular working day, it is highly probable that the combination of having breakfast, travel and working is observed together. On the contrary, if a sports activity is carried out in the morning, there is a small chance that it is performed again in the evening. The interdependencies of daily activities have been considered as a crucial factor in the activity-travel decision making process (e.g. Delafontaine et al., 2012; García-Díez et al., 2011). A modeling process, which takes into account the sequential information and generates activity-travel sequences that are consistent with the sequential constraints observed from real travel behavior, is thus important. The existing activity-based models have integrated the sequential information of daily activities into the modeling process. But as previously described, the activity-based models are constructed based on a small set of activity-travel sequences from travel surveys, thus subject to the shortcomings that are inherent to the traditional data collection methods. A model, which is based on massive mobile phone data while taking into account the sequential aspect of activity-travel behavior, has so far been lacking.

1.4. Research contributions
Extending the current studies on the application of mobile phone data to transportation research, and particularly addressing the above mentioned limitations in the development of travel demand models, our study proposes a new approach which is based on the phone data and considers the sequential information imbedded in activity-travel patterns. Specifically, this study is to build a workers’ travel demand model based on mobile phone data using Hidden Markov Modeling (HMM) techniques. The derived model characterizes the probabilistic distribution of activities and their related travel on a day among workers. The models can be used to simulate new activity-travel sequences across the whole employed population. The synthesized sequences can be subsequently aggregated into certain travel measures which serves as important input for travel demand analysis in the region.

Compared to existing activity-based models, this approach offers the following advantages. (i) This method is built upon the observed current activity-travel behavior of a large proportion of population, thus providing a more representative and up-to-date modeling process. (ii) Through a long period of mobile phone data records, inter- and intra- personal variations of travel behavior as well as weekday, weekend and seasonal deviations are captured. (iii) The use of mobile phone data generates no extra financial cost in terms of data collection, making it a cost-effective approach. This is particularly practical in developing countries where, as stated before, the high cost of traditional travel data collection mechanisms combined with other disadvantages of the methods have deterred the much needed development of a new, effective and cheaply realized travel demand modelling technique. With the use of the large-scale mobile phone data, the proposed method can be regarded as a reality mining approach which places the realized trips of travellers in daily life directly at the centre of the analytical process. (iv) When this method is compared with the OD-based modeling approach, the OD-based method analyzes travel behavior in terms of the distribution of all individual trips over different pairs of origin-destination locations; it is an aggregated modeling process. While the approach developed in this study examines the entire activity-travel sequences and focuses on the sequential aspect of travel behavior. In this new
approach, the locations which are accessed by an individual on the same day are viewed and tackled as a whole, rather than an isolated participation in activities. Both methods analyze activity-travel behavior from different perspectives, thus providing a complementary means of modeling travel demand based on mobile phone data. In addition, while the OD-based approach is just an end product of the observed behavior from the phone data, and reflects the current mobility phenomena; the model proposed in this study is able to predict travel demand in regions where no phone data is provided or in future scenarios, e.g. the displacement of residential areas or the establishment of new industrial sites.

The remainder of this paper is organized as follows. Section 2 introduces the mobile phone data and Section 3 details the proposed modeling approach. A case study is conducted in Section 4, and a comparison of the modeling results against the data in the validation set is carried out in Section 5. Finally, Section 6 ends this paper with major conclusions and discussions for future research.

2. Mobile phone data description

The mobile phone dataset consists of full mobile communication patterns of around 9 million users in Senegal between January 1, 2013 to December 31, 2013 (de Montjoye et al., 2014). The dataset contains the location and time when each user conducts a call activity, including initiating or receiving a voice call or text message, enabling us to reconstruct the user’s time-resolved call location trajectories. The locations are represented with the identifications of base stations (cells) in a GSM network; the radius of each of the stations ranges from a few hundred meters in metropolitan to a few thousand in rural areas, controlling our uncertainty about the user’s precise location. Despite the low accuracy of users’ exact locations, the massive mobile phone data represents a significant percentage (i.e. 69%) of this country’s total population, providing a valuable source and opportunity for the analysis on human travel behavior and for drawing relevant inferences that can be statistically sound and representative. In order to address privacy concerns, the original dataset has been split into consecutive two-week periods. In each period, users are randomly selected and assigned to anonymized identifiers. New random identifiers are chosen for re-sampled users in different time periods. The data process results in totally 25 randomly sampled datasets, each of which contains communication records of 300,000 users over two weeks. One of these datasets is selected for this study. Table 1 illustrates typical call records of an individual identified as user20 on Thursday, January 24th, 2013.

<table>
<thead>
<tr>
<th>Time</th>
<th>Cell_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:57:00</td>
<td>751</td>
</tr>
<tr>
<td>13:40:00</td>
<td>749</td>
</tr>
<tr>
<td>16:59:00</td>
<td>177</td>
</tr>
<tr>
<td>17:43:00</td>
<td>751</td>
</tr>
<tr>
<td>21:28:00</td>
<td>751</td>
</tr>
</tbody>
</table>

3. Methodology

3.1. Overview of the approach
The method is composed of 4 major steps. (i) Home, work and other stop locations for each user are identified, based on their mobile phone records. (ii) The obtained location trajectories are clustered according to the travel features encoded in the sequences. (iii) In each cluster, a Hidden Markov Model is constructed, which characterizes the probabilistic distribution of the corresponding sequences. (vi) The obtained models are used to simulate activity-travel sequences across the whole employed population in the study region. The overall structure of the approach is shown in Fig. 2, and the detailed procedure is elaborated as follows.
3.2. Home, work and other stop location identification

3.2.1. Mobile phone call location trajectories

A call location trajectory from a mobile phone user during a day, i.e. call-location-trajectory, is defined as a series of locations where the user makes calls when traveling or doing activities, as the day unfolds. It can be formulated as a sequence of $l_1 \rightarrow l_2 \rightarrow \ldots \rightarrow l_n$, where $n$ is the length of the sequence, i.e. the total number of locations that the user has travelled to when making calls that day, and $l_i$ ($1 \leq i \leq n$) is the identification of the locations, e.g. cell IDs in this study. At each $l_i$, there could be multiple calls $k_i$ ($k_i \geq 1$), referred as call-frequency; the time for each of the calls is denoted as $T(l_i,1), T(l_i,2), \ldots, T(l_i,k_i)$, respectively. The time interval between the first and the last call time in the set of consecutive calls, i.e. $T(l_i,k_i) - T(l_i,1)$, is defined as call-location-duration. Accommodating the time signatures of the multiple calls, a call-location-trajectory can be represented as $l_1(T(l_1,1),T(l_1,2),\ldots,T(l_1,k_1)) \rightarrow \ldots \rightarrow l_n(T(l_n,1),T(l_n,2),\ldots,T(l_n,k_n))$, simplified as $l_1(T(1),T(2),\ldots,T(k_1)) \rightarrow \ldots \rightarrow l_n(T(1),T(2),\ldots,T(k_n))$. Given the above call-location-trajectories constructed from the mobile phone data, the home and work locations are first predicted. This is followed by the identification of stop locations where activities are carried out.

3.2.2. Prediction of home and work locations

Various methods have been proposed to derive home and work locations from mobile phone data, mainly based on the visited frequency of a location during a particular time period (e.g. Becker et al., 2011; Calabrese et al., 2011). However, different time windows have been specified in these methods, depending on the context of the study area. In this study, a similar approach is adopted, but the time windows are empirically estimated from the mobile phone data as follows. The time period when call activities start to increase considerably in the morning during weekdays is chosen as the work start time, denoted as work-start-time. Similarly, the moment when the second peak of call activities start to appear in late afternoon is considered as the work end time, referred as work-end-time. Around this time, it is assumed that people start to communicate for off-work activity engagement.

Based on these two temporal points, a location is defined as the home location if it is the most frequent stop throughout the weekend period as well as during the night-time interval on weekdays between work-end-time and work-start-time. On the contrary, a location is considered as a work place if it satisfies the following criteria. (i) It is the most common place for call activities in the perceived work period between work-start-time and work-end-time on weekdays. (ii) It is not identical to the previously identified home location for the user. (iii) The calls at the location are not limited in only one day, they should occur at least 2 days a week.
With the above-defined identification criteria, we assume that people have only one home location and at most one work location. The additional locations, which are occasionally accessed for home or work activities, are regarded as a stop for non-mandatory activities. In addition, only individuals, who work in areas different from their home locations and who work at least two days per week, are included for the analysis of workers’ travel behavior.

3.2.3. Identification of stop locations

After the identification of the distinct home and work locations for each user, the remaining locations in the call-location-trajectories are either stop-locations where people pursue non-mandatory activities or non-stop-locations. Each of these non-stop-locations can be further divided into either a trip-location where the user is traveling, or a false-location that is wrongly documented due to location update errors. The location update errors normally occur when call traffic is busy in the user’s real location area, and consequently this location is shifted to less crowded cells for short time periods, causing location area updates, without the users’ actual moving (e.g., Calabrese et al., 2011).

In addition, for the identified home or work locations, some occurrences of the locations could also be caused by non-stop reasons, e.g., people travelling in the same area as their home locations when making calls. Therefore, each location occurrence in the call-location-trajectories will be classified into stop-locations and non-stop ones, regardless its activity type.

The scenarios, where the two types of non-stop-locations could occur, can be illustrated with the call records of two typical users. The trajectory from the first user, identified as user265, is \( l_1(17:06,17:43) \rightarrow l_3(17:51) \rightarrow l_4(17:56,19:41) \rightarrow l_d(21:55) \), where 4 locations are observed, with the call-location-duration as 37, 0, 105 and 0 min respectively. Each of these locations needs to be identified as either a stop visit or just a passing-by place. The trajectory of the second user, i.e. user72, is \( l_1(13:21,20:11) \rightarrow l_2(22:00) \rightarrow l_3(22:02) \rightarrow l_4(22:05) \rightarrow l_2(22:07,23:12) \). This user has 5 location updates, with the call-location-duration as 410, 0, 0, 0 and 65 min respectively. It should be noted that the time interval between the first and second visit to location \( l_2 \) is only 7 min. Although there is a possibility that this user may have travelled at a high speed during this period, the temporary interruption of \( l_2 \) by the extra locations \( l_3 \) and \( l_4 \) is such a short interval is most likely resulted from the location update errors. Consequently, locations \( l_3 \) and \( l_4 \) are falsely connected to the user’s mobile phone at 22:02 pm and 22:05 pm although he/she had been actually remaining at location \( l_2 \) during this period.

In order to identify the stop-locations, the approach proposed in the study (Liu et al., 2014) is used, which consists of the following steps. (i) For each location \( l_i \), the call-location-duration is first examined. If it is longer than a certain time limit, denoted as \( T_{\text{call-location-duration}} \), this location is considered as a stop-location. (ii) Otherwise, if the condition does not hold (e.g., only a single call made at \( l_i \)), and if the location appears in the middle of a daily sequence of \( n \), i.e. \( 1 < i < n \), a second parameter, namely \( \text{maximum-time-boundary} \), defined as the time interval between the last call time at \( l_i \)’s previous location and the first call time of its next location, i.e. \( T(l_{i+1},l_i) - T(l_{i+1},l_{i-1}) \), is computed. If this time period is longer than a threshold value, defined as \( T_{\text{maximum-time-boundary}} \), \( l_i \) is perceived as a stop visit. (iii) When \( l_1 \) is in the first or last position of a trajectory and the call-location-duration is shorter than \( T_{\text{call-location-duration}} \), there is no sufficient information to estimate maximum-time-boundary for this visit. Thus, all the distinct locations, where the user has stayed at least once for conducting an activity over the entire survey period, are collected. These locations are considered as potential stop locations that are on the user’s daily activity agenda and that are visited either routinely or once in a while. If \( l_i \) is one of these locations, it is assumed to be a stop for activity purposes. In contrast, if \( l_i \) is the place where the individual has not been observed doing activities, it is
then considered as a passing-by place or being recorded as a localization error and therefore removed.

After the removal of locations that are either trips or stem from localization errors, all the remaining locations from a call-location-trajectory are regarded as stops and formed a stop-location-trajectory. Based on the above described identification process, if a duration of 30 and 60 min are used for $T_{\text{call-location-duration}}$ and $T_{\text{maximum-time-boundary}}$ respectively, as set up in our experiment described in Section 4, the obtained stop-location-trajectories for user265 and user72 are $l_1 \rightarrow l_3 \rightarrow l_4$ and $l_1 \rightarrow l_2$ respectively.

3.3. Stop-location-trajectory classification

Each location $l_i$ in the previously obtained stop-location-trajectories is complemented with its function, denoted as activity($l_i$), categorized into home, work and non-mandatory activities, represented as ‘H’, ‘W’ and ‘O’, respectively. While H and W encapsulate all activities performed at home and work (including school) places respectively; O refers to all activities undertaken outside home and work places, differentiated between maintenance activities (e.g. shopping, banking or visiting doctors) and discretionary activities (e.g. social visits, sports or going to restaurants) (e.g. Arentze & Timmermans, 2004). Travel is implicit in between each two consecutive locations of the sequences.

Various methods have been used to classify activity sequences, mainly based on either a priori scheme or a numerical distance measure. A priori scheme aims to cluster the sequences according to predefined variables, e.g. socio-demographic factors of respondents or activity-travel features of the sequences. For example, researches (Spissu et al., 2009) first extract activity sequences of all employed people and then divide the sequences into HWH, HOH, HOWH, HWOH and HWOWH, depending on whether non-mandatory activities are involved, and if so, on when these non-mandatory activities are conducted. This classification method provides a simple way to build the clusters and to analyze the correlation between the behavior of each cluster and the socio-demographic characteristics of the corresponding individuals. Numerical distance measure methods, on the other hand, classify activity-travel sequences based on some measures of distances between the sequences, such as the number of identical activities (e.g. Roorda & Miller, 2008) or the similarities of the activities and their sequential order derived using sequence alignment methods (SAM) (e.g. Joh et al., 2008; Saneinejad & Roorda, 2009).

In this study, the stop-location-trajectories are classified based on the travel features of the sequences, i.e. the number of home based tours on the days. Two types of home-based tours, including home-based-work-tour and home-based-non-work-tour, are defined as a chain of locations (trips) that starts and ends at home and accommodates at least one work or one non-mandatory location visit, respectively. Based on this definition, a stop-location-trajectory for a working day can be classified into 1-home-based-work-tour (e.g. HWH), 2-home-based-work-tours (e.g. HWHWHH), or 3 (or more)-home-based-work-tours (e.g. HWHWHHH), referred as 1_HBWT, 2_HBWT or 3_HBWT, respectively. While for a non-working day, the trajectory can be assigned into 1-home-based-non-work-tour (e.g. HOH), 2-home-based-non-work-tour (e.g. HOH), or 3 (or more)-home-based-non-work-tour (e.g. HOHOHOH), namely 1_HBNT, 2_HBNT or 3_HBNT, respectively. Apart from the above 6 classes, the weekday days when an individual does not make any trips are characterized into an additional class, represented as the single letter of H.

Given a group of users along with the distances between the home and work locations of the individuals, referred as $d$, their stop-location-trajectories can be attributed to the above corresponding classes. The relative frequencies of the trajectories in each of the 7 clusters over the total number of the sequences, in each particular range of distance $d$, is referred as
characterizes the observed probabilities of the sequences in each tour class with respect to the home-work distances.

3.4. Hidden Markov Model construction

3.4.1. Model configuration

A pHMM is a probabilistic representation that can capture statistical relevant information implicit in a group of related sequences. It was introduced into bio-informatics in the 1990s (Krogh et al., 1994) and has since been widely used for large-scale protein sequence analysis (e.g. Finn et al., 2014). The information extracted from a group of sequence includes: (i) a sequence of positions, each with its own distribution overall all possible letters; (ii) the possibility for either skipping a position or inserting extra letters between consecutive positions.

In this study, the HMM building process for the two classes, including 1_HBWT and 1_HBNT, are described. The similar process applies to the remaining tour classes including 2_HBWT, 3_HBWT, 2_HBNT and 3_HBNT.

A HMM for the 1_HBWT class is designed as follows (see Fig. 3). It divides a sequence into four different parts, including: (i) before-going-to-work sub-sequences which represent the activities and travel undertaken before leaving home to work, e.g. HOH; (ii) commute sub-sequences which account for the activities and travel pursued during the home-to-work and work-to-home commutes respectively, e.g. HOW or WOH; (iii) work-based sub-sequences which accommodate all activities and travel conducted from work, e.g. WOW; (iv) after-work sub-sequences which comprises the activities and travel engaged after arriving home from work, e.g. HOH.

Based on the above segmentation of the sequences, a total of 8 states is defined, including the start home, work and end home locations, defined as m1, m2 and m3 respectively, and the other stop locations corresponding to each part of the sequences, defined as m1,1, m1,2, m2,1, m2,2 and m3,l, respectively. Each of these states can emit an letter, i.e. x, from all possible types of x governed by a distinct emission probability distribution, defined as p_{em}(x|state).

At each of the states, maximum 3 possible transition probabilities $\pi$s are assigned to describe the likelihood of movement between each two connected states as follows. (i) Transitions linking state $m_k$ ($k=1, 2$) to the other 3 possible states, including: to state $m_{k-1}$, i.e. $\pi(m_{k-1}|m_k)$, when a trip is made in the morning before going to work ($k=1$) or at noon during work period ($k=2$); to state $m_{k-2}$, i.e. $\pi(m_{k-2}|m_k)$, when an activity is conducted during the commuting way from home to work ($k=1$) or from work to home ($k=2$); to state $m_{k+1}$, i.e. $\pi(m_{k+1}|m_k)$, when no stops occur on the commuting ways from home to work ($k=1$) or from work to home ($k=2$). (ii) Transitions from state $m_3$ to only a state $m_{3,l}$, i.e. $\pi(m_{3,l}|m_k)$, when a trip is made in the evening after coming back from work. (iii) Transitions from state $m_{k,l}$($k=1, 2, 3$) to state $m_k$, i.e. $\pi(m_k|m_{k,l})$, when the person returns back home after finishing all activities outside in the morning or in the evening ($k=1$ or 3), or when the person returns to work after finishing activities outside at noon ($k=2$); or to itself, i.e. $\pi(m_{k,l}|m_{k,l})$, when an extension of multiple activities is done in the respective periods. (iv) Transitions from state $m_{k,2}$($k=1, 2$) to state $m_{k+1}$, i.e. $\pi(m_{k+1}|m_{k,2})$, when all the activities are finished on the commuting way from home to work ($k=1$) or from work to home ($k=2$); or to itself, i.e. $\pi(m_{k,2}|m_{k,2})$ when an extension of multiple activities is done on the commute trips.
Apart from the above 8 states for stop locations, an additional \textit{End} state is added to the end of the model, allowing transitions from \(m_3\) to the end of the sequence; the corresponding transition probability is defined as \(\pi(\text{End}|m_3)\).

![Diagram](attachment:image-url)

**Fig. 3.** The HMM for a home-based-work-tour

The above-defined model configuration thus turns the home-based-work-tours into a network system of a set of states. States \(m_k\) (\(k=1, 2, 3\)) underline the basic structure of the sequences, i.e. the home and work locations, while the introduction of the remaining states accommodates the situation where activities are conducted at different periods that are formed based on the home and work places. The transition probabilities \(\pi\) reveal the intensity of the conversion between different states (situations). Alongside the transition probabilities, the model also accommodates the emission probability of letter \(x\) at each state, i.e. \(p_{\text{em}(x|\text{state})}\). In the current study, variable \(x\) represents the type of different activities; however, it can also be used to characterize other dimensions of the sequences, e.g. travel start time, distances and travel modes, thus capable of modeling multiple aspects of activity-travel behavior.

Fig. 4 illustrates the HMM for the 1\_HBNT class. It has only 3 states, including the states for start and end home locations, i.e. \(m_1\) and \(m_2\), respectively, and the third one, i.e. \(m_{1,2}\), representing locations for non-work activities conducted during the home-based tour.

![Diagram](attachment:image-url)

**Fig. 4.** The HMM for a home-based-non-work-tour
3.4.2. Model parameter estimation

After the model structure is defined, the next step involves the estimation of the specific parameters including the transition probabilities and emission probabilities. The probabilities \( \pi \) and \( p_{\text{emit}}(x|\text{state}) \) can be obtained by the observed frequencies of the letters at the corresponding periods of the sequences (e.g. Durbin et al., 1998). Let \( A(r|q) \) as the frequencies of the transitions from a state, denoted as \( q \), to another state, denoted as \( r \), and \( E(x|\text{state}) \) as the frequencies of letter \( x \) at state \( \text{state} \), respectively. The estimators for the parameters are given by the following formula.

\[
\pi(r|q) = \frac{A(r|q)}{\sum_r A(r|q)}, \quad p_{\text{emit}}(x|\text{state}) = \frac{E(x|\text{state})}{\sum_r E(x'|\text{state})}
\]

Where, \( x, x' \in \{\text{set of all letter types at the state}\} \).

In the parameter estimation process, a pseudocount is set, which is a small value added to \( A(r|q) \) or \( E(x|\text{state}) \) if the instances of the corresponding observed cases are zero. This is to adjust the probability of rare but not impossible events so that the events are not completely excluded. The relative values of pseudocounts represent the prior knowledge on the expected probabilities of the corresponding events.

3.5. Monte Carlo simulation

3.5.1. The whole process of the simulation

Using the constructed HMMs and the distance between the home and work locations of an individual, the Monte Carlo method can be used to generate a new sequence. Monte Carlo simulation is a process that approximates solutions to quantitative problems, e.g. determining the properties of some phenomenon or behavior, through repeated statistical sampling. In this process, the investigated system is simulated a large number of times; for each simulation, all of the uncertain parameters in the system are sampled according to their respective probabilistic distribution. The simulation results are a large number of separate and independent realizations, each representing a possible “future” for the system. The results can be used for subsequent statistical analysis on the properties of the system.

In the simulation process, we first generate a tour class according to the probabilistic distribution characterized in the distance-based-tour-class-distribution. From this selected class, an entire daily sequence for this individual is then simulated based on the HMM derived from the specific class. The detailed simulation procedure based on the HMM for 1_HBWT class is described in the following section; a similar process can be applied to other classes using the respective models.

3.5.2. HMM simulation

Given distance \( d \) and the HMM as demonstrated in Fig. 3, the new sequence, i.e. \( s \), is generated as follows. (1) Sequence \( s \) is initiated by the start home activity at state \( m_1 \) (i.e. \( s=H \)). (2) The next state is decided among the three states of \( m_{11}, m_{12} \) and \( m_2 \), according to the corresponding transition probabilities of \( \pi(m_{11}|m_1), \pi(m_{12}|m_1) \) and \( \pi(m_2|m_1) \). (3) If \( m_{11} \) is chosen, activity \( x \) emitted from probability distribution \( p_{\text{emit}}(x|m_{11}) \) is added to the sequence (i.e. \( s=Hx \)). At this state, a next transition needs to be chosen between going back to \( m_1 \) (i.e. \( s=HxH \)) or continuing on this state (i.e. \( s=Hxxx \)), based on \( \pi(m_{11}|m_1) \) and \( \pi(m_{12}|m_1) \) respectively. If the latter situation is selected, the loop at \( m_{11} \) continues until a transition to the
home location at \( m_1 \) occurs (i.e. \( s=Hxx..xH \)). (4) If \( m_{12} \) is selected, \( x \) is added to the sequence (i.e. \( s=Hx \)). At \( m_{12} \), a new transition is decided to either move to \( m_2 \) (i.e. \( s=HxW \)) or remain on this state (i.e. \( s=Hxx \)), governed by probabilities \( \pi(m_2|m_{12}) \) and \( \pi(m_{12}|m_{12}) \) respectively. The remaining on this state continues until a transition to \( m_2 \) is chosen (i.e. \( s=Hx..xW \)). (5) If \( m_2 \) is selected, activity \( W \) is added to the sequence (i.e. \( s=HW \)). (6) The similar procedure described in steps 2-5 is repeated for next states including \( m_2 \) and \( m_3 \), using the corresponding transition probabilities. The simulation process finally stops when the transition from \( m_3 \) to the \textit{End} state of the model is realized based on \( \pi(\text{End}|m_3) \).

4. Case study

In this section, adopting the proposed approach and using the mobile phone dataset described in Section 2, we carry out a case study. In this process, a set of stop-location-trajectories for workers are first identified. The corresponding individuals are then randomly divided into two parts with the ratio as 4 to 1, for model training and validation, respectively. From the training set, the stop-location-trajectories are classified; in each cluster, a HMM is constructed. Based on the derived HMMs, new activity-travel sequences for individuals in the validation set are simulated.

4.1. Stop-location-trajectory construction

4.1.1. Work-start-time and work-end-time

Fig. 5 describes the distribution of the frequencies of calls made in each hour of the weekdays, showing that from 8am in the morning, calls start to increase considerably and reach their peak at noon; while at 20pm in the evening, a second climax of call activities starts to occur. These two morning and evening temporal points are chosen as the work-start-time and work-end-time, respectively.

<table>
<thead>
<tr>
<th>Frequencies (%)</th>
<th>Call time (hour)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
</tr>
</tbody>
</table>

*Fig. 5. The distribution of the time of calls*

Based on the pre-defined criteria for home and work location identification, 319,492 users (i.e. 99.9% of the total users in the mobile phone dataset) have their home locations discovered. The remaining 0.1% are those who made no calls at weekend or in the night period from 20pm to 8am across the two surveyed weeks. As a result, their homes cannot be spotted by these rules. Meanwhile, 89,643 users are screened out as employed people, if they work between 8am and 20pm at least two weekdays per week. By contrast, those who work in the same location as their homes, who work at night shifts or at weekends, who work less than
two days a week, or who make few calls at work, are left out. Although the final obtained workers account for only 28.1% of the total users in the selected dataset, they represent the part of population who regularly travels to work during the day time period among weekdays, thus they are an important target group for travel behavior analysis and transport network management. All the 7,897,854 call records of these individuals during weekdays are extracted, and the consecutive calls made at a same location are aggregated. This reduces the records to 3,479,532 locations. The locations for a same user on a same day are linked according to the temporal order, resulting in total 781,817 call-location-trajectories that will be used for further analysis.

4.1.2. \( T_{\text{call-location-duration}} \) and \( T_{\text{maximum-time-boundary}} \)

For each location in the call-location-trajectories, a distinction must be made between stop-locations and non-stop ones which include trip- and false-locations. Two parameters characterize this identification process. The first one \( T_{\text{call-location-duration}} \) defines the minimum time interval at a location, above which the location is considered as a possible stop. The other parameter \( T_{\text{maximum-time-boundary}} \) estimates the total time that is required to travel from the previous cell to the current one and from the current one to the next cell. In addition, it should also be able to detect location update errors which usually occur in a short time interval. In this experiment, \( T_{\text{call-location-duration}} \) and \( T_{\text{maximum-time-boundary}} \) are set as 30 min and 60 min respectively. Under these thresholds, 33.3% of all the locations from the call-location-trajectories are removed; the remaining locations in these sequences form the set of stop-location-trajectories. The average length of these trajectories is 2.97. Based on the assumption that a user starts and ends a day at home, the stop-location-trajectories are added with a home activity at the beginning and/or end of the sequences if the home activity is absent from these two positions. All the obtained stop-location-trajectories are divided into training and validation sets.

4.2. Stop-location-trajectory classification

The obtained stop-location-trajectories from the training set are classified according to the number of home-based-work-tours and home-based-non-work-tours accommodated in the sequences. The average frequencies of sequences in each class relative to the total number of the sequences are 63.05%, 5.29%, 0.84%, 22.31%, 1.86%, 0.26% and 6.39% for classes 1_HBWT, 2_HBWT, 3_HBWT, 1_HBNT, 2_HBNT, 3_HBNT and H, respectively. The sequences in each class are further split based on distance \( d \) of the corresponding users. Fig. 6 shows the distribution of the sequence frequencies in each class, across each kilometer of \( d \). In this figure, each curve represents a particular class. It is noted that, as \( d \) increases, most of the curves do not remain constant; variation in the distribution of the frequencies within each of the classes is observed. For instance, for the top curve representing the most typical class 1_HBWT, the frequencies increase as \( d \) gets larger but starts to decrease when \( d \) reaches a certain distance, e.g. 11km. While for the second top curve featuring class 1_HBWT, the frequencies show a stable rising trend as \( d \) increases. It suggests that, given a certain distance \( d \), the observed sequence probabilities of each tour class slightly differ from the average frequency over all distance values in the class.
Fig. 6. The distribution of sequence frequencies in each class over home-work distances

Based on the observation from Fig. 6, we thus divide \(d\) into 4 intervals including \(\leq 2km\), 2-6\(km\), 6-11\(km\), and >11\(km\). The frequencies of each class in each of these intervals characterize the distance-based-tour-class-distribution. Table 2 lists the obtained results; the average over all distance values are also presented as a comparison. This table further demonstrates the variations among different distance intervals. For instance, for class 1\_HBWT, when \(d\) increases, the frequencies become higher, implying that more people conduct one home-based-work-tour for more days. However, when \(d\) is larger than a certain value, e.g. 11\(km\), people start to perform less home-based-work-tours. Instead, they tend to stay at home or only conduct 1 tour for non-work purposes, as reflected from the frequencies of 28.62\% and 7.87\% in the interval of \(d\) >11\(km\) for classes 1\_HBNT and H which are the highest probabilities over all distance intervals in these two classes.

A further test on this table obtains a statistics of 30569.65 with a significant p-value (i.e. <0.0001), signaling considerable differences in the frequencies across various distance intervals.

Table 2. The sequence frequencies of each class in each of the distance intervals (%)

<table>
<thead>
<tr>
<th>Distance(d)</th>
<th>1_HBWT</th>
<th>2_HBWT</th>
<th>3_HBWT</th>
<th>1_HBNT</th>
<th>2_HBNT</th>
<th>3_HBNT</th>
<th>H</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\leq 2)</td>
<td>61.00</td>
<td>9.76</td>
<td>1.70</td>
<td>19.42</td>
<td>1.57</td>
<td>0.15</td>
<td>6.41</td>
<td>100</td>
</tr>
<tr>
<td>2-6</td>
<td>66.69</td>
<td>5.14</td>
<td>0.77</td>
<td>20.05</td>
<td>1.68</td>
<td>0.21</td>
<td>5.47</td>
<td>100</td>
</tr>
<tr>
<td>6-11</td>
<td>70.05</td>
<td>2.69</td>
<td>0.31</td>
<td>20.15</td>
<td>1.60</td>
<td>0.25</td>
<td>4.95</td>
<td>100</td>
</tr>
<tr>
<td>&gt;11</td>
<td>58.80</td>
<td>1.61</td>
<td>0.20</td>
<td>28.62</td>
<td>2.46</td>
<td>0.45</td>
<td>7.87</td>
<td>100</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>63.05</td>
<td>5.29</td>
<td>0.84</td>
<td>22.31</td>
<td>1.86</td>
<td>0.26</td>
<td>6.39</td>
<td>100</td>
</tr>
</tbody>
</table>

4.3. Hidden Markov Model construction

From all the trajectories in each cluster, a HMM is constructed and the corresponding parameters are estimated. Table 3 presents the transition probabilities for the model derived from the 1\_HBWT cluster, with parameter \(\text{Pesucount}\) being tuned as 0.02. Based on the structure of the model defined in Fig. 3, at the End state \(m_3\), transitions including \(\pi(m_3|m_0)\), \(\pi(m_3|m_1)\) and \(\pi(m_3|m_2)\) are not expected, they are thus represented with ‘Null’.
Table 3. Transition probabilities of the HMM derived from the 1_HBWT cluster

<table>
<thead>
<tr>
<th>Locations</th>
<th>$\pi(m_1, m_2)$</th>
<th>$\pi(m_1, m_3)$</th>
<th>$\pi(m_2, m_2)$</th>
<th>$\pi(m_2, m_3)$</th>
<th>$\pi(m_3, m_2)$</th>
<th>$\pi(m_3, m_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start home (m_1)</td>
<td>0.02</td>
<td>0.29</td>
<td>0.72</td>
<td>0.02</td>
<td>0.02</td>
<td>0.38</td>
</tr>
<tr>
<td>Work (m_2)</td>
<td>0.18</td>
<td>0.31</td>
<td>0.51</td>
<td>0.24</td>
<td>0.76</td>
<td>0.41</td>
</tr>
<tr>
<td>End home (m_3)</td>
<td>0.05</td>
<td>Null</td>
<td>0.95</td>
<td>0.24</td>
<td>0.76</td>
<td>Null</td>
</tr>
</tbody>
</table>

Regarding the emission probabilities $p_{\text{emit}}(x|\text{state})$, in this study, as all activities at the other stop locations except the home and work places, are classified into a single type O, thus $x='O'$ and $p_{\text{emit}}(x|\text{state})=1$ for all activities generated at these locations.

4.4. Monte Carlo simulation

Based on the derived distance-based-tour-class-distribution and HMMs, new sequences for users from the validation set who consist of different workers from those included in the training set, are simulated. In this process, the home-work distance $d$ is first derived from each of the users, and a tour class is chosen based on the probabilities described in the distance-based-tour-class-distribution. In this case study, only when the 1_HBWT class is selected, an entire sequence for the particular user is then further generated according to the HMM derived from the corresponding cluster.

5. Comparison of the simulation results with the validation set

To examine the performance of the proposed modelling approach, we compare the sequences simulated from the models with the original stop-location-trajectories drawn from the validation set. The comparison is carried out in two aspects, including the aspect of individual locations, e.g. the average number of locations visited each day, and the sequential aspect of the locations.

5.1. The average number of locations each day

Among all 156374 stop-location-trajectories observed from 18284 users in the validation set, 61.91% of them fall into the 1_HBWT cluster. The average length of the sequences from the considered cluster is 2.79, and it increases to 4.55 after H is added to the two ends of the sequences.

For all the 18284 users, the tour class is first simulated based on their home-work distances. This results in 62.92% of the users falling into the 1_HBWT cluster. For the obtained users, the entire sequences are generated according to the HMM built from this cluster; the average length of the simulated sequences is 4.72, a close match to the average length of the sequences in the validation set.

5.2. The sequential aspect of the locations

From all the validation sequences in the 1_HBWT cluster, 677 types which are formed by the various combinations of activity locations in particular orders, are found. While for the simulated sequences, 948 types are generated; 520 of them are also observed among the validation sequences. Table 4 lists the sequence frequencies for the 13 most prevalent types, each of which accounts for more than 1% of the total number of sequences in the corresponding sets.
The relationship of the sequence frequencies over all the types between the two data sets is shown in Fig. 7, with the coefficient R as 0.93. The high value of R suggests that the derived HMM model is able to capture the probabilistic distribution of the activity locations and their temporal sequencing revealed by the mobile phone data, and can properly represent workers’ travel behavior in a study area. As a result, the sequences generated from the derived models can accurately reflect the travel demand in the region.

### Table 4. The sequence frequencies for the 13 most prevalent types in each set (%)

<table>
<thead>
<tr>
<th>Types</th>
<th>Validation</th>
<th>Simulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>HWH</td>
<td>41.38</td>
<td>35.09</td>
</tr>
<tr>
<td>HWOH</td>
<td>11.38</td>
<td>12.29</td>
</tr>
<tr>
<td>HOWH</td>
<td>7.97</td>
<td>8.39</td>
</tr>
<tr>
<td>HWOWH</td>
<td>4.60</td>
<td>4.79</td>
</tr>
<tr>
<td>HWOOH</td>
<td>3.53</td>
<td>4.92</td>
</tr>
<tr>
<td>HOWOH</td>
<td>3.26</td>
<td>2.99</td>
</tr>
<tr>
<td>HOOOH</td>
<td>2.21</td>
<td>3.13</td>
</tr>
<tr>
<td>HWOWOH</td>
<td>1.83</td>
<td>1.75</td>
</tr>
<tr>
<td>HWOOOH</td>
<td>1.34</td>
<td>2.01</td>
</tr>
<tr>
<td>HWOOH</td>
<td>1.33</td>
<td>1.42</td>
</tr>
<tr>
<td>HOWOOH</td>
<td>1.21</td>
<td>1.24</td>
</tr>
<tr>
<td>HOWOWH</td>
<td>1.12</td>
<td>1.12</td>
</tr>
<tr>
<td>HOOOWH</td>
<td>1.01</td>
<td>1.18</td>
</tr>
</tbody>
</table>

Fig. 7. The correlation of sequence frequencies for each type between actual phone location sequences and simulated ones

6. Discussions and conclusions

In this paper, we have developed a new method of modelling workers’ travel demand based on mobile phone data. The advantage of this approach is that it does not depend on conventional travel data survey methods. The data requirement is fairly simple and its collection cost is low. In addition, the massive mobile phone data monitors current travel behavior in a large proportion of the population over a long time period. The models derived from the data are thus capable of providing a more general and objective representation of current mobility demand. Apart from the benefits that are realized by the use of the mobile phone data, this approach also provides added value in taking into account the sequential constraints of activity-travel patterns into the modelling process. Once the models are developed in a region, they can be used to simulate activity-travel sequences for each of the employed people in the whole area, given the home and work locations of the individuals. The generated sequences can then be aggregated and subsequently be employed for travel demand analysis, e.g. the average number of trips made in the morning before going to work, on the commuting way, or in the evening after arriving at home. The models can also be utilized to forecast travel demand for future scenarios, e.g.
the displacement of residential areas or the establishment of new industrial sites, which could cause changes in the home-work distances of the workers. Furthermore, travel sequences in a new region, where no phone data is available, can also be predicted by the models, under the assumption that these two regions share similar activity-travel patterns of individuals, e.g. regions from a same country.

With respect to the performance of the approach, data collected from people’s natural mobile phone usage in Senegal in the whole year of 2013 has been used, and the test results show the following major strengths of the proposed method. (i) While the average length of daily sequences from the 1_HBWT cluster in the validation set is 4.55, a close average value of 4.72 is achieved for the simulated sequences. (ii) Among all the 677 different types of the validation sequences, 520 (e.g. 76.8%) are also observed from the simulated sequence set. Particularly, the distribution of sequence frequencies over the 13 most prevalent types shares a high level of similarity between these two sequence sets. (iii) An overall comparison on the frequency distribution over all the 677 sequence types between these sets reveals a correlation of 0.93. All the above results suggest that the derived HMM model is able to capture the probabilistic distribution of activity locations and their sequential orders revealed by the mobile phone data. As a result, the sequences generated from the models can properly represent workers’ travel behavior and lead to an accurate travel demand estimation in the region.

Despite the promising experimental results, the method could be enhanced and extended in the future research in terms of data processing, sequence clustering and model building. Concerning data processing, firstly, by using a fixed work period (e.g. 8am-20pm on weekdays in this experiment), individuals who work during night shifts are ignored. The prediction accuracy of home and work locations could be improved by taking into account the detailed information on individuals’ work regime. Secondly, in the process of stop location identification, two parameters, namely $T_{\text{call-location-duration}}$ and $T_{\text{maximum-time-boundary}}$ are used. $T_{\text{call-location-duration}}$ defines the maximum time duration needed to traverse a single cell area; while $T_{\text{maximum-time-boundary}}$ estimates the total time required for the travel from a previous cell to the current one and from the current one to the next cell. Instead of using overall threshold values of 30 min and 60 min for these two parameters respectively, the settings could be tailored to each particular individual and cells, through the use of the individual’s travel speed and the size of the cell areas.

In terms of sequence clustering, the number of home-based tours encoded in the sequences as well as the home-work distances of the corresponding individuals are used as the classifiers. However, travel behavior is shaped by a range of multiple factors including the conditions of land use and transportation network as well as the social-economic characteristics of individuals. The social-economic information of the phone users could be inferred based on the mobile phone data, and the information could be integrated into the clustering process.

As to model building, improvement can also be made in terms of the following aspects. Firstly, in the designing of the HMM (see in Fig. 3), locations among different parts of the sequences are modelled independently, the correlation between these parts is thus unaccounted for. The interdependencies of activities performed on a day should be integrated in the modeling process, e.g. through conditional probabilities. Secondly, instead of considering only one-dimensional location sequences consisting of home, work and other stop locations, more dimensions of activity-travel patterns could be characterized using the emission probabilities $p_{\text{em}(x|\text{state})}$ at each state of the HMM, thus modelling the multiple aspects of travel behavior. For instance, the locations for other activities O can be distinguished among detailed activity categories. A number of research has been dedicated to annotating activity purposes on the mobile phone locations (e.g. Liu et al., 2013). Similar to activity types, other dimensions, e.g. travel start time and travel distances, can also be
incorporated into the models. In particular, the travel distance at a stop location should be measured relative to the home or work place, and the distribution of the travel distances at this stop is characterized with the emission probability, i.e. \( p_{\text{emit}}(x|\text{state}) \). Once the model is built and a new sequence is simulated for an individual, the specific geographic position of this stop location can be derived based on the obtained distance value, the home or work position of the corresponding individual, as well as the land use data describing the distribution of activity locations surrounding the home or work place.

When being faced with the challenge of acquiring both mobile phone data and real travel survey data from a same or similar study region, in this study the modeling results are tested using mobile phone data of users who are different from those involved in the model training process. However, due to the event-driven nature of the data collection, mobile phone data only reviews the presence of a user at a certain location and time point when his/her phone device makes GSM network connections. The places, where the individual has stayed but no calls were made, are missed. Thus, in the future research, the proposed method must be compared against a real travel survey from the study region or from a region with a similar context. The discrepancies between the simulated sequences and the actual travel sequences could be examined and handled e.g. through an overall scaling factor used by the research (Shan et al., 2011) described in Section 1. Alternatively, the technique developed in the study (Liu et al., 2014), which transforms each of the stop-location-trajectories into actual travel sequences, could be adopted. The obtained actual travel sequences can subsequently be used for the construction of the HMMs.

With the rapid development of mobile phone based services in the future (e.g. Liu & Chen, 2013; Monares et al., 2013), the amount of location data, which is recorded not only when people make calls but when they use the application services on their phones, will continuously grow. The data will reveal more activity locations and travel episodes, thus providing another prospect of improving the model performance and leading to an even better travel demand estimation.

Acknowledgements
The authors would like to acknowledge the support of the European Union (EU) through the project grant of Data Science for Simulating the Era of Electric Vehicles (DataSim). We also want to thank the Second Mobile Phone Data for Development Challenge (D4D-Senegal) committee for the provision of mobile phone data.

References


1 Introduction

A lack of transport infrastructure has been identified as a key constraint on growth in Africa. The World Bank estimates that, if road infrastructure in Sub Saharan Africa would be at levels as in Korea, economic growth rates could be 2.6 percent higher; in Senegal, they estimate the impact to be even larger.

In developed countries, regular survey of commuting times are carried out in order to assess the quality of the existing transport infrastructure and to inform policy makers about where and when to build new roads or upgrade existing capacity. As many African countries experience rapid urbanization, upgrading of public transport infrastructure is key in order to ensure the smooth flow of goods and services.

The existing economic literature has focused on the transport cost of moving mainly physical goods (see e.g. Atkin and Donaldson 2014; Donaldson 2013; Clark et al. 2004), rather than studying the cost of moving individuals.

The economic rationale behind the desire of low transport cost is due to the concept of opportunity cost. A good that is in transit is binding capital that is not being economically used. The cost of transport for physical goods is thus the financial cost that accrues during the time the good is in transit, which is typically measured by an interest rate. However, this concept applies for the movement of humans as well. Here the opportunity cost of transport is the time that individuals spend commuting to work, rather than actually working. As labor costs are typically higher than interest rates, this suggests that the time cost for moving individuals to the place of economic activity are substantial.

This note provides some preliminary evidence on whether the construction and opening of a toll road in August 2013 has had an impact on human mobility in Dakar, Senegal. We use a coarse approach to study whether the number of individuals whose location of economic activity has changed since the introduction of the toll-road. In particular the question is whether more people shift their physical location of economic activity to the new toll-road.
activity, i.e. where they work, in a way that is correlated with locations that disproportionately benefit from the toll road.

The toll road connects Dakar and Diamniadio. It is estimated that the journey between Diamniadio should now take no more than 15 minutes during normal traffic conditions; in the past, this journey could take as long as 90 minutes. The toll fee is set by the government and not the toll operator. It was estimated through a set of surveys to identify a toll fee that would be acceptable to the average Senegalese. The first foundation stone was laid back in 2005, in Malick Sy. Some eight years later, the toll highway between Dakar and Diamniadio is finally complete. It was officially opened on August 1, 2013, following the completion of the second and final road section.

The project is unique as it is one of a few highways in Africa built in urban areas. Diene Farba Sarr, Managing Director of the Senegalese investment promotion agency, claims that the toll road is “a highway that will help to boost development.”

This paper will present some preliminary evidence on whether the opening of the toll road had an impact on human commuting behavior. We rely on mobile phone usage data to identify where individuals work relative to where they live. The empirical design estimates whether people, who live in places that benefit from improved accessibility due to the toll-road see more people working elsewhere relative to where they live.

In order to study this question, two things are required. First, we need to assign stationary locations to individuals whose mobile phone call-detail records we observe. Secondly, we need a measure of a locations exposure to the opening of the toll road. The next two sections describe how we proceed.

2 Stationary Clusters

The individual mobility data is aggregated at a polygon level. It is not clear what is the right approach to do so. One option is to divide the study area into grid cells; the appropriate cell size is not easily arrived at. Since the number of antennas varies across cells, this could result in an overestimation of mobility for areas with higher antenna density. Despite this concern, we use a fairly coarse regular grid to begin with. The regular quadratic grid has grid cell size of 0.03 x 0.03 degrees. This corresponds to an area of around 3 square kilometers at the equator. An alternative approach that will be used in a revised version of this note uses grids that flexibly vary in size based on Voronoi diagrams as used in Yuan and Raubal (2013); Ji (2011).

The study area is broken up into 52 of these 0.03 x 0.03 degree grid cells. If we consider all possible pairs of cells, this would correspond to 52 x 52 = 2704 pairs. Most of these pairs will have never any associated mobility among one another. The analysis is restricted to include a set of 44 grid cells with some population associated and the corresponding set of 1936 pairs for which some corresponding cross-cell intra day mobility will be obtained. The choice of a limited set of coarse cells makes the analysis
more tractable. We proceed as follows.

We use the individual level fine mobility data for two week time windows in order to construct mobility measures at the grid cell level. For every mobile phone user, we attempt to identify clusters of mobility over the two week time window. We compute the average location of the phone at the antenna level during day-time. We categorize daytime as the hours between 9 AM and 4 PM. The other hours are classified to be non-day time hours. We hypothesize that the average location during daytime is a good proxy for the location of economic activity of the average mobile phone user, while the average location early mornings and evenings is a good indicator for the location of non-economic activity. In future work, we will consider different methods to cluster individuals to grid cells, in particular, kulldorf- or DBSCAN clustering methods. We set out using simple means as it is the quickest method of any clustering routine. A concern is that the mean may be distorted by individuals traveling into Dakar by airplane from within Senegal. In order to reduce this concern, we focus our analysis on individuals whose movements fall exclusively into the broad Dakar region depicted in Figure 1.

We compute the individual level day- and night-time locations across all 2 week time window datasets in each grid-cells. Based on this, we can compute the number of individuals that economically active during daytime in cell $B$, but are usually observed at night-time in a separate cell $A$. This is indicative of the number of individuals that make the journey between cells $A$ and $B$, denote this as $X_{AB}$, on a day to day basis. We can think of this as the “export” of laborers from location $A$ to location $B$. Since we observe mobility over time as we have data pertaining to 25 samples covering 2 week time windows, we can add a time index so that we observe flows between $A$ and $B$ at time $t$. Our goal is to study whether $X_{ABt}$ changes once the toll road has opened.

This is a very crude approach and provides a first pass for one important margin of mobility: does improved transportation access allow workers that live further away to participate in relatively distant labor markets? This addresses an extensive margin as improved transportation infrastructure can increase labor supply. In order to study this, we need a measure of a locations exposure of access to the toll road.

\section{Exposure to the Toll Road}

We compute the “exposure to the toll road” as a simple reduced form measure of the travel time or distance travelled between two grid cells $A$ and $B$. We contrast two states of the world: one where this distance or travel time is computed imposing the constraint that the toll road can not be used and once, allowing the toll road to be used. The routes are computed using Google Maps Direction via its application programming interface. An example is illustrated in Figure 3. This provides two alternative routes from Keur Massar, in the north east of Dakar, into the commercial center near Boulevard Dial Diop. One of the routes uses the toll-road. The travel time is estimated to be 25 minutes for 22.7 km, while the travel time on the route avoiding the toll road is 30 minutes for
a distance of 23.2 km. The time difference arises due to the assumed different travel speeds on the legs of the journey that the two routes have not in common. The assumed time savings due to traveling on the toll-road for this journey is almost 17%.

We compute all routes from the centroids of the coarse grid cells to all other grid cells under two regimes: once allowing for the use of the toll road, and once no allowing the use of the toll road. Clearly, this approach provides, at best, only a very coarse reduced form measure of the location specific time saving from having access to the toll road, as it does not take into account for congestion that may occur along the different path. If congestion is unevenly spread and more likely to occur on routes that do not use the toll road, then the estimated reductions in travel time are a lower bound for the true reductions in travel time. Nevertheless, this provides for a first pass on whether the toll road did have an effect on the coarse measure of mobility described in the above section.

The result is a reduced form measure $\Delta d_{AB}$ of the distance reduction measured in terms of “time savings” or “distance savings” on the route between cells $A$ and $B$.

In order to get a sense of the spatial distribution of the computed travel time reductions, we can compute a weighted average of the measure $\Delta d_{AB}$. For every cell $A$, we compute the weighted average $\Delta d_A = \sum_{b \in B} \omega_{Ab} \Delta d_{Ab}$. $\omega_{Ab}$ is the share of people who live in cell $A$ but work in cell $b$ before the opening of the toll road. The result is a measure of the weighted average time savings by grid cells; these are plotted out in Figure 4. It becomes evident that the spatial distribution is concentrated along the route of the toll-road. The weighted travel time reductions can be as high as 12.1% but also negligibly small.

The next section presents the simple empirical specification we estimate.

4 Empirical Specification

The empirical specification is motivated by the trade literature studying gravity equations; these are commonly derived from models of trade. The model was first studied by Tinbergen (1962); many trade models yield empirical specifications that yield trade flows following a gravity type equation. Gravity models, despite several limitations, have been used extensively by geographers, urban planners and economists (see Simini et al., 2012).

The simplest specification relates directed flows between two locations $A$ and $B$ to a measure of the size of population in location $A$ and that of location $B$.

$$X_{AB} = G(M_A^{\beta_1}M_B^{\beta_2}/D_{AB}^{\beta_3})$$

(1)

For a given size of population in location $A$ $M_A$, an increase in the population of a location $M_B$ leads to increasing flows between $A$ and $B$, all else constant. Distance between two locations decreases the flow of people.
In our framework, we observe mobility over time as we have data pertaining to 25 samples covering 2 week time windows. In addition we have time-variation in the measure $D_{AB}$ as the travel distance or time may change with the opening of the toll road. Hence, we are interested in comparing the flows between $A$ and $B$ in these two states of the world: $E[x_{AB} | \text{toll road open}]$ and $E[x_{AB} | \text{no toll road}]$. We can obtain an estimate of this by running the following flexible linear regression:

$$X_{ABt} = \alpha_{AB} + \delta_t + \sum_{t=1}^{25} \gamma_t \times \Delta d_{AB} + \epsilon_{ABt} \tag{2}$$

This specification controls for shifters that are specific to a location pair through the fixed effects $\alpha_{AB}$ as well as general time trends $\delta_t$. These may capture sampling effects due to the repeated cross-sectional sampling for the mobile phone data. The coefficients of interest is the evolution of the coefficients $\gamma_t$. These measure the effect of the reduction in travel time $\Delta d_{AB}$ at different points in time $t$. We expect that these coefficients to change somewhere around the opening date of the toll road near August 1st. The idea is to estimate the above specification on a balanced panel at the grid-cell level, plot out the coefficients $\gamma_t$ and perform a structural break analysis in the estimated coefficients. The interesting question is going to be, whether there is any change in the coefficients $\gamma_t$ and whether the date of the change is correlated with the actual date of the opening of the toll road.

An alternative specification estimates a single coefficient to get the increase in average level of commuting activity across all periods following the opening of the road. The specification becomes:

$$X_{ABt} = \alpha_{AB} + \delta_t + \text{Post}_t \times \Delta d_{AB} + \epsilon_{ABt} \tag{3}$$

Where $\text{Post}_t = 1$ for all periods following the opening of the toll road. We can use this estimated coefficient to obtain a sense of the magnitude of increased labor mobility due to the toll road.

In order to get a sense of the spatial heterogeneity of the effect, i.e. which locations see more out-commuting activity we can estimate an interaction of a grid-cell indicator with a post August 1st, 2013 dummy variable. The estimated coefficient then provides the average increase in out commuting activity across all destination grid cells. The specification is:

$$X_{ABt} = \alpha_{AB} + \delta_t + \sum_{a \in A} \text{Post}_t \times \gamma_a + \epsilon_{ABt} \tag{4}$$

This provides a set of coefficients $\gamma_a$ for every grid cell identified as place of home;
the estimated coefficients can be visually presented.

5 Results

The first results are presented in visual form by plotting out the sequence of estimated coefficients $\gamma_t$ obtained from a linear regression of specification 4. The results are presented in Figure 5. The individual estimated coefficients are plotted out in black; the dashed grey lines correspond to 90% confidence intervals obtained when accounting for two-way clustering on the day-time grid cell and the night-time grid cell. The date of the opening of the toll-road is indicated as the vertical red-line.

The coefficient pattern suggests that in the time-window in which the introduction of the toll-road falls, there is a significant spike in mobility. Grid-cell pairs that experienced a significant reduction in travel times saw a significant increase in the number of individuals who work elsewhere during day-time. The estimated effects become smaller but continue to be, on average, significantly larger compared to the estimated coefficients for the whole 8 months prior to the opening of the toll road. This is indicated by the blue line, which provides the estimated means in the two states of the world. The estimated coefficient for the 8 months prior to the opening of the toll road is 2.53, while it is, on average 7.32 for the months following the opening of the toll road. The variation we exploit comes from the reduction in travel times that was simulated using Google Maps.

The timing of the increase in commuting activity across cells corresponds very well with the date of the opening of the toll road on August 1st, 2013. In order to obtain a better estimate of the overall effect in levels, we can simply estimate a version of the first specification that averages the pre- and post opening sequence of coefficients. These are presented in Table 1.

The coefficients across the Table suggest that average commuting behavior increased statistically significantly. This is robust to including more demanding time fixed effects, that are specific to the origin and destination grid cell in columns (2) and (4). Columns (3) and (4) restrict the analysis to the grid cell pairs in which there is strictly greater than zero commuting.

The unweighted average travel time reduction, across all pairs, is 7.8%. The coefficient in column (1) suggests that, on average, mobility across cells increased by 0.39 individuals. That is, on average, 0.39 individuals see their stable location during the day be a different one as during night following the opening of the toll road. This is very small; the average cross cell mobility is only 29.11 individuals; so in relative terms, the effect is just around a 1.34% increase in mobility.

This hides the significant spatial heterogeneity of the effect. Of the 1936 pairs, only 858 see any predicted change in travel times. These pairs are driving the estimated effect. These pairs see a reduction in travel time by around 17.3%; on average, these pairs had only 7.29 individuals moving usually across cells over time. The relative
effect is much larger for this set. In order to get a sense of the spatial incidence of the opening of the toll road, we can simply estimate the pre-and post effects specific to the location in which individuals cluster in their assumed home grid cell. We simply estimate the specification replacing the $\gamma_i$ with an indicator that is equal to 1 in for the period following August 1st, 2013. We replace the measure $\Delta D_{AB}$ with a set of indicator variables for the different grid cells.

The result is again best visually presented; they can be found in Figure 6. The cells colored in shades of blue benefit from the toll road by seeing increases in the number of people who live there and commute to work into other grid cells. Locations in the commercial center seem to lose out. They see fewer people commuting out in response to the introduction of the toll road. This could be due to three things: first, it could be an artifact of the choice of day- and night time time windows; these may not be adequate anymore following the opening of the toll road. In particular, places that experience significant reductions in travel times may now experience different “normal” working hours. Second, it could reflect individuals moving out of the city center; lastly, it could reflect a genuine reduction in commenting behavior. Another concern is that the density of mobile phone masts is higher in the city center; this makes the estimated locations day-time and night time locations there a lot more precise. It is unclear whether one should expect this measurement error to be varying over time in a way that is correlated with the opening of the toll road. This could be the case, if the mobile phone tower network has changed with the construction and opening of the toll road. An approach using Voronoi cells and a different spatial clustering method may improve on this margin.

Despite these concerns, it is instructive to illustrate the overall effect and the relative effect sizes. Downtown Dakar experiences a drop in out-commuting activity by around -38.87 individuals. On average, the downtown grid cell saw cross-cell mobility of around 191.33 individuals per day. The relative decrease is thus significant around 20%. For the grid cell which mainly covers the administrative area of Les Parcelles Assainies, the estimated effect is an increase in cross cell mobility around 40.01. The average cross cell mobility for this part of Dakar before the opening of the toll road is 153.33. Hence, the absolute change translates into a relative increase by around 26%.
Figure 1: Toll Road Opened in August 2013 marked in red; other main roads are marked in. 0.03 x 0.03 degree grid cells are outlined. Data from Open Street Map.
Figure 2: Study Area - Grid Cells included in the Analysis of Cross Cell Intra Day Mobility Over Time.
Figure 3: Example Computation of Travel Time Reduction due to Toll Road Access. The route using the toll road is highlighted in blue, while the alternative route is highlighted in light grey. Data from Google Maps.
Figure 4: Spatial Distribution of Travel Time Reductions. Travel time reductions are computed as weighted averages, weighted by the total number of people who commute across grid cells before the opening of the Toll Road.
Figure 5: Estimated Effect of Travel Time Reduction on Inter Cell Mobility Over Time. Dashed lines are 90% confidence intervals; the blue line corresponds to the estimated structural break in the series of estimated coefficients. The vertical line corresponds to the official opening of the toll road.

Table 1: Increase in Commuting Activity following Opening of Toll Road

<table>
<thead>
<tr>
<th></th>
<th>Including zeroes</th>
<th>Excluding zeroes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Post x Δd_{AB}</td>
<td>4.958**</td>
<td>8.466**</td>
</tr>
<tr>
<td></td>
<td>(2.004)</td>
<td>(3.489)</td>
</tr>
<tr>
<td>Mean of DV</td>
<td>29.1</td>
<td>29.1</td>
</tr>
<tr>
<td>Time FE</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Pair FE</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Time x Origin/ Destination FE</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Clusters</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>Observations</td>
<td>47300</td>
<td>47300</td>
</tr>
<tr>
<td>R-squared</td>
<td>.993</td>
<td>.994</td>
</tr>
</tbody>
</table>

Notes: All regressions are simple linear regressions. Time x Origin/ Destination FE are a separate set of time fixed effects for origin grid cells and destination grid cells. Excluding zeroes estimates the specifications only on the set of cross grid cell pairs with strictly positive commuting. Robust standard errors clustered at the origin grid cell level with stars indicating *** \( p < 0.01 \), ** \( p < 0.05 \), * \( p < 0.1 \).
Figure 6: Estimated Effect of Toll Road Introduction on Inter Cell Mobility By Non-Day Time Cells. The incidence of increased inter cell mobility is observed in locations along the route.
References


Yuan, Y. and M. Raubal (2013). Extracting dynamic urban mobility patterns from mobile phone data.
Travel demand analysis with differentially private releases

David Gundlegård*
david.gundlegard@liu.se
Clas Rydergren
clas.rydergren@liu.se
Jaume Barcelo
jaume.barcelo@liu.se

Department of Science and Technology, Linköping University

Nima Dokoohaki
nima@sics.se
Olof Görnerup
olofg@sics.se
Andrea Hess
andreah@sics.se

SICS Swedish ICT

Abstract — The use of mobile phone data for planning of transport infrastructure has been shown to have great potential in providing a means of analyzing the efficiency of a transportation system and assisting in the formulation of transport models to predict its future use. In this paper we describe how this type of data can be processed and used in order to act as both enablers for traditional transportation analysis models, and provide new ways of estimating travel behavior. Specifically, we propose a technique for describing the travel demand by constructing time sliced origin destination matrices which respect the level of detail available in Call Detail Records (CDR) from mobile phone use.

When analyzing large quantities of human mobility traces, the aspects of sensitivity of traces to be analyzed, and the scale at which such analysis can be accounted for is of high importance. The sensitivity implies that identifiable information must not be inferred from the data or any analysis of it. Thus, prompting importance of maintaining privacy during or post-analysis stages. We aggregate the raw data with the goal to retain relevant information while at the same time discarding sensitive user specifics, through site sequence clustering and frequent sequence extraction. These techniques have at least three benefits: data reduction, information mining, and anonymization. Further, the paper reviews the aggregation techniques with regard to privacy in a post-processing step.

The approaches presented in the paper for estimation of travel demand and route choices, and the additional privacy analysis, build a comprehensive framework usable in the processing of mobile phone data for transportation planning.

The project presented in this paper a part of the D4D-Senegal challenge.

Keywords — mobility, mobile phone call data, transportation, travel demand, privacy, differentially private releases.

I. INTRODUCTION

Investments in transport infrastructure have been identified to have a positive effect on the economic growth. Since large transport infrastructure investments are very costly, it is important to make careful analysis of the cost-benefit-ratio for each potential investment. The use of mobile phone data for planning of transport infrastructure has been shown to have great potential (see e.g. Berlingerio et al. 2013 and Blondel et al 2013).

By mapping the cell phone data to the transport infrastructure it has been shown to be possible to estimate the current use of the transport system with high accuracy. Based on the estimations, suggestions for improvements to the existing transport system can be generated, for example by using transportation models for scenario evaluations. Decisions taken today on infrastructure development and urban planning can lock cities into mobility behavior patterns for the next 30 to 50 years. Improvements to the infrastructure would result in more efficient mobility and, in the long run, increased economic growth.

The benefit of using cellular network data over traditional sensors, like link counts and manual travel surveys, is a much better coverage. From travel demand estimation based on cellular network signaling data we get direct observations of the generated trips and the distribution of trips for a sample of the population. Dynamic origin and destination matrices can be constructed using techniques for assigning trips into time periods, which takes into account the uncertainty in time stamps of trip start and end, related to the poor sampling in time. To enable detailed transportation analysis, based on the travel demand description, mode choice, route choice and a temporal distribution of travels is also needed.

The approach used in this paper will cover the travel demand description, temporal travel distribution and route choice. The temporal distribution and route choice analysis is addressed by filtering out trips that are suitable for the different purposes. Due to the huge sample of trip observations that are included in the data set we can still get a statistically interesting number of observations.

When analyzing large quantities of human mobility traces, two research aspects are often overlooked: sensitivity of traces to be analyzed, and the scale at which such analysis can be accounted for. On one hand, sensitivity implies that identifiable information must not be inferred from the data or any analysis of it. Thus prompting importance of maintaining privacy during or post-analysis stages. At the same time adapting to such requirements during or post analysis is also an issue, as time and scale factors often justify the true business value of providing mobility analytics to the masses. While to
address the former, often researchers tend to take into consideration data analysis under assumptions that privacy is either preserved while data is being analyzed, or when the results of analysis is to be published and released to data scientists. And to address the latter, means that such requirements are to be held during or after analysis on a scalable and reasonable manner. Maintaining such requirement is often difficult as each dataset may contain millions of trajectories and each analysis task might need to be repeated several times for fine-tuning, regardless of the overhead of corresponding task of course.

Although the D4D dataset is anonymized to a certain degree by reduced spatial and temporal resolutions, de Montjoye et al. (2013) have demonstrated that traces of human mobility are quite unique, to the extent that even a handful of data points, possibly acquired from limited external knowledge, is sufficient to re-identify the trace of an individual even in a sparse and coarse mobility dataset as the one at hand. Our approach to address this challenge consists of two steps. Firstly, we aggregate the raw data with the goal to retain relevant information while at the same time discard sensitive user specifics. Put differently, we merge individual sequences of sites used into aggregates that should reflect collective mobility behavior rather than the whereabouts of individuals. Aggregating the data, which is here done through site sequence clustering and frequent sequence extraction, has at least three benefits: data reduction, information mining, and anonymization. The second part of our approach is to review resulting models with regard to privacy in a post-processing step. On a general note, maintaining privacy during and post analysis phases, encompass two main categories of approaches commonly referred to as Privacy-preserving data analysis (PPDA) and Privacy-preserving data publishing (PPDP) (Fung et al. 2010). Our focus for privacy-preservation is on the latter part, which is maintaining published results indistinguishable.

A. Aim and key paper outcome

The key outcomes of the paper are a set of transport demand indicators, based on the concepts of trajectories and trips, which are measured using the present type of mobile phone usage data. Based on these indicators, we present demand and route estimations for the case of Senegal, but applicable also to other regions where the same type of data is available. Furthermore, we analyze how data of this type can be distributed maintaining privacy during or after the different analysis stages. This is done by aggregating the data through site sequence clustering and frequent sequence extraction, which has the benefits of data reduction, information mining, and anonymization.

Analyses presented in this paper are based on the mobile phone data from the country of Senegal, presented in detail in de Montjoye et al. (2014). The project presented in this paper a part of the D4D-Senegal challenge.

B. Outline

The rest of the paper is organized as follows. In Section II the background to the studied case, Senegal, is presented in terms of the mobile phone data used, and previous work on this type pf mobile phone data for travel demand and privacy are reviewed. In Section III the mobility characteristics derived from the data is shown, Home and POI, trip definitions are given, the processes of sequence clustering, frequent sequence extraction and travel demand matrix construction are presented. Section IV discusses the scalability of the presented processes. Section V contains the privacy analysis based on trajectory and travel demand data. The paper is concluded in Section VI, where the results are summarized and use cases are presented.

II. BACKGROUND

A. Cellular and transportation infrastructure

Senegal is located in the west of Africa and has about 12 million inhabitants. The capital of the country is Dakar in far west part of the country and close to the Atlantic Ocean. Dakar has 1.1 million inhabitants, with about 2.7 million inhabitants in the urban area close to the city.

An overview of the road infrastructure, as presented in the Open Street Map, is presented in Figure 1, where also the distribution of the mobile antennas is shown, represented by the red dots.
It should also be mentioned that this dataset only contain data from CDR, which is a subset of the mobility data that is available in a cellular network. Other types of data that can be collected from the cellular network, e.g., location updates, handover events or measurement reports, would affect the results of the analysis. A detailed description of the data available in cellular networks for the purpose of traffic management and planning can be found in Gundlegärd and Karlsson (2006).

C. Previous work

In this section we review previous work related to travel demand estimation from mobile phone data, and privacy aspects of releasing mobile phone data of CDR type for use in, for example, travel demand estimation.

Techniques for using data from communication networks to understand the interactions among people has been developed both in academia, such as MIT Sensible City Lab and Pisa KDD Laboratory, starting with Ratti et al. (2006), and commercial institutes, such as IBM Smarter Planet Initiative and Microsoft Research during the recent years. One of the first large scale data evaluation of mobile phone data is described in González et al. (2008).

In Fiadino et al (2012) the problems of bias from using CDR data is discussed, pointing out that the accuracy of the mobility analysis can be low for datasets where the majority of the users make only a small number of calls per day, and, hence, a known approximate location. The estimation of individual’s home and work locations has been treated in several papers, see e.g. Ratti et al. (2006), Vieira et al. (2010), Isaacman et al. (2011) and Csáji et al. (2012). They apply techniques of varying complexity, based on clustering the locations for day time and night time separately. A source of many results from experiments to identify important places from CDR data of the same type as in this paper, only a slightly smaller sample and for the Ivory Coast is available in Blondel et al. (2013). Despite the shortcomings in the level of accuracy of CDR data, mode choices have been analyzed in e.g. Wang (2010), although only individuals with a high call frequency was used for the analysis. From articles with the aim of estimating origin destination demand matrices from CDR data, it is possible to identify differences in both the time resolution and spatial resolution. The aggregation in the spatial dimension can, for example, be built up using known traffic analysis zones when known (as in Calabrese et al. 2011) or as zones of fixed size squares (as in Wang et al. 2010) or individual antennas (as in Nabouli et al. 2013). The time resolution varies from a static matrix for one hour (as in Csáji et al. 2012) to several time periods, see Wang et al. (2013).

A more advanced approach of utilizing mobile phone data for both travel demand estimation and route choice is presented in Iqbal et al (2014). Their technique makes use of an additional source of data, link count from specific streets in the transportation network. In Cáceres et al., (2012) the correlation between number of calls and the actual traffic flow is investigated. A more detailed overview of the use of mobile phone data for transport analysis can be found in Rajna (2014).

Among the approaches for privacy-preserving data analysis, one approach that has gained increasing attention in recent years is Differential Privacy (Dwork 2006). Differentially Private algorithms allow for results of the computation to be perturbed in a way that has small impact on aggregates, and at the same time de-individualizes the data of individual s present within the data and in turn minimizes their privacy loss. One of the reasons for popularity of Differential Privacy (DF) is that it considers no assumptions about the capabilities of a potential adversary, whilst other approaches need explicit adversarial modeling for similar tasks (Mir et al. 2013).

Given the two main challenges of privacy-preserving data analysis, the state of the art of differentially private mobility analytics tend to address either one corresponding issues at hand, whilst those differently private solutions that consider scale and accuracy at the same time are more recent(Mir et al. 2013; Acs and Castelluccia 2014; Fan and Xiong 2012). Mir et al., aim at maintaining a reasonable accuracy for modeling mobility of users (Isaacman et al. 2012), through generation of a synthetic population based on existing traces at metropolitan scales (Mir et al. 2013). Fan et al., present an approach that adaptively trades-off the accuracy to utility in the release of real-time sensitive time series data (Fan and Xiong, 2012). Acs and Castelluccia, present an anonymization technique for releasing spatio-temporal density of a metropolitan scale trace input data. They argue that even with large dimensional sensitive data, differential privacy can provide practical utility with meaningful privacy guarantee (Acs and Castelluccia, 2014).

III. MOBILITY AND TRAVEL DEMAND ANALYSIS

Travel demand analysis for transportation planning is traditionally performed using the classical four-step model, which divides the problem into 4 different sub-problems: trip generation, trip distribution, mode choice and finally route assignment. From cellular network data we get direct observations of combined trip generation and trip distribution for the users in the data set, but the poor resolution in time and space in CDR data causes problems to relate antenna movements to physical movements. Also overlapping antenna coverage in the cellular network causes problems related to movements that are purely an artifact of cellular network characteristics.

The poor resolution in time and space is even more problematic in the last two steps, mode choice and route assignment. To be able to make analysis on route choice and also temporal demand characteristics, we have used an approach where different types of trip definitions are used for the different steps in the analysis. For route choice we have filtered trips that have good resolution in space and for temporal analysis we have filtered trips with good resolution in time. Due to the large amount of trips in the whole data set, we can still get enough observations to enable also analysis of dynamics that is seldom captured in the majority of user trajectories.
A. Mobility characteristics

The resolution in space and time of user location sampling is a key component in determining which type of mobility analysis that can be made with the data set. To enable comparison of the results based on this data set we have calculated average inter-event statistics, see Figure 2, which can be compared with Figure 1 in Calabrese et al. (2013). Calabrese et al. analyze data that not only include call and SMS connections, but also connections to the Internet over the cellular network. They report an arithmetic average of 84 minutes for the medians (corresponds to the blue group). They conclude that the average of 84 minutes allows the detection of changes in locations where the user stops for as little as 1.5 hours. The corresponding values for our dataset is an arithmetic average for the medians (blue group) of 308 minutes which indicates that it would be possible to detect stops which are about 5 hours and longer.

By viewing the distribution of inter-arrival times as in Figure 3, we see that a significant fraction of users use their terminals comparably infrequently. Note also that there are local maxima separated by 12 and 24 h intervals, indicating that a comparably large number of terminals are used once per day, every second day, and so forth.

B. Home and POI

Calling activity is correlated to the users’ points of interests (POI). Therefore, it is feasible to estimate the home location of users based on a sequence of location observations, see e.g. Dash et al. (2014). Since POIs, especially home and work, are very important for a user’s trip generation and distribution we have used the estimated home and work location of users as input to one type of trip generation, see section III.C. We have simply used the call events to estimate the home and work location of users, based on the frequency of calls from different locations during daytime and during nighttime. The home and work location identified as locations with a minimum distance of three kilometers, and not belong to neighboring antennas in the Voronoi graph. By aggregating home and work locations for all users, we can get a technique for identifying residential and industrial or public areas. In Figure 5 a heat map of the difference between home and work locations is shown. Blue indicates more home locations than work locations, and red indicates more work locations than home locations. The red area in the middle of the figure is the International Blaise Diagne airport, located south-east of Dakar. The red area in the lower part of the figure most likely indicates an industrial area located along Route Sindia-Thies. The small red area north-west of Thies shows an area with a university and an airport.
Figure 5: Heat map of difference of number of home locations and number of work locations. Blue indicates potential residential areas and red indicates areas with large daytime activity. The red area in the middle of the figure is an airport.

Since trips generally are generated from residential areas to industrial or public areas in the morning and the opposite in the evening, this kind of map can directly give a rough understanding of travel patterns in an area.

C. Trip definitions

In order to analyze travel demand and mobility, individual movements need to be identified. In this section we define two ways of describing movement, trips and sequences. Trips are here related to movements between activities and are only defined by start and end location, referred to as origin and destination. A sequence is a series of locations consecutively visited by a user and can be extracted from trajectories.

We extract sequences by processing the data in the following way (default values are given in parenthesis):

1. Repeating sites are discarded. If at the beginning of the sequence, the last site is kept, else the first site is kept.

2. All records are divided into sequences, where a new sequence starts if
   a. There is a new user or
   b. The time between two records is larger than a time-out value (60 minutes)
   c. The distance between two locations is larger than twice the maximum cell range (50 km)

3. Sequences that contain less than a given number of sites (5) - assumed to be generated by stationary users - are discarded.

The processing of the user trajectories into sequences enables a level of anonymization, since they do not include the time stamp of the locations. We have explored two approaches to model common user routes as represented by sequences: clustering and frequent sequence mining, see section III.D and III.E. Sequences are in Wang et al. (2012) described as transient OD observations (t-OD). This process of extracting sequences filters out many movements by the thresholds in inter-event time (step 2b) and sequence length (step 3).

In order to study travel demand, it is important to capture as many movements as possible from the CDR data, even with poor resolution in time and space. This is done using assumptions on travel behavior related to predefined POIs (here, home and work location) and by relaxing the constraint on inter-event time compared to the sequence definition.

In the trip definition we assume that all movement start from the home location in the morning and end in the home location in the evening, unless the user’s distance to home is larger than a threshold value $d_{\text{max}}$. Furthermore, a threshold value, $d_{\text{min}}$, is used as a minimum movement distance to identify the start of a trip as well as snap the origin or destination location to any of the user’s POI:s. One of the rationales for this trip definition is that it is relatively easy to estimate the home location of a user, given that the user trajectories cover a sufficient period of time.

The algorithm for generating trips is divided into three functions, main, detect_trip_end and detect_trip_start. The functions are presented in Algorithm 1a, 1b and 1c, respectively.

\begin{verbatim}
main()
for all u in U
    for each day d in D
        for all user positions day d, p_udk
            if(trip_active == false)
                trip_active = detect_trip_start()
            end
            if(trip_active == true)
                trip_ended = detect_trip_end()
            end
            if(trip_ended)
                store_trip()
            end
        end
    end
end

detect_trip_end ()
    if(p_udk == workbase or p_udk == homebase)
        destination = p_udk
    else
        if(p_udk+1 exists)
            if(p_udk == p_udk+1)
                destination = p_udk
            end
        else
            if(d(p_udk,homebase) < d_{max})
                destination = homebase
            else
                destination = p_udk
            end
        end
end
\end{verbatim}

Algorithm 1a: Main function for the trip generation.

Algorithm 1b: Function for detecting trip end.
Algorithm 1c: Function for detecting trip start.

```
detect_trip_start()
if (trip_set empty)
    if(pudk != homebase and \(d(p_{udk},homebase) < d_{max}\)
        and \(d(p_{udk},homebase) > d_{min}\))
        trip_active = true
        origin = homebase
    end
    if(pudk != homebase and \(d(p_{udk},homebase) > d_{max}\))
        trip_active = true
        origin = pudk
    end
    if(pudk == workbase and \(d(p_{udk},homebase) > d_{max}\))
        trip_active = true
        origin = homebase
        destination = workbase
    end
else
    if(pudk != previous_trip_start(trip_set) and \(d(previous\_trip\_start(trip\_set),\ pudk) > d_{min}\))
        origin = previous\_trip\_start(trip\_set)
    end
```

The number of trips generated using this trip definition is approximately 0.7 trips per day and user, with \(d_{max}\) set to 3 km and \(d_{min}\) set to 100 km. This can for example be compared to the number of sequences generated, which is approximately 0.06 sequences per user and day.

The distance distribution for the trips generated is shown in Figure 7. It can be noted that the number of trips tend to follow the decay of the distance with a negative exponential; similar to what is common in gravity models for trip distribution (Wilson, 1967).

![Figure 7: Trip distance distribution for the generated trips.](image)

D. Sequence clustering

With the goal to extract common transient mobility patterns of users, we have developed a clustering algorithm that aggregates sequences with respect to their site-to-site transitions. The approach is stochastic and based on locality-sensitive hashing (min-hash) (Broder et al. 1998), which is used to generate groups of sequences that are likely to be similar. More specifically, each sequence is mapped to its set of site transitions (constituting pairs of site id:s). The sequences are then clustered with respect to the similarity between their respective site-pair sets in terms of the sets’ Jaccard distance.

The algorithm is similar to the approach proposed in (Görnerup 2012), with the difference that we do not employ graph clustering to group sequences, but instead iteratively merge them into aggregates. In practice this is done by calculating two min-hash values per site transition set and then initialize clusters as singleton aggregates that each contains a single sequence. We then group the aggregates that have the same min-hash values – constituting candidate sets of similar clusters – and merge the aggregates within groups that have a Jaccard distance below a given threshold. The procedure is iterated multiple times, resulting in successive merging of aggregates into coarser aggregates. The algorithm terminates when the procedure converges (typically after approximately 20 iterations in our experiments).

When applying the algorithm on a year’s worth of sequences, the resulting aggregate sizes follow a power law
distribution, see Figure 8, where, at the one end, most of the aggregates are small (only containing a handful of sequences) and, at the other end, there is a long tail of few large aggregates. The distribution indicates that most of the sequences are unique and may therefore not be aggregated.

Figure 9, showing an example of aggregates that each consists of at least 50 sequences, indicates that the clustering approach captures comparably (cf. Section III.E) large-scale patterns that cover distances on the order of kilometers rather than meters. Note, however, that these results are tentative and that no quantitative evaluation of resulting clusters has been performed due to current lack of ground truth.

Figure 9: Example of sequence aggregates illustrated as site-to-site transitions color coded by aggregate.

E. Frequent sequence extraction

Sequential pattern mining is a well-established technique in data mining to detect partially or totally ordered subsequences of items in series of actions or events (c.f. Mooney and Roddick 2013). Applied to movement trajectories, commonly taken routes – i.e., location sequences – can be discovered in a scalable manner. We outline here the experiments conducted with two kinds of sequential pattern mining algorithms: (i) the Seqwog algorithm is based on the frequent itemset mining mechanism Relim (Borgelt 2005); however, the algorithm preserves the sequence of locations. (ii) the MG-FSM algorithm (Miliaraki et al. 2013) is a scalable frequent sequence mining algorithm built for MapReduce, which tolerates “gaps” between consecutive locations. Preserving the sequence of locations is important in our use case since it requires storing the direction of movements. Allowing gaps in the sequence is beneficial; on the other hand, to cope with changes in the base station topology that occur during the time span the trajectories have been captured. For example, if a new base station is deployed, the algorithm would in any case identify the frequent path despite an additional location exists within the particular sequence.

Configurable parameters of Seqwog are support threshold $s$, giving the fraction of sequences that need to contain a subsequence to count it as “frequent” ($s = 0.0075$), and minimum sequence length $m$, which is set to 3 to capture both the direction users are typically coming from and they are heading to, given a particular location. For MG-FSM, the support threshold $s$ is the absolute number of trajectories and has been set to 500. The minimum sequence length is set as well to 3, the maximum length to 6. The maximum allowed gap $\gamma$ is set to 1, i.e., only subsequences are considered that can be built without omitting more than one consecutive location. Figure 10 visualizes the results for these parameter settings for an excerpt of Senegal’s map. It has to be noted that density and length of the sequences depend significantly on the support threshold configuration, which has to be hence chosen carefully. Furthermore, the setting might be adapted to the actual user density in different areas of the country since a low default threshold for all areas leads to a consideration of merely common routes starting in the major city.

Figure 10: Sample result for the capital area for the algorithm Seqwog (left) and MG-FSM (right).

F. Travel demand

The travel demand is one essential input to models for transportation analysis. The travel demand is normally described in an origin-destination matrix. Given a division of a geographical area and a division of the area into zones, the origin-destination matrix describe the number of trips from each pair of zones, e.g. from zone A to zone B for each pair (A, B). The origin-destination matrix describe the demand in a given time interval, for example one hour. Normally, the origin-destination matrix describes the number of trips that starts at zone A during the specified time interval, going to zone B.

Cellular network data is interesting from demand modelling perspective, since we can get direct observations of the travel demand for all transport modes, see Angelakis et al. (2013).
Input to the time-sliced OD-matrix are the trips generated by the trip definition described in Algorithm 1a-c. These trips give direct observations of trip generation and distribution for the sample of users in the data set. However, it should be noted that in this data set, the physical process of interest here, i.e. the human travel behavior, is under-sampled, which gives an uncertainty in which trips that has been made by a specific user.

The spatial resolution of the data set is limited by antenna density, since only antenna ID:s are available in the data set. The antenna density is strongly correlated to population density and hence we get a better spatial resolution of trips in areas with denser population. However, the main problem when generating travel demand from CDR data might not be the spatial resolution, but rather the overlapping coverage of antennas, which makes the standard Voronoi representation of cell coverage a poor representation. This problem becomes worse in areas where macro cells with large transmission power in elevated positions are used for coverage and micro cells with low transmission power are used for capacity. We try to cope with the antenna oscillations by only considering trips longer than a minimum distance $d_{min}$ and not consider trips between antennas that are Voronoi neighbors.

Since users are sampled only during phone activity in terms of calls and SMS, there is a large uncertainty in the temporal domain for the start and end of each trip. Since we want to include as many trips as possible to get a good estimate of the travel demand, we need to include trips with poor temporal resolution. We assign each trip to a time period according to the probability of the trip being started in each time period.

For an individual that makes a trip, as defined by the trip definition, corresponding to a CDR at location A at 7:00 and a CDR at location B at 10:45, the contribution to the demand matrices will be computed as follows. First, we estimate a travel time based on the Euclidean distance from A to B and a travel speed of 50km/h. Let us, as an example, assume that the distance between A and B is 50 kilometers, then we deduce that the trip has started sometimes between 7:00 and 9:45. By assigning equal probability to all start times during this time interval, the contribution from this specific trip will be 1/2.75 to the demand matrix holding the demand from 7:00-8:00, 1/2.75 to the demand matrix holding the demand from 8:00-9:00 and 0.75/2.75 to the matrix holding the demand from 9:00-10:00, for the element representing the travel relation A-B. The trip weights assigned is illustrated in Figure 11. The method can easily be modified to use other than uniform probability distributions, taking into account more information about trip departure times.

This type of weighted OD matrices is calculated for both antenna level and arrondissement level. In Figure 12 both antenna level (blue) and arrondissement level (red) OD is shown for the city of Dakar, filtered for the pairs with largest number of trips. Due to the large number of antenna pairs, it is difficult to see any general trends in the visualization for antennas, however, at least in this example, it is easier to capture in the arrondissement level OD.

In Figure 13 arrondissement level OD is shown for the whole country. It can be seen that most of the travel demand is located in the Dakar area and along the north border of the country.

In Figure 14 the arrondissement level OD is shown for the Dakar region and it can be seen that most of the trips are made within the city, but Dakar also attracts trips to and from the larger cities in the region.
In order to get a potentially higher temporal resolution for trips, we have further analyzed trips that have a small difference in estimated travel time based on origin and destination location compared to the timestamps of the start and end observations. Due to the large data set it is still possible to get a large number of travels in each arrondissement OD pair. Figure 15 shows the distribution of start times for all travels (blue) and for one specific OD pair (red). The specified trip definition in combination with this filtering of well-defined start times indicates that there is peak in travels that start around 12 and 21. However, one should note the strong correlation with the number of events shown in Figure 4, indicating a bias due to bias in location sampling.

The transport route choice, and possibly also mode choice, can be studied by filtering out a subset of the trips that are well suited for each task. Once again, due to the large sample size, we can get enough number of travels to gain understanding of both route choice and mode choice. We have filtered out sequences with short inter-event time and assigned them to routes in the following way:

1. Simplify: Find the set of links on the road network that intersects with the Voronoi polygons of the antennas included in the trajectory and compute a shortest path within each polygon.

2. Construct route: Calculate the shortest path that passes through every Voronoi polygon associated with the antenna in the sequence. In each Voronoi polygon, a shortest path computed in Step 1 is used.

Doing this for all sequences between a specific antenna pair generates a route probability for a set of routes in the given antenna pair. This route probability can then be used together with the travel demand between the specific antennas to estimate a travel flow distribution on the computed routes in for the specific OD pair. By summing all the OD flows that pass a given link, we can get a very rough idea of the flow on that link. Figure 16 shows an example of route probabilities for a given antenna pair calculated using the procedure described above. Although the results from this technique may be affected by oscillations from connecting to different antennas even if the user stay at the same location, the technique may have some potential for generating choice sets as input to more advanced route choice models (see, Bekhor et al. 2006).

IV. SCALABILITY

When developing methods for analyzing and modeling cellular network data it is necessary to address scalability due to the potentially huge volumes of data that needs to be processed. In particular, when considering low-latency online applications, scalability is a prerequisite for any algorithm to be applicable in a real-world scenario.

The clustering algorithm presented in Section III.D is well suited for parallelization. It is implemented in Scala using the
Spark framework (Zaharia et al. 2010) and can therefore readily be used in a cluster environment. As seen in Figure 17, the algorithm may also be applied to large data sets using more modest computational resources.

We have also evaluated the performance of the frequent sequence mining algorithms discussed in Section III.E in terms of scalability by measuring the time needed to process different sizes of data. Figure 18 visualizes the runtime performance of Seqwog and MG-FSM. While Seqwog performs better on small-scale datasets (less than 2 Mio. sequences), whereas MG-FSM yields more favorable runtime values as the data size increases.

![Figure 17: Runtime of clustering algorithm for different number of sequences on an Apple MacBook Pro with a 2.8 GHz Intel Core i7 processor and 16 GB of RAM.](image)

![Figure 18: Runtime of frequent sequence mining algorithms Seqwog and MG-FSM for different number of sequences.](image)

V. PRIVACY ANALYSIS

As stated previously, within our work we focus on maintaining the results of the analysis, differentially private. To do so, we are using a generic differential privacy system that makes two assumptions: first, any individually-sensitive analysis has been tailored with little or no assumption of privacy in mind, thus demanding the release and publication point to be perturbed accordingly. And on the other hand, given the size of the data and its corresponding analytical tasks, framework of needs to scale the analysis to large quantities of mobility traces. Both requirements are satisfiable through the framework used, which is called GUPT (Mohan et al. 2012). The framework is generic in the sense that its external hooks and data transformers can execute any computational and analytical tasks. For which we have used to develop our queries and data transformers with. GUPT is also equipped with components allowing least-trusted analyses and data releases to be made within and from its boundaries, which authors refer to as isolated execution chambers.

A. Privacy Budget Management

Differentially Private frameworks realize the concept of privacy in terms of a utilitarian concept referred to as budget (which we denote as Epsilon from this point forth). Such formulation specifies how many queries a data owner allows towards their data and in turn analysts can spend to consume the released data. Similar differential privacy systems such as PINQ (McSherry, 2009) realize utilitarian budget provision but did not explicitly provide any mechanisms allowing programmers to correlate distribution of the queries to amount of budget.

This is while GUPT provides algorithmic means allowing programmer’s to find and estimate efficient budget per data and task, which in turn allows for automated mechanisms to distribute the limited privacy budget between queries.

A.1 Budget Estimation for Non-Interactive Queries

We hereby present the results of experimental budget estimation with respect to analyses and data sizes at hand. To measure the budget we use a median computation that finds the average site id. Motivation for using such task is first, reading all input bounded values so we can retrieve and analyze all input values. At the same time, since the outputs are already processed such task resembles the type of statistical analysis that data analysts will run on the results to be released.

Figure 19 illustrates the mean relative error for two sets of budget estimates for statistical analysis for a sequence clustering input. Two sets of outputs were analyzed; normal differentially released output and filtered outputs. Using filtering outputs enables us to make sure irregular values (that do not fit the norm of the output) are not released with results. Each computation is generated using three iterations minimum to make sure ranges of Epsilon converge. In this figure only best and converged results are shown.

![Figure 19: Convergence of privacy budget estimations for a site sequence cluster input. Epsilon values taken from range of [0,1].](image)
Figure 20 illustrates mean relative error for two sets of results for two frequent set clustering techniques. To illustrate convergence of optimal budget estimations, three iterations and their corresponding budgets and relative errors are visualized. As visualized budget estimations tend to become more accurate as the range of relative errors reduce and become static helping us to provision corresponding privacy budget according to task and data at hand.

![Figure 20: Convergence of privacy budgets for two frequent sequence mining route aggregation techniques. Each aggregate set were analyzed using an Epsilon from range of \([0.2, 1]\). Estimator was run for three iterations.](image)

A.2 Budget Estimation for Interactive Queries

Results discussed so far, are the results of analysis across the whole output data, thus releasing the value of differential privacy in terms of non-interactive queries. In addition to analyzing the budget per width of the released data, we also decided to analyze the budget variation per longitude of the data released. Such analysis is often in high demand if interactive queries are formulated. With respect to this part of the study we used our mean estimator and we executed it on a set of nominal features of two of origin destination matrix and sequence clustering techniques.

The data unit studied at each analysis session is an aggregate in terms of site sequence clustering and a trip in terms of origin-destination techniques. For each technique a fraction of unit were chosen. To make sure accuracies are representative, a privacy budget range between 0.1 and 1 were chosen. Figure 21, depicts the budget ranges for site sequence clustering on the left side and origin destination matrix on the right side. As shown the projected accuracies can show how sensitive each approach is with respect to handling interactive data release.

![Figure 21: Comparative analysis of privacy budget ranges for longitudinal data units.](image)

B. Calibrating Noise using Output Range Estimation

In addition to estimating optimal privacy budget with respect to data and analysis requested, we also need to stress out the importance of the amount of noise that is added to the released outputs. Another important aspect of the differentially private system at hand is estimating the amount of noise to be added.

To do so, we need to select an optimal output block size (which we refer to as Gamma \(\gamma\)) that will allow us to balance the estimation error and the noise. Since optimal block size varies from problem to problem. Getting the optimal block size based on the analysis task helps to reduce the final error to a large extent. Needless to say, the larger the block size, the smaller the estimation error.

As a result, to be able to estimate the correct output range for the tasks at hand, we also experiment with variations of Gamma values. Same experiment setting as described in privacy budget part is used on three different site sequence cluster outputs. Figure 22 shows how variations of Gamma affect the amount of the noise and thus mean relative error of the released outputs. To also study the impact of privacy budgets, we also changed the variation of the Epsilon values step-wise. Gradual increase of relative mean error with respect to size of the cluster input data is visible throughout the visualizations.
Figure 22: Output range estimations for statistical release of three experimental site sequence clusters. Gamma values are taken from the range of [0,4] and Epsilon values are taken from the range of [1,2].

Since the release sizes can vary depending on the cluster size outputs, it is important that we can leverage an optimal Gamma range in addition to Epsilon to tradeoff the utility of data release to loss of accuracy as well. This is worth mentioning that the optimal tradeoff between block size and number of data blocks can vary for different queries executed on the dataset.

VI. USE CASES AND CONCLUSIONS

The travel demand is a crucial input to infrastructure and transportation planning and is traditionally estimated using census data, travel surveys and models for trip generation and trip distribution. The travel surveys includes detailed travel patterns for a small percentage of the travelers and are relatively expensive to collect, hence the travel surveys are typically updated with a very low frequency. However, cellular network signaling data enables direct observations of trips for a large number of travelers in a cost efficient way and this will change our understanding of human mobility and travel demand fundamentally.

Since the traditional way of estimating travel demand depends on census data that lacks a temporal component and static models for trip generation, travel demand dynamics has not been studied in great detail. Most of the efforts have been made related to road traffic demand, where dynamic demand estimation has been performed by fusing static demand with sensors that has high temporal resolution, e.g. traffic counts in the road network. However, the demand estimation problem based on selected link flows is severely underdetermined and the estimates include a lot of uncertainty. Furthermore, road traffic counts are only measuring vehicles and not travelers, which for some applications are less suitable.

The spatial and temporal resolution that is possible to achieve with cellular network signaling data depends on the cellular network infrastructure, but also on which interface in the cellular network the data is collected from as well as any preprocessing that is made on the data. CDR data based on SMS and call activities typically suffers from a relatively poor temporal resolution which needs to be compensated for in the estimation procedure.

We have noticed mobility activities that we believe are related to the use of a hierarchical cell structure with macroscopic (umbrella) cells in combination with microscopic cells. More meta data about the cellular network, such as transmission power and antenna height would enable more detailed analysis of this, and potentially better mobility estimates for shorter movements.

In the travel demand estimation from cellular network signaling data we get direct observations of combined trip generation and trip distribution for a sample of the population. In this paper we use a trip definition that generates trips for users based on travel assumptions related to points of interests, which has the potential to generate more accurate travel demand estimates compared to approaches where trips filtered by the low resolution sampling are not considered. A dynamic OD matrix is estimated using a new way of assigning trips into time periods, which takes into account the uncertainty in time stamps of trip start and end, related to the poor sampling in time.

To enable more detailed transportation analysis, based on the travel demand we also need mode choice, route choice and a temporal distribution of travels. All these three are possible to do with cellular network data, by filtering out trips that are suitable for the different purposes. Due to the huge sample of trip observations that are included in the data set we can still get statistically interesting number of observations. We have demonstrated this for simple examples related to route choice and temporal distribution, but the principle holds also for mode choice. This type of observations are very interesting since many traditional models that are used for mode choice, route choice and temporal distribution rely on basic assumptions that may not always be valid and can also contain a very large set of model parameters that are difficult to calibrate. For example, many route choice models rely on the assumption that each user has perfect knowledge of the traffic situation and requires volume-delay functions for each link in the network.

An interesting side result from the trip generation based on users’ home and work location is that we have calculated a generic metric of the relationship between residential and commercial/industrial activity in an area. This metric is interesting for transportation analysis, but potentially also for other application areas.
We have outlined a framework for differentially private release of arbitrary mobility analytics allowing us to forecast ranges of optimal privacy budgets and output ranges for public data publication. For the experiments presented, we have also showed that albeit the size of data and queries at hand utility could be maintained at decent levels.

We have proposed two approaches for aggregating cell id sequences - through clustering or frequent sequence mining - and that this results in a certain degree of anonymization due to that mobility information of individuals is replaced by descriptions of collective movements of users. The degree of anonymization may be adjusted by the coarseness of acquired aggregates, where we face a trade-off between accuracy and sensitivity, where the latter may be quantified within the differential privacy formalism. At one extreme, all sequences are grouped into a single aggregate that provide maximum privacy protection but essentially no mobility information. At the other extreme, all sequences form singleton aggregates that give maximum information, but no privacy protection. The point to choose between these two extremes is not only dictated by privacy requirements, but also by computational constraints e.g. pertaining to storage and computational capacity and available bandwidth, that require that the data is distilled - in other words, aggregated.

The presented approaches build a comprehensive framework usable in offline processing or in real-time applications for, e.g., telecommunication operators or transportation planners. Several considerations have to be taken into account for real-time applications. The clustering algorithm (see Section III.D) is highly suitable for this mode as the clustered sequences can be incrementally updated. However, updating the set of frequent location sequences identified by sequential pattern mining (see Section III.E) would require scanning the whole, now extended dataset D'} to enable two types of operation, namely (i) deletion of sequences which have been frequent in D, but are infrequent in D', and (ii) insertion of sequences that have been infrequent in D but frequent in D'. Improving the efficiency of updates might be done for example by caching semi-frequent sequences based on a threshold lower than the actual support threshold.

Cellular network signaling data will change how we understand travel demand dynamics and human mobility in general. In developing countries, the cellular network is typically much more developed than the traffic and transport sensor infrastructure, which will make it an extremely valuable source of information for strategic, tactic, and maybe also in the future, operational planning of transportation networks. Efficient algorithms and models that utilize the characteristics of the underlying cellular network data while maintaining the personal integrity of users in the system will have a large potential in improving transportation and environmental quality in many large cities in the world.

ACKNOWLEDGEMENTS

The two first authors thank Nils Breyer, Carl Johansson, Tao Peng and Samyar Ravanbakhsh for implementing the Matlab code for producing Figure 16.

This work was supported by the Swedish Governmental Agency for Innovation Systems (VINNOVA).

REFERENCES


ABSTRACT

Official traffic counts approximate the amount of traffic observed in roads. These counts are computed by local authorities to model traffic and understand transportation needs. Although there exist automatic traffic collection techniques such as CCTV or road sensors, these tend to be highly expensive. Thus, countries with limited resources typically compute the official traffic counts with manual approaches including individuals counting the number of cars that go through different roads. Given the ubiquity of cell phones in Senegal, we propose the use of cell phone data as a proxy for modeling traffic. Specifically, we design a technique to automatically compute official traffic counts using mobility features extracted from Call Detail Records. We evaluate the technique against official traffic count numbers in Senegal and obtain correlation coefficients between real and predicted values of $r = 0.889$. Our approach provides a reliable technique to measure traffic counts at large-scale and in an affordable manner.

1. INTRODUCTION

Traffic counts approximate the amount of traffic observed on a particular road. These counts are critical for transportation planning and transportation analysts since many policy decisions, including the construction of new roads, are based on such numbers. Currently, there exist two types of data collection techniques to gather information regarding traffic counts: manual and automatic. Manual approaches involve hiring a number of observers who manually count the number of vehicles that drive through specific roads. This approach is not only expensive, but also very hard to scale to cover the bulk of highways and roads in a country. On the other hand, there exists a handful of automatic techniques for monitoring traffic including road sensors, cameras, or toll information among others. These techniques automatically gather road and traffic information with high frequency and generally report average values through specific periods of time e.g., daily averages. Although automatic techniques are easier to scale than manual approaches, they tend to be expensive due to the high costs involving traffic monitoring systems. Another important flaw of these automatic techniques is that due to its cost, only specific critical road segments tend to be monitored e.g., segments with traffic jams or road segments with high percentages of traffic accidents.

In this paper, we explore novel automatic techniques to approximate traffic counts that might alleviate the limitations surrounding both manual and traditional automatic techniques: costs and scalability. This is specially true for emerging economies with limited resources such as Senegal that are looking for alternative ways to measure traffic at large-scale and in an affordable manner. Specifically, we propose the use of Call Detail Records (CDRs) to evaluate traffic volumes on roads. Our proposal is based on the fact that cell phones have become a pervasive sensor of human behavior due to its penetration rates. In Senegal, cell phone penetration rates are well over 94% and the telecommunications infrastructure covers all the country with different levels of granularity mostly between rural and urban environments. Since this infrastructure collects information about the location of the cell phones, we propose to this information as a proxy for traffic count estimation.

The exists an important body of work in the general area of traffic estimation using different sources of sensing data including cameras [9], cell phone data (mostly handover information) [5] or passive data [10] [14]. However, these approaches typically focus on identifying a plethora of traffic issues including road congestion or traffic routes. In this paper, we propose an approach to a problem that although simple, would be very helpful for developing countries with limited resources and for whom traffic counts are a very relevant measure. Additionally, this approach opens the door for the development of cost-effective measures to estimate road conditions in areas where a cell phone network is deployed but no infrastructure to estimate traffic has or can be developed. The rest of the paper is organized as follows: section 2 explains the type of datasets used in our project; section 3 describes the methodology we propose to approximate traffic counts from CDR data and section 4 presents our results. We finalize covering related work and drawing our conclusions in sections 5 and 6.

2. DATASETS

2.1 Traffic Counts

We make use of the latest collection of official traffic counts in Senegal computed in 2002 by national authorities with support from the African Development Bank [2][1]. The official traffic counts contain information for four types of roads: national, regional, departmental and provincial (see Table 1). National roads communicate large administrative regions (Régions) and neighboring states and around 60% of them are paved; regional roads connect the fourteen Régions in Senegal; departmental roads communicate departments (Départements) and around 18% of them are paved; and finally provincial roads connect all other minor urban and production centers. Figure 1 shows a map with the National roads and their official traffic counts (color-coded).

<table>
<thead>
<tr>
<th>Road Type</th>
<th>Network Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>National</td>
<td>3351km, 59.2% paved</td>
</tr>
<tr>
<td>Regional</td>
<td>12056km, 72.0% paved</td>
</tr>
<tr>
<td>Departmental</td>
<td>5667, 18.4% paved</td>
</tr>
<tr>
<td>Provincial and others</td>
<td>4450, 5% paved</td>
</tr>
</tbody>
</table>

Table 1: Types of roads and lengths in Senegal.
To compute the official traffic counts, each road is divided into a set of segments (links) and traffic counts are provided for each of the individual segments. Each traffic count represents the annual average daily traffic (AADT) or average number of daily cars observed in any given road segment throughout a year. Figure 2 presents a partial view of several road segments in National one showing high and medium traffic count values.

2.2 Call Detail Records

Call Detail Records (CDRs) are collected by telecommunication companies for billing purposes. Every time a phone call is made or received, a set of variables are saved including the anonymized cell phone numbers, date and time as well as information regarding the latitude and longitude of the cellular tower that handled the service. The dataset used for this challenge contains CDRs from Senegal collected for a year: from January 1st to December 31st, 2013. In this project, we make use of Dataset 2 which contains fine-grained mobility data on a rolling 2-week basis at an individual level with 300,000 randomly sampled users whose cellular activity spans at least 75% days of the whole year and whose interactions are kept under 1,000 per week.

Figure 3 shows the map of Senegal with the Voronoi polygons approximating cellular coverage areas in different parts of the country. As observed in other countries [8] [12], the density of towers increases in urban areas i.e., the coverage areas are much larger and less granular as we move from urban to rural areas. It is important to highlight that the real position of the cellular towers is unknown since noise has been added due to commercial and privacy concerns. However, the noise added is proportional to the density of cellular towers and preserves the overall structure of the mesh. Next section also covers how this limitation might affect our methodology. Finally, combining both traffic counts and CDR data, the coverage area of a cellular tower can be traversed by one or multiple road segments as seen in Figure 5.

3. METHODOLOGY

In this paper, we propose to use the cellular activity observed in any given tower as a proxy to predict the traffic that goes through the road segments that traverse the tower’s coverage area. In other words, we use the cellular activity computed from the CDR data in Dataset 2 to predict the official traffic counts per road segment collected by the African Development Bank. However, the cellular activity might be due to individuals who are walking by the coverage area or to individuals who are driving by. Since traffic counts report the daily average number of cars observed in a road segment, we need to envision techniques that compute the cellular activity exclusively as a measure of the motorized traffic. In remote areas of Senegal, located far away from cities and towns, it is highly probable that the cellular activity will be mostly due to individuals driving by the cellular towers. However, in urban environments both types of motorized and non-motorized traffic will take place and will be captured as cellular activity. To disentangle the non-motorized traffic from the overall cellular activity we propose two different approaches: Filter Regions and Filter Users.

3.1 Filter Regions

This approach only considers the cellular activity at towers that give coverage to geographical areas that are mostly inhabited. The idea is that by focusing on rural and remote areas of Senegal with very little residential life with respect to its total geographical area, the cellular activity will mostly represent individuals driving on road segments that traverse the coverage area of the cellular tower. For that purpose, we eliminate from our analysis all coverage areas that are located within and nearby towns with populations larger
than 1,000. We have observed that towns with populations smaller than 1,000 are typically covered by a cellular tower that spans a geographical area much larger than the towns themselves. Thus, the amount of people possibly walking by remains very small with respect to the total activity observed for the coverage area, making our approach sound for the identification of motorized traffic. Figure 4 shows an example of the Filter Regions approach. Each circle surrounding a population represents the regions (coverage areas) that we filter from our analysis. The circle size is proportional to the population of the town. With this approach, only the coverage areas not filtered are used to approximate the official traffic counts per road segment. Finally, it is important to clarify that this approach offers the advantage that only aggregate information per cellular tower is used without the need to access individual location patterns from users. However, this approach is only useful for computing traffic counts in roads that cover regions that are mostly inhabited. For this filter, we define the cellular activity in a coverage area $i$ as the average daily number of calls detected in that tower ($\text{ActivityCA}_i$).

![Figure 3: Map of Cellular Coverage in Senegal with Voronoi Polygons.](image)

![Figure 4: Coverage areas eliminated from analysis with the Filter Regions approach.](image)

### 3.2 Filter Users

Instead of filtering out highly populated areas, this second approach focuses on filtering out users that are not driving. To accomplish this, we compute the daily distances traveled by each individual and the time it takes them to cover those distances. With these two values in hand, we can compute speed and filter out users whose speeds are lower than a given value that clearly differentiates walking from driving. In this paper, we fix that value at a safe 10mph since humans have been reported to be able to walk at maximum speeds of 5.6mph [11]. Formally, we can express this second approach as:

$$s_{i,d} = \frac{D(t_j, t_k)}{\text{time}}$$

where $s_{i,d}$ represents the speed of individual $i$ for a given day $d$, $D(t_j, t_k)$ the distance between the two farthest away towers for that user in day $d$ and $\text{time}$ the time it took the user to cover that distance; $\text{driving}_d$ represents the set of individuals $i$ in Dataset 2 that were driving in day $d$. In this second approach, we define cellular activity in a coverage area $i$ as the daily average number of users driving through that area ($\text{ActivityCA}_i$). As a result, the approach requires the access to individual CDR data but has the advantage of being applicable to any urban or rural environment.

### 3.3 Predictive Models

Our goal is to explore predictive models that allow us to use the cellular activity of a coverage area $\text{ActivityCA}_i$ as a proxy for road segment traffic counts. However, coverage areas and road segments do not necessarily cover exact geographical regions. As a result, we cannot simply assign the cellular activity of a coverage area to a road segment that traverses it. Figure 5 shows the representation of part of National 3 with a set of road segments that traverse the coverage areas of several cellular towers (represented by their Voronoi polygons). We observe that different scenarios can occur: part of a road segment traverses the full coverage area, a complete road segment traverses it, or multiple road segments traverse a unique coverage area. To account for all these scenarios, we approximate the cellular activity for a given road segment $j$, $\text{ActivityRS}_j$, as the weighted average of the cellular activity of all the coverage areas that are traversed by the road segment, weighted by the length of the road segment within each coverage area:

$$\text{ActivityRS}_j = \sum_{i=1}^{n} \text{ActivityCA}_i \times \frac{l_{j,i}}{l_j}$$

where $\text{ActivityRS}_j$ represents the cellular activity for road segment $j$, $n$ the total number of coverage areas traversed by the road segment, $\text{ActivityCA}_i$ is the cellular activity in coverage area $i$ computed from CDR data and $l_{j,i}$ is the percentage of the length of the road segment $j$ that traverses coverage area $i$. This is based on the assumption that the cellular activity of a road segment will depend on the cellular activity of the coverage area it traverses weighted by the length of the road segment in that area i.e., the longer the segment, the more activity from the coverage area it captures. It is important to recall that $\text{ActivityCA}_i$ can represent the daily average number of calls or the daily average number of drivers in the coverage area depending on whether we use the Filter Regions or Filter Users approach.

At this point we have, for each road segment in our dataset, a ground truth value with the official traffic counts as well as the approximate cellular activity for that road segment, $\text{ActivityRS}_j$. By combining together these two values, we frame the prediction problem as a supervised problem with the official traffic counts as the dependent variable and the cellular activity per road segment as the independent variable. For this paper, we explore two types of predictive approaches: linear regression models and Support Vector Regressions. We describe each approach in detail. Formally, the linear regression seeks a model such that:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$
where \( \hat{tc}_i \) are the predicted traffic counts for road segment \( i \), \( \alpha \) and \( \beta \) are parameters of the linear regression and \( \text{ActivityRS}_i \) is the cellular activity for that road segment. As explained earlier, that cellular activity can be computed using either the Filter Regions or the Filter Users approach. We fit this model with the Ordinary Least Squares approach.

In an attempt to evaluate non-linear models, the second predictive model we explore is Support Vector Regressions (SVR) with a Radial Basis Function (RBF) kernel. SVR with RBF first maps the input to a higher dimensional feature space using a nonlinear mapping (kernel) and then a linear model is constructed in that feature space. Formally,

\[
\hat{tc}_i = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b
\]

\[
K(x_i, x_j) = \exp\left(\frac{\| x_i - x_j \|^2}{2\sigma^2}\right)
\]

where \( K \) is the kernel function i.e., an RBF in our case and \( x \) represents the cellular activity for road segment \( i \) (\( \text{ActivityRS}_i \)). Similarly to Support Vector Machines (SVM), SVR looks for the hyperplanes that maximally separate the samples in the projected space although a margin of error tolerance is accepted. SVR uses parameters cost \( (C) \) and \( \epsilon \) to apply a penalty to the optimization for points which are not correctly predicted. In our experiments, we use a grid search to look for the optimal values. Finally, given that we use regressions, we describe the performance of both predictive approaches using the correlation coefficients between the predicted \( tc_i \) and the official \( tc_i \) traffic counts. The resulting correlation coefficient will measure the similarity between the real official counts and the predicted counts using the proposed predictive approaches. Next section describes results using the two predictors and the two filtering approaches.

### 3.4 Limitations

As discussed in Section 2, the location of the cell towers in the dataset is not the real one. Some noise has been added to avoid publishing the real locations due to commercial and privacy issues. This might affect our methodology since coverage areas are computed based on the cellular tower locations provided by the challenge organizers. If the cellular towers are noisy, the coverage areas will not necessarily represent the real coverage areas and traffic counts associated to those might be noisy as well. However, as reported in the official challenge paper [6], the noise has been added in such a way that the whole mesh structure for the coverage areas is maintained. Thus, although traffic counts associated to the coverage areas might not be the real ones, overall they do represent general traffic trends across different geographical areas. These trends are the ones that we want the prediction models to capture. Although not ideal, the models will still be able to cope with and distill general trends from the noisy cellular tower locations. Similarly, as stated in [6] not all individuals are reported in Dataset 2 but rather only those with certain lower and upper bounds of activity. Although this could also be seen as a limitation, it is important to highlight that the filter is common across all coverage areas, thus again allowing us to preserve the general trends in our predictive models. Finally, a common limitation in projects that use datasets from different sources is the difference in data collection years: the official traffic counts were collected in 2002 whereas the cellular data is from 2014. Ideally, both datasets should be collected during the same time period. However, the results we report are quite promising despite the time difference in data collection.

### 4. RESULTS

In this section, we report the accuracy of using cellular activity as a proxy for official traffic counts. Specifically, we describe results for the two measures of cellular activity: Filter Regions and Filter Users and the two types of regression: linear and SVR.

To test the accuracy of the prediction algorithms, we divide each dataset into randomly selected training and testing subsets with a distribution of 80% – 20% and repeat the random selection 100 times. For each prediction technique and filtering approach, we report average performance results over all 100 runs. Specifically, we report two measures: the correlation coefficient \( r \) and the root mean squared error (RMSE). Correlation coefficients report the correlation between the predicted and the official traffic counts whereas the RMSE reports the mean error between the two. In the case of SVR, we use a grid search approach to tune the values for epsilon \( (\epsilon) \) and cost \( (C) \) that give best performance in terms of prediction results. The grid search uses a 10-fold cross-validation over the training set to approximate best values.

Table 2 shows prediction performance for both techniques and filtering approaches during training and testing. Parameter values for SVR are \( C = 512 \) and \( \epsilon = 0.3 \) for the Filter Regions approach and \( C = 8, \epsilon = 0.4 \) for Filter Users. First, we observe that SVR obtains better predictive results than Linear Regressions for both types of filtering approaches \( (r = 0.698 \text{ vs. } r = 0.512 \text{ in the best case}) \). This result probably reveals that a non-linear approach...
provides a better fit for the official traffic counts that we want to approximate. To further understand this, Figures 6 and 7 show the official traffic counts (x-axis) and the predicted values during training (y-axis) using linear regression and SVR with the Filter Regions and Filter Users approaches, respectively. As can be seen, the SVR (red crosses in the plot) does a better job at approximating the training samples for both filtering approaches, which also translates into better prediction results during testing as observed in Table 2. The correlation coefficients during testing decrease a little bit probably due to overfitting during training (from $r = 0.704$ to $r = 0.698$).

A second important observation is that the Filter Users approach shows consistently better results than the Filter Regions approach when compared across regression techniques. In fact, the best predictive results were observed for the Filter Users approach and SVR with a correlation coefficient between official and predicted counts of $r = 0.698$ during testing ($r = 0.704$ for training). However, it is also fair to say that the accuracy for SVR and Filter Regions approach had a correct performance with a correlation coefficient of $r = 0.572$ during testing. From these results, it appears that filtering users based on their speed is a more robust approach than just filtering out road segments in urban-like environments.

In an attempt to improve our results, we explored the set of roads for which the traffic count predictions were the worst. We observed that both predictive approaches were doing a poor job with some Regional, Departmental and Provincial roads. As explained in Table 1 these roads are mostly unpaved and the traffic conditions tend to be really poor. Thus, we recomputed the linear regression and SVR for both filtering approaches but only considering National roads. Table 3 shows the prediction accuracy across all scenarios. In general, we observe similar trends to our analysis with all roads:

(1) SVR shows more predictive power than Linear Regression for both filtering approaches and (2) Filter Users is more predictive than the Filter Regions approach. However, the main change is that the correlation coefficients between the official and predicted traffic counts were highly improved. In fact, the most accurate results are reported for SVR and the Filter Users approach with correlations of $r = 0.905$ during training and $r = 0.889$ during testing.

Overall, our results indicate that traffic counts can be predicted from cellular activity with good accuracy for analysis including all types of roads and with high accuracy for analysis that only focus on National Roads.

### 5. RELATED WORK

The existence of a wide range of papers exploring the identification of traffic issues including road congestion or traffic routes using different sources of information such as cameras or GPS information [9][13]. Focusing on cell phone infrastructures (GSM and UMTS), Bar-Gera used cell phone data to measure travel speeds and travel times in Israel [3]. The author compared cell phone-based measures with those obtained through dual magnetic loop detectors and showed that there were significant correlations between the two. A similar approach was presented by Chi et al. to extract traffic information exploiting handover patterns from UMTS signals [5]. Moving on to traffic congestion, Janecek et al. used data from a cellular mobile network to compute road congestion [10]. They validated their approach against four different traffic monitoring datasets and showed that their methods can detect traffic congestion very accurately and in a timely manner. Other interesting work has also been done in the area of route classification [4]. Becker et al. showed how to use handoff patterns from cellular phone networks to identify the routes that people take through a city. The authors presented two classification algorithms to match cellular handoff patterns to routes and validated their methods using statistics provided by a state transportation authority. A similar approach to ours was presented in [7], however the authors used passive network data instead of simply CDRs, making our approach much more challenging. Overall, these approaches typically focus on identifying a plethora of traffic issues in developed countries with plenty of CDR information across customers. Our work aims to develop simple mechanisms with high social impact to be able to approximate travel counts in emerging regions with few resources.

### 6. CONCLUSIONS

In this paper, we have presented a technique to automatically approximate official traffic counts using mobility features extracted from Call Detail Records. We have evaluated two approaches to detect motorized traffic versus individuals walking: Filter Regions and Filter Users. The former focuses on eliminating highly populated areas where it might be harder to differentiate walking from driving whereas the latter directly eliminates individuals who are not driving. Our results show that SVR together with the Filter Users approach are highly predictive of the traffic counts with correlation coefficients between real and predicted of $r = 0.889$ during testing. As a result, we believe that this approach provides a reliable technique to measure traffic counts at large-scale and in affordable manner.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Filter Regions</th>
<th>Filter Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>$(r=0.434, RMSE=1676)$</td>
<td>$(r=0.588, RMSE=1391)$</td>
</tr>
<tr>
<td>Test</td>
<td>$(r=0.402, RMSE=1701)$</td>
<td>$(r=0.512, RMSE=1427)$</td>
</tr>
<tr>
<td>SVR</td>
<td>$(r=0.684, RMSE=1221)$</td>
<td>$(r=0.704, RMSE=915)$</td>
</tr>
<tr>
<td></td>
<td>$(r=0.572, RMSE=1376)$</td>
<td>$(r=0.698, RMSE=1218)$</td>
</tr>
</tbody>
</table>

Table 2: Prediction correlation coefficients $r$ and RMSE for linear regression and SVR: All Roads.
<table>
<thead>
<tr>
<th>Technique</th>
<th>Filter Regions</th>
<th>Filter Users</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>( r = 0.604 ), ( \text{RMSE} = 1243 )</td>
<td>( r = 0.584 ), ( \text{RMSE} = 1321 )</td>
</tr>
<tr>
<td>SVR</td>
<td>( r = 0.865 ), ( \text{RMSE} = 803 )</td>
<td>( r = 0.748 ), ( \text{RMSE} = 906 )</td>
</tr>
</tbody>
</table>

Table 3: Prediction correlation coefficients \( r \) and RMSE for linear regression and SVR: Only National Roads.

7. REFERENCES

Towards Enabling Mobile Social Crowd-Sensing for Unstructured Transport Information Management: Performance Evaluation of Large-Scale End-to-End Publish/Subscribe Interaction

Georgios Bouloukakis\textsuperscript{a}, Nikolaos Georgantas\textsuperscript{a}, Rachit Agarwal\textsuperscript{a}, Animesh Pathak, Valérie Issarny\textsuperscript{a}

\textsuperscript{a}Team MiMove, Inria Paris - Rocquencourt, France

Abstract

Developing countries are characterised of chaotic large-scale traffic, especially in major cities. Robust mobile systems for information concerning transport services are of critical importance. Additionally, although most of the people have mobile phones, even smartphones, a large part of the population rely on SMS data access only. The situation of Senegal reflects the above. In this paper, we take a first step towards enabling an application platform for citywide and countrywide transport information management relying on ‘mobile social crowd-sensing’. To inform the stakeholders of expected loads and costs, we model a large-scale mobile publish/subscribe system as a queueing network. We introduce additional timing constraints such as (i) mobile user’s intermittent connectivity period; and (ii) data validity lifetime period (e.g. that of sensor data). Using our MobileJINQS simulator, we parameterize our model with realistic input loads derived from the D4D dataset and varied lifetime periods in order to analyze the effect on response time. This work provides system designers a coarse grain design time information when setting realistic loads and time constraints.

Keywords: Publish/Subscribe, Mobile Social Crowd-Sensing, Large Scale, Queueing Networks, Transportation Systems

*Email addresses of the authors are \{firstname.lastname\}@inria.fr

Preprint submitted to D4D Challenge December 31, 2014
1. Introduction

A reliable and efficient transportation system is a key growth indicator for a country. The transportation system of Senegal, although developing, still consists of many unplanned and informal settlements with unreliable services and infrastructure [11, 2]. This applies to both countrywide transportation system (e.g., road network and buses, railway) and urban transportation system (e.g., taxis and buses in Dakar). Particular conditions include high traffic of people and merchandise between Dakar and the rest of the country, and movement of population during religious festivals. In this context, the native population and the visitors greatly benefit from information concerning transport services and schedules, as well as timely reports about transport conditions and unexpected events. However, use of the Internet by transport providers and travellers in Senegal remains limited. Indicatively, recent reports show that 31% of SMEs in Senegal have Internet access, with only 7% having a website [1]. Additionally, despite the fact that most of the people have mobile phones, only 26% of them have mobile Internet access [7] in Senegal. The above percentage is almost equal to the total Internet access, as use of fixed Internet remains very low. Hence, for a large part of the population, the only alternative for data access is SMS.

Considering the above limitations in Senegal, we propose the development of an application platform for citywide and countrywide transport information management relying on ‘mobile social crowd-sensing’. Leveraging the massive adoption of mobile phones by the population, mobile social crowd-sensing enables a user to ‘sense’ and report on the environment with respect to personal, social or public context. This complements the objective and the authoritative information coming from structured information sources and compensates for the lack of such information. Such a rich data, after careful processing, can be used not only to directly inform the population, but also to develop advanced services for local stakeholders and to drive studies and policies for state authorities. Our application platform would enable development of mobile applications
and related support systems for transport information in Senegal, hence contributing to the improvement of travel experiences in the country.

In our approach, we study and experiment with appropriate interaction styles (based on message-passing, events, data sharing) on top of 3G/2G/SMS data connections (or even future 4G [10]), further depending on the specific application and data. For this, we rely on our ongoing work on Extensible Service Bus, (XSB) [5], a framework for handling interconnection between interaction protocols applying the above-mentioned interaction styles. We are particularly interested in interaction adaptation depending on the network conditions (e.g., switching to SMS-based protocol when the 3G/2G network is unavailable). Furthermore, we deal with large-scale sensing interactions, as we are targeting large-scale deployments. For this, we build upon our recent work on MobIoT [6], a service-oriented middleware enabling efficient sensing over ultra-large population of mobile sensors.

In this paper, we take a first step towards enabling an application platform for transport information management based on mobile social crowd-sensing keeping in mind the particular context and constraints of Senegal. This consists of evaluating the publish/subscribe interaction style in a large-scale setting where resources of mobile users are limited, which translates into limited and intermittent connectivity in the system. We have opted for the publish/subscribe paradigm, as it is deemed appropriate for spatio-temporal interaction between mobile entities.

More specifically, we introduce a queueing network model for the end-to-end interaction within a large-scale mobile publish/subscribe system. We leverage the dataset provided by Orange Labs to parametrize this model. We refer to this dataset as the D4D dataset [9]. We then develop a simulator named MobileJINQS that implements our model and uses the dataset traces as realistic input load to the system model over the time span of a whole year. Prior to this, we extensively analyze the D4D dataset in order to identify the data

that we are interested in and infer primary results. Based on the results of our simulation-based experiments, we thoroughly evaluate the behavior of the publish/subscribe system and identify ways of tuning the system parameters in order to satisfy certain design requirements.

The rest of the paper is structured as follows: in Section 2, we summarize our analysis of the D4D dataset. In Section 3, we introduce our model, its parametrization based on the dataset, and our simulator. Then, we present, in Section 4, our simulation experiments and their outcomes. We finally conclude this paper in Section 5.

2. D4D Dataset Analysis

In this section, we provide a brief description of how we analyze the real data provided to us by Orange Labs in order to parametrize and feed our simulation model described in the next section. The dataset or the D4D dataset contains Call Detail Records (CDR) of the users that are subscribed to the Sonatel services in Senegal. This data is collected over a period of 1 year (Year 2013). The D4D dataset consists of 3 sub-datasets: Dataset1 consists of number of calls and duration of calls made during an hour, Dataset2 consists of fine grain spatial user mobility trace, while dataset3 consists of coarse grain spatial user mobility trace. It should be noted that as the dataset collected is a CDR, the logs in Dataset2 and Dataset3 are made only when the user initiates a call or sends an SMS.

From the D4D dataset we chose Dataset2 for the realization of the model on a large scale. We choose Dataset2 due to the fine spatial granularity of the dataset. Further, properties of Dataset2 include a) temporal granularity of 10mins b) observations spanning over the year 2013 c) track of almost 300,000 unique users. Dataset2 reflects the mobility pattern of a single user and also the number of users associated to a given antenna in a given time interval. This lead us to first extract number of people associated to a given antenna in a given time interval.
Let $A$ be the set of Antennas, $T$ be the set of unique time intervals over which the data is collected and $U$ be the set of users tracked in the Dataset2. Let $P$ be a $|T| \times |A|$ matrix.

**Definition 1.** Let $N_t^i$ be the number of people associated to antenna $i \in A$ at a given time $t \in T$ then

$$P_t^i = N_t^i$$

Note that, $N_t^i$ also denotes number of connections made to the antenna $i \in A$ at a given time $t \in T$. From $P$ we extract how many users are present at a given time interval on a country scale.

**Definition 2.** Let $N_c$ be the matrix of size $|T| \times 1$ such that element $N_c^t$ is the number of users in a country at time interval $t \in T$. Then

$$N_c^t = \sum_{i \in A} P_t^i$$

Let $MaxN$ be a matrix of size $|A| \times 365$ where $MaxN^d_i$ represents the max number of people in an antenna $i$ on a given day $d$.

**Definition 3.** Let $Day$ be the set of unique days of the year. Let $TimeSlot$ be the set of all unique time intervals on a given $d$, then for $d \in Day$

$$MaxN^d_i = \max_{t \in TimeSlot} (P_t^i)$$

Note that $T \supset TimeSlot$.

We provide visualization of $P$ from Jan 07, 2013 00:00:00 until Jan 20, 2013 23:50:00 (Cf. Fig. 1). An animated version of the same is available here[^visual]. From the visualization we observe that most of people are located near Dakar region and in the major cities in the west and north west of Senegal (north of Gambia). The population graphs also show less utilization of infrastructure during the night hours. It should be noted that the antenna IDs are marked from west to east (Cf. Fig. 2). The Fig. 3 shows $N_c^t$.

[^visual]: http://xsb.inria.fr/d4d#visualization
From the Def. 3 we extract \( \max_{t \in T} (N_t^c) \), the max number of people in the whole system (Country) at any given time interval \( t \) such that \( t \in T \). From the Dataset 2 \( \max_{t \in T} (N_t^c) = 112,937 \) observed on Thu, 08 Aug 2013 from 23:00:00 to 23:10:00 GMT, i.e., on the End of Ramadan, a public holiday in Senegal. Note that this number is different from the number of unique users tracked because some users might appear only at some distinct times while being inactive for other times. The Fig. 4 provides the visualization of \( MaxN \). Fig. 4 shows that over a day, antenna IDs in \( A' = [1, 580] \) have lot of users as compared to users in \( A - A' \). It should also be noted that there are some antennas where the users were not tracked over the entire period of the year 2013. This is also represented in the Fig. 4.
3. Publish/Subscribe Performance Model

Our main objective in this work is to enable large-scale sensing interactions that are part of a mobile social crowd-sensing application. Particularly, we target an application platform for citywide or countrywide transport information management based on unstructured information posted by mobile users. In terms of communication infrastructure support, we rely on the publish/subscribe interaction paradigm that provides loosely coupled form of required interaction, especially in large-scale environments. Additionally, such an application platform must guarantee that the sensing data is processed and delivered to the corresponding mobile users on-time, despite the intermittent connectivity of the latter for resource-saving purposes. To ensure the freshness of the delivered data, events are characterized by a validity period, after the expiration of which
they are discarded by the system. In the following sections, after a brief presentation of the publish/subscribe interaction paradigm, we introduce our model that encompasses the above concerns. We then detail the parameterization and simulation of the model based on the D4D dataset.

3.1. Publish/Subscribe interaction paradigm

In the publish/subscribe interaction paradigm, multiple peers either take the role of publisher or a subscriber and interact via an intermediate broker. Publishers produce events characterized by a specific topic to the broker. While subscribers subscribe their interest for specific topics with the broker, who maintains an up-to-date list of subscriptions. The broker matches received events with subscriptions and delivers a copy of each event to each interested subscriber. This style of interaction is decoupled. In terms of space coupling, interacting
peers do not need to know each other. Events are diffused to subscribers only based on topics (one-to-many interaction). In terms of time coupling, peers do not need to be present at the same time during interaction. Subscribers may be disconnected at the time when the events are published to the broker, who then keeps the events in a dedicated buffer for each subscriber. Thus, subscribers receive the pending events when reconnected.

3.2. Large-scale mobile publish/subscribe system

A large-scale publish/subscribe system relies on a network of brokers. It may be the case that all brokers are accessible to publishers and subscribers, or that there is a distinction between edge brokers (brokers at the periphery of the network) and brokers situated in the network backbone (core brokers). In both cases, each broker propagates the subscriptions it receives to its neighbors,
thereby populating the routing table of each broker.

In a large-scale mobile publish/subscribe system, publishers and subscribers are mobile entities. Interactions then rely on volatile network access to the nearest edge broker. Publishers connect to the system to publish events. They may set a lifetime limit to these events, depending on the nature of the specific application and data. Subscribers connect occasionally to the system to receive new events, and disconnect to save energy. While moving, a subscriber may hand off between brokers. This results in each broker updating their subscriptions and routing table as well as transferring between them the events stored for the subscriber but not yet delivered.

3.3. End-to-end interaction model of a large-scale mobile publish/subscribe system

In our study, we focus on end-to-end interaction between publishers and subscribers going through two access points of the system, an input access point and an output access point. The input access point is provided by an edge broker. Geographically close mobile publishers generate input flow to the whole system through this access point. The output access point is also provided by an edge broker. Geographically close mobile subscribers receive output flow from the whole system through this access point.

Figure 5: End-to-end interaction between publishers and subscribers through two system access points

The input access point is modeled as a queueing system where the service
delay $\mu_{in}$ represents matching incoming events (input flow $\lambda_{in}$) to subscriptions and routing them to corresponding subscribers and/or brokers (part A of Fig. 5). This may involve replication of some events. This may also involve dropping of some events, either due to no corresponding subscription or due to expiration.

The output access point is modeled as a queueing system where the service delay $\mu_{out}$ represents the transmission of buffered events to the intermittently connecting subscribers previously determined by routing (output flow $\lambda_{out}$), as depicted in part C of Fig. 5. Here again, some events may expire before being delivered to their destination.

We focus only on the input and output processes. We do not consider, for the moment, the rest of the publish/subscribe system (i.e., multiple input and output access points, routing of events among multiple brokers). Additionally, we link directly the output of the first queueing system to the input of the second one, considering the path through the publish/subscribe system as a simple wire. This is equivalent to having a single centralized broker for the whole publish/subscribe system. Again, in this paper, we intend to restrict our model to only certain features.

3.4. Use of the dataset to parametrize the model

For parametrizing our model, we rely on our D4D dataset analysis presented in Section 2. More particularly, we are interested in $P_t^i$, the number of users associated to a given antenna $i \in A$ (or number of connections to this antenna) in a given time interval $t \in T$ over the whole recording time period $T$ of Year 2013. We select from the set $A$, $A'$ such that $A \supset A'$ where $A'$ consists of antennas $i$ with large $\text{Max}N_d^i$. Note that, these antenna IDs are between 1 through 579. These antennas are located mainly in the Dakar region and in the major cities in the west and north west of Senegal. While loads of the selected antennas are quite different in terms of mean value and variance, almost all of them present their peak around the End of Ramadan, as identified in Section 2.

We map the two access points of our publish/subscribe system model to two antennas identified in the dataset. We use, in particular, the trace providing the
number of connections to an antenna every 10 minutes for a period of 50 weeks. Each such connection corresponds to the initiation of a call or the emission of an SMS from the specific antenna. We make the assumption that this trace can equally represent the reception of a call or an SMS through the antenna. For the input access point, we map the number of connections per 10 min interval at the selected antenna to an equal number of events published over the same time interval.

**Definition 4.** Let $\lambda_{in}$ be the input process at the input access point associated to the antenna $i \in A$, and $P^t_i$ be the number of connections to the antenna $i$ in each interval $t \in T$, as defined in Def. 1. Then $\lambda_{in}$ is a non-homogeneous Poisson process with rate parameter $\lambda(t)$ piecewise constant in each interval $t \in T$: 

$$\lambda(t) = \frac{P^t_i}{|t|}$$

For the output access point, we map the number of connections per 10 min interval at the selected antenna to an equal number of events delivered to subscribers over the same time interval, provided that there are enough events in the queue. This relates to the number of subscribers that are connected to the access point during the specific interval, i.e., that are available to receive new events waiting for them if any. To provide this effect, we model the service process at the output access point accordingly.

**Definition 5.** Let $\mu_{out}$ be the service process at the output access point associated to the antenna $j \in A$, and $P^t_j$ the number of connections to the antenna $j$ in each interval $t \in T$, as defined in Def. 1. Then $\mu_{out}$ is a non-homogeneous Poisson process with rate parameter $\mu(t)$ piecewise constant in each interval $t \in T$: 

$$\mu(t) = \frac{P^t_j}{|t|}$$

This is equivalent to having service time of events that follow an exponential distribution with mean equal to $\frac{1}{\mu(t)}$ for the interval $t$.

The output $\lambda_{out}$ of the queueing system modeling the input access point is a process similar to the input $\lambda_{in}$, provided that the queueing system does not
saturate. Then, the $\lambda_{\text{out}}$ is fed to the input of the queueing system modeling the output access point. We also need to take into account the possible replication and dropping of events; this results in $\lambda_{\text{out}} \neq \lambda_{\text{in}}$. Without loss of generality, we consider that in our setup there is no replication of events or dropping due to the absence of corresponding subscription; however, events may still be dropped due to expiration.

Since interaction in the publish/subscribe system is time-decoupled, any two antennas can be selected. Nevertheless, having $\lambda_{\text{out}}$ (or $\lambda_{\text{in}}$) $> \mu_{\text{out}}$, in terms of overall mean values or over specific time intervals, will result in high numbers of expired events. This is when the event input flow to the system is higher than the event delivery flow constrained by the limited availability of the subscribers.

3.5. Simulation of the model

We have developed a simulator that implements our model and uses the dataset to parametrize the model. The results of our experiments enable thorough analysis of the behavior of the system and can assist a system designer in the configuration of the system parameters in order to satisfy a set of design requirements.

Our simulator, MobileJINQS, is an open-source library for building simulations encompassing constraints of mobile systems. MobileJINQS is an extension of JINQS, a Java simulation library for multiclass queueing networks [4]. JINQS provides a suite of primitives that allow developers to rapidly build simulations for a wide range of stochastic queueing network models. However, JINQS only supports the base characteristics of the queueing networks.

MobileJINQS retains the generic model specification power of JINQS, while providing additional features of interest to mobile or other systems such as: (i) lifetime limitation for each customer entering a queue, (ii) intermittently available (on-off) queue server or server with variable service rate over time to represent mobile users’ behavior, and (iii) input flow with variable customer arrival rate over time to represent real input dataflow traces. In particular, a system designer is able to set lifetimes, on-off intervals, as well as variable
service rates and arrival rates following well-known probability distributions.

By relying on the D4D dataset traces as detailed in the Section 2, we setup and execute a set of experiments with MobileJINQS. Our experiments and their results are presented in the following section.

4. Simulation Results

In this section, we provide results of simulations using MobileJINQS of our publish/subscribe system with varied incoming loads, service delays and lifetime periods. We use the dataset to derive realistic traces for incoming loads and service delays. System designers are able to tune the system by selecting appropriate lifetime periods. We demonstrate that varying incoming loads and service delays has a significant effect on response time. In the case of varying lifetime periods, the tradeoff involved between the rate of successful event transactions and response time is also evaluated.

4.1. Selecting representative input load

As identified in Section 3.4, we select from the set $A$, set $A'$ such that $A \supset A'$ and $A'$ has antennas with large $MaxN_i$. Then, traces are derived through number of connections to an antenna at each 10 min interval for a period of 50 weeks. To perform our simulations with varied traces, we classify the load of each selected trace into three main categories: (i) low load antenna; (ii) medium load antenna and (iii) high load antenna. Fig. 6 depicts four antennas used for the experiments of this section. Antenna 9 has a low load trace with overall average rate 0.04 number of connections for a period of 50 weeks. Antennas 24 and 14 have medium load traces with overall average rates 0.075 and 0.082, respectively. Finally, antenna 161 has a high load trace with overall average rate 0.129.

4.2. Response time of our simulation model

Fig. 5 represents the model used for the simulation. To get end-to-end response time between input and output access points, we tune the system with
parameters: (i) input flow $\lambda_{in}$; (ii) in service delay $\mu_{in}$; and (iii) out service delay $\mu_{out}$. In this experiment, to avoid high response time or event expiration rate we choose $\lambda_{in} < \mu_{out}$.

At the input access point, we map the load of Antenna 9 to the input flow $\lambda_{in}$. Particularly, we correspond the number of connections of an antenna to the number of events for publishing at each 10 min interval over the whole period. The service delay $\mu_{in}$, that is responsible for the transmission of events into the output access point, is set to 1 event per second. This means that we consider rapid event transmissions to avoid event expirations at the input access point.

At the output access point, the output flow $\lambda_{out}$ is equal to the input flow $\lambda_{in}$, provided that our queueing system does not saturate. For the service delay $\mu_{out}$, we map the load of antenna 161. Particularly, we correspond the number of connections of an antenna to the number of connected subscribers at each 10 min interval over the whole period. At a 10 min interval each connected subscriber is receiving one single event (i.e., subscribers’ connectivity rate = 1). Finally, in this experiment we consider infinite lifetime period for each event transaction.

Transaction response times are shown in Fig. 7 over a period of 50 weeks where users’ connectivity rate equals to 1. We depict samples of response times every 2-weeks period. The minimum response time for all end-to-end transac-
Figure 7: End-to-end transaction Response Times from low load Antenna 9 to high load Antenna 161

From 1st week to 4th week, the response times is 40 mins due to the fact that subscribers may be disconnected for a long period. Between 5th and 15th week, we get a higher response times (73 mins). This implies that, rate of incoming events is much more higher than the rate of subscribers' connectivity.

Consequently, getting lower response times depends on subscribers’ behavior. Fig. 7 shows response times if we change the subscribers’ behavior by multiplying connectivity rate with 10. This means that and they connect 10 times more for receiving events. By this way, the maximum and minimum response times are 3.5 mins and 1.8 mins respectively.

4.3. Response time vs Success Rate with varying lifetime periods

Based on the previous experiment, the only way to off tune the system in order to satisfy certain design requirements (lower response time), is to change the subscribers’ behavior. In fact, a system designer cannot change this. Therefore, it is essential to keep subscribers’ behavior invariant and find other ways to satisfy our design requirements. Data sent from input access point is valid for a limited lifetime period (in case of transportation, data of a traveler reporting some incident is valid for a lifetime period, for instance).

In this simulation experiment, we keep the same settings as previously and we introduce the lifetime period at each input event. Lifetime period is following an
exponential distribution with mean values of 1, 10, 20 and 30 mins. End-to-end response times and success rates are shown in Fig. 8. As expected, with higher levels of lifetime periods, we notice high success rates, but also much higher response time over the 50 weeks period. More specifically, by increasing the lifetime period, response time gets higher quickly, but the success rate increases very little. When lifetime period equals to 1 min, we get the minimum response time equals to 9.5 seconds with message success rate 62%.

Then, we simulate different pairs of antenna loads at the input/output access points. We use the antennas shown in Fig. 6 and perform experiments for 3 different types of end-to-end transactions: input access point with medium load to output access point with (i) low load; (ii) medium load; and (iii) high load antennas.

Fig. 9 shows the mean value of response time and success rate for the 50 weeks period. For the first transaction, success rate is very low (0.04 %) even by using 30 min of lifetime period. In the other hand, response time is much higher comparing to the other pairs. For the second transaction, success rates are between 10 - 20 % with response times between 2 - 7.5 mins. Finally, for the third transaction, we get 35 - 50 % success rate with reasonable response times of 0.3 - 2.2 mins.
5. Conclusions

In this work, to enable mobile social crowd-sensing for unstructured transport information management in Senegal, we study the behavior of the underlying communication infrastructure under load. We rely on the publish/subscribe interaction style to provide loosely coupled form of the interactions with additional constraints, in large scale environments. Properties such as intermittent connectivity of mobile users and freshness of delivered events, are modeled using a queueing network for an end-to-end interaction. MobileJINQS, implements our model and leverages incoming loads and service delays derived from the D4D dataset. We demonstrate that varying incoming loads and service delays have a significant effect on response time. Furthermore, by introducing varied lifetime periods in the published events, we evaluate the trade-off between the rate of successful event transactions and response time. Our future work includes comparison of publish/subscribe interaction paradigm with other interaction paradigm (message-passing, data sharing), in relation with the network access capacity and the application requirements. Also, we intend to study the response time and success rate for the various combinations of antennas in more fine-grained scales (e.g., check what their evolution is over one day).
6. Acknowledgement

The authors thank the D4D organizers for providing the data.

7. References


Abstract — In the following work we present a short description of data mining methods & techniques applied to creating an infrastructure development and planning methodology using the data cleaved from cell phone usage patterns. The tower-to-tower communication data was utilized to derive clusters in which people commute across Senegal. In these clusters dominant user trails were identified. Time variant statistics on these trails across the year, highlighting congestion, and inferences for one communication cluster are presented in detail.

Index Terms — Spatial Temporal Clustering, Markov Cluster Algorithm, Sequence Pattern Mining

I. INTRODUCTION

In the ‘Data for Development Senegal’, participants have been provided with cell phone usage data at various level of granularity. Cell phone usage logs from 1666 cell phone towers were provided in 3 data sets containing different sets of information. Dataset 1 contained tower to tower traffic for one year for each of the 1666 sites on an hourly basis. Dataset 2 contained tower level mobility data on a rolling 2-week basis, with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled individuals for each 2 week period. Dataset 3 contained coarse-grained (123 arrondissement level) mobility data with bandicoot behavioral indicators at individual level for approximately 150,000 randomly sampled users over the duration of one year.

In a recent report [1] on progress related to infrastructure improvement for Senegal released by World Bank, one of key challenges identified is increasing rural accessibility beyond the centers close to existing network of roads. This is pertinent given the population density per KM, as shown in Figure 1, and the ongoing investments in Senegal related to infrastructure development.

We propose an ensemble of clustering and sequential pattern recognition to identify the most frequently travelled routes by the residents in Senegal, and measuring the congestion across all the significant routes. In order to highlight rural infrastructure needs, our clustering schema identifies high communicating and closely knit towers across Senegal. Within these clusters, as previously mentioned, we identify common trails and highlight statistics on congestion observed across the year.

II. ASSUMPTIONS & HANDLING DATA UNCERTAINTIES

In computing the usage and travel patterns using the provided data sets a few assumptions were made for pattern detection. Since GPS (and therefore velocity information) was unavailable, it is assumed that all cell phone data is from users that were on the move in using transport systems (public and private). Due to the truncated time intervals in data set 2, all the users in dataset 2 are assumed to be independent of one another. In other words, no user is repeated in each 2-week window. Additionally, the data set also does not contain information of tower switching in overlap areas, therefore the data from towers within a 5 kms of each other were considered as single towers. Finally, since the data only tracks the point at which a phone call was made, the information on the individual’s position between phone calls is imprecise. A cut-off point was assumed for cases that appeared to involve air travel.

III. METHODOLOGY

The process flow for the analysis is depicted in Figure 2. As the first step, we cumulated the dataset 1 to generate the number of communications or pings between each pair of towers, both for call and number of texts over the course of the year. For every pair of nodes $i$ and $j$, we come up with $C_{ij}$.
and $S_i$ for number of calls and number of texts from tower $i$ to tower $j$ respectively. We then normalize the two functions to arrive at the total association between the two towers. However, for simplicity, we have just considered the number of calls as our measure of association.

The association matrix derived as a result, is not symmetric but a directed graph. Using the Markov Cluster Algorithm [2, 3], a variation of a graph clustering algorithm, on the association matrix derived, we identified tower clusters indicating social groups within which mobile communication was observed to be high across the year. Prior to mining dominant user trails within these communication clusters, a hierarchical clustering within these communication clusters was performed to reduce the unwanted noise that may crop up due to proximity between the towers.

We allocated each of these users to one of the communication cluster where their activity (presence) was a maximum. Further, we masked user trails in dataset 2 with the results of hierarchical clustering. Finally for each communication cluster we identified the dominant trails shown by users for a 2-week period. We used Fourier-Closed+time algorithm from the open source tool Sequential Pattern Mining Framework [SPMF, 4] to mine these sequence patterns with time constraints.

The rules obtained were cleaned to a desired format and were later converted into tuples and those repeating were retained only once with the maximum support. Within these tuples, we retained the ones where the Greater Circle Distance is greater than a predefined threshold to reduce noisy patterns generated due to proximity between towers. Figure 2 describes the process graphically. Dominant tuples found using pattern mining consist of two towers each and further the tuples indicate a frequent path where the users are frequently moving from one tower in tuple to another tower in the same tuple.

For identifying congestion in the extracted tuples, from dataset 2 we mine out biweekly time series on total number of people using the tuple route. Similarly we also mine out the time series of median time (and speed) taken for the tower change in a tuple path across users in the two week period across the year. From these time series we draw out further inferences on how congestion was on these tuple paths across the year.

IV. RESULTS

Using the methodology presented in the previous section on datasets 1 and 2, a brief summary of our results include:

1. A total of 62 communication clusters were identified post application of a graph clustering algorithm.
2. We identified 60 tuple paths using a threshold of 5KM significance.
3. In 9 communication clusters, no dominant trails or tuple paths were observed. In these cases, though user trails were observed, these trails did not qualify with minimum support levels to be labeled as a dominant path. We present these 9 communication clusters on the Senegal map in Figure 3, these results have to be interpreted in conjunction with Figure 1 and further analysis needs to be done with respect to sensitivity of threshold level of 5 KM.

4. Of the 60 tuple paths – 27 of these were observed within clusters and the rest across clusters indicating movement of users not only within their communication cluster but across neighboring clusters as well.
5. The dominant paths mined covered about 23 communication clusters – coverage of 43% (excluding the 9 clusters where dominant trails were found) – indicating a good coverage of the rural areas in Senegal. The coverage can be increased by reducing the significant threshold distance of 5 KM. However this has to be done with guidance from domain to reduce noise.
6. Histograms of median speeds observed for tower change on tuple paths is shown in Figure 4. From Figure 4, we clearly observe 3 distinct modes of travel, 2 airplane modes in Figure 4a and, a mode corresponding to walking and driving in Figure 4b. We would like to point to readers that these plotted speeds are confounded with the effect of times when the call is made prior to leaving a location and arriving at new location on the tuple path. Due to this confounding the estimates could be conservative; nonetheless the inference on the modes is valid.

7. In certain cases we also observed dominant tuples resulted in closed path structures too.
In Figure 4 we demonstrate the working of our methodology for the Fatick region in Senegal.

- a. Communication Cluster for Fatick Region
- b. Combining Towers in Close Proximity within a Communication Cluster – Hierarchical Clustering
- c. Identifying Dominant Trails within a Communication Cluster

Fig 5. Identifying Dominant Trails within a Communication Cluster

After finding a communication cluster corresponding to this region in Figure 3a, we combined clusters occurring in 5 KM radius in Figure 3b. In Figure 3c we highlight the dominant trails that were identified. The trails identified in this case are bidirectional.

Further exploring issues related to congestion in Figure 6 we plot the bi-weekly time series plot of total users travelling from b to a (b-a) and b to c (b-c).

From Figure 6a we infer that traffic volumes from b-a are significantly high compared to b-c throughout the year. Further in Figure 6b we also find that the instantaneous point wise correlation between both the series are also high.

Next we further investigate on the median velocity observed on b-a & b-c over the biweekly periods of the year in Figure 7 and find the median velocities over the year have been similar. Overall we infer given low volume flows on b-c, infrastructure can be improved to decrease the travel time. Following this improvements on trail path b-a are also warranted, in comparison to the existing conditions. Herein too we would point to our readers that we assume the median velocity is obtained using the median time for transfer between towers on a trail path. We assume the median of this time distribution corresponds to average travel time between the two locations.

In concluding the analysis for the selected region, we also performed a spectrum analysis on both volume flows and velocity time series for each trail path b-a & b-c. A sample spectrum graph is presented in Figure 8 from which we observed that for all the time series considered a cycle of 10 week corresponding to 2 months was prevalent.
This phenomenon requires further investigation by domain specialists, and contrasted with the recommendations suggested to arrive at optimal infrastructure decisions for the Fatwick region. Similar analysis can be performed for the other dominant trail paths across other regions too.

V. SUMMARY

A combination of data mining techniques and visual analytics can help in understanding the data better and help identify regular patterns and abnormalities in data. The findings from the data mining exercise showcase that with further granularity in the data along with the GPS information will enable the development of infrastructure by determining the high traffic regions. The information may also be used to determine the optimal location of emergency services like fire trucks, ambulances, police etc.

ACKNOWLEDGMENT

The work for this effort was conducted as part of the Corporate Research Robert Bosch Engineering and Business Service. We would like to acknowledge Dr. Rajesh Nagraja Rao’s efforts towards meeting our project goals. Herein we would also like to highlight the valuable suggestions by Manik Singhal, Prakash Lekkala and other team members.

REFERENCES


The municipality and the territory: two scales of understanding the city
Project Summary

Working on phone calls and their intensities, it is possible to show the network of cities of Senegal and therefore allow a diagnosis of the situation of the urban system (how it works?). Should result a series of recommendations for land planning, public policy development, development of future infrastructures and thus tender to a sustainable land use.
Possible use for development

The project aims to develop new public policies and provide a diagnosis for the development of a sustainable land use. The hierarchy of cities provides guidance on the planning of networks (roads, electricity, water supply, …). It shows the imbalances between the territories (Dakar and the rest of the country) and give keys to correct it.
Main Results

1. The calls and their intensities clearly show how the urban system works, they also show the economic role of the different regions, cities or municipalities.

2. Dakar unbalance the network, in terms of land use that would argue for a greater balance between Dakar and the "desert" of Senegal.

3. The important role of natural barriers, like the Gambia, but also the areas of "fossil valleys".

4. Calls are close to relatives, from one city to another with the exception of Dakar.
Methods

The three maps were produced with the same set of data, but by three different methods:

Map 1: Calls and intensity.
The thickness of the lines is the number of calls between antenna A and B. The colour corresponds to the intensity (intensity = (calls between A and B) / (total calls A + total calls B)) on the basis of 1666 antennas.

Map 2: Mutual information
Map of mutual information is based on probability to receive calls using the following formula: Inf Mutual Pij = * Ln (P ij / (Pi Pj *)) * Total. Pij is the probability of ux i to j ux, so ux i to j divided by the number of total ux. Pi. correspond for the probability of ux i to X therefore the sum of ux i to X divided by the number of total ux.

Map 3: Behaviour of antennas
The colours show the antennas, which are called as each other and have the same behaviour (cluster)
Introduction

The process of metropolisation promotes the development of a number of major cities to the detriment of national territories. This study seeks to show the roles of cities inside the urban system they form.

Balanced urban system should allow giving an answer to the current problems, for instance the development of uncontrolled urbanization, which generates a high consumption of natural resources, spatial and social segregation and growing disparities between regions.

A coherent system of cities should enable cities to work together jointly and increase their chance to play a role to face the urban problems in a few regions. The main challenge in Senegal is to counterbalance the hegemony of Dakar strengthening the urban framework in all the country.
Introduction

A sustainable land management should maintain at first the network of cities and should eventually lead to a limitation of urbanization in a second time.

The idea of the urban framework is based on the theories of Christaller and Lösch of the polycentric city. It can be analysed as a functional system or as a strategic network.
The literature often makes the difference between network and system (Dupuy, 1972) while for others like Offner and Pumain (1996) the concept is superimposed. We retain from Dupuy (1992) the three types of territories it offers for areas as open systems:

- Homogeneous regions: system limits are clear and the internal flows are low.

- Polarized regions: the centres of spatial organization are the main centres serving secondary centres, which themselves have relationships with a hinterland.

- Anisotropic regions: urban forms are arranged along one or more axes

Analysis of the call flow will allow us to understand the types of areas and the analyse the behaviour of networks. The way the system works should then provide us the basis for land use policies based on the real behaviour of the network and not political or administrative borders as usual.
National Territory
Results

The maps we draw the analyses were produce with various methods; they are described in the text below.

Analysis of the maps shows a number of important issues for understanding the country and the role of municipalities in the national territory.

The following results are based on calls, their intensities. They clearly shows how the cities network is working.

1. Dakar is the only city in Senegal! This provocative statement shows that we are dealing with an urban macrocephaly and the network of Senegal's cities are maintained only through the relationship with the Dakar region. This region with 3.137196 inhabitants according to ANSD (National Agency of Statistics and Demography) in 2013 is 23% of the country's population which is over 13 millions inhabitants. The number of calls and their intensities indicate that the role of Dakar is even greater when it comes to phone calls instead of population. There is a clear relationship between economic role, presence of infrastructures and big company and the number of phone calls. The relationship is not only based on the number of inhabitants but on the economical and social activity.
Results

2. The role played by the rural community of Touba (department of Mbaké) is also symptomatic. This makes Touba (although institutionally it is not a city), the pilgrimage town of Mouride community, the second city of the country. It plays a central role and is connected with its hinterland as well as with the biggest cities in the West of Senegal. In a secular Republic, the role of religion is very important as we can see with Touba.

3. While this may seem anecdotal, there is a strong connection between Touba and Tivaouane, the two cities of two most important Religious Communities of Senegal (Tijane and Mouride). The religious aspect plays an important role here and the overlap of the two community is important.

4. Kaolack is an important polarity in the center of the “peanut area”. It takes its role during the French colonization and still retains its importance for the area it covers.
5. St. Louis, the former capital of French West Africa (AOF), and Senegal and Mauritania even before 1958, is no longer the first city in the country, even the second one. Its network head role of the Senegal River makes its relatively small role despite its border position with Mauritania to the north.

6. The network of cities of the Senegal River is remarkable and it is an entity which takes place along the river and not just at the delta. We can assume an even larger network with the cities of Mauritania on the other side of the river that are not on the map.

7. The Casamance is not connected with the rest of the country and develops its own network between Lower and Upper Casamance.

8. Tambacounda is quite independent and acts as a regional centre with strong connections with the River, with High Casamance and Kedougou close to the border.

9. Finally, there is an important connection between Dakar and the « Petite Côte » with Mbour and Sally showing the connexion between the city and the touristic area of Senegal.
Map n° 1: Intensity
Map n° 2: Flow inter-antennas
Map n° 3: Co-clustering
Key lessons for the country:

1. The calls and their intensities clearly show how the urban system works in Senegal and also they show the economic role of the different regions, cities or municipalities.

2. Dakar unbalance the network, in terms of land use that would argue for a greater balance between Dakar and the "desert" of Senegal. From a regional perspective, or even African perspective, the dominance of Dakar is an asset to the economy and to the role the city plays in an international perspective. Competition of world cities requires strong polarity on a city, but it must be connected to its own national network.

3. The important role of natural barriers, like the Gambia, but also areas of fossil valleys.

4. Calls are close to relatives. Except for Dakar phone the whole of the country, call practices are clearly in close relatives, that is to say, from town to town.
Implementation

Our research results provide leads for future implementation of the project or implementation of new planning policies. Our results should allow to shows how the cities works and thus provides guidance for:

1. Determine a hierarchy of the national road system
2. Plan the development of power systems and water supply.
3. Develop regional balance
4. Anchor Dakar in a larger territory
Focus on Dakar
Map Dakar n°1: Intensity
Results II

1. Natural barriers play an important role as the great Niayes or the airport. The phone calls draw a realistic map of urban areas and green space or major equipment.

2. Calls are gradually, there is no relation centre - suburbs but some relatively autonomous territories. Rufisque is symptomatic of this, as well as the centre of Dakar.

3. Current areas under urbanization (East of the city) have relatively low intensity and calls on the map showing the little economic activity and few people staying in this area. The population of the suburbs are still strongly connected to the centre.

4. For the older suburbs such as Pikine and Guedewaye, they have a relatively high autonomy. The center-periphery picture is no longer so obvious.
Implementation

1. The activity in the suburbs is important, it must be supported by the infrastructure that enhance autonomy;

2. The greatest potential is in the East of the city in recently urbanized areas. This is where the balance with the city center should be, not necessarily farther to the southeast as is currently expected;
Ecole Polytechnique Fédérale de Lausanne – Switzerland
Dr Jérôme Chenal, Dr Mariano Bonriposi
www.epfl.ch

Groupe 8 - France
Guillaume Josse, Pedro de Oliveira
www.groupe8.com

PredicSis – France
Bertrand Grèzes-Bisset, Hadrien Chicault, Cedric Thao
www.predicsis.com
Sanitization of Call Detail Records via Differentially-private Summaries

Mohammad Alaggan
HCI Lab,
Computer Science Department,
Helwan University,
Helwan 11795, Egypt
malaggan@fci.helwan.edu.eg

Sébastien Gambs
Université de Rennes 1 - Inria
/ IRISA,
Avenue du Général Leclerc
35042 Rennes Cedex, France
sgambs@irisa.fr

Stan Matwin
Institute for Big Data Analytics,
Faculty of Computer Science,
Dalhousie University,
Halifax, Canada
stan@cs.dal.ca

Erico N de Souza
Institute for Big Data Analytics,
Faculty of Computer Science,
Dalhousie University,
Halifax, Canada
erico.souza@dal.ca

Mohammed Tuhin
IBM Canada and
Faculty of Computer Science,
Dalhousie University,
Halifax, Canada
mohammed.tuhin@dal.ca,
mtuhin@ca.ibm.com

ABSTRACT
In this work, we initiate the study of human mobility from sanitized call detail records (CDRs). Such data are extremely valuable to solve important societal issues such as the improvement of urban transportation or the understanding on the spread of diseases. One of the fundamental building block for such study is the computation of mobility patterns summarizing how individuals move during a given period from one area (e.g., cellular tower or administrative district) to another. However, such knowledge cannot be published directly because this type of data raises serious privacy issues due to its inference potential, such as the (re-)identification of individuals. To address these issues and to foster the development of such applications in a privacy-preserving manner, we propose in this paper a novel approach in which CDRs are summarized under the form of a differentially-private Bloom filter for the purpose of privately counting the number of mobile service users moving from one area (region) to another in a given time frame. Our sanitization method is both time and space efficient, and ensures differential privacy while solving the shortcomings of a solution recently proposed to tackle this problem. We also report on experiments conducted with the proposed solution using a real life CDRs dataset. The results obtained show that our method maintains a high utility while providing strong privacy guarantees. In addition, our method achieves - in most cases - a performance similar to another method (linear counting sketch) that does not provide any privacy protection.

Keywords
Differential privacy, sanitization, call detail records, mobility data, location privacy.

1. INTRODUCTION
One of the key ingredients of the digital economy of the future is the opportunity to exploit large amounts of data. In particular, Call Detail Records (CDRs) that are generated by users of mobile devices and collected by telecom operators could potentially be used for the socio-economic development and wellbeing of populations. For instance, such data can be used for scientific research (e.g., the study of the human mobility), but also for practical social goals in general (e.g., to find the best place to build an infrastructure such as a bridge or to create a new bus line). However, learning the location of an individual is one of the greatest threats against his privacy because it can be used to derive other personal information. For example, from the movements of an individual it is possible to infer his points of interests (such as his home and place of work) [25, 30, 18, 39], to predict his past, current and future locations [22, 36], to infer his social network [23] or even to conduct a de-anonymization attack [19]. Thus, it is of paramount importance to develop new methods that can be used to mine CDRs while preserving the privacy of the individuals contained in this data.

In this work, we address the issue by proposing a novel data sanitization method that learns a privacy-preserving data structure out of CDRs that can be used to identify global mobility patterns in the population while hiding the individual movements of users. For instance, consider the setting of a communication network composed of a large number of cellular antennas. In this scenario, each cellular antenna has a local knowledge of the activities conducted in its area (i.e., calls and SMS) as well as the identities of the users. If each cellular tower would publish regularly (e.g., every hour, every 12 hours or each day), the detailed CDRs of these activities, then it would be very easy to merge these partial
CDRs into a global data set and then to run a data mining algorithm on them. For example, it would be possible to count the number of different users observed in a cellular antenna over some period of time or the intersection between the set of users of different antennas for consecutive periods (e.g., to compute the so-called origin-destination matrices commonly used in urban planning). However, making such a data publicly available is not possible both for the privacy issues mentioned previously as well as for obvious commercial reasons (i.e., CDRs constitute a high-value asset of the telecom operator). In addition, storing the data in a centralized database only solves the problem partially as there is always a risk that a malicious attacker can exploit a security weakness to get access to this data or that an insider can cause a privacy breach.

**Contributions.** To counter these threats, in this paper we propose a method in which each cellular tower regularly publishes a differentially-private data structure summarizing in a compact manner information about users that have received or made a call (or SMS) in this tower. Once a differentially-private summary has been released, the cell tower can erase from its memory all information that it has recorded during the considered period. Afterwards, these summaries can then be combined using set operations tailored to this structure, such as counting the number of elements and size of set intersection, to identify generic mobility patterns. However, at the same time these summaries are designed to hide the actions and movements of a particular user. We also explore the achievable trade-off between utility (measured in terms of the accuracy of the derived mobility patterns) and privacy for the sanitized location data.

The outline of the paper is the following. First in Section 2, we introduce the concept of CDRs, as well as the notions of sketches and differential privacy that are central to our approach. Afterwards in Section 3, we present a motivating example of a privacy mining task before in Section 4 reviewing the related work including privacy-preserving methods developed to perform data mining on CDRs as well as differentially-private variants of sketches. Then in Section 5, we describe our novel method for performing data mining operations on CDRs, which is based on differentially-private summaries, focusing in particular on the private computation of intersection. Finally, we experiment this method on both real and synthetic data in Section 6 before concluding with some future work in Section 7.

## 2. PRELIMINARIES

In this section, we review the concept of Call Detail Records (CDRs), which correspond to the data on which we work, as well as the notion of summary, a data structure that can be used to represent in a compact and concise manner global properties of a CDRs dataset. Finally, we also briefly introduce the notion of differential privacy, which is the privacy model considered in this paper.

### 2.1 Call Details Records

Call Detail Record (CDR) data are generated by users of mobile services and gathered by telecom operators. Basically, CDRs record the time, the cell tower as well as the identity of a user involved in a voice call or a text message. CDRs datasets are collected primarily for pricing purposes. However in addition, these datasets have a high value, not only for scientific research (e.g., the study of the human mobility), but also for society in general (e.g., to find the best place to build an infrastructure such as a bridge or to create a new bus line) and for the economy (e.g., to identify the best neighborhood to open a new shop targeting a particular group of the population). In particular, CDRs are a major source of information for sociology, transportation, environmental monitoring and social networks [31]. For instance, performing data mining on CDRs help to understand how mobility patterns influence the spread of diseases (e.g., malaria) or how to optimize a bus system in a city of a developing country [27].

However, CDRs are sensitive with respect to privacy as they contain spatio-temporal information about the positions of users. Thus, CDRs provides a spatio-temporal fingerprint of the users within a defined region, which in turn makes easy identification of an individual possible by linking CDRs with other sources of information. For instance, if an adversary observes that a person moves from location \(A\) to location \(B\) each day of the week between 7AM and 9AM and in the opposite direction between 4PM and 6PM, then one can trivially deduces that this person lives in location \(A\) and that location \(B\) is his work place. In addition, the inference potential of CDRs is huge as such data can be used to infer other personal information about a user such as his religion, his political belief or his medical status, just to name a few. In particular, the Electronic Frontier Foundation has recently published a list of sensitive personal information (e.g., attendance of a particular church, an individual’s presence in a motel or at an abortion clinic) inferred about an individual knowing only his mobility traces [8].

### 2.2 Summaries and sketches

With the widespread use of Internet, tons of data are produced every second, which gives rise to the notion of Big Data. For such huge volume of data that are generated at a high velocity in applications like Internet traffic analysis and monitoring contents of massive databases, streaming is one of the techniques for low latency computing. In particular, if the dataset is too large to fit in memory, one might want to process the data sequentially with a limited amount of memory. Streaming algorithms achieve this with a single (or a small number of) pass(es) through the data. In addition to save on storage space, it might be necessary to transform the data into a more compact form, which gives rise to the notion of summaries or synopses. Basically, a summary represent statistical properties of the original data set with a storage cost much lower than the size of the original data at the expense of small inaccuracies in query results.

Sketch, a specific type of summary, is a data structure employed in the domain of stream processing. Sketches are usually simple to implement and enable to perform data mining over huge streams of data using limited space and time. More precisely, a sketch is a compact representation of data that uses only sublinear (in data and input size) space and is highly parallelisable in practice. A sketch is computed in an online and incremental manner, in contrast to sampling techniques that consider only a fraction of the whole data. The concept of hashing is employed in most
sketching algorithms in order to ensure a good accuracy with high probability. Thereafter, we briefly review some of the standard sketching approaches.

- **A Linear Counting Sketch (LCS)** [38] is a data structure enabling to probabilistically count the number of distinct item, even in the presence of duplicates. Basically, a LCS is an array of bits initially set to 0. An element is inserted in an LCS by setting to 1 the bit whose position is determined by the output of a hash function.

- **A Bloom filter** [7] is a widely used data structure originally designed for membership testing (i.e., testing whether a particular element is contained in the set considered). A Bloom filter can summarize a large dataset using linear space, is very simple to construct and inserting an element or testing its membership can be done in constant time. Briefly, a Bloom filter is composed of an array of $m$ bits (like a LCS) and equipped with a set of $k$ independent hash functions. An element can be inserted into a Bloom filter by passing it as input to each of the hash functions. The $k$ outputs of the hash function correspond to $k$ positions of the Bloom filter, which are all set to one independently of their previous values. Testing the membership is done in a similar manner by considering that an item is contained in the filter only if the $k$ corresponding bits are all set to 1. Due to the collisions generated by the hash functions, false positives can arise leading to considered that an element can be erroneously considered as being a member of the filter. The accuracy of a query to a Bloom filter depends on the number $k$ of hash functions used, the size $m$ of the filter as well as the number of items inserted. Bloom filters are often used for distributed applications requiring limited storage space or low communication cost [10]. Since the seminal paper of Bloom in 1970, many researchers have designed extensions and generalizations of Bloom filter that can answer other types of queries.

Bloom filters and sketching algorithms have wide variety of applications including compressed sensing [20], networking [17, 40, 11, 33], databases [9, 32], privacy [16] and machine learning [34].

### 2.3 Differential privacy

In this paper, we are interested in a strong privacy notion called **differential privacy** [19]. Differential privacy aims at providing strong privacy guarantees with respect to the input of some computation by randomizing the output of this computation, and this independently of the auxiliary information that the adversary might have gathered. In our setting, the input of the computation is a summary of the CDRs activity observed by a cellular tower and the randomized output will be a perturbed version of this summary (e.g., a Bloom filter or another type of sketch).

Two databases $\mathbf{x}$ and $\mathbf{x}'$ are said to **differ in at most one element** or said to be **neighbors** if they are equal except for possibly one entry.

---

**Definition 1 (Differential privacy [15]).** A randomized function $F : D^n \rightarrow D^m$ is $\epsilon$-differentially private, if for all neighboring databases $\mathbf{x}, \mathbf{x}' \in D^n$ and for all $t \in D^m$:

$$\Pr[F(\mathbf{x}) = t] \leq e^{\epsilon} \cdot \Pr[F(\mathbf{x}') = t].$$

This probability is taken over all the coin tosses of $F$ and $e$ is the base of the natural logarithm.

The privacy parameter $\epsilon$ is public and may take different values depending on the application (for instance it could be 0.1, 0.25, 1.5 or even 10). The smaller the value of $\epsilon$, the higher the privacy but also as a consequence the higher the impact on the utility of the resulting output. A relaxed concept of differential privacy called $(\epsilon, \delta)$-differential privacy [14] is a probabilistic variant in which the guarantees of differential privacy hold with probability of $1 - \delta$.

Originally, differential privacy was developed within the context of private data analysis and the main guarantee is that if a differentially private mechanism is applied on a dataset composed of the personal data of individuals, no output would become significantly more (or less) probable whether or not a single participant contributes to the data set. This means that an adversary observing the output of the mechanism only gains a negligible information about the presence (or absence) of a particular individual in the database. This statement is a statistical property about the behavior of the mechanism (i.e., function) and holds independently of the auxiliary knowledge that the adversary might have gathered. More specifically, even if the adversary knows the whole database but one individual row, a mechanism satisfying differential privacy still protects the privacy of this individual row. In our setting, the database that we want to protect is a CDRs dataset and the objective of a differentially private mechanism is to hide the presence or absence of a particular user in these CDRs.

Dwork, McSherry, Nissim and Smith have designed a generic technique, called the **Laplacian mechanism** [15], that achieves $\epsilon$-differential privacy for a function $f$ by adding random noise to the true answer of $f$ before releasing it. The amount of noise that needs to be added in proportional to the desired level of privacy $\epsilon$ and to the **sensitivity** of the function $f$, which measures the maximal possible change of the output when removing or adding a row in the database. Subsequently, McSherry and Talwar have proposed the **exponential mechanism** [28] which unlike the Laplacian mechanism that works only for functions with numerical output, provides differential privacy for functions whose output is more structured (e.g., graphs or trees). Both previous mechanisms (i.e., Laplacian and exponential mechanisms) work for queries that have some pre-defined form (e.g., such as the **counting** query that asks “how many rows satisfy the predicate $P$”), based on which they randomized the answer to that query to achieve differential privacy. These mechanisms are interactive as they require a two-way communication protocol between the curator (the entity in charge of the database) and the client performing the query. Thus, during this computation, the curator has to be online in order to receive the query and prepare the associate response to this query.

On the other hand, a **non-interactive** mechanism computes...
some function from the original database and releases it once and for all, which corresponds to a one-way communication protocol. The output released by the non-interactive mechanism can later be used by anyone to compute the answer to a particular class of queries (usually not just a single specific query), without requiring any further interactions with the curator. It is important to understand that the answer is not computed by the non-interactive mechanism, but rather that the answer is computed from the output released by the non-interactive mechanism. Thus after publishing this output the curator can go offline. One particular type of non-interactive mechanism is the generation of a synthetic dataset that allows the answer to certain class of queries (but not necessarily all) to be approximated. Examples of non-interactive mechanisms for differential privacy include \cite{5, 26}.

3. MINING MOBILITY DATA: A MOTIVATING EXAMPLE

Within the context of the D4D (Data for Development) project, CDRs can be leverage on to build a predictor of people's gatherings. The ability to predict that people will meet (in large numbers) in a given geographic location at a given period can be used to help the organization of such gatherings, including its security aspects, as well as provision of services with peak demand (e.g., ensuring the availability of enough bandwidth for phone calls). This task can be recast as a time-series prediction, in which a time series represents the number of users at a cellular antenna close to the event, and peaks and variations in the time series are indicative of the event itself. This task can be conducted by focusing on the dynamics of the changes of the number of users at a given antenna but of course other auxiliary information could be used. Techniques issued from the area of financial time series analysis can be used to compute these additional attributes, such as the Bollinger Band, the MFI and the EMA indicators \cite{37}.

This task can be recast as a classification task in which the objective is to detect human gatherings at prayer times. This classification problem is unbalanced: considering it as a two class problem, in which one class are the times of gatherings (prayers), while the other class are all other times. We observe that the number of instances of the former class is significantly smaller than the number of instances of the latter class. To deal with this issue, a modified version of Adaboost M1 \cite{Scha} was used as the classifier. This version considers a small change in the weight calculation to improve more detections on the prayer times. Using this variant of Adaboost enables to improve on the recall of the minority class (i.e., prayer time detection), with reducing only slightly the recall of the majority class (i.e., all other activities). At the same time, we observed that the precision of the majority class did not decrease while the precision of the minority class has improved. Satisfactory sensitivity increases the likelihood that the classification algorithm can detect the event before it happens. In our application this means predicting correctly a gathering of people in the area of a given antenna, giving a chance to the network managers to react by allocating the necessary bandwidth.

Note that mining task described only uses aggregate data representing the number of devices in the vicinity of a given tower. As the number of users at a particular at a given time in our data are quite large, there are no ethical issues related to potential privacy disclosure of individuals. However, privacy concerns would legitimately apply if one releases directly the raw CDRs. Instead in this paper, we propose to rely on a summary of CDRs based on the Bloom filter. Consequently, we are currently investigating the use of data summaries, as the ones described thereafter, for the task of gathering prediction. If this approach is successful, it will become possible to predict gatherings from mobile data using purpose-built, high-level data summaries that guarantees with a high level of confidence the impossibility to re-identify a particular user. Furthermore, we believe that publishing mobility data in a summary, de-identified form, that still enables useful computation, will enable easier sharing of mobility data for research, decision making and policy purposes.

4. RELATED WORK

The recent trend of publishing or sharing CDRs with third parties, has raised many privacy concerns. The most widely used approach to thwart these issues is to anonymize CDRs by techniques such as replacing user identifiers with random identifiers, making the call time approximate, . . . However, such anonymized CDRs cannot prevent privacy breaches in which by linking different databases, an adversary can target a small subset of persons \cite{12}. Thus, it is necessary to apply privacy-preserving methods along with data mining. For instance, differential privacy \cite{13}, which has now become the standard privacy model, has recently been applied on the result of data mining conducted on CDRs \cite{29}. Thereafter, we some recent related work on privacy-preserving data mining of CDR data.

Data mining on anonymized CDRs. In the past, anonymized CDR data have been used to estimate the presence of visitors \cite{21, 1}. In particular, the objective of these studies is to enable the linking of routes used by tourists with the places they visited such as the connection of sites of interest with the points of entry or final destinations within the country. Anonymized CDR data can also be used to study the extent and impacts of events such as fairs, concerts, sports events and also to manage crowds of visitors. In addition, the study of aggregated mobile phone data from Ivory Coast has lead to the discovery of features that have a strong correlation with poverty indicators \cite{35}. In particular, these features can be used to provide poverty estimates at a spatial resolution finer than previously possible.

Differentially-private summaries. Even if publishing a summary or sketch seems to be more privacy-preserving than directly releasing the original data set, there is still the possibility that the personal information contained in a sketch might be partially extracted or that one can cause a privacy breach by linking this data with other information. For instance, disclosing directly a plain Bloom filter is a very bad idea with respect to privacy as it is possible for an adversary observing this Bloom filter to query exhaustively for all possible items in order to reconstruct the set encoded in this Bloom filter. Thus, it is necessary to ensure that the summary released also ensure strong privacy.
guarantees (e.g., differential privacy). This can be done for instance by perturbing the output of a sketching algorithm before releasing it.

For instance to compute privately the similarity between profiles in a social platform, Alaggan, Gams and Kermarrec [2] have introduced a novel privacy-preserving sketching technique called BLIP (Bloom-then-Flip). To the best of our knowledge, BLIP is the only data summary solving the above problem. We briefly review it below. In the context considered, the profile of each user is represented compactly using a Bloom filter and the main objective of BLIP is to prevent an adversary with unlimited computation power from learning the presence or absence of an item in the profile of a user by observing the Bloom filter representation of this profile. BLIP ensures \( \epsilon \)-differential privacy [13] by flipping each bit of Bloom filter with some probability before publishing it. BLIP has the advantage of having the same communication cost as a Bloom filter, while guaranteeing privacy at the expense of a slight decrease of utility.

Kamp and co-authors [24] have designed a generic set of primitives for privacy-preserving mobility monitoring and modeling using stationary sensors (e.g., Bluetooth scanners). In particular, they considered two fundamental mobility monitoring tasks: (1) to track the number of distinct persons that are present at a location of interest (crowd monitoring) and (2) to reconstruct flows of persons between two or more different locations (mobility monitoring). Their method is based on the use on linear counting sketches. To solve these two problems, the authors compute the intersection between two sketches. With respect to privacy, one of the main shortcomings of their approach is to assume that the underlying hash function is kept private, which is a form of privacy through obscurity in direct opposition with Kerckhoffs’s principle. Our work is directly inspired from this work, for which we rely on Bloom filter by extending a recently proposed differentially-private Bloom filter by enabling the computation of the size of the intersection between two such Bloom filters. Note that in contrast to [24], we do not assume the privacy of the hash functions used in the construction of Bloom filters.

5. MINING CDRS VIA DIFFERENTIALLY-PRIVATE SUMMARIES

System model. Our main objective is to perform data mining operations on differentially-private summaries that are learnt directly from CDRs. For instance, we would like to be able to count the number of distinct users represented in a particular summary or to compute the size of the intersection between two summaries. To realize this, we will leverage on the BLIP mechanism described in the previous section. More precisely, we consider as our model a cellular network composed of a large number of cellular antennas that are responsible for recording the CDRs related to their neighborhood (basically the events generated by calls or text messages sent or received in their corresponding cellular towers). Each of this cellular tower publishes at a regular interval (e.g., every 6 hours or each day) a summary of the user it has seen during this period. More precisely, this summary takes the form of a Bloom filter in which each element that is inserted is actually the identifier of a user associated to a CDR. One of the advantage of using a Bloom filter is that even if a user is inserted several times (e.g., if he is responsible for generating several CDRs during the same period), his impact on the Bloom filter is the same. In addition, Bloom filter can be updated incrementally and in an online manner. Once the end of the period is reached, all cellular towers BLIPed their summaries before publishing them and then erase their memories. Indeed, there is no need for them to remember any information as the main objective of this work is to be able to perform operations on CDRs directly from the differentially-private summaries computed. One of the benefit of erasing their memories is that there is no risk later than private information leak from cellular towers due to a security breach.

Computing the size of the intersection between two summaries. Broder and Mitzenmacher [10] described a method to approximate the intersection of two sets, \( S_1 \) and \( S_2 \), given their Bloom filter representation, \( B_1 \) and \( B_2 \). Basically, they provided a relationship between the inner product of two Bloom filters and the cardinality of the set intersection of the two sets encoded in those Bloom filters. This expected value for the inner product \( \sum B_1[i]B_2[i] \) corresponds simply to the probability that a particular bucket is simultaneously set to one in both Bloom filters, multiplied by the total number of buckets: \( m(1 \times \Pr[B_1[i] = 1 \wedge B_2[i] = 1]) \). We extended their method to the case in which the bits of both Bloom filters are flipped with some probability (such as in BLIP) prior to computing the inner product, which enables to estimate the cardinality of set intersection in a privacy-preserving way.

Our main result is summarized by the following theorem.

**Theorem 1.** Given two sets \( S_1, S_2 \) with cardinalities \( n_1, n_2 \), respectively, and their flipped Bloom filters \( B_1, B_2 \) whose bits were flipped independently with probability \( p \), let \( Q \) be the inner product between \( B_1 \) and \( B_2 \). Let also

\[
g(x) = -\ln(x/m) - C_1/C_2 + C_3, \tag{1}
\]

in which \( q = 1 - p \), \( k \) is the number of hash functions employed by the Bloom filter, \( m \) is its number of buckets, \( \phi = 1 - 1/m \), \( C_1 = (pq - q^2)(\phi^{k+1} + \phi^{kn_2}) - q^2 \), \( C_2 = k \ln \phi \) and \( C_3 = \ln((1 - q^2)/k)\ln \phi + n_1 + n_2 \). Then \( W = g(Q) \) is an estimator for \( |S_1 \cap S_2| \) with expected value:

\[
E[W] \approx |S_1 \cap S_2| + \operatorname{var}(Q)/2C_2(\E[Q] - C_1m)^2. \tag{2}
\]

**Proof.** Consider \( E[W] = E[g(Q)] \). The expectation of the second degree Taylor expansion (i.e., truncated after the third term\(^1\)) around \( E[Q] \) of \( g(Q) \) is

\[
g(E[Q]) + \operatorname{var}(Q)/2C_2(E[Q] - C_1m)^2 \tag{3}
\]

There is a bias term \( \operatorname{var}(Q)/2C_2(E[Q] - C_1m)^2 \) that we will further detailed later. We now compute \( E[Q] \) and prove that \( g(E[Q]) = |S_1 \cap S_2| \). Consider two standard unflipped Bloom filter representations of two sets \( S_1 \) and \( S_2 \). The \( j \)-th bit of both Bloom

\(^1\)We talk about the error introduced by this truncation in the next section.
filters will be set to 1 simultaneously if it was set by an element in \( S_1 \cap S_2 \) or by some element in \( S_1 - S_2 \) and a different element in \( S_2 - S_1 \). Consider the probability space of the choice of the hash functions used and let \( A_j \) be the event that the \( j \)-th bit is set in the first Bloom filter by an element of \( S_1 - S_2 \), \( B_j \) be the event that the \( j \)-th bit is set in the second Bloom filter by an element of \( S_2 - S_1 \), and \( C_j \) as the event that the \( j \)-th bit in both Bloom filters was set by an element in \( S_1 \cap S_2 \). The event that the \( j \)-th bit is set in both Bloom filters is \( (A_j \cap B_j) \cup C_j \) and its probability is \([10]:\)

\[
\Pr[C_j] + (1 - \Pr[C_j]) \Pr[A_j] \Pr[B_j] = 1 - \phi^{kn_1} - \phi^{kn_2} + \phi^{kn_1+n_2-|S_1\cap S_2|}.
\]

For the case of flipped Bloom filters let \( A = 1 - \Pr[A_j], B = 1 - \Pr[B_j], \) and \( C = 1 - \Pr[C_j] \), such that

\[
\Pr[C_j] + (1 - \Pr[C_j]) \Pr[A_j] \Pr[B_j] = (1 - C) + C(1 - A)(1 - B).
\]

For a bit to be set to 1 in the flipped Bloom filter it either had to be 1 in the unflipped Bloom filter and remained unchanged despite the flipping or the converse. Therefore, it is easy to verify that the probability that the \( j \)-th bit is set to 1 is both flipped Bloom filters is

\[
q^2(1 - C) + C[q^2(1 - A)(1 - B) + pqA(1 - B) + pqB(1 - A) + p^2AB],
\]

which is equal to

\[
(pq - q^2)(\phi^{kn_1} + \phi^{kn_2}) + q^2 + (p - q)^2\phi^{kn_1+n_2-|S_1\cap S_2|}.
\]

Now, considering that the inner product \( Q \) between two flipped Bloom filters is the sum of \( m \) binary random variables, each of which has the probability of being 1 given by Equation (4). One can easily verify by simple algebraic manipulation that \( g(\mathbb{E}[Q]) = |S_1 \cap S_2| \). We also have that \( \mathbb{E}[Q] \) itself is the same as (4) multiplied by \( m \) by independence. Hence, the result follows.

In the proof of Theorem 1, we truncated the Taylor expansion after the third term, which naturally introduces an error. However, we demonstrate that for standard values of the parameters, the error is negligible. The values of the parameters considered is \( \epsilon = 3, m = 187500, k = 2, \) and \( n_1 = n_2 = 50000 \). Then, with probability at least 99.9%, when \( Q \) is in the range \( \mathbb{E}[Q] \pm 3\sqrt{m} \) [2], the absolute error introduced by truncating beyond the third term is at most 0.003 for all values of \( |S_1 \cap S_2| \). The error is clearly negligible compared to \( |S_1 \cap S_2| \). Moreover, for the same settings, the ratio of the fourth term to the third term is 0.002, which justifies its truncation.

The bias of the estimator as in Equation (2) can be bounded using the Bhatia-Davis inequality [6] \( \text{var}(Q) \leq (m - \mathbb{E}[Q])\mathbb{E}[Q] \).

Using this bound, the absolute ratio of the bias term

\[
\text{var}(Q)/2C_2(\mathbb{E}[Q] - C_1 m)^2
\]

to the set intersection \( |S_1 \cap S_2| \) is at most 30% for \( |S_1 \cap S_2| \leq 12000 \), for the same set of values used in the previous paragraph. We will see in the next section that this value represents a mean relative error (MRE) of 0.3 on the \( x \)-axis, which the figures show to be very small.

### Counting the number of users in a summary.

Note that the computation of the intersection requires knowledge of \( n_1 \) and \( n_2 \), the true cardinalities of the two sets under consideration. Bahl, Furon, and Gambos [4] have recently proposed a way to estimate the set cardinality directly from BLIP. Another possibility is to release directly the set sizes by applying the \( \epsilon \)-differentially private Laplacian mechanism [15] with error \( O(1) \), which is much smaller than the error \( O(\sqrt{m}) \) introduced by BLIP itself. In that case, the total privacy budget will be \( 2\epsilon \) instead of \( \epsilon \). However, different values for \( \epsilon \) can be chosen independently for BLIP and for the release of \( n_1 \) and \( n_2 \), to reduce the privacy budget.

### Intersection of three sets.

Given three flipped Bloom filters of three sets \( A, B, \) and \( C \), all of which are flipped with the same probability, we can predict the number of elements in the intersection of the three sets. For this, we chose to use the regression through random forests for prediction [9]. The features use to train this classifier are the following: (1) the number of elements in the three sets, (2) the number of elements in the intersection of every pair of sets and (3) the weight of the average of the three flipped Bloom filters, namely \( w = \sum_{j=1}^{m} \frac{1}{3} (B_j[j] + B_j[j] + B_j[j]) \), in which \( B_j \) is the flipped Bloom filter for the \( j \)-th set. To summarize, each data point in the training set is described by the following seven features (\( |A|, |B|, |C|, |A \cap B|, |B \cap C|, |C \cap A|, w \)). When any of these cardinalities are not known, they can be easily estimated. For instance, the cardinality of \( A, B, \) or \( C \) can be estimated from their flipped Bloom filter directly using the method of §3.3 in [4], or they can be released in a differentially private manner using the Laplacian Mechanism. In addition, the cardinality of their pairwise intersection can be deduced using the estimator proposed in the previous section. Intuitively, the \( w \) feature contains information on whether the three Bloom filters had \( \text{jointly} \) a 1 in the corresponding bucket before being flipped, which is the case when they share an item that maps to this bucket. The pairwise intersection cardinalities are needed to give information on whether the buckets were set to 1 after the flipping due to an item shared by at most two of the sets. The experiments in Section 6 demonstrate that this method accurately estimates the number of elements in the intersection of three sets.

### Towards more complex data mining tasks.

The computation of the size of the intersection between two different CDRs is a fundamental building block to the design of more complex data analysis. For instance, in the context of mobility mining one may want to compute more sophisticated data structures out of the CDRs such as trajectories and mobility models. We leave as future work the possibility of performing other operations on differentially-private summaries as well as the design of other building blocks for conducting privacy-preserving data mining algorithms on CDRs. In addition, to complement the mobility analysis with a social one requires the development of other data structures that could
capture this aspect of data while preserving the privacy of users as well as their social ties.

6. EXPERIMENTAL RESULTS

Experimental setting. In this section, we report on the results obtained by applying our method on a real dataset from a telecom operator. This large dataset includes coarse-grained phone call data for users. More precisely for each call made by the user, his location (with granularity up to district) as well as the time of the call (with granularity up to 10 minutes) are recorded. In our experiments, we have first extracted the set of users for each district and for each month and encode them in a Bloom filter. Afterwards, the Bloom filter was flipped according to the recipe of BLIP before it is released. Afterwards, we use our algorithm to estimate the intersection between each pair of district given their flipped Bloom filters. To observe the behavior of our algorithm under different conditions, we specifically selected four sets of district couples. The choice of these four sets was based on the different number of users in the districts. For instance, the first set contains two districts with 57000 and 42000 users, and we call it “two large” in the figures while the next set contains two districts with 600 and 1000 users, and we call it “two small”. For the two other sets, there is a small district and a large one, but one set has a small intersection between the two districts while the other one has a (relatively) big intersection. In particular, the first set has 15000 and 1200 users and an intersection of 180 users, thus we call it “small intersection”. The second set has 3400 and 39000 users with an intersection of 3339 users and we call it “big intersection”. All the experiments reported involving BLIP are averaged over 100 independent trials.

Evaluation. To evaluate the utility of our method, we compute the Mean Relative Error (MRE) of the estimated intersection. The MRE is defined as the absolute difference between the estimated intersection and the true intersection, divided by the true intersection. We also plot the standard deviation of the MRE for selected parameters. In addition, we give as a baseline the result using linear counting sketches, with no privacy guarantees, as a way to estimate intersection between two sets in Figure 1.

In Figure 2, we show the MRE for different values of Bloom filter parameters: $k$, the number of hash functions, as well as $m$ the number of buckets, for a privacy level $\epsilon = 3$. This value of epsilon is rather strict and provides high privacy guarantees. We can observe how the MRE changes based on the properties of the districts. In particular, if the two districts are big, or when the intersection is large, the MRE is less than 12%, regardless of the choice of parameters. However, when the two districts are small or when the intersection is small, the choice of parameters becomes critical. Generally, a lower value for $k$ and a higher value for $m$ is better, except when the two districts are small. We also see in Figure 3 that when $\epsilon = 11$ (which corresponds to low privacy guarantees) that utility gets even better.

Afterwards, we analyze the situation when $k = 2$ as we have seen this is the best value for $k$ in Figure 4. The good utility obtained for low $k$ may not be a property of the Bloom
filter, but due to BLIP, the flipping probability \( p = \frac{1}{1 + \exp(\epsilon/k)} \) depends significantly on \( k \). We plot the MRE against \( \epsilon \) for each set of districts and investigate how the MRE reacts with the value taken by \( m \). We observe that beyond a certain value for \( \epsilon \), a higher value of \( m \) gives a better utility. However, for the case of two small districts, the variance of the results is high as indicated by the 95% confidence interval surrounding the MRE value.

Observe that for big districts or big intersections, the utility obtained is very close to that of the baseline of linear counting sketch. However, the lower utility attained for small districts and small intersections seems necessary to protect the privacy of individuals. Recall that differential privacy ensures that the amount of information contributed by each user is limited, hence, a smaller number of users cannot contribute with much information if their privacy is to be protected with differential privacy guarantees.

**Neighboring Districts.** Finally, we have also investigated the efficiency of our approach on four neighboring districts. Indeed, neighboring districts are expected to have higher intersection values, and thus higher utility for our method. The four neighboring districts are labeled 1, 2, 3, and 4 and are respectively of sizes 48697, 38770, 55393, and 57678 users. The edges (in graph notation) representing their neighborhood relation are

\[ \{(1, 2), (1, 3), (2, 3), (3, 4)\} , \]

with respective intersections of sizes 28863, 37851, 30478 and 42528. As shown in the following figures, the results
obtained for these districts are very accurate. In particular, for $\epsilon = 3$, the MRE ranges from 20% to 40% (Figure 7), with standard deviation below 0.04 (Figure 8).

**Intersection of three sets.** As explained in the previous section, to predict the intersection of three sets a random forest regression model was trained, with each point of the training set being described by seven attributes: the set cardinalities, the cardinalities of their pairwise intersection and the weight of the average of the three flipped Bloom filters. We carried an experiment using $k = 2, m = 187500$ and $\epsilon = 3$. We sampled 50 districts for the training set and 50 districts for the testing set. For each of them we enumerated all possible subsets of three districts and choose 50 district triples (25 for training and 25 for testing) with an intersection cardinality higher than 4. The median intersection cardinality in both the training and testing sets is 10. The prediction quality is measured in terms of the Mean Relative Error (MRE). For $\epsilon = 3$, the MRE is 0.53 (with median 0.26), which means that relative error for half the testing set was at most 26%, and never more than 53% for the rest of the testing set.

**Summary.** From these results, we conclude that when $k = 2$ and $m = 187500$, we can privately estimate the intersection of two and three districts with a high degree of accuracy as shown in Figure 5. The standard deviation can also be seen in Figure 6, which demonstrate that the estimation is not only accurate but also displays a high degree of confidence. We believe that the relatively high error for small districts

---

**Figure 6: Standard deviation of the mean relative error for different levels of the privacy parameter $\epsilon$ when the number of hash functions $k = 2$ and the number of buckets is $m = 187500$ bits.**

**Figure 7: Neighboring districts: MRE for different levels of the privacy parameter $\epsilon$ when the number of hash functions $k = 2$ and the number of buckets is $m = 187500$ bits.**

**Figure 8: Neighboring districts: standard deviation of the MRE for different levels of the privacy parameter $\epsilon$ when the number of hash functions $k = 2$ and the number of buckets is $m = 187500$ bits.**
and small intersection is necessary to protect the individual users and is unavoidable. In addition, our method is very efficient in terms of computation and communication. Indeed, this method requires only to compute two hash functions per user in the district, which is extremely efficient. For the communication cost, only 183 kilobytes are sufficient for summarizing the data of an entire district. The communication cost of our method is exactly the same as needed for a plain Bloom filter, which is a constant number O(1), of bits for each user per the Bloom filter. More precisely, this communication cost is around 187500/50000 = 3.75 bits per user for a district of 50000 users.

7. CONCLUSION
In this work, we have initiated the study of human mobility from anonymized call detail records (CDRs). More precisely, we have proposed a method by which CDRs are sanitized in the form of a differentially-private Bloom filter for the purpose of privately counting of the number of mobile service users moving from one area (region) to another in a given time frame. Our sanitization method is both time and space efficient and solves the shortcomings of a recent solution proposed by Kemp and co-authors to this problem. The results obtained show that our method achieves - in most cases - a performance similar to another method (linear counting sketch) that does not provide any privacy guarantees. Thus, we conclude that our method maintains a high utility while providing strong privacy guarantees. We leave as future work the design of other privacy-preserving building blocks for other data analysis tasks.

Acknowledgments
We would like to acknowledge the Call Detail Records datasets provided by the ‘D4D Challenge’ in Senegal.

8. REFERENCES


A Cell Dwell Time Model for Synthetic Population Generation from Call Detail Records

Thierry Derrmann, Raphaël Frank, Foued Melakessou, German Castignani and Thomas Engel
Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg, 4 rue Alphonse Weicker, L-2721 Luxembourg
firstname.lastname@uni.lu

Abstract—In this work we propose a novel Cell Dwell Time Model that can be used to generate a synthetic population. We introduce two new metrics to define the attractiveness of cell sites based on global and individual parameters obtained via the analysis of the Data For Development (D4D) Call Detail Records (CDR). We rely on the shortest path road network to interconnect two distant cell sites. The resulting dwell time model can be used to compute accurate user trajectories even with partial information. This work represents a first step towards the generation of a synthetic population that can be used to perform a wide range of simulative studies to evaluate and optimize transportation networks.

Keywords—D4D Challenge, Mobility Analysis, Cell Dwell Time modelling, Synthetic Population

I. INTRODUCTION

The study of human mobility is of great importance for a number of applications in different fields. It can be used to facilitate urban planning [1], improve the transportation networks [2] [3] and also provide important insights on epidemic spreading of diseases [4]. Until recently, the data available to conduct those studies was limited and laborious to retrieve and process. Today, with the ubiquity of mobile phone networks new data can be made available to the research community around the world to better understand human mobility. Those datasets are usually composed of Call Detail Records (CDR) that list all transactions of mobile users on the network. Most importantly, by knowing the geographical location of cell towers, it allows the reconstruction of the spatiotemporal activities of users on a countrywide scale, providing important insights on human mobility. This is especially relevant in developing countries, as often alternative data sources are very limited. Further, the mobile phone penetration in these countries is usually very high. As an example, in Senegal the mobile phone penetration rate recently exceeded 100% [5].

However, the CDR datasets only provide information about the location of a user when a transaction is being made, i.e. during call or messaging activity. Also, the coverage area of a cell can vary between several hundreds meters up to a couple of kilometers. As a consequence, user trajectories are often partial and thus have to be approximated. In order to overcome those limitations, CDR data is often combined with other available data sources (e.g. census information) to better model human behaviour [6]. Those models, called Synthetic Populations are very useful for policy makers to test how mobility flows are affected by environmental changes and other events (e.g. modification of the transportation networks).

In the context of the Data For Development (D4D) Senegal Challenge we propose a novel Cell Dwell Time Model based on CDR data that can be used to generate synthetic populations. By analyzing the available datasets, we introduce a global and individual attractiveness metric to model dwell times for the different cell sites. Further, we rely on the shortest road path to interconnect two distant cell sites. This allows us to better estimate the path of a user between call or messaging activities even with partial information. Our results show that our model follows a realistic behaviour, which can be applied to generate synthetic populations.

The reminder of this paper is structured as follows. In Section II, we review relevant related work on cell dwell time modelling and synthetic population generation. In Section III we describe the dataset and motivate our assumptions. Our model is introduced in Section IV. In Section V, we present our preliminary results and indicate directions for future work. A conclusion is provided in Section VI.

II. RELATED WORK

Over the past decades different probability distributions have been used to model cell dwell times [7]. Fang [8] found that Phase-Type (PH) distributions provide a very good description of the dwell times. Among PH distributions, Coxian and Hyper-Erlang show to provide a good fit because of their universality property. Similarly, in [9], the authors found that there is a relationship between channel holding time and cell dwell time by taking the assumption that cell dwell times are Coxian or Hyper-Exponentially distributed. They also presented Extreme Value Distributions (EV) for modelling cell dwell times, such as the Generalized EV Distribution and the Weibull Distribution. These are heavy-tailed probability distributions that can express both transit and long cell dwell times of network users.

The work presented by Apolloni et al. [6] at the D4D Ivory Coast Challenge provided some interesting insights on the limitations regarding the generation of synthetic populations using only CDR data. They make the assumption that the inter-site time is proportional to the lengths of the intersections of the straight line between the two cell sites. Further, they used simplistic assumptions regarding the composition of households and mobility patterns. Due to the lack of alternative data sources their results could not be validated.

The work of Kung et al. [10] studies the commuting behaviour between home and work locations using, amongst others, CDR datasets. In order to identify the home and work
location they first constructed and Individuals’ Travel Portfolio, which consists of a list of frequently visited cells. This and other parameters and then used to identify the home and work cells. For simplicity they used the great circle distance to estimate the commuting distance between the home and work location.

An interesting application where CDR based synthetic population could be useful is traffic optimization via iterative simulations. Zilske at al. [3] show that mobile phone transactions without any layer of interpretation provide plausible traffic patterns. It has however been pointed out that further verifications would be needed to validate their assumption.

Our approach defines a first step towards a realistic synthetic population that can be used as an input to perform traffic simulations and optimizations. Compared to existing works, we add new attractively metrics used as an input of our cell dwell time model. Further, we rely on publicly available roadmap data to better estimate distance and travel time between distant cell sites.

III. DATA PREPROCESSING

This section presents the different computational steps that were performed before the model could be estimated. First, we will present the datasets used, followed by a presentation of the methods that have been taken into account in order to obtain the required metadata.

A. The 2014 D4D Challenge Datasets

The work proposed in this paper is based on the 2014 D4D Challenge Datasets provided by Orange Senegal. In particular, we consider the two datasets detailed below:

1) SET1 (Inter-site communications): It represents call activities per hourly time slot and source/destination antenna.

2) SET2 (User trajectories): It represents successive user activities with truncated timestamps (rounded to 10 minutes) and antenna.

B. Inter-Antenna Road Paths

Since we want to estimate the intermediate movements of mobile phone users, we imported the Senegalese road network from OpenStreetMap and added the provided cell sites into a spatial database. We precomputed the closest nodes and inter-antenna routes, distances and temporal costs using the Dijkstra algorithm within the spatial database framework.

The inter-antenna routes are represented as geometries, i.e. the consecutive road network point sequence. Using these route geometries, we computed their intersection with the Voronoi polygon geometries of the cell sites. Thus, we obtained the sequences of expected sites visited along all the inter-antenna routes (i.e. about 2.5M routes).

C. SET1 and SET2 Data Extraction

In order to create realistic movement patterns, it is necessary estimate the intermediate movements of users between their call and messaging activities. We propose to combine our knowledge about the general population from the aggregate dataset (SET1) with the knowledge about individual users (SET2). For SET1, the necessary precomputations involve computing the total call count and duration per each site and hour of day (cf. Algorithm 1). For SET2, successive user locations and timestamps were put inside a zip list (cf. Algorithm 2), in order to prepare the subsequent intermediate location and dwell time imputation.

D. Assumptions

We assume that the amount of call activity in a site reflects the amount of people and activity present in the site. That means that the probability of a user being located in a site \(s\) at time \(T\) \(P_T(s)\) is proportional to the aggregate amount of call activity in that site at time \(T\). The aggregate amount of activity is denoted as global (objective) site attractivity \(A_G\) and is computed using the metadata generated using Algorithm 1.

\[
P_T(s) \propto A_G(s)
\]  

\(1\)

Algorithm 1 SET1 Global Attractivity Computation

1: procedure PREPROCESS SET1
2: for all Month do
3: group by sitesource, hourofday
4: aggregate callcount, callduration
5: end for
6: end procedure

Algorithm 2 SET2 User Trajectory Extraction

1: procedure PREPROCESS SET 2
2: for all 2WeekSlice do
3: group by userid
4: aggregate ziplist(siteid, timestamp)
5: end for
6: end procedure

---

\(1\)The computations were performed using PostGIS and pgRouting
We also assume that users travel on the shortest road network path between two consecutive call locations. During our research on the Senegalese transportation infrastructure, we found that the prevalent means of transportation is via the road network using private cars and taxis.\(^2\)

Note that these assumptions influence the parameter estimation. Some transportation models allow for dynamic route choice. This allows the movement of some individuals without using the shortest path (e.g. onto a less occupied path to the same destination).

Furthermore, we introduce a measure of individual (subjective) attractivity, i.e. how often a user has performed a call or messaging activity from a site. This serves to increase the estimated dwell time in frequent locations, such as home and work. We distinguish between daily and nocturnal favorite sites (the details of this will be discussed in the following section), and we use the metadata output from Algorithm 2.

\[
P_I(s) \propto A_I(s) \tag{2}
\]

IV. CELL DWELL TIME MODEL

In this section, we will introduce the definitions of the attractivity measures and link them to a dwell time model.

A. Attractivity

The attractivity measures serve to estimate the amount of time a user has spent within a certain site. We propose two different measures, i.e. the global attractivity (spanning all users of the network) and the individual attractivity (concerning a single user). The purpose of these metrics is to reflect both the whole population’s behaviour as well as the individual user’s behaviour.

According to the assumptions from the previous section, we propose the following expressions for the attractivity measures:

1) Global Cell Site Attractivity: The global attractivity of a site in a given hour of day (see Equation IV-A1) represents the global system activity that exists in this site, i.e. the sum over all destination antennas (d) from a given site (s) and in a given hourly timeslot (t). We base this computation on the data obtained from the pre-computation Algorithm 1.

\[
A_G(s, t) = \sum_{d \in \text{Antennas}} n_{\text{calls}}(s, d) \tag{3}
\]

The histogram in Figure 2 shows the distribution of the global attractivities. Most sites correspond to lower attractivity values, and there are gradually less sites of higher attractivities. This corresponds to the fact that the activity on the network concentrates in some sites of high population density, while other sites (e.g. transition sites) tend to have lower call activity.


2) Individual Cell Site Attractivity: We distinguish night and day attractivity according into two different time categories. We consider as a user’s (u) daily attractivity of a site (s) the amount of activities \((n_{\text{activities}})\) by this user in the given site between 7:00 and 21:00. The nightly attractivity corresponds to the amount of activities between 10:00 and 6:00. This computation is based on pre-computation Algorithm 2.

\[
A_I(u, s)_{\text{Day}} = \sum_{t \in [7, 21]} n_{\text{activities}}(u, s) \tag{4}
\]

\[
A_I(u, s)_{\text{Night}} = \sum_{t \in [22, 23, 0, \ldots 6]} n_{\text{activities}}(u, s) \tag{5}
\]

According to the starting time of the two successive activities (between which we want to impute locations), we choose the according attractivity measure. Note that we normalize all the attractivity measures, in order to be able to use them as weights for estimating the times spent in intermediate sites.

In Figure 3, we show the distribution of the individual attractivities. As with the global attractivity, most sites correspond to lower attractivity values, and only few sites have increased attractivity values. These sites, however, correspond to places of high importance to the individual (where calls are initiated or terminated frequently).

In the following section, we explore different functions combining both attractivity types above into weights that split up the dwell times associated with an inter-call time.

B. Cell Dwell Time Probability Density Function

There are various probability density functions that have been used to model dwell times in mobile communication, in particular heavy-tailed distributions (extreme value distributions) [9]. A constant rate of departure, represented by an
exponential dwell time distribution does not fit the observations for different attractivity weight functions. This is the case because there is a tendency to stay longer in a cell if the user has already stayed for longer before. The rate of departure is higher at the beginning (short transit stay) than after some dwell time (longer activity in site). The rate of departures we observed are decreasing with increasing dwell time. For that reason, we model the dwell times as Weibull random variables for each site \( s \) and hour of day \( t \) (Equation IV-B).

\[
T_{\text{Dwell}}(s, t) \sim \text{Weibull}(\lambda^t_s, k^t_s)
\]

The parameters \( \lambda \) and \( k \) are estimated for each antenna and hour of day using Maximum Likelihood Estimation as shown in Equation 7 [11].

\[
\begin{align*}
\hat{\lambda}^t_s &= \frac{1}{N} \sum_{i=1}^{N} (x_i^t - x_N^t) \\
\hat{k}^{-1} &= \frac{\sum_{i=1}^{N} (x_i^t \ln x_i - x_N^t \ln x_N)}{\sum_{i=1}^{N} (x_i^t - x_N^t)} - \frac{1}{N} \sum_{i=1}^{N} \ln x_i
\end{align*}
\]

\[ F(A_G, A_I) := (A_G)^2 A_I \]

\[ \sum F(A_G, A_I) = 1 \]

\[ \sum F(A_G, A_I) = 1 \]

We chose this function mainly based on the resulting P.Values when fitting the Weibull distribution and using a Kolmogorov-Smirnov-Test on a sample of the fitted distribution against the observed data (cf. Table I). This produces a bowl-shaped weight distribution as depicted in Figure 4. The leftmost (majority) of weights corresponds to transit cells of low relative attractivity, whereas the right tail represents the most attractive sites, that are more numerous than sites of average attractivity.

As we can see from the quantiles in Table I, we obtain about 80% short dwell times (less than 1 hour or 3600 seconds duration), and about 10% dwell times greater than around 7 hours. This matches the fact that in SET2_P02.CSV, the mean amount of sites visited between 2 activities amounts to 3.85. If we assume that on average, one of these sites is the site of a longer stay, about 75% of sites visited will only be transitional sites, and thus of short dwell time.

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Formula} & Q(0.8) & Q(0.9) & Q(0.95) & P - \text{Value} \\
\hline
A_G & 13580 & 24578 & 28886 & 0.299 \\
A_G A_I & 8466 & 29604 & 36806 & 0.292 \\
A_G^2 A_I & 3499 & 25882 & 33707 & 0.5837 \\
\hline
\end{array}
\]

TABLE I: Quantiles and mean P-Values for antenna 559 in SET2_P02.CSV
**Algorithm 3** SET2 Trajectory Completion

1: procedure **SET2 IMPUTATION**
2: for all site1, site2 in successive Set 2 Activ. do
3:   ΔT ← TimeDifference(site1, site2)
4:   sites_intermediate ← Intersected(site1, site2)
5:   for all site in sites_intermediate do
6:     AG ← GetAttrGlobal(site)
7:     AI ← GetAttrIndiv(site)
8:   end for
9:   AG ← AG / ∑AG
10:  AI ← AI / ∑AI
11:  Weights ← (AG)^2 AI
12:  Weights ← Weights / ∑Weights
13:  for all site in sites_intermediate do
14:    DwellTime_site ← Weights_site ΔT
15:  end for
16: ...
17: end for
18: end procedure

**D. Cell Dwell Time Computation**

In Algorithm 3, we make use of the pre-processed data and perform all the computations introduced above. For each pair of successive user activity positions in SET2, the time difference between both activities is computed.

The corresponding travel route is queried from the spatial database (i.e., the intersection of the shortest road path with the Voronoi tessellation of the network sites), yielding a sequence of sites (sites_intermediate). Then, both attractivity measures are evaluated for all of these sites and the time difference ΔT is split according to the attractivity weights. Thus, we obtain the dwell times for each individual site and can compute the arrival times in sequence by summing up the dwell times.

Note that we did not exclude any kind of users or trips, which allows us to base our model on the entire user base of the network.

**V. RESULTS AND FUTURE WORK**

In this section, we present the resulting fitted dwell time distributions obtained from the imputed SET2 using the proposed imputation Algorithm 3.

**A. Resulting Dwell Times**

As a first result, we can see in Figure 5 that most interactivity-times correspond to short trips. This can either be due to high call activity or to a low mobility among the observed population.

Figure 6 shows the empirical CDFs of the estimated dwell times and the corresponding fitted Weibull variates given arrival times between 2:00 and 3:00 (at night) and 12:00 and 13:00 (daytime) in site 559. We can see that at night, about half of the estimated dwell times are longer than 5000 seconds, while at daytime, this proportions only makes up around 20% of the users visiting this site.

Figure 7 represents quantiles 0.5, 0.8, 0.9 and 0.95 of dwell times in SET2_P02.CSV for all antennas over an entire day. We can see that dwell times are much longer at night, while between 10:00 and 19:00, they are shorter due to the increased daytime mobility. This corresponds to the intuitive understanding of long, nocturnal stays that are represented by the higher quantiles.

Figure 8 represents the median cell switch rate over the
course of a day, i.e. the median amount of cell switches per hour. There is a significant slow-down of mobility at night (between 0:00 and 6:00), and an increase in mobility at daytime, with two peaks around 11:00 and 19:00, which indicates the rush hour commuter traffic.

Figure 9 shows Quantile-Quantile plot of the fitted Weibull variates against the estimated dwell times for site 559. Generally speaking, we obtain better fits at daytime, and slightly worse fits at nighttime, but still acceptable results, as indicated by the Kolmogorov-Smirnov-Test we performed (cf. Table I).

In Figure 10, we can see what the mean dwell times between 11:00 and 12:00 are on a country-wide scale. They appear to be higher in the remote, central are of the country, while there are lower dwell times along the main routes. The Dakar area exhibits behaviours of both kinds. Generally speaking, zones with both activity and transit appear in orange, while pure transit is displayed in a lighter color (corresponding to shorter mean dwell times).

B. Synthetic Population Generation

In future work, we plan to create a generative population model representing user mobility as sequences of visited sites combined with the dwell times at these sites.

It is possible to define such a synthetic population mobility model using two components:

1. The dwell time distribution parameters per site and arrival hour (time slot).
2. A trajectory catalog that includes the observed and inferred single-day site sequences and the corresponding average starting time and distance covered.

The model proposed in this paper provides these necessary inputs. This would allow the creation of an artificial population by sampling site sequences and corresponding times of stay.

In order to reproduce the target demographic statistics, we use census data to estimate the amount of users living in each site area and sample the corresponding amount of trajectories starting there. Potential sources for this could be the CIESIN project (as proposed in [6]), or the official Senegalese census.

Figure 11 shows a possible methodology for the generation of a synthetic population. For each site in the network, an amount of users according to the census data is generated from the CDR data using our model.

Figure 11: Proposed Scheme for Synthetic Population Generation

As the model results look promising, we want to evaluate it using different simulation frameworks (such as MATSim) and for different applications. Ideally, we would like to compare the modeled dwell times to the actual observed dwell times in the mobile phone network sites. This would allow us to further refine our definition of the function linking to the different attractivity measures. Ultimately, as mobile phone operators in more countries could provide anonymous and geospatially anonymised data, the model proposed in this paper would allow the creation of synthetic populations from the data for various purposes.

Furthermore, in order to obtain an entirely generative model, it would be of interest to replace the trajectory catalog...
Fig. 9: QQPlot of Fits for Site 559 in SET2_P02.CSV
Estimated Mean dwell times in Senegal given arrival between 11:00 and 12:00, thousand seconds

Fig. 10: Heatmap of estimated dwell time means given arrival between 11:00 and 12:00
by a Markov model of cell transitions. This would yield a fully generative mobility model. It remains to be investigated whether there is significant statistical bias among the mobile phone user vs. non-user population, and to what degree this would impact the resulting simulations.

VI. Conclusion

In this work we have presented a model of mobile network cell dwell times combining the different datasets of the D4D challenge. We have introduced a novel way of estimating cell dwell times based on cell attractivity factors (both on single-user and network scale). Also, we found a weight function that produced dwell times that can be fitted using a Weibull distribution. Furthermore, we proposed as future work a way of generating a synthetic population using this model, which can be used as an input to perform various simulative studies (e.g. for transportation planning).

ACKNOWLEDGMENT

The authors would like to thank the National Research Fund of Luxembourg (FNR) for providing financial support through the CORE 2013 MAMBA project (C13/IS/5825301).

REFERENCES


Improving Mobility Predictability based on User Clustering on D4D Dataset

Jaeseong Jeong†, Mathieu Leconte† and Alexandre Proutiere†

Abstract

We study a mobility prediction problem, where we want to select a next location based on the mobility patterns learned from the history trajectories. In practice, the mobility predictors can suffer from the lack of training data. Motivated by the empirical study on the large-scale D4D dataset, where the similarities between mobility patterns of users are exhibited, we develop an advanced mobility predictor which can improve predictability by replenishing its training data with the data of other users selected by clustering. Through trace-driven simulation using D4D SET2 data, we show the high performance gain of the proposed predictor against the existing predictors.

I. INTRODUCTION

Predicing human mobility has received a great deal of attention, strongly motivated by wide range of applications. We can imagine several examples which can be enabled by the mobility prediction: 1) location-based services can be provided to users in advance of their movement (e.g., mobile advertisement, recommendation, risk alarm), 2) urban traffic engineering and forecasting can be more precise, 3) the protocols in wireless mobile networks (e.g., scheduling and handover management, data prefetching) can be improved based on the prediction of mobile users’ cell associations [1]. Thus, it becomes of great importance to increase the prediction accuracy.

The mobility trajectories generated by users contain several inherent patterns due to their regular behaviors such as commuting to an office or visiting preferable restaurants. If those patterns can be recognized by training from the history trajectory, a prediction becomes available. However, there exist several challenges associated with gathering the history trajectories: 1) detecting the current location with sensors (e.g., GPS, Wi-Fi and cell tower) consumes non-negligible energy, 2) the training duration for collecting the sufficient trajectory for the prediction is long (e.g., typically, longer than one weeks), 3) users may hesitate to log their trajectories due to privacy leakage. If the trajectory is not gathered sufficiently due to the above reasons, the patterns can not be recognized or contain high noise, which in turn induces an incorrect prediction.

There have been extensive studies on mobility predictors [2]–[5]. The most popular one is order-\(k\) Markov predictor [2], which predicts the location towards which the transition is the most likely given its current location with respect to a kernel estimated by the gathered trajectory. When the mobility follows order-\(k\) markov model, it provably achieves the optimal performance at the asymptotic region where the history trajectory of the user is infinitely long [5]. However, when the length of gathered trajectory is short, Markov predictor suffers from incorrect predictions due to the lack of trained patterns. Since other predictors [3], [4] are trained by the same trajectory, they also fails to predict when the length of gathered trajectory is short.

In this paper, we develop an advanced predictor which performs well irrespective of the amount of training data, by incorporating a non-parametric bayesian approach [6]. The main idea behind this approach is to suppress the noise from the small amount of individual training data (i.e., short length of the collected trajectory), by merging the trajectories of a group of users whose inherent patterns are figured out to be similar. We refer to such groups of users as clusters. Accordingly, the proposed predictor newly includes a cluster learning layer. This idea is validated by our empirical study with a large-scale dataset provided by Orange D4D, where similarities between many pairs of users in their mobility patterns are exhibited. Then, we compare the performance of the proposed predictor with other existing predictors under the large-scale dataset, and show up to 65% prediction gain against the existing predictors, especially when they don’t have sufficient training data.

†: KTH Royal Institute of Technology, EE School/ACL, Sweden, email: {jaeseong, mleconte, alepro}@kth.se
A. System model

Data collection on users. We consider a set of users denoted by \( \mathcal{U} \). All users are moving on a common set of locations \( \mathcal{L} \) whose element is base station (BS) ID. We denote by \( x^u_t \) BS ID associated by user \( u \) at time \( t \), time being discrete here. We also write \( x^u_{1:t} \) for the trajectory of user \( u \) up to time \( t \), and \( x^{\mathcal{U}} \) for the collection of trajectories of a set of users, \( \mathcal{U} \). Throughout this paper, we denote random variables associated to locations as an uppercase letters \( X \), whereas the realizations are denoted by lowercase letters \( x \).

Prediction on a server. Whenever users update its BS association, they report their BS ID to the server. The server is responsible to collecting \( x^{\mathcal{U}} \) and predicting users’ next locations. \( c^u \) denotes a cluster in which user \( u \) is involved, and \( \hat{c}^{\mathcal{U}} \) denotes a cluster assignment of a set of users, \( \mathcal{U} \).

User mobility prediction model. Throughout this paper, we assume that the users are moving according to a Markovian model of order 1 (i.e., each user \( u \) is associated to a mobility kernel \( \theta^u \) which is a probability transition matrix over the set of locations \( \mathcal{L} \)). In other words, if user \( u \) is at location \( i \in \mathcal{L} \) at time \( t \), then the probability that \( u \) is at location \( j \in \mathcal{L} \) at time \( t + 1 \) is given by \( \theta^u_{i,j} \). We denote by \( \Theta \) the set of such mobility kernels.

B. Problem statement

Given the history \( x^{\mathcal{U}} \) of all users, we let \( f^u : \mathcal{L}^{t+1} \times T \rightarrow \mathcal{L} \) be a predictor for \( u \)’s next location, where \( T \) is the length of trajectories. The predictability of a predictor \( f^u \) denoted by \( \pi^u(f^u) \) is given by the probability that it accurately predicts \( u \)’s next location, (i.e., \( \pi^u(f^u) = \mathbb{P}(X^u_{t+1} = f^u(X^{\mathcal{U}}_t)) = \mathbb{E}\left[\theta^u X^{\mathcal{U}}_t, f^u(X^u_t) \middle| X^{\mathcal{U}}_t\right] \)).

Consequently, the objective of the predictor \( f^u \) throughout this paper should be

\[
\max_{f^u} \pi^u(f^u) = \max_{f^u} \mathbb{E}\left[\theta^u X^{\mathcal{U}}_t, f^u(X^u_t) \middle| X^{\mathcal{U}}_t\right] \quad (1)
\]

Then, the optimal prediction \( f^{u*} \) at time \( T \) which achieves the objective in (1) is given by

\[
f^{u*}(x^{\mathcal{U}}) = \arg \max_i \mathbb{E}\left[\theta^u X^{\mathcal{U}}_t, f^u(X^u_t) \middle| x^{\mathcal{U}}_t\right] \quad (2)
\]

But, unfortunately, the optimal prediction in (2) is infeasible, because the prior distribution of \( \theta^u \) which is necessary in computing the bayes estimator in (2) (i.e., \( \mathbb{E}[\theta^u | x^{\mathcal{U}}] \)) is unknown. Thus, we instead take a non-parametric bayesian approach to approximate the bayes estimator.

III. PREDICTION ARCHITECTURE

A. Overall summary

The proposed predictor of user \( u \) predicts the location towards which the transition is the most likely for \( u \) given its current location with respect to a kernel \( \hat{\theta}^u \) computed by a non-parametric bayesian approach. In particular, \( \hat{\theta}^u \) is drawn from Dirichlet process (DP) mixture model [6], [7] which is a well-known non-parametric bayesian model. It is worth to note that a sample from posterior distribution under DP mixture model is realized as a cluster structure (i.e., \( \hat{c}^{\mathcal{U}} \)), and thereby, a user clustering is essentially involved in the computation of \( \hat{\theta}^u \).

We aim at computing the bayes estimator under DP mixture model for \( \hat{\theta}^u \) as follows.

\[
\hat{\theta}^u = \mathbb{E}_g[\theta^u | x^{\mathcal{U}}] = \int_{\hat{c}^{\mathcal{U}}} \mathbb{E}_g[\theta^u | c^{\mathcal{U}}, x^{\mathcal{U}}] d\pi_g(c^{\mathcal{U}} | x^{\mathcal{U}}) \quad (3)
\]

where \( \mathbb{E}_g[\cdot] \) and \( \pi_g(\cdot) \) denote the expectation and probability distribution under DP mixture model, respectively. To handle (3), the prediction procedure is composed of three components (i.e., mobility learner, cluster learner and mobility predictor) as illustrated in Fig 1. We now elaborate each of them.

B. Mobility learner

The role of mobility learner is to monitor and collect the trajectory of users (i.e., \( x^{\mathcal{U}} \)). Each trajectory is obtained by logging the sequence of associated BS IDs of the user. Typically, the association information is reported to the server during the voice communication. The mobility learner delivers the trajectories of all users to the cluster learner and mobility predictor.
C. Cluster learner

Cluster learner is responsible to sampling cluster assignments of users from the posterior distribution of DP mixture model conditioned on \( x^L \) (i.e., \( p_g(c^L|x^L) \) in (3)). Sampling \( c^L \) multiple times and averaging them can replace the integration over \( c^L \) in (3) by the law of large number. To sample a cluster assignment, we use Gibbs sampling algorithm [8] with parameters \( G_0 \) and \( \alpha \). Using this sampling algorithm, cluster learner parallelly samples \( c^L \) \( B \) times and delivers the sampling results to mobility predictor. We denote the cluster assignment of \( b \)-th sample by \( c^{U,b} \). After the end of \( B \) sampling, the parameters \( G_0 \) and \( \alpha \) are updated for the next sampling phase as follows.

\[
G_0(\theta) = \frac{1}{B} \sum_{b=1}^{B} \sum_{c=1}^{K_b} \frac{|\mathcal{U}_{c,b}|}{|\mathcal{U}|} p(x^{\mathcal{U}_{c,b}}|\theta)
\]

\[
\sum_{n=1}^{\lfloor K_b \rfloor} \frac{\alpha}{\alpha + n} - 1 = \frac{1}{B} \sum_{b=1}^{B} K_b
\]

where \( K_b \) is the number of clusters in \( b \)-th sampling, and \( \mathcal{U}_{c,b} \) is a set of users whose associated cluster at \( b \)-th sample is \( c \).

\(^1\)Conditioned on the trajectories of users, Gibbs sampling algorithm builds a markov chain whose state is a set of feasible cluster assignments, and run the markov chain until it converges to its stationary distribution. In previous studies, this sampling algorithm is also described as a clustering algorithm when the number of clusters is unknown.
D. Mobility predictor

Given each sample of cluster assignment \( \hat{c}^{U,b} \), mobility predictor computes the expectation of \( \hat{\theta}^u \) in (3) based on the trajectory aggregated over users in a same cluster as follows.

\[
\mathbb{E}_g[\theta^u | \hat{c}^{U,b}, x^T] = \frac{\int_{\theta} \theta \cdot p(x^V_b | \theta) dG_0(\theta)}{\int_{\theta'} p(x^V_b | \theta') dG_0(\theta')}
\]

(5)

where \( V_b \) is a set of users who are in a same cluster with \( u \) at \( b \)-th sample (i.e., \( V_b = U_{c_u,b,b} \)). By averaging the results of (5) over \( B \) samples, we arrive at an approximation of \( \hat{\theta}^u \) in (3).

\[
\hat{\theta}^u \approx \frac{1}{B} \sum_{b=1}^{B} \frac{\int_{\theta} \theta \cdot p(x^V_b | \theta) dG_0(\theta)}{\int_{\theta'} p(x^V_b | \theta') dG_0(\theta')}
\]

(6)

Consequently, computing \( \hat{\theta}^u \) by (6), the predictor of user \( u \) at time \( T \) predicts the next location by \( \arg \max_j \hat{\theta}^u_{x^u,T,j} \).

IV. EXPERIMENTAL RESULT

A. Simulation setting

We use SET2 data [9] which includes individual trajectories (timestamps and associated base station (BS) IDs). To reduce the complexity in the experiments, we select 50 mostly crowded BSs among 1666 BSs in terms of the number of associated users during 2 weeks. Then, trajectories of all users are generated over a set of selected BSs, \( L \). Also, we randomly pick 200 users among the set of users who had associated more than 10 selected BSs during 2 weeks (i.e., users who had moved around more than selected 10 BSs).

To evaluate the performance of the proposed predictor (Bayes), we test order-1 Markov predictor (Markov) and aggregated predictor (Agg) for comparison. Markov predictor estimates the kernel based only on its own trajectory, and predicts the most likely location. Aggregated predictor estimates the kernel with the combined trajectories of all users, and predicts the most likely location. In the implementation of Bayes, the number of samples \( B \) is set to be 12. We further introduce an ideal predictor of user \( u \), \( f^{u,I} \), which gives an upper bound of predictability. The ideal predictor of \( u \) predicts the most likely location in terms of the ground-truth kernel, \( \theta^u \). Note that, since \( \theta^u \) is impossible to be known in advance, the ideal predictor cannot be implemented in practice. In our simulation, we set \( \theta^u \) as a kernel estimated by an entire trajectory of \( u \) (2 weeks trajectory).

B. Similarity with respect to predictability

We evaluate the similarity between users’ mobility in the dataset. Similarity between user \( u \) and \( v \) is defined as the normalized predictability of user \( u \) achieved by the ideal predictor of user \( v \), (i.e., \( \pi^u(f^{v,I}) / \pi^u(f^{u,I}) \)). Note
that this metric is asymmetric, and the high value of similarity between \( u \) and \( v \) means that, if the predictor of user \( u \) has less training data, it can replenish the training data with the trajectory of user \( v \) without hurting the prediction performance.

Fig. 2 plots the similarity between 200 users in the dataset. To show the underlying structure of similarity, users are clustered and re-arranged in \( x \) and \( y \) axes, accordingly. We observe that similarities between many pairs of users are almost 1, and some of users are clustered. Motivated by this observation, the predictor can circumvent the lack of training data by merging with the training data of other users who turn out to be similar (or in a same cluster).

### C. Predictability

Fig. 3 shows the fraction of accurate predictions of all users over time. \((i.e., \frac{\text{(# of correct prediction of users at time } t)}{\text{(total prediction of users at time } t)})\). At the start of the simulation, all users don’t have any training data. The proposed predictor (Bayes) outperforms other predictors over all time span. At the first phase between 0 and 30 hours, an oscillation is observed due to the small number of total prediction counts. After 30 hours has passed, the performances of all predictors start to increases over time, as the training data is accumulated. During the first and second days at which the individual training data is insufficient, Bayes outperforms Markov by up to 65%, because Bayes successfully use more training data merged over clusters. Then, as the training period increases, the performance of Bayes and Markov keep increasing, whereas the performance of Agg does not increase. At the end of simulation, the gain of Bayes over Agg becomes 17%. We also observe the distinction between the upper bound and other predictors even at the end of simulation. This is because the given trace is only for 2 weeks, which is not enough to show asymptotic performance of predictors. Note that, if the time goes to infinity, Markov and Bayes provably achieves the performance of Upper bound.

### V. Conclusion

In this paper, we develop a mobility predictor based on a non-parametric bayesian approach, motivated by the empirical study on the large-scale dataset, where the similarities between users are observed. The main contribution of the proposed predictor is that it manages to find a prediction gain by learning the trajectories of other selected users. However, this paper only discusses the prediction under a simple mobility model (order-1 Markov model). Therefore, it would be interesting to extend the predictor to account for more complicated mobility models as a future work.
REFERENCES

From Digital Footprints to the Dynamic Population Distribution and Road Network Efficiency
Lei Dong 1, Ruiqi Li 2*, Jiang Zhang 2
1. School of Architecture, Tsinghua University, 100084, Beijing, China
2. School of Systems Science, Beijing Normal University, 100875, Beijing, China

Abstract
High resolution of dynamical population is of great importance for policy, operational decisions, and research across many fields, but remains constrained by the logistics of censuses and surveys. And how efficient the road network of a city is crucial to its development. In this article, we mainly focus on these two aspects. We introduce a method to identify users’ home and work places from mobile phone data, and then we analysis the day and night, local and non-local population distribution in Dakar, the capital of Senegal. We show the dynamical patterns of human mobility and ‘hotspots’ defined by population density. We combine the users’ records with road network data, and get the correlation coefficient and unmatched region between population density and road density in Dakar. We also introduced a road network efficiency indicator, which provides a quantitative measure to guide transportation infrastructure development. Above all, the traffic flow patterns, ‘hotspots’ and unmatched regions could all be scientific basis for the government to make future urban planning policies.

Introduction
With accumulation of metadata, especially mobile phone data, there rises a research trend about human mobility [1], related processes (such as epidemic spreading [2, 3]) and cities, as well as urban dynamics [4, 5]. Compared to other data such as car GPS [6], RFID [7], social networks data such as Twitter or Foursquare [5, 8], mobile phone data provides a much more complete picture of the human activity [9] especially for the mobility, which pave a better way for understanding cities. Mobile phone data is just like some digital footprints left by the people living in a city, which gives us a trace and clue to get to know a city better and more objective, not only the human dynamics but also the structure or even efficiency of a city.

Meanwhile, the urbanization problem of Africa has attracted more and more researchers’ attention. However, for African studies, the lack of reliable data has always been a big obstacle to do sound empirical analysis. Therefore, many researchers used proxy variables to estimate the distribution of the population and socio-economic variables for countries in Africa, including AfriPop project (www.worldpop.org.uk/), GPW, GRUMP, LandScan, UNEP [9] and VIIRS (http://ngdc.noaa.gov/eog/viirs/). However, as far as we know, there’s little high-resolution dynamical population data within city scale in Africa.

How people spatially distribute in cities? How they interact with each other? These questions are crucial to socio-economic development. The density (e.g. agglomeration [10], scale economies), distance to resources and spatial division all influence the structure and economic output of cities [9, 11]. Planning development and policy-making also require the data of spatial distribution of populations, settlements. For example, with the population and road network data, we could detect the unmatched areas between roads and residents, which could be helpful to urban planning in the future.

Lei Dong and Ruiqi Li contributed equally.
* Email: rickylee3380@gmail.com
What’s more, residents in cities are not static, they are moving, interacting, which can be appropriately depicted by network flows [12]. So we also want to know how the people distribute during the day and night. The mixed population (weighted by population density during different time) could be an important indicator for the city’s activities and some local performance of a region.

In this paper, we propose a method to detect a user’s home and work place (i.e. OD) from their mobile phone records, and then with all the users’ data, we can get this dynamical population. Our method also paves a way to estimate the working population and resident population in a specific region, which can be useful to some policy making.

Since we obtain all the OD pairs, we may wonder how efficient the commuting process is, which may be quite crucial for personal feeling and the overall efficiency of a city. There’s some research work shows that distance to work place can be import explanatory factors in socio-economic outputs. On nation scale, the GDP seems correlates with the travel times, longer the travel time is, lower the GDP per capita is [9]. With better data, the detection of anomaly routes may provide suggestions to infrastructure development and improvement. For example, the OD pairs with high volume and abnormal long route distance would be considered as a road need improvement. We also introduced a road network efficiency indicator to describe the commuting efficiency quantitatively, which could be applied to other cities to do comparative study.

With the population and road network data, we detect the unmatched areas between roads and residents, which may give us a clue for road network development. We show the unmatched areas between top-down planning such as road networks and bottom-up growing such as population dynamic distribution, which could be helpful to society development, urban planning policy, and well-being of the residents in Senegal.

Results

Data description. Our analysis is based on an anonymous mobile phone dataset of Call Detail Records (CDR) of phone calls and text exchanges between more than 9 million of Orange's customers in Senegal from January 1, 2013 to December 31, 2013 [13].

The original dataset contained more than 9 million unique aliased mobile phone numbers. When preparing datasets, Orange, Inc. retained only users meeting both of these criteria: 1. users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset); 2. users having had an average of less than 1000 interactions per week. The users with more than 1000 interactions per week were presumed to be machines or shared phones. [13]

The provided datasets under secrecy agreement are: (a) one year of antenna-to-antenna traffic for 1666 antennas on an hourly basis, (b) one year of fine-grained mobility data (site level) on a rolling 2-week basis for a year for about 300,000 randomly sampled users meeting the two criteria mentioned before, (c) one year of coarse-grained mobility data at arrondissement level with bandicoot behavioral indicators at individual level for about 150,000 randomly sampled users. [13]

Research region selection and general features. We use the first two dataset for our research and mainly focus on the region of Dakar, because Dakar City is not only the capital but also the largest city of Senegal. And according to Dataset2 (fine-grained mobility data on every 2-weeks), we find
that there are more than 170,000 users appeared in the region of Dakar, which accounts almost 60% of total users provided (300,000). And from the traffic volume obtained from Dataset2, we can obtain all their mobility traces (see Fig.1) and we can clearly find that the region of Dakar is of highest volume.

![Fig. 1](image)

**Fig. 1** (a) the cumulative traffic volume of one year of all the mobile phone users. Once his/her site location ‘B’ is different from previous site ‘A’ and the time interval is not too long, then we identify it as a direct mobility event. (b) the stable dominant origins-destinations (ODs) pairs extracted from mobility with timestamp information (see OD detection and Method). And there’s a significant ‘twin-city’.

Dakar is on the western edge of African mainland, and the special location helps it to be an important regional port for African-European trade. The Dakar region includes four departments: Dakar, Guediawaye, Pikine and Rufisque. The research area and the call density of Dakar are shown in Fig 2. If we regard the call density as a proxy of activity of a specific area, then we can observe an interesting ‘breathing’ phenomenon of the city (see SI). We find that plateau and Medina is the most active region which is also in agreement with common sense.
**Fig 2.** The main research region and the average daily call density of this area. We perform the Voronoi tessellation with the given distribution of antenna, and calculate the area of the polygons.

For a closer look of mobile phone call’s pattern, we find that quite different from some other cities in developed country (such as Madrid, Barcelona) [4], the number of mobile phone calls is still very high during the weekend and the average call durations are almost the same (see Fig. 3), which may indicates that they tend to have a busy weekend for entertainment or other business. Another guess is that they don’t clearly differentiate weekday and weekend, and just enjoy a leisure life style. What’s more, people in Dakar seems quite enjoy the nighttime life, they are quite active during the night and make a lot of long time phone calls, and the calling duration is very high in the night (especially the late night). Besides, they also text a lot in the night.

We also find out some common feature of the calling pattern with developed cities, that there are two peaks in phone calls activity, one in midday and another in the night (about 21-22 pm), and the time for second peak is a little later than most Spanish cities [4].

Another interesting unexpected finding is that the average call time during the day is about 60 seconds (especially of January, it’s exactly 60 seconds, see SI), which may be influenced by the charges mode or may be just a reflection of daytime work character.

**Fig. 3** (color online) (a) the averaged duration of all the mobile phone calls (b) the number of mobile phone calls (c) the average call duration (d) the number of SMS for weekdays and weekends.
**Day-night population distribution (OD detection) and the matching indicator between road and population.** According to the time of mobile phone call activities, we are able to determine the hotspots of day and night, which could give us the estimated working population during day and residence population at night. The detection of daytime and nighttime population can be useful to medical and educational resources attributing.

The day and night sampled users’ distribution (see section Methods for the detailed information) provides an interesting indicator about the organization of Dakar. Fig. 3(a) and (b) show that most of the users are distributed in Dakar Plateau, Grand Dakar, Médina, and Parcelles Assainies. Among them, the Dakar Plateau on the south of Dakar is the historical heart of the city.

Considering the sample accounts for less than 10% of the total population in the city, so the averaged weighted population (see section Methods for the definition) could be more than 16,000 person/km$^2$, much larger than most of big cities around the world. And the population density in central area is over 40,000 person/km$^2$.

We also calculate the road network density under the same grid (Fig. 3(c)) to test the correlation between population and roads. In theory, the higher the population density is, the more roads should be constructed to meet people’s needs. And as expected, the road network density $\rho_R$ and population density $\rho_P$ have a very strong positive correlation. For Dakar, the relation is $\rho_R \sim \rho_P^{0.257}$. The correlation coefficient 0.257 is lower than cities of developed countries, which is 0.5 as previous research shows [14]. It means that compared with the high population density, the road network density is not enough. That may partly because the Open Street Map (OSM) data are incomplete in some developing countries, but more importantly, it reveals that the urban infrastructure in Africa is inadequate during the fast process of urbanization. This statistical method could also help us find the unmatched areas, such as site 2, 79, 314 shown on Fig. 3(c) and 3(d). These results will provide a scientific basis for the city governors and urban planners to make the plans for the regional development.
**Fig. 4** (a) Day (b) night mobile phone users’ density, which could be an estimation of population density there, (c) road density, (d) scaling relation between road density and mixed population density according to the time a user spent in that region.

**Fig. 5** (a) daytime and (b) nighttime density of Local users; (c)(d) for the nonlocal users. The spatial distribution pattern and magnitude level are quite different from each other.

**Spatial analysis of OD and Road network structural efficiency.** After obtaining the OD pairs, we want to know how people commute from one place to another. Then naturally comes to an important question: is this city efficient enough? Does the people have a good experience when commuting to work places? To answer these questions quantitatively, we need to know how efficient a specific route is. If two places are geometrically close (with a small Euclidean distance, assign as $d_e$) but the route distance ($d_r$, see Methods for how to obtain $d_r$ between two places) is large, then we can assume this route’s structural efficiency ($\eta$) is low and needs improvement. For the overall city, we can get the overall structural efficiency weighted by the population proportion commuting on each route. Then we can get overall structural efficiency

$$\eta = \frac{d_r(O_i, D_j)}{d_e(O_i, D_j)} \cdot \frac{P(O_i, D_j)}{\sum_{O_i,D_j} P(O_i, D_j)},$$

(1)

where $P(O_i, D_j)$ stands for the number of persons commuting from place $i$ to $j$.

This structural efficiency provides quantitative measures to guide transport infrastructure development strategies. For Dakar, $\eta = 1.04$, which indicates a good structural efficiency. With more data of other large cities, we can compare the overall commuting efficiency of a city.

What’s more, we also find that in Dakar, the linear relation by Ordinary Least Squares (OLS)
between route distance and euclidean distance is
\[ d_r = 1.18d_e + 458. \] (2)
there are several anomaly pairs which may indicates this route needs improvement.

<table>
<thead>
<tr>
<th>O</th>
<th>D</th>
<th>Euclidean Distance</th>
<th>Users</th>
<th>OD Distance</th>
<th>DO Distance</th>
<th>Route distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>367</td>
<td>367</td>
<td>636.46</td>
<td>651</td>
<td>636.46</td>
<td>636.46</td>
<td>636.46</td>
</tr>
<tr>
<td>401</td>
<td>401</td>
<td>928.37</td>
<td>632</td>
<td>928.37</td>
<td>928.37</td>
<td>928.37</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>413</td>
<td>412</td>
<td>1708.46</td>
<td>20</td>
<td>2104</td>
<td>2104</td>
<td>2104</td>
</tr>
<tr>
<td>460</td>
<td>437</td>
<td>3423.96</td>
<td>20</td>
<td>4104</td>
<td>4013</td>
<td>4058.5</td>
</tr>
</tbody>
</table>

Table 1 OD analysis. The Euclidean distance is computed by the coordinates of O and D sites.

Users is the number of people travel by this route. Route distance is the average of OD and DO distance, which are returned by Google Map API (see methods for detailed information).

Fig. 6 (a) the averaged OD pairs in Dakar region extracted from the mobility events. (b) Scaling relation between road distance (i.e. real routing distance) and Euclidean distance, a superlinear relation with exponent 1.08 is observed, and there are several anomaly pairs which may indicates this route needs improvement. And the size of a node represents the OD volume. The gray line is of slope 1.

Discussion
In this paper, we present a method to detect users’ home and work places (i.e. OD) from mobile phone data, and with users’ data in Dakar, we analysis the day and night, local and non-local population distribution. This result reveals not only the dynamic process of human mobility, but also could be used to discover “hotspots” and “backbones” in cities. Combined with other datasets such as road network data, we get the correlation coefficient and unmatched region between population and infrastructure in Dakar region. The “hotspots” and unmatched region could both be scientific basis for the government to make future urban planning.

We also introduced a road network efficiency indicator, which provides a quantitative measure to guide transportation infrastructure development strategies.

There are also a lot of subjects that could be addressed in future research:
First, the determination of local and nonlocal users require further study, at present, the selection of the threshold introduces some arbitrariness, even though it may sounds make sense.
For example, maybe we can set the threshold as the average time spent by all the persons appeared in one place. And maybe we can borrow some idea from the Lorenz curve, a standard object in economics. What’s more, local and nonlocal maybe a relative concept, especially in the marginal region or between some “twin city”, which also observed in this study as shown in Fig. 1(b).

Second interesting question maybe the natural city detection based on commuting patterns, which may go beyond the administrative boundary, since the commuting bonding the city together. The determination of city boundary is of great importance about determining scaling laws [19], policy making. And what’s more, it may reveals how our city really looks like: there may be some region geometrically located within the city, but seldom visited or interacted with other components, then it’s actually not an active region exclusive from the city. In that sense, we may see a real fractal city [20].

What’s more, we also wonder whether commuting tells us more. Since we have the OD matrix, then we can get the commuting volume. We may be able to get an Engel coefficient about travel behavior. The ratio between commuting volume and total traffic volume maybe is an interesting indicator for how wealth this person is. We can imagine that once a person go out is all about going to work, then this guy must be not that rich, he/she may suffers from hard and long time work. If we have this average ratio citywide, then it may tell us some interesting story about how the people are in this city. We may also be able to get a corresponding Gini curve for the whole city.

Urbanization will last for decades in Senegal. Only with better understanding of what is happening in the cities, can people enjoy the optimal socioeconomic growth and innovation in urban societies.

Methods

Voronoi tessellation. We divide Dakar into sub-region by Voronoi grids, which are calculated by the geographical coordinates of antennas. As mentioned in D4D data materials, for commerical and privacy reasons, the organizer assigned a new position to each site uniformly in its original Voronoi cell.

Weighted population density. The standard definition of population density is total population divided by total land (or urbanized) area. While, a more meaningful method is “weighted population density” (assign as $\rho_{w,p}$), which is computed by:

$$\rho_{w,p} = \frac{P_i}{\sum_i P_i} \rho_i,$$

where $\rho_i$ is the density of a sub-region $i$, which is divided by antenna Voronoi grids in this paper, and $P_i/\sum_i P_i$ is the weight assigned to corresponding region. This discounts large, barely populated areas, and gives more weight to densely-populated areas. It’s much closer to the real population density in cites of unbalanced population distribution.

Determination of duration of day and night. By integrating all the records of one users during two week in the dataset 2 provided by Orange, Inc., we can get a clear list places he/she visited with information about both when and how many times. Then we can count the locations he/she appears at night and day, and regard the place appeared with highest frequency as home and workplace, respectively. So the determination of duration of day and night is crucial (especially the night, due to abundant night life activities). As shown in Fig. 1(b), we can clearly see two peaks in phone call activities, which we assume correspond to work and entertainment. So after the second peak till next morning, we assume it’s time of staying at home. The duration of day is the work
hour without commuting time (i.e. daytime without the period of rapid increasing of activities and the second riseup).

Then we can get all users’ OD. For the one who fail to detect an O or D, we assign it as NAN and didn’t draw it out on Fig. 6. For Fig. 4 and 5, if O is unknown but D is known (or vice versa), we still accumulate it as daytime population and draw it out together.

**Road network data.** The data of road networks are accessible in [http://metro.teczno.com/](http://metro.teczno.com/) (maps extracted from Open Street Map [http://www.openstreetmap.org/](http://www.openstreetmap.org/)). And we firstly convert the Mercator coordinates to the longitude and latitude, and then we calculate the road length density of each Voronoi cell by QGIS.

**Road network efficiency.** With Google map’s API ([http://maps.googleapis.com/maps/api/distancematrix/json?origins=]&[destinations=]&[mode=driving&sensors=false]), we can get the travel time and route distance between two locations. Since we are already able to calculate the Euclidean distance between these two locations, we can get the ratio of real route distance and Euclidean distance as the structural efficiency of this road.

And there’s also amount of OD pair is a self-loop, which means that people live here and here. Then the route distance is assigned as average length of that region. We actually omit these self-loops in the analysis.

What’s more, we also want to know the time efficiency which can be regarded as the ratio of route distance and travel time, however, this is limited by Google’s API, which seemingly cannot return the real time data of travel time.

**References**


Supporting Information.
All the supporting information could be downloaded from http://www.idonglei.github.io/d4d

Acknowledgment
Ruiqi Li acknowledges Guoyu Wu from Beijing International Studies University for the helpful discussion about the general habits and behaviors of residents in Dakar. Lei Dong acknowledges Jianzhu Wang for discussion about the transportation situation in Africa.
Using mobile phone data for Spatial Planning simulation and Optimization Technologies (SPOT)

Serigne Gueye¹, Babacar M. Ndiaye³, Didier Josselin³, Michael Poss⁵, Roger M. Faye², Philippe Michelon¹, Cyrille Genre-Grandpierre¹, and Francesco Ciari⁴

¹ Laboratoire d’Informatique d’Avignon (LIA), Université d’Avignon, France
² Laboratoire de Traitement de l’Information (LTI), ESP- Cheikh Anta Diop University, Dakar, Sénégal
³ Laboratoire de Mathématiques de la Décision et d’Analyse Numérique (LMDAN) FASEG-University of Cheikh Anta Diop, Sénégal
⁴ Institute for Transport Planning and Systems (IVT), Zurich, Switzerland
⁵ UMR Etude des Structures, des Processus d’Adaptation et des Changements de l’Espace (ESPACE), Avignon, France

Abstract. We propose in this paper a methodology to find locations or relocations of some Dakar region amenities (home, shop, work, leisure places), that may reduce travel time or travel distance. The proposed methodology mixes multi-agent simulation with combinatorial optimization techniques; that is individual agent strategies versus global optimization using Geographical Information System. We use MATSim as a multi-agent simulator system, and need for that to generate agent plans. Some additional methods are thus proposed to generate representative agent plans from mobile phone data provided by Orange. Some preliminary numerical results are presented on the Dakar region showing the potential of the approach.

Keywords: amenities location, multi-agent simulations, combinatorial optimization, local search, clustering, GIS, planning

1 Introduction

Many urban areas in the world, especially in developing countries, are faced to a rapid population density increase, that generates a transport demand that cannot be supported by transport infrastructures. Between 1976 and 2005, the population in the Dakar region had been multiplied by approximately 6⁷ while

⁷ source : Enquête ménage CAUS/2001/PDU Dakar horizon 2025
in the same time the transportation network and the urban design was not sufficiently adapted to this change. It leads to congestion problems and a reduction of the urban accessibility defined as the capacity to reach some given resources or activities, within a given time. As a quantitative measure of the accessibility in a time interval, we call global accessibility in an urban area, the sum of the whole travel times or distances (for all the people) between the urban amenities.

When thinking about suitable actions to improve the accessibility, two dimensions are usually taken into account by planners: the transportation network design and the location of amenities. Indeed, people uses the transportation network motivated by activity objectives and places located somewhere in the urban area. Thus, to improve the accessibility to the facilities to allow these activities, one should improve both dimensions of the problem.

In 2007, a planning of the Dakar urban areas over the horizon 2025 had been performed by GMAT (Groupe Métropolitain en Aménagement des Transports) and CETUD (Conseil Exécutif des Transports Urbains de Dakar) (see [6]). This study, called “Plan de Déplacement Urbain de l’agglomération de Dakar-Horizon 2025 (PDUD)” contains a series of future projects or recommendations, concerning each of this dimensions. For instances, among a very large list of projects, let us cite the construction of the highway “Patte d’Oie - Diamniadio” opened in 2013, that strongly improves the transportation network, the Diamniadio urban pole (4000 ha) whose construction started in 2014, located at 30 km of Dakar downtown, the closure of an old important inter-regional bus (Gare Pompiers), relocated in a a new more suitable and non-occupied place (Baux maraichers) in the suburb of Pikine (10 km of dakar downtown). Notice that the new activities that should take place in the old location is (to our knowledge) not yet clearly defined.

We observe that another possible relocation decision may be, instead of relocating this station in a non-occupied place, to exchange it with another existing amenities, thus solving in the same time the question to know what activities should be carried out in the former inter-regional bus station. For instance, switching with a significant commercial or shopping amenity with the inter-regional bus station would be possible. One may also consider not only relocating a single amenity, but rather finding the “best” relocation decisions, according to an objective of global accessibility optimization, that is to say relocating several amenities, in various non-occupied or occupied sites. A simple method could be to analyze all possible relocation scenarios. Nevertheless, as the number of amenities linearly increases, the number of possible scenarios increases exponentially, making intractable such an approach. This paper proposes a methodology by which a very large set of relocation possibilities can be simulated, analyzed, and the “best” one can be found, according to some quantitative measures. The methodology was coded in a prototype software called SPOT, that originated from two projects DAMA [1] and ORTRANS [12], and which is operating as
follows.

Finding good geographical locations of amenities that optimize the global accessibility measure, supposes to be able to foresee, as realistically as possible, the trip flows induced by the users moving on the transportation network, between all amenities. In this task, SPOT uses the multi-agent simulator MATSim (see Balmer et al. [2]). In MATSim, the actors of the modelled system are the agents (i.e. the city residents). The agents act according to given “realistic” rules. They try to perform some activities at different places and have learning capabilities. The overall traffic observed in the urban area emerges from the simulation as a consequence of individual agents behaviour, each pursuing his/her individual interests. MATsim basically needs three data to perform a simulation: the transportation network (network.xml), the amenities location (facilities.xml) and the initial agent plans (plans.xml). At the first MATSim iteration, each agent follows one or several possible initial plans contained in the agent plan file. A plan takes place on one day. It is defined, at least, by a sequence of activities (with their geographic locations), and a list of traveling modes (car, bus, walk, bike,...) between all successive activities. For example, an agent can initially be at home, then goes to work by car, then goes shopping by walk and finally reaches a leisure activity by car before coming back home by car. Each agent initially choose a plan. All plans are then simulated by a traffic simulation module, that computes the different routes in the transportation network. Then, agents learn about the travel time or distance experienced during the chosen plans, and try, in the subsequent iteration, to optimize his/her plan (if necessary). He can for example change the transportation mean (car, public transportation, walk and bike), the activity schedules within a certain margin, or the locations of some leisure or shopping available places. The plan optimization is simulated by a genetic algorithm [9], that in fact only concerns 10% of the population. For each agent, some possible new plans are generated, and viewed as the components of a “genetic” population. As in any genetic algorithm, the population components (here the plans) can be crossed (cross-over), muted, and each solution is then evaluated (fitness function). The evaluation consists of giving a score to a plan, called the utility. Roughly, the utility is a function defined by the sum of the utilities to perform activities minus the disutility associated to the transportation cost (see Charypar and Nagel [7]). When new plans (eventually similar to the previous ones) are chosen by an agent, a new traffic simulation is performed. Then agents learn again from the new experiences, try to find other better plans, and so on... until a fixed number of iterations is reached. In theory, for an infinite number of iterations, the system converges in a Nash equilibrium state where each agent will choose a definitive plan (see Horni et al. [10]). That is a state where no agent will have some interest to change again its plan for increasing its individual utility. In practice, for a fixed number of iterations, the system has already been tested in more than 7 large cities (Zurich and complete Switzerland, Berlin and Munich, Padang, Gauteng, Toronto, Tel Aviv, Kyoto) and show a certain ability to reproduce real-life observations.
Following a complete MATSim simulation, in SPOT we adopt a global (or collective) view which contrasts with the individual behaviour of the agents in the simulation. Given the total amount or a very large ample of flows of Origin-Destination (O-D) trips observed between all amenities, our problem is indeed slightly different: it aims at finding some suitable relocations to increase the global accessibility for a set of selected amenities. Let us remark that the MATSim simulations are operating on only one day, such as the global accessibility we seek to improve. So, to be pertinent, the simulated plans should be as representative as possible of what the agent do most frequently.

The problem of finding a good relocation is viewed as a combinatorial optimization problem and solved using a local search algorithm. The new locations provided by the algorithm are then used to update the facility file, as well as the plan provided. A new MATSim simulation is performed, followed by a new step of location optimization and so on... until a fixed number of iterations. Contrary to the MATSim simulation current process, no theoretical results guarantee that the whole iterating process handling individual agent interests (in MATSim) and the collective global optimization of the amenity (re)locations can converge to an balanced state. The figure 1 summarizes the SPOT methodology.

Fig. 1. SPOT

In the sequel, we detail in section 2 how the network and facilities files have been generated. In section 3, we show how the D4D challenge data were exploited to derive a representative initial plans for the agents. Section 4 deals with the computation of the O-D flows, the selection of the amenities to relocate and the local search procedure. Some preliminary numerical results are given in section 5. We then conclude this work and give some perspectives in section 6.
2 Network and Facilities Files

The network file is generated using Open Street Map (OSM) resources \(^8\), in particular the OSM data for Senegal provided by the Humanitarian OpenStreetMap Team (HOT) \(^9\). Using the tool Osmosis \(^10\) and the opensource Geographical System QuantumGIS, we separately extract the roads and the highways, and also a list of identified amenities with their geographical locations. Roads and highways populate the file network.xml, and the amenities are used for the file facility.xml.

Most of the time, the type of amenities in the list was not correctly annotated. We processed a semi-automatic assignment using specific requests in the QGIS database. Thus, when necessary, the activity types was fixed to home, work, shop or leisure. In particular, for the “home” type, the amenities obtained from the OSM provide district names (as Fann, Point E, HLM,...) without (of course) indicating precise individual home location in these districts, as required in the MATSim plan file. For these districts, a spatial sampling constrained by resident area boundaries was then necessary to randomly generate a large set of home locations, respecting the density distribution of the Dakar region population in the different urban districts. Some information about this distribution had been provided in the CETUD and GMAT document \([6]\). For instance, we learn in this study that, in 2007, the Dakar region working population was distributed as follows

<table>
<thead>
<tr>
<th>Ville</th>
<th>Arrondissement</th>
<th>Population</th>
<th>Pourcentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakar</td>
<td>Plateau</td>
<td>215.343</td>
<td>8.71</td>
</tr>
<tr>
<td></td>
<td>Grand-Dakar</td>
<td>254.434</td>
<td>10.25</td>
</tr>
<tr>
<td></td>
<td>Almadies</td>
<td>121.006</td>
<td>4.90</td>
</tr>
<tr>
<td></td>
<td>Parcelles Assainies</td>
<td>237617</td>
<td>9.61</td>
</tr>
<tr>
<td>Pikine</td>
<td>Thiaroye</td>
<td>239.053</td>
<td>9.67</td>
</tr>
<tr>
<td></td>
<td>Dagoudane</td>
<td>461.648</td>
<td>18.68</td>
</tr>
<tr>
<td></td>
<td>Niayes</td>
<td>209.859</td>
<td>8.49</td>
</tr>
<tr>
<td>Guédiawaye</td>
<td></td>
<td>435.350</td>
<td>17.61</td>
</tr>
<tr>
<td>Rufisque</td>
<td></td>
<td>160.860</td>
<td>6.51</td>
</tr>
<tr>
<td>Bargny</td>
<td></td>
<td>41.220</td>
<td>1.67</td>
</tr>
<tr>
<td>Sébikotane</td>
<td></td>
<td>19.400</td>
<td>0.78</td>
</tr>
<tr>
<td>Zone rurale</td>
<td></td>
<td>76.940</td>
<td>3.11</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>2,471,730</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.** Extracted from the CETUD and GMAT report \([6]\)

---

\(^8\) https://www.openstreetmap.org
\(^9\) http://wiki.openstreetmap.org/wiki/WikiProject_Senegal
\(^10\) http://wiki.openstreetmap.org/wiki/Osmosis
Although the spatial 2014 distribution probably differs from this old one, we used the same proportion of locations, due to the fact it was our unique source of reliable information. Our goal being to process simulations with a maximum of 25000 agents, we distributed 25000 points (supposed homes with an agent) according to the previous percentages in the Dakar districts contained in the file SENEegal-ARR.csv of the challenge data.

Fig. 2. Homes (grey points), facilities (yellow points), antennas (crosses), and district locations

The figure 2 was designed with the free Geographical Information System QGIS 11. It uses shape files of the whole Senegal, as well as the different CSV files provided in the challenge. A new layer composed of the transportation network was added. The figure shows (yellow circles) several location types: amenities different from home, mobile phone antennas (red cross) and home locations (grey circles).

Finally and in order to construct the plan file, for each antenna we computed the list of all the amenities within a maximal distance of a given threshold. Those amenities are then easily accessible by an agent detected within a region covered by an antenna.

11 http://www.qgis.org/en/site/
Generating initial plans is an important step generally performed using household surveys and population census. A basic MATsim plan for an agent looks like this:

```
< person id="pid120" employed="no" >
  < plan selected="yes" >
    < leg mode="car" >
      < act type="home" facility="1202" x="-17.446145" y="14.7382253" end_time="11:10:00" />
      < act type="shop" facility="476" x="-17.388572" y="14.7697" end_time="15:30:00" />
    </ leg >
    < leg mode="car" >
      < act type="home" facility="1202" x="-17.446145" y="14.7382253" end_time="23:59:59" />
    </ leg >
  </ plan >
</ person >
```

This example tells that the agent “pid120” lives at “1202”, located at the geographic coordinates $x = -17.446145$ and $y = 14.7382253$. He leaves his house at 11:10:00, by car, for shopping at the facility “476”. He then leaves the shopping place at 15:30:00 for coming back home where he stays until the end of the day.

The methodology presented in this paper is an attempt to substitute to the surveys and censuses, thanks to the exploitation of mobile phone data available in real time, while surveys require longer updating periods and are expensive in financial and human resources. For such a purpose, we are particularly interested in the challenge data set named SET2. Let us recall that the data are organized in 25 files, each file containing the list of visited antennas, over a period of 2 weeks, for 320,000 individuals, randomly selected. For each file, the sample of 320,000 individuals is renewed to ensure anonymity. Reading the files, it appears that the user detections had been made with a frequency of 10 mn. Let us notice that from time to time, several antennas can be co-located nearby, so that a call can be supported over a short period, by several antennas.

Each file contains:

- **user id**: the identifier of the person;
- **timestamp** (format YYYY-MM-DD HH:MM:00): the date and time during which the connection was made;
- **site id**: the identifier of the antenna. A second file (SITE_ARR_LATLON.csv) allows to find the associated geographic coordinates.

A short example extracted from the file SET2_P01 is given below:

```
1,2013-01-07 13:10:00,461
1,2013-01-07 17:20:00,454
1,2013-01-07 17:30:00,454
1,2013-01-07 18:40:00,327
1,2013-01-07 20:30:00,323
1,2013-01-08 18:40:00,323
```
From these data, we aim at generating some representative plans of daily trips of the Dakar (and suburbs) inhabitants during a working day. The methodology we propose is divided in several steps detailed below. Notice that in some steps (in particular the step 2), we introduce some concepts closed to a previously contribution in this topic (see Berlingerio et al. paper in [3]).

### 3.1 Step 1: Clustering of antennas

The step 1 deals with the problem of antennas co-localization, each being capable to detect, at almost the same time, a (or many) user(s). Thus it gives the illusion of a aggregated movement. We tackle this issue, by grouping antennas in clusters using a standard hierarchical ascendant clustering algorithm (see [13] for a survey on clustering methods) applied to the file SITE_ARR_LATLON.csv. At the end of this algorithm, in each cluster, the maximal distance between any couple of antennas does not exceed a given threshold. Thus an agent successively detected by two antennas in the same cluster will be then considered as motionless.

An illustration of this process is provided with the agent 2 in the file SET2_P01 who was detected, almost at the same time, by three different antennas, as described by the following lines:

<table>
<thead>
<tr>
<th>Time</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-01-10 19:30:00 408</td>
<td></td>
</tr>
<tr>
<td>2013-01-10 19:30:00 416</td>
<td></td>
</tr>
<tr>
<td>2013-01-10 19:30:00 419</td>
<td></td>
</tr>
</tbody>
</table>

Computing the geographical distances between each of these locations gives a maximum distance of approximately 4.1 km. Taking as threshold the value 5 km will have the effect to put the three antennas in the same cluster. Thus, in this case, we consider that the agent was “stopped” somewhere in the area covered by these antennas.
3.2 Step 2: Generating individual trajectories

**Definition 1.** For any agent $j$, we call “stop”, noticed $p^j = (c^j, s^j, e^j, a^j, b^j, m^j)$ a time interval where $j$ stay on a region covered by the antennas of a cluster $c^j$.

A stop is characterized by a starting date ($s^j$), an ending date ($e^j$), the type of activity performed ($a^j$: home, work, shop, leisure,...), the geographic location of the performed activity, and the transportation mode used to leave the stop ($m^j$).

**Definition 2.** For one day, we define a “trajectory” $T^j$, for an agent $j$, as an ordered sequence of stops: i.e $T^j = <p^j_1, p^j_2, ..., p^j_n>$. 

The step 2 consists initially of finding, for any SET2 file, the trajectories of the agents in each day, of the 15 possible ones, but without indicating the information at this step $a^j, b^j, m^j$. For a given SET2 file, finding the cluster ($c^j$), the starting and ending dates ($s^j, e^j$) is done by reading the file line by line. As long as two successives lines involve some antennas located in the same cluster, we consider that the agent “stops” in the corresponding area. For instance, coming back to our first example, the algorithm will find that the agent $j = 1$ “stopped” in the region defined by the cluster of the antenna 323 between 2013-01-08 18:40:00 and 2013-01-09 11:00:00.

Potentially, applying this procedure to the whole 25 files, may result in $320000 \times 25 \times 15$ trajectories. In sequel, we decided to build more reasonable set of trajectories that will considerably reduce the quantity of simulated plans used to derive a MATSim aggregated plan file.

3.3 Step 3: Finding stop activity

The step 3 tries to assign an activity type ($a^j$) to each stop. The available activity types are the following: home, work, shop, leisure. This assignment is done in a precise order: home first, then work, then shop and finally leisure.

For home and work activities, we adopt a process closed to the one used in Berlingerio et al. paper [3]. For each agent in a SET2 file, and each stop of the agent, we compute the number of hours passed in this stop during the night. If this number exceeds a given threshold, then we consider that the agent passed one night in the cluster associated to the stop. We then compute the total number of nights of each agent in each visited cluster, and retain the cluster with the highest number of nights. If this number exceeds a certain threshold, and if some home facilities exist “around” this cluster, then it is identified as its home location. Let us recall that when generating facilities files (see section 3), each antenna has been associated to a list of amenities at a maximal distance of a given threshold. By “around”, we include all the amenities belonging to, at least, one amenity list of the cluster. A “home” amenity is then randomly chosen in these lists and its geographical location assigned to the attributes $b^j$. After that,
all “sufficiently” (i.e. exceeding a given threshold) long stop of the agent detected in this cluster will be considered as a “home” activity.

We proceed similarly for identifying the work activities, taking into account the cluster with the highest total number of working hours. The working hours for a given stop must be in a given fixed time period (between 6:00:00 and 18:00) where work is supposed to start and end. It must also exceed a certain given minimal threshold supposed to be a minimal amount a working times. Working activities must also exist around the cluster where the stop is located. We additionally check that the best “working” cluster has not been yet fixed as a “home” location, assuming that (in general) working place and home are not co-located. If this case happens, we take the second cluster with the highest number of working hours. The remaining stops which are not identified as “home” or “work” are then fixed, when possible, as shop or leisure as follows.

We start by shopping activities. All stops in a cluster for which shopping amenities exist are assigned to shopping activities if the duration of the stop is large enough (in respect with a given threshold), and if the start and end of the stop is included in a given interval representing shop opening and closing times in Dakar. After this step, the last stops that are not identified as shops are considered to be “leisure”, using the same criteria as before with different duration thresholds, and different start and end intervals. At the end of the process, all the stops without any assigned activity are erased. If by removing these stops, a trajectory of an agent becomes empty, we also erase the corresponding list. All these deleted data represent a significant reduction of combinatorial in our numerical experiments.

Having obtained these purged data sets, since we seek to do a one day simulation, we choose for each agent a single trajectory among the existing ones, using different possible rules: randomly, from the longest list, from the longest list starting by a home activity (if it exists), from the longest list containing home and work activities. Let us notice that proceeding this way may lead to choose two lists dealing with two different days (for two agents). But what we want is something representative of the plans that the transport infrastructure should support. By choosing the longest list, for instance, we are interested in a kind of “worst case”.

For each file SET2, we potentially obtain 320000 trajectories, each corresponding to one agent plan. This number, although being far away from the maximal number of trajectories of one file (320000 × 15), remains too high for our purpose. We drastically reduce it using a clustering steps detailed in the subsection 3.5. However, prior to this step, we assign a mode for each stop.
3.4 Step 4: Mode assignment

The goal of the mode assignment step is to fix the mode \((m_l)\) that the agent was supposed to use for leaving a stop to reaching the next one. We consider three possible modes: car, public transport (pt) and walking. For each stop to which we want to assign a mode, we compute the agent \textbf{minimal speed} from a stop to the next one. This can be done by dividing the maximal distance between the origin cluster and the destination one, by the time difference between the instant where the agent leaves the stop, and the instant where he enters in the next stop. “This process gives one speed by stop, except for the last one of the moving chain. If the speed is greater than a given threshold, we consider that the mode type is “motorized” without precising at this step if it is “car” or “public transport”. A speed below the same threshold is consider as “walk” only if the distance between the two stop clusters are “reasonable”. That is to say below another threshold.

To determine the precise “motorized” mode, we associate to each agent an \textbf{average speed}, (i.e the sum of all speeds of its trajectory divided by the number of stops), and we use statistic information. We know, using a survey performed by CETUD [5], that in 2000 the number of car owners for 1000 Dakar habitants was 20. By which we can evaluate that in 2014 this number has been approximately increases to 30 cars for 1000 habitants, thus giving a percentage of 3%. The agents are then sorted in the decreasing order of their average speeds. We assign the mode “car” to the 3% faster motorized agents, and “pt” to the remaining motorized ones. All stops with no assigned mode are erased. If it happens that the trajectory list of an agent becomes empty afterwards, the agent himself is erased which leads to another reduction in the data.

3.5 Step 5: Trajectories clustering

The aim of this step is to select a significantly reduced sample of plans which are, as much as possible, representative of the whole trajectories in the SET2 files.

Thus, we try to group trajectories in clusters, each cluster being composed of plans “closed” to each others, according to a given distance measuring the similarity between two plans. Then, in each cluster, only one trajectory, representing all the others, will be chosen for the simulation. For instance, if two agents live in the same area and have the same sequence of activities in a similar cluster, we expect the two trajectories to be grouped in the same trajectory cluster and we only consider a single trajectory in the simulation. At the opposite, two different sequences of activities should be placed in different clusters and analyzed separately.

Following the observation made on files SET2, i.e. the agent detections are made every 10 mn, we associate to the trajectory of an agent \(j\), a vector \(v_j = (v^j, w^j)\) of dimension 288, where \(v^j\) and \(w^j\) are vectors of dimension 144 (i.e \(24\ h / 10\ mn\)).
Each component of $v^j$ and $w^j$ represents a detection instant, in a day period. For each $i$, $v^j_i$ is the cluster where the agent is located in the instant $i = 1, 2, ..., 144$, eventually “unknown” if no detection have been made. And $w^j_i$ is the type of activity made by the agent at the instant $i$, eventually “unknown”. $v$ and $w$ can be computed from the trajectory lists.

For two vectors $t^j$ and $t^k$ of two agents $j$ and $k$, we define the distance between them as follows:

$$d(t^j, t^k) = \sum_{i=1}^{144} \chi(v^j_i, v^k_i) + \sum_{i=1}^{144} \chi(w^j_i, w^k_i)$$

where, in general, $\chi(a, b) = 1$ if $a = b$ and 0 otherwise. In other words, this distance gives the sum of the cluster differences, and activity differences. It can be proved to be a metric in the mathematical sense. Using this metric, the same hierarchical clustering algorithm used for antennas clustering are performed for trajectories within a given threshold for the maximal distance between two plans.

After this step, we judiciously have to choose one trajectory in each cluster that will represent each class. We choose in each cluster the trajectory minimizing the total distance to the other trajectories of the same cluster. That is the so-called 1-median problem optimal solution (see Daskin [8]) computed in each cluster. Notice that in the hierarchical clustering algorithm, fixing a high threshold will have the effect to obtain large clusters, thus in turn to reduce significantly the number of plans to simulate, since only one plan is chosen by cluster. But when the threshold increases, the plans chosen become less representative of the whole set including those erased. In this case, numerical observations provide some relevant indications on the suitable value.

At the end of the step 3, we have a list of agent trajectories (one for each agent) supposed to be “representative” of the population. This list is transformed in a MATSim xml plan file and used for simulation with the previously generated network and facility files. At the end of the simulation, the optimization of amenities (re)locations starts (steps 6, 7, 8 of the figure 1). Below, we detail how this process works.

4 O-D flows, Amenities Selection and Local Search Algorithm

At the end a the MATSim simulation, each agent performed a plan in the transportation network, thus generating some flows between the amenities. We aggregate these flows to have a global view of the traffic. More precisely, between all couples of amenities, we compute the total number of trips during the simulation time. This gives an Origin-Destination flows matrix ($F$) between the amenities. We also compute the distances (in kilometers and in time) between each couple
of amenities giving us two matrices \((D\) and \(S)\). For each amenity, we store also the sum of the incoming and outgoing flows, giving us a view of the **traffic intensity** in each amenity.

Defining the global accessibility (see the introduction section 1) as the sum of the travel time, or travel distance, between couple of amenities and for all agents, the data generated above allow us to evaluate this global accessibility as follows:

\[
\sum_{i=1}^{n} \sum_{j=1}^{n} F_{ij} D_{ij}
\]

for the travel distance, or

\[
\sum_{i=1}^{n} \sum_{j=1}^{n} F_{ij} S_{ij}
\]

for the travel time, where \(n\) is the number of amenities.

The next step consists in selecting a set of amenities to relocate them in the best way. Two mechanisms are possible. Either we give (by hand) a list of amenities to study, or the code automatically computes one as follows.

The automatic amenities selection starts by sorting the amenities in the decreasing order of their traffic intensities. Then \(x\%\) (for a given \(x\)) of the amenities with the highest traffic intensity, and of a certain given types of activity, are chosen to be candidates for relocations or spatial switching. The goal here is to search how the locations of the selected amenities should be exchanged in order to reduce the global accessibility cost. The exchange of the position of two amenities can be mathematically formalized by a permutation \(\pi\) defined in the set of amenities. More precisely, \(\pi(i) = j\) means that the position of the amenity \(i\) is exchanged with the position of the amenity \(j\). We thus search for a permutation \(\pi\) involving only the selected amenities and minimizing one of the following value:

\[
(V_1) : \sum_{i=1}^{n} \sum_{j=1}^{n} F_{ij} D_{\pi(i)\pi(j)}
\]

or

\[
(V_2) : \sum_{i=1}^{n} \sum_{j=1}^{n} F_{ij} S_{\pi(i)\pi(j)}
\]

We add a constraint in the optimal permutation, that consists in accepting only the exchange of amenities of different types, i.e. exchanging two amenities of the same type do not impact at all the global accessibility.

Let us notice that in the current state of the code, moving an amenity in an non-occupied place, (as done for “Baux maraîchers”: see the introduction) is not
possible but will be included in the next version. Such an option can be viewed as an extension of this work. Indeed, non-occupied place can be represented by a set of possible available locations in which “fictive” activities can take place, with 0 incoming and outgoing flows.

The problem consists in minimizing \( V_1 \) (or \( V_2 \)) is a well-known problem in the combinatorial optimization litterature. It is called the Quadratic Assignment Problem [11]. We solve it using a standard local search procedure, also know as 2-opt neighborhood search (see [4] or [14] for a survey).

5 Preliminary Numerical Results

The aim of this section is to see how the methods we proposed behave in a complete round of SPOT simulations. We would like to know if the locations optimization may contribute to improve the overall utility or to decrease the global travel time.

It is important to state that these tests are preliminary. MATsim simulations have included a large set of scenario configurations, impacting the way that the utility function is computed, the way that each agent choose new locations for activity in the replanning process, the replanning strategy, the computation of the score of the plan which determines the utility values. For each of them, we have done some arbitrary choices that should be fit more correctly, considering real-life observations. We have also made many choices in the different steps of the plan construction. Some of them being open to criticism, considering the way the agents are detected by mobile phone antennas. Indeed, agents being detected when a call occur, the concept of “stop” does not rigorously correspond to a real-life stop, since we don’t know what the agent really does between two calls. Moreover, some activities (work for instance) may occur during travelling times. Using another detection technology, more accurate than a time granularity of 10 mn, may make more realistic the “stop” concept. Because of all of this drawbacks, the results reported here should be seen as an illustration of the “potential” of our method to contribute to urban planning process and planning. Further research will be necessary to make it more “operational”.

We launched two sets of simulations with an increasing number of agents in each, to assess the scalability. We ran the simulations using a DELL R510 server equipped with 125GB of memory and an Intel® Xeon® 64-bit processor with two cores of 2.67GHz each.

The first set has been performed by generating agent plans from the trajectory file SET2_P01.CSV, dealing with the first two weeks of January 2013. Using the methodology explained above, we generate 4693 agent plans and simulate the plans with MATSim with a scale factor of 100. “Scale factor” is a MATSim parameter by which each agent will represent, in our case, 100 others.
We fixed to 100 the number of MATSim iterations, and to 5 the number of iterations of the whole SPOT loops. At the end of the 5 iterations, some statistics have been performed to analyze the variation of utility value and travel time following each relocation. The pictures 3 and 4 show the mean utility values, and mean travel times for all agents, after each relocation iteration.

![Graph showing mean utility values and travel times](image)

**Fig. 3.** Utilities Mean through simulation iterations *SET2_P01.CSV*

The statistics are computed using a toolbox coded with the SPOT software. The utility means are obtained by computing the sum of the selected plan scores of all the agents at the end of MATSim iterations, divided by the number of agents. One can see that the utilities at the first iteration of the SPOT method are very low (even negative), showing that many agents perform long trips to realize their activities. Let us notice that the utility value is roughly the difference between the utility to perform the activities minus the disutility of the trip to reach these activities. Hence, longer the trip to perform few activities, lower the score. But whereas we propose some relocations, the average utility increases until it becomes positive. In the same time, the mean travel times decrease until a certain point where it increases. Notice that the utility function is more complex than the rough explanation above. We should intuitively expect that while utility increases, travel time decreases. However, some agents may realize more important activities, explaining this counter-intuitive variation in the last iterations.

Instead of tracing the mean travel times, it is also possible to plot the maximum travel times. In this case, the maximum travel times experienced by all agents are extracted after the MATSim iterations. We obtain the picture shown
in fig 5. We also observe a non-monotone variation, however we can see that maximum travel times experienced by the travelers tend to decrease with the relocations proposed.

To visualize the agent vehicle moves, the events of the last MATSim simulation, following the last relocation, have been displayed using the Senozon software. This movie is available in the dropbox link given in footnote. It can be observed, in this movie, that the plans generated initially with the SPOT methodology, and simulated in MATSim, are able to capture a simple fact observable in the Dakar region. The habitant trips from the popular east districts to the west, centre, and south areas where the majority of working, commercial and administrative activities are concentrated. And the trips in the opposite direction where probably the agents go back to home.

The second set of numerical tests concerned the trajectory file dealing with the first two week of March 2013. We generate in this case 6356 agent plans with the same scale factor of 100. Due to limited computational times, we ran 4 iterations of the complete SPOT method (instead of 5 as before).

The same variation as in the previous experiments can be observed in figure 6, showing that (at least for these two cases) the relocation contributes in increasing the utility mean. Moreover, at the opposed of the previous graph, the variation of the travel times mean (figure 7) here is monotonic, decreasing

---

13 https://www.dropbox.com/s/syxgkn9w7w69q12/SPOT-SENOZON.mov?dl=0
Fig. 5. Travel Times Max $SET2_P01.CSV$

Fig. 6. Utilities Mean for $SET2_P05.CSV$
iteration by iteration.

Fig. 7. Travel Times Mean $SET_{2.05}.CSV$

Thus these experiments give promising indications on the ability of the method to proposed representative plans, and on the capacity of the relocation algorithm to improve, globally, people moves. However, as point out in the beginning of this section, further investigations are evidently needed to validate the approach.

6 Conclusion and Perspectives

We presented in this paper a set of techniques used for generating agent plans from mobile phone data, and for automatically proposing suitable relocations of some amenities within a large set of spatial opportunities, by simulating urban trips. This work is based on a research developed in agent based systems and operation research by an inter-disciplinary team composed of computer scientists and geographers. It opens on hard scientific and operational issues including representative spatial resampling, suitable activity assignment in time geography, agent utility modelling and optimisation. It also opens, after validating the methodology with further improvements, some perspectives on technological developments.

Acknowledgement

This work has been possible with the suggestions, advices, code contributions, implications in other projects of: Rosa Figueiredo, Mohammed A. Ait Ouahmed,
Mouhamadou A.M.T. Baldé, Diaraf Seck. We gratefully thanks all. We also thanks Sonatel and Orange Group for providing us with the data.

References

Abstract

The rate of urbanization in developing countries, defined as the speed with which a population shifts from rural to urban areas, is among the highest in the world. The disproportionate number of citizens that live in a small number of cities places incredible pressure on the largest cities in these countries, which may already be faced with limited resources, weak industrialization, and underdeveloped infrastructures. Urban planning researchers as well as policy makers have suggested that governments in developing countries make capital investments within and surrounding smaller cities to attract citizens away from large urban centers, thereby lowering the pressure placed on overpopulated urban centers and making it more attractive for citizens to migrate to the smaller cities. This paper proposes a methodology that maps signals in mobile phone usage data to longstanding urban planning theories. These signals are subsequently combined in an unsupervised learner to discover regions within which city investments should be made. Qualitative evaluations of the selected arrondissements illustrate the promise of our approach.

1 Introduction and Motivation

A virtually universal trait across developing countries is the extraordinarily high rate of urbanization, which is defined as the migration of citizens from traditional, tribal, and rural regions to large city centers [14]. Ever increasing political turmoil in rural or tribal towns, ecological breakdowns, and the romantic (and often unrealistic) notion held by citizens that great opportunity exists in urban areas as compared to rural towns all contribute to such high urbanization rates [23]. However, urbanization is one of the most pressing challenges that faces developing nations. This is because urbanization leads to very poor living conditions as a city’s population exceeds its capacity with respect to infrastructure and available jobs. They also encourage an unstable bazaar economy that is impossible for the country’s government to tax or regulate, high rates of crime, and pollution. Urbanization is also causally related to the fact that developing countries exhibit a distorted and interdependent economy that produces products specifically for developed countries, and has large population growth and widespread poverty. The country of Senegal is no exception to the urbanization phenomenon; over 42.5% of the population lives in urban areas\(^1\) and over 71.9% of citizens living in the country’s 50 most popular cities reside in Dakar and Grand Dakar. To further demonstrate the intensity with which urbanization occurs in Senegal, Figure 1 shows how the majority of Senegal’s population is concentrated on its West coast, and the top quarter of cities with highest population density primarily lying in a region to the east of Dakar and Grand Dakar.\(^2\)

Urban planning researchers and policy makers concur that an effective way to reduce the negative effects of urbanization is to encourage a country’s citizens to migrate out of, rather than into, overpopulated urban centers by investing in the rapid development of promising towns and cities in alternative areas of the country [16]. Doing so simultaneously relieves the pressure applied to large central cities while investing dollars into the development of new cities that will add to the power of the country’s economy. The ideal town or city for rapid investment is one that already has an established local economy, has a developed infrastructure that supports its present inhabitants, and is self-sustaining; that is, it is located sufficiently far from existing large urban centers so that it does not rely on their economy, people, or ser-

\(^1\)http://www.indexmundi.com/senegal/urbanization.html
\(^2\)http://en.wikipedia.org/wiki/Template:Largest_cities_of_Senegal
services to thrive [1]. However, the socioeconomic data about towns and cities in developing countries that is required to measure economy, self-sustainability, and infrastructure development is understandably unreliable, dated, and difficult to collect [6]. This makes it all but impossible for researchers and government officials to identify promising locations for investment, and hence reduce urbanization in a developing country. Rather than relying on empirical data, we may alternatively rely on theoretical models of urban development as exhibited by developed and emerging countries, on theories developed by geographers and urban planners that explain how and why cities within a country essentially ‘self-organize’ into predictable patterns according to universally applicable geographic, economic, and social constraints. \textit{Central Place Theory (CPT)} is a long-standing, hotly debated, and recently more accepted theory explaining such self-organization of cities across a country [2]. It hypothesizes that some cities in a country are ‘Central Places’ that carry a very high population and produce a disproportionately large number of goods. Other types of communities, namely small ‘villages’ and middle-sized ‘towns’, naturally develop at different distances from central places depending on their reliance to its goods, people, and economy, with ‘middle towns’ being self-sustaining yet less developed compared to central places. The more recent \textit{Central Flow Theory (CFT)} postulates that cities develop in a cooperative manner by sharing information and interests using modern technology so that distance is not a constraining factor. Intriguingly, there is almost no work towards operationalizing these concepts to quantitatively assess the degree to which geographical areas follow this pattern. Such an operationalization would be immensely beneficial; identifying locations that central place or central flow theory identifies as a self-sustaining middle town would strongly suggest that, with appropriate investment, it could one day become a central place that relieves the urbanization effect of closely connected, overpopulated urban centers.

This paper proposes an innovative approach that uses mobile phone data to operationalize concepts from CPT and CFT to identify locations in Senegal where increased investment is most likely to (theoretically) reduce the migration of citizens to the large over-populated urban cities and instead make it more attractive to migrate to the newly emerging urban areas. Our approach is unique in its: (i) ability to identify promising locations for urban development without needing to rely on detailed socioeconomic data; and (ii) quantification of geographic and urban planning theories through the use of mobile phone data. Given the fact that mobile phone data is collected across many of developing countries already [3], our approach may be applicable for any nation facing intense urbanization.

The layout of this paper is as follows: Section 2 introduces Central Place and Central Flow Theory, concepts on which our methodology is based. Section 3 identifies fea-
2 Central Place and Central Flow Theory

Geographers have developed two spatial theories that attempt to explain how and why urban centers are distributed across a geographic space. This section describes these two theories in more detail, and through a preliminary analysis of regional data over Senegal, finds evidence that supports these theories within the country.

2.1 Central place theory

Central Place Theory (CPT) is a method to explain the tendency of villages, towns, and cities to self organize according to a cascading spatial hierarchy [7]. It proposes a spatial organization illustrated in Figure 2 where small villages and towns (low places) and secondary centers (middle places) lie in regions where larger urban centers (central places) carry their influence. The hierarchical structure is centered at an urban center or Central Place - a large population zone able to supply goods and services (low-level outputs) as well as knowledge and culture (high-level outputs) to its surrounding area. Thus, a necessary requirement for a Central Place to thrive is sufficient distance from other Central Places, so that neither offers a redundant and wasteful outputs that a nearby Central Place would already satisfy. Low Places, manifested as towns and and villages, live within the sphere of influence of a Central Place. Due to their strong reliance on the nearby Central Place for low- and high-level outputs, they may have low population, have an underdeveloped local economy, and carry a weak infrastructure. We define communities living on the periphery of a Central Place’s influence as a Middle Place. Middle places are by necessity partially self-reliant due to the larger geographic distance between them and the Central Place. They are able to produce some, but not all of the low-level outputs provided by Central Places and remain reliant on Central Places for high-level outputs. Being located at the periphery of regions of influence, Middle Places are by definition situated between a number of other Central Places and may exert a pressure on all of them simultaneously. Despite their less developed infrastructure, the ability for these self-sustaining Middle Places to agglomerate resources from a number of independent Central Places [28] places them in a unique position to integrate knowledge and resources that would otherwise be separate from each other [27].

While the hierarchical signature of CPT can be seen across many landscapes [8, 20, 24, 13, 32], there has been limited work towards operationalizing or modeling the phenomenon so that it may be applied to geographic datasets. These limited contexts include mathematical models based on CPT to predict city population growth [31], understand the hierarchical organization of cities over a geographic area [18], evaluate the way CPT interplays with economic growth over a spatial area [19], and to help explain geographical factors impacting phenomena such as sports tourism [9]. CPT has undergone a recent resurgence in popularity given its complementary relationship with modern urban economic geography theories, and is accepted as a reasonable model for explaining the spatial patterns of city development [29].

To evaluate the degree to which CPT is exhibited across Senegal, we scraped detailed population and location data across 6,135 cities, towns, and villages over Senegal from
version 2.2 of the Global Gazetteer \(^3\). As expected by CPT, major population centers are located far away from each other, as seen in Figure 4 that plots the locations of Dakar, Louga, and Thies, which are among the most populated cities in Senegal. These 3 cities, as expected according to CPT, are located far enough away from each other so that their population, economy, cultures, and support provided to their immediate regions do not interfere with each other. Figure 3 explores how the mean population and percent change in population among cities that lie within 5km bands radiating from the center of these three cities changes with distance. We identify a pattern where populations quickly drop near a Central Place, and then remain steady or slowly rise for cities ever farther away. Spikes may signal a Middle City that can sustain a larger population. To better identify population increases that may represent a Middle Place, the blue plots on the bottom row of Figure 3 compare the percent change in city populations as a function of distance. The drastic downward spikes seen within 20km from the Central Place, and again at approximately 60km and 170km as we move away from Louga, and at 60km and 80km as we move away from Thies, correspond to the big population declines between the other Central Places within 200km and the Low Places that immediately surround these Central Places. For example, Dakar is approximately 178km away from Louga and 77km from Thies by road, while Louga is 114km away from Thies. These fluctuations in population as a function of distance from a Central Place are a strong signal that CPT may explain the distribution of cities in Senegal.

According to CPT, Middle Places should have a high potential to evolve to become economic and cultural drivers for a country by developing into new Central Places. This is because Middle Places are already positioned in between the influences of existing Central Places, thus minimizing the disturbance of their evolution into a Central Place on the economies of neighboring cities. They are also already self-sustainable, with an infrastructure in place that supports a moderate population and production of goods and services. Finally, Middle Places have the ability, in the future, to create new low- and high-level outputs by agglomerating the outputs provided by nearby Central Places. We therefore hypothesize that such Middle Places are the most promising locations for economic and infrastructure investment in a developing country to mitigate the negative effects of increasing migration to existing large urban centers (Central Places).

2.2 Central flow theory

Central Flow Theory (CFT) is a recently proposed theory for explaining urban development that is complementary to CPT \([35]\). Whereas CPT is anchored around the spatial influence of Central Places, CFT describes non-local interactions among places without regard for physical distance. It also emphasizes the cooperative aspects of place interactions where information, ideas, specialists and other ‘foreign’ commerce are exchanged for mutual economic benefit rather than an organization of places into a dependency hierarchy. The complementary nature of the two theories have been seen in studies on the historical development of various urban places. Large Central Places interact with their geographic surroundings and nearby cities (CPT) \([17]\) to provide outputs that drive their economy, but their further development hinges on the free exchange of ideas and integration.
of ‘foreign’ commerce (CFT) [34, 33]. Agent-based simu-
lations further explain the interlocking relationship between
CPT and CFT for Central Place development [21]. We hypo-
thesize that places performing such exchanges occurring
at a low to moderate rate (compared to the level of exchanges
occurring among Central Places) signal a willingness to in-
tegrate foreign commerce, and already have the capacity to
share new ideas and information with places they may not be
dependent on according to the CPT hierarchy. These are all
desirable properties that would magnify the effects of eco-
nomic and infrastructure investments.

To evaluate the degree to which CFT holds across Sen-
gal, we use a (meta) dataset consisting of all mobile phone
calls in the time period between January and December
2013 [10]; the data is at the level of cell phone towers. Fig-
ure 5 plots the distribution of the total duration of all con-
versations made between the 1,666 towers in the country
over this time. The distribution exhibits a clear power-tailed
shape as seen in the distribution of calling activity across
many other mobile phone datasets [11, 30, 5]. We seek to
use this mobile phone communication data as a proxy for
the amount of information or ideas exchanged between indi-
vidual places. Towards this end, we only consider communi-
cation between towers whose cumulative duration of all
conversations fall in the top 1.5% of this distribution, which
translates to an average of 2,739 minutes of conversations
per day. This filtering step leaves 38,613 flows of communica-
tion that fall in the tail of the distribution in Figure 5,
where statistically significant calling activity occurs.

We subsequently form an undirected graph where nodes
represent towers and edges correspond to the flows of activ-
ity as described above. To evaluate the popularity of a call-
ing tower (e.g. the extent to which information and ideas
are exchanged within places nearest to it) we consider the
PageRank centrality of towers in this graph. PageRank con-
siders the popularity $p_i$ of calling tower $i$ to be propor-
tionate to the popularity of the towers it communicates with. It is
defined by:

$$p_i = \alpha \sum_j A_{ij} \frac{p_j}{g_j} + (1 - \alpha) \frac{1}{N}$$

where $A$ is a matrix with $A_{i,j}$ given as the cumulative length
of all conversations between towers $i$ and $j$, $g_j$ is the de-
gree of node $j$, $g_j = \max(1, k_j)$, $N$ is the number of towers,
and $\alpha = 0.87$ is a damping parameter set according to
the recommendations based on earlier work [4]. In Fig-
ure 6, we compare the location of the 10 most populated
cities in Senegal in the Global Gazetteer against the loca-
tion and PageRank centrality of calling towers (larger ver-
tices correspond to higher PageRank). We identify a strong
correlation between the position of the most popular cities
(Central Places) and the location of call towers that exhibit
the largest amount of activity, as predicted by CFT. We also
observe, even though the distribution of PageRank centrali-
ties is skewed, many call towers with high PageRank lying

![Figure 5: Distribution of total calling times across towers](image)

between these most populated cities as seen in Figure 6(a).
Although these locations may not have a high population,
large PageRank centrality suggests that places around these
towers are undergoing significant exchange of ideas and in-
formation with Central Places. According to CFT, such ex-
changes are positive indicators for these places becoming
Central Places in the future.

3 Identifying Middle Places for Development

Recognizing the fact that CPT and CFT may help explain
urban development in Senegal thus far, we consider unsu-
ervised methods for identifying areas in the country most
likely to correspond to Middle Places. We intentionally de-
cided not to focus on supervised methods for this problem
as there is virtually no ground truth data available for what
is considered to be the ‘best’ place for urban investment.
Instead, our unsupervised approach considers a number of
features from a dataset of mobile phone calls that are the-
oretically linked to CPT and CFT, and combines these in a
methodology that classifies arrondissements by the types of
places (Central, Middle, or Low) they support. We chose
to classify arrondissements rather than individual towns be-
cause: (i) high-resolution data expressing the calls made be-
tween villages, towns, and cities are unavailable; (ii) govern-
ment investments in urban development can likely be more
easily be budgeted for an administrative unit, rather than for
a specific city; and (iii) arrondissements that support Mid-
dle Places may be prime areas for infrastructure investment,
and for making modern investments such as development of
planned communities or technology parks. In this section,
we present the features we consider for modeling and the
classification methodology.

3.1 Features considered

We consider four different features of a dataset consisting
of mobile phone calls made between call towers in each
arrondissement of the country. We chose features that, ac-
cording to CPT and CFT, should take on an extreme value
if an arrondissement supports the development of Middle
Places over Low or Central Places. These features include:
• **Total call volume**: This is defined as the total number of calls placed by mobile users in an arrondissement.

• **Distance of Calls**: This feature is defined as the $X^{th}$ percentile of the distribution of the geographic distance calls placed by cities in an arrondissement travel. This feature provides consideration for the geographic component of CPT, where Middle Places tend to find themselves far away from the Central Places they contact for information and knowledge. The best value of $X$ is found during model selection.

• **Demand-weighted distance of calls**: This is defined as the sum of call durations weighted by the physical distance that each call traveled in kilometers.

• **Self-Sufficiency**: This is defined as the percentage of an arrondissement’s calls that occur between mobile cell towers within the same arrondissement. This percentage reflects the “locality” of calls made within an administrative region; areas with strong internal communication suggest a weaker reliance on the information provided by people located in other arrondissements.

• **Partnership**: This is defined by counting the number of unique arrondissements that comprise the top 80% most active connections (in terms of call volume) from an arrondissement. Noting that Central Places combine information from a number of other places in order to create new products and knowledge, it should be the case that the most active communications from an arrondissement supporting Middle Places connect to as many external locations as possible.

• **Centrality**: We represent all calls between arrondissements as a graph, with an edge feature as the total number of calls between two arrondissements. We then consider the eigenvector, PageRank, and betweenness centrality. The betweenness centrality of arrondissements in this graph measures the number of shortest paths that pass through the arrondissement being measured. Betweenness centrality thus reflects the ability of cities in an arrondissement to connect to other locations in Senegal, thus acting as a broker of information and resources, and as a place where ideas and knowledge meet. Eigenvector and PageRank metrics score an arrondissement on the graph based on the scores of other arrondissements it is strongly connected to; thus Middle Places may take on high values due to their (theoretically) strong connections to many Central Places.

### 3.2 Methodology

We classified arrondissements by the degree to which they support Middle Places by clustering over a vector that represents an arrondissement and whose components are defined as the value of the features presented above. $K$-means clustering is a standard algorithm for clustering such vectors, however it is very sensitive to initialization and the distance measure used. Instead, we work with Finite Mixture Models (FMM) that search for a best fitting mixture of probabilistic data distributions that explain the total distribution of values exhibited in the entire dataset. FMM relaxes many of the constraints imposed by $k$-means clustering and is less sensitive to the scale and range of values of the features [15]. Relaxation of these assumptions is suitable to the research objective of operationalizing CPT/CFT because a larger proportion of places should be characterized into a Low Place cluster, followed by Middle Places, and finally Central Places. Cluster sizes should also follow this pattern. We used the mclust Finite Mixture Modeling software package in R to search for clustering solutions where the mixed models were part of the exponential family. The package reports results from many combinations of hyperparameter settings that encode assumptions about the types.
### Table 1: Correlation between distance of calls and self-sufficiency features

<table>
<thead>
<tr>
<th>Distance Traveled by X% of Calls</th>
<th>Correlation with Self-Sufficiency</th>
<th>Variance of Distance Traveled</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% (median)</td>
<td>-0.58</td>
<td>454</td>
</tr>
<tr>
<td>60%</td>
<td>-0.37</td>
<td>854</td>
</tr>
<tr>
<td>70%</td>
<td>-0.17</td>
<td>3,581</td>
</tr>
<tr>
<td>80%</td>
<td>0.10</td>
<td>10,872</td>
</tr>
</tbody>
</table>

Table 2: Clustering solutions with different variable settings

<table>
<thead>
<tr>
<th>Solution</th>
<th>Variables</th>
<th>BIC</th>
<th>Pseudo-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>Self-Sufficiency, Partnership, Betweenness Distance ($X = 60%$)</td>
<td>-1.288</td>
<td>41.6</td>
</tr>
<tr>
<td>Alt. A</td>
<td>Self-Sufficiency, Partnership, Betweenness, Distance ($X = 50%$)</td>
<td>-1.297</td>
<td>46.2</td>
</tr>
<tr>
<td>Alt. B</td>
<td>Self-Sufficiency, Partnership, Betweenness, Demand-Weighted Dist.</td>
<td>-1.342</td>
<td>31.0</td>
</tr>
<tr>
<td>Alt. C</td>
<td>Self-Sufficiency, Partnership, PageRank Centrality, Demand-Weighted Dist.</td>
<td>-1.316</td>
<td>38.4</td>
</tr>
</tbody>
</table>

3.3 Model selection

Model selection criteria in unsupervised learning has an inherent level of subjectivity due to the latency of the dependent variable, and no observable outcome exists to compare model validity against [26]. We adopt the following process to evaluate a potential solution that classifies arrondissements by the degree to which they support Middle Places in terms of the following criteria, in order of priority:

1. **Multicollinearity**: prior to introducing independent variables into a clustering model, high correlations between variables inhibit variable selection. Correlations of $>0.5$ are considered high, and correlations between 0.3 and 0.5 are monitored as we evaluate the solution using criteria (2) through (4). If two variables are highly correlated then these are not introduced into the models because they overstate the impact of their phenomena on the solution.

2. **Actionability**: In this criteria, we ask if cluster variables and boundary values allow for a governing body to take action on the results. For instance, if Middle Places, as defined by the CPT and CFT features, fall entirely within Grand Dakar or if they comprise a large proportion of Senegal’s cities, the ability for an organization to take action on the results is limited. This is a logical and subjective, yet necessary, criterion.

3. **Bayesian Information Criteria (BIC)**: Finite Mixture Modeling, the primary clustering technique used in our work, utilizes BIC as the key statistic for comparing solutions [12]. It is defined as:

   \[ B = 2 \log P(X|M, \Theta) - d \log n \]

   where $X$ is the set of observed data vectors, $M$ is the fitted clustering model with maximum likelihood parameters $\Theta$, $d = |\Theta|$, and $n = |X|$. Models with larger $B$ tend to be better models, since if the data $X$ fits the model $M(\Theta)$ well, its log-likelihood should be higher.

4. **Pseudo-F Statistic**: The Pseudo-F statistic is a measure of the efficiency of a clustering result. It is defined as the ratio of the mean sum of squares distance between vectors in different clusters to the mean sum of squares distance between vectors in the same cluster [22]. Larger Pseudo-F scores correspond to ‘tighter’ clusterings where intra-cluster distances between vectors is small and inter-cluster distances are high.

In our analysis we found that total call volume, demand-weighted distance of calls, weighted average distance of calls, Eigenvector centrality, and PageRank centrality were heavily skewed to a very small number of well developed cities including Dakar. This skewness reduces the actionability of results; they would consistently suggest that Dakar and other well developed cities should be further developed, but it is difficult to channel resources into these complex cities due to their skewed distribution.

Figure 8: Correlations between transformed and standardized features. PageRank shows high ($|r| > 0.5$) correlations with call distance and self-sufficiency, while betweenness is only moderately correlated with Self-Sufficiency.
urban spaces. Some of these features also caused multicollinearity issues; for example Figure 8 identifies how PageRank centrality exhibits high correlation with call distance and self-sufficiency.

Table 2 enumerates through FMM solutions using features such as self-sufficiency, partnership, betweenness, call distance, PageRank, and demand-weighted distance. We found that the best solution given in the first row identifies 4 clusters using the self-sufficiency, partnership, betweenness centrality, and distance at the $X = 60^{th}$ percentile. Besides exhibiting the highest BIC and nearly highest Psuedo-F, we found that setting the call distance feature using the $60^{th}$ percentile of the distribution minimized the correlation between this feature and self-sufficiency. As seen in Table 1, the $60^{th}$ percentile is an approximate elbow point that reduces correlation while maintaining a small amount of variance that does not heavily skew this feature value to Central Places that almost the entire country contacts (e.g. Dakar).

Figure 7 uses a Self-Organizing Map to visualize the distribution of the features used in the best clustering solution across the arrondissements of Senegal. The colors of the nodes in each map represent the scaled values of the features from low (cool colors) to high (hot colors). The number of nodes of some color is proportional to the number of arrondissements whose value is in the range represented by the color [25, 36]. Note that each map is initialized with a random assignment of arrondissements to nodes. The maps identify how the distance of calls, partnership, and self-sufficiency metric exhibit a small skew towards a small number of arrondissements (those that host Middle Places) while betweenness centrality is better distributed. The more even distribution of betweenness centrality is likely due to the fact that both Middle and Central Places have important brokerage locations for information and communication across the country, hence both types of Places may be represented by the hotter nodes. The large number of cool betweenness centrality and partnership nodes capture the Low Places that do not serve as brokers of any kind of information nor do they communicate with a large number of external places.

Figure 7: Feature values for best clustering solution

![Figure 7: Feature values for best clustering solution](image)

4 Cluster results and discussion

Figure 9 uses a dot plot to present the centroid positions of the four clusters in the best FMM solution. We subjectively map these values to being relatively LOW, MODERATE, or HIGH for each cluster in Table 3. We label each of the four clusters as:

- **Cluster 1: Dakar and its Suburbs.** Cluster 1 identifies 8 arrondissements that, as visualized in Figure 10(a), contain Dakar and its suburbs. These arrondissements show high betweenness, meaning they are hubs for calls throughout the country. Yet, their low call distance and partnership implies exclusivity; information flow passes primarily through partners within the same cluster. This cluster is grouped by both geography and numerical values of the features, supporting the theoretical definition of a Central Place.

- **Cluster 2: Middle Places.** The nine arrondissements placed in Cluster 2 quantitatively support the definition...
<table>
<thead>
<tr>
<th>Cluster Description</th>
<th>Self-Sufficiency</th>
<th>Partnership</th>
<th>Betweenness</th>
<th>Distance</th>
<th>Cluster Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Places: Dakar and Suburbs</td>
<td>MODERATE</td>
<td>LOW</td>
<td>HIGH</td>
<td>LOW</td>
<td>8</td>
</tr>
<tr>
<td>Middle Places: Emerging Opportunities</td>
<td>HIGH</td>
<td>MODERATE</td>
<td>HIGH</td>
<td>LOW</td>
<td>9</td>
</tr>
<tr>
<td>Low Places: Villages Between</td>
<td>LOW</td>
<td>MODERATE</td>
<td>LOW</td>
<td>MODERATE</td>
<td>37</td>
</tr>
<tr>
<td>Middle-Low Places: Common Towns</td>
<td>MODERATE</td>
<td>MODERATE</td>
<td>MODERATE</td>
<td>MODERATE</td>
<td>69</td>
</tr>
</tbody>
</table>

Table 3: Cluster labels and features

For Middle Places. They have high self-sufficiency, and when calls do leave these areas, they are reaching a large number of other arrondissements. The low distance that calls travel may be due to short connections with other proximate cities, which CPT supports.

- **Cluster 3: Low Places.** The 27 arrondissements in this cluster exhibit a low degree of self-sufficiency and betweenness, and a moderate level of partnership and call distance. Low self-sufficiency is an indicator of a Low Place which needs to strongly rely on nearby other places for resources and information. Similarly, a low betweenness value indicates that the location is not a broker of information, and that these arrondissements are not of interest to most other arrondissements. In Figure 10 (a), the Low Places (blue positions) tend to be surrounded by a number of other nearby arrondissements, further supporting the notion that they rely on nearby Central Places, Middle Places or Low-Middle Places (Cluster 4).

- **Cluster 4: Low-Middle Place.** Finally, the majority of arrondissements fall into a cluster with moderate self-sufficiency and betweenness as well as partnership and distance, suggesting that they support a mixture of Low and Middle places. The positions of such arrondissements in Figure 10 (a) find them to be near Dakar and its suburbs, (b) by the border of the country, (c) in remote regions, or (d) immediately surrounded by arrondissements that only support Low Places.

Because arrondissements in Cluster 2 support Middle Places much more strongly as compared to those in Cluster 4, we further investigate the cities seen in Cluster 2 arrondissements to validate that they exhibit features that make them promising opportunities for urban development:

(i) **Thies.** Thies is one of Senegal’s largest cities and sits in an area considered to be a transportation hub that services routes between St Louis, Dakar and Bamako. It is also a major producer of peanuts and fertilizer that are among the country’s top exports, and host reserves of important metals. It thus has the potential to become an even stronger economic hub for the city under further investment.

(ii) **St Louis.** St Louis is the capital of Senegal’s St Louis arrondissement and is located in the northwest of the country near the mouth of the Senegal river on the Mauritanian border. It was a capital of Mauritania which at the time was a neighboring colony. It has a heavy tourism based economy, has a high rate of sugar production, fishing irrigated alluvial agriculture, pastoral farming, trading and exportation of peanut skins. The city was listed as a UNESCO World Heritage Site in 2000 and cultural tourism has become an engine

---

4 [http://www.aljazeera.com/indepth/features/2012/02/201222695110410730.html](http://www.aljazeera.com/indepth/features/2012/02/201222695110410730.html)

of growth\textsuperscript{6}. (iii) Mbour. Mbour is a city in the Thies Region of Senegal. It lies on the Petite Cote 80km south of Dakar. The city’s major industries are tourism, fishing and peanut processing. It is Senegal’s fifth largest city and, by some indicators, is among one of the fastest growing\textsuperscript{7}. (iv) Ziguinchor. Ziguinchor is a river-port town in southwestern Senegal lying along the Casamance River. It is one of the largest cities in Senegal, but is largely separated from the north of the country by The Gambia\textsuperscript{8}. Ziguinchor remains economically dependent on its role as a cargo port, transport hub and ferry terminal. A primary highway crosses the Casamance River just east of the city, linking the region with Bignona about 25km to the north, and (via The Gambia), the rest of Senegal. It features a large peanut oil factory and is also known for producing great quantities of rice, oranges, mangoes, bananas, cashews, tropical fruits and vegetables, fish, and prawns. Ziguinchor is also home to a new University which opened in 2007\textsuperscript{9}.

5 Conclusions and Future Work

In this paper we introduced a data driven methodology to identify the most promising areas in Senegal for economic investment. We identified features, using mobile phone data, that speak to Central Place and Central Flow Theory, which are important geographic and urban planning theories that explain the way cities in a country naturally develop. To the best of our knowledge, this paper is the first attempt made to operationalize these theories for forecasting the places in a country where investments should be made, and to quantify CPT/CFT concepts in a dataset of mobile phone records. Future work will examine alternative clustering methods and distance metrics that define similarity, formulate other data features that are related to CPT and CFT, and reformulate our idea as an optimization problem that ranks arrondissements in order of the ‘best’ places in Senegal for investment.

References

\begin{enumerate}
\end{enumerate}


Understanding Traffic Matrix for Transportation Planning with Interregional Connectivity

ABSTRACT
As modern information society with Internet and mobile devices create much datasets, Cellular phone records had made it available to perform large scale of studies on social problem and human mobility. In this paper, using CDRs in Senegal, we evaluate the effect of constructing “Dakar Diamniadio Toll Highway” by showing connectivity between regions in nearby the road and user movement patterns. Obviously, changes of traffic volume and time exist after constructing new highway among 10 arrondissements in Darkar. Users are also affected by the new highway in detail analysis based on weekend and weekday. The results could be used for reevaluation of the new highway for supplement. Other developing countries refer to the result with considering country-specific characteristics of Senegal while planning for new highway in the future.

Categories and Subject Descriptors
[Human-centered computing]: Ubiquitous and mobile computing – Empirical studies in ubiquitous and mobile computing.

General Terms
Experimentation, Human Factors, Measurement

Keywords
Data for development, Human mobility

1. INTRODUCTION
As modern information society with Internet and mobile devices create much datasets. Analyzing large datasets including potentially spatial and temporal information about individuals gives us opportunity to understand society. Moreover, cellular phone records had made it available to perform large scale of studies on social problem and human mobility.

By utilizing call records from mobile phone, several research issues in social and economic aspects are introduced as follows: In urban planning or transportation planning, how can we estimate the effect of changing road network? What people displacement patterns can be measured with the highway traffic or trajectory? How do people travel from a city to another, or commute from a region to another in a city?

Our focus in the paper is accessibility or connectivity between regions including urban and suburban areas. Urban cities generally provide jobs to suburb. Moreover, accessibility is capitalized into land values and also affects population distribution. When the road network is changed, the accessibility is also changed between regions and affects various aspects from citizens’ life style to social and political event.

Dakar is known to central area in social and economic perspective in Senegal. While infrastructure of Dakar is well developed, suburban area is not developed and has low accessibility to Dakar. To solve above problems, governments of Senegal launched DDTH (Dakar Diamniadio Toll Highway) in June 2009, which constructs a highway connecting the city center of Dakar to the Diamniadio suburb (see Figure 1). The effect of constructing the highway is known to 1) improvement of mobility within Darkar urban area, 2) decrease of Dakar’s congestions from about 90 to 30 minutes, and 3) higher accessibility to social and economic services of Dakar in suburban area.

However, government needs expensive cost and time to evaluate the effect of DDTH because it needs statistical survey. Traditional ways are to depend on census data, surveys, vehicle counting and so forth. These methods are time consuming and their expenses are high in order for collecting data. In addition, small population samples only are available.

To evaluate effect of constructing new highway using data mining with big data, we leverage datasets of CDRs (anonymized Call Detail Records) from January to December in 2013, which are provided by Data for development (D4D) challenge from Sonatel and Orange group. The data were extracted from the mobile network in Senegal. Because the data is spatially coarse, users’ trajectories can be dealt with by analyzing the data. Because the road connecting between Pikine and Diamniadio opens August in 2013, we focus the effect of the road.

Previous status report of DDTH did not show changes of connectivity of regions in Dakar and user movement pattern in detail level. Because the highway highly affects not only life style (e.g., work place or home) but also economic aspects, for instance, unemployment rate, land values, between regions in nearby the highway, evaluating interregional connectivity with moving pattern is important. Moreover, well-developed countries already built highway and the effect of highway is not clearly specified with big data. In developing countries, predicting effect of new highway to be built is important for maximizing their benefits.
In this paper, we show changes of interregional connectivity and movement patterns before and after constructing the new highway. The result also shows potential to be used for planning new next new highway in political level.

2. RELATED WORKS
Current research appears to validate the view that mobile phones and economic development in Africa are available enough to cover trajectories of many individuals. Further research in this area may include regional gap or differential discovery. In detail, mutual interaction or exchange in country should be considered when the policy makers determine the policy, and city connectivity is also critical for improving inside-region connectivity. For example, economic development includes not just projection and consumption but also the exchange activity to allow the city’s growth. For these reasons, we perform the experiment on the relationship between interregional connectivity and traffic on CDR data.

3. ANALYSIS OF CALL DATA
Before stating full-scale analysis, we conduct several fundamental experiments on traffic. The displacement of 14 individual patterns in Dakar with SET2 [Figure 2]. Density of CDR data has availability to understand population density with most frequent visits in Figure 3. Furthermore, as for preparing the experimentation of looking into datasets, we have preprocessing phase to deal with better understanding traffics throughout regions. With SET2 regarding fine-grained mobility at site level, we did not count the cellular site if the user’s site is the same as next to before site so that the traffic matrix is built with counts of each site. The number of all cell sites nearby cell tower is 1666, and traffic matrix size is 1666 by 1666 as well. By creating traffic matrix with 25 sets from every two-week, we filtered out other users outside of Dakar, if a user does not include any cell tower site in Dakar.

3.1 Interregional Connectivity
In suburban regions, accessibility is one of critical factors to get a life of ease. Issues related to geographic location among regions can be called interregional connectivity. The importance of regional connectivity is no doubt to be emphasized at country level in that mutual interaction is the power of information as well as flow of goods. In other words, interregional connectivity ensures the demand for increasing distribution system and trade economy.

Table 1. Dakar versus Suburban regions

<table>
<thead>
<tr>
<th>ID</th>
<th>Department</th>
<th>ARR</th>
<th>Division</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DAKAR</td>
<td>PARCELLES ASSAINIES</td>
<td>Left side of the new highway</td>
</tr>
<tr>
<td>2</td>
<td>DAKAR</td>
<td>ALMADIES</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>DAKAR</td>
<td>GRAND DAKAR</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>DAKAR</td>
<td>DAKAR PLATEAU</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GUEDIAWAYE</td>
<td>GUEDIAWAYE</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>PIKINE</td>
<td>PIKINE DAGOUDANE</td>
<td>Right side of the new highway</td>
</tr>
<tr>
<td>7</td>
<td>PIKINE</td>
<td>THIAROYE</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>PIKINE</td>
<td>NIAVES</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>RUFISQUE</td>
<td>RUFISQUE</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>RUFISQUE</td>
<td>BAMBILOR</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. User displacement patterns (14 users)

Figure 3. Density graph of traffic (Whole users of SET2 P01)

Figure 4. Inter-regional traffic changes. The red arrows show the traffic increases, and blue ones show traffic decreases. The orange line is the new highway constructed on August 1st. The tendency of increase and decrease are along the road.
In this case of new road network, it may identifying the interregional trajectories are along highway where ~~~ We concentrate on different geographic footprints in Dakar with arrondissement level in that regional difference is pretty big with respect to concentration resources.

The objective of cell towers in city is to aggregate individual specific trajectory into clustered region. From the regions left-side of highway to the regions on the right-side of highway Figure 1. Division of two area on the left and right side of highway. We delete the arrondissement ID 5 in that the area is not passing through the highway. Left side of the new highway is 1,2,3,4 arrondissement (PARCELLES ASSAINIES, ALMADIES, GRAND DAKAR, DAKAR PLATEAU), the right side is 6, 7, 8, 9, 10 arrondissement (GUEDIAWAYE, PIKINE DAGOUDANE, THIAROYE, NIAYES, RUFISQUE, BAMBILOR)

### 3.1.1 Comings and goings among arrondissements

By targeting the two selected parts, we use the cellular antenna across the whole Dakar and selected parts of arrondissements. Coming from and going to Dakar is significantly inclined as seen red arrows in Figure 4, which shows interregional exchange traffic in Dakar with weight (Traffic amount). Enhance the exchange and communication each other. Even the information to other cities are limited in some ways. The thing to note is the tendency of increases is along the road. A closer look at the data indicates that the highway plays a role as a communication facility between two parts. Connectivity between arrondissement 8 and 10 are much stronger relation on traffic.

We confirm the differences are focused on the left side of the new highway in Figure 7, which means that a lot of people would like to change their displacement patterns to near Dakar. Interregional traffic in Dakar are definitely affected by this highway project in that it raises the accessibility to central Dakar. In Figure 6, most traffics on each region are focused on several close region. In addition, a sharp increase in traffic is after set 15. Upper plots are similar each other, and vice versa on the bottom in Figure 5. This represents that more high traffic occurs in nearby region, but sometimes it can be changed by road network, enhancing connectivity or familiarity. In this sense, this study create the value to understand regional trade and communication levels.

#### 3.1.2 Travel time decrease

People tend to commute their company or school which are located in central Dakar. After utilizing the set of cellular cities, we also estimate the speed of interconnecting regions using the centers of their cell towers in the arrondissement. We confirm time decrease before and after. Since highway construction is starting from the middle of Dakar department, we need to divide the department and regions into two area.
3.2 User Pattern Analysis

3.2.1 Important places

We look into collection of datasets first with visualizing each user’s trajectory according to timestamp. By using spatial-temporal data by pattern mining, it may be needed to find the interesting locations e.g., home, workplace, grocery store, gym. That is because finding those places can enhance understanding of general human movement pattern, and improve modeling of human mobility. In the figure 1, cell towers are depicted as hexagons, and the hexagons with red color represents the weight of importance. Pattern comparison between weekdays and weekend

Identifying whether or not the change of traffic network has some potentials to affect people's mobility or their trajectory patterns is worthy to analyze. To be specific, policy maker in transportation planning is interested in how the highway influences on the country's economy, trade, tourism and so forth. Our basic idea is to compare these traffic matrix on 25 sets for every two-week, the cosine similarity shows very low value on weekend, comparing to other weekdays. That is one of reasons why we divide the time into two (weekdays, weekend).

3.2.1.1 Weekdays Only, Commute time only

In the local context, most people who have their own car take the highway to commute Dakar for their job. Generally people think that it is better to take the highway for going faster towards Dakar. Even they commute from faraway area such as Pikine or Thiaroye. To read timestamps in an efficient way, we separate timestamp for three hours. We assume that about 60 percent of people commute from other regions to central Dakar comparison between two cases in table 3. Usual commuting time in Dakar is at 8 AM and 5 PM. Even though we recognize the most frequent places, and they are the favorite haunt for users, we cannot say that it is house in that some people spend much time in workplace. Assumptions are

![Figure 9. Inter-regional traffic changes to other regions with time series.](image)

<table>
<thead>
<tr>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>The most frequent cell site</td>
<td>House</td>
</tr>
<tr>
<td>The second frequent cell site</td>
<td>Workplace (Dakar)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Timestamp and expected user pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timestamp</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. Cases of assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
</tr>
<tr>
<td>The most frequent cell site</td>
</tr>
<tr>
<td>The second frequent cell site</td>
</tr>
</tbody>
</table>
experimented as follows. They show the quite big time lag with at most 20 minute.

3.2.1.2 Weekend Only, Active time only
Active time is 12 PM to 15 PM. On weekend, trajectories of individuals may be likely to go other cities or other regions.

3.3 Monthly User Rate
In order for better understanding users’ geographical footprints, we show the rate per month with different features. At first, comparison between entropy of places and number of places is conducted. With SET3 indicator among 146,352 users, the entropy of places has similar inclination to growth in Figure 1. They have similar patterns and increase in number. Total distance of traveling from individuals are calculated in Figure 12.

We expect that the total distance from individuals increases in a way that people may want to access or visit farway regions by taking highway. Although the new highway has toll, the fast accessibility is advantageous. We discover the ratio to in that every set has different user ID. In other words, following users’ trajectories are not available. The importance of regional connectivity is demanding for increasing distribution system and trade economy.

4. CONCLUSION
Using CDRs in Senegal, we evaluate the effect of constructing new highway by showing connectivity between regions in nearby the road and user movement patterns. Obviously, changes of traffic volume and time exist after constructing new highway among 10 arrondissements. Users are also affected by the new highway in detail analysis based on weekend and weekday. The results could be used for reevaluation of the new highway for supplement. Other developing countries refer to the result with considering country-specific characteristics of Senegal while planning for new highway in the future.

5. REFERENCES
Figure 13. Total distance of users in Dakar for every two-week

(x: Time, y: Total distance of users)


High Resolution Traffic Maps Generation Using Cellular
Big Data

El-Mahdy, Ahmed, Dr, E-JUST, Egypt
Algizawy, Essam, Eng., E-JUST, Egypt
Ogawa, Tetsuji, Dr, E-JUST, Egypt
Shishiny, Hisham, Dr, IBM, Egypt
Baddar, Mohamed, Eng., IBM, Egypt
Kimura, Keiji, Prof., Waseda University, Japan

December 31, 2014
Abstract

We consider, for the first time, utilising the mobile big data for microscopic level traffic analysis. The project develops a HMM formulation, and uses Viterbi decoding to discover actual road segments of trips. This facilitates road-level analysis without the need for high cost traffic on-road traditional sensors. We then generate Dakar traffic intensity maps for main roads, for every hour in the covered 50 weeks, of the year 2013. Moreover, we develop and apply a traffic prediction model to the data and identify significant traffic seasonality patterns.
Contents

1 Introduction 5
  1.1 Motivation ............................................ 5
  1.2 Problem Statement .................................... 5
  1.3 Objectives ........................................... 6
  1.4 Relevant Previous Work ............................... 6
  1.5 Report Organisation .................................. 9

2 CDR and Map Generation 10
  2.1 CDR from D4D Orange Challenge ....................... 10
  2.2 Map generation ...................................... 10
  2.3 Zone representation .................................. 11
  2.4 Exit point extraction ................................ 11

3 Traffic Monitoring Using Hidden Markov Model 13
  3.1 Overview ............................................. 13
  3.2 Trip Detection ....................................... 15
  3.3 Road And Traffic Modeling Based on HMM ............. 16
    3.3.1 Model Representation ............................ 16
    3.3.2 Parameter Initialization .......................... 16
  3.4 Trip Trajectory Estimation ........................... 17
    3.4.1 Forced Alignment Using Viterbi Algorithm ...... 17
    3.4.2 Accurate Road-Level Route Estimation .......... 18
  3.5 Statistics of Trip Trajectories And Traffic Intensity Map Generation .......... 18
  3.6 Adaptive model estimation ........................... 18

4 Optimising the Viterbi Decoder 23
  4.1 Viterbi Algorithm .................................... 23
    4.1.1 Eliminating Multiplication by Zero Computations .. 24
    4.1.2 Parallelisation of the Decoding Process .......... 25
    4.1.3 Performance Analysis ............................ 25

5 Validation and Analysis 29
List of Figures

2.1 Voronoi partition for 300 BTSs .................................................. 12

3.1 Conceptual image of (1) road network and (2) HMM structure applied. Blocks $e_i$ in upper figure express exit points, which is defined as states in HMM. ... 14

3.2 Diagram of adaptive traffic monitoring system. .............................. 15

3.3 Comparing between various exit point choices. (1) road network with four zones; (2) junction exit points route; (3) edge exit points route; and (4) combined junction and edge exit points route. ......................... 19

3.4 Examples of road-level, fine-grained trip trajectory. ....................... 20

3.5 Example of traffic intensity map for one hour. ............................... 22

5.1 Hourly distribution of flows for each edge. ................................... 30

5.2 Correlation between flows of each origin/destination pair and corresponding gravity model. .......................................................... 31

6.1 Sample TS plot for a road edge (the day of the peak traffic in 2013 is 2013-8-9, which is the first day of “Eid el Fetr” for muslims). .......................... 33

6.2 Monthly aggregated traffic data .................................................. 34

6.3 TS hourly seasonal variation as depicted by the Seasonal Adjustment Factor (SAF) ......................................................... 34

6.4 TS daily seasonal variation as depicted by the Seasonal Adjustment Factor (SAF) -Note that 7/1/2013 is Monday, and 13/1 is Sunday ................. 35

6.5 A sample TS model for modeling and forecasting daily traffic flow on one of the road edges ...................................................... 36

6.6 MAPE frequency distribution for the generated TS models ............... 36
List of Tables

1.1 Urban road and traffic network research. . . . . . . . . . . . . . . . . . . . . . 8
Chapter 1

Introduction

1.1 Motivation

Traffic monitoring can provide accurate information to both public users and decision makers. That can provide for better commute time and planning for new roads, thereby decreasing current road congestions; it also includes traffic prediction for the next coming hours and days.

Such analysis is currently cost prohibitive requiring installing traffic sensors such as inductor loops and video camera all over the cities. Recently, location information obtained from mobile sensors and crowd sourcing have been shown to be effective for that purpose. We consider, for the first time, utilising the mobile big data for ‘microscopic’ level traffic analysis. The project develops a hidden Markov model (HMM) formulation, and uses Viterbi decoding to discover actual road segments of trips. This facilitates road-level analysis without the need for high cost traffic on-road traditional sensors. We then generate Dakar traffic intensity maps for main roads, for every hour in the covered 50 weeks, of the year 2013. Moreover, based on time series, analysis, modeling and forecasting techniques, we develop and apply a traffic prediction model to the data and identify significant traffic seasonality patterns.

1.2 Problem Statement

We attempt to solve the following two problems as

- how to develop a real-time traffic monitoring system that can yield fine-grained traffic intensity map from sparse observations of base transceiver station (BTS) fingerprints and an algorithm to achieve time-evolution of the system; and

- how to develop techniques for analyzing, modeling, and forecasting time series of traffic flows in future horizons, from the past historical traffic data with additional road network attributes, e.g., road types.

E-JUST and Waseda University are in charge of the former problem and IBM Cairo is in charge of the latter.
1.3 Objectives

To address the problems described in Sect. 1.2, the following points are considered:

- to define a model representation suitable for modelling the traffic phenomena on the road network;
- to define a model structure to cope with missing (i.e., very sparse) observations as it is not guaranteed that users receive very frequent calls for consecutive regions;
- to develop accurate and efficient trip trajectory mapping techniques;
- to develop traffic intensity map generation techniques;
- to develop an algorithm to achieve efficient time-evolution of the system;
- to establish a validation framework in traffic monitoring research especially in the case where there are no ground truth (like D4D Orange Challenge); and
- to develop techniques for analysing, modelling, and forecasting time series of traffic flows in future horizons, from the past historical data, with additional road network attributes (e.g., road types).

- to discover hidden patterns in the traffic flow data, such as traffic seasonality patterns

1.4 Relevant Previous Work

Recently, several researchers have made a lot of efforts on estimation of human mobility [1] using a set of visited locations for each person [2, 3, 4, 5, 6].

Becker et al. [7] have attempted to identify the route and estimate the traffic intensity from cellular handoff patterns and signal strength. 15 routes were identified using nearest neighbor classifier based on earth mover distance (EMD) [8] or logarithmic signal-to-noise ratio (SNR)-based classifier. The former system is not feasible to apply to a large-scale data as both nearest neighbor-based classification and EMD calculation are significantly time-consuming and the latter does not assume to conduct efficient time evolution of the systems since it is memoryless (i.e., does not use any models).

Krumm et al. [9] applied HMMs to map-matching problem. Thiagaran et al. [10] developed VTrack to address the same problem, which carried out mobile phone localization using WiFi and GPS followed by mapping the location estimates onto the road segments using HMMs. They have also developed CTrack [11], which is well organized system for GSM-based map-matching and designed on the basis of similar concept to our work. CTrack is able to match a set of GSM fingerprints to road segments using HMMs with an accuracy of about 75%. It should be noted that main focus of CTrack is accurate trajectory mapping of each trip while our focus is not only trajectory mapping but also efficient adaptive framework i.e., development of the framework for efficient time evolution of the system. In CTrack, the emission and state transition scores from HMM are heuristic and non-parametric (i.e., explicitly using data to calculate the scores). This
system, therefore, could need troublesome work such as data selection in the big-data tasks and make system adaptation difficult. In contract, our system calculates the emission and state transition scores from probabilistic density functions trained using accumulated statistics to achieve efficient time evolution of the system. Especially, in our method, intensity map generation and model parameter estimation are fully incorporated into time evolution of road and traffic network model based on segmental $K$-means clustering algorithm (or EM algorithm).

In addition, the systems provided by Becker et al. [7] and Thiagarajan et al. [11] ($CTrack$) assume rich information such as high-resolutional observation sequences (e.g., one sample per second and one sample per 10 meters) that consist of the base transceiver station (BTS) fingerprints and signal strengths with time stamps. $CTrack$ can utilize further information such as an accelerometer and a compass to improve the accuracy. In contrast, our system is applicable even under more restricted and challenging condition assumed in the D4D Orange Challenge task, i.e., only very sparse BTS fingerprints with non-constant intervals of timestamps are provided. Our HMM-based system developed in the present work is designed so as to overcome this severe requirements.

Berlingerio et al. [12] have developed $AllABoard$, which estimates origin/destination flows and traffic volumes under the constraints in the D4D Orange Challenge, to optimize public transport. $AllABoard$ consists of 1) stop extraction conducted by each user; 2) aggregated origin/destination flows between those stops; and 3) shared route patterns extraction from the sequences of stops visited by each user. This system does not assume any parametric models to estimate the origin/destination flows and did not consider efficient time evolution of the system using observed data. This work focuses on the flows based on the stops. The yielded routes, therefore, are coarse-grained. In contrast, an attempt is made in our work to estimate the fine-grained routes.

Many efforts have been made to address the problem of forecasting human mobility in the future [13, 14, 15].
## Table 1.1: Urban road and traffic network research.

<table>
<thead>
<tr>
<th>Method</th>
<th>Problem</th>
<th>Dataset</th>
<th>Feature parameter</th>
<th>Sampling rate</th>
<th>model</th>
<th>model parameter</th>
<th>trajectory mapping / classification</th>
<th>model update</th>
<th>ground truth</th>
<th>validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTrack (MIT) [11]</td>
<td>trip trajectory mapping</td>
<td>small scale (4 months)</td>
<td>BTS fingerprints, timestamps, RSSI (up to 7 BTSs per location); sensor hints (accel, compass)</td>
<td>high resolution (sample per second)</td>
<td>HMM</td>
<td>heuristic (non-parametric)</td>
<td>Viterbi search</td>
<td>non-parametric</td>
<td>yes</td>
<td>Precision and recall</td>
</tr>
<tr>
<td>(AT&amp;T) [7]</td>
<td>route identification and traffic intensity estimation</td>
<td>small scale (8 months; 15 routes; only 8 drives for each route)</td>
<td>BTS fingerprints, duration</td>
<td>high resolution (sample per 10m)</td>
<td>Nearest neighbour (NN)</td>
<td>NN (non-parametric)</td>
<td>NN-based classifier (earth mover distance)</td>
<td>N/A</td>
<td>yes</td>
<td>classification accuracy</td>
</tr>
<tr>
<td>AllABoard (IBM Dublin) [12]</td>
<td>trip trajectory mapping (coarse grained) and traffic intensity estimation</td>
<td>D4D Orange Challenge 2013</td>
<td>BTS fingerprints, timestamps</td>
<td>very sparse (with missing observations)</td>
<td>N/A</td>
<td>N/A</td>
<td>Linear optimization</td>
<td>N/A</td>
<td>no</td>
<td>correlation with gravity model</td>
</tr>
<tr>
<td>Proposed (E-JUST)</td>
<td>trip trajectory mapping (fine grained); traffic intensity estimation and model adaptation</td>
<td>D4D Orange Challenge 2014</td>
<td>BTS fingerprints, timestamps</td>
<td>very sparse (with missing observations)</td>
<td>HMM</td>
<td>multinominal distribution for state transition and emission</td>
<td>Viterbi search</td>
<td>segmental $K$-means clustering + MAP-like adaptation</td>
<td>no</td>
<td>correlation with gravity model</td>
</tr>
</tbody>
</table>
1.5 Report Organisation

The rest of the present report is as follows.

Chapter 2 describes data used in the present work, i.e., call detail records; CDR, and pre-processing for traffic monitoring such as extraction of road network attributes, e.g., Volonoi regions, edge, junctions, and exit points.

In Chapter 3, a traffic monitoring system based on hidden Markov models is proposed. This chapter describes as follows: 1) trip detection; 2) trip trajectory mapping of sparse base transceiver station (BTS) fingerprints onto the road segments using Viterbi alignment of exit point sequences and road attributes such as junctions; 3) accumulation of statistics of trip trajectories using segmental $K$-means clustering, yielding road-level intensity map; and 4) adaptive model update algorithm. This section also mention that the proposed model can naturally handle the missing observations.

Chapter 4 describes optimization of Viterbi decoding in trip trajectory mapping. In this chapter, the implementation is optimized significantly, the traits of the model is exploited, and the algorithm is parallelised to scale over the parallel cluster system.

In Chapter 5, the experimental validations are carried out in terms of the accuracy of traffic flows estimated using the developed system under the D4D Orange Challenge task. Since the ground truth information are not provided, the criteria such as the gravity and equilibrium model, which transportation behaviours generally follow, are applied for evaluation.

Chapter 6 describes the use of time series analysis, modelling and forecasting techniques to discover hidden patterns, such as traffic flow seasonalities, and to forecast road traffic in future horizons.

Finally, summary and future work are presented in Chapter 7.
Chapter 2

CDR and Map Generation

This chapter provides background on the provided call detail records (CDR) for the Senegal and our methodology of generating a corresponding traffic map. The map will be used in the following chapter for identifying route trips, given the observation sequences of the BTS.

2.1 CDR from D4D Orange Challenge

In 2014, Orange has held a competition, where they provided various sets of CDR data for processing. In this work we focus on the finest grain set of CDR providing for timestamp in 10 minutes granularity, a corresponding BTS (antenna) ID, and an anonymised user mobile phone numbers. The set is grouped into two weeks buckets, where the users are not changed. The CDR data is collected from about 300,000, randomly chosen users. The CDRs are filtered such that only users who interact more than 75% of the bi-week durations are kept. Also, users with more than 1000 interactions per one week are filtered out, as those are more likely to be machines or shared phones.

In Senegal, the BTSs, the mobile network antennas, are located in various places, with many are co-located. The meta data provided includes distorted BTS locations so as to preserve privacy and for other commercial reasons; the located are changed so that they belong in a random location inside its Voronoi cell. We restrict our analysis to the Dakar area, considering from up to 600 antenna BTS.

In general, a CDR is generated when a user initiates or receives a phone call or an SMS. Nothing is known about the duration of the phone call.

2.2 Map generation

We have used the open street maps (OSM) for Dakar, supplied the competition. The road network is a graph of nodes. A road is represented as a sequence of nodes. The nodes are an important primitive in describing the map; a road is a sequence of nodes, where a road segments is a straight line between two consecutive nodes; we refer to road segments as ‘edges’, using the same terminology as the Simulation of Urban MObility (SUMO) simulator system. A junc-
tion is a special node, where it joins two or more roads. A node can have a ‘shape’ attribute to describe its geometrical structure (e.g., a circle in case of a junction).

Another important structure is the ‘way’ where it is a sequence of nodes that represents a whole road. A relevant attribute, for the purpose of this work, is the type of the way, called ‘highway’. The highway can types in ‘importance’ descending order are: motorway, trunk, primary, secondary, tertiary, unclassified, residential, and service. The ‘unclassified’ is considered the least important in the road network.

Each node is associated with a unique id and an x/y location in a cartesian coordinate system. We utilise the cartesian space of the SUMO simulator system, but without normalisation and shifting (offsetting).

2.3 Zone representation

To represent an BTS coverage area, we partition the space with Vironoi algorithm. Each BTS is thus associated with a Vironoi cell. A zone is thus defined as a convex set of vertices, associated with a zone id. Figure 2.1 shows a typical partition for 300 antenna, covering Dakar city. The rationale behind utilising Vironoi partitioning is that each cell would be closest to the cell’s BTS than any other BTS. Therefore, it is more likely that a mobile user will be associated with that particular BTS. However, as will be explained in the next chapter, that we model the fact that a user can be associated with not necessary the closest BTS; we model that by setting the corresponding emissions probabilities.

From CDR predictive, we set the zone id with the CDR’s BTS id. We also call the value of the zone’s id as an emission. Also the sequence of BTS id’s obtained for a particular users is called a sequence of observations.

2.4 Exit point extraction

As will be discussed in the next chapter, the exit-points represent the state of our model. An exit-point is an intersection of an edge with a cell boundary. To simplify the implementation, we consider the corresponding edge (the one that contains the exit-point) as it is small (in the order of 100 meters) and the SUMO simulator system simplifies computing them; we refer to an edge that contains an exit point as an exit edge.

It is also worth noting that an edge is associated with a direction and number of lanes. Therefore, and exit-edge/point is labeled ‘exit’ as such point is considered an exit from current Voronoi cell.

We summaries all definition in the next section.
Figure 2.1: Voronoi partition for 300 BTSs
Chapter 3

Traffic Monitoring Using Hidden Markov Model

3.1 Overview

A typical urban road network is modeled as a graph where each node is a junction, and an edge represents the road edge; the road edge consists of one more straight line segment; the latter essentially provide a shape for the road. In the present study, hidden Markov models (HMMs) are exploited to represent the traffic phenomena on that road network. In this model, each state is defined to be the exit edge point. Figure 3.1 illustrates the conceptual image of the urban road network and HMM structure applied.

Figure 3.2 illustrates the diagram of the system applied after generating the map. The system mainly consists of four-stage processing as follows:

1. **trip detection**: Base transceiver station (BTS) fingerprints for each trip (only the mobility) are extracted from all the BTS fingerprints including stops.

2. **trip trajectory mapping using Viterbi search**: A trajectory for each trip is estimated by mapping BTS fingerprints onto the road segments. This procedure is achieved by the following two steps:
   - **Viterbi alignment**: An optimal exist point sequence is estimated from BTS fingerprints for each trip using Viterbi algorithm.
   - **accurate route estimation (post-processing of Viterbi decoding)**: More accurate route for each trip is estimated using not only the Viterbi outputs (i.e., exit point sequences) but also other road network attributes such as edges and junctions.

3. **statistics accumulation of trip trajectories and traffic intensity map generation**: The statistics of trip trajectories are accumulated. The traffic intensity map can be yielded using the statistics in terms of state transitions. The statistics of exit point sequences are also accumulated for model adaptation.
Figure 3.1: Conceptual image of (1) road network and (2) HMM structure applied. Blocks $e_i$ in upper figure express exit points, which is defined as states in HMM.
4. **model update using accumulated statistics**: Efficient time evolution of the system is conducted by adaptively updating the model parameters using the previous model and aforementioned statistics.

### 3.2 Trip Detection

HMM used does not explicitly incorporate time into its model as will be discussed in the following sections. It is therefore important to preprocess the observations so as to filter out the ones that are mainly due to the user arrived to his/her destination and not in transit. We base our approach on that proposed by Berlingerio et al. [12]: the approach computes a set of stop points for each user, where every two consecutive (in time) stops constitutes a trip. We therefore extend the approach so as to allow for keeping the intervening observations.

We utilise the same notation as Berlingerio et al. such that $H$ is the historical activity of a particular user; $H = < a_1, a_2, \ldots, a_n >$ where $a_i$ is the $i$-th CDR observation given by $(b, t)$, $b$ is the BTS id, and $t$ is the timestamp.

We define a stop similar to Berlingerio et al. where a stop is the maximal subsequence

$$ s = < a_m, a_2, \ldots, a_k > $$  \hspace{1cm} (3.1)

s. t. $0 < m \leq k \leq n$,  \hspace{1cm} (3.2)

$$ \max_{\forall m \leq i \leq j \leq k} \text{Distance}(a_i, b_j) < \text{th}_s, $$  \hspace{1cm} (3.3)

$$ \text{Duration}(a_m, a_k) \geq \text{th}_t, $$  \hspace{1cm} (3.4)

where $\text{th}_s$ is spatial threshold (set to 1 km) and $\text{th}_t$ is temporal threshold (set to 1 hour).
In our formulation, we differ in that we reduce the stop sequence into the last element \( a_k \) as well as extracting the sequence \(< a_1, \ldots, a_k >\) as a trip. Moreover, we remove successive repeated observations with the same timestamp (the resolution of the data is in 10 min). The rationale here is that such small sequence is not caught by the stopping definition above and it can lead to false counting of the user as moved.

3.3 Road And Traffic Modeling Based on HMM

3.3.1 Model Representation

We define the states here to be the exit edge points. In the present work, both state transition and emission probability in HMMs are represented by multinomial (discrete) distribution. \( a_{ij} \) denotes the probability of the state transition \( S_i \) to \( S_j \); \( b_i(c) \), the probability of the signal from the BTS \( c \) being received at the state \( S_i \); \( N_S \), the number of states; and \( Z_i \), the zone where \( S_i \) is included. In the road and traffic network model, same state transitions \( a_{ii} \) (\( 1 \leq i \leq N_S \)) and adjacent state transitions \( a_{ij} \) (\( i \neq j; 1 \leq j \leq N_S; \) and \( S_i \) is adjacent to \( S_j \)) are dominant among all state transitions. It, however, is much likely to have missed observation as it is not guaranteed that users receive very frequent calls for each region. Small probabilities, therefore, should be assigned to unadjacent-state transitions in order to cope with the missed observations. As for the emission probability, observations from neighboring zones should be considered. This accounts for error in the communication model representation, where a call is not necessary associated with the nearest BTS. To handle this phenomena, the multinomial (discrete) distributions for all BTSs being observed are assigned to the states.

3.3.2 Parameter Initialization

The state transition probabilities are initialized on the basis of the distance between the states as

\[
a_{ij} = \frac{1/D(S_i, S_j)}{\sum_{1 \leq k \leq N_S} 1/D(S_i, S_k)},
\]

where \( D(S_i, S_j) \) denotes the Euclidean distance between the exit edge points (i.e., states) \( S_i \) and \( S_j \).

The initial emission probability of the BTS \( c \) being observed at \( S_i \) can be calculated as

\[
b_i(c) = \begin{cases} 
0.8 & (c \sim Z_i) \\
0.2 \cdot w_{i,k} & (c \sim Z_{k \neq i})
\end{cases},
\]

where \( c \sim Z_i \) expresses that the BTS \( c \) is in the zone including \( S_i \) and \( w_{i,k} \) denotes the weight of the signal from \( Z_{k \neq i} \) being received at \( S_i \), which is computed as

\[
w_{i,k} = \frac{1/\|p_i - p_k\|}{\sum_u 1/\|p_i - p_u\|},
\]

where \( p_i \) denotes the location of the BTS in \( Z_i \).
3.4 Trip Trajectory Estimation

The road-level trajectory of a trip is estimated by the forced alignment using Viterbi algorithm followed by post processing using road network attributes.

3.4.1 Forced Alignment Using Viterbi Algorithm

Forced alignment using Viterbi algorithm estimates an optimal state sequence (exit point sequence) $\hat{S} = \hat{s}_{n,1}, \cdots, \hat{s}_{n,T_n}$ for the observation sequence of the $n$-th trip $x_{n,1}, \cdots, x_{n,T_n}$.

Let $v_n(t, i)$ be the emission probability of $x_{n,1}, \cdots, x_{n,t}$ being generated along the optimal state sequence and $x_{n,t}$ being emitted at $S_i$; $\psi_n(t, i)$, the pointer to the previous state; $T_n$, the number of samples in the $n$-th trip; and $N_S$, the number of states. In this case, the the optimal state sequence can be obtained by the following procedure:

1. Initialization ($1 \leq i \leq N_S$)

$$v_n(1, i) = \pi_i \cdot b_i(x_{n,1}) \quad (3.8)$$
$$\psi_n(1, i) = 0 \quad (3.9)$$

2. Iteration ($2 \leq t \leq T_n, 1 \leq i \leq N_S$)

$$v_n(t, i) = \max_{1 \leq k \leq N_S} \left[ v_n(t-1, k) \cdot a_{ki} \right] \cdot b_i(x_{n,t}) \quad (3.10)$$
$$\psi_n(t, i) = \arg \max_{1 \leq k \leq N_S} \left[ v_n(t-1, k) \cdot a_{ki} \right] \cdot b_i(x_{n,t}) \quad (3.11)$$

3. Convergence

After iterative calculation of $v_n(t, i)$ from $t = 1, \cdots, T_n$, the emission probability of $x_{n,1}, \cdots, x_{n,T_n}$ being generated from the optimal state sequence can be obtained as follows:

$$P(x_{n,1}, \cdots, x_{n,T_n}) = \max_{1 \leq k \leq N_S} \left[ v_n(T_n, k) \right] \quad (3.12)$$
$$\hat{s}_{T_n} = \arg \max_{1 \leq k \leq N_S} \left[ v_n(T_n, k) \right] \quad (3.13)$$

4. Backtracking ($t = T_n - 1, T_n - 2, \cdots, 1$)

The optimal state sequence of the $n$-th trip can be obtained by tracking back the nodes giving the maximum likelihood as

$$\hat{s}_{n,t} = \psi_n(\hat{s}_{n,t+1}, t + 1) \quad (3.14)$$
3.4.2 Accurate Road-Level Route Estimation

Viterbi search can generate the optimal state sequence, which consists of a set of edge exit points. Then, a post-processing of Viterbi decoding is carried out: more fine-grained road-level route for each trip can be determined using additional road network attributes such as junctions as well as the edge exit points yielded by Viterbi decoding. Figure 3.3 illustrates the road-level route (trip trajectory) for various exit point choices; (1) the basic road network with four zones; (2) the route from junction exit points; (3) the route from edge exit points (Viterbi outputs); and (4) the route from junctions and edge exit points (Viterbi outputs), which is applied to the proposed system. It should be noted that the junction and edge exit point route has substantially smaller state space and can be appropriate to consider depending on the problem scale. Figure 3.4 illustrates an example of the resulting route.

3.5 Statistics of Trip Trajectories And Traffic Intensity Map Generation

Two sorts of statistics are accumulated in the developed system: the statistics for road-level, fine-grained trip trajectories and those of the BTS fingerprints for exit edge point (state) sequences. The former yields the traffic intensity map and the latter is utilized for model adaptation.

The traffic intensity, which represents the number of vehicles passing through the specific road segments during a given period with \( N \) trips, can be computed as statistics of trip trajectories that are yielded as described in Sect. 3.4. In this case, the number of trip trajectories are accumulated for \( N \) trips for each road segment. Figure 3.5 shows an example of the yielded traffic intensity map for one hour.

Forced alignment using Viterbi algorithm can assign an observation \( x_{n,t} \) to a state in HMM. As aforementioned, this assignment yields an optimal state (exit point) sequence for a trip. Let \( \xi_n(t, i, j) \) be an assignment of the observations \( x_{n,t-1} \) and \( x_{n,t} \) to the state \( S_i \) and \( S_j \), respectively; and \( \gamma_n(t, i) \) be an assignment of \( x_{n,t} \) to \( S_i \). In the case of Viterbi alignment, \( \xi_n(t, i, j) \) takes one if the state transition \( s_{t-1} = S_i \) to \( s_t = S_j \) is on the optimal state sequence for the \( n \)-th trip and otherwise takes zero. In addition, \( \gamma_n(t, i) \) takes one if the state \( s_t = S_i \) is on the optimal state sequence and otherwise takes zero. Noted that the statistics of \( \xi_n(t, i, j) \) and \( \gamma_n(t, i) \) accumulated along the optimal state sequences are used also for updating the model parameters under the maximum likelihood criterion.

3.6 Adaptive model estimation

Time evolution of the system can be achieved by adaptively updating the model parameters in HMMs. The maximum likelihood estimates of the state transition and emission probability distribution can be calculated using accumulated statistics about the BTS fingerprints for the state sequences, \( \sum_{n=1}^{N} \sum_{t=1}^{T_n} \xi_n(t, i, j) \) and \( \sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_n(t, i) \), obtained as the results of Viterbi
Figure 3.3: Comparing between various exit point choices. (1) road network with four zones; (2) junction exit points route; (3) edge exit points route; and (4) combined junction and edge exit points route.
Figure 3.4: Examples of road-level, fine-grained trip trajectory.
alignment, as follows:

\[
\begin{align*}
    a_{ij}^{\text{ML}} &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \xi_n(t, i, j)}{\sum_{n=1}^{N} \sum_{j=1}^{N_S} \sum_{t=1}^{T_n} \xi_n(t, i, j)} \quad (3.15) \\
    b_{i}^{\text{ML}}(c) &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_n(t, i) \cdot \delta(x_{n,t}, c)}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_n(t, i)} \quad (3.16)
\end{align*}
\]

where \(a_{ij}^{\text{ML}}\) denotes the ML estimate of the probability of the state transition \(S_i\) to \(S_j\); \(b_{i}^{\text{ML}}(c)\), the ML estimate of the probability of the signal from BTS \(c\) being received at the state \(S_i\); \(N\), the number of training data sequences (i.e., trips); \(T_n\), the number of samples in \(n\)-th trip; and

\[
\delta(x_{n,t}, c) = \begin{cases} 
1, & \text{if } x_{n,t} = c \\
0, & \text{otherwise}
\end{cases} \quad (3.17)
\]

For example, Eq. (3.15) represents the rate of the vehicles receiving the signal while passing through the road \(S_i\) to \(S_j\) among the vehicles passing through \(S_i\) for \(N\) trips.

The maximum likelihood estimates described in Eqs. (3.15) and (3.16) are possible to be unreliable for the states and state transitions with only small amount of data assigned. To solve this problem, an attempt has been made to linearly interpolate the ML estimates of the model \(a_{ij}^{\text{ML}}\) and \(b_{i}^{\text{ML}}(c)\) with the previous model parameters \(a_{ij}^{(r)}\) and \(b_{i}^{(r)}(c)\) yielding the new estimates \(a_{ij}^{(r+1)}\) and \(b_{i}^{(r+1)}(c)\) as follows:

\[
\begin{align*}
    a_{ij}^{(r+1)} &= \alpha \cdot a_{ij}^{(r)} + (1 - \alpha) \cdot a_{ij}^{\text{ML}} \quad (3.18) \\
    b_{i}^{(r+1)}(c) &= \alpha \cdot b_{i}^{(r)}(c) + (1 - \alpha) \cdot b_{i}^{\text{ML}}(c) \quad (3.19)
\end{align*}
\]

where \(r\) denotes the number of updates and \(\alpha\), the interpolation coefficient.
Figure 3.5: Example of traffic intensity map for one hour.
Chapter 4

Optimising the Viterbi Decoder

The execution time of the Viterbi algorithm is critical, owing to the large number of states and observation sequences. A naive implementation of the algorithm results in around 80 seconds to reconstruct a single observation sequence (running on a recent iDataPlex, Intel Xeon CPU E5-2660, 2.20GHz based processing node). For the 300,000 users provided, the number of observation sequences per day is in the order of 100,000; thus this will require 92 days to process a single day.

Sampling of the observations is a key to decreasing such number, however, that would trade off with accuracy of the obtained results. In this chapter, we therefore, optimise the implementation significantly, exploiting traits of the model, and parallelising the algorithm to scale over our parallel cluster system.

4.1 Viterbi Algorithm

Algorithm 1 shows the Viterbi algorithm. The algorithm takes the following as input:

- Set of states, ‘$S$’; it contains the HMM states.
- The transition probabilities among states, ‘$P_{i,j}$’ (where $i, j \in S$).
- A set of all possible emissions, ‘Em’, that can be emitted by all the states.
- The emission probabilities for each state, ‘$E_{i,j}$’ (where $i \in$ Em, $j \in S$).
- An initial starting probabilities for each state, ‘StartProb$_i$’ (where $i \in S$).
- The given observation sequence, ‘Obs’, which the Viterbi algorithm decodes, finding the corresponding hidden states that emits such sequence with maximal probability.

The algorithm starts by the initialisation of the Viterbi array, $V$, to null (line 2); $V$ is a two dimensional array where the first index is the time and the second is the state. The array holds the best (maximum) Viterbi probability for a given state. The algorithm also initialises the Path,
$P$, array to null (line 3). The path is a one-dimensional array indexed by the state and contains the sequence of states constituting the best current path to that particular state.

The algorithm then initialises the V array to initial probabilities, and P array to the corresponding state (lines 5 – 8).

The algorithm main loop (lines 8–17) iterates over each time step (corresponding to an observation), and for each state it checks the best predecessor state, that has a maximum probability (line 12). As we have $n$ states, and $l$ observations (or time steps), the complexity of such operation is $O(n^2l)$.

The last part of the algorithm selects the V entry with highest probability and emits it and the corresponding path (lines 18–20).

4.1.1 Eliminating Multiplication by Zero Computations

One obvious optimisation is eliminating zero transitions from calculation. This can be easily done by providing a linked list representation of all predecessor states to a particular state, thus potential improving the asymptotic complexity by $n$ for sparse transitions.

A more difficult optimisation is to eliminate multiplication with zero emission probabilities, i.e. for states that have zero emission probabilities. The problem here is that the zero has to be written into the corresponding entry in the V array, requiring the same order of computation. However, we propose a simple data structure that allows for avoiding the zero writes.

A key observation with the Viterbi algorithm is that at a given time step, $t$, all states values (in the V array) are updated, based on reading values from the previous time step $t−1$. We thus associate a timestamp, ‘$T$’, with each value indicating the time of the last update, as well as keeping the last two written values, at $T$ and $T − 1$; when reading and writing, the current time step value, $t$, is given, and the value is either one of the stored values or zero if $T < t − 1$.

Algorithms 2 and 3 below provide the corresponding algorithms for writing and reading. The V array dimensionality is reduced to a one dimension, such that each entry includes:

- $T$: time stamp of the last write operation.
- $v[0..1]$: an array of two Viterbi values, organised as stack.
- $h$: in the index of the head of the stack, indicating the most recent value.

The write operation (Algorithm 2) writes the value onto the oldest entry, updating the head index, $h$, to point to the other entry. If the ‘WriteTimestamp’ is not the next $T + 1$ (i.e. not the next, expected, time stamp), the oldest entry is nulled; in this case we maintain the condition that the last entries are recorded, and since no write is done for $T − 1$, it is set to zero.

The Read operation (Algorithm 3) takes the read timestamp, ‘ReadTimestamp’, and the corresponding entry values, and returns the correct value; as the entry includes values stored into $T$ and $T − 1$, a read to an earlier entry returns zero.

The full algorithm after optimisation is given in Algorithm 4; the modified lines are shown in ‘bold’.
4.1.2 Parallelisation of the Decoding Process

The parallelisation of the process is trivial as every observation sequence can be processed inde- cently. We have utilised a 3-node IBM iDataplex cluster for conducting our experiments. Each node is equipped with Intel Xeon CPU E5-2660, 2.20GHz processor, and 32GB RAM. The cluster also includes Nvidia Kepler GPUs and Intel Xeon Phi accelerator (though the accelerators are not used in this project).

4.1.3 Performance Analysis

It is worth noting that when considered the total number of exit-edges considers (order of 6000) and a sequence of 16 observation the execution time of the base algorithm is approximately 8000ms. Our new optimisation results in around 40ms, given a speed up of 200x. Moreover, with increasing the number of roads (e.g. when considering a larger area), the number of transitions and observations per state is likely to be constant as it depends on the local traffic network topology and antenna. Therefore the computation complexity becomes $O(n)$ (in both time and space).
Input: Emissions, \( E_m = \{e_0, e_1, \cdots, e_{m-1}\} \) where \( m \) is the number of all possible emissions.

Observation sequence, \( O = \langle o_0, o_1, o_2, \cdots, o_{l-1} \rangle \) where \( o_i \in E_m \) and \( l \) is length of the sequence.

Transition Probabilities, \( P_{i,j} \) such that \( 0 \leq i, j \leq n - 1 \).

States, \( S = \{s_0, s_1, \cdots, s_{n-1}\} \).

Emission Probabilities, \( E_{i,j} \) such that \( i \in S \) and \( j \in \text{obs} \).

StartProb \( \in S \).

Output: Max. Probability, OutputProb

Edge sequence, OutputPath = \( \langle e_0, e_1, \cdots, e_{l-1} \rangle \)

\begin{algorithm}
\begin{algorithmic}
\State \( V \leftarrow \phi \);
\State Path \( \leftarrow \phi \);
\Comment{Initialise the case for the observation 0}
\For{all \( y \in S \)}
\State \( V[0][y] \leftarrow \text{StartProb}[y] \times E_{y,O_0} \);
\State Path\[y\] = \( y \);
\EndFor
\Comment{Proceed with dynamic programming for the rest of observations}
\For{\( i \leftarrow 1 \) to \( l - 1 \)}
\For{all \( y \in S \)}
\State \((\text{prob}, \text{state}) = \max((V[t-1][y_0]P_{y_0,y}E_{y,O_1}, y_0 \in S)) \);
\State \( V[t][y] = \text{prob} \);
\State new\[y\] = path\[state\] + \[y\];
\EndFor
\EndFor
\State path = newpath;
\State \((\text{prob}, \text{state}) = \max((V[n][y], y \in S)) \);
\State OutputProb = \text{prob};
\State OutputPath = path\[state\];
\end{algorithmic}
\end{algorithm}

Algorithm 1: Viterbi Algorithm
Algorithm 2: Write Algorithm

Input: WriteTimestamp
Output: $T, v[2], h$

begin

1. $h \leftarrow h$
2. $v[h] \leftarrow$ value;
3. if $(T+1) \neq$ WriteTimestamp then
4. $v[h] \leftarrow 0$
5. $T \leftarrow$ WriteTimestamp;

end

Algorithm 3: Read Algorithm

Input: ReadTimestamp
Output: Value

Input: $T, v[0..1], h$

begin

1. $\Delta T \leftarrow T - \text{ReadTimestamp};$
2. if $(\Delta T = 1)$ then
3. Value $\leftarrow v[\bar{h}]$
4. else
5. if $(\Delta T = 0)$ then
6. Value $\leftarrow v[h];$
7. else
8. Value $\leftarrow 0;$
9. end
10. end
11. end

end
**Input:** Emissions, \( \text{Em} = \{e_0, e_1, \cdots, e_{m-1}\} \) where \( m \) is the number of all possible emissions.

Observation sequence, \( \text{Obs} = <o_0, o_1, o_2, \cdots, o_{l-1}> \) where \( o_i \in \text{Em} \) and \( l \) is length of the sequence.

Transition Probabilities, \( P_{i,j} \) such that \( 0 \leq i,j \leq n - 1 \).

States, \( S = \{s_0, s_1, \cdots, s_{n-1}\} \).

Emission Probabilities, \( E_{i,j} \) such that \( i \in S \) and \( j \in \text{obs} \).

StartProb\(_i \in S\).

**Output:** Max. Probability, \( \text{OutputProb} \)

Edge sequence, \( \text{OutputPath} = <e_0, e_1, \cdots, e_{l-1}> \)

```java
begin
1 V ← [φ];
2 Path ← φ;
3 /* Initialise the case for the observation 0 */
4 for all \( y \in S \) do
5     /* Initialise the Veterbi entry to have a timestamp \( T \) with zero, \( v[0] \) with StratProb, \( v[1] \) with zero, and head index \( h \) with zero */
6     V[\( y \)] ← (0, \text{StartProb}[y] \times E_{y,o_0}, 0, 0);
7     Path[\( y \)] = \( y \);
8 end
9 /* Proceed with dynamic programming for the rest of observations */
10 for \( i \leftarrow 1 \to l - 1 \) do
11     for all \( y \in S \mid E_{y,o_t} \neq 0 \) do
12         (prob, state) = \text{max}(\text{Read}(V[\( y_0 \)], t - 1)P_{y_0,y}E_{y,o_t}, y_0 \in S \mid P_{y_0,y} \neq 0);
13         Write(V[\( y \)], t, prob);
14         newpath[\( y \)] = path[state] + \[ y \];
15     end
16     path = newpath;
17 end
18 OutputProb = prob;
19 OutputPath = path[state];
end
```

**Algorithm 4:** Optimised Viterbi Algorithm for Sparse Transitions and Emissions
Chapter 5

Validation and Analysis

In this chapter we evaluate the accuracy of the trip trajectories generated by our developed system. The generation is based on the D4D Data Set 2, provided by the D4D Orange Challenge 2014. The effectiveness of adaptive training is left for future work. In the challenge, the ground truth information (i.e., correct labels) about the routes of trips are not provided. We, therefore, attempt to validate our developed system in terms of our estimated traffic flows, as conducted by Berlingerio et al. [12].

Figure 5.1 shows the hourly flow distribution for each edge-exit point, together with the mean hourly flow. Here, the flow is defined as the number of trips at each edge during a given period of one hour. This distribution shows a clear peak in the evening of the day, indicating the trend that congestion is worse in the evening; it is similar trend to the urban mobility patterns reported in [16].

The estimation process of trip flows can be validated by comparing the results from the developed system and those from the gravity model, which is the basic model used in transportation and defined in [12] as

\[ \text{Gravity}(O, D) = \frac{O_{\text{out}} \times D_{\text{in}}}{\text{Distance}(O, D)^2}, \]  

(5.1)

where \( O \) and \( D \) denote the consecutive edge points corresponding to the origin and destination; \( O_{\text{out}} \), the sum of all the flows going out from the edge \( O \); and \( D_{\text{in}} \), the sum of all the flows ending in \( D \). Figure 5.2 shows the correlation between the flows yielded by the developed system and corresponding gravity model. The result shows that the flows from the developed system gave a high correlation (0.93) with the gravity model, indicating that the developed system follows well to general traffic phenomena.
Figure 5.1: Hourly distribution of flows for each edge.
Figure 5.2: Correlation between flows of each origin/destination pair and corresponding gravity model.
Chapter 6

Traffic Time Series Analysis, Modelling and Forecasting

6.1 Introduction

Congestions in urban roads is a major problem in cities, negatively affecting population mobility and may result in deteriorating and delaying their daily business activities, while at the same time increasing both fuel consumption and traffic-related air pollution. To mitigate congestion in urban roads and related consequences, local city urban planners and decision makers need to have a thorough detailed insight, comprehensive knowledge and analysis of the city road traffic patterns. This could include the ability to predict future traffic flow and potential congestions, on every road during every hour of the day, throughout the year. Correlation of road traffic patterns with weather conditions such as rain and fog and road surface conditions, is also very important in assessing future road potential traffic congestions and the influencing factors.

In order to understand, analyze and predict road traffic conditions in a city, instead of using conventionally expensive and quickly outdated travel surveys, we use in this study records of daily mobility, based on large-scale mobile phone data. We approach the problem by observing the number of individuals moving from one location to another, on the city roads, based on their calls and text messages. We base our study on the following generated data set from CDR data:

Computed accumulated number of moving mobile users per unit of time, flowing on the city road segments (Dakar in this case), for every hour and every day during the year 2013

For this data set, we used Time Series analysis, modeling and forecasting techniques, in order to identify the seasonality in the traffic flow, on road segments (categorized by road type). Hourly and daily seasonality were investigated. Then we forecast the traffic intensity in future time horizons.

It is to be pointed out that traffic flow rates of mobile users on road segments based on the provided CDR data, has to be corrected to take into consideration:

- percentage of mobile penetration in Senegal population
- market share of Orange operator in Senegal (as CDR data is related to Orange subscribers only)
probability of a mobile user making a call while he is on the road

probability of a mobile user drive alone or share a car

Finally we need also to estimate the actual number of mobile users from the 300000 randomly sampled Orange operator given users.

From figures provided by ARTP regulator in Senegal for the last quarter in 2012 (source: BIZTECH AFRICA, May 2013), mobile penetration in Senegal is estimated to be 94.24%, and Orange market share in Senegal is estimated to be 62%.

6.2 Road traffic Time Series analysis, modeling and Forecasting

6.2.1 Time Series analysis

Around 608 road edges (segments) were selected from Dakar road network for this analysis. For each edge/segment, we have examined the associated Time Series (TS) for the accumulated hourly and daily mobile users flow, during 350 days (from Jan 7th to Dec 22nd) in 2013. Some of the TS contained missing data. The missing data were handled using a weighted local average imputing method.

In Figure 6.1 we notice that there is no trend (since a constant random sample of 300 000 Orange mobile users is always used). However a strong seasonality is apparent.

From the monthly aggregated traffic data in Figure 6.2, we have identified a noticeable increase in traffic flow during summer (July, August and September), that reaches its peak in August, then starts to decrease slowly until it reaches its minimum in December.
6.2.2 Time Series Decomposition

This was performed using IBM SPSS Statistics 22 software. We performed TS decomposition for the following types of seasonality, using the SEASON function of IBM SPSS Statistics:

- 24 hours per day
- 7 days per week

Additive seasonality decomposition, instead of multiplicative, was used since TS examination showed no increasing seasonal variance.

Figure 6.3 shows the seasonal hourly variations for the hourly road accumulated flow, as depicted by the TS Seasonal Adjustment Factor (SAF). We notice that the road traffic flow starts
to build up from 7 am until 11 am, then stabilizes between 11 am and 6 pm. After 6 pm, the flow increases rapidly until it reaches its peak at around 1 am, then decreases to its minimum at 5 am.

Figure 6.4 shows the seasonal daily variations for the daily road accumulated flow, as depicted by the SAF. We notice that the traffic increases gradually during the week starting Mondays, and reaches its maximum on Sundays.

6.2.3 Time Series modeling and forecasting

For TS modeling, We have applied the Expert Modeler of IBM SPSS Modeler on the daily accumulated traffic flow on the 608 road edges, with 7 days per week seasonality. The data were divided into 80/20 for training/testing. Then we forecasted 2 months = 56 days ahead. The generated TS models by IBM SPPSS Modeler Expert Modeler that best fit the data were mostly simple seasonal models, as most of the time series have no trend. But some were ARIMA models. Figure 6.5 shows one of the automatically generated TS forecasting model associated with one of the road edges, and the generated forecast.

To measure the accuracy of the forecast, we used the Mean Absolute Percentage Error (MAPE) measure. The average MAPE for all generated TS models was 27.4, which is acceptable, and the frequency distribution of the MAPE error for the generated Time Series for all road edges, is given in Figure 6.6.
Figure 6.5: A sample TS model for modeling and forecasting daily traffic flow on one of the road edges

Figure 6.6: MAPE frequency distribution for the generated TS models
Chapter 7

Conclusions and Future Work

7.1 Summary and Conclusion

This project has developed a HMM-based model for identifying trips on the road-level using big cellular phone’s CDR. A key problem with this approach is it is not guaranteed that a user will generated a contagious trajectory of observations. Thus modelling this problem is important to incorporate into our model. We have provided this functionality through allowing for not only next neighbours exists, but also further on neighbours. Results indicated a possible solution for are around 58% for all generated trips.

The project preprocess the CDR sequences (for a particular user) to eliminate stopping states, where the user is not in a ‘trip’ anymore. This filter operation is done subject to a distance and time criteria.

The model parameters for the main HMM-based method are computed such that the transition probabilities are functions of their shortest path distance; the self transition has given the value of the furthest point so as to decrease emphasising staying in a current zone. Also, in a preprocessing stage, we eliminated repeated observations with zero time differences (in 10m resolution). This removes a case that is not modelled by our HMM, where we assume the mobile is in consistent trip mode (moving), whereas a repeated observation with very short time indicates a stop. In general we do not eliminate repeated states as that can potentially lose information (such as if the mobile moved to a next neighbour zones while the reported observation int eh CDR is still the same previous zone. This can happen especially due to the imprecise BTS locations.

We also consider the problem of validating results even though there is not ground truth. We, therefore, resorted to visual hourly flow outs over the whole year and show similar trend of edges individual flow to that or aggregate and typical analysis, elsewhere [16]. We also considered the traffic gravity model, and achieve high coloration between measured using our and gravity model values. Moreover, we consider the more strict equilibrium based modelling. All validation methods have shown strong coloration, indicating the accuracy of the approach is acceptable.

In this study, we used, as well, Time Series analysis, modelling and forecasting techniques, in order to identify the seasonality in the traffic flow, on road segments. Hourly and daily sea-
sonalities were investigated. Then we forecast the traffic intensity in future time horizons.

This study can lead to generating traffic intensity map for Senegal’s major roads. Together with the provided traffic analysis and forecasting component, the project can supply accurate information to both public users and decision makers, helping to improve a traffic metric and support ‘new road?’ planning, thereby decreasing road congestions in Senegal.

7.2 Future work

There are many research points made open by this project; a key element is further validation of the obtained trajectories to the ground truth; this can be achieved by conducting limited experiments on a number of volunteers.

Another research point is provide for probability adaption so as to learn the transition probability better; machine learning methods can be potentially investigated. Moreover, the more elaborate EM can be utilised to allow for more modelling power and complexity. In addition, with the seasonality effects, even on the daily basis (weekend and weekdays), different sets of probabilities can be trained accordingly.

Another important aspect is identifying vehicles. This work considers the mobility of users. An important future work would be to cluster the locations so as to identify the vehicle as well as its type.

Also, it would be interesting to consider traffic speeds as well as flows in traffic analysis and forecasting. Average velocities on road segments depend mainly on the road traffic conditions and congestions. Therefore, average velocities can better represent the traffic flow on each road segment, at a specific hour of a day of a year, and can be used in this study to predict the road segment flow status at future dates.

To provide an even higher value for the end user, a system-optimal routing can be investigated given the detailed data we obtain; the aim of which is to provide for a global routing problem for all cars so as to maximise a specific metric, such as commute time.

The use of GPU and accelerators is an important aspect to pursue in the future. This will provide for most cost effective scaling up, especially when considering all roads, and requiring real-time processing.

Finally, integrating all the above with into a system and deploy in the Senegal.
Bibliography


1 Introduction

It is widely understood that the communication infrastructure is a critical societal infrastructure, and forms the connecting fabric for a country. One of the challenges during a disaster event is that cell towers might become in-operational either due to direct damage, or due to loss in power. This can cause severe problems if there are critical towers or regions of the city, where failures can disconnect the connectivity of the network—identifying them can help city planners understand the inherent vulnerability of the communication infrastructure, and develop contingency plans before a disaster.

We study this problem for the Dakar metropolitan area in Senegal in two network abstractions of the call data records: the cell-tower call network (referred to as the call network, and denoted by $G_c$) and a network induced by the mobility data (referred to as the mobility network, and denoted by $G_m$). We study resilience of these networks to failures and the underlying community structure. Our main observations are summarized below.

1. Resilience to failures: We find that the call network $G_c$ is quite resilient to geographically correlated failures of cell towers. In contrast, the mobility network, $G_m$, has relatively small separators, and is much more vulnerable to such failures.

2. Community structure: We examine the community structure in both networks, and find a surprising degree of spatial correlations within communities. There are fairly distinct communities, which are located in different regions of the city. Some of the individual communities appear to be more vulnerable to loss of connectivity due to failures in cell towers.

3. Regions of significant variations in mobility and call volume: we use an approach based on spatial scan statistics and find a small region of the city where such variations are significant. Increasing the resilience of the towers in this area would be important for handing such spikes.

2 Background

We consider two networks based on the D4D call data records:

1. Call network $G_c = (V, E_c)$, where $V$ denotes the cell towers in this region, with an edge $e = (u, v) \in E_c$ if there is a call from a user connected to cell tower $u$ to a user connected to cell tower $v$. We also have a weight $w_e$ associated with each such edge, which indicates the number of calls between these towers.
2. Mobility network $G_m = (V, E_m)$, where $V$ denotes the cell towers in this region, with an edge $e = (u, v) \in E_m$ if there is a mobility trace in which a user is recorded at cell tower $u$ at some time and is later recorded at cell tower $v$. Similarly, we have weight $w_e$, which corresponds to the number of such movements. $G_m$ is actually a directed network, indicating the mobility flux across regions; in our analysis, we ignore the directionality of the edges.

Both these networks can be constructed for any time period. In our analysis, we consider them for each month separately. The weight distribution in both networks is very skewed, as shown in Figure 1.

3 Community structure and resilience

We use the Louvain clustering method of [1] on the two networks constructed for each month, as shown in Figure 2. Because of the highly skewed weight distribution (Figure 1), we consider the subgraph with weights more than 1000.

We observe in Figure 2(a) that the call network has four major communities, which are geographically laid out. This is somewhat surprising, since most of this region is fairly dense, with dense urban land use [2]. The spatial structure is partly the result of strong spatial correlation between the end points of the edges of the call network, as shown in Figure 3. Further, the entire network is very well connected, making it quite resilient to any spatially localized failures in the region. In contrast, the mobility network $G_m$ has many more and smaller communities, as shown in Figure 2(b). It can also be shattered relatively more easily with smaller number of spatially localized failures.

The networks show some variation over time, as shown in Figures 4 and 5, though the salient features remain consistent.

We note that the mobility network $G_m$ only considers incomplete information, because a mobility trace from tower $u$ to tower $v$ might not record the intermediate towers during a person’s movement. However, such movement is constrained to happen on the transportation infrastructure; considering such mobility will introduce more edges to the set $E_m$ in graph $G_m$. However, these edges are likely to be short range, and will cause some adjoining communities to merge, but will still retain some of the small separators.
Figure 2: Community structure for the networks restricted to calls in January: (a) $G_c$ and (b) $G_m$. Colors show the different communities.

Figure 3: Frequency distribution of the distance between the end points of edges in $G_c$ for the month of January.

Figure 4: Community structure for the networks restricted to calls in February: (a) $G_c$ and (b) $G_m$. 
4 Regions of significant variations in mobility and call volume

We now study the problem of identifying regions of significant variation in call and mobility patterns. We use an approach based on spatial scan statistics developed by Neill [3]—this approach involves defining a score function $F(S)$ for a region $S$, and identifies the region with the maximum score. We take $F(S)$ to be the number of calls at towers in the region $S$ for a specific time interval. In the results shown in Figure 6, we consider a time duration of 1 day. Further, we consider spatial regions in the form of disks of radius 1 km, and consider the maximum over all such regions. Our implementation identifies the region $S$ with the maximum score, compared to the scores within the past 28 days.

Figure 6 shows regions for two different periods. We find the same set of towers in the middle of the Dakar region as having the highest scores. This suggests that the towers in this region are crucial components of the cellular infrastructure. Some of these towers lie in regions with high risk of flooding or coastal erosion, as identified in [2].
References


Exploring relationships between human mobility motifs and rainfall for development

Aki-Hiro Fujihara*
Department of Management Information Science,
Fukui University of Technology, 3-6-1 Gakuen Fukui 910-8505, JAPAN

Daisuke Nogiwa
Department of Sports and Health Sciences,
Fukui University of Technology, 3-6-1 Gakuen Fukui 910-8505, JAPAN

Toshihiro Kasai
Department of Management Information Science,
Fukui University of Technology, 3-6-1 Gakuen Fukui 910-8505, JAPAN

Shota Maegawa
Graduate School of Management Information Science,
Fukui University of Technology, 3-6-1 Gakuen Fukui 910-8505, JAPAN
(Dated: January 2, 2015)
Abstract

In recent years, studies on human mobility patterns using Big data, such as mobile phone records, GPS data, geotag data of SNS, have been performed. Although studies on human mobility patterns are eagerly undertaken, there seems to be no study to investigate the possibility that human mobility patterns can be changed by weather conditions. As we can know intuitively, rainfall can affect daily human mobility patterns in general. But, it is important to explain how rainfall changes daily human mobility patterns using some statistical properties. In our project, we investigate the difference of daily human mobility motifs [12] between rainy and dry (no rain) days. To investigate the mobility patterns, we use high-resolution movement routes dataset made available for the D4D Senegal Challenge. By analyzing the dataset, we can explain with the difference of the size distribution of human mobility motifs and the concept of the center of gravity of human visitation frequency that daily human mobility motifs change by rainfall in general. We also consider a random effect model and find that visitation frequency tends to negatively depend on rainfall.
I. INTRODUCTION

With each rainy season in Senegal, estimated 100,000 Senegalese people are reportedly affected by large-scale flooding across the country [1–3]. Especially in the largest urban area in Senegal, Dakar, flows of rainwater accumulate on the inside of the area in the rainy season because the height above sea level is little more than zero meter, which also causes large-scale damage by floods. By the flooding in 2013, there was a considerable number of people who lost all of their belongings, food stocks, and even homes. In addition, farming lands were inundated and public buildings, such as houses and schools, were severely damaged, which causes declines of productivity and education. Furthermore, accumulated rainwater for long time gives off a terrible smell, and also produce large numbers of malaria vector anopheles mosquitoes. Managing rainwater flows by urban planning is a serious problem in Dakar. Therefore, this is one of important problems to solve not only for reducing damages by floods, but also for enhancing the health and quality of life in Senegal.

On the other hand, In recent years, studies on human mobility patterns using Big data, such as mobile phone records, GPS data, geotag data of SNS, have been performed by some research groups [4–7]. These studies have revealed that human mobility patterns is far from random, but is ordered. The first important finding of these studies is that human behavior is easily predictable because humans frequently go to same visited locations with high probability. This fact indicates that it is difficult to preserve human’s privacy from analyzing Big data in general. The second important finding is that humans travel long distance with high probability. Based on these recent new findings, mobility models for explaining human mobility patterns have been also studied [8–10]. These models are constructed with taking into consideration statistical properties of human mobility traces and serendipitous encounters.

Although studies on human mobility patterns are eagerly undertaken, there seems to be no study to investigate the possibility that human mobility patterns can be changed by weather conditions. As we can know intuitively, rainfall can affect daily human mobility patterns in general. For example, when it rains, humans tend to avoid going outside because someone may think that they don’t want to get their clothes or shoes wet or others are affected by a kind of psychological effect with thinking it isn’t the mood to go outside today. It is reported that there are some relationships between mood and weather condition by
analyzing Twitter dataset [11]. Furthermore, when rainwater flows accumulate to create a big rainwater puddle on the road, it can physically blocks out foot or vehicle traffic, which sometimes causes a disruption of efficient logistics networks. Therefore, it is important to understand how rainfall changes daily human mobility patterns in general.

In our project, for the first step toward the goal of managing rainwater flows by urban planning, we investigate the difference of daily human mobility motifs [12] between rainy and dry (no rain) days. If we can develop some method to automatically detect any big difference of the mobility patterns between rainy and dry days, it is useful for some applications in urban planning to estimate potential flooding locations and the efficiency of logistics networks. To investigate the mobility patterns, we use high-resolution movement routes dataset made available for the D4D Senegal Challenge. This dataset created by records of mobile phone calls and SMS contains high-resolution human mobility patterns especially in Dakar because about a quarter of the total cell towers in Senegal is densely distributed around the city of Dakar and the positions of humans are recorded as a trace information during two-week periods by detecting which cell tower is used for communication by mobile phones.

By analyzing the dataset, we find that daily human mobility motifs shrinks on rainy day compared to sunny day in general. This means that the number of visited locations decreases and the center of gravity in the sequences of visited locations becomes close to the most frequently visited place in the majority of people in Dakar. We also consider a random effect model to explain whether people visit less places by the influence of rainfall. By using Restricted Maximum Likelihood (REML) approach, we find that visitation frequency tends to decrease depending not only on the event that it is rainy today, but also on the event that it was rainy yesterday.

The rest of the paper is organized as follows. In Section II, we explain the dataset that we used and methods that we investigate to show the finding that human mobility motifs shrink because of rainfall. Section III, we show some results about the analysis of human mobility motifs. We also demonstrates a visualization image of analyzed human mobility motifs by qgis and results of the random effect model. We summarize our project, discuss future directions, and comment in Sections VI.
II. MATERIALS AND METHODS

About the dataset we use in this project

We use high-resolution movement routes dataset made available for the D4D Senegal Challenge (SET2 dataset). This dataset consists of human ID, timestamp, and cell tower ID. This 3-tuple information is recorded as a call or SMS of mobile phones occur for a 12-month period of 2013 all over in Senegal. The whole dataset is separated into 25 CSV files and for each file there are tens of millions of the tuple and the total number of human IDs reaches 320,000. Each cell tower ID is associated with location information (longitude and latitude) of the corresponding cell tower. Because cell towers are densely distributed around the city of Dakar, therefore, the location information of cell towers show which cell tower humans calling or sending SMS are close to with high resolution, but it isn’t enough to find which road or street humans are walking with having a call or sending SMS).

We extract all the weather information (including temperature, precipitation, humidity, and wind direction and speed) in 2013 observed at a weather-monitoring station at Yoff in Dakar from the webpage of Météociel [13]. To investigate the difference of daily human mobility motifs between rainy and dry days, we use the above data of precipitation and the SET2 dataset. Since Senegal has a tropical climate with a single short rain season between June and September, the amount of precipitation on most of days in a year are zero. So, we sample some rainy days in August to study the difference between rainy and dry days and also the effect between weekday and holiday. The examined August days (which is all included in a CSV file named “SET2_P16.CSV”) in this study is summarized in Table I. To segregate human mobility patterns in Dakar, we extract recorded entries with cell tower ID less than 402 from the above CSV file.

About methods we apply in our project

In our project, we execute the following three methods to investigate the difference of human mobility motifs between rainy and dry days.

(i) We firstly analyze the difference by observing the size distribution of human mobility motifs, which is defined by the difference of statistical distributions of the total number of different visited locations in the focusing period [12]. Some statistical values including mean,
TABLE I. Relationships of the examined August days in a two-week period from 8/5 to 8/18 and precipitation conditions of these days. \(^a\) It shows the focusing date. \(^b\) It shows the amount of precipitation in the focusing day.

<table>
<thead>
<tr>
<th>Weekday</th>
<th>Dry</th>
<th>Rain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>8/5 (^b) 0</td>
<td>8/12 (31)</td>
</tr>
<tr>
<td>Tuesday</td>
<td>8/6 (0)</td>
<td>8/13 (26)</td>
</tr>
<tr>
<td>Wednesday</td>
<td>8/14 (0)</td>
<td>8/7 (77)</td>
</tr>
<tr>
<td>Friday</td>
<td>8/9 (0)</td>
<td>8/16 (2)</td>
</tr>
<tr>
<td>Holiday</td>
<td>Thursday</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8/15 (0)</td>
<td>8/8 (28)</td>
</tr>
</tbody>
</table>

variance, skewness, and kurtosis are calculated to evaluate the difference.

(ii) We secondly analyze the difference to investigate the spatial property of human mobility motifs. To do this, we calculate the center of gravity in human mobility motifs, which is defined by the weighted average of visited locations. By comparing the difference of the centers of gravity between dry and rainy days and the distances from them to the most visited locations, we explore a hidden patterns in human mobility motifs. We also visualize the spatial patterns of both the centers of gravity in rainy and dry days and the most visited locations.

(iii) We finally formulate a random effect model using Restricted Maximum Likelihood (REML) approach. The detailed description of the model is given by the following equation.

\[
\log(y_{it}) = \alpha_i + \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \epsilon_{it},
\]

where the notations are listed as follows.

- \(i = 1, \ldots, 402\): place index in Dakar city only
- \(t\): time index (SET2_P16.CSV only)
- \(y_{it}\): number of persons visited in place \(i\) at time \(t\)
- \(x_{1it}\): dummy binary variable whether it is rainy (indicated by 1) or not (indicated by 0) in place \(i\) at time \(t\).
• $x_{2it}$: dummy binary variable whether it was rainy (indicated by 1) or not (indicated by 0) yesterday in place $i$ at time $t$

• $x_{3it}$: dummy binary variable whether the focusing day is a holiday (indicated by 1) or not (indicated by 0) at time $t$

• $\alpha_i$: parameter for random effect in place $i$

• $\beta_0$: parameter for intercept

• $\beta_1, \beta_2, \text{and} \beta_3$: parameter for explanatory variables of $x_{1it}$, $x_{2it}$, and $x_{3it}$

In addition, there are assumptions on probability distributions for $\alpha_i$ and $\epsilon_{it}$ as follows.

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2). \quad (2)$$

By analyzing this model, we examine how the visitation frequency of humans are correlated with the condition of rainy day and the amount of precipitation.

### III. RESULTS

We investigate the statistical property of the number of different visited locations by plotting the size distribution of the daily human mobility motifs by comparing the difference between dry and rainy days. The obtained distributions are illustrated in Figs. 1 and 2. As reported in [12], the shape of the size distribution of the daily human mobility motifs $f(N)$ can be approximated by a lognormal distribution described by

$$f(N) \sim \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln N - \mu)^2}{2\sigma^2}\right), \quad (3)$$

where $\mu$ is the scale parameter and $\sigma$ is the shape parameter. As shown in Fig. 1, the shape of the size distribution $f(N)$ doesn’t seem to change by the precipitation condition. We also calculate some statistical values including mean, variance, skewness, and kurtosis to compare them. The calculated values are summarized in Table II. As shown in Table II, mean, variance, skewness, and kurtosis of rainy day with both weekday and holiday increase a little compared to those of dry day. The most increased statistical value is kurtosis. This is reasonable because as shown in Figs. 1 and 2 the height of the size distribution at the peak clearly increase, which indicates that kurtosis seems to be the best estimator of the four to
distinguish rainy day from dry day. However, there seems to be little difference between the precipitation condition as we observed the probability distribution function of the size distribution.

The normalized distribution function of the difference of the number of visited locations between dry and rainy days, that is for \((N_{dry} - N_{rain})/(N_{dry} + N_{rain})\) is also illustrated in Fig. 3. In the plot of the complementary cumulative distribution function, it is easily found that the majority of the difference distribution is biased to the positive side, although there
TABLE II. Scale and shape parameters of the size distributions of human mobility motifs $f(N)$ in dry and rainy days with weekday and holiday, and their mean, variance, skewness, and kurtosis.

<table>
<thead>
<tr>
<th></th>
<th>Weekday</th>
<th></th>
<th>Holiday</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dry</td>
<td>Rainy</td>
<td>Dry</td>
<td>Rainy</td>
</tr>
<tr>
<td>scale</td>
<td>1.655395</td>
<td>1.579093</td>
<td>0.820745</td>
<td>0.891451</td>
</tr>
<tr>
<td>shape</td>
<td>0.779740</td>
<td>0.790790</td>
<td>0.704617</td>
<td>0.722916</td>
</tr>
<tr>
<td>mean</td>
<td>1.355265</td>
<td>1.367077</td>
<td>1.281771</td>
<td>1.298622</td>
</tr>
<tr>
<td>variance</td>
<td>1.536886</td>
<td>1.623885</td>
<td>1.056304</td>
<td>1.157589</td>
</tr>
<tr>
<td>skewness</td>
<td>3.509614</td>
<td>3.606385</td>
<td>2.921028</td>
<td>3.054212</td>
</tr>
<tr>
<td>kurtosis</td>
<td>27.895265</td>
<td>29.733207</td>
<td>18.252982</td>
<td>20.212776</td>
</tr>
</tbody>
</table>

FIG. 3. The normalized distribution function of the difference of the number of visited locations between dry and rainy days for $(N_{dry} - N_{rain})/(N_{dry} + N_{rain})$. The left is the probability distribution function and the right is the complementary cumulative distribution function.

is a considerable probability in the negative side. It should be noted that this bias seems to disappear into random noise when the distribution is plotted with accumulated data of some weeks.

Next, we define the center of gravity of human mobility motif to investigate the spatial difference between dry and rainy days. The center of gravity of human mobility motif is defined by the weighted sum of visitation locations. It is generally different between them.
FIG. 4. The spatial difference of the center of gravity of human mobility motif between dry and rainy days. The black lines indicate the coastline of Dakar and areas, and also the blue circles show the locations of the cell tower. The green marked lines indicate the distance between the two centers of gravity. The red lines indicate the distance between the most visited location and the center of gravity with dry day. The blue lines indicate the distance between the most visited location and the center of gravity with rainy day.

even though the most visited location is the same. We visualize the difference of the center of gravity of human mobility motif using a open source GIS called QGIS in Fig. 4. It is difficult to see any clear pattern in Fig. 4, but the complementary cumulative distribution function of the spatial difference can be shown a pattern. The distribution is illustrated in Fig. 5. As shown in Fig. 5, the difference shifts to the positive side, especially around the higher difference values.

Finally, we consider a random effect model with REML approach to examine how human’s visitation frequency is correlated with the precipitation condition and the amount of precipitation. The final result is calculated and the covariance parameter of the model becomes 0.3819, which means that the model has a good agreement with the dataset. More
detailed results are shown in Table III. As shown in Table III, today’s and yesterday’s rainfall is negatively correlated with human’s visited population. This concludes that rainfall does influence daily human mobility motifs.

IV. DISCUSSIONS

We investigated how rainfall affects the human mobility motifs using Big data made available by D4D challenge and weather information in Dakar. By analyzing the difference

| Parameter          | Estimate value | Std. Dev. | Degree of Freedom | t-value | Pr > |t| |
|--------------------|----------------|-----------|-------------------|---------|------|---|
| Intercept: $\beta_0$ | 6.7925         | 0.03198   | 383               | 212.39  | <0.0001 |
| Rain (Today)$^a$: $\beta_1$ | -0.02497 | 0.008712  | 4971              | -2.87   | 0.0042 |
| Rain (Yesterday)$^b$: $\beta_2$ | -0.05778 | 0.01001   | 4971              | -5.77   | <0.0001 |
| Holiday$^c$: $\beta_3$ | 0.01019        | 0.008712  | 4971              | 1.17    | 0.2421 |

$^a$ 1 if it rains today 0 otherwise  
$^b$ 1 if it rains yesterday 0 otherwise  
$^c$ 1 if it is a holiday 0 otherwise
of the size of the motifs and the spatial difference of the centers of gravity, we find some
evidences that rainfall can change the details of daily human mobility motifs. Moreover, we
also consider a random effect model with REML approach to show that human’s visitation
frequency is correlated with the precipitation condition and the amount of precipitation.

As I mentioned before, there are some possible reasons that rainfall changes the patterns,
that is, mainly categorized into physical and psychological reasons. The segregation of the
two reasons is at this time difficult because of lack of information like traffic conditions,
relationship between altitude and rainwater puddle formation, and more detailed human
mobility data who walks which road when. If more detailed relationship between rainfall
and human mobility patterns can be found, we might inversely estimate where it rains from
the difference of the size distribution and the center of gravity of some humans living in
areas. This could become a distributed weather-monitoring system available with low cost
without deploying any weather-monitoring station.

At this time, because we couldn’t obtain information of altitude and the area of roofs
in Senegal, we couldn’t propose a location planning of low-cost rainwater harvesting tanks
[14, 15] based on the estimation of the places where big rainwater puddles easily are formed
in flooding. If Synthetic CDR which is unavailable this time in D4D Challenge have more
detailed information of human mobility patterns, it might be possible to observe much clearer
difference of human mobility patterns between dry and rainy days. If so, the method for
detecting the difference would be useful for urban planning to estimate flood damage and
to avoid building houses in areas of potential flooding, and to propose cost-effectiveness of
the deployment of low-cost rainwater harvesting tanks for the development of Senegal.

ACKNOWLEDGMENT

Studies and Researches were performed using mobile communication data made available
by SONATEL and Orange within the D4D Challenges. The authors thank SONATEL and
Orange for giving us the opportunity to use a wonderful big dataset to proceed this project.
This work is partially supported by Japan Society for the Promotion of Science through
Grant-in-Aid for Grant-in-Aid for Young Scientists (B) (25870958).


TEXT ME - IF YOU CAN:
LITERACY, NETWORKS AND MOBILITY IN DAKAR

Martina Kirchberger*
Christopher Small

Columbia University

31 December 2014

ABSTRACT

We analyze mobility data from calls and text messages from 3 million users in the Dakar region over the entire year 2013. We show that that individuals residing closer to the central business district of Dakar are not only geographically more centrally located but also have a higher number of contacts, communicate from a larger number of towers, and have a higher entropy of contacts (relatively more weaker ties). We then investigate the use of the text to call ratio as a proxy for literacy. We show that there is a difference between individuals residing in areas where calls dominate relative to texts: over a substantial range of distances from the CBD, individuals have fewer contacts they text and call with, and they have fewer weak ties. Finally, we do not find evidence that individuals who are residing in these areas compensate the lack of mobile phone activity with a higher level of mobility.

Keywords: networks, mobile communication, mobility.

JEL Classification: D8, J6, O1.

*Correspondence: 475 Riverside Drive, Suite 253, New York, NY 10115, UK; Email: mk3759@columbia.edu.
1 Introduction

One of the main reasons for cities to exist is to connect people with each other. Networks are important for disseminating labor market information, such as job openings and candidates (Jackson 2009). Weak ties, contacts with whom we interact infrequently, have been found to be important determinants of labor market outcomes (Granovetter 1973). Helsley and Zenou (2014) show that in equilibrium agents who are more central in the network tend to locate closer to the geographical center. While we have a substantial amount of information on urban residents in developed countries, cities in developing countries are far less well understood and this is particularly true for African cities. It is likely that networks matter even more in contexts where information asymmetries are more abundant and labor markets possibly less transparent. We propose to use mobile phone communications data to understand how Dakar’s residents communicate with each other how social and geographical distance are related.

Until recently, the data to answer such questions was not available at the necessary spatio-temporal scale. Surveys at the city level often do not contain enough observations to estimate consistent averages across space, and they are not able to record behavior with the frequency required to understand complete networks over time. The data analyzed in this note allows us to overcome these difficulties. We have a large number of individuals (about 7 million in total) spread out over an entire year. The timing of mobile communications and their trajectories across space allow us to determine their home location. We then use behavioral indicators to understand how networks, communication patterns and mobility vary across space. Proposing a novel proxy for literacy, the text to call ratio of a particular location, which we validate at the national level, we show that these patterns differ for individuals depending on the level of texts to calls in their home location. Individuals who reside in areas where text to call ratios are low have a lower number of contacts, and fewer weak ties. They also do not seem to be compensating by higher individual level mobility.

There are at least two potential application of this analysis: first, using readily available mobile phone data could help policy makers to identify areas across space where literacy rates are lower and at the same time which tend to be less integrated into networks to promote the dissemination of labor market information in these locations; second, the provision of information
via text messages is increasingly used as a means to disseminate information (Fafchamps and Minten 2012). Having data on mean text to call ratios in particular areas could be an important source of information on literacy and consequently whether text messages or calls are the more appropriate method of dissemination of information.

This research relates to a recent literature in economics providing evidence on the link between mobile phone usage and economic development in Africa (Aker and Mbiti 2010). More specifically, this includes studies on the adoption of network goods (Björkegren 2014), prices obtained by farmers and fishermen (Aker and Fafchamps 2014; Jensen 2007), risk sharing (Blumenstock et al. 2014; Jack and Suri 2014) and violence (Shapiro and Weidmann 2013).

This note is structured as follows: section 2 presents the data and methodology; section 3 presents the main results; 4 concludes.

2 Data and Methodology

In the context of the Data for Development Challenge, Orange Senegal has made available three datasets: (i) antenna-to-antenna traffic for 1666 antennas on an hourly basis, (ii) tower level mobility data on a rolling 2-week basis for a year with behavioral indicators about 300,000 randomly sampled users, and (iii) one year of arrondissement level mobility data with monthly behavioral indicators for about 150,000 randomly sampled users. To understand the link between individuals’ residence, communication patterns and mobility in the Dakar region we mainly use the first dataset provided as part of the Orange Data for Development Challenge. Due to its size, the dataset it is split among 25 distinct datasets, each having between 40 and 50 million rows. Each row contains the anonymized user id of the individual, time stamp of the communication, and the tower the individual connected to during that communication.

We define an individual’s home location as the tower that is recorded most often between 7pm and 7am in this two week window. We then compute the travel distance of the individuals’ “home tower” to the Central Business District (CBD) of Dakar. To do this, we extract the Open StreetMap extents for Dakar, build a network dataset and use the Origin Destination Cost Matrix tool in ArcGIS 10.2.2 to compute the network distances between Dakar’s CBD at a longitude of 17.44667 and a latitude of 14.69278.
these two locations. As features are snapped to the nearest location on the network, we define the total distance to the CBD as the travel distance on the network to the center of Dakar plus the straight line distance of the tower to the nearest link on the network.\textsuperscript{2} Figure 1 shows the distance from the city center, illustrating the importance of using the distance on the network rather than straight line distances between towers and the CBD.

Figure 1: Distance from Dakar’s CBD

Notes: Warmer colors represent higher distances from the city center of Dakar. The graph highlights the importance of using the distance on the network rather than the straight line distances for cities that do not have a circular shape.

We also compute the total distance travelled by our individuals in the sample by computing the Origin Destination Cost Matrix for each pair of towers located within the Greater Dakar region, again adding the snapping distance to the origin and destination tower. Whenever we have a communication record from a tower that is different to tower the individual was connected during the previous communication, we count the movement as distance travelled. Finally, we combine our data with a range of behavioral indicators for each individual provided by Orange such as the number of con-

\textsuperscript{2}Three towers (13, 289 and 333) are located close to non-transversable network element positions so we match them with the closest tower in the dataset, adding the distance in the direction of the city center; this affects about 3,000 individuals. We have 2,162 individuals (or 0.07\% of the sample) who don’t undertake any mobile communication between 7am and 7pm, so we exclude them from the sample.
tacts and Shannon’s entropy of calls and texts. This means that we end up with a dataset of about 3 million individuals over the course of 2013.

For the generalizability of the results it is important to keep in mind that our sample is selected along several dimensions: first, individuals have to use a mobile phone to be in our sample; second, we only have individuals who chose Orange as their network provider; third, Orange additionally selected users based on two criteria: (i) users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset), and (ii) users having had an average of less than 1000 interactions per week (users with more than 1000 interactions per week were presumed to be machines or shared phones); finally, in this note we focus on individuals residing in the Dakar region.\(^3\)

### 3 Results

Table 1 shows some basic descriptive statistics of the sample. Our restriction to only use individuals residing in Greater Dakar means that we include about 38% of individuals of the sample.\(^4\) Individuals are active on average during 10 days of our 14 day window, with a mean of about 170 connections to towers (so roughly 17 per active day), and connecting to about 14 distinct

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion Dakar</td>
<td>0.38</td>
<td>0.03</td>
<td>0.29</td>
<td>0.42</td>
<td>2914369</td>
</tr>
<tr>
<td>Active days (all)</td>
<td>10.39</td>
<td>3.24</td>
<td>0</td>
<td>14</td>
<td>2914369</td>
</tr>
<tr>
<td>Activities recorded</td>
<td>173.74</td>
<td>212.95</td>
<td>10</td>
<td>2000</td>
<td>2914369</td>
</tr>
<tr>
<td>Distinct towers</td>
<td>14.34</td>
<td>12.89</td>
<td>2</td>
<td>230</td>
<td>2914369</td>
</tr>
<tr>
<td>Total individual distance</td>
<td>145.14</td>
<td>165.58</td>
<td>0.17</td>
<td>12158.46</td>
<td>2914369</td>
</tr>
<tr>
<td>Distance to the CBD</td>
<td>10.23</td>
<td>6.71</td>
<td>0.25</td>
<td>29.78</td>
<td>2914369</td>
</tr>
<tr>
<td>Entropy of contacts (text)</td>
<td>1.45</td>
<td>0.94</td>
<td>0</td>
<td>5.91</td>
<td>2551774</td>
</tr>
<tr>
<td>Entropy of contacts (call)</td>
<td>2.83</td>
<td>0.65</td>
<td>0</td>
<td>6.31</td>
<td>2914099</td>
</tr>
<tr>
<td>Entropy of places (all)</td>
<td>1.4</td>
<td>0.65</td>
<td>0</td>
<td>4.89</td>
<td>2914369</td>
</tr>
<tr>
<td>Number of contacts (call)</td>
<td>33.65</td>
<td>25.84</td>
<td>1</td>
<td>580</td>
<td>2914099</td>
</tr>
<tr>
<td>Number of contacts (text)</td>
<td>11.23</td>
<td>14.52</td>
<td>1</td>
<td>475</td>
<td>2551774</td>
</tr>
</tbody>
</table>

Notes: Summary statistics based on distinct users in Set 2 dataset who remained in the Dakar Region for the respective 2 week windows, residing at a distance of 30km or less from Dakar’s center.

\(^3\)We exclude individuals residing at a distance of more than 30km from the city center as only 1.4% of our sample is living there and there are only 15 towers which makes it difficult to interpret the averages for different distances.

\(^4\)The analysis of call and text volumes in our companion paper shows that these 38% account for 62% of calls and 82% texts.
towers in the 2 week period. Individuals traveled on average 145km during the two week period, and reside about 10km away from Dakar's CBD. The entropy index of calls is significantly higher compared to texts and entropy of places. This indicates that the predictability of contacts in texting is higher than the predictability of contacts in calls. In our analysis we interpret the entropy measure as a measure of weak ties. The entropy measure will be high when users are in touch with a large number of distinct users, such that an additional user provides a high information content. We also look at the number of different contacts which does not take into account the frequency of interaction. Users called on average 34 different contacts, and texted 11 different contacts in the 2 week window.

Figure 2 shows individual's behavior and distance from the city center of Dakar as a kernel-weighted local polynomial regression with 95 percent confidence intervals marked by the vertical bars. The graphs show a remarkable drop-off in all the variables measuring mobile phone activity and network size as one moves away from the city center. Individuals contact fewer numbers, have fewer communications, and have fewer weak ties (lower entropy of calls and texts). A central question surrounding urbanization in Africa is whether the poor are able to take advantage of the benefits of cities. Ideally,
we therefore would like to investigate whether this pattern of higher connectivity and larger networks is the same for the rich and the poor. A drawback of the current dataset is that we do not have any background information on the users in our sample such as their age, education, income or occupation. However, we know a lot about their calling and texting behavior. While calling and texting seem to provide similar information at first sight (individuals using their phones), there is one key difference between sending a text versus calling a person: the former requires one to be literate, while the other does not. This matters in a context where literacy rate is about 35% at the national level and 65% for the Dakar region (DHS 2013).

To see whether the text to ratio is correlated with literacy rates we briefly turn to the SET1 data and compute the average text to call ratio of the entire year for each of the towers. We then draw Voronoi cells around each tower and spatially link the data with the location of survey clusters of the 2010 Demographic and Health Survey (DHS), which contains information on 20,617 males and females. About 29% of the individuals in the DHS sample are able to read the whole sentence presented in the interview. Since DHS clusters are offset by random amounts (0-2km for urban clusters, 0-5km for rural clusters, 0-10km for a random 1% sample or rural clusters), we draw a 5km buffer around the cluster. We then determine the total area of overlap of any of the DHS cluster buffers and the Voronoi cell, and allocate these parts in proportion of the area that is overlapping. When multiple buffers intersect a Voronoi cell over the same location, we allocate the proportions equally among these cells. Finally, we compute the mean of males and females in the sample who are able to read an entire sentence in every cluster. Figure 6 in the Appendix shows a scatterplot and local polynomial regression indicating that indeed it appears that the text to call ratio is increasing with literacy.

Under the assumption that this ratio indeed contains information on the socio-economic characteristics of the underlying population, we next explore whether there are differences in the behavior of our urban sample along this dimension by splitting along the median text to call ratio of one. Figure 7 in the Appendix shows the distribution of call to text ratios in our sample and 3 shows its spatial distribution.\textsuperscript{5}

\textsuperscript{5}The correlation between the text to call ratio from the usable (as defined in our companion paper) tower level aggregates from SET1 and the residence based individual level tower aggregates from SET2 is about 0.24 and is statistically different from zero at the 1 percent level.
Figure 3: Map of tower level text to call ratio

Legend
Text to call ratio
- 0.4 - 0.8
- 0.8 - 1.0
- 1.0 - 1.4
- 1.4 - 1.9
- 1.9 - 3.2
- CDB Dakar

Note: Text to call ratios in Dakar. Warmer colors represent higher levels of text to call ratios.

Figure 4: Number of contacts and entropy by text to call ratio

Notes: The figure shows individual’s behavior and distance from the city center of Dakar as a kernel-weighted local polynomial regression with 95 percent confidence intervals marked by the vertical bars, split along the median text to call ratio.

Figure 4 splits the sample into individuals residing in areas which are
above and below the median text to call ratios. The panels show that individuals residing in areas where calls dominate over texts, over a substantial spatial distance from Dakar, are in touch with a significantly lower number of contacts. So even though they call relatively more compared to texting, they call fewer people and have fewer weak ties over a substantial range of the distribution. They also do not only text less compared to calling, they text fewer people.

Finally, we ask whether poorer individuals compensate for the lower number of contacts and weak ties by higher mobility.

Figure 5: Total distance travelled

![Mobility graph](image)

Notes: The figure shows the total distance traveled as a function of the distance from the city center of Dakar as a kernel-weighted local polynomial regression with 95 percent confidence intervals marked by the vertical bars, split along the median text to call ratio.

Figure 5 plots the total distance travelled illustrating quite the opposite pattern. Up to 25km from the city center, individuals living in areas where texts are more prevalent than calls travel significantly more over the two-week window than those living in areas where calls are more prevalent than texts. Thus, there is no evidence that individuals compensate for the lower level of mobile phone usage by a higher level of spatial mobility.
4 Conclusion

This note illustrated the use of call and text volumes as well as mobility patterns to understand the level of integration of individuals into networks. We find that individuals residing closer to the central business district of Dakar are not only geographically closer but also have a higher number of contacts, call from a larger number of towers, and have a higher entropy of contacts. We then use the text to call ratio as a proxy for literacy. We show that there is a gap for individuals residing in areas where calls dominate relative to texts: these individuals have fewer contacts they text and call with, and they have fewer weak ties. Finally, we do not find evidence that individuals who are residing in these areas compensate the lack of social integration with a higher level of mobility.
References


A Appendix

Figure 6: Text to call ratio and average literacy rate

Notes: Kernel-weighted local polynomial regression with 95 percent confidence intervals of the mean literacy rate in a tower's Voronoi cell and the text to call ratio of a tower and scatterplot of the raw data. The graph shows that there is a positive relationship between mean literacy rate and the text to call ratio of a tower.
Notes: Average text to call ratio for each of the towers located 30km or less from the city center of Dakar. We first compute individual level text to call ratios and then aggregate at the level of the tower.
Cellphone’s Users Trajectories Clustering from Call Detail Records

Paulin Melatagia Yonta\textsuperscript{1,2}, Blaise Ngonmang\textsuperscript{1,3}, Romaric Meleu\textsuperscript{1,4}, Vanessa Kamga\textsuperscript{1}, Armel Nzekon\textsuperscript{1}, Claude Tinku\textsuperscript{1}

\textsuperscript{1} LIRIMA-IDASCO, Université de Yaoundé I, Faculté des Sciences, Département d’Informatique, B.P. 812 Yaoundé, Cameroun

\textsuperscript{2} IRD, UMI 209, UMMISCO, 32 avenue Henri Varagnat, F-93143, Bondy, France

\textsuperscript{3} Université Paris 13, Sorbonne-Paris-Cité, L2TI (EA3043), F-93430, Villetaneuse, France

\textsuperscript{4} Université de Ngaoundéré, Faculté des Sciences, Département de Mathématique-Informatique, B.P. 454, Ngaoundéré, Cameroun

E-mail: paulinyonta@gmail.com, blaise.ngonmang@univ-paris13.fr, meleu.romaric@gmail.com, ansylvania.kamga@gmail.com, armeljanz@gmail.com, tinkuclaude@gmail.com

Abstract. In this paper we propose a $k$-medoid like algorithm to cluster the trajectories of cellphone’s users. The trajectories are deduced from Call Detail Records and are clustered to generate spatio-temporal trajectories which allows to browse the maximum locations of each cluster in accordance with their timestamps. The clustering algorithm used an new distance function based on sequences alignment and called Alignment distance with Threshold Penalty. We have conducted some experimentations on D4D Senegal Challenge datasets which highlighted three classes of cluster’s centroids : daily trip centroids, circular trajectories centroids and short trajectories centroids. These results, combined to the the spatial-temporal representation of the centroids can be used to improve transportation management in cities and villages.

Keywords : Trajectories Clustering, CDR, K-medoid, Transportation Management, Alignment distance with Threshold Penalty, D4D Senegal Challenge.

1. Introduction

The extensive generation of telecommunication call detail records boost the requests of a diverse range of services that involve the exploit of knowledge stored in these data. This has posed new challenges in datamining. Among the datamining tools, clustering analysis, which groups similar patterns to reveal overall distribution and interesting correlations in datasets, has a number of applications in data compression, image processing, pattern recognition, and market research. In this paper we study the problem of clustering the trajectories of cellphone’s users generated from theirs Call Detail Records (CDR) of phone calls and text exchanges. The goal of this work is to unveil some interesting and important patterns about how cellphone’s users move. These patterns can be useful for governments especially for public transportation management and urban planning. They can also be used by the transport companies, taxi companies to improve their services. Indeed, one can use the accepted cluster centroids, obtained as the output of our analysis, as the movement profiles patterns of the customers.
Trajectories of mobile phone’s users generated from CDR are spatio-temporal data that indicate the successive positions of a user. The spatial component relates to the ID or the geographical coordinates of the site where the transmitting BTSs, the mobile network antennas, are located and temporal aspect is the timestamp at which the user make a phone call or send a text message.

Many work have been done in trajectory clustering [9, 5, 6, 8] but few of them use CDR of phone calls and text exchanges as input data. According to [8], Reades et al.[11] were one of the first who attempted to analyse urban dynamics on a city level using Erlang data. Erlang data is a measure of network bandwidth and indicate the load of cellular antenna as an average number of calls made over specific time period (usually hour). Kisilevich et al. [8] propose a classification of trajectory clustering methods. The three main groups are : model-based, distance based and visual aided. Model-based methods attempt to describe the whole dataset by generating a suitable function of time. Distance-based methods use specially designed distance functions that are meant to show similarity between objects and visual-aided methods rely on human expert’s judgement, who will interactively change clustering settings to achieve the desired clustering result. In our work we will use the distance based approach for our findings; this allows breaking the whole trajectory clustering process into two steps: 1) calculation of distances between trajectories according to the defined distance function and 2) actual clustering using a known clustering algorithm. Compared to existing algorithms, we make the following changes in this work: 1) we use a compact representation of a spatio-temporal trajectory by replacing a succession of positions on the same site by a single one 2) we define a new similarity measure based on sequences alignment that allow the discovery of recurring trajectory patterns 3) we define a new formula to calculate cluster’s centroids.

This paper is structured as follows. Next, we formally describe the clustering trajectory problem by giving some definitions. Section 3 details each of the steps that comprise our proposed algorithm. In Section 4 the experimentation are and preliminary results are presented. Some conclusions are offered in Section 5.

2. Definitions and Problem Statement

The CDR-based trajectories clustering problem can be formulated as follows. The movement of a user $u$ is tracked as a sequence $S(u)$ of spatial locations, each associated to a timestamp, of the form $R = ((l_0, t_0), (l_1, t_1), ...)$, where $l_i$ is the user’s location at time $t_i$ ($\forall i \geq 0, t_i \leq t_{i+1}$). The location can be the geographical coordinates of a site or its ID. Let $[L](R)$ denote the sequence of the locations of a trajectory $R$, i.e. $[L](R) = (l_0, l_1, ...)$; $L$ is the set of all possibles locations. The goal here is to partition the monitoring area $A$ with is the set of all couples $(u, S(u))$ for the monitored users into clusters $C_1, C_2, ..., C_k$ such that all the trajectories of a cluster can be traversed by the same object (a user, a taxi, a bus, ...). The monitored history delay for trajectories can be one day, one month or any fixed duration. In this work, we consider daily trajectories. Note that the original dataset $D$ has to be transformed to a new one, $D'$, according to the chosen monitored history delay; the same user $u$ may appears in many couple $(u, S'_d(u))$ (where $S'_d(u)$ is the sequence of spatio-temporal CDRs of user $u$ for the date $d$).

**Definition 1.** Let $R = R_u = ((l_1, t_1), (l_2, t_2), ...)$ and $R' = R'_u = ((l'_1, t'_1), (l'_2, t'_2), ...)$ be two trajectories of length $n$ and $m$ respectively. Let $A^*$ the monitoring area with spatio-temporal tuples of the form $(-, 0)$, i.e. a spatial location ‘-’ called gap, is added to the set of locations. An alignment is a function $f : A \times A \rightarrow A^* \times A^*$ such that $f(R, R')$ is a global alignment of sequences $[L](R)$ and $[L](R')$ where every location is replaced by the corresponding spatio-temporal data. The definition of
the global alignment considered here is the same as that of the Needleman-Wunsch [10].

**Definition 2.** An alignment is valid for the CDRs trajectory clustering if it verifies the following properties; let \( R = R_n = ((l_1, t_1), (l_2, t_2), \ldots) \) and \( R' = R'_m = ((l'_1, t'_1), (l'_2, t'_2), \ldots) \) be two trajectories of length \( n \) and \( m \) respectively, for \( 1 \leq i \leq n \) and \( 1 \leq j \leq m \):

- only two similar sites can be aligned, i.e. \( (l_i, t_i) \) and \( (l'_j, t'_j) \) are aligned iff \( l_i = l'_j \)
- \( (l_i, t_i) \) is aligned before \( (l'_j, t'_j) \) iff \( t_i \leq t'_j \)
- \( (l'_j, t'_j) \) is aligned before \( (l_i, t_i) \) iff \( t'_j \leq t_i \)

When the alignment of two trajectories is valid, they can be merge according this alignment to form a new trajectory. The merging of two alignments produces an alignment which is consistent with the two inputs; i.e. if three people performs separately the three trajectories, they will move forward together with high probability.

**Definition 3.** Let \( R = R_n = ((l_1, t_1), (l_2, t_2), \ldots) \) and \( R' = R'_m = ((l'_1, t'_1), (l'_2, t'_2), \ldots) \) be two trajectories of length \( n \) and \( m \) respectively and \( f(R, R') \) a valid alignment. The merging of \( f(R, R') \) produces a trajectory from the alignment by performing the following operations:

- if \( (l_i, t_i) \) is aligned with \( (l'_i, t'_i) \), merging them produces \( (l_i, (t_i + t'_i)/2) \)
- if \( (l_i, t_i) \) is aligned with \( (-, 0) \), merging them produces \( (l_i, t_i) \)

The merging of an alignment is valid if the timestamps of the output trajectory are strictly increasing.

3. Proposed algorithm

Our algorithm finds the centroids of each cluster, where the coordinate of each centroid is a sort of mean of the coordinates of the objects in the cluster. This centroid is the movement profile pattern of the trajectories of the cluster. However instead of deriving our algorithm from \( k \)-means, we proposed a \( k \)-medoids like clustering algorithm. Indeed, \( k \)-means clustering is sensitive to the outliers and a set of objects closest to a centroid may be empty, in which case centroids cannot be updated. In \( k \)-medoids, representative data points called medoids are considered as centers of clusters instead of centroids. Among many algorithms for \( k \)-medoids clustering, Partitioning Around Medoids (PAM) proposed by Kaufman and Rousseeuw [7] is known to be most powerful.

In our algorithm, we define a new distance function for dissimilarity measure between trajectories and a last step is added to the PAM algorithm to compute the centroids of the clusters.

3.1. The Distance Function

Dynamic Time Warping (DTW) [1], Longest Common Subsequence (LCSS) [12], Edit Distance on Real sequence (EDR) [4] and Edit distance with Real Penalty (ERP) [3] are the most used distance functions for trajectory clustering. For our work LCSS and EDR are not considered because they do not take into account the time differences between locations. Moreover LCSS focuses only on the matched parts and ignores all the unmatched portions; this is not convenient for our study
Penalty (A TP). Let $T$ define the following distance function called Alignment distance with Threshold $\theta$

evaluate the distance between each aligned pair of sites. With a valid alignment, we
calculate the discrepancy and added to the penalty. Our hypothesis is that a good
not used because mapping between all objects in both trajectories means every dis-
crepancy is considered and added to the penalty. Our hypothesis is that a good
distance function for the studied problem must align the two trajectories and evaluate
the distance between each aligned pair of sites. With a valid alignment, we
define the following distance function called Alignment distance with Threshold
Penalty (ATP). Let $T$ be the matrix $|L| \times |L|$ of the time taken to move between
two locations.

$$
ATP(R_n, R'_m) = \begin{cases} 
\sum_{i=1}^n t_i & \text{if } m = 0 \\
\sum_{i=1}^m t'_i & \text{if } n = 0 \\
\min\{ATP(R_{n-1}, R'_{m-1}) + h_\theta((l_n, t_n), (l'_m, t'_m)) , \\
ATP(R_n, R'_{m-1}) + h_\theta((-0), (l'_m, t'_m)) , \\
ATP(R_{n-1}, R'_m) + h_\theta((l_n, t_n), (-0))\} 
\end{cases}
$$

where

$h_\theta((l_i, t_i), (l'_i, t'_i)) = \begin{cases} 
0 & \text{if } d((l_i, t_i), (l'_i, t'_i)) \leq \theta \\
|t_i - t'_i| & \text{otherwise} 
\end{cases}$

and

$$
d((l_i, t_i), (l'_i, t'_i)) = \begin{cases} 
|t_i - t'_i| & \text{if } l_i, l'_i \text{ are not gaps} \\
|t_i - T[i-1][i] - (t_{i-1} + t'_{i-1})/2| & \text{if } l'_i \text{ is a gap; } l_{i-1}, l'_{i-1} \text{ not gaps} \\
|t_i - T[i-1][i] - t_{i-1}| & \text{if } l'_i, l'_{i-1} \text{ are gaps} \\
|a'_i - a_{i-1}| & \text{if } a'_i, a_{i-1} \text{ are gaps} \\
|t'_i - T[i-1][i] - (t_{i-1} + t'_{i-1})/2| & \text{if } l_i \text{ is a gap; } l_{i-1}, l'_{i-1} \text{ not gaps} \\
|t'_i - T[i-1][i] - t_{i-1}| & \text{if } l_i, l'_{i-1} \text{ are gaps} \\
|t'_i - T[i-1][i] - t'_{i-1}| & \text{if } l_i, l_{i-1} \text{ are gaps} 
\end{cases}
$$

Algorithm 1 PAM Algorithm

1: Initialize: randomly select (without replacement) $k$ of the $n$ data points as the
2: medoids
3: for each medoid $m$ do
4: for non-medoid data point $o$ do
5: Swap $m$ and $o$ and compute the total cost of the configuration
6: end for
7: end for
8: Select the configuration with the lowest cost.
9: Repeat steps 2 to 8 until there is no change in the medoid.

where two geographically closed sites can be considered as identical. DTW is also
not used because mapping between all objects in both trajectories means every dis-
crepancy is considered and added to the penalty. Our hypothesis is that a good
distance function for the studied problem must align the two trajectories and evaluate
the distance between each aligned pair of sites. With a valid alignment, we
define the following distance function called Alignment distance with Threshold
Penalty (ATP). Let $T$ be the matrix $|L| \times |L|$ of the time taken to move between
two locations.
ATP is close to ERP according Eq. (1). The difference between ATP and ERP is the distance function $h_{θ}$ which is close to one of DTW. Intuitively $h_{θ}$ is used to define a threshold below which two spatio-temporal coordinates are considered similar (distance = 0). Moreover to approximate the temporal component of a gap at position $i$, we use the mean of the timestamps of the coordinates at position $i - 1$ and the distance from location at position $i - 1$ to location at position $i$ in the alignment. Since we use only valid alignments, only the following configurations may occur, either $l_i = l'_i$ or $l_i = l'_i$ and $l_i = -$ or $l'_i = -$. 

3.2. Cluster’s centroid calculation

While using a $k$-medoids algorithm for the clustering, it is important at the end of the algorithm to determine the centroid of each cluster. A centroid of a trajectories cluster should represent all trajectories that belong to that cluster in some summarized manner and can be seen as the spatio-temporal trajectory which allows to browse the maximum locations of the cluster in accordance with the timestamps at these locations. We propose the below algorithm for the centroid calculation. The algorithm start by sorting the trajectories in descending order of their length. The goal of this step is to maximize the number of trajectories that will be merge to the medoid to obtain the centroid. Step 2 of the algorithm set the first value of the centroid to the medoid. In the for loop, each trajectory is aligned with the current centroid and merged. If the merging is valid according to Definition 3, the resulting trajectory is set as the new centroid.

Algorithm 2 Calculation of the centroid

Require: $p$ trajectories $R^i$ ($1 \leq i \leq p$) that are in the same cluster $C$ and the trajectory $R^{i*}$ which is the medoid of $C$.

Ensure: The centroid $R^{ic}$ of the cluster $C$.

1: Sort the trajectories $R^i$ in descending order of their length. Let $σ$ be the sorting order.
2: $R^{ic} ← R^{i*}$
3: for all $i$ from 1 to $p$ do
4:  $R ←$ merging $R^{ic}$ and $R^{i(1)}$ according to a valid alignment.
5:  if $R$ is valid then
6:    $R^{ic} ← R$
7:  end if
8: end for

4. Experimentations

4.1. The D4D Senegal Challenge datasets

D4D Senegal Challenge datasets are generated from the Call Detail Records (CDR) of the customers of Sonatel collected for a year, from January 1 to December 31, 2013. There is three datasets: Dataset 1 is a one year of site-to-site traffic for 1666 sites on an hourly basis, Dataset 2 is a fine-grained mobility data (site level) on a rolling 2-week basis with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled users for each 2 week period and Dataset 3 is a one year of coarse-grained (123 arrondissement level) mobility data with bandicoot behavioral indicators at individual level for about 150,000 randomly sampled users for a year. In the dataset, only users meeting both of the following criteria where retained:
1. users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset).

2. users having had an average of less than 1000 interactions per week. The users with more than 1000 interactions per week were presumed to be machines or shared phones.

Each dataset has been designed to balance utility with privacy, utility being the research that can be done with the data while privacy is the potential risk of re-identification of users.

4.2. Data preprocessing

For the trajectories clustering, we have used the Dataset 2. The algorithms presented in this paper for site level can also be applied to arrondissement level after a suitable preprocessing of Dataset 3.

The files SET2_P01.csv through SET2_P25.csv of Dataset 2 contain the user_id, timestamp, and site_id for each of the 25 two-week periods. For the clustering task, we have transform each file so that each line of the new file represents a compact daily trajectory of a user in the form:

```
user_id,site_id_1,timestamp_1,site_id_2,timestamp_2,...
```

The trajectory is said compact because every sequence of spatio-temporal coordinates with the same site_id: site_id,timestamp_1, site_id,timestamp_2,...,site_id,timestamp_q was replaced by site_id,timestamp_1, site_id,timestamp_q, i.e. only the first and the last coordinates are kept in the output sequence. The matrix $T$ of the means of the estimated time taken to move between locations is constructed during this preprocessing step. This matrix is the adjacency matrix of the network of movements of cellphone’s users (the movements network for short). Note that the user_id is not used by our clustering algorithm. To analyse these trajectories and provide representatives trajectories for each cluster, we rely on the time dilation assumption that states that similar trends might take place over different periods of time; so we do not consider the date part of the timestamps (YYYY-MM-DD).

4.3. Structure of the movements networks

We found that degree distribution of the movements networks are bimodal with high transitivity (a mean of 0.705 for the 25 files) and high average clustering coefficient (a mean of 0.732 for the 25 files). Bimodal degree distribution is characteristic of complex networks which contain two types of nodes. Nodes in the distribution with the higher mode are called hubs, those in the other distribution peers. That is, as shown in Figure 1(a) the movements network deduced from CDR of files SET2_P04 contains hubs sites of degree $\approx 900$ (sites of ID 3, 6, 22, 34 for example) and peers of degree $\approx 200$ (sites of ID 980, 888, 832 for example). An observation of the types of the activities around these sites can help to link these hubs and these peers to areas from which or to which people move a lot (for example markets or administrative areas for hubs and residential areas for peers).

Figure 1(b) and Figure 1(c) that represent the degree distribution of a network of call and text messages between sites (obtained from Dataset 1) show that from most of the sites, the cellphone’s users call or send messages to all other sites. These networks are dense as shown by the first line of Figure 2(b) and (c). From the same figures we can also observe that the greater the ID of a site is, the less the site interacts with others; the adjacency matrices are sparse for high ID. Moreover
Figure 1: Degree distribution of (a) a movements network (b) a networks of calls (c) a network of text messages

Figure 2: Adjacency matrices before and after communities detection (a) and (d) the movements network of SET2_P02 (b) and (e) the network of calls of SET1V_02 (c) and (f) the network of text messages of SET1S_03

the sites whose ID are > 600 are less connected to the other sites in the movements networks.

We have used the Louvain algorithm [2] to extract the communities of theses networks and the main observations are the following:

- Among the 1666 sites, there are 166 isolated ones which do not interact (no call and no message) with the other.

- In the movements networks, there are two dense clusters. The first one (represented by the first blue square in Figure 2(d)) contains ≈ 500 sites. These sites are mostly those of ID < 600. It’s the cluster of hubs sites. The second dense cluster identified by the Louvain algorithm is formed by ≈ 250 sites (the smallest blue square in Figure 2(d). The ID of the sites of this cluster are in the range [600, 1050] and they are quite well connected to the sites of the hubs cluster. This second cluster is formed by peers sites.

- Apart from the hubs and the peers clusters of the movements networks, the
other ones are formed by sites of ID > 1050. These sites are less connected to the other.

The communities in the networks of calls and the networks of text message exchange were not clearly identifiable by the Louvain algorithm.

4.4. Results

The proposed algorithm was used for trajectories clustering of the Dataset 2 of the D4D Challenge datasets and 3 classes of clusters emerged:

1. There were some clusters with centroids representing movements along the roads linking distant cities and villages. Figures 3(a-c) illustrate these centroids. The trajectories in this class of clusters represent trips made in a day. We observed that most of the centroids of this form had one of their ends in or around the city of Dakar. The timestamps and the number of trajectories on such cluster can interest inter-city transport agencies for the optimization of their resources.

2. The second class of clusters are those with circular trajectories centroids as shown in Figures 3(d-f). The trajectories in such clusters mostly cover two main areas: the first sites are close and grouped in the first area, then there is a move to a second area in which are grouped the other sites of the trajectories and finally, there is a return to the first area. Using Google Maps API, we found that such trajectories are very present in major cities of Senegal. So these trajectories may interest taxis and bus companies to improve the schedules and the paths of their rounds.

3. The third class of trajectories may also interest taxis. Indeed, clusters of this class consists of movements of relative short distances in a city or village (see Figures 3(g-i)).

Our clustering algorithm outputs the centroids with the timestamps associated to each site. This information is very important for traffic scheduling not only the transport companies but also for an efficient management, by the government, of the cities and villages. Indeed an aggregation of the timestamps of the centroids weighted by the size of each cluster can help to identify the periods of high traffic, places of high traffic demand, ...

5. Conclusion

We proposed a $k$-mediod like algorithm for cellphone’s users trajectories clustering. The algorithm which is based on PAM used a new distance function. This function is called Alignment distance with Threshold Penalty (ATP) it try to capture the fact that two trajectories can be merged to form a valid one. At the end of the PAM algorithm, we propose a cluster’s centroid calculation algorithm that produces the centroid of a cluster. The centroid is important for the studied problem since it represent the profile of the movements of some users. The experimentations conducted on the Dataset 2 of the D4D Senegal Challenge highlighted three classes of cluster’s centroids: daily trip centroids, circular trajectories centroids and short trajectories centroids. The spatio-temporal representation of the trajectories and the clustering algorithm proposed in this paper can be used to improve the scheduling of transportation companies in particular and the transportation management strategies in cities and villages in general.
References


Generating anonymous sequence-preserving datasets

Shaked Sigal and Rokach Lior

Ben-Gurion University of the Negev, Israel

ABSTRACT

Over the years, together with the evolution of services like navigation and telecommunication, wide amounts of data have been collected. This data could be analyzed in order to learn behavior of diverse individuals or populations. Unfortunately in most cases the collected data cannot be delivered to third parties for privacy issues.

In this work we suggest a new methodology for generating a synthetic dataset based on a given dataset, while preserving k-anonymity. The algorithm first groups similar sequences within the origin data. Each group's sequences are then modeled according to Markov assumption, and a new dataset is generated according to the behavior of groups with more than k objects. To the best of our knowledge this is the first work that preserves sequences within anonymized data. We will demonstrate how this technique yields high similarity between sequences in the origin and synthetic datasets.

*This work was supported in part by Deustche Telekom AG Labs.

1 INTRODUCTION

The generation of anonymous datasets that preserve sequences, or movement patterns in our context, will facilitate the release of sensitive data like call records and GPS data for research community. This will enable analyzing populations as well as individual profiles and better fitting transportation and infrastructures to population needs.

We suggest a novel technique for generating synthetic anonymous datasets, while preserving sequence patterns within the data.

Sequences are first partitioned into groups with similar behavior. Then a Markov model is trained for each group and a synthetic dataset is generated according to the ensemble of models.

This paper describes our work in more details. In section 2 we present some preliminary definitions, the algorithm is described in details in section 3, some initial experimental results are presented in section 4. We conclude current work and present our next steps on section 5. Finally, a wide literature review is presented in section 6.
2 Problem Definition

A sequential dataset $D_{\text{orig}}$ contains records of the form $<$obj, t, e$. Each record indicates an event $e$ that occurred at time $t$ and is attributed to object obj.

A window size $w$ is defined for separating one object’s events into several sequences. In this manner we can avoid long sequences which lead to over-fitted results which hardly could be anonymized. In the current work we treat each 24-hour period as a separated sequence of events ($w = 24$ hours). A sequence $s_i$ is viewed as a series of events $e_i$ assigned to the same object, ordered by event times $t$, for the duration of a single window size $w$. A sequence is denoted by $<$e$_1$,.., e$_n$>, where $e_i$ is the $i^{th}$ event in the sequence. A sequence of cell-IDs, for example, can be constructed by first sorting one object’s events according to ascending record time. Then for each different date, all ordered events, or cell-IDs in this example, are concatenating into sequences $(123\rightarrow 45\rightarrow 45\rightarrow 345\rightarrow 67)$.

The entire dataset $D_{\text{orig}}$ is represented as a set of sequences $s_i$ performed by objects obj, denoted by $<$obj, $s_1$, .., $s_n$>.

**Definition 1 (similarity of sequences):** we measure similarity between two sequences as the normalization of longest common subsequence (LCSS $[\text{Vlachos et al. (2002)}]$).

Let two sequences be defined as follows: $s_1 = (e_1, e_2,.. e_m)$ and $s_2 = (e_2, e_3,.. e_n)$. The longest common subsequence length is given as:

$$\text{LCSS}(S_1,S_2) = \begin{cases} 
\emptyset, & m = 0 \text{ or } n = 0 \\
\text{LCSS}(\text{rest}(S_1),\text{rest}(S_2)) + 1, & x \geq 0 \\
\max(\text{LCSS}(\text{rest}(S_1),S_2),\text{LCSS}(S_1,\text{rest}(S_2))), & \text{otherwise}
\end{cases}$$

where "rest" refers to the prefix of the sequence.

The similarity is normalized by deviation with the geometrical mean of the two sequence lengths.

**Definition 2 (sequences Group):** A group $G$ of sequences, denoted by $<$s$_1$, .., s$_n$>, contains similar sequences (as measured by the LCS measure). An object obj can be assigned to several sequences and therefore can be attached to several different clusters.

**Definition 3 (sequences k-anonymity):** A sequence satisfies k-anonymity if it was generated according to a model that was trained based on a group $G$ of sequences that are assigned to at list k different objects.

**Definition 4 (sequential datasets similarity):** Similarity between origin dataset $D_{\text{orig}}$ and synthetic dataset $D_{\text{syn}}$ is measured as the average distance between each sequence in $D_{\text{orig}}$ and its nearest neighbor in $D_{\text{syn}}$.

$$\text{sim}(D_{\text{orig}},D_{\text{syn}}) = \sum_{s_i \in D_{\text{orig}}} \frac{\text{LCSS}(s_i, \text{nearest neighbor}(s_i, \text{nearest} - \text{neighbor}(s_i)))}{|D_{\text{orig}}|}$$

where nearest. neighbor$(s_i, \text{nearest} - \text{neighbor}(s_i))$ returns sequence in $D_{\text{syn}}$ with maximal LCSS-similarity to $s_i$.

**Problem 1 (The anonymous sequences generation problem):** let $s_i = \{ e_i, ... e_n \}$ with $e_i \in \Sigma^*, \forall i \in \{1, ..., n\}$ be the set of sequences in dataset $D_{\text{orig}}$ whose sequential patterns must be learned and maintained in a synthetic created dataset $D_{\text{syn}}$. The anonymous sequences generation problem requires the transformation of $D_{\text{orig}}$ into $D_{\text{syn}}$ such that:
1. \( \forall s_i \in D_{\text{syn}}, \text{objects.amount}(G(s_i)) \geq k \)
2. \( \text{sim}(D_{\text{orig}}, D_{\text{syn}}) \leq l \)

The first requirement guarantees that generated sequences satisfy k-anonymity, while the second requirement ensures the sequential similarity between the origin and synthetic datasets.

3 AN ALGORITHM FOR GENERATING ANONYMOUS SEQUENTIAL DATASETS

3.1 PREPROCESSING STEPS
As a preprocessing step the algorithm identifies sequences according to the defined window size. In the current work we focus on 24-hours window size, therefore each combination of date and object is counted as a separate sequence.

3.2 GROUPING SIMILAR SEQUENCES
In order to generate k-anonymous sequences we first divide sequences into groups. Then a model is trained for each group that satisfies k-anonymity. Each such model will serve the sequence generation process by representing a certain type of behavior. An agglomerative hierarchical clustering approach is used for this task. In the beginning of the process, each sequence is in a cluster of its own. The clusters are then sequentially combined into larger clusters, until all sequences are grouped within a single cluster. At each step, the two clusters separated by the shortest distance are combined. In the complete-linkage clustering that we used, the distance between two clusters is made by a single sequences pair, namely those two sequences (one in each cluster) that are furthest from each other according to the LCSS measure described in previous sections. The separation into \( n \) different groups is then conducted by setting a level in the hierarchy where \( n \) groups exist. Groups that contain sequences that are assigned to less than \( k \) objects are suppressed in order to preserve k-anonymity.

3.3 LEARNING SEQUENTIAL MODEL
We model each group of sequences using Markov model, which is a stochastic model that assumes that future states depend only on the present state. Generally, this assumption enables reasoning and computation with the model that would otherwise be intractable. We refer to the event hour as an influencing factor and therefore combine it in the calculated statistics. For each group we collect the following statistics:

1. Probability to start a sequence from each state at each hour.
2. Probability for transition between a pair of states at each hour.
3. Probability to start a sequence at each hour.
4. Average time between states at each hour from each state.

In addition, the following cross clusters statistics are collected:
1. Probability of an object to start first sequence according to each cluster
2. Probability for transition of object between a pair of clusters.

3.4 Generating a Synthetic Sequential Dataset

An algorithm for generating a synthetic sequence preserving dataset that satisfies k-anonymity while preserving sequential patterns observed in a given dataset is presented below.

**Generate-sequence-preserving-dataset (D\textsubscript{orig}, k, n)**

Input: $D\textsubscript{orig}$ – a dataset; $k$-anonymity threshold, $n$ – groups amount, $w$- window size.
Output: a dataset $D\textsubscript{syn}$.

1. $D\textsubscript{orig}.sequences$ $\leftarrow$ partition data into $w$ size sequences
   //Clustering data
2. clusters $\leftarrow$ complete-linkage-hierarchical-clustering ($D\textsubscript{orig}.sequences$)
   //Collecting statistics
3. For each $c$ in clusters
4.   Start.cluster.prob $\leftarrow$ count objects with first sequence in $c$, normalize by objects amount
5.   For each dest.$c$ in clusters
6.     Transition.cluster.prob $\leftarrow$ count how many pairs of sequential sequences of a given object transit from $c$ to dest.$c$, normalize by sequences amount
7.   For each hr in 1:24
8.     start.hour.prob $\leftarrow$ count sequences in $c$ that starts within $hr$, normalized by sequences amount
9.   For each state $s$
10.    Start.state.prob $\leftarrow$ count sequences in $c$ within $hr$ that starts within $s$, normalized by sequences amount
11.    Avg,tba $\leftarrow$ calculate average time between records in $s$ within $hr$ and its following records
12.   For each state dest.$s$
13.     transition.state.prob $\leftarrow$ count how many sequences in $c$ within $hr$ transit from $s$ to dest.$c$, normalize by sequences amount
14.   Min.time, max.time, objects.amount, sequence.per.object.amount, end.seq.hr $\leftarrow$ aggregate $D\textsubscript{orig}$ for setting these parameters

   //Generating dataset
15. For $obj$ in 1: objects.amount
16. For $seq$ in 1:sequence.per.object.amount
17.   If ($seq==1$)
18.      Then Cur.cluster$\leftarrow$ sample(Start.cluster.prob)
19.      Else Cur.cluster$\leftarrow$ sample(Transition.cluster.prob[c= Cur.cluster])
20.   Cur.hour$\leftarrow$ sample(start.hour.prob[c= Cur.cluster])
21.   Cur.state$\leftarrow$ sample(Start.state.prob [c= Cur.cluster, hr= cur.hour])
22.   Cur.time$\leftarrow$ set-hour(Cur.hour ,Min.time)
23.   $D\textsubscript{syn}.add(obj, Cur.time, Cur.state)$
24.   While (cur.time< max.time) & (cur.hour< end.seq.hr)
25.      Cur.time$\leftarrow$ Cur.time+ Avg,tba[c= Cur.cluster, hr= cur.hour, s=cur.state]
26.      Cur.hour $\leftarrow$ hour(cuur.time)
27.      Cur.state$\leftarrow$ sample (transition.state.prob[c= Cur.cluster, hr= cur.hour, s=cur.state])
28. While is.null(cur.loc) & (cur.time< max.time) & (cur.hour< end.seq.hr)
29.       Cur.time <- Cur.time+ Avg.tba[c= Cur.cluster, hr= cur.hour, 
                  s=cur.state]
30.     Cur.hour <- hour(cur.time)
31.   Cur.state <- sample (transition.state.prob[c= Cur.cluster, hr=
                        cur.hour, s=cur.state])
32.       If (!is.null(Cur.state))
33.          Then D_syn.add(obj, Cur.time, Cur.state)
34. return D_syn

The algorithm accepts as parameter an input dataset $D_{orig}$, which will be used for training a model according to which new dataset $D_{syn}$ will be generated. Other input parameter are the amount of groups ($n$), the amount of minimum objects in each group ($k$) and the window size ($w$) according to which sequences will be divided.

The algorithm first clusters $D_{orig}$ (line 2) with complete linkage hierarchical clustering algorithm, using sequences as elements and LCS as similarity measure. Then (lines 3-14), statistics are collected as follows: Start cluster probability is collected, transition between clusters probability is collected per each pair of clusters, starting hour probability is measured per each cluster, starting state probability and average time between arrivals are measured per each combination of cluster, hour and state, and state transition probability is measured per each combination of cluster, hour and a pair of states.

Other global statistics are collected from $D_{orig}$: minimal and maximal date-time, objects amount, averaged amount of sequences per object, and the last hour of a sequence (can also be set as input parameter).

The generation process then takes place (lines 15-33): Objects are created according to their amount in $D_{orig}$, per each object several sequences are generated according to the average sequences amount per object in $D_{orig}$. Per each sequence a starting record is calculated (lines 17-23). A cluster is first chosen according to collected statistics. If the current sequence is the first sequence of that object, the cluster will be chosen according to starting cluster statistics, otherwise, it will be chosen according to cluster transition statistics. Then a starting hour is chosen according to statistics in that cluster, and a starting state is chosen according to statistics in the current cluster and hour. The first record of the sequence is then appended to $D_{syn}$.

The rest of the records in that sequence are added iteratively while not reaching the end of the sequence (lines 24-33) by first progressing in time according to average time between arrivals within the current cluster, hour and state. The progress in time will continue until a transition between states is possible according to statistics within the current cluster, time and state. A new state is then chosen and the record is appended to $D_{syn}$.

The algorithm returns the created synthetic dataset ($D_{syn}$) as an output.
4 EXPERIMENTS

We run a set of experiments on samples extracted from the project's second dataset, which is a fine-grained mobility data (site level) on a rolling 2-week basis with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled users.

Our current implementation is written in R, based on the TraMineR and cluster packages. Further effort is required in order to handle big-data demands. We are currently working on a Java implementation that issues big data needs.

We ran 10 simulation, each processing a 50,000 records sample of the data, with 390 different objects and 4904 different sequences on average. The algorithm's input parameters were set to 250 clusters and k=2 anonymity.

In order to evaluate sequence preservation within the synthetic dataset, we measured similarity between each sequence in the origin dataset and its nearest neighbor in the synthetic dataset (using LCSS measure). Figure 1 exhibits the number of sequences per each level of similarity preservation percentage in each of the simulations.

**Figure 1: Preserved similarity within sequences**

As can be seen, most of the sequences in the origin dataset preserve high similarity to their equivalent in the synthetic dataset. The averaged similarity between sequences in the origin
and fabricated data is 69%. This score is damaged mainly for sequences that were
suppressed from the synthetic dataset due to k-anonymity requirements.

5 CONCLUSIONS AND FUTURE WORK

The suggested algorithm seems to preserve sequences within synthetic datasets, with about
70% averaged similarity between origin and synthetic sequences. 
Further effort is currently placed at tuning the algorithm's parameters, as well as fitting the
algorithm for big data needs, including the adoption of an incremental learning approach that
enables new datasets to utilize earlier knowledge. We also plan on expanding the algorithm to
handle additional influencing factors (for example week day versus holiday).
Tuning the algorithms parameters (k for anonymity and n for clusters amount) is also of great
importance. Other clustering approaches could also be estimated, as well as other similarity
measures.
Other solutions for preserving k-anonymity should be examined (not only suppression of small
groups) that seemed to damage similarity between original and synthetic sequences.
A version of the algorithm that specializes on spatio-temporal data could be developed. In that
case sequence events similarity should consider spatial closeness.

6 RELATED WORK

We survey here a wide range of related issues. As can be seen in this section, different aspects
of anonymizing datasets were treated before, but with three main distinctions: 1. Other
techniques embed slight changes in the origin dataset instead of generating a whole fabricated
dataset with the same patterns. 2. They do not focus on the maintenance of sequences within
the data. 3. Closest solutions only handle spatio-temporal trajectories and do not supply
solutions the generic sequence data.

6.1 SYNTHETIC DATA GENERATION

Synthetic data refers to data which was not directly measured but instead was generated,
usually based on a statistical or a probabilistic model. Data generation is the process of storing
this data in order to enable performing different business processes like research, software
testing or analysis, assuming that the synthetic data is similar to the real data and therefore will
yield the same inferences. First the real data is analyzed in order to build a model and then a
synthetic data set is built by sampling records from the created model. Privacy is gained by
noise created by the model bias as well as by the sampling randomness.
Most of the existing generators are based on statistical models and attend issues like data
dependencies and inter-attribute correlations, but to the best of our knowledge none of them
preserves sequences within the data. Wu et al. (2007) address the problem of testing databases
applications. They create synthetic data using the general location model which is built using
rules and statistics extracted from the production data and its DDL definitions. They also
conduct a disclosure analysis in order to avoid providing traces of sensitive rules. Bruno and
Chaudhuri (2005) suggest a C-like Data Generation Language (DGL) that requires user's
definition of data types and distributions in order to generate data. Yahalom et al. (2010) suggest creating a set of rules according to production DB and generating a test DB using a constraint satisfaction problem solution of these rules. Young et al. (2009) demonstrate generating synthetic institutional care datasets using both Bayesian network and hierarchical Bayesian models. They conclude that with a hierarchical model it is easier to create synthetic data that have a relatively low risk of disclosure. They also claim that the Bayesian model is more sensitive to prior information, which is critical for selecting an appropriate network.

Vreeken et al. (2007) generate data using models created by the KRIMP algorithm. This model maintain data distribution of patterns within the data, but these patterns refer to combinations of different attributes within a single record (the pattern length is limited by the amount of dimensions) while we seek for patterns of events created by several records. MacHanavajjhala et al. (2008) generated a synthetic dataset according to real data that describe employees' home and working places. They use multinomial model with dirichlet prior and show that generation of synthetic data can achieve privacy with little noise. Their solution however does not handle data with large domains where the data tend to be sparse. Aggarwal et al. (2004) condense data into multiple groups of pre-defined size, for each group they maintain statistical information about the mean and covariance across the dimensions. The minimum size of a group represent its indistinguishability level. Each group is created by sampling a record and adding its k nearest neighbors. They also suggest an incremental technique for updating the model. It is worth mentioning that this approach only deals with numeric attributes.

Theodoridis et al. (1999) deal with the generation of spatio-temporal data. They propose an algorithm, called Generate_Spatio_Temporal_Data (GSTD), for generating timeevolving (i.e., moving) point or rectangular objects. GSTD generates records by randomly generating spatial and temporal attributes within given minimum and maximum thresholds according to a given distribution. Tuning these parameters can lead to different types of movements (different direction, velocity, agility ect.). Giannotti et al. (2005) generate synthetic trajectories based on CDRs using an extension of GSTD algorithm. They use several probability distributions to generate different group behavior defined by the user. They also define, for each group, an infrastructure, a collection of rectangles that must be avoided by group objects, during generation process. This spatial-temporal data generator, however, relies on user definitions and does not learn a model based on real data.

Isaacman et al. (2012) use CDRs or census data to generate sequences of locations and associated times that capture how individuals move between important places in their lives, particularly home and work. Spatial home-work densities are estimated relying on three distributions: 1.probability for home according to the likelihood of where people live in the simulated city; 2.probability of having different commute distances, conditioned on that given home location; 3. the probability of different work locations around a circle of the chosen commute distance from the home location. They maintain calling patterns by creating a 24-dimensional vector per each user, with calling probabilities calculated by the amount of calls made by the user at each time of the day. Then the user vectors are clustered (using X-means algorithm). Finally per-minute call probability distribution is calculated per each group. A synthetic user is generated by first tying its home and working location. For each day, calls are made in the amount and hours chosen according to the calling distribution of the user's class. When a “call” is made, the location of the call is determined to be either home or work according
to the probability of a person being in the location at that time of day. A more complex algorithm that can supports more than two visited locations per user is also suggested. According to the authors it is highly unlikely that any of the synthetic people exhibits a mobility pattern that identifies a single real person, however, this technique mainly preserves densities over time, but does not maintain sequences and routes of the underlying users.

6.2 Privacy preservation

Several models are commonly used to prevent privacy attacks on relational data. K-anonymity [Sweeney (2002)] prevents linkage attacks by requiring every set of indistinguishable records (with respect to certain identifying attributes) to contain at least K records. L-diversity [Machanavajjhala et al. (2007)] requires that the distribution of a sensitive attribute in each set of indistinguishable records has at least l well-represented values. However, these models are not effective on trajectory data due to its high dimensionality, sparseness, and sequentiality [Ghinita et al. (2008)]. Dwork (2006) proposes $\epsilon$-differential privacy, according to which given two different datasets that differ on exactly one record, a data analysis algorithm that satisfies differential privacy will output randomized results with probability distributions that are far by $\epsilon$ from being identical. Therefore, an adversary will not be able to guess whether a given record appears in the origin dataset no matter how much he knows in advance. A randomized algorithm satisfies $\epsilon$-differential privacy if the ratio between the probability that the algorithm outputs any output on a dataset and the probability that it outputs the same output on a dataset that differ in exactly one record, is bounded by a constant. Typically, differential privacy is achieved by adding noise to the outcome of a query. One major limitation of the Laplace mechanism is that it requires the output of the algorithm to be real numbers. Another way to obtain differential privacy is through the exponential mechanism [McSherry and Talwar (2007)]. Given a quality function that scores outcomes of a calculation, where higher scores are better, for a given database, the quality function induces a probability distribution over the output domain, from which the exponential mechanism samples the outcome. This probability distribution favors high scoring outcomes (they are exponentially more likely to be chosen), while ensuring $\epsilon$-differential privacy. In spite of the privacy guarantee provided by differential privacy, it has been criticized for not being able to achieve usable information quality in some data analysis tasks, especially on complex data models and datasets with highly correlated records [Yang et al. (2012)]. It is also more difficult to design differential privacy protocols to handle arbitrary updates.

Ghinita et al. (2008) propose an anonymization technique which addresses the challenge of high dimensionality. First, they organize the data as a band matrix by performing permutations of rows and columns in the original table such that most non-zero entries are near the main diagonal, taking advantage of its sparseness and preserve correlation. Next, nearby transactions are grouped according to QID similarity. In each such anonymized group that preserve correlation, sensitive items are separated from the QID, and published in a separate summary table, that expose only probabilities per each value of the sensitive attribute per each group. Their technique, however, does not handle sequences (ignores the order of different sensitive values). Moreover, it separates data into QID fields and sensitive fields, whereas in trajectories the time and place serves both as QID and sensitive fields.
Some studies suggest algorithm for the anonymization of sequential data, in which detailed timestamps are ignored. However, in many cases the time information is essential. Atzori et al. (2007) formalize the threats to anonymity by means of inference channels through frequent itemsets. They exclude itemsets with less than the minimal required support. Ghasemzadeh et al. (2014) propose a method for anonymizing trajectory data. They adopt the LK-privacy model, and therefore demand that any subsequence with length at most L in a trajectory database must be shared by at least K records. LK-privacy guarantees that the probability of a successful identity linkage attack is at most 1/K. They aim at preserving the information quality for supporting effective passenger flow analysis. First a passenger probabilistic flowgraph is extracted, which is a tree where each node is a pair (location, time) and each edge represent the transition between two nodes, with a certain probability. Then they identify violating sequences and remove it (completely or partially) in order to transform the database into another version that satisfies the LK-privacy requirement with maximal similarity between original and anonymized flowgraphs. Fung et al. (2009) propose LKC-privacy for anonymizing high-dimensional, sparse RFID data. They ensure not only LK-anonymity but also that the confidence of inferring any sensitive values S from the L-length trajectory is not greater than C. LKC-privacy bounds the probability of a successful record linkage attack to be ≤ 1/K (LK-anonymity) and bounds the probability of a successful attribute linkage attack to be ≤ C (LK-dilution). They propose a greedy algorithm to anonymize the dataset by a sequence of suppressions (removing violating subsequences). Chen et al. (2012) propose a sanitization algorithm to generate differentially private trajectory data by making use of a noisy prefix tree based on the underlying variable-length n-gram model behind the data. Each node holds the count of trajectories described by the nodes in the current branch and Laplace noise is added to these counts. They argue that their approach leads to better quality in terms of count query and frequent sequential pattern mining. However, these two approaches are limited to relatively simple data mining tasks and preserve relatively short patterns. Pensà et al. (2008) use similar trajectories' prefix tree and pruning technique for the anonymization of sequential data, while each pruned trajectory propagates an increase in the support of the most similar trajectory in the prefix tree (using the edit distance for measuring similarity between trajectories). This work does not consider the information carried by temporal annotations as well as the geographical proximity of locations/regions.

Some recent works study anonymization of trajectory data from different perspectives, while assuming the data is in the form of continuous GPS data. These techniques, however, cannot be used for the anonymization of sequential data that lacks accurate positions. Abul et al. (2008) propose (K,δ)-anonymity, where K different trajectories should exist in a cylinder of radius δ. In the first step they use greedy clustering to group trajectories in clusters of at least k elements by first initializing centroids with far trajectories, then adding k nearest neighbors to each cluster, and repeatedly adding the rest of the trajectories to the nearest cluster with the split of clusters with more than 2k-1 members. Afterwards a minimum space translation is performed in order to achieve (k, δ)-anonymity (each trajectory is translated to the cluster's centroid). Terrovitis and Mamoulis (2008) suggest an algorithm that iteratively suppresses selected locations from the original trajectories until a privacy constraint is satisfied. Their greedy algorithm unifies unsecured trajectories according to the magnitude of information loss, calculated as distance between the unified secured and origin trajectories minus the distance between the separated
secured and origin trajectories. Nergiz et al. (2007) apply spatial discretization into two dimensional grids. They use a condensation based approach to form groups of similar trajectories (modifying the clustering algorithm for k-anonymity by enforcing that the size of the clusters should be more than k). The algorithm specifies the matching point pairs between the trajectories and anonymize the paired points w.r.t. each other (by replacing the points with the minimum bounding box that covers the points). Any unmatched points are suppressed. Finally they reconstruct trajectory by switching each generalized point with an atomic point (one of the grouped points). Andrienko et al. (2009) anonymize spatio-temporal datasets while preserving the k-anonymity property by the generalization of movement data. Specifically, replacing exact positions in the trajectories by approximate positions. Starting and ending movement locations as well as changing direction spots are clustered into groups of close locations, each group is represented by its centroid location and then a trajectory is represented as a sequence of centroid locations, with its corresponding time intervals. A prefix tree is built according to the trajectories and then pruned in order to preserve k-anonymity.

6.3 PATTERN MINING
Agrawal et al. (1995) introduce the basic problem of mining sequential patterns (FSP) over a large database of customer transactions. They view all the transactions of a customer as a sequence of items ordered by increasing transaction-time. A sequence \{a_1, a_2, ..., a_n\} is contained in another sequence \{b_1, b_2, ..., b_m\} if there exist integers \(i_1 < i_2 < ... < i_n\) such that \(a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, ..., a_n \subseteq b_{i_n}\). The problem of mining sequential patterns is to find the maximal sequences among all sequences that have a certain user-specified minimum support. Each such maximal sequence represents a sequential pattern. Based on this work, Giannotti et al. (2007) introduce trajectory patterns as concise descriptions of frequent behaviors, in terms of both space and time. First they extract popular locations (neighborhoods visited by at least 10% of the objects or locations extracted from a gazetteer), then they extract trajectories that contain these locations within their neighborhood (close time and location). T-patterns are represented as a pair of locations and durations. They use a step-wise heuristic that search for increasing size of patterns at each step. Each step only examine patterns that contain prefixes and suffixes recognized along previous steps. This work concerns frequent sequential pattern and ignores simultaneousness of the patterns. Peng et al. (2006) predict movements of objects in order to reduce the energy consumption for object tracking. They generate variable memory Markov (VMM) model for each object in each cluster head (requiring a single scan of the sequence). To facilitate the learning and maintenance of each VMM model, they use a variation of suffix tree and one corresponding buffer in a cluster head is used to hold the most recently segment of moving records. VMM model is trained on the fly and not all the tree nodes are stable for predicting. Mature nodes are those tree nodes in which a certain amount of records are collected and the conditional probabilities are stable. When predicting the next movement, only mature node is participated in prediction. If the average prediction hit rate is larger than \(\delta\), the cluster head will be in the prediction phase. Tsai et al. (2007) also utilize the characteristic of the group movement of objects to achieve energy conservation in the inherently resource-constrained wireless object tracking sensor network (OTSN). They use VMM for learning statistic of object moving sequences. Patterns are stored as a data structure called Probabilistic
Suffix Tree (PST) for mining significant moving patterns. PST building algorithm is a compression scheme that converts a large data set into a variable length tree. The tree represents a dictionary of significant sequences that are meaningful for predicting next location. Similarity between pairs of PSTs is calculated as a combination of condition probability and Euclidean distance over the union of node symbol strings of the two PSTs. The problem is then transformed into a graph connectivity problem, where highly connected components are extracted. After groups are found, the most representative PST named GPST is heuristically selected for each group such that the storage cost is reduced. Ron et al. (1996) suggest using a PST which generates an equivalent distribution to that of a probabilistic suffix automata (PSA). They apply the algorithm to correct corrupted English text and to build a stochastic model for E.coli DNA. It seems, though, that using PSTs is only efficient for problems with small domains. Spampinato and Palazzo (2012) propose an automatic system for the identification of anomalous fish trajectories. Hidden Markov Models (HMMs) are used to represent and compare trajectories. Multi-Dimensional Scaling (MDS) is applied to project the trajectories onto a low-dimensional vector space, while preserving the similarity between the original data. Normal events are then defined as set of trajectories clustered together, on which HMMs are trained and used to check whether a new trajectory matches one of the usual events, or can be labeled as anomalous.

6.4 SIMILARITY BETWEEN SEQUENCES

Keogh and Pazzani (2000) suggest the DTW (Dynamic Time Warping) distance. It is an alignment based distance. The basic idea behind DTW is to find out the warping path between two trajectories that minimizes the warping cost. However, DTW requires continuity along the warping path, which makes it sensitive to noise and it is unable to find trajectories that have similar shapes but with dissimilar gaps in between. Nergiz et al. (2007) also define a distance metric between trajectories based on the alignment problem. The alignment problem has been well studied for strings, the goal there is to find an alignment of strings such that the total pairwise edit distance between the strings is minimized. Chen and Ng (2004) introduce a metric distance function, ERP (Edit distance with Real Penalty), to measure the similarity between time series data. ERP introduces a constant value as the gap of edit distance and uses real distance between elements as the penalty to handle local time shifting. However, it does not consider variations in gap sizes between two similar subsequences of the trajectories. Like DTW, ERP can handle the time series data with local time shifts. However, because it takes the differences of real values as distance, ERP is also sensitive to noise. Vlachos et al. (2002) use LCSS (Longest Common Subsequence) to compare two trajectories with the consideration of spatial space shifting. The LCSS distance finds the alignment between two sequences that maximize the length of common subsequence. LCSS requires a threshold $\epsilon$ to be established. This threshold is used to determine whether or not two elements match and allows LCSS to handle noise by quantizing the distance between two elements to two values, 0 and 1, to remove the larger distance effects caused by noise. Compared with DTW and ERP, LCSS is robust to noise. However, LCSS allows gaps with various sizes to exist between similar shapes in the sequences, which cause its inaccurate. Chen et al. (2012) introduce a new distance function, Edit Distance on Real sequence (EDR) which is robust against noise that usually appear in trajectories. EDR is based on edit distance on strings, and removes the noise effects by
quantizing the distance between a pair of elements to two values, 0 and 1. Seeking the minimum number of edit operations required to change one trajectory to another offers EDR the ability to handle local time shifting. Furthermore, assigning penalties to the unmatched parts improves its accuracy. Their compare of EDR with other popular distance functions such as Euclidean distance, DTW, ERP and LCSS, indicate that EDR is more robust than Euclidean distance, DTW and ERP, and it is on average 50% more accurate than LCSS. They also suggest several pruning techniques to improve the retrieval efficiency, among which is pruning by histograms. They map strings to their frequency vectors (FV). It is proven that the frequency distance (FD) between the FVs of two strings is the lower bound of the actual edit distance. In fact, frequency vectors are one-dimensional histograms over strings, where each bin is a character in the alphabet. Porikli and Haja (2004) propose HMM-based distance. Each trajectory is fitted by an HMM (Hidden Markov Model). The HMM-based distance is the sum of likelihoods for each trajectory to be generated by its own model minus the sum likelihoods for each trajectory to be generated by the other trajectory’s model. This measure was criticized by Zhang et al. (2006) for suffering from over-fitting due to the small training data that trains a statistical model for each trajectory.

Fu et al. (2005) cluster vehicle trajectories using similarity measure according to the average Euclidean distance between corresponding points of two trajectories. This solution is not robust to noise or time shifting which often appear in trajectory data. Moreover, it suffers from the limitation that the length of both trajectories must be the same. In order to overcome the last limitation Vlachos et al. (2002) apply the strategy according to which the shorter of the two trajectories slides along the longer one and the minimum distance is recorded. Bashir et al. (2003) propose using the Euclidean distance after performing PCA (Principle Components Analysis). A trajectory is first represented as a 1-D signal by concatenating the x and the y projections. Then the signal is converted into the first few PCA coefficients. The trajectory similarity is computed as the Euclidean distance between the PCA coefficients. Zhang et al. (2006) demonstrate that in outdoor surveillance scenes, the simpler PCA+Euclidean distance is competent for the clustering task even in case of noise, as more complex similarity measures such as DTW, LCSS are not efficient due to their high computational cost.

Abraham and Sojan Lal (2012) apply a spatiotemporal similarity measure which is a combination of structural and sequence similarities calculated based on three components. Phase I is based on a similarity measure that checks the number of common locations visited by two trajectories. In Phase II, the similarity is measured using the sequence alignments of travel locations. Finally the temporal similarity is calculated as the difference in visiting times at each POI (point of interest) with Times of Interest (TOIs). POIs and TOIs are derived from the query trajectory. These three similarity components are combined into one score.

6.5 EVALUATION METHODOLOGY FOR ANONYMIZATION PROCESS

Different techniques have been used in the literature for estimating the quality of the anonymization process. Aggarwal et al. (2004) show that the synthetic data preserve the inter-attribute correlations of the data, and they also demonstrated high accuracy when using the synthetic data for a classification task, even higher from classification results of the origin data because the condensation removes anomalies from the data. Isaacman et al. (2012) assess the dissimilarity of two population density patterns using the Earth Mover’s Distance (EMD), which
attempts to find the minimum amount of energy required to transform one probability distribution into another. To determine the average error between real and synthetic location density patterns, they generate location probability distributions for each hour of the day from synthetic CDRs. Then they calculate the EMD between their synthetic distributions and the reference probability distribution Random Waypoint and Weighted Random Waypoint. In the Random Waypoint (RWP) model each user selects a random destination from all possible destinations in the area to be simulated and waits for a random amount of time. The Weighted Random Waypoint (WRWP) model behaves similarly, however, in this variant, the destinations are not chosen from a uniform distribution. Instead, a location probability distribution is used to weight possible waypoints. The distribution used here is obtained by combining all of the hourly probabilities into a distribution that gives the popularity of the location for making calls over the whole day. Bayardo and Agrawal (2005) suggest measuring discernibility that measures the data quality of the anonymized dataset based on the size of each anonymity set, since data quality shrinks as more data elements become indistinguishable. Given a clustering $P = \{p_1, \ldots, p_n\}$ of dataset $D$, where $p_n$ represents the trash bin, the discernibility metric is defined as:

$$DM(D) = \sum_{i=1}^{n} |p_i|^2 + |p_n||D|$$

Few works measure the similarity between sequences that appear in the origin and anonymized datasets. Ghasemzadeh et al. (2014) try to preserve data quality along the anonymization process by measuring similarity between passenger probabilistic flowgraphs extracted from the original and anonymized data. The similarity is measured using their Info measure which regards four attributes of the graph with pre-chosen weights per each attribute. They calculate per each node $d$: the number of instances of $d$ in the flowgraph, the total number of child nodes of $d$ in the flowgraph, the number of root-to-leaf paths containing $d$ in the flowgraph, the number of trajectories in the database that contain $d$. For each pair of identical nodes in the original and anonymized flowgraphs, these attributes are computed and the ratios among them is summed up. Pensa et al. (2008) compare the collections of pattern extracted before and after the anonymization process. They define two metrics for measuring similarity between two collections of patterns: 1) Frequent Pattern Support Similarity that calculates per each pair of trajectories the support similarity (minimal support divided by maximal support), and calculates average similarity according to all the generalized trajectories. 2) Frequent Pattern Collection Size Similarity which is achieved by dividing the minimal dataset size with the maximal dataset size. Fung et al. (2009) use the distortion ratio that measures the percentage of pair instances suppressed for achieving a given LKC-privacy requirement, measured by

$$\frac{|Pairs_{orig}| - |Pairs_{anonymized}|}{|Pairs_{orig}|}$$

Terrovitis and Mamoulis (2008) measure utility by the average difference between the original trajectories and the corresponding published trajectories. The difference between two trajectories is calculated as the sum of distances of the contained points. When a point in one trajectory is before the start or after the end of the compared trajectory, the distance is calculated according to the start or end point of the compared trajectory. Otherwise, the distance is calculated between the point and its projection on the compared trajectory. Several studies compare the results of queries or data mining tasks performed on both the origin and anonymized datasets in order to compare between the both. Young et al. (2009) estimate the results by prediction task using logistic regression model on both real and synthetic
datasets. Andrienko et al. (2009) apply cluster analysis to the original and modified trajectories using their density-based clustering OPTICS algorithm. They found that the results of clustering the original and the generalized trajectories are very similar when the distance threshold for the generalized trajectories is about one half of the distance threshold for the original trajectories. Chen et al. (2012) evaluate their data generator by measuring the relative error of count queries' answer on the sanitized sequential database with respect to the true answer on the original database. They also compare errors between top K most frequent sequential patterns in the dataset. Abul et al. (2008) use an extension of distortion, based the sum of pointwise distances between the original and translated trajectories that also consider distortion inserted along the preprocessing and space translation phases. They also compare the results between queries evaluated on the original and anonymized datasets. Nergiz et al. (2007) measure the difference between the original and anonymized datasets by the number of removed points as well as the distortion. In order to measure distortion they perform a clustering process and measure the deviation from the original clustering results. They used a bottom-up complete-link agglomerative clustering algorithm, coupled with the ERP distance metric, which has been specifically developed for trajectories. In order to estimate the distortion they consider every pair of trajectories and verify whether both are in the same cluster in the reference partition and whether they are in the response partition.

7 References


On the anonymizability of mobile traffic datasets

Marco Gramaglia
CNR - IEIIT
Corso Duca degli Abruzzi 24
10129 Torino, Italy
marco.gramaglia@ieiit.cnr.it

Marco Fiore
CNR - IEIIT
Corso Duca degli Abruzzi 24
10129 Torino, Italy
marco.fiore@ieiit.cnr.it

ABSTRACT
Preserving user privacy is paramount when it comes to publicly disclosed datasets that contain fine-grained data about large populations. The problem is especially critical in the case of mobile traffic datasets collected by cellular operators, as they are prone to subscriber re-identifiability and they are resistant to anonymization through spatiotemporal generalization. In this work, we investigate the anonymizability of two large-scale mobile traffic datasets, by means of a novel dedicated measure. Our results are in agreement with those of previous analyses, and provide additional insights on the reasons behind the poor anonymizability of mobile traffic datasets. As such, our study is a step forward in the direction of better dataset anonymization.

1. INTRODUCTION
Public disclosure of datasets containing micro-data, i.e., information on precise individuals, is an increasingly frequent practice. Such datasets are collected in a number of different ways, including surveys, transaction recorders, positioning data loggers, mobile applications, and communication network probes. They yield fine-grained data about large populations that has proven critical to seminal studies in a number of research fields.

However, preserving user privacy in publicly accessible micro-data datasets is currently an open problem. Publishing an incorrectly anonymized dataset may disclose sensitive information about specific users. This has been repeatedly proven in the past. One of the first and best known attempts at re-identification of badly anonymized datasets was carried out by then MIT graduate student Latanya Sweeney [1, 2] in 1996. By using a database of medical records released by an insurance company and the voter roll for the city of Cambridge (MA), purchased for 20 US dollars, Dr. Sweeney could successfully re-identify the full medical history of the then governor of Massachusetts, William Weld. She even sent the governor full health records, including diagnoses and prescriptions, to his office. A later, yet equally famous experiment was performed by Narayanan et al. [3] on a dataset released by Netflix for a data-mining contest, which was cross-correlated with a web scraping of the popular IMDB website. The authors were able to match two users from both datasets revealing, e.g., their political views.

Recently, severe concerns have been raised by privacy breaches in mobile traffic datasets. These datasets are collected at different locations of the cellular network infrastructure, and contain information about movements and traffic generated by millions of subscribers, typically for long time spans in the order of months. Mobile traffic datasets have become a paramount instrument in large-scale analyses across disciplines such as sociology, demography, epidemiology, or computer science. Unfortunately, they are also extremely prone to attacks on individual privacy. Namely, mobile traffic datasets suffer from the following issues:

1. Elevate re-identifiability. Mobile subscribers have very distinctive patterns that make them easily identifiable even within a very large population. Zang and Bolot [4] showed that 50% of the mobile subscribers in a 25 million-strong dataset could be uniquely detected with minimal knowledge about their movement patterns, namely the three locations they visit the most frequently. The result was corroborated by de Montjoye et al. [5], who demonstrated how an individual can be pinpointed among 1.5 million other mobile customers with a probability almost equal to one, by just knowing five random spatiotemporal points contained in his mobile traffic data.

2. Low anonymizability. The legacy solution to re-identifiability is generalization and suppression of data. However, both studies above proved that blurring users in the crowd, by reducing the spatial and temporal granularity of data, is hardly a solution in the case of mobile traffic datasets. Zang and Bolot [4] found that reliable anonymization is attained only under very coarse spatial aggregation, namely when the mobile subscriber location granularity is reduced to the city level. Similarly, de Montjoye et al. [5] explained that a power-law relationship exists between re-identifiability and spatiotemporal aggregation of mobile traffic. This
Table 1: Standard micro-data database format.

<table>
<thead>
<tr>
<th>Pseudo-id</th>
<th>Gender</th>
<th>Age</th>
<th>ZIP</th>
<th>Degree</th>
<th>Income</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>00015701</td>
<td>Male</td>
<td>21</td>
<td>77005</td>
<td>Bachelor</td>
<td>13,000</td>
<td>...</td>
</tr>
<tr>
<td>089036402</td>
<td>Male</td>
<td>37</td>
<td>77065</td>
<td>Master’s</td>
<td>90,000</td>
<td>...</td>
</tr>
<tr>
<td>42330327</td>
<td>Female</td>
<td>60</td>
<td>89123</td>
<td>High School</td>
<td>46,000</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 2: Mobile traffic database format.

<table>
<thead>
<tr>
<th>Pseudo-id</th>
<th>Spatiotemporal samples (fingerprint)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>c1,8, c14, c17</td>
</tr>
<tr>
<td>b</td>
<td>c14,8, c15, c16, c17, ... c14,15, c14,16, c15,17</td>
</tr>
<tr>
<td>c</td>
<td>c16,8, c17,20</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

implies that privacy is increasingly hard to ensure as the resolution of a dataset is reduced. In conclusion, not only mobile traffic datasets are easily re-identifiable, but they are also hard to anonymize. Ensuring individual privacy risks to lower the level of detail of such datasets to the point that they are not informative anymore.

In this work, we aim at better investigating the reasons behind such inconvenient properties of mobile traffic datasets. We focus on anonymizability, since it is a more revealing feature: multiple datasets that are all re-identifiable may be more or less difficult to anonymize. Attaining our objective brings along the following contributions: (i) we define a measure of the level of anonymizability of mobile traffic datasets, in Sec. 2; (ii) we provide a first assessment of the anonymizability of two large-scale mobile traffic datasets, in Sec. 3; (iii) we unveil the cause of naive re-identifiability and poor anonymizability in such datasets, i.e., the heavy tail of the temporal diversity among subscriber mobility patterns, in Sec. 4. Finally, Sec. 5 concludes the paper.

2. HOW ANONYMIZABLE IS YOUR MOBILE TRAFFIC FINGERPRINT?

In this section, we first define in a formal way the problem of user re-identification in mobile traffic datasets, in Sec. 2.1. Then, we introduce the proposed measure of anonymizability, in Sec. 2.2.

2.1 Our problem

In order to properly define the problem we target, we need to introduce the notion of mobile traffic fingerprint that is at the base of the mobile traffic dataset format. We also need to specify the type of anonymity we consider – in our case, $k$-anonymity. Next, we discuss these aspects of the problem.

2.1.1 Mobile traffic fingerprint and dataset

Traditional micro-data databases are structured into matrices where each row maps to one individual, and each column to an attribute. An example is provided in Tab.1. Individuals are associated to one identifier, i.e., a value that uniquely pinpoints the user across datasets (e.g., his complete name, social number, or passport number). Since identifiers allow immediate cross-database correlation, they are never disclosed. Instead, they are replaced by a pseudo-identifier, which is again unique for each individual, but changes across datasets (e.g., a random string substituting the actual identifier). Then, standard re-identification attacks leverage quasi-identifiers, i.e., a sequence of known attributes of one user (e.g., the age, gender, ZIP code, etc.) to recognize the user in the dataset. If successful, the attacker has then access to the complete record of the target user. This knowledge can directly include sensitive attributes, i.e., items that should not be disclosed because they may pertain to the personal sphere of the individual (e.g., diseases, political or religious views, sexual orientation, etc.). It can also be exploited for further cross-database correlation so as to extract additional private information about the user.

The same model directly applies to the case of mobile traffic datasets. However, the database semantics make all the difference here: while mobile users are the obvious individuals whose privacy we want to protect, attributes are now sequences of spatiotemporal samples. Each sample is the result of an event that the cellular network associated to the user. An illustration is provided in Fig. 1a, which portrays the trajectories of three mobile customers, denoted with pseudo-identifiers $a$, $b$, and $c$, respectively, across an urban area. User $a$ interacts with the radio access infrastructure at 8 am, while he is in cell $c_1$ along his trajectory. Then, he triggers additional mobile traffic activities at 2 pm, while located in a cell $c_2$ in the city center, and at 5 pm, from a cell $c_3$ in the South-East city outskirts. The same goes for users $b$ and $c$. All these spatiotemporal samples are recorded by the mobile operator\footnote{The actual precision of the information recorded, both in space and in time, can depend significantly on the nature of the probes used by the operator. Typically, probes located closer to the radio access can capture more events at a finer granularity, but require more extensive deployments to attain a similar coverage than lower-precision probes located in the mobile network core. In all cases, our discussion is independent of the mobile traffic data collection technique, and all the analyses performed in this work can be applied to any type of mobile traffic data.} and constitute the mobile traffic fingerprint of the user. The resulting database has a format such as that in Tab.2, where subscriber identifiers are replaced by pseudo-identifiers, and each element of a user’s fingerprint is a cell and hourly timestamp pair.

2.1.2 $k$-anonymity in mobile traffic

In order to preserve user privacy in micro-data, one
Figure 1: Example of mobile traffic fingerprints of three subscribers. (a) Initial dataset granularity: user locations are recorded at cell level, and the temporal information has a hourly precision. (b) First aggregation level: positions are recorded at each neighborhood, and the time granularity is reduced to two hours. (c) Second aggregation level: location data is limited to Eastern or Western half of the city, and the time information is merged over 12 hours.

has to ensure that no individual is uniquely identifiable in a dataset. This principle has led to the definition of multiple notions of non-uniqueness, such as $k$-anonymity [1], $l$-diversity [6] and $t$-closeness [7]. Among those, $k$-anonymity is the baseline criterion, to which $l$-diversity or $t$-closeness add further security layers that cope with sensitive attributes or cross-database correlation. More precisely, $k$-anonymity ensures that, for each individual, the set of attributes (or its quasi-identifier subset) is identical to that of at least other $k-1$ users. In other words, each individual is always hidden in a crowd of $k$, and thus he cannot be uniquely identified among such other users.

Granting $k$-anonymity in micro-data databases implies generalizing and suppressing data. As an example, in order to ensure 2-anonymity on the age and ZIP code attributes for the first user in Tab. 1, one can aggregate the age in twenty-year ranges, and the ZIP codes in three-number ranges: both the first and second user end up with a $(20,40)$ age and $770**$ ZIP code, which makes them both 2-anonymous. Clearly, the process is lossy, since the information granularity is reduced. Many efficient algorithms have been proposed that achieve $k$-anonymity in legacy micro-data databases, while minimizing information loss [8].

Also in mobile traffic datasets, $k$-anonymity is regarded as a best practice, and data aggregation is the common approach to achieve it [4, 5]. In this case, one has to ensure that the fingerprint of each subscriber is identical to that of at least other $k-1$ mobile users in the same dataset. We remark that previous works have typically considered a model of attacker who only has partial knowledge of the subscribers’ fingerprints, e.g., most popular locations [4] or random samples [5].

In order to counter such an attack model, a partial $k$-anonymization, targeting the limited information owned be the attacker, would be sufficient. However, we are interested in a general solution, so we do not make any assumption on the precise knowledge of the attacker, which can be diverse and possibly broad. Thus, $k$-anonymizing the whole fingerprint of each subscriber in the mobile traffic dataset is the only way to deterministically ensure mobile user privacy.

Both spatial and temporal aggregations can be leveraged to attain this goal. Examples are provided in Fig. 1b and Fig. 1c. In Fig. 1b, cells are aggregated in large sets that roughly map to the nine major neighborhoods of the urban area; also, time is aggregated in two-hour intervals. The reduction of spatiotemporal granularity allows 2-anonymizing mobile users $a$ and $b$: both have now a fingerprint composed by samples $(V,8-9)$, $(III,14-15)$, and $(VII,16-17)$. User $c$ has instead a different footprint, with samples $(IV,6-7)$ and $(III,20-21)$. If we need to 3-anonymize all three mobile customers in the example, then a further generalization is required, as in Fig. 1c. There, the metropolis region is divided in West and East halves, and only two time intervals, before and after noon, are considered. The result is that all subscribers $a$, $b$, and $c$ have identical fingerprints $(West,1-12)$ and $(East,13-24)$. Clearly, this level of anonymization comes at a high cost in terms of information loss, as the location data is very coarse both in space and time.

This is precisely the problem of low anonymizability of mobile traffic datasets unveiled by previous works [4, 5]: even guaranteeing 2-anonymization in a very large population requires severe reductions of the spatiotemporal granularity, which limits the usability of the data.
2.2 A measure of anonymizability

We intend to devise a measure of anonymizability that is based on the k-anonymity criterion. Thus, our proposed measure evaluates the effort, in terms of data aggregation, needed to make a user indistinguishable from k-1 other subscribers.

We start by defining the distance between two spatiotemporal samples in the mobile traffic fingerprints of two mobile users. Each sample is composed of a spatial information (e.g., the cell location) and a temporal information (e.g., the timestamp). The distance must keep into account both dimensions. A generic formulation of the distance between the i-th sample of a’s fingerprint, \((s_i^a, t_i^a)\), and the j-th sample of b’s fingerprint, \((s_j^b, t_j^b)\), is

\[
    d_{ab}(i, j) = w_s \delta_s (s_i^a, s_j^b) + w_t \delta_t (t_i^a, t_j^b). \tag{1}
\]

Here, \(\delta_s\) and \(\delta_t\) are functions that determine the distance along the spatial and temporal dimensions, respectively. The former is based on the spatial information in the two samples, \(s_i^a\) and \(s_j^b\), and the latter on the temporal information, \(t_i^a\) and \(t_j^b\). The factors \(w_s\) and \(w_t\) weight the spatial and temporal contributions in (1). In the following, we will assume that the two dimensional have the same importance, thus \(w_s = w_t = 1/2\).

We shape the \(\delta_s\) and \(\delta_t\) functions by considering that both spatial and temporal aggregations induce a loss of information that is linear with the decrease of granularity. However, above a given spatial or temporal threshold, the information loss is so severe that the data is not usable anymore. As a result, the functions can be expressed as

\[
    \delta_s (s_i^a, s_j^b) = \begin{cases} 
    \frac{\text{dist}(s_i^a, s_j^b)}{\delta_s^{\max}} & \text{if dist}(s_i^a, s_j^b) \leq \delta_s^{\max} \\
    1 & \text{otherwise},
    \end{cases} \tag{2}
\]

and

\[
    \delta_t (t_i^a, t_j^b) = \begin{cases} 
    \frac{|t_i^a - t_j^b|}{\delta_t^{\max}} & \text{if } |t_i^a - t_j^b| \leq \delta_t^{\max} \\
    1 & \text{otherwise}.
    \end{cases} \tag{3}
\]

In (2), \text{dist}(s_i^a, s_j^b) = |s_i^a.x - s_j^b.x| + |s_i^a.y - s_j^b.y| is the Taxicab distance [9] between the spatial components of the samples, whose coordinates are denoted as \(x\) and \(y\) in a valid map projection system. Both functions fulfill the properties of distances, i.e., are positive definite, symmetric, and satisfy the triangle inequality. They range from 0 (samples are identical from a spatial or temporal viewpoint) to 1 (samples are at or beyond the maximum meaningful aggregation threshold). Concerning the values of the thresholds, in the following we will consider that the aggregation limits beyond which the information deprivation is excessive are 20 km for the spatial dimension (i.e., the size of a city, beyond which all intra-urban movements are lost) and 8 hours (beyond which the night, working hours, and evening periods are merged together).

The sample distance in (1) can be used to define the distance among the whole fingerprints of two mobile subscribers \(a\) and \(b\), as

\[
    \Delta_{ab} = \begin{cases} 
    \frac{1}{n_a} \sum_{k=1}^{n_a} \min_{h=1}^{n_b} d_{ab}(h, k) & \text{if } n_a \geq n_b \\
    \frac{1}{n_b} \sum_{h=1}^{n_b} \min_{k=1}^{n_a} d_{ab}(k, h) & \text{otherwise}.
    \end{cases} \tag{4}
\]

Here, \(n_a\) and \(n_b\) are the cardinalities of the fingerprints of \(a\) and \(b\), respectively. The expression in (4) takes the longer fingerprint between the two, and finds, for each sample, the sample at minimum distance in the shorter fingerprint. The resulting \(\Delta_{ab}\) is the average among all such sample distances, and \(\Delta_{ab} = \Delta_{ba}, \forall a, b\).

The measure of anonymizability of a generic mobile user \(a\) can be mapped, under the k-anonymity criterion, to the average distance of his fingerprint from those of the nearest \(k\)-1 other users. Formally

\[
    \Delta_a^k = \frac{1}{k-1} \sum_{b \in N_{a}^{k-1}} \Delta_{ab}, \tag{5}
\]

where \(N_{a}^{k-1}\) is the set of \(k-1\) users \(b\) with the smallest fingerprint distances to that of \(a\).

The expression in (5) returns a measure \(\Delta_a^k \in [0, 1]\) that indicates how hard it is to hide subscriber \(a\) in a crowd of \(k\) users. If \(\Delta_a^k = 0\), then the user is already \(k\)-anonymized in the dataset. If \(\Delta_a^k = 1\), the user is completely isolated, i.e., no sample in the fingerprints of all other subscribers is within the spatial and temporal thresholds, \(\delta_s^{\max}\) and \(\delta_t^{\max}\), from any samples of \(a\)’s fingerprint.

3. TWO MOBILE TRAFFIC USE CASES

We employ the proposed measure to assess the level of anonymizability of fingerprints present in two mobile traffic datasets released by Orange in the framework of the Data for Development Challenge. In order to allow for a fair comparison, we preprocessed the datasets so as to make them more homogeneous.

- Ivory Coast. Released for the 2012 Challenge, this dataset describes five months of Call Detail Records (CDR) over the whole the African nation of Ivory Coast. We used the high spatial resolution dataset, containing the complete spatio temporal trajectories for a subset of 50,000 randomly selected users that are changed every two weeks. Thus, the dataset contains information about 10 2-weeks periods overall. We performed
a preliminary screening, discarding the most disperse trajectories, keeping the users that have at least one spatio-temporal point per day. Then, we merged all the user that met this criteria in a single dataset, so as to achieve a meaningful size of around 82,000 users. This dataset is indicated as d4d-civ in the following.

- **Senegal.** The 2014 Challenge dataset is derived from CDR collected over the whole Senegal for one year. We used the fine-grained mobility dataset, containing a randomly selected subset of around 300,000 users over a rolling 2-week period, for a total of 25 periods. We did not filter out subscribers, since the dataset is already limited to users that are active for more than 75% of the 2-week time span. In our study, we consider one representative 2-week period among those available. This dataset is referred to as d4d-sen in the following.

In both the mobile traffic datasets, the information about the user position\(^2\) is provided as a latitude and longitude pair. We projected the latter in a two-dimensional coordinate system using the Lambert azimuthal equal-area projection. We then discretize the resulting positions on a 100-m regular grid, which represents the maximum spatial granularity we consider\(^3\). As far as the temporal dimension is concerned, the maximum precision granted by both datasets is one minute, and this is also our finest time granularity.

4. RESULTS

The measure of anonymizability in (5) can be intended as a dissimilarity measure, and employed in legacy definitions used to understand micro-data database sparsity, e.g., \((\epsilon, 5)\)-sparsity [3]. However, these definitions are less informative than the complete distribution of the anonymizability measure. Thus, in this section, we employ Cumulative Distribution Functions (CDF) of the measure in (5) in order to assess the anonymizability of the two datasets presented before.

4.1 The good: anonymity is close to reach

Our basic result is shown in Fig. 2. The plot portrays the CDF of the anonymizability measure computed on all users in the two reference mobile traffic datasets, d4d-civ and d4d-sen, when considering 2-anonymity as the privacy criterion.

We observe that the two curves are quite similar, and both are at zero in the x-axis origin. This means\(^2\) that no single mobile subscriber is 2-anonymous in either of the original datasets, which confirms previous findings on the elevate re-identifiability of mobile traffic datasets [4, 5]. However, the probability mass gathered in both cases in the 0.1-0.2 range, i.e., it is quite close to the origin. This is good news, since it implies that the average aggregation effort needed to achieve 2-anonymity is not elevate.

As an example, 50% of the users in the d4d-civ dataset have a measure 0.09 or less, which maps, on average, to a combined spatiotemporal aggregation of less than one km and little more than 20 minutes. In other words, the result seems to suggest that half of the individuals in the dataset can be 2-anonymized if the spatial granularity is decreased to 1 km, and the temporal precision is reduced to around 20 minutes. Similar considerations hold in the d4d-sen case, where, e.g., 80% of the dataset population has a measure 0.17 or less. Such a measure is the result of average spatial and temporal distances of 1.7 km and 41 minutes from 2-anonymity.

One may wonder how more stringent privacy requirements affect these results. Fig. 3 shows the evolution of the anonymizability of the two datasets when \(k\) varies from 2 to 100. As expected, higher values of \(k\) require that a user is hidden in a larger crowd, and thus shift
the distributions towards the right, implying the need for a more coarse aggregation. However, quite surprisingly, the shift is not dramatic: 100-anonymity does not appear much more difficult to reach than 2-anonymity.

4.2 The bad: aggregation does not work

Unfortunately, the easy anonymizability suggested by the distributions is only apparent. Fig. 4 depicts the impact of spatiotemporal generalization on anonymizability: each curve maps to a different level of aggregation, from 100 meters and 1 minute (the finer granularity) to 20 km and 8 hours. As one could expect, the curves are pushed towards smaller values of the anonymizability measure. However, the reduction of spatiotemporal precision does not have the desired magnitude, and even a coarse-grained citywide, 8-hour aggregation cannot 2-anonymize but 30% of the mobile users.

This result is again in agreement with previous studies [4, 5], and confirms that mobile traffic datasets are difficult to anonymize.

4.3 The why: long-tailed temporal diversity

We are interested in understanding the reasons behind the incongruity above, i.e., the fact that spatiotemporal aggregation yields such poor performance, even if the average effort needed to attain $k$-anonymity is in theory not elevate.

To attain our goal, we proceed along two directions. First, we separate the spatial and temporal dimensions of the measure in (5), so as to understand their precise contribution to the dataset anonymizability. Second, we measure the statistical dispersion of the fingerprint distances along the two dimensions: the rationale is that we observed the average distance among fingerprints to be quite small, thus the reason of the low anonymizability must lie in the deviation of sample distances around that mean.

4.3.1 Impact of space and time dimensions

Formally, we consider, for each user $a$ in the dataset, the set $N_a^{k-1}$ of $k-1$ other subscribers whose fingerprints are the closest to that of $a$, according to (5). Then, we disaggregate all the fingerprint distances $\Delta_{ab}$, between $a$ and the users $b \in N_a^{k-1}$ into sample distances $d_{ab}$, as per (4). Finally, we separately collect the spatial and temporal components of all such sample distances, in (1), into ordered sets $S_a^k = \{w_a \delta_a\}$ and $T_a^k = \{w_a \delta_t\}$. The resulting sets can be treated as disjoint distributions of the distances, along the spatial and temporal dimensions, between the fingerprint of a generic individual $a$ and those of the $k$-1 other users that show the most similar patterns to his.

Examples of the spatial and temporal distance distributions we obtain in the case of 2-anonymity are shown in Fig. 5a-5e. Each plot refers to one random user in the $d4d$-civ or $d4d$-sen dataset, and portrays the CDF of the spatial ($w_a \delta_a$) and temporal ($w_a \delta_t$) component distance, as well as that of the total sample distance ($d$). We can remark that temporal components typically bring a significantly larger contribution to the total fingerprint distance than spatial ones. In fact, a significant portion of the spatial components is at zero distance, i.e., is immediately 2-anonymous in the original dataset. The same is not true for the temporal components.

A rigorous confirmation is provided in Fig. 5f, which shows the distribution of the temporal-to-spatial component ratios, i.e., $\sum_{T_a} w_a \delta_t / \sum_{S_a} w_a \delta_s$, for all subscribers $a$ in the two reference datasets. The CDF is skewed towards high values, and for half of mobile subscribers in both $d4d$-civ or $d4d$-sen datasets temporal components contribute to 80% or more of the total sample distance. We conclude that the temporal component of a mobile traffic fingerprint is much harder to anonymize than the spatial one. In other words, where an individual generates mobile traffic activity is easily masked, but hiding when he carries out such activity it is not so.

4.3.2 Dispersion of fingerprint sample distances

Not only temporal components weight much more than spatial ones in the fingerprint distance, but they also seem to show longer tails in Fig. 5a-5e. Longer tails imply the presence of more samples with a large distance: this, in turn, significantly increases the level of aggregation needed to achieve $k$-anonymity, as the latter is only granted once all samples in the fingerprint have zero distance from those in the second fingerprint.

We rigorously evaluate the presence of a long tail of hard-to-anonymize samples by means of two complementary metrics, still separating their spatial and temporal components. The first metric is the Gini coefficient, which measures the dispersion of a distribution around its mean. Considering an ordered set $S = \{s_i\}$,
The former show cases where no dispersion at all is observed around 0 (coefficient close to zero), and cases where the distribution is very sparse. The latter has the same behavior as the overall distance, with values clustered around 0.5. The result (i) corroborates the observation that the overall anonymizability is driven by distances along the temporal dimension, and (ii) imputes the latter to the complete absence of easy-to-anonymize short tails in the distribution of temporal distances. Fig. 6d and Fig. 6b show instead the CDF of Tail weight indices. Here, the result is even more clear: the tail of temporal component distances is typically much longer than that of spatial ones, and in between those of exponential and heavy-tailed distributions\(^4\). Once more, the temporal component tail fundamentally shapes that of the overall fingerprint distance.

5. DISCUSSION AND CONCLUSIONS

At the light of all previous observations, we confirm the findings of previous works on user privacy preservation in mobile traffic datasets. Namely, the two datasets we analysed do not grant \(k\)-anonymity, not even for the minimum \(k = 2\). Moreover, our reference datasets show poor anonymizability, i.e., require important spatial and temporal generalization in order to slightly improve the findings of previous works on user privacy preservation in mobile traffic datasets. Namely, the two datasets we analysed do not grant \(k\)-anonymity, not even for the minimum \(k = 2\). Moreover, our reference datasets show poor anonymizability, i.e., require important spatial and temporal generalization in order to slightly improve the findings of previous works on user privacy preservation in mobile traffic datasets.

\(^4\)As a reference, an exponential distribution with mean equal to 1 has a Tail weight index of 1.5, and a Pareto distribution with shape 1 has an Tail weight index of 14.
prove user privacy. The fact that these properties have been independently verified across diverse datasets of mobile traffic suggests that the elevate re-identifiability and low anonymizability are intrinsic properties of this type of datasets.

In our case, even a citywide, 8-hour aggregation is not sufficient to ensure complete 2-anonymity to all subscribers. The result is even worse than that observed in previous studies: the difference is due to the fact that we consider the anonymization of complete subscriber fingerprints, whereas past works focus on simpler obfuscation of summaries [4] or subsets [5] of the fingerprints.

Our analysis also unveiled the reasons behind the poor anonymizability of the mobile traffic datasets we consider.

On the one hand, the typical mobile user fingerprint in such datasets is composed of many spatiotemporal samples that are easily hidden among those of other users in the dataset. This leads to fingerprints that appear easily anonymizable, since their samples can be matched, on average, with minimal spatial and temporal aggregation.

On the other hand, mobile traffic fingerprints tend to have a non-negligible number of elements that are much more difficult to anonymize than the average sample. These elements, which determine a characteristic dispersion and long-tail behavior in the distribution of fingerprint sample distances, are mainly due to a significant diversity along the temporal dimension. In other words, mobile users may have similar spatial fingerprints, but their temporal patterns typically contain a non-negligible number of dissimilar points.

It is the presence of these hard-to-anonymize elements in the fingerprint that makes spatiotemporal aggregation scarcely effective in attaining anonymity. Indeed, in order to anonymize a user, one needs to aggregate over space and time, until all his long-tail samples are hidden within the fingerprints of other subscribers. As a result, even significant reductions of granularity (and consequent information losses) may not be sufficient to ensure individual privacy in mobile traffic datasets.

6. REFERENCES

Human Mobility during Religious Festivals in Senegal: An Orange Mobile Dataset Analysis

Christelle Scharff, Ph.D. (Principal Investigator)
Pace University, New York, NY, USA
Seidenberg School of Computer Science and Information Systems
Department of Computer Science
cscharff@pace.edu

Khadidiatou Ndiaye, Ph.D. (co-PI)
George Washington University, Washington DC, USA
Milken Institute School of Public Health & Health Services
Department of Global Health
kndiaye@email.gwu.edu

Aminata Niang Diene, Ph.D. (co-PI)
University of Cheikh Anta Diop of Dakar (UCAD), Dakar, Senegal
Geography of Health, Geography Department
aminaniang@orange.sn

Fatou Maria Drame, Ph.D. (co-PI)
Gaston Berger University, St Louis, Senegal
Geography of Health, Social Science Department
fatou-maria.drame@ugb.edu.sn

Meghan Jordan, MS Computer Science Candidate
Pace University, New York, NY, USA
Seidenberg School of Computer Science and Information Systems
Department of Computer Science
meghan.jordan@pace.edu

Briana Vecchione, BS Computer Science Candidate
Pace University, New York, NY, USA
Seidenberg School of Computer Science and Information Systems
Department of Computer Science
bv26460n@pace.edu
Table of Contents

1. Introduction
2. Religious Festivals in Senegal
3. Methodology
3.1. Orange Data
3.2. SNEIPS Health Data
3.3. Tools
3.4. Methodology
4. Findings Related to the Orange Data
4.1. Gamou of Tivaouane
4.2. Magal of Touba
4.3. Analysis
5. Findings Related to the SNEIPS Health Data
5.1. Calls to the Numéro Vert in 2011
5.1. Calls to the Numéro Vert in 2012
5.3. Analysis
6. Conclusions and Discussions

Abstract

Mobile phone records are particularly useful to study human mobility. We use the 2013 mobile phone records released by the Orange operator in Senegal as part of the 2014 Data for Development Challenge to model human mobility before, during and after two key religious festivals in Senegal that are attended by millions of Muslim Senegalese: Gamou of Tivaouane for followers of the Tijaniyya brotherhood and Magal of Touba for the followers of the Mouridiyya brotherhood. To study the correlation between human mobility and health issues we also secured access to health data collected through a toll free hotline service (Numéro Vert) of the Ministry of Health. We found interesting structures in the human mobility patterns showing that these festivals imply massive movements of population from different parts of Senegal depending on the festival and permitting to identify the interconnectedness of communities. Our analysis also showed the main routes used by the pilgrims and their travels’ times. While we did not find significant patterns in the health data, it is important to understand the reasons and origins of the calls to the Numéro Vert and integrate it as a tool for awareness health campaigns before and during the festivals and beyond. These findings will be worthwhile for numerous structures, including the ministries of Transport, Health, and Hydraulic, as well as other stakeholders planning future religious festivals. They have important implications ranging from resource management to service allocation and awareness campaigns during religious festivals in Senegal.
1. Introduction

Availability of large datasets of mobile phone records inspires researchers from different domains including Physics, Computer Science, Economics, Public Health, Anthropology and Sociology, to study human dynamics, a branch of Complex Systems [1]. Mobile phone records are particularly useful to study human behavior by providing temporal and spatial information at scale and different levels of granularity [2]. They can be used to model mobility and migration patterns [3]. Understanding these patterns is crucial for urban planning, transport infrastructure design, analysis of communities’ networks, spreading and mitigation of diseases, disaster management, etc.

We use the 2013 mobile phone records released by the Orange operator in Senegal as part of the 2014 Data for Development Challenge (D4D 2014, http://www.d4d.orange.com) to model human mobility before, during and after two of the most celebrated religious festivals in Senegal that are attended by millions of Senegalese: Gamou of Tivaouane and Magal of Touba. Our aim is to investigate mobility and transport patterns by studying the changes of communication volume between Tivaouane and Touba respectively and other locations in Senegal.

We also secured access to data collected through a toll free hotline service (Numéro Vert) of the Senegalese National Health Service of Education and Information (Service National de l’Education et l’Information pour la Santé, SNEIPS) from the Ministry of Health. Senegalese call this toll free hotline service to get information about health issues they face or need advice about on a daily basis. We obtained the data for 2010, 2011 and 2012. The Ebola situation in West Africa and the complete mobilization of the Ministry of Health to control the disease prevented us from obtaining the data for 2013 on time for analysis in this report. We explored the correlation between human mobility and health issues during the period surrounding the festivals, which are often subject to health issues such as cholera and typhoid outbreaks. The last cholera outbreak happened during the Magal of Touba in 2008 with a total of 2,054 cases and 8 deaths reported by the Senegalese Ministry of Health [4]. This is due to important crowding of population, poor sanitation, lack of storage for clean water and food, and lack of septic systems in the cities hosting the festivals. While we did not find significant patterns in the health data, it is important to understand the reasons and origins of the calls to the Numéro Vert and integrate it as a tool for awareness health campaigns before and during the festivals and beyond.

We found interesting structures in the human mobility patterns showing that these festivals imply massive movements of population from different parts of Senegal depending on the festival and permitting to identify the interconnectedness of communities. We provide animated map data visualizations depicting these movements of population that can be easily interpreted and conveyed to any audience, from the Senegalese population to data scientists. These visualizations are the first ones available for understanding the scope of festivals in Senegal. We believe that our results will contribute to the multiple ways Big Data and data science can be used to understand a range of experiences. These findings will be worthwhile for numerous structures, including the ministries of Transport, Health, and Hydraulic, as well as other stakeholders planning future religious festivals. These findings have important implications ranging from resource management to service allocation and awareness campaigns during religious festivals in
Senegal. Understanding how and when people are leaving for and returning from festivals is important to plan awareness campaigns (e.g., health and road safety) and target messages types and contents (e.g., TV, radio and print). Our visualizations can be used to raise awareness amongst the population and plan actions based on tangible evidence. Planning is currently not based on accurate data availability and analysis. With the ubiquity of mobile phones in Senegal (mobile penetration of 93% in 2013) and the use of mobile phone records planning can be done in a more systematic manner. Our visualizations provide present data in a stimulating way to captivate and reach people from different backgrounds.

This report is organized as follows. Section 2 presents the two considered religious festivals: Magal of Touba and Gamou of Tivaouane, and provides the background necessary to understand the relevance of the choice of this study. Section 3 describes the methodology we used to analyze the data. Section 4 and Section 5 present our findings based on the Orange and health data respectively. Section 6 concludes, discussed our results, and presents our future work.

2. Religious Festivals in Senegal

Our interest in the topic is based on the importance of the two considered religious festivals in the life of Senegalese in country and abroad. Senegal has a population of 14.13 million in 2013. The total number of mobile subscribers is 12.661 million with a penetration rate of around 93% and a total volume of 2.3 billion minutes [5]. Three operators, Orange (58.34%), Tigo (20.92%) and Expresso (20.74%), share the mobile subscribers. Internet penetration is around 9% and Internet usage is dominated by mobile data with over 70% users via 3G since 2008 and 4G more recently in 2014. In this study we are interested in Tivaouane (92 km from Dakar) and Touba (182 km from Dakar), the two cities of the considered religious festivals. The populations of Tivaouane and Touba are 40,500 and 620,500 in 2010, respectively, compared to 2,396,800 in Dakar [6]. Islam is the predominant religion, practiced by 95.4% of the population. Christians (4.2%) include Roman Catholics and diverse Protestant denominations. There is also a 0.4% population who maintain traditional African religions [7]. Tijaniyya (4 to 6 million) and Mouridiyya (3 to 5 million) are the two largest Muslim Sufi brotherhoods in Senegal. Sufism distinguished itself from other branches of Islam by the presence of spiritual guides called Marabouts in Senegal. The Tijaniyya brotherhood is the largest in number and Mouridiyya is the most active. The Gamou of Tivaouane and the Magal of Touba are the main festivals of these brotherhoods. Their dates depends on the Muslim calendar and, thus, change every year. Pilgrims, called Talibés, join from different places in Senegal and abroad, and travel by car, bus, motorbikes, trucks, carriages and foot. They are hosted in the holy cities by local inhabitants. The festivals revolve around prayers and meditations, but also official ceremony with government representatives. Mobile service operators have strategic roles during these festivals: they are present to advertise and sell their products and responsible in providing quality services. For example, in 2013, Orange deployed 40 temporary 3G stations and 187 voice and SMS stations in Touba and its neighbor cities and ADSL lines to transmit the festival live on TV and the Internet [8]. ARTP, the regulator, is also monitoring the quality of service. Festivals have an important economic impact at the micro and macro levels in Senegal. In Touba a pilgrim usually spends around 92,000 FCFA for the Magal including transport, phone, religious items, clothes and hair, and contributions and Touba contributes 250 milliards of FCFA (around $450 million).
to Senegal economy [9]. The informal sector benefits greatly from the religious festivals (e.g., tailors, hairdressers, mattress sellers, house renovations). The agro-alimentary industry is greatly involved as water and food need to be available for pilgrims [10,11]. Banks and financial institutions are also part of the equation. Money is sent to pilgrims by family in Senegal and abroad (e.g., through mobile money mechanisms).

In 2013, the 111th edition of Maouloud (Gamou in Wolof, the main language of Senegal), took place in Tivaouane on January, 23rd 2013. The Maouloud celebrates the birth of the prophet Muhammad (571 in Mecca). In Senegal, the Tijaniyya brotherhood was spread by the savant and religious leader Cheikh El Hadji Malick Sy (1855-1922), who created the celebration of the Gamou of Tivaouane. After it was initiated in Saint Louis, it has been taking place in Tivaouane since the beginning of 1900. At that time, Tivaouane became a center for Islamic education and culture. Others Gamous are celebrated around the country, in particular Dakar, Kaolack and Thies.

The Grand Magal, also called the Magal of Touba, is a highly regarded Muslim festival followed by Mourides that takes place in the city of Touba. Magal means to pay respect, celebrate and magnify in Wolof. People attend the festival in the name of Cheikh Amadou Bamba (1853-1927), a powerful Islamic leader who founded the Mouridiyya brotherhood, preached non-violence, and returned to the city of Touba after being exiled due to colonial rules. The festival is celebrated on the day of the exile of Cheikh Amadou Bamba to Gabon (1895). Touba is in the department of Mbacke in the Region of Diourbel. Two Magals took place in 2013, in January and December. We only considered the Magal of December 22, 2013 in this study.

Movements of population during festivals create important problems with crowded and insecure roads with numerous accidents and deaths. They create unsanitary conditions that are subjects to epidemics. In terms of transportation, while it normally takes around 4 hours to reach Touba from Dakar (192 km), it may take more than 12 hours close to the festival day. This is due to traffic congestion and accidents. Accidents and deaths occur due in general to traffic violations, lack of vehicle and road maintenance, two-way roads, etc. During festivals, firefighters, ambulances, police and security agents are mobilized to improve the situation. In terms of health, campaigns on TV and radio focus on hand washing, especially after numerous Salaam Aleykum greetings involving handshaking, a practice in Senegal, and avoidance of street food, food conservation for prolonged periods, and avoidance of water storage [10,11].

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tivaouane</td>
<td>40,500</td>
<td>01/23/2013</td>
<td>1 - 2 million</td>
<td>2 deaths, 70 injured, 38 interventions, 36 accidents within the city [12]</td>
</tr>
</tbody>
</table>
3. Methodology

Our analysis focused on the 2013 Call Detail Record (CDR) data released by Orange for the Data for Development Challenge (D4D, http://d4d.orange.com) for Senegal and the 2011 and 2012 health data provided by the National Service of Education and Information (SNEIPS) of the Ministry of Health and collected through a toll free hotline service (Numéro Vert). We did not study the 2010 health data. We subsetted the data based on the dates and cities of the festivals we considered.

- Cities: Tivaouane and Touba
- Festivals: Gamou of Tivaouane (01/23/2013) and Magal of Touba (12/22/2013)

We modeled human mobility before, during and after two of the most celebrated religious festivals in Senegal based on volumes and changes of volumes of communications between Tivaouane and Touba respectively and other locations in Senegal. We used plots to show the frequencies of calls and SMS during specific periods and animated map visualizations to show human mobility. We explored the correlation between human mobility and health issues during festivals using the SNEIPS health data by looking at the calls during the period surrounding the festivals, the reasons of the calls, and the cities where callers were located.

3.1. Orange Data

We used SET1 files related to voice and SMS. These Orange data show the calls between two sites and provide the number of calls and total call duration on an hourly basis. In the case of SMS, it provides the number of SMS on an hourly basis. One year of site-to-site traffic for 1666 sites on an hourly basis are provided. The headers of SET1 data are provided below (Tables 2 and 3). We subsetted the data based on the dates and cities of the festivals we considered. We considered seven days before and after the festivals. For the cities we considered, the SITE_IDs are described on the maps below (Figures 1 and 2). We used the 4 sites of Tivaouane and 6 sites for Touba. For Touba, we chose the 6 sites in the center of city out of the 47 potential sites for the city. Table 4 summarizes the dates, cities, data sets, SITE_IDs, and number of records we considered.

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Outgoing_site_id</th>
<th>Incoming_site_id</th>
<th>Number_of_calls</th>
<th>Total_call_duration</th>
</tr>
</thead>
</table>

Table 2. Header of the SET1 Orange Data (Voice)
Table 3. Header of the SET1 Orange Data (SMS)

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Outgoing_site_id</th>
<th>Incoming_site_id</th>
<th>Number_of_SMS</th>
</tr>
</thead>
</table>

Figure 1 - Sites in Tivaouane (SITE_ID: 604, 605, 606, 609)
<table>
<thead>
<tr>
<th>Festival</th>
<th>Date</th>
<th>7 Days before Festival</th>
<th>7 Days after Festival</th>
<th>Orange Datasets</th>
<th>Number of Sites</th>
<th>Number of Orange Records +/- 7 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamou of Tivaouane</td>
<td>01/23/2013</td>
<td>01/16/2013</td>
<td>01/30/2013</td>
<td>Calls SET1/SET 1V_01.CS</td>
<td>4</td>
<td>Calls 839,248</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SMS</td>
<td></td>
<td>SMS 398,273</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SET1/SET 1S_01.CS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(604, 605, 606, 609)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Orange Data Used to Analyze the 2013 Senegalese Religious Festivals

| Magal of Touba | 12/22/2013 | 12/15/2013 | 12/29/2013 | Calls SET1/SET 1V_12.CSV | SMS SET1/SET 1S_12.CSV | 6* (1043, 1046, 1049, 1050, 1054, 1055) * Center of Touba only | Calls 708,476 SMS 216,343 |

3.2. SNEIPS Health Data

The health data were obtained from the National Service of Education and Information (SNEIPS) of the Ministry of Health. They were collected through a toll free hotline service (Numéro Vert). The calls were recorded by agents in charge of providing information and advice based on the questions and conditions described by callers. We provide the header of the SNEIPS data in Table 5. We formatted (e.g., dates) and corrected (e.g., grammatical and orthographic mistakes) the data. We also translated them from French to English. Table 6 provides a snapshot of the data for 2011 and 2012 and describes the number of female and male callers, as well as the average and median age of the callers. The Ebola situation in West Africa and the mobilization of the Ministry of Health prevented us from obtaining the data for 2013.

Table 5. Header of the SNEIPS Health Data

<table>
<thead>
<tr>
<th>Region</th>
<th>District</th>
<th>Age</th>
<th>Gender</th>
<th>Marital Status</th>
<th>Date</th>
<th>Reason</th>
<th>Verbatim</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Number of Calls</th>
<th>Calls by Females</th>
<th>Calls by Males</th>
<th>Average Age of Callers</th>
<th>Median Age of Callers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>20,106</td>
<td>7115</td>
<td>12991</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>2012</td>
<td>11,944</td>
<td>4017</td>
<td>7927</td>
<td>21</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 6. Snapshot of the SNEIPS Health Data for 2011 and 2012

3.3. Tools

We used Amazon Cloud (http://aws.amazon.com/ec2) for storage and processing of our data. Our language for data manipulation was R (version 3.1.2, http://cran.r-project.org) and we
We explored the cloud-based analysis the CartoDB georeferencing. Screenshots of mobility population between the festivals. Numéro Vert is often subject to health issues. The last cholera outbreak happened during the Magal of Touba in 2008 with a total of 2,054 cases and 8 deaths reported by the Senegalese Ministry of Health [4]. This is due to overcrowding, poor sanitation, lack of storage for clean water and food, and lack of septic systems in the cities hosting the festival.

We created plots of the SNEIPS health data using R to show the frequency of calls to the Numéro Vert in 2011 and 2012. We also determined the top reasons and regions of the calls seven days before, during and after the festivals to observe differences. We used bubble map

3.4. Methodology

We modeled human mobility during two of the most celebrated religious festivals in Senegal using the changes of communication volume between Tivaouane and Touba respectively and other locations in Senegal.

We created plots of the Orange data using R to show the frequency of calls and SMS during the months of the festivals to observe changes. We distinguished incoming and outgoing calls and SMS in / to Tivaouane and Touba, and compared and contrasted the phenomenon in the two targeted cities. We looked at the difference between the volume of calls and SMS.

Our goal was also to provide visualizations to present data in a stimulating way to captivate and reach people from different backgrounds. CartoDB permitted us to produce and publish animated map visualizations that showed the mobility of the population based on dates. We manipulated the data using SQL to get desired views in CartoDB.

We used torque map visualizations of the Orange data. Torque visualizations display temporal data on maps and permit to see the progression of points (numbers of calls per sites) based on dates. Screenshots of the animated visualizations are provided in this report. We produced screenshots of numbers of calls for two chosen days before, during and after the festivals.

We explored the correlation between mobility of the population and health issues during festivals. Festivals are often subject to health issues. The last cholera outbreak happened during the Magal of Touba in 2008 with a total of 2,054 cases and 8 deaths reported by the Senegalese Ministry of Health [4]. This is due to overcrowding, poor sanitation, lack of storage for clean water and food, and lack of septic systems in the cities hosting the festival.

We created plots of the SNEIPS health data using R to show the frequency of calls to the Numéro Vert in 2011 and 2012. We also determined the top reasons and regions of the calls seven days before, during and after the festivals to observe differences. We used bubble map

We received a CartoDB grant (http://cartodb.com) for this project that consisted of 1GB storage, unlimited tables and georeferencing. CartoDB is a cloud-based mapping, analysis and visualization tool that generates spatial and temporal applications for both mobile and the web.

3.4. Methodology

We modeled human mobility during two of the most celebrated religious festivals in Senegal using the changes of communication volume between Tivaouane and Touba respectively and other locations in Senegal.

We created plots of the Orange data using R to show the frequency of calls and SMS during the months of the festivals to observe changes. We distinguished incoming and outgoing calls and SMS in / to Tivaouane and Touba, and compared and contrasted the phenomenon in the two targeted cities. We looked at the difference between the volume of calls and SMS.

Our goal was also to provide visualizations to present data in a stimulating way to captivate and reach people from different backgrounds. CartoDB permitted us to produce and publish animated map visualizations that showed the mobility of the population based on dates. We manipulated the data using SQL to get desired views in CartoDB.

We used torque map visualizations of the Orange data. Torque visualizations display temporal data on maps and permit to see the progression of points (numbers of calls per sites) based on dates. Screenshots of the animated visualizations are provided in this report. We produced screenshots of numbers of calls for two chosen days before, during and after the festivals.

We explored the correlation between mobility of the population and health issues during festivals. Festivals are often subject to health issues. The last cholera outbreak happened during the Magal of Touba in 2008 with a total of 2,054 cases and 8 deaths reported by the Senegalese Ministry of Health [4]. This is due to overcrowding, poor sanitation, lack of storage for clean water and food, and lack of septic systems in the cities hosting the festival.

We created plots of the SNEIPS health data using R to show the frequency of calls to the Numéro Vert in 2011 and 2012. We also determined the top reasons and regions of the calls seven days before, during and after the festivals to observe differences. We used bubble map

We received a CartoDB grant (http://cartodb.com) for this project that consisted of 1GB storage, unlimited tables and georeferencing. CartoDB is a cloud-based mapping, analysis and visualization tool that generates spatial and temporal applications for both mobile and the web.

3.4. Methodology

We modeled human mobility during two of the most celebrated religious festivals in Senegal using the changes of communication volume between Tivaouane and Touba respectively and other locations in Senegal.

We created plots of the Orange data using R to show the frequency of calls and SMS during the months of the festivals to observe changes. We distinguished incoming and outgoing calls and SMS in / to Tivaouane and Touba, and compared and contrasted the phenomenon in the two targeted cities. We looked at the difference between the volume of calls and SMS.

Our goal was also to provide visualizations to present data in a stimulating way to captivate and reach people from different backgrounds. CartoDB permitted us to produce and publish animated map visualizations that showed the mobility of the population based on dates. We manipulated the data using SQL to get desired views in CartoDB.

We used torque map visualizations of the Orange data. Torque visualizations display temporal data on maps and permit to see the progression of points (numbers of calls per sites) based on dates. Screenshots of the animated visualizations are provided in this report. We produced screenshots of numbers of calls for two chosen days before, during and after the festivals.

We explored the correlation between mobility of the population and health issues during festivals. Festivals are often subject to health issues. The last cholera outbreak happened during the Magal of Touba in 2008 with a total of 2,054 cases and 8 deaths reported by the Senegalese Ministry of Health [4]. This is due to overcrowding, poor sanitation, lack of storage for clean water and food, and lack of septic systems in the cities hosting the festival.

We created plots of the SNEIPS health data using R to show the frequency of calls to the Numéro Vert in 2011 and 2012. We also determined the top reasons and regions of the calls seven days before, during and after the festivals to observe differences. We used bubble map
visualizations of the health data, which scale the data and permit to compare values (numbers of calls) based on locations (cities). We looked at trends in the 2011 and 2012 data to see if we could establish patterns for 2013.

4. Findings Related to the Orange Data

In this section, we provide plots and screenshots of the animated map visualizations showing human mobility during the Gamou of Tivaouane and the Magal of Touba. We also present our findings and compare the patterns we extracted from the two religious festivals.

4.1. Gamou of Tivaouane

*Calls from Tivaouane*

The CartoDB animated torque visualization showing calls from Tivaouane is available at [http://cdb.io/13s0Kn6](http://cdb.io/13s0Kn6). Screenshots are presented in Figure 3. They show the volume and intensity of phone call communications originating from Tivaouane seven days before, the day of, and seven days after the Gamou of Tivaouane. Darkest colors show where there is the highest concentration of phone calls.

The 2013 Gamou of Tivaouane took place on January 23rd 2013. We can see an increase of calls from Tivaouane starting 2-3 days before the festival. Another increase appears on the day of the festival and decrease just after. The festival takes place on the night between January 22nd and January 23rd. Communication on January 29th is similar to communications on January 18th. During peak, communications are intense all over Senegal with the main cities of Senegal: Dakar, Thies, Touba, Kaolack, Tambacounda, Louga, Saint Louis, Mbour, Diourbel, and Ziguinchor. The roads used by pilgrims are visible: Diourbel - Bambey - Thies - Tivaouane, Dakar - Thies - Tivaouane, Saint Louis - Louga - Tivaouane, Mbour - Thies - Tivaouane, Kaolack - Fatick - Mbour - Thies - Tivaouane, Tambacounda - Fatick - Mbour - Thies - Tivaouane, and Ziguinchor - Kaolack - Fatick - Mbour - Thies. Thies, Mbour, Saint Louis, Kaolack and Louga are traffic hubs. Most of the traffic is on the axis Dakar - Thies - Tivaouane. Senegalese pilgrims from the borders with Mauritania and Mali seem to use the National 2 road that reaches Saint Louis to attend the festival. The phone calls permit to observe human mobility. This hypothesis is confirmed by further investigations. Pilgrims are attending the festival in Tivaouane (See Figure 4) and call people from their regions and people who are on their way. They also call within the city of Tivaouane itself (See Figure 7).
Figure 3 shows the frequency of calls originating from Tivaouane during January 2013. It indicates clearly that the highest concentration of phone call communications out of Tivaouane take place on the day of the festival. Communications are more than ten times higher during the festival. This is due to the pilgrims who are attending the festival in Tivaouane.

Figure 4 shows the frequency of calls originating from Tivaouane during January 2013. It indicates clearly that the highest concentration of phone call communications out of Tivaouane take place on the day of the festival. Communications are more than ten times higher during the festival. This is due to the pilgrims who are attending the festival in Tivaouane.
Calls to Tivaouane

The torque visualizations for calls to Tivaouane can be found at: http://cdb.io/1zPc5tf. Figures 5 show the volume of phone call communications to Tivaouane seven days before, during and seven days after the Gamou of Tivaouane. Communications to Tivaouane are following the same patterns as communications from Tivaouane.
Figure 6 shows the frequency of calls to Tivaouane during January 2013. It indicates clearly that the highest concentration of phone call communications to Tivaouane takes place on the day of the festival.

**Figure 6. Frequency of calls to Tivaouane (Gamou of Tivaouane on 01/23/2013)**
Calls within Tivaouane

During the Gamou of Tivaouane, communications within Tivaouane are also very high as indicated on Figure 7. The number of calls is around four times higher than before and after the festival. Pilgrims begin to be present in Tivaouane two days before the festival and leave right after.

**Figure 7. Frequency of calls to Tivaouane (Gamou of Tivaouane on 01/23/2013)**

SMS from Tivaouane

In addition to phone call communications, we examined SMS communications. The following three figures (Figures 8, 9 and 10) show SMS communications with SMS originating from Tivaouane, sent to Tivaouane, and exchanged within Tivaouane. We notice that the highest number of SMS is again on the day of the festival.
Figure 8. Frequency of SMS from Tivaouane (Gamou of Tivaouane on 01/23/2013)

SMS to Tivaouane

Figure 9. Frequency of SMS to Tivaouane (Gamou of Tivaouane on 01/23/2013)
SMS within Tivaouane

Figure 10. Frequency of SMS within Tivaouane (Gamou of Tivaouane on 01/23/2013)

4.2. Magal of Touba

Calls from Touba

Figures 11 were obtained from the animated CartoDB torque visualization available at: http://cdb.io/13q15Ha. They show the volume of phone call communications to Touba seven days before, during and seven days after the Magal of Touba. Communications begin to increase on December 18th, four days before the festival and decrease on December 23rd. The highest number of communications takes place on the day of the festival.

The roads used by pilgrims are visible: Dakar - Thies - Diourbel - Touba, Saint Louis - Louga - Thies - Diourbel - Touba, Kaolack - Dioubel - Touba, and Fatick - Bambey - Diourbel - Touba. Some smaller roads are also used by pilgrims. The phone calls permit to observe human mobility during the festival of the Magal of Touba. Thies, Diourbel and Louga are hubs with heavy traffic volumes.
Figure 11 - Calls from Touba in CartoDB (Magal of Touba on 12/22/2013)

Figure 12 shows phone call communications from Touba. The highest frequency of calls takes place on the day of the festival, December 22nd 2013. Communications are about seven times higher than during regular days. This shows that there is a massive number of pilgrims who attend the Magal of Touba.

**Frequency of calls with Touba as originating city**

![Bar chart showing frequency of calls](Image)

Figure 12. Frequency of calls from Touba (Magal of Touba on 12/22/2013)
**Calls to Touba**

The animated Torque visualization of the calls to Touba is available at: [http://cdb.io/1vcpYgV](http://cdb.io/1vcpYgV). In the Figures 13, we show the volume of phone call communications to Touba seven days before, during and seven days after the Magal of Touba. The patterns are similar to the calls from Touba. Figure 14 shows that lots of calls are made to Touba on 12/22/2013.
Figure 13 - Calls to Touba in CartoDB (Magal of Touba on 12/22/2013)

Figure 14. Frequency of calls to Touba (Magal of Touba on 12/22/2013)
**Calls within Touba**

This plot shows phone call communications within Touba. The highest frequency of calls takes place on the day of the festival, December 22nd 2013. Please note that we are only considering 7 sites in Touba (out of the 47 possible ones).

![Frequency of calls with Touba as receiving and originating city](image)

**Figure 15. Frequency of calls within Touba (Magal of Touba on 12/22/2013)**

**SMS from Touba**

We examined the SMS communications from, to and within Touba, and noticed that the highest number of messages was sent on the day of the festival.
Figure 16. Frequency of SMS from Touba (Magal of Touba on 12/22/2013)

SMS to Touba

Figure 17. Frequency of SMS to Touba (Magal of Touba on 12/22/2013)
SMS within Touba

**Figure 18. Frequency of SMS within Touba (Magal of Touba on 12/22/2013)**

**4.3. Analysis**

The plots of the frequency of phone calls and SMS to, from and within Touba and Tivaouane show that the number of phone calls and SMS increases dramatically during the festivals (from four to ten times more phone calls and SMS on the days of the festivals). The range of the numbers indicates that the number of people increases in the cities themselves. Pilgrims are attending the festivals massively.

It is also noticeable that there are more calls than SMS to, from and within Touba and Tivaouane. This may be due to the fact that communications are discounted during the festivals and, also, that literacy rate is low in Senegal (39.30% in 2013).

Pilgrims are contacting people from their cities and villages of origin and people who are on their ways to the festivals. Their journey is clearly visible on the torque map visualizations. These animated visualizations permit to observe when pilgrims are initiating their travel and returning, the routes they used to reach Touba and Tivaouane, and what regions they come from. Thies, Mbour, Saint Louis, Kaolack and Louga are traffic hubs of the Gamou of Tivaouane. Diourbel, Thies, and Louga are traffic hubs of the Magal of Touba.
Communications during the Gamou of Tivaouane are much more scattered than the communications during the Magal of Touba suggesting that pilgrims of the Gamou of Tivaouane come from across Senegal. Communications during the Magal of Touba are from the wide part of Senegal on the West of Touba and above Gambia. Communications during the Gamou of Tivaouane come from the West of Senegal as well as from the center and from the borders with Mauritania and Mali.

<table>
<thead>
<tr>
<th>Festival</th>
<th>Dates (/-7 Days)</th>
<th>Number of Outgoing Calls</th>
<th>Duration of the Outgoing Calls (seconds)</th>
<th>Number of Incoming Calls</th>
<th>Duration of the Incoming Calls (seconds)</th>
<th>Number of Incoming SMS</th>
<th>Number of Outgoing SMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamou of Tivaouane</td>
<td>01/23/2013</td>
<td>2,759,807</td>
<td>134,220,442</td>
<td>2,877,134</td>
<td>144,732,852</td>
<td>1,647,497</td>
<td>1,605,929</td>
</tr>
<tr>
<td>Magal of Touba</td>
<td>12/22/2013</td>
<td>1,477,090</td>
<td>79,254,650</td>
<td>1,226,886</td>
<td>62,734,055</td>
<td>486,145</td>
<td>458,237</td>
</tr>
</tbody>
</table>

Table 7. Number of Calls & SMS and Duration of the Calls during the Religious Festivals in 2013

5. Findings Related to the SNEIPS Health Data

In this section, we analyze the SNEIPS health data we obtained from the Ministry of Health. We used different ways to analyze the health data for 2011 and 2012. We provide:

- Scatterplots that show the number of calls for 2011 and 2012
- Plots of the number of calls to the Numéro Vert seven days before, during and seven days after the Gamou of Tivaouane and the Magal of Touba in 2011 and 2012
- Maps of the number of calls to the Numéro Vert in 2011 and 2012
- Tables showing the main reasons and regions of the calls in 2011 and 2012 and seven days before, during and seven days after the Gamou of Tivaouane and the Magal of Touba in 2011 and 2012

5.1. Calls to the Numéro Vert in 2011

In 2011, the calls to the Numéro Vert are mainly coming from Dakar, Pikine, Rufisque, Mbacke, Diourbel, and Touba. The number of calls from Tivaouane is less than 1% of the calls of 2011 and thus not noticeable on the map (Figure 20). During the Gamou of Tivaouane and the Magal of Touba we cannot notice any particular increase of calls to the Numéro Vert. There are more calls during the Magal of Touba than during the Gamou of Tivaouane. The main reasons of the calls are to get information about stomach ache, headache, HIV/AIDS and family planning. After the festivals we can see that there is a higher number of calls regarding stomach ache, but it is not really significant.
Figure 19. Frequency of calls to the Numéro Vert for 2011
Figure 20. Mapping of the calls to the Numéro Vert for 2011 (See http://cdb.io/1zqjufE)

Frequency of calls per day for the 2011 Gamou
(Gamou of Tivaouane on 02/15/2011)

Figure 21. Frequency of calls to the Numéro Vert -/+7 days of the 2011 Gamou of Tivaouane
Figure 22. Frequency of calls to the Numéro Vert +/-7 days of the 2011 Magal of Touba

<table>
<thead>
<tr>
<th>Festival</th>
<th>Date</th>
<th>7 Days before Festival</th>
<th>7 Days after Festival</th>
<th>Number of calls -7 Days</th>
<th>Number of calls +7 Days</th>
<th>Number of calls during festival</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamou of Tivaouane</td>
<td>02/15/2011</td>
<td>02/08/2011</td>
<td>02/22/2011</td>
<td>395</td>
<td>515</td>
<td>59</td>
</tr>
<tr>
<td>Magal of Touba</td>
<td>01/06/2011</td>
<td>12/30/2010</td>
<td>01/13/2011</td>
<td>744</td>
<td>735</td>
<td>143</td>
</tr>
</tbody>
</table>

Table 8. Numéro Vert Data for the Religious Festivals in 2011
## Table 9. Snapshot of the Reasons of the Calls to the Numéro Vert in 2011

<table>
<thead>
<tr>
<th>Total phone calls</th>
<th>Top 5 phone calls reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>20,106</td>
<td>Family planning</td>
</tr>
<tr>
<td></td>
<td>1,753</td>
</tr>
<tr>
<td></td>
<td>Other*</td>
</tr>
<tr>
<td></td>
<td>1,427</td>
</tr>
<tr>
<td></td>
<td>Stomach ache</td>
</tr>
<tr>
<td></td>
<td>1,410</td>
</tr>
<tr>
<td></td>
<td>Headaches</td>
</tr>
<tr>
<td></td>
<td>1,334</td>
</tr>
<tr>
<td></td>
<td>HIV/AIDS</td>
</tr>
<tr>
<td></td>
<td>1,055</td>
</tr>
<tr>
<td></td>
<td>* reported by the agent as Other</td>
</tr>
</tbody>
</table>

*Table 9. Snapshot of the Reasons of the Calls to the Numéro Vert in 2011*

<table>
<thead>
<tr>
<th>Top reasons &amp; regions of the calls</th>
<th>-7 days before the festival</th>
<th>+7 days after the festival</th>
<th>During the festival</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamou of Tivaouane 02/15/2011</td>
<td>other* : 30</td>
<td>stomach ache: 44</td>
<td>stomach ache: 7</td>
</tr>
<tr>
<td></td>
<td>stomach ache: 25</td>
<td>headaches : 37</td>
<td>kidney pain : 5</td>
</tr>
<tr>
<td></td>
<td>aids : 24</td>
<td>other : 37</td>
<td>other : 4</td>
</tr>
<tr>
<td></td>
<td>pain : 22</td>
<td>aids : 29</td>
<td>aids : 3</td>
</tr>
<tr>
<td></td>
<td>headaches : 19</td>
<td>information : 28</td>
<td>joint pain : 3</td>
</tr>
<tr>
<td></td>
<td>dizziness : 13</td>
<td>diarrhea : 24</td>
<td>pain : 3</td>
</tr>
<tr>
<td></td>
<td>Diourbel:135</td>
<td>Diourbel :183</td>
<td>Diourbel :23</td>
</tr>
<tr>
<td></td>
<td>Thies : 67</td>
<td>Dakar :124</td>
<td>Dakar : 8</td>
</tr>
<tr>
<td></td>
<td>Dakar : 66</td>
<td>Thies : 57</td>
<td>Kaolack : 7</td>
</tr>
<tr>
<td></td>
<td>Louga : 61</td>
<td>Kaolack : 49</td>
<td>Louga : 7</td>
</tr>
<tr>
<td></td>
<td>Kaolack : 22</td>
<td>Louga : 48</td>
<td>Thies : 5</td>
</tr>
<tr>
<td></td>
<td>Fatick : 15</td>
<td>Saint-Louis: 23</td>
<td>Ziguinchor: 3</td>
</tr>
<tr>
<td></td>
<td>(Other)** : 29</td>
<td>(Other) : 31</td>
<td>(Other) : 6</td>
</tr>
<tr>
<td></td>
<td>* reported by the agent as Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>** All other cases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magal of Touba 01/06/2011</td>
<td>other : 30</td>
<td>stomach ache: 7</td>
<td>other : 72</td>
</tr>
<tr>
<td></td>
<td>stomach ache: 25</td>
<td>kidney pain : 5</td>
<td>stomach ache: 50</td>
</tr>
<tr>
<td></td>
<td>aids : 24</td>
<td>other : 4</td>
<td>headaches : 46</td>
</tr>
<tr>
<td></td>
<td>pain : 22</td>
<td>aids : 3</td>
<td>aids : 38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Headaches</td>
<td>Joint Pain</td>
<td>Pain</td>
<td>Information</td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>28</td>
<td>27</td>
</tr>
<tr>
<td>Dizziness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Other)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>District</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Diourbel</td>
<td>341</td>
<td></td>
<td>Diourbel:308</td>
<td></td>
</tr>
<tr>
<td>Dakar</td>
<td>121</td>
<td></td>
<td>Dakar:134</td>
<td></td>
</tr>
<tr>
<td>Thies</td>
<td>83</td>
<td></td>
<td>Thies:83</td>
<td></td>
</tr>
<tr>
<td>Louga</td>
<td>71</td>
<td></td>
<td>Louga:69</td>
<td></td>
</tr>
<tr>
<td>Kaolack</td>
<td>52</td>
<td></td>
<td>Kaolack:54</td>
<td></td>
</tr>
<tr>
<td>Saint-Louis</td>
<td>19</td>
<td></td>
<td>Fatick:29</td>
<td></td>
</tr>
<tr>
<td>(Other)</td>
<td>57</td>
<td></td>
<td>(Other):58</td>
<td></td>
</tr>
</tbody>
</table>

Table 10. Snapshot of the Reasons and Districts of the Calls to the Numéro Vert -/+7 days of the Religious Festivals in 2011

5.2. Calls to the Numéro Vert in 2012

In 2012, we obtained the same findings as in 2011.

Figure 23. Frequency of calls to the Numéro Vert for 2012
Figure 24. Mapping of the calls to the Numéro Vert for 2012 (See http://cdb.io/1zqjX1t)

Frequency of calls per day for the 2012 Gamou
(Gamou of Tivaouane on 02/04/2012)

Figure 25. Frequency of calls to the Numéro Vert -/+7 days of the 2012 Gamou of Tivaouane
Figure 26. Frequency of calls to the Numéro Vert -/+7 days of the 2012 Magal of Touba

Table 11. Numéro Vert Data for the Religious Festivals in 2012

<table>
<thead>
<tr>
<th>Festival</th>
<th>Date</th>
<th>7 Days before Festival</th>
<th>7 Days after Festival</th>
<th>Number of calls -7 Days</th>
<th>Number of calls +7 Days</th>
<th>Number of calls during the festival</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamou of Tivaouane</td>
<td>02/04/2012</td>
<td>01/28/2012</td>
<td>02/11/2012</td>
<td>419</td>
<td>398</td>
<td>58</td>
</tr>
<tr>
<td>Magal of Touba</td>
<td>01/12/2012</td>
<td>01/05/2012</td>
<td>01/19/2012</td>
<td>308</td>
<td>477</td>
<td>53</td>
</tr>
</tbody>
</table>
### Table 12. Snapshot of the Reasons of the Calls to the Numéro Vert in 2012

<table>
<thead>
<tr>
<th>Total phone calls</th>
<th>Top 5 phone calls reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>11,944</td>
<td>Stomach ache 942</td>
</tr>
<tr>
<td></td>
<td>Headaches 857</td>
</tr>
<tr>
<td></td>
<td>HIV/AIDS 785</td>
</tr>
<tr>
<td></td>
<td>Other 698</td>
</tr>
<tr>
<td></td>
<td>Family planning 622</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top reasons &amp; regions of the calls</th>
<th>-7 days before the festival</th>
<th>+7 days after the festival</th>
<th>During the festival</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamou of Tivaouane 02/04/2012</td>
<td>other : 37</td>
<td>stomach ache : 37</td>
<td>headaches : 7</td>
</tr>
<tr>
<td></td>
<td>stomach ache : 32</td>
<td>other : 27</td>
<td>stomach ache : 5</td>
</tr>
<tr>
<td></td>
<td>headaches : 31</td>
<td>aids : 25</td>
<td>bleeding : 2</td>
</tr>
<tr>
<td></td>
<td>aids : 21</td>
<td>headaches : 18</td>
<td>dizziness : 2</td>
</tr>
<tr>
<td></td>
<td>family planning: 16</td>
<td>painful periods: 16</td>
<td>information : 2</td>
</tr>
<tr>
<td></td>
<td>foot pain : 14</td>
<td>information : 14</td>
<td>knee pain : 2</td>
</tr>
<tr>
<td></td>
<td>(Other) : 268</td>
<td>(Other) : 261</td>
<td>(Other) : 38</td>
</tr>
<tr>
<td></td>
<td>Diourbel : 126</td>
<td>Dakar : 72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dakar : 83</td>
<td>Thies : 59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thies : 73</td>
<td>Louga : 53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Louga : 47</td>
<td>Kaolack : 46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kaolack : 36</td>
<td>Fatick : 17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fatick : 30</td>
<td>(Other) : 25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Other) : 24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Diourbel : 29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dakar : 8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thies : 7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Louga : 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kaolack : 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Louga : 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Saint-Louis : 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Other) : 3</td>
<td></td>
</tr>
<tr>
<td>Magal of Touba 01/12/2012</td>
<td>family planning: 25</td>
<td>stomach ache : 46</td>
<td>stomach ache : 6</td>
</tr>
<tr>
<td></td>
<td>stomach ache : 23</td>
<td>other : 32</td>
<td>aids : 4</td>
</tr>
<tr>
<td></td>
<td>headaches : 20</td>
<td>headaches : 31</td>
<td>family planning : 4</td>
</tr>
<tr>
<td></td>
<td>information : 19</td>
<td>information : 28</td>
<td>information : 4</td>
</tr>
<tr>
<td></td>
<td>pain : 11</td>
<td>aids : 25</td>
<td>information on</td>
</tr>
<tr>
<td></td>
<td>aids : 9</td>
<td>family planning : 23</td>
<td>sexuality : 2</td>
</tr>
<tr>
<td></td>
<td>(Other) : 201</td>
<td>(Other) : 292</td>
<td>joint pain : 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Other) : 29</td>
</tr>
</tbody>
</table>
Table 13. Snapshot of the Reasons and Districts of the Calls to the Numéro Vert -/+7 days of the Religious Festivals in 2011

<table>
<thead>
<tr>
<th>Districts</th>
<th>Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diourbel</td>
<td>123</td>
</tr>
<tr>
<td>Dakar</td>
<td>116</td>
</tr>
<tr>
<td>Louga</td>
<td>75</td>
</tr>
<tr>
<td>Thies</td>
<td>61</td>
</tr>
<tr>
<td>Kaolack</td>
<td>39</td>
</tr>
<tr>
<td>Fatick</td>
<td>31</td>
</tr>
<tr>
<td>Kolda</td>
<td>2</td>
</tr>
<tr>
<td>(Other)</td>
<td>32</td>
</tr>
<tr>
<td>Diourbel</td>
<td>18</td>
</tr>
<tr>
<td>Dakar</td>
<td>9</td>
</tr>
<tr>
<td>Thies</td>
<td>8</td>
</tr>
<tr>
<td>Louga</td>
<td>3</td>
</tr>
<tr>
<td>Kaolack</td>
<td>7</td>
</tr>
<tr>
<td>(Other)</td>
<td>4</td>
</tr>
</tbody>
</table>

5.3. Analysis

The toll free hotline calls that we analyzed were for the years of 2011 and 2012. Callers are mainly from Dakar, Pikine, Rufisque, Mbacke, Diourbel and Touba. Less than 1% of the calls are from Tivaouane. We extracted the primary reasons for the toll free hotline calls seven days before, during and seven days after the festivals to determine if there was a pattern in the reasons of the calls. Callers have mainly questions or need information about stomach ache, headache, HIV/AIDS and family planning. We were interesting in seeing if people had specific health issues after the festivals. Even if we can see that there are calls about stomach ache after the festivals we are not sure that this is relevant and need to investigate this conclusion further.

6. Conclusions and Discussions

Mobility and Communications

The literature on analysis of call details records (CDR) shows that we can estimate the flow of traffic between areas from the flow of communications [1,2,3]. In this study we used CDR to show human mobility during two of the key religious festivals in Senegal: Gamou of Tivaouane of the Tijaniyya brotherhood and Magal of Touba of the Mouridiyya brotherhood. The data revealed important patterns of festival mobility through the analysis of calls and SMS. It showed considerable amount of communications occurring during the festivals due to human mobility. The analysis also showed the main routes used by the pilgrims and their travels’ times, and provided insights into the areas most pilgrims are coming from.

Importance of Open Data for Senegal

Advances in technology provide opportunities to collect, store and analyze extraordinary amounts of data. Exploring this data is useful not only to advance science but also to understand a range of human experiences. In this study, we focused on an interdisciplinary project lead by a team of researchers in computer science, public health and geography of health. Planning of festivals is based on estimations, not on data analysis and tangible evidence. Getting accurate data is very difficult. With the ubiquity of mobile phones in Senegal (mobile penetration of 93% in 2013) using CDR can play an important role for planning festivals in a more systematic way.
Structural and Health Implications

The findings on mobility have a wide range of implications and opportunities. Understanding the patterns of mobility, the timeline for pilgrims and technological needs can help for better planning of religious festivals and other gatherings. Government agencies, telecommunication organizations and other stakeholders can essentially plan for the following years by looking at the current data. They can anticipate technological, transportation and health needs, and prepare to respond to these massive movements of population, including during public safety and emergency situations.

The considerable amount of use of mobile phone before, during and after the festivals provide great opportunities for the use of mobile health. Prevention messages can be designed and implemented in areas where pilgrims are coming from, going back to and throughout the festival roads. Advertisements for resources such as the Numéro Vert are already set up in preparation for the festivals. However, such work could go further to include more SMS, voice and even video messages promoting healthy behaviors. Operators like Orange advertise their services in Tivaouane and Touba during the festivals, but are not really involved in awareness campaigns. In addition, our findings could be used in controlling and stopping the spread of infectious diseases by providing information about health structures available during the festivals and sending SMS.

Sensitive Data Concerns

A good understanding of the local context is important when reaching to conclusions. We realize that our study addresses sensitive cultural practices in Senegal and that religion is a delicate topic. Our study has nothing that is subject to polemic. It focuses on showing the mobility of pilgrims based on brotherhood belongings, and implications ranging from resource management to service allocation and awareness campaigns.

Recommendations

Given these implications and opportunities afforded by this data analysis, we offer the following recommendations for situations of massive movements of population due to the Gamou of Tivaouane and Magal of Touba festivals. We understand that some of these efforts are already in process but believe it is important to cover the wide range of uses of Orange CDR data.

1. Integration of data from different sources to fully understand the scope and opportunities of Orange CDR data
2. Open data to all stakeholders so that informed planning can be fully coordinated
3. Setting up an emergency plan incorporating mobile components (e.g., SMS and voice messages on health and transportation)
4. Coordination of pilgrims’ transportation plan using data on travel locations and times
5. Continuous review of data after large religious festivals to identify lessons learned and plan for following years
Limitations

We faced difficulty in obtaining the proper data from local agencies. The team in Senegal had access to the Numéro Vert data of 2010, 2011 and 2012 before the beginning of this study. We sent emails with questions and requests for material to agencies listed on the web site of Orange Data for Development Challenge (D4D 2014). None of the agencies answered emails. Local contacts preferred to interact with the Senegalese partners. Data on sites such as ANSD (Senegalese National Agency of Statistics and Demographics) are available in an aggregate form. With more data and more communications with local agencies, we could have gone further in our analysis.

The Orange data we used for the Magal of Touba were collected from 6 sites (out of the 47 possible sites for Touba). We focused on the use of CartoDB through a paid account and a CartoDB grant that had only 1 GB of storage. With more storage we could have explored more sites and have a better vision of the scope of the Magal of Touba for a more comprehensive comparison with the Gamou of Tivaouane.

The health data from the SNEIPS for 2011 and 2012 did not permit us to distinguish patterns of phone calls during the religious festivals. We did not get the data for 2013 due to the focus of the Ministry of Health on keeping the Ebola epidemic under control in Senegal. We noticed that the data had different issues related to the way calls are recorded by agents. Numerous entries contain orthographic errors and did not mention the reasons of the call (e.g., the reason is reported as other in that case). More categorization of the reasons is also required for analysis of the data at a higher level. Storage of the entries seems another issues as no standard is in place to save the data for further analysis.

Future Work

If granted further access to the data, we would be interested to map the complete data of the Magal of Touba by using the 47 Orange sites of Touba. We would like also to analyze the data of the Magal of Touba of January 2013. To further explore human mobility during festivals and to draw stronger conclusions, we would be interested using the SET2 and SET3 datasets of the Orange Data for Development Challenge.

We are planning a complete analysis of the Numéro Vert data by categorizing the data and using data mining techniques. A significant amount of data recorded and collected from the Numéro Vert concerns family planning and HIV/AIDS. We will compare our conclusions with the ones of the Infoado SMS service of the One World NGO that was used to raise awareness on sexuality and reproduction amongst adolescents in Senegal [13].

Acknowledgment

We thank the Orange operator and the Data for Development (D4D) committee for the access to the mobile phone dataset of Orange in Senegal for 2013. We are grateful to the Service National de l’Education et l’Information pour la Santé (SNEPIS) for access to the Numéro Vert Health
Data for 2010, 2011 and 2012. This work was supported by a 2014-2015 CartoDB grant, the Seidenberg School of Computer Science at Pace University, and the 2013-2015 Pace University Undergraduate Research Initiative.

References


Data for Development Reloaded: Visual Matrix Techniques for the Exploration and Analysis of Massive Mobile Phone Data

Stef van den Elzen, Martijn van Dortmont, Jorik Blaas, Danny Holten, Willem van Hage, Jan-Kees Buenen, Jarke J. van Wijk, Robert Spousta, Simone Sala, Steve Chan, Alison Kuzmickas

Abstract—We present visual analytics techniques for the exploration and analysis of massive mobile phone data. We use a multiple coordinated view approach with a scalable and flexible visual matrix as a central element to our solution. In addition we provide different views to explore both space, time and structure in one unified framework. We discuss various visualization and interaction techniques that enable users to identify both temporal and structural patterns, including normal behavior, outliers, anomalies, periodicity, trends and counter-trends. The visualization and interaction techniques are applied to the data for development challenge data containing phone calls and text messages between more than 9 million people in Senegal collected over the course of an entire year. From this data we extract and discuss global events, weekly recurring events, regional patterns and outlier events. The insights gained by identifying and analyzing the patterns can be used for better policy decision making.

Index Terms—Mobile Phone Data, Visual Analytics.

1 INTRODUCTION

Big data can be leveraged to advance the growth and socio-economic state of developing countries. It is, however, unclear how to achieve this. Therefore, the Orange group organized the first Data for Development (D4D) challenge in 2012 focusing on Mobile Phone Data in the form of anonymized Call Detail Records (CDR) collected over a period of 5 months in Ivory Coast [12]. This led to 260 creative and innovative applications and 80 research papers [17]. After this success, Orange in collaboration with Sonatel Senegal launched a second D4D challenge. Three datasets were made available based on the collection of CDRs of mobile phone calls and text messages between more than 9 million people in Senegal over the course of the year 2013 [17]. Senegal consists of 14 regions that are divided into 45 departments, which are again divided into a total of 123 arrondissements. These arrondissements contain a total of 1666 towers. In this paper we focus on tower-to-tower communication. Data collection and initial preprocessing was performed by Orange Labs and Sonatel. Preprocessing consisted of caller anonymization and aggregating CDRs on an hourly basis. For each hour we are given the total number of calls, total number of text messages and durations of the calls, between each pair of towers. Furthermore, only people having more than 75% days of the year with interactions were included. Persons with more than 1000 interactions per week were presumed to be machines or shared mobile phones and are excluded from the data. Additionally, exact locations of towers were not given, but locations were slightly jittered for commercial and privacy reasons.
The visual matrix proved an ideal candidate and we significantly enhanced our visualization tool of the previous challenge with the visual matrix as an anchor point. The first color-shaded matrix display is presented by Loua [31] and dates back to 1873. Reordering of the visual matrix to reveal structure and patterns is discussed by Brinton [14] and Bertin [10]. More detailed related work on ordering the visual matrix is described in Section 3.4. A historical overview of the history of the visual matrix is provided by Wilkinson and Friendly [49].

3 Visual Analytics Approach

The main goal is to help with the growth and improvement of the socio-economic state of Senegal. It is believed that effective use of big data can help to achieve this, therefore, mobile phone communication data was collected and released to the research community. A first step is to gain insight and understand the massive amounts of mobile phone traffic. The insights, provided in the form of patterns, are next transformed into knowledge by domain experts. This knowledge can ultimately be used by policy makers to make better decisions. Understanding of the data can be achieved by revealing hidden patterns in the data, in network structure, time, and space, such as:

- **normal behavior**: the neutral or average state (number of calls, durations and number of text messages) for specific towers or tower-tower combinations;
- **outliers**: sudden singular large deviations from the normal behavior;
- **anomalies**: missing or incorrectly collected data;
- **periodicity**: periodic repetition of number of calls, duration, number of text messages or a combination of both;
- **trends**: increase or decrease in number of calls, duration, number of text messages over time between two or more towers;
- **counter-trend**: deviating trend pattern by showing mirrored behavior.

More complex patterns are possible by (non)linear combinations of the above defined basic building blocks. Therefore we cannot rely on purely automatic methods, such as statistics, data mining or machine learning, to detect these (combined) patterns, since we do not know what patterns are present in the data. Important patterns may therefore be missed and also context for verification or further analysis is neglected. To overcome these problems we utilize a human-in-the-loop visual analytics approach. We provide a combination of visualization and automatic methods that enables users to iteratively explore, analyze and refine the data to find complex patterns.

We implemented exploration and analysis methods in a prototype developed using Qt/C++ that runs on Windows, Linux and Mac operating systems. Initial data preprocessing was done by Orange and...
Sonatel. We further processed the data by computing relevant aggregates, such as the daily traffic and the traffic aggregated to the arrondissement, department or region level, using SAP HANA (High-Performance Analytic Appliance). SAP HANA is an in-memory, column-oriented, relational database management system designed to handle both high transaction rates and complex query processing [36].

To support the exploration and analysis of the massive mobile phone data, it is important to have different perspectives on the data; we are not only interested in how many calls were made between two towers, but also when they happened and where the towers are located. We need to be able to simultaneously inspect all three aspects, structure, time, and space, in order to reveal complex correlations. Such a holistic view is achieved by providing a multiple coordinated view setup with specialized visualizations for the different aspects, coupled by linking and brushing techniques [15, 25]; if items in one view are highlighted or selected, the associated items in all different views are highlighted as well.

Following the information-seeking mantra [42]: overview first, zoom and filter, then details-on-demand, we start with an overview of the number of calls over time per cell tower. As a central element in our solution, from which the exploration workflow is started, we use a visual matrix.

### 3.1 Matrix visualization

The matrix visualization consists of colored square cells [31]. The number of cells is determined by the attributes projected on the horizontal and vertical axes. On each intersection of horizontal and vertical attributes a cell is drawn. This cell represents a value and is colored according to a user selected colormap. Initially, a black-body colormap is used which is perceptually best if no information on the data is assumed [13].

Here we choose a visual matrix as a starting point for the exploration process because it has a number of advantages:

- **flexible**: Different attributes can be used for the axes and cells to support a wide range of analyses. Projecting towers vertically and time horizontally and number of calls in the cells shows tower call-intensity over time. Projecting towers both horizontally and vertically with deviation of number of calls compared to average in the cells allows for outlier analysis on the communication channel level. Also, both the axis and cells allow for filtering and (hierarchical) aggregation for a high-level overview as well as a more detailed low-level analysis.

- **scalable**: To provide the highest information density, each individual screen pixel can represent a cell value in the matrix. Combined with hierarchical aggregation on both axis and cell level as well as computer-graphics interpolation techniques a high level of scalability is achieved which is desirable when dealing with massive amounts of data.

In addition, matrices are more readable with respect to dense graphs and medium to large graphs for most tasks [21]. Here we are dealing with highly dense graphs.

The matrix view can also show a linked histogram for both rows and columns and an icicle plot [8] of the hierarchy on the cell towers (region, department, arrondissement, tower) when the matrix is sorted based on the cell tower hierarchy. The visual matrix itself can visualize the absolute values of a selected attribute as well as delta values relative to either a specific day, the average day of the year or the average day of the week for each cell. This enables users to detect different patterns as described in Section 3.

Figure 3 shows the visual matrix with all different configurable elements. Interaction enables the exploration of data contained in the visual matrix. We elaborate on the different interaction techniques in the next section. Next to the visual matrix we provide different linked visualizations such as geographical views and line charts, which are explained in Section 3.3.

![Visual matrix with Highlighted Items](image)

**Figure 3.** Schematic representation of visual matrix with all different configurable elements; icicle plots at right and bottom for hierarchical grouping of the towers based on geography, histograms at top and left for aggregated information about the values in the cells, and cells are colored with a user selected colormap according to the value they contain. Dark cells represent little communication and white cells represent a high number of calls. Users are enabled to select different measures for the cell coloring, e.g., number of calls, duration, difference compared to average, ratio text versus voice, etc.

### 3.2 Interaction Techniques

All visualizations are coupled with each other through standard linking and brushing techniques [15, 25]; if items in one view are highlighted or selected, the associated items in the other views are also highlighted or selected. Each view has a number of general interaction elements that are uniformly implemented. In addition, there are interaction techniques that are specific to a view. These are described in Section 3.3 which discusses the linked views. The highlighting and selection of items is implemented in all views; cells and rows/columns in the matrix view, points on the line in the line chart, and areas (regions, departments, arrondissements, towers) in the geographical and icicle plot views. Furthermore, for both the matrix view and the geographical view, zooming and panning techniques are implemented; users are able to freely zoom and pan the visualization. This enables both detailed inspection of elements as well as a high-level overview for the detection of global patterns. In addition, the linking and brushing enables analysis of different perspectives on the data for correlation detection, while simultaneously providing a context. Extra information about the selected or highlighted items is given via tooltips and the status bar (details on demand).

Once an interesting event has been detected, the tool can automatically open a browser window or use an existing instance, to launch a search engine, such as Google, with the event data as a search query: a combination of date, time, region, department, and arrondissement. This allows for quick confirmation of the detected event by cross-checking publicly available (news) sources for information on the selected locations and times.

### 3.3 Coupled Visualizations

While matrices are a powerful tool in their own right, it is difficult to detect geographical and temporal patterns in a matrix alone. To over-
come these shortcomings we have integrated a coupled geographical view (see Figure 4) as well as a time series line chart (see Figure 5). Again, highlighting or selecting in any of the views will update all other views through standard linking and brushing techniques. This places interactions in the matrix in a geographical context via the map and provides insight in the temporal behavior of selected items by displaying the associated time series information in a line chart, while interactions on the geographical view can be placed in structural context in the matrix, as can be seen in Figure 11.

The matrix view can be set to display all tower-to-tower communication for the entire year or for a specific time period by displaying an adjacency matrix with attributes, such as number of calls or text messages, mapped to a color. In this mode, histograms of the activity on each row and column will be displayed to the left and top of the matrix, respectively. The geographical view shows a map of Senegal, with the tower positions overlayed as colored dots. When no nodes are selected, the dots are colored according to their global activity level using the selected color map. If a specific tower is highlighted or selected, the total activity from that tower to all other towers is mapped to the tower color. When a tower is highlighted or selected in the map or in the matrix view, the temporal view displays the activity of the selected tower for the entire year.

For both the matrix view and the map view, the color mapping can be set to either map to the absolute attribute values (see Figure 6(a)) or map to a delta value versus either a specific day, the yearly average for that tower-tower pair (see Figure 6(c)) or the average week day, matching activity for a specific tower-tower pair on a Monday, for example, to the average activity on all Mondays throughout the year for that tower-tower pair (see Figure 6(d)).

### 3.4 Normalization and Clustering

Initially the rows in the matrix (representing towers) are ordered by tower identifier, see Figure 7(a). To see more structure in the data we sort the rows and columns according to the given hierarchical order of the towers by geographical location: region, department, arrondissement, and tower. However, towers with similar call behavior over time may be scattered across the matrix. Naturally towers with similar call behavior are of interest and we would like to group them. To achieve this, several techniques for matrix reordering [30] can be used.

We implemented different methods for matrix reordering to reveal similarity, structure and outliers. First, users are enabled to select a distance measure between two towers, or pair of towers (communication channels). We implemented Euclidean, Manhattan, Pearson [34], Spearman [43], and, Kendall [26]. After distance computation, we initially apply hierarchical clustering [24]. Users are free to choose a different clustering method like k-means [33] or k-medians [23].

After hierarchical clustering, we use the resulting dendrogram to sort the items in the matrix. Sorting is performed in a depth-first-search [16] order. This results in similar items, either towers, or communication channels, to be placed close to each other, see Figure 7(b). Besides the clustering algorithm, and the distance measure, users can define the number of clusters and which time-span to take into account. After a parameter change, the new sorting order is directly reflected in the matrix view.

By applying a colormap to the matrix view, each cell is colored according to its representative value. If a few values in the matrix are high compared to the other values in the matrix little variation in color is shown, because the majority of values is mapped to a small color range. Therefore, before the computation of the distances between items in the matrix view, users are enabled to normalize the data. One way to achieve this is to clamp the data to a certain range or use a log scale, rather than a standard linear scale. Both are offered as an option to the user, however, a more appropriate way to solve this problem is normalization. Normalization is performed based on the entire matrix, or normalization is applied separately to each row or column of the matrix. This enables users to see global patterns, local patterns, similar call behavior independent of scale, and emphasizes peaks in the matrix, see Figures 7(c) and 7(d).
Fig. 6. Zoomed section of matrix of Touba and Dakar using various display modes. (a) total calls for 2013. (b) December 22nd absolute calls. (c) December 22nd delta with yearly average per link. (d) December 22nd delta with average day of week per link.

4 RESULTS

Below we first describe the workflow we utilized to discover and interpret patterns in the Senegal D4D challenge tower-to-tower communication data. Next, for a number of patterns, a general description is given followed by one or more instances found in the data.

4.1 Workflow

A typical exploration is started with the visual matrix showing call intensity over time; we refer to this as the temporal matrix. On the vertical axis each tower is represented and the horizontal axis represents time. Next, a black body colormap is applied to each cell of the matrix. Then clamping and normalization are applied followed by hierarchical clustering to group towers with similar call behavior. Different normalization methods as discussed in Section 3.4 provide multiple perspectives that each highlight certain patterns. Next, interesting patterns are inspected in more detail using the linked views to place them into a structural and geographical context. If the pattern is difficult to interpret, the columns and rows of the matrix are adjusted to, e.g., tower versus tower, for a specific date. Furthermore, nearby towers are inspected for similar behavior. If, after zooming in on the details, the pattern is still unexplainable a search string is automatically constructed that includes the time and geographic (hierarchical) position of the tower (or tower-tower combination). This search string is translated to French and an Internet search engine (like Google) is opened in a browser showing the search results. We then manually check the search results for a source that clearly explains the pattern. Furthermore, we look for correlation with external data using the International Disaster Database [1] and historical weather data [3].

With this workflow we were able to find many interesting patterns in the data that are clearly related to events. However, we also found patterns that we were unable to explain (or correlate with external data), for these patterns, we provide hypotheses.

4.2 Global Patterns

Several global trends become clear when observing the data. A high portion of activity for most towers is very localized, to the tower itself and its nearest neighbors. Dakar and its surrounding region have a central role in Senegalese society, as the region is responsible for much of the traffic, though this should come as no surprise as nearly 30% of the towers are located in this region (see Figure 8) and a large portion of the population lives here.

Fig. 7. Matrix visualization with time horizontal (one year, aggregated by days) and towers vertically ordered by (a) tower identifier (b) hierarchical clustering with clamping, and additional normalization by (c) row, and (d) column. By clustering and normalization several patterns a clearly visible in the data. These patterns are discussed in more detail in Section 4.

Fig. 8. Tower density in the Dakar region is particularly high; 489 cell towers of the total 1666 are located here.
Fig. 9. Temporal matrix (right) with annotated events (left). A more detailed description is given in Sections 4.2, 4.3, 4.4 and 4.5.
One of the strongest globally occurring patterns is a peak on December 31st. This is most likely related to New Years Eve. One notable exception is Touba where many towers show a significant decrease in activity compared to the local average. Perhaps the most salient pattern is the clearly visible dip in communications on the 29th of March, see Figure 9(a). We could not find a clear reason for this dip. A possible explanation is a global disruption in the electricity network or bad weather with thunderstorms all over the country hindering cell tower communication, however, we were not able to verify this using historical weather data [3]. It might be related to Easter, which is celebrated the subsequent weekend, however, this is unlikely in the largely Islamic Senegal.

Finally, we see that several towers only show activity towards the end of the year. These towers were most likely only activated during the year. Furthermore, some towers seem to be deactivated during the year and others are only active for a short period of about a month, see Figures 9(b) and (c). Also, a small number of towers show no activity whatsoever during the year.

4.3 Weekly Recurring Patterns

While many towers show a weekly recurring pattern, which shows up clearly in the temporal matrix due to clearly periodical color contrast, it is mostly the towers located in urban areas that show this pattern stronger. Many towers seem to adhere to a weekly cycle where weekdays show higher activity compared to the weekends, as shown in Figure 9(d). The fact that these patterns show up more clearly in urban areas can most likely be attributed to the nature of labor in cities versus labor in rural areas.

Certain towers show a different pattern where one day in particular shows higher activity compared to the rest of week, as can be seen in Figure 9(e) where different towers peak at Monday, Tuesday, Thursday, Friday or Saturday. While we could not find a clear explanation, the recurring nature of the pattern might indicate a market day or a recurring religious or sports event. Insight from these surges in localized call activity can help to inform the planning of road construction or other infrastructure maintenance, so as to avoid such work during times when an area tends to be heavily crowded.

4.4 Events

Detection of outliers related to events is performed with the temporal matrix. Outliers are displayed as clearly contrasting brighter or darker cells in the matrix, see Figure 9(f). Once a potential outlier has been detected, users can zoom in on the towers involved with the event by opening the associated timeline to get a better impression of the overall activity of the according tower. For a date of interest in the tower-to-tower matrix, towers with unusual activity show up as differently colored rows and columns compared to the background. This effect is particularly visible when visualizing the difference of the activity compared to the average day of the week activity, shown in Figure 11.

There are two public holiday events [2] that clearly stand out in the visual matrix; the end of Ramadan (Thursday, August 8, 2013 to Saturday, August 10, 2013), and the Feast of Sacrifice (Tuesday, October 15, 2013 to Thursday, October 17, 2013), see Figure 9(g) and (h). For these holidays there is a global increase of call intensity. Due to the strong pattern, we can also spot deviations due to disruption of the global pattern. Several towers in Dakar show a peak two days before and a strong decrease during the Feast of Sacrifice. This could be an increase to finish work before the holidays, see Figure 9(i). We also see a cluster of towers concentrated around Touba with increased call intensity the entire week after the Feast of Sacrifice (Figure 9(j)). Most likely, the festivities there are longer compared to other areas. Another pattern that deviates from the global increase is clearly visible. Here, there is no increase on both holidays, however, there is a strong increase in call intensity a week after the end of Ramadan on both August 14 and 15. On August 15, it is a feast day of the Assumption of Mary, one of the Catholic holy days, which is also a public holiday in Senegal [48]. Most likely, these towers are positioned in areas where Christian belief is practiced. In the linked geographic view we observe that the locations of the involved towers are all in the Casamance region, where many of Senegal’s Christians reside [32].

Although not as visible as the end of Ramadan and the Feast of Sacrifice, we do see a global increase in the number of calls on November 13, also a public holiday due to Tamkharit, the celebration of the Islamic new year [2], see Figure 9(k). Also, the Prophets Birthday, another public holiday on January 24 is clearly visible in the data due to an overall decrease of the number of calls, especially in commercial areas (e.g., Dakar) by a significant increase in other less commercial urban areas.

Another event related to religion that clearly stands out due to an increase in the number of calls is the Grand Magal [39], a major annual pilgrimage to Touba from December 20th to December 23rd, see Figure 9(l). Also, the yearly pilgrimage to Popenguine [41] is a clear outlier visible on the timeline. Held on May 19th, it shows a tenfold increase in traffic on the 19th and 20th, see Figure 9(m). There are several other Magal (Wolof word for celebration) that correlate to spikes in the number of calls. For example, on August 22 (Figure 9(n)), the Magal of Touba Réfane [44], a small town located in the department of Bambey and the Magal de Porokhane, on March 14 (Figure 9(o)), in the village Porokhane, near Niourou d’Rip in the Kaolack region [40]. Using normalization by column reveals an outlier on March 10, in the Tivaouane region, where La Ziarra Generale is celebrated [29].

On April 10, there is a significant increase in call intensity in the Matam region. This correlates with the threat of flooding of the Sénégal River due to heavy rainfall [37].

During April 11 to April 25 we see an overall increase in the number of calls with two clear spikes in the Kolda region, see Figure 9(p). We found that this correlates with an investigation of the Global Water Initiative on the living conditions of farmers. In the plan of action report for this operation [22] it is stated that the investigation will consist of two parts, the first from April 11 till 15, followed by a second phase from April 21 till 25; this probably explains the two spikes in the overall traffic.

On June 8 and 9, an increased number of calls is clearly visible at a tower in Dakar near the Parcelles Assainees region. This correlates with a festival on culture, artistic expression and sports for children that is held there [18], see Figure 9(q). Also, from April 14 until 22 we see a general increase in the number of calls in the Dakar region, this pattern stands out because it breaks the periodic week-weekend pattern. The increase in this week correlates with the death of the old and reassignment of the new religious leader, the Grand Serigne of Dakar [38], see Figure 10.

Fig. 10. Death of the old and reassignment of the new religious leader the Grand Serigne de Dakar. (top) pattern stands out due to breaking the week-weekend pattern; (bottom) more detailed line graph showing call intensity for the selected cell tower for the entire year. (right) linked geographical visualization that shows the location of the tower providing context.
Fig. 11. Combination of linked views showing heightened activity in Touba around the 22nd of December compared to the average day of the week using a red (below average) to blue (above average) colormap. (bottom left) tower to tower matrix visualization; (bottom right) geographical visualization showing the location of the selected tower and the according arrondissement. (top) line chart showing call intensity for the selected tower separated by incoming calls (green) and outgoing calls (red) or overlapping (orange). A clear spike with above average number of calls is visible for December 22 with also heightened activity around this day. Other spikes are New Years Day, End of Ramadan, and Feast of Sacrifice.

4.5 Regional Patterns

Starting in October until the end of the year we note an increase in tower activity across several towers, mainly located in the southern parts of the country. Given that this correlates to the harvesting season [20] for many of the countries agricultural products, such as rice, cotton and peanuts, it seems likely that the noted trend is related to activities surrounding the harvesting of these products.

For the towers that were not active at the beginning of the year, as mentioned in Section 4.2, we see that activation of these towers is usually correlated to a drop-off in activity for a nearby tower of similar magnitude to the activity of the newly activated tower.

Weather also plays a role in call activity. Several drops in activity around towers can be correlated to known thunderstorm activity [3] in the region at the time of the drop-off in activity.

The temporal matrix also highlights unusual activity around towers that normally show little or no activity during the year. Several towers show unusual spikes, over one or two days, many orders of magnitude larger than the normal activity on these towers. In extreme cases towers that have no activity throughout the year suddenly spike up to around 1M calls. For many of these spikes we could not find a satisfactory reason in the data available to us, so it is unclear whether these spikes are merely anomalies in the data or whether they are correlated to some unusual event.

One interesting observation we have made is that the ratio between voice and text messages seems to differ for communication to rural and urban areas. Communication to and from rural areas seems to be dominated by voice traffic, while traffic in urban areas seems more balanced, see Figure 12. We suspect this might be related to differences in literacy levels between urban and rural areas. With a literacy rate of around 50% [4], it is one of the challenges facing Senegal. This difference in the use of text messages between urban and rural areas could help to inform government and other public assistance programs which utilize text messages as a way to disseminate information. The efficiency of such short message service (SMS) or text campaigns could be maximized by focusing on areas which most routinely communicate via SMS, or crafting text messages to specifically address lower levels of literacy (i.e. through pictographs or other means).

5 Conclusions

We developed a highly interactive prototype system for the exploration of massive mobile phone data in the context of the D4D challenge. As a central element and starting point of the exploration process we implemented a scalable and flexible visual matrix. We define a number of patterns that are of interest to understand the data both on a global and low-level detailed scale. The patterns of interest are normal behavior, outliers, anomalies, periodicity, trends, and, counter-trends. For the visual matrix we provide and discuss techniques for the discovery of these patterns. The most important features are flexibility of attribute projection on both axes, color-mapping, hierarchical aggregation, summarizing histograms, interaction, coupling with other visualization to provide context and coupling with external applications (browser) to retrieve content from the Internet.

These techniques are then applied to the D4D data of Senegal and we were able to identify and explain a number of patterns clearly present in the data. We found an increase in number of calls correlating with local religious events such as the Pilgrimage to Touba and Popenguine. Also, an increased number of calls correlating with global religious events such as the end of Ramadan and the Feast of Sacrifice are clearly visible. Using the visual matrix we could effortlessly identify towers that were activated or deactivated throughout the
Fig. 12. Differences in urban and rural areas for voice versus text ratios. Communication from and to a selected typical (a) urban and (b) rural tower is shown. A black (balanced voice and text) to red, yellow, white (predominantly voice traffic) colormap is used.

5.1 Future work

We unfortunately did not have access to rich external data sources or local domain experts. For future work it would be interesting to enable cross-correlation by also loading external socio-economic data. Also, we believe that by looking at data at an even finer scale, for example, individual cell phone calls, or content of SMS messages, more patterns are revealed, providing a higher level of insight useful for decision making policy. However, we do understand this poses difficulties due to privacy and commercial concerns.

Besides tracking Islam versus Christian via holiday events as described in Section 4.4, we could also track controversy/trends as pertains to the African Renaissance Monument, if we are provided with the content of (anonymized) SMS messages. In previous work for U.S. Africa Command (AFRICOM) [5] we found interesting trends in April and December for Dakar (the location of the monument), and the religious center of Senegal Touba. Senegal’s African Renaissance Monument typically generates a heightened level of communication/controversy during the months April and December. The monument is controversial for 3 principal reasons: 1) cost (given the backdrop of Senegal’s economic crisis); 2) religious controversy regarding the “un-Islamic representation” as well as several statements made by the principal advocate, President Abdoulaye Wade; and 3) the connection/motivation behind utilizing North Korean architects/designers. Construction had begun in April (and unveiled in April) and its anticipated completion date was December, hence the significance of those two months.

Acknowledgments

We would like to thank Niels Willems, Thomas Ploeger, and Bart van Arnhem for support on hardware and software (SAP HANA appliance) and valuable advice and comments during discussions.


ABSTRACT

Load balancing in mobile cellular networks is an important mechanism that enables distribution of demand across neighboring cells, which is critical for better resource utilization and user satisfaction. Current approaches for load balancing are reactive, redistributing users only when the offered load approaches the cell capacity. This approach can lead to deteriorated network performance and user experience. In order to better cater for their users, mobile networks need to be proactive and provision resources based on expected demand. To this end we propose a load balancing mechanism that allows for proactive network configuration based on prediction of traffic load. Our approach makes use of power control mechanisms to reconfigure the coverage of a mobile base station and thus control the amount of users and offered load at that base station. We apply our method on a real-world cellular network in Senegal and show that it enables better distribution of load in Orange Telecom’s network in Senegal.

1. INTRODUCTION

Recent projections on Internet traffic demand predict that the traffic generated in 2018 alone will be as much as the total traffic from the period of 1984-2013 [9]. Majority of this traffic will be generated by mobile devices and will be handled by various wireless networks, including mobile cellular networks. This increase will lead to a wireless capacity crunch, which will require novel techniques to accommodate the ever-growing demand.

The demand in mobile cellular networks is often times unevenly distributed throughout the network causing some cells to be overloaded, while others - underutilized. Traditionally, this problem has been addressed by the use of load balancing techniques that aim to redistribute the user load among neighbouring cells in order to unify their utilization [6]. Current methods for load balancing utilize handset handover from more to less loaded cells. Along with the handover, association parameters are set in both the source and the target cell to ensure that the handset does not return to the congested cell. A major drawback of this technique is that it is reactive: users are redistributed only when the cell reaches capacity. This can be detrimental to user experience of already associated users since no capacity is left to accommodate the dynamically-changing demand of those users.

To address this problem we propose a novel technique for proactive user redistribution that achieves load balancing through Adaptive Power Assignment (APA). In cellular networks, users’ handsets associate with the cell with highest perceived signal quality, thus by manipulating the emitted power at each base station, one can mandate the user association across cells.

In this preliminary study we first analyze the existence of load imbalance in neighboring cells in the cellular data from Senegal made available as part of the Data for Development (d4d) challenge. We demonstrate that in the densely populated area of Dakar (Figure 1) there are regions of local load imbalance, e.g. cells that are loaded significantly higher/lower than their neighbors in terms of number and duration of calls. We model the network coverage of individual cellular towers as a power diagram [8], a generalization of a Voronoi diagram in which cells are defined in terms of a generalized notion of distance that involves a power parameter that can model the relative power of emission. Within this model we can manipulate the area in which a given tower has maximum power.

Our analysis is in two parts. First, we analyze the load imbalance in Senegal at different times of the day with a closer look at the capital area of Dakar. We demonstrate that there are a number of instances in which a cell is sig-
nificantly more (or less) loaded than its neighbors (load dis-
parity). Next, we introduce and study a local heuristic that
manipulates the relative power of cells in order to improve
the spatial load balance, and thus improve the quality of
service for individual users. Our approach can be easily dis-
tributed (i.e. no need of global synchronization) and can
be implemented in both commercial cellular networks and
remote-area/disaster-relief networks using alternative tech-
nology such as cognitive radios. We provide an extensive
discussion of the implications of our load-balancing design
for alternative technology remote area network design as well
as disaster relief scenarios.

Our contributions in this study are as follows:
1. Using real-world data from Senegal, we demonstrate
the existence of significant disparity of load of towers
in a cellular network over time.
2. We introduce a novel approach based on power dia-
grams for proactive re-distribution of users, designed
to minimize cell tower overload and maximize utiliza-
tion.
3. We perform extensive evaluation of our spatial load
balancing approach and demonstrate that it has the
potential of improving the operation of the existing
network in Senegal as a concrete example.
4. We provide an extensive discussion of the implication
of our approach to both commercial cellular networks,
but also ones deployed in remote areas and for the
purposes of disaster relief.

2. RELATED WORK
Traditionally, load balancing in cellular networks has been
performed via manipulation of cell parameters such as the
cell pair specific offsets [1] that trigger users to associate
with one cell or another [5]. The downside of this approach
is that it is based on the currently-offered network load and
requires knowledge of user identity in order to perform load
balancing. The latter does not allow for development of pre-
dictive methods that balance the network proactively rather
than reactively.

In this work we consider a methodology that relies on
manipulation of the transmit power at a base station. This
method is user-oblivious and thus allows for predictive load
balancing. Several previous works have considered manipu-
lation of base station emitted power for improved network
performance [7, 3]. Niu et al. [7] propose cell zooming in
order to minimize the energy consumption in a cellular net-
work. Chen and Baccelli [3] develop a combined user as-
sociation and power control technique that considers SINR
and interference in order to minimize the delay to send in-
formation to associated users. None of these works directly
considers power manipulation for load balancing. Further-
more, these previously proposed methods perform evalua-
tion in simulated scenarios. In contrast, our work directly
tackles the problem of load balancing through manipulation
of emitted power. Furthermore, through analysis of real-
world traffic, we demonstrate the feasibility of the problem
as well as that the proposed technique successfully balances
the offered load in a cellular network.

3. IMBALANCE IN A REAL-WORLD NET-
WORK
We first study the operation of the Senegal Orange net-
work in order to understand the extent to which imbalance
in the load of operating cellular towers exist. We perform
both whole-network analysis as well as focused analysis of
the capital region of Dakar.

3.1 Data Preparation
We employ the data on antenna-to-antenna traffic (Set
1) for January, 2013 [4]. Our analysis is both at the coun-
try level (including all towers), as well as focused on the
densely populated urban area of Dakar. In order to select
towers for the latter, we filter cells based on their position (as
provided in SITE_ARR_LATLON.csv) including only those
within the Dakar region. While the provided coordinates are
perturbed for privacy and commercial purposes [4], the new
locations are within the Voronoi cells of the original antenna
locations, so we do not expect the findings of our analysis
to be affected significantly by this data anonymization step.
The total number of sites in the whole country is 1666, while
the number of cells in our analysis of the Dakar region is 489.
We define the load of a cell tower as either the number
of calls within an hour or the total call duration (measured
in seconds\(^1\)). This information is reported in the antenna
to antenna traffic data on an hourly basis. While load bal-
ancing can essentially be done by re-assigning users, we use

\(^1\)The document [4] accompanying the data does not explic-
tly specify the measurement unit, however, based on the
numbers reported seconds is the most likely unit.
number of calls and call duration as proxy variables for the number of associated users. Thus, our analysis assumes uniform user activity and uniform positions of users within the cell, since those are not available at a fine-grained level. While users’ activity and position may not be uniform in a real world scenario, our proposed approach can easily be adapted to handle skewed distributions of user activity and non-uniform user positions within cells.

3.2 Cell Tower Load in Time

Our first goal is to assess the extent to which cells in a real world system are imbalanced in the amount of load they observe and the corresponding trends at different times of the day. We divide the day into 4 periods: Morning (5am-11am), Midday (11am-5pm), Evening (5pm-12pm) and Night (12am-5am). For each of those periods, we compute the average hourly number of calls in a given cell, as well as duration of calls. We use data from the month of January, 2013.

Figure 2 shows the distributions of average hourly number of calls and duration for different times of the day. The cell towers (sites) from the whole country are included in the first two Figures 2(a) and 2(b), while only cells in the Dakar region are used for Figures 2(c) and 2(d). Within the whole country, there are close to 200 locations that have no activity (Fig. 2(a)). On the other side of the spectrum there are around 500 sites that observe more than 10 thousand calls per hour in the midday and evening hours. Many of those high-load cells are in the Dakar region (Fig. 2(c)). The vast majority of the sites observe between one and ten thousand calls per day, with less loaded sites observed in the morning and night hours.

Similar trends can be observed in the distribution of average call duration per hour (Fig. 2(b) and 2(d)). Based on this metric, the most loaded cells often observe more than 3 hours of talk time on average, meaning that there are at least that many individuals talking simultaneously while associated with the cells. While the latter may be a conservative estimate, the important conclusion is that there is significant range of loads that individual cell towers observe and this disparity is evident in the whole country Fig. 2(b), as well as in the region of Dakar Fig. 2(d).

Beyond the distributions of cell tower loads, we are interested in understanding how those towers are spatially arranged. We aim to eventually answer the question: Are there cells of different load located next to each other? An affirmative answer to this question would mean that there is space for load balancing based on power manipulation. Ideally, we will observe cells of different loads situated next to each other in which case we can re-distribute associated users by decreasing the power of overloaded cells and increasing those of the neighboring idle cells.

Figure 3 shows heatmap-like diagrams depicting the load of cells in the region of Dakar at different times of the day. Tower locations are depicted as dots and the borders of the cells in which a tower’s power is maximum are denoted by black segments. Here we assume that all towers emit with the same power resulting in a Voronoi diagram of the cells. Furthermore, cells are colored according to their load, where darker red denotes a more loaded cell (in terms of number of calls per hour) while pale color denotes less loaded cells. There are multiple instances of overloaded cells that have less overloaded neighbors and similarly idle cells surrounded by more loaded ones. Such cases will allow adaptive re-

assignment of users based on careful manipulation of the power. Such manipulation and successful reassignment will rely on an “associate-with-the-strongest-signal-tower” policy for mobile users, which is commonly adopted by the industry [6].

In order to quantify the amount of disparity in neighboring cells and hence the existence of spatial imbalance, we define the notion of local disparity in load δ. This measure ranks high cells whose load makes them outliers in their neighborhood. More quantitatively, we measure the average disparity δ_i of a cell i’s neighborhood as:

$$\delta_i = \frac{\sum_{j \in N_i} |V_i - V_j|}{|N_i|},$$

where N_i is the set of neighboring cells and V_i is the measured quantity in a cell (average call duration or number of calls).

We plot the distribution of the load disparity δ_i for the Dakar region towers in Fig. 4. During the night and morning time intervals, almost all towers have disparity between 0 and 500. This is likely due to the fact that there is less overall load during these times. Therefore, the largest of the average call values, those which cause the most disparity, will not usually occur during these intervals, but rather during the midday and evening hours. This is supported by the figure where we observe that in the midday and evening intervals, the bulk of the disparities lie between 500 and 1500.

4. TOWARDS SPATIAL LOAD BALANCING OF MOBILE CELL LOAD

In this section, we propose the problem of spatial load balancing and experiment with the utility of simple local heuristic to improve the cell load based on cell power adjustment.

4.1 Adaptive Power Assignment (APA)

Mobile network users associate with the base station from which they receive the strongest signal. This method of user association often leads to uneven distribution of load across the network, whereby some cells become overloaded, while others remain underutilized. As we demonstrated in the previous section using real world data from Senegal, the uneven
load is a phenomenon particularly exacerbated by the density of the user population and peak hours of the working day. In order to achieve better balance and eventually improved quality of service when the system operates at peak capacity, mobile cellular networks make use of load balancing mechanisms. This further helps utilizing the available resources and achieve a more uniform distribution of the load.

Currently, load balancing is achieved by performing handover of users from overloaded cells to adjacent cells with spare capacity. This approach limits the ability of the network to provision for expected bursts in traffic load. In order to tackle this problem we propose a method that relies on power control in order to balance the load spatially in the network. Particularly, by manipulating the emitted power at a given base station we can change the signal level received by users’ handsets and thus switch handsets across cells. Since this approach does not require knowledge of user identity, it allows for proactive configuration of network resources that is informed by predicted network load.

While being able to predict user mobility will enable a fully flexible and predictive approach to network resource utilization, we first evaluate the possible gain of a data-driven user redistribution approach if the loads at a given time point are known. A natural next step, that we discuss briefly in the discussion, is to add a predictive model for user mobility that will allow the network to “foresee” the future resource requirements and reconfigure in a proactive manner based on future estimates.

Let $\mathcal{S} = \{S_i\}$ denote the set of antenna sites, $\mathcal{V} = \{V_i \in \mathbb{R}^2\}$ denote their corresponding loads and $P = \{P_i \in \mathbb{R}^+\}$ denote the power with which the cell towers emit signal. The area in which a given tower’s signal is the strongest is called the tower cell. When all towers emit at the same power level, the set of all cells and antenna locations correspond to a Voronoi diagram of the space.

When towers emit at different power, the resulting cells start deviating from a regular Voronoi diagram. For example, a tower emitting a signal at a higher power than its neighbors will have a larger surface cell than its counterpart if all towers emit at the same power level (i.e. the Voronoi diagram case).

Power diagrams are natural extensions of Voronoi diagrams in which cells are similarly defined as the set of points that are closest to a given site, however the distance function is generalized based on an additional power parameter \[P \geq 0\]. Depending on the type of generalized distance function, power diagrams can be multiplicative (i.e. the distance is normalized by the power values) or additive \[8\]. We resort to the additive definition in our model according to which a new distance from a site $S_i$ to a point $X$ in the 2-D space is computed as

$$d_P(S_i, X) = d_{EUC}(S_i, X)^2 - P_i,$$

where $d_{EUC}$ is the euclidean distance and $P_i$ is the power associated with the site $S_i$. With this definition, the boundaries between cells are shown to remain polygonal \[8\]. A power diagram $\mathcal{D}$ for the sites $S$ with power weight $P$ is then defined as the polygonal partitioning of the space such that every cell is the set of points closest to a corresponding site $S_i$ according to $d_P(\cdot)$.

Constructing a power diagram, just like Voronoi diagrams, takes $O(n \log n)$, where $n$ is the number of sites \[2, 8\]. We adopt the methods used by Nosaj and colleagues \[8\] and refer to the computation of a power diagram given specified powers as $\text{PowerDiagram}(S, P)$.

The notion of disparity $\delta$ from the previous section is similarly defined for power diagrams. Modifying the power assignment $P$ results in changes in the area of cells. We assume that users have uniform activity and that they are located uniformly within the space of a cell of the power diagram. Hence, we associate a unit surface of the cell with a corresponding fraction of load that the cell observes.

Let the original uniform power assignment be $P^0$ and the corresponding area of various cells be $\text{Area}^0(S)$. Let us also denote the original load of the various sites as $V^0$. Given a new power assignment $P$, the new loads of sites are redistributed according to the changes in area, assuming uniform

![Figure 3: Voronoi diagrams showing cells in the region of Dakar during four periods of the day. The diagrams are colored based on the average number of calls in each cell during each time period compared to the maximum average call value observed thus far. Darker red color in a cell corresponds to a higher number of calls. From this visualization, it is evident that there are multiple instances of neighboring cells with different loads.](image)
the load balancing problem in terms of the resulting loads $V_i$ and $V_j$, according to the power diagram $\mathcal{D}^{OPT}(S, P)$. The intuition is to find cells whose load is very different from their neighbors and update their power such that the load is spread within the neighborhood. Our goal is to measure how their neighbors and update their power such that the load is spread within the neighborhood. Our goal is to measure how much can be gained in terms of load balance using a simple possibly suboptimal greedy algorithm, that also lends itself to distributed implementations in real networks.

Algorithm 1 Greedy APA

Require: $S, V, \Delta, \maxit$
Ensure: $P$
1: Initiate $P$ uniformly
2: $\mathcal{D}^{old} \leftarrow \text{Power-Diagram}(S, P)$
3: while iterations $\leq \maxit$ do
4: // Find highest discrepancy site
5: $S_m = \text{MaxDiscSite}($\mathcal{D}^{old}, V)$
6: $D_m = \sum_{j \in N_m, V_j \geq V_m} V_j - V_m$
7: $D_v = \sum_{j \in N_m, V_j < V_m} V_m - V_j$
8: if $D_m \geq D_v$, then
9: $P_m = P_m + \Delta$
10: else
11: $P_m = P_m - \Delta$
12: end if
13: $\mathcal{D}^{new} \leftarrow \text{Power-Diagram}(S, P)$
14: $V \leftarrow \text{UpdateLoad}(m, \mathcal{D}^{old}, \mathcal{D}^{new}, V)$
15: end while

Alg. 1 outlines our overall procedure for computing a power assignments of cells that achieves a good balance of load. The input to the algorithm includes the set of sites $S$, the load values $V$ associated with each sites which may correspond to call duration or number of calls, an atomic unit of increment of power $\Delta$ and the maximum number of iterations $\maxit$. We first initialize the power of all sites to be equal and compute a power diagram, which for this assignment is equivalent to a Voronoi diagram (Steps 1 and 2). Next, we perform $\maxit$ iterations of updates (Steps 3-15). In each iteration we first identify the highest discrepancy site $S_m$ (Step 5), and depending on the predominant source of discrepancy, relatively more loaded of relatively less loaded neighbors, we correspondingly increase (Step 9) or decrease (Step 11) the power of site $S_m$. Next, we compute a new power diagram $\mathcal{D}^{new}$ that takes into account the update to $P_m$ (Step 13) and lastly update the load $V_m$ of $S_m$ and its neighbors based on the changes in the respective areas. Note, that our assumption is that load (or users) are distributed uniformly within the cell’s area of dominance and hence reduction of area leads to proportional reduction of load. We can compute a more precise power assignment in the presence of actual user positions, however, when positions are unknown a spatially uniform assumption suffices.

Algorithm 2 UpdateLoad

Require: $m, \mathcal{D}^{old}, \mathcal{D}^{new}, V^{old}$
Ensure: $V^{new}$
1: if $\text{Area}(S_m, \mathcal{D}^{old}) < \text{Area}(S_m, \mathcal{D}^{new})$ then
2: $V_m^{new} = V_m^{old}$
3: for $\forall j \in N_m$ do
4: $\Delta V = \frac{\text{Area}(S_j, \mathcal{D}^{old}) - \text{Area}(S_j, \mathcal{D}^{new})}{\text{Area}(S_j, \mathcal{D}^{old})}$
5: $V_m^{new} = V_m^{old} - \Delta V$
6: $V_m^{new} = V_m^{new} + \Delta V$
7: end for
8: else
9: // compute reduction of $V_m$
10: $\Delta V = \frac{V_m^{old} \text{Area}(S_m, \mathcal{D}^{old}) - \text{Area}(S_m, \mathcal{D}^{new})}{\text{Area}(S_m, \mathcal{D}^{old})}$
11: $\Delta A = \text{Area}(S_m, \mathcal{D}^{old}) - \text{Area}(S_m, \mathcal{D}^{new})$
12: $V_m^{new} = V_m^{old} - \Delta V$
13: // distribute $V_m$ reduction to neighbors
14: for $\forall j \in N_m$ do
15: $V_j^{new} = V_j^{old} + \frac{\text{Area}(S_j, \mathcal{D}^{new}) - \text{Area}(S_j, \mathcal{D}^{old})}{\Delta A}$
16: end for
17: end if

The update of neighborhood loads upon adjusting the power of a given cell $S_m$ is described in Alg. 2. The input includes the index of the site to be updated $m$ and the power diagrams corresponding to the previous and current power assignments: respectively $\mathcal{D}^{old}$ and $\mathcal{D}^{new}$. If the area of $S_m$ has increased, we initiate the new value of the load $V_m$ with its old value (Step 2) and apply updates for every neighbor (Steps 3-7). For each neighbor we compute the decrease of load $\Delta V$ as a fraction of the difference of its surface to its old surface (Step 4) and update the new load values accordingly (Steps 5,6). Alternatively, we also consider the case in which the power $P_m$ and hence the surface of the cell corresponding to $S_m$ has shrunk (Steps 9-16). In this case, we first compute the reduction of its load value $\Delta V$ (Step 10) and its area $\Delta A$ (Step 11). Next, we reduce the load of $S_m$ (Step 12) and increase the loads of all neighbors proportionally to the fraction of their surface gain and the total surface loss of $S_m$ (Steps 14-16).

4.3 Evaluation of the proposed system
We implemented our proposed approach and tested its effectiveness on the real-world data from Senegal. Figure 5 shows the reduction of the maximum disparity $\delta$ as a function of the number of iterations, when applied to the average number of calls in the morning period in the Dakar region. We plot separate curves for different settings of the atomic unit of increment of power $\Delta$. More aggressive power updates (higher $\Delta$) achieve a faster reduction of the maximal disparity reducing it by almost half of its original values in 500 iterations. One interesting aspect is the observed spikes in the various curves. The reason for spikes is the greedy selection for sites whose power is to be updated. It is possible that directly updating the site of largest disparity may increase the disparity of neighboring cells. A more exhaustive probing for an optimal site to update will avoid this problem, but add more complexity to the algorithm and possibly require more site coordination in a distributed version of the algorithm.

Figure 6 shows the behavior of our algorithm when applied to the average number of calls in the four different periods of the day. The original disparity in the midday and evening periods are much higher than the those in the morning and at night. Our algorithm was able to reduce the disparity in half (from over 4000 to 2000) in less than 100 iterations of the greedy APA algorithm ($\Delta = 15.0$). We observed the greatest disparity in the morning and evening periods as the network usage and hence the disparity in those periods is relatively lower.

Finally, Fig. 7 presents the loads of the original Voronoi diagram and the power diagram computed by our greedy APA algorithm after 500 iterations. The figure “zooms in” on an area with several cell whose load drastically changed resulting in the reduction of the overall maximum disparity. Some of those cells are annotated by black arrows in both the initial and final diagram. As expected, overloaded cells become less loaded, while underutilized ones receive more load from their neighbors.

5. DISCUSSION

The analysis in this paper makes a number of assumptions about uniformity of user activity and position. A fine-grain information on user positions (or their estimates) can enable better redistribution schemes and more realistic estimation of the balance introduced by our scheme. Our approach can potentially affect centralized commercial cellular network, by allowing better load balancing and quality of service. Our long-term goal however is to evaluate the application of such scheme for ad-hoc disaster relief networks and deployments in remote areas [10]. For the latter scenarios, we will need to develop a energy-minimizing version of APA as consumed energy will be another important optimization criterion along with allowing connectivity and acceptable QoS.

Another future direction is to incorporate user mobility models and adaptively perform APA for a future period based on predicted cell occupation.

6. CONCLUSION

We introduced a novel data-driven approach to load balancing in mobile cellular networks as an important mechanism that enables distribution of demand across neighboring cells. While existing approaches for load balancing are reactive, redistributing users only when the offered load approaches the cell capacity, we focus on a proactive data-driven solution that provisions resources based on expected demand at a given time of the day. We demonstrated the existence of imbalance in the data from Senegal and develop an approach capable of improving the maximal imbalance of cells using manipulation of emitted power. We applied
our method on a real-world cellular network in Senegal and showed that it enables better distribution of load in Orange Telecom’s network in Senegal.

7. REFERENCES


Spatiotemporal dynamics of mobile communication volumes can be used as a proxy for ambient population distribution and activity. Spatiotemporal analysis of hourly call and text volumes from 1666 cell towers (sites) throughout Senegal in 2013, provided by the Orange Data for Development (D4D) challenge, depicts both daily and weekly cycles as well as spatial and temporal disruptions to these cycles. We use spatial and temporal correlation matrices and Empirical Orthogonal Function (EOF) analysis to characterize spatiotemporal dynamics of communication and infer collective behaviour from call and text volume data. The spatial and temporal correlation matrices identify spatiotemporal disruptions to normal communication patterns. Specifically, spatial disruptions associated with holiday travel and temporal distinctions between urban and rural communication patterns. The EOF analysis identifies the spatial and temporal patterns that most concisely represent the dominant features in the data. The temporal feature space of low order Principal Components provides a simple representation of the diversity of temporal patterns of call and text communication. We illustrate the use of these analytical tools to highlight spatiotemporal differences between calling and texting in urban and rural Senegal. The topology of the feature space clearly distinguishes between high volume weekly periodicities associated with developed urban areas and low volume non-periodic communication associated with rural populations. Similar analyses of longer time series of call and text volumes could potentially distinguish effects of interannual variability of seasonal climatic parameters (e.g. timing of precipitation) on rural populations and evolution of transportation networks on urban and peri-urban populations.

Introduction

Most current knowledge of collective human behavior comes from field surveys and relatively small numbers of retrospective observations – often qualitative with unknown accuracy or degree of representation of the impacted population. In contrast, mobile communication data can provide pervasive, quantitative observations of human communication patterns. With sufficient spatial and temporal detail, mobile communication data can even be used as proxies for other
types of collective behavior (e.g. mobility (Becker et al. 2013; González et al. 2008)). In this study we use a spatially and temporally extensive collection of voice call and SMS text message volumes to quantify spatiotemporal communication patterns in Senegal in 2013. Spatiotemporal analysis of hourly call and text volumes from 1666 cell towers (sites) throughout Senegal in 2013, provided by the Orange Data for Development (D4D) challenge, depicts both daily and weekly cycles of movement and communication - as well as spatial and temporal disruptions to these cycles.

We use spatial and temporal correlation matrices combined with Empirical Orthogonal Function (EOF) analysis to characterize spatiotemporal dynamics of communication and infer collective behaviour from call and text volume data. The spatial and temporal correlation matrices identify spatiotemporal disruptions to normal communication patterns. Specifically, spatial disruptions associated with holiday travel and temporal distinctions between urban and rural communication patterns. The EOF analysis identifies the spatial and temporal patterns that most concisely represent the dominant features in the data.

The objective of this analysis is to illustrate a novel approach to characterization of spatiotemporal regularities and irregularities in communication patterns. By characterizing the regularities in collective behavior, a standard is established against which irregularities can be identified. This two stage characterization provides an simple and objective way to identify and quantify the dominant spatial and temporal patterns present in spatiotemporal observations with a minimum of assumptions about the form of the patterns. We introduce the concept of the temporal feature space to the analysis of mobile communication data, and use sets of temporal feature spaces to identify spatial and temporal disruptions to regular patterns of communication. We then illustrate the utility of the spatiotemporal characterization and the temporal feature space to map locations of high communication volume with strong weekly cycles. These tools have potential application to economic development in their ability to characterize some aspects of collective behavior that cannot easily be captured using traditional tools (e.g surveys).

**Spatial and Temporal Correlation Matrices**

The temporal periodicity and geographic variability of mobile communication activity can be illustrated efficiently using Time-Space maps in which call and text volumes are shown as a function of Julian day and geographic latitude of site (Fig. 1). At the scale of Fig. 1, the structure of each Time-Space map is dominated by the weekly cycle and the contrast between higher and lower volume sites. In addition, the Time-Space map reveals the presence of numerous zero-volume sites and large gaps of missing data at specific towers. We visually inspected each of the 1666 time series from each geographic site (tower) and eliminated those with zero call or text volume or gaps larger than 7 days. This reduced the 1666 sites to 1331 usable sites.
The regularities in the Time-Space maps become more apparent when the data are aggregated in time and space. Averaging call and text volumes in time provides spatial maps of activity (Figure 2). At national scales, the greatest contrast is seen between Dakar and the rest of the country – both in call and text volumes. Similarly, averaging call and text volumes in space provides time series of activity. To better represent the dominant daily and weekly periodicities, we use calendar plots in which spatially aggregated call or text volume is shown as a 2D function of hour of day, hour of week, and week of year (Figure 3). On the annual time scale the bimodal daily cycle of call activity and seasonal variations in text activity are readily apparent. The distinction between daily call and text cycles is apparent with the emergence of a daily bimodal peak in call volumes contrasting a single early evening peak in text activity.
volumes. A seasonal distinction between call and text volumes also appears during the month of Ramadan and the weeks following Ramadan. Call volumes drop during the day in Ramadan while night time text volumes increase. The increase in night time text volumes is more apparent in the Log$_{10}$ of the text volumes. These aggregate measures show the dominant spatial and temporal patterns but they do not capture spatiotemporal variations in these patterns.

Figure 2. Call and text volume maps. Both call and text volumes have heavy tailed distributions in space. Most sites have relatively low volumes while a minority of sites have much higher volume. Both the ratio of texts to calls, and their temporal correlation are higher in Dakar.
The Time-Space maps show all the data without the need for interpolation but some of the spatial and temporal relationships among sites are not immediately apparent. Spatial and temporal disruptions to the dominant geographic and cyclic structure are more clearly illustrated using spatial and temporal correlation matrices (Figures 4 & 5). Spatial correlation matrices show the temporal evolution of spatial patterns. The spatial correlation of temporal patterns among all sites in the study area changes in time as daily and weekly patterns repeat. Higher spatial correlations among individual site volumes occur at times when spatial distributions of call or text volumes are more similar; lower correlations occur at times when the spatial patterns are less similar. At the scale of the entire year shown in Figure 4, the weekly cycle is apparent as the regular grid pattern of higher correlations between pairs of weekdays and lower correlations between weekdays and weekends. Darker bands of lower spatial correlation occur most prominently on holidays when spatial patterns differ most strongly from the normal patterns occurring throughout the year. The lowest spatial correlations (hence largest spatial disruptions) of the year

![Calendar plots of call and text volumes. Daily and weekly cycles dominate both call and text volumes. Call volumes drop during Ramadan (weeks 27 to 30) but night text volumes increase and persist for weeks after the end of Ramadan.](image)

Figure 3: Calendar plots of call and text volumes. Daily and weekly cycles dominate both call and text volumes. Call volumes drop during Ramadan (weeks 27 to 30) but night text volumes increase and persist for weeks after the end of Ramadan.
occur on the days of and preceding the Prophet’s Birthday (jd 24) and Christmas (jd 359). Note also the more blocky structure and generally lower correlations for text volumes compared to call volumes. The higher correlations for calls indicate that call volume patterns are more spatiotemporally regular than text volume patterns over the course of the year. The blockier structure of the text volume correlation matrix suggests that spatial patterns of text volumes change more abruptly throughout the year than spatial patterns of call volumes.

Figure 4  Spatial correlation matrices and distributions. Correlations range from 0 (black) to 1 (white) on each matrix. Each matrix shows the spatial correlation coefficient of site volumes between every pair of days for 1331 usable sites. Darker bands correspond to times that have spatial patterns of site volume that are relatively dissimilar to spatial volume patterns of other times. Both call and text volume distributions are dominated by high (> 0.8) spatial correlations, indicating persistence of spatial patterns of volumes through time. Julian days of lower correlation correspond to days preceding holidays.

The complement to the spatial correlation matrix is the temporal correlation matrix. Temporal correlation matrices show the spatial variations (similarity and difference) in temporal patterns among each pair of sites at different geographic locations (Figure 5). Higher correlations occur between pairs of sites that have more similar temporal patterns. Here the corrupting effect of unusable sites with missing data is apparent as these sites have low correlations with all other sites. Removing these unusable sites has a profound effect on both the temporal
correlation matrix and the distributions of correlations. With unusable sites removed, it is much easier to see the distinction of the temporal patterns in Grand Dakar and Touba, as well as many other smaller locations with distinct temporal patterns. While spatial and temporal correlation matrices offer a simple way to identify spatial and temporal similarities among times and places (respectively), they give no indication what these spatial and temporal similarities are. To resolve these patterns, we can exploit information contained in the eigenstructure of the correlation matrices.

Figure 5 - Temporal correlation matrices and distributions. Correlations range from -1 (black) to 1 (white) on each matrix. Each matrix shows the temporal correlation coefficient of volume time series between every pair of 1666 sites (top) and 1331 usable sites (bottom). Darker bands correspond to sites that have temporal patterns of daily volume that are relatively dissimilar to volume patterns of other sites. Note generally higher correlations among temporal patterns of text volume compared to call volume - both before and after unusable sites are removed.
Principal Components and Empirical Orthogonal Functions

Principal Component (PC) transformations are commonly used to represent uncorrelated modes of variance in high dimensional data. Because spatial patterns are often correlated in time, PC transforms provide an efficient low dimensional projection of the uncorrelated components of the spatial and temporal patterns in spatiotemporal data. The same property applies to temporal dimensions. In meteorology and oceanography the PC transformation provides the basis of Empirical Orthogonal Function analysis; a standard tool for analysis of spatiotemporal patterns and processes. (see Bretherton et al. 1992; Preisendorfer 1988; von Storch and Zwiers 1999) for overviews). PC transformations have been used by (Eagle and Pentland 2009) to characterize vectors of longitudinal behavioral data for the purpose of classification and prediction of “eigenbehaviours” in individuals daily routines. PC transformations have also been used to characterize patterns in mobile communication data. (Reades et al. 2009) used PC transformations with mobile phone data in Rome to identify “eigenplaces” as recurring patterns of mobile phone usage in geographically specific contexts in Rome. (Calabrese et al. 2010) used PC transformations to represent temporal patterns in wireless network data as input to a clustering algorithm used to identify locations with common temporal patterns. The EOF analysis we perform here is similar to these studies in that we seek to use the PC transform to minimize redundancy of temporal patterns and identify the spatial distribution of key temporal patterns. Our analysis extends previous work in this area by using the temporal feature space (Small 2012) of the low dimensional PCs to represent continuous variations in temporal patterns discontinuously across geographic space. The temporal feature space is a Euclidean space in which the dimensions correspond to the temporal patterns that represent the spatiotemporal variance most efficiently in terms of the smallest number of dimensions.

The PC transform provides a very convenient tool for identification of spatiotemporal patterns. By rotating the coordinate system to align with orthogonal dimensions of uncorrelated variance, any location-specific time series $P_{xt}$ contained in an $N$ geographic time series can be represented as a linear combination of temporal patterns, $F$, and their location-specific components, $C$, as:

$$P_{xt} = \sum_{i=1}^{N} C_{ix} F_{it}$$  \hspace{1cm} (1)

where $C_{ix}$ is the spatial Principal Component (PC) and $F_{it}$ is the corresponding temporal Empirical Orthogonal Function (EOF) and $i$ is the dimension. The EOFs are the eigenvectors of the covariance matrix that represent uncorrelated temporal patterns of variability within the data. The PCs are the corresponding spatial weights that represent the relative contribution of each temporal EOF to the corresponding pixel time series $P_{xt}$ at each location $x$. The relative contribution of each EOF to the total spatiotemporal variance is given by the eigenvalues of the covariance matrix. $N$ is the number of...
discrete dimensions represented by the data; which may be greater, or less, than the true physical dimensionality of the process(es) measured. Principal Components are uncorrelated but not necessarily independent – unless the data are jointly normally distributed. In systems where the same deterministic processes are manifest at many locations, but stochastic processes are uncorrelated, the variance of the spatiotemporal structure of the deterministic processes can be represented in the low order PC/EOF dimensions while the stochastic variance is represented in the higher order dimensions (Preisendorfer 1988). When a clear distinction can be made, this can provide a statistical basis for separation of deterministic and stochastic components of an image time series. However, the transformation is purely statistical so there is no guarantee that the separation is physically meaningful.

We characterize the dominant spatial and temporal patterns in the call and text data by applying a PC transformation to the Time-Space maps of the 1331 usable site time series of daily call and text volumes. The result is a set of temporal EOFs representing the canonical temporal patterns and a corresponding set of spatial PCs representing the relative importance of each temporal EOF to each geographic location (site). The five low order EOFs of the call and text data are shown in Figure 6. While each EOF shows some distinct feature(s) related to specific dates of the year (e.g. holidays when communication volumes increase), it is also apparent that the transformation has not completely separated different processes into different dimensions. This mixing of processes is one of the primary challenges to the use of EOF analysis in Oceanography and Meteorology. This comes about because the EOFs are purely statistical constructs, effectively independent of the physical processes responsible for the observations. Similarly, the spatial patterns of the individual PCs show multiscale heterogeneity and
defy simple interpretation (Figure 8). We attempt to resolve this dual conundrum by considering combinations of temporal EOFs as they are represented in the temporal feature space of the PCs.

**Temporal Feature Spaces**

The temporal feature space of the PCs reveals the relationships among the different uncorrelated patterns in the EOFs. Figure 8 shows three orthogonal projections of the 3D space of low order PCs – as well as a 2D projection of PCs 4 and 5. For simplicity, we focus only on the temporal EOFs and spatial PCs of the call volume data for the remainder of this analysis. The temporal feature space in Figure 7 can be visualized as a 3D cloud of points in which each point corresponds to a geographic location (site) with its own (possibly non-unique) time series of call and text volumes. Hence the location of
Each point in the temporal feature space gives some indication of the relative importance of the corresponding temporal EOF to the temporal pattern of that site. Different temporal patterns occupy different parts of the temporal feature space. Points closer together in the space have more similar temporal patterns. Several examples of individual point time series of call volumes are shown in Figure 7. As would be expected from comparison with the temporal patterns in EOFs 1-3, the 3D temporal feature space of call volumes is dominated by patterns associated with the Prophet’s Birthday and Christmas, as well as the overall geographic distribution (PC/EOF 1) with much higher volumes occurring in Dakar than elsewhere in Senegal. In contrast, the temporal feature space of PCs 4 and 5 is dominated by weekly cycles with high volumes on weekdays and low volumes on weekends. We illustrate how this 2D projection of the temporal feature space can be used to identify high volume sites dominated by weekly cycles.

Because temporal EOFs 4 and 5 are both dominated by weekly cycles (rather than holidays), these two EOFs can interfere constructively and destructively according to the sign of their corresponding PCs. Points occupying quadrant I of the PC 4/5 space correspond to sites where the EOFs interfere constructively to give high amplitude weekly cycles, like those shown in examples g, h and i in Figure 8. Note that examples g and h correspond to sites with high weekday volumes and low weekend volumes while site i is just the opposite with higher weekend volumes. These sites with high volume weekly cycles are all located in Grand Dakar as shown in the map in Figure 9. Many of these sites are also characterized by high temporal correlations of call and text volumes (Figure 2). We conjecture that these locations with high call volumes and strong weekly cycles as well as in-phase text volume cycles correspond to areas with large transient populations of commuters with strongly synchronized work schedules. The fact that text volumes are strongly synchronized with call volumes suggests that much of the text communication is related to work activities because it decreases in parallel with call volumes on weekends.
Figure 8 Temporal feature spaces and temporal endmember time series of call volume. Orthogonal projections of PCs 1, 2 and 3 show outlier time series associated with religious holidays (a, b, c). The space of PCs 4 and 5 shows strong weekly periodicities.
Figure 9  Strong weekly cycles in time and space. Sites with strong weekly cycles (top, red) occur entirely within Grand Dakar - but are not a majority of sites. Base levels of call volumes (bottom) vary considerably from site to site.
Conclusions

The combined use of spatial and temporal correlation matrices, EOF analysis and the temporal feature space provides a relatively simple and easily implemented set of tools for characterization of spatial and temporal patterns in mobile communication data. One of the primary benefits of this suite of tools is the minimum of assumptions required for its implementation. The principal assumption of EOF analysis is that variance represents information and that correlation represents redundancy. While there are clearly many situations where this is not true (e.g. anomaly detection), it is a reasonable assumption for a wide range of problems. In these cases, the sequence of analyses illustrated here provides a relatively straightforward way to quickly characterize the structure of large, high dimensional data sets.

The relevance of these preliminary results to development is multi-fold. Rapid, objective identification of distinct temporal patterns allows analysts to draw inferences and hypotheses about the activities causing the patterns. Our conjecture about high volume weekly cycles and commuters could have implications for development of transportation corridors. Spatial and temporal consistencies in the magnitude and timing of holiday

Figure 10 Night lights and call volumes. Night light composite showing 2013 VIIRS brightness (gray) + OLS decadal change (color). Color implies change. Warmer colors indicate brightening. Locations of example time series from Figure 8 shown by letters. Most time series correspond to areas with sufficiently lighted development to be imaged from space - but not all. Smaller dimmer lights in southeast likely due to fire.
travel could also be useful for planners and transportation engineers. High call volumes that do not show strong periodicities may correspond to economic activities not dictated by the 5 day work week (e.g. farming). Low text to call ratios may give some indication of literacy of users in specific locations (See Kirchberger and Small, this volume, for details). Similar analyses of longer time series of call and text volumes could potentially distinguish effects of interannual variability of seasonal climatic parameters (e.g. timing of precipitation) on rural populations and evolution of transportation networks on urban and peri-urban populations.

Acknowledgements CS gratefully acknowledges the support of the NASA Socioeconomic and Data Applications Center (SEDAC) and the Doherty Foundation.

References

1. Introduction

Even anonymized and space- or time-coarsened records of cell phone use by individual subscribers provide, in principle, timeseries of snapshots of those individuals’ coarse locations. To the extent that their locations between these snapshots can be inferred, and to the extent that we have a representative sample of users, these data can provide a high resolution estimate of population mobility. Such estimates have a wide range of potential development applications: for individuals they reflect socio-economic factors [1]; on small scales they can predict the draw on power grids; they can support transportation and disaster planning [2], or parameterize epidemiological models [3,4], over a range of space- and time-scales; and on large scales they can reveal circulatory migration patterns or urbanization trends [5]. Previous efforts to analyze human mobility using telecommunications data have not attempted sub-daily estimates [4]; in this preliminary report we propose a novel method for extracting mobility information at the same time resolution as it is provided in telecommunications data.

2. Background

As part of the Data for Development (D4D) Sénégal Challenge, Orange provided 25 fortnightly datasets spanning almost all of 2013, each including approximately 300,000 (anonymized) users, and the locations (cell towers) and times (in 10 minute intervals) of their calls [6]. (There is also a dataset consisting of about 150,000 users for the whole year, but their calls are located only to the much coarser level of an arrondissement, of which there are 123, compared to 1666 cell towers; hence we focus on the fortnightly data.) Orange’s Sénégal partner is Sonatel, which has a market share of about 2/3; the total number of mobile subscriptions is slightly greater than the population. In 2014, there has been an Ebola outbreak in West Africa [4]; while there has been to date only a single case in Sénégal, it has emphasized the importance of knowing the real-time movement of persons and motivated the development of the algorithm described here.

3. The model

Consider a set of $N = 1666$ locations and a timeseries $\{X_t \in [N] \mid t \in [T]\}$, where $T = 6 \times 24 \times 14 = 2100$. At each timestep there is an observed signal (emission) $Y_t \in [N] \cup \{0\}$, where 0 indicates that no call was made. (We assume that in each 10 minute time interval
only one location is observed for the user. If there is more than one, we use the one which has the maximum number of calls.) The distribution of these signals is:

\[ b_j(y_t) = \Pr(Y_t = y_t \mid X_t = j) = \begin{cases} p_j(t) & \text{if } y_t = j; \\ 1 - p_j(t) & \text{if } y_t = 0; \\ 0 & \text{otherwise.} \end{cases} \]

It will be convenient to define \( Z(t) \) to be a diagonal matrix with \( z_{jj}(t) = 1 - p_j(t) \). We will model mobility by a sequence of Markov transition matrices:

\[ A(t) = (a_{ij}(t) = \Pr(X_{t+1} = i \mid X_t = j)) \].

Define the initial distribution over locations to be \( \pi_j = \Pr(X_1 = j) \).

If there were no time dependence for the \( p_j \)'s and the \( a_{ij} \)'s, this would be a (special case of a) hidden Markov model. But we want to keep time dependence because we expect the probability of making a call, or moving from one location to another, to depend at least on the time of day. Figure 1 shows the total call volume as a function of time in January 2013; the daily periodicity is clear, with a huge dynamic range, and there seem to be weekly and monthly trends as well. To incorporate this phenomenon, we define functions \( f_j(t) \) which we think of as the “active fraction of the population at location \( j \) at time \( t \)”. Then we set

\[ p_j(t) = E_j[f_j(t)], \]

and, for \( i \neq j \),

\[ a_{ij}(t) = T_{ij}[f_j(t)]. \]

Here \( E_j[\cdot] \) and \( T_{ij}[\cdot] \) depend on some parameters \( \epsilon_j \) and \( \tau_{ij} \), respectively, but we expect \( E_j' \geq 0, T_{ij}' \geq 0 \). This makes our model different than the typical hidden Markov model, which presumes constant emission and transition probabilities.

Figure 1. Total call volume as a function of time for January 2013.
4. Estimation

If this were a typical hidden Markov model we could estimate its parameters using the Baum-Welch algorithm [7,8]. Since it is not, we develop a novel expectation maximization algorithm to estimate the parameters $\Theta = (\pi_j, \epsilon_j, \tau_{ij})$, given $y$ and $f_j(t)$, as follows:

Suppose we have computed $\gamma_j(t) = \Pr(X_t = j \mid y, \Theta)$. Then we would update $\pi_j$ to be $\gamma_j(1)$, and $\epsilon_j$ so that the expected number of emitted $j$s is the observed number. More precisely, the expected number of emitted $j$s is:

$$\sum_{t=1}^{T} p_j(t) \gamma_j(t) = \sum_{t=1}^{T} E_j[f_j(t)] \gamma_j(t).$$

Let $k_j = |\{t \in [T] \mid y_t = j\}|$. Then we would like to solve

$$k_j = \sum_{t=1}^{T} E_j[f_j(t)] \gamma_j(t)$$

for $\epsilon_j$, or at least find values for $\epsilon_j$ that minimize $\sum_j (k_j - \sum_t E_j[f_j(t)] \gamma_j(t))^2$.

Similarly, suppose we have computed $\xi_{ij}(t) = \Pr(X_t = j, X_{t+1} = i \mid y, \Theta)$. Then we would update $\tau_{ij}$ so that this agrees with the expected number of $j \rightarrow i$ transitions:

$$\sum_{t=1}^{T-1} \xi_{ij}(t) = \sum_{t=1}^{T-1} T_{ij} E_j[f_j(t)] \gamma_j(t),$$

or at least find values of $\tau_{ij}$ that minimize $\sum_{ij} (\sum_t \xi_{ij}(t) - T_{ij} E_j[f_j(t)] \gamma_j(t))^2$.

To compute $\gamma_j(t)$ and $\xi_{ij}(t)$ we first compute two sequences of (co)vectors, $\alpha(t) \in \mathbb{R}^N$ and $\beta(t) \in (\mathbb{R}^N)^t$, recursively, as follows:

$$\alpha(1) = \begin{cases} \hat{e}_{y_1} & \text{if } y_1 \neq 0; \\ Z(1) \pi & \text{otherwise}; \end{cases}$$

$$\alpha(t) = \begin{cases} \hat{e}_{y_t} Z(t) A(t-1) & \text{if } y_t \neq 0; \\ Z(t) \pi(t-1) & \text{otherwise}; \end{cases}$$

and

$$\beta(T) = \begin{cases} \hat{e}_{y_T}^T & \text{if } y_T \neq 0; \\ 1^T Z(T) & \text{otherwise}; \end{cases}$$

$$\beta(t) = \begin{cases} \hat{e}_{y_t}^T Z(t) & \text{if } y_t \neq 0; \\ \beta(t+1) A(t) Z(t) & \text{otherwise}, \end{cases}$$

where $1 \in \mathbb{R}^N$ is the vector of all 1s and $\hat{e}_j \in \mathbb{R}^N$ is the $j^{th}$ standard basis vector. Here the $A(t)$ and $Z(t)$ depend on the current estimate $\Theta$ of the parameters. Then

$$\gamma_j(t) = \frac{\beta_j(t) Z_j(t)^{-1} \alpha_j(t)}{\beta(t) Z(t)^{-1} \alpha(t)},$$

$$\xi_{ij}(t) = \frac{\beta_i(t+1) A_{ij}(t) \alpha_j(t)}{\beta(t+1) A(t) \alpha(t)}.$$
Thus the algorithm proceeds by initializing the parameters $\Theta$, computing $\gamma_j(t)$ and $\xi_{ij}(t)$, and then solving (1) and (2) to update $\Theta$. Iterating until some standard of convergence is achieved, we then output the estimated $\Theta$.

5. The ensemble

We run the estimation algorithm described in §4 for each of the approximately 300,000 users in a fortnightly dataset. The result is a set of parameters for each, on which we can use the Viterbi algorithm to compute the most likely sequence of locations, one per time step [9]. Doing this gives us an estimated population fraction at each location $j \in [N]$ at each time $t$. Using these estimates we can compute $f_j(t)$, the fraction of users at $j$ at time $t$ who make a call (or who transition to another location at $t + 1$), i.e., the “active fraction”. Notice that before analyzing the whole ensemble of users, we had no way to know how many were at location $j$ at time $t$, so we had no access to $f_j(t)$. Thus for the first pass through the algorithm we use $f_j(t) = f(t) =$ the fraction of users who make a call at time $t$. On the second pass through the algorithm we can differentiate between locations. And we can iterate the individual to ensemble to individual to ensemble sequence until some standard of convergence is achieved. (Although completely different in details, this use of the ensemble information is reminiscent of Blumenstock’s analysis of Rwandan telecommunications data [5].)

6. Applications

Having estimated the locations and the transition matrices for each user at each time, we compute the (population-weighted) average transition matrix at each location and each time. This gives us an almost instantaneous (10 minute intervals) picture of the flow of population from one location to another. Computing powers of this transition matrix gives the flow on larger timescales (e.g., the 6th power for hourly flow). As noted in §1-2, we are particularly interested in using this to parameterize epidemiological models, but also in using it to reveal patterns of circular migration of agricultural (peanut farming, specifically) labor, and of urbanization in Sénégal.

References

“Commentary: Containing the Ebola outbreak—the potential and challenge of mobile network data”, *PLoS Currents Outbreaks* (29 September 2014) Edition 1; doi: 10.1371/currents.outbreaks.0177e7f52217b8b634376e2f3efc5e.


INTRODUCTION

Most probability theory relies upon geometrical methods for analyzing data. For instance, a statistical distance must be defined so that two statistical objects can be quantified as being either close or far apart in some statistical measure. So, probability theory fundamentally encodes some type of length information. But, what if you want to concern yourself with a more fundamental property of the statistical objects: how are they structured? Topological data analysis, specifically the persistent homology method, accomplishes this. It determines the global structure of a set of data rather than its metric properties.

Topological data analysis is a new approach to analyzing the structure of high dimensional datasets. Persistent homology, specifically, generalizes hierarchical clustering methods to identify significant higher dimensional properties which are out of reach of any other approach. In this project, our goal is to analyze mobile network data from Senegal to determine whether significant topological structure is present. We investigate two independent questions: whether the introduction of the Dakar motorway has any significant impact on the topological structure of the data; and if communities can be detected using this method. With the level of resolution provided, this analysis does not detect changes in the invariant topological structure under any metric. For clustering, we find that modularity represents a better-founded approach than considering components of persistent $H_0$ groups constructed via the Vietoris-Rips complex.

A simplicial $k$-chain ($c_k$) is a sum of $k$-simplices ($\sigma_k$):

$$c_k = \sum_i \alpha_i \sigma_k^i, \quad \alpha_i \in \mathbb{F}$$

(1)

where $\mathbb{F}$ is some field. Each $k$-simplex can be thought of as a $k$-dimensional polytope. Thus, a 2-simplex represents a triangle; a 3-simplex represents a tetrahedron, etc.

Thus, various $k$-chains define a free Abelian group which is denoted as $C_k$ - i.e. $c_k \in C_k$. The boundary operator $\partial_k : C_k \rightarrow C_{k-1}$, is a linear homomorphism defined to act on $\sigma_k = [v_0, v_1, \ldots, v_k]$

$$\partial_k \sigma_k = \sum_i (-1)^i [v_0, v_1, \ldots, \hat{v}_i, \ldots, v_k] \in C_{k-1}$$

where “$\hat{v}_i$” means this element is removed from the simplex. This definition forces the condition used to compute homology: $\partial^2 \equiv 0$. This definition allows a flow of information in the various chain groups:

$$\ldots \rightarrow C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \rightarrow \ldots$$

Various subgroups of this map can be defined. In particular, the cycle group $Z_k \equiv \ker \partial_k$ and the boundary group $B_k \equiv \text{im} \partial_{k+1}$. Because $\partial^2 \equiv 0$, this implies $B_k \subseteq Z_k \subseteq C_k$. This condition is necessary so the homology group can be defined as the quotient group:

$$H_k \equiv Z_k/B_k = \ker \partial_k/\text{im} \partial_{k+1}$$

Each homology group, $H_k$, contains information about the existence of $k$-dimensional holes in the space. For instance, the torus has $H_0 = \mathbb{Z}$, $H_1 = \mathbb{Z}^2$, $H_2 = \mathbb{Z}$ and all the remaining homology groups vanish. Refer to Hatcher’s text [4] for a full treatment of the subject.

PERSISTENT HOMOLOGY

The previous discussion on homology requires the spaces to be triangulable, that is able to be thought of
as a sum of k-simplices. For an arbitrary data set, there is no fundamental procedure to triangulate this space. Various ways do however exist, each with their own distinct set of rules, that can be used to construct simplices from data. For each of these procedures, we choose the coefficients in equation 1 to be in $\mathbb{Z}_2$.

We use the terms point cloud and data set interchangeably. Let $d(a, b)$ denoted the distance in a metric space between points $a$ and $b$. Let $Z$ denote the point cloud. We refer to $\epsilon$ as the filtration value, or simply the filtration. Note that for a large enough filtration, the complex will become one connected component. For a small enough filtration, there will be as many connected components as there are vertices. A vertex set consists of the base set of points used to construct higher dimensional simplices. Refer to [5], [6], [7] for overviews of persistent homology.

**Vietoris-Rips Complex**

Given a point cloud, the Vietoris-Rips Complex $(R_\epsilon)$ defines k-simplices as being determined by $(k+1)$-tuples of points whose balls of radius $\epsilon/2$ pairwise intersect [5]. The balls are drawn around each point in the point cloud, and the radius can be computed with an arbitrary metric. Specifically, to construct $R(Z, \epsilon)$:

1. The vertex set is $Z$
2. Edge $[a, b]$ is in $R(Z, \epsilon)$ iff $d(a, b) \leq \epsilon$
3. Higher dimensional simplices are in $R(Z, \epsilon)$ if all of its edges are in $R(Z, \epsilon)$

One of the motivating reasons for this construction is that the union of the balls, which we interpret as being fundamentally representative of whatever topology the points came from, has a homotopy type that is closely related to the homotopy type of $R(Z, \epsilon)$ (see [8]).

**Lazy Witness Complex**

The construction of $R(Z, \epsilon)$ is computationally expensive because the entire point cloud is included as the vertex set. Choosing the vertex set as a subset of $Z$ reduces the computation necessary to construct the simplex set over the range of all filtration values. $L \subset Z$ is called the landmark set, and points in it are chosen in one of two ways by selecting from $Z$ [9]:

- **random** point selection: select points randomly from $Z$, the resulting set being $L$
- **maxmin** point selection: first, choose a random point in $Z$ to serve as the first point in $L$. Each additional point in $L$ is inductively chosen from $Z$ by maximizing $d(z, l_i) \forall l_i \in L, z \in Z$

The size of $L$ is variable depending on how large a vertex set is needed. Specifically, to construct $LW(Z, L, \epsilon, \nu)$:

1. The vertex set is $L$
2. Edge $[a,b]$ is in $LW(Z, L, \epsilon, \nu)$ iff $\exists z \in Z$ such that $\max\{d(a, z), d(b, z)\} \leq D_\nu(z) + \epsilon$
3. All higher dimensional simplices are in $LW(Z, L, \epsilon, \nu)$ if all of its edges are in $LW(Z, L, \epsilon, \nu)$

$D_\nu(z)$ is defined to be the distance from $z$ to its $\nu$th closest neighbor. A feature of the lazy witness complex is that it behaves like a Delauney triangulation of the space when $\nu = 1$; for $\nu = 0$, the complex behaves similarly to $R(Z, \epsilon)$ [9].

**Core Subsetting**

Core subsetting is a procedure that helps uncover statistically significant topological structure in data. Since the data in this analysis does not come from a pure topological structure, but rather from data in a real world process, a priori we do not expect the entire data set to have an interesting topological structure. Rather, we expect subsets of the data to have the interesting structure. The procedure of core subsetting follows. First, start with an arbitrary $n \times n$ metric space $A$:

$$A = \begin{pmatrix}
a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\
a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{n,1} & a_{n,2} & \cdots & a_{n,n}
\end{pmatrix}$$

The second step is to produce the density vector:

$$\Delta_k = \begin{pmatrix}
\delta_1^k \\
\vdots \\
\delta_n^k
\end{pmatrix}$$

Where $\delta_i^k$ is defined to be the inverse of the distance from the $j$th point in the metric space (i.e. the $j$th row of $A$) to the $k$th closest neighbor. Hence, $k$ is a parameter we scan over. A large $k$ can be thought of as giving a more global estimate of the topology; similarly, a small $k$ gives a more local estimate. The final step is to select a percentage (later on referred to as “P”) of the densest points in $\Delta_k$. The points from the metric space that give these densest points are then chosen to form a smaller metric space:
There are various quantities we examine when determining the fundamental topological structure of the point cloud. These quantities are computed with javaPlex - a software built to construct persistent homology from an arbitrary point cloud [10]. The quantities we examine are:

- **Barcode Plots**: These depict the various generators of the different \( \text{LW}(Z, L, \epsilon, \nu) \) or \( \text{R}(Z, \epsilon) \). The x-axis represents the filtration value; the y-axis represents, in no physically significant ordering, the different homology generators. A barcode exists for each \( H_n \).

- **Betti Numbers**: Are integers that count how many generators of a specific dimension exist at a specific filtration. For example,

\[
[H_1(\text{R}(Z, \epsilon), \epsilon = 3)] = 2
\]

means the first dimensional homology group for a Vietoris-Rips complex at a filtration of \( \epsilon = 3 \) has Betti Number equal to 2. In other words, it has 2 one-dimensional holes at this filtration.

- **Relative Dominance** [9]: A few definitions are necessary:

1. \( R_0 \) = The filtration value at which a certain topological structure appears
2. \( R_1 \) = The filtration value at which a certain topological structure disappears
3. \( R_0 \) = The filtration value at which the complex becomes one connected component

Relative Dominance is defined as \( \delta_R = \frac{R_1 - R_0}{R_n} \). By a topological structure appearing, we mean any combination of Betti numbers in the various dimensions. So, a large \( \delta_R \) corresponds to the topological structure being physically significant; a small \( \delta_R \) corresponds to the topological structure potentially being noise or a statistical fluctuation.

**METRICS FOR ANALYSIS**

We consider three metric spaces which arise from initially constructing a complete weighted graph, or equivalently a complete weighted adjacency matrix \( W_{ij} \), whose vertices correspond to the 1666 cell phone towers (or a subset thereof). In two of the three cases, this adjacency graph does not immediately define a metric space, as the triangle inequality is not satisfied. However, through the use of an algorithm determining the shortest path between any two points in a complete weighted graph, a genuine metric space can be constructed. This procedure was carried out to obtain two of the metrics defined below, while the third can be defined directly from its complete weighted graph.

**The Floyd-Warshall Algorithm**

Our computation of persistent homology requires an underlying metric space from which simplices and chain complexes can be defined. Given a complete weighted adjacency graph on a set of vertices \( X = \{1, 2, 3, ..., n\} \) with edge weights \( w_{ij} \), one can construct a metric space \((X, d)\) by the following construction, known as the Floyd-Warshall Algorithm. It is constructed recursively:

\[
\text{Path}_0(i, j) = w_{ij}, \\
\text{Path}_k(i, j) = \min\left( \sum_{\gamma \in \text{edges}(lm) \in \gamma} w_{lm} \right)
\]

where the minimum is taken over all paths \( \gamma \) in the adjacency graph from vertex \( i \) to vertex \( j \), using only vertices in the set \( \{1, 2, ..., k\} \) as intermediate vertices.

We then define a metric space \((X, d)\) by:

\[
d(i, i) = 0 \ \forall \ i, \\
d(i, j) = \text{Path}_n(i, j)
\]

One can verify that this satisfies the axioms of a metric space, since we begin with a complete weighted adjacency graph. Note: there are other possible constructions in going from a weighted adjacency graph to a metric space.

**Data Aggregation**

For each of the metrics we consider, the explicit construction of the metric from the data depends on a choice of aggregation period from the data provided. Let \( T \) denote an arbitrary set of time intervals during the year; for example, \( T \) could be the entire month of July, or the set of time intervals corresponding to hour 5 from every day of the year. We define below three metrics which depend explicitly on the choice of aggregation period \( T \).

**Inverse Call Duration Metric**

The first metric we consider is a metric on the set of the 1666 cell phone towers, or a subset thereof, which...
is determined solely by the call volumes between towers. For a given choice of aggregation period $T$, and choice of a subset of the towers, let $C(T)_{ij}$ be the total duration of calls made between tower $i$ and tower $j$ during the aggregation period $T$. Note that our definition includes contributions from both $i$ to $j$, and from $j$ to $i$, so that $C(T)_{ij} = C(T)_{ji}$. These quantities were obtained from the call data provided in SET1V. Using these quantities, we then define a complete weighted adjacency matrix $w(T)_{ij}$ by:

$$w(T)_{ij} = \begin{cases} C(T)_{ij}^{-1} & \text{if } C(T)_{ij} \neq 0 \\ 1 & \text{else} \end{cases}$$

Finally, we take the complete weighted adjacency matrix obtained above, and run the Floyd-Warshall Algorithm on it to obtain the Inverse Call Duration Metric (ICD Metric) for the aggregation period $T$.

**Gravity Model Call Metric**

Given an aggregation period $T$ and a choice of subset of the towers, let $C(T)_{ij}$ be as above, and let $C(T)_i$ be the total duration of calls made to or from tower $i$ during the aggregation period $T$. Let $d_{ij}$ be the geographical distance between the towers $i$ and $j$. This distance matrix was computed using the (altered, and thus slightly inaccurate) latitudes and longitudes provided, via the length of a spherical geodesic. Consider the following model, which is often described as a gravity model:

$$\log(C(T)_{ij}) = b + \log\left(\frac{C(T)_i C(T)_j}{d_{ij}^a}\right)$$

Here $a$ and $b$ are parameters of the model. Since the call data will of course not fit the model exactly for any particular choice of $a$, $b$ and $T$, we performed a linear least squares fit. In doing so, we fit only the subset of the data aggregated over the period $T$ for which $C(T)_{ij}$, $C(T)_i$, $C(T)_j$, and $d_{ij}$ were nonzero. The questions of how well the data fits the model, and whether a linear least squares fit is the best methodology to be using, were both considered, but will not be addressed here.

We then constructed the complete weighted adjacency graph $w_{ij} = d_{ij}^a$, where $a_0$ is the parameter value arising from a linear least squares fit. Finally, we ran the Floyd-Warshall algorithm on this adjacency matrix to produce a metric, which we will call the Gravity Model Call metric (GMC metric).

**Dot Product Call Metric**

Given an aggregation period $T$ one can construct for each tower, with label $i$, the vector:

$$v_i = \begin{pmatrix} C(T)_{i1} \\ \vdots \\ C(T)_{i1666} \end{pmatrix}$$

Then define the Dot Product Call Metric (DPC metric) $r$ to be the metric

$$r_{ij} = \arccos\left(\frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}\right)$$

It is easy to check that the angle between two vectors in a Euclidean space does indeed define a metric space.

**Illustrative Example**

Visualization and verification of these techniques is easily done by considering the geometrical, and in this case geographical, structure of the most persistent generators of the first homology group $H_1$. One can consider the most trivial example of what such persistent generators might mean by using the metric space arising from the physical geographical distance between any two cell towers. The persistent generators of the first homology group arising from a Vietoris-Rips complex correspond in general to long-lived cycles in the set of towers for which certain constituents are not directly connected over a long range of the filtration parameter. In the case of using the geographical metric arising from a Vietoris-Rips complex on the set of towers, the persistent generators correspond to the largest geographical voids of cell phone towers.

To demonstrate the technique, we first inspect Figure 1 which maps the approximate tower locations within Senegal.

![FIG. 1: The approximate location of all 1666 towers mapped onto Senegal where each tower is represented by a light-blue pin.](image)

Visually, it is clear that a large void in the towers exists in the northeastern region of Senegal as a result of the
Ferlo Nord Wildlife Reserve. In order to more clearly see this void or cycle, we can consider subsets of these towers. In particular by choosing 800 towers via the sequential maxmin approach and then selecting the 750 densest points, we arrive at the following barcodes in Figure 2.

FIG. 2: Rips barcodes for the simplicial complexes created via the geographical distance metric.

Now note that one generator of the first homology group $H_1$ persists over the entire range of filtration values. If we then map the towers and overlay the generator of this persistent $H_1$ homology group, we find the following result in Figure 3. Additionally, the relative dominance of this persistent $H_1$ homology group generator is $0.7321$, indicating the significance of the feature.

FIG. 3: The approximate location of the 750 towers selected via sequential maxmin and then subsetting based on density as well as the most persistent $H_1$ homology group generator in dark blue.

This example, though trivial, demonstrates a successful application of topological data analysis as we have identified the large void in towers resulting from Senegal’s extensive Wildlife Reserve. As we move to more useful metric spaces, the ability for simple visualization is lost, but the principles of the analysis remain the same.

RESULTS

We aim to use the formalism of persistent homology applied to the three particular metrics on the set of towers described above to address the thematic development issues outlined by the D4D Development team. Our hope is that this novel method of telecom data analysis will allow certain objectives to be accomplished in a relatively easy way which will provide new insight into the structure of the data.

Dakar Motorway

The first issue we intend to address is the facilitation of useful development of transportation and infrastructure systems in Senegal. The analytical framework we work with seems well suited to addressing the problem of identifying regions which would benefit the most from the development of new local transport methods, as well as identifying the effects of installment of new local transport methods which have already occurred. In particular, we considered the opening of the Pakine to Diamniadio section of the Dakar Motorway in 2013, with the goal of identifying signatures and impact of the new section of road via local changes in the persistent homology in the regions most directly affected.

For the three metrics described above, we considered the persistent homology of the Vietoris-Rips complex arising from the aggregation periods $T_1$ and $T_2$ corresponding to the entire months of July and August in 2013, respectively. Additionally, we limited the points in our metric space to be among the cell phone towers with labels between 1 and 500, which correspond to the western region of Senegal potentially impacted by the Dakar Motorway opening. These choices of aggregation periods and subset were made with the goal in mind of identifying a substantial change in the generators of the first homology between these months, indicating a signature of the construction of the section of the Dakar motorway between Pikine and Diamniadio, which opened on August 1st of 2013.

In Figures 4, 5, and 6, barcode plots are displayed for the 0th and 1st dimensional homology for Vietoris-Rips complexes arising from the Inverse Call Duration metric, the Gravity Model Call metric and the Dot Product Call metric respectively. For each of the barcode plots shown, we considered the relative dominance of the three most persistent generators of the one dimensional homology $H_1$. These relative dominance measures are listed in Table I.

Note that the relative dominance varies widely over the three metrics. While there is no precise lower bound on the relative dominance necessary to deem that a generator is indicative of a genuinely persistent topological structure, it is apparent that the Inverse Call Duration
FIG. 4: Rips Barcodes for the simplicial complexes created from the DPC metric for the aggregation periods $T_1$ and $T_2$ respectively.

<table>
<thead>
<tr>
<th>GCM $T_1$</th>
<th>GCM $T_2$</th>
<th>DPC $T_1$</th>
<th>DPC $T_2$</th>
<th>ICD $T_1$</th>
<th>ICD $T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2308</td>
<td>0.2895</td>
<td>0.2632</td>
<td>0.3846</td>
<td>0.4615</td>
<td>0.5714</td>
</tr>
<tr>
<td>0.1795</td>
<td>0.1842</td>
<td>0.2368</td>
<td>0.3333</td>
<td>0.4359</td>
<td>0.4571</td>
</tr>
<tr>
<td>0.1538</td>
<td>0.1579</td>
<td>0.2368</td>
<td>0.2821</td>
<td>0.4103</td>
<td>0.3714</td>
</tr>
</tbody>
</table>

TABLE I: The relative dominance for Rips of the three most persistent generators of $H_1$ for the three metrics over the aggregation periods $T_1$ and $T_2$.

metric and the Dot Product Call metric give rise to simplicial complexes whose most persistent generators are more highly dominant than those of the Gravity Call Model metric.

In addition to the fact that the Gravity Call Model metric has a low relative dominance for its most persistent generators, each of the three metrics yields barcodes which look similar in structure for the aggregation periods of July and August. In the case of the GCM metric, this similarity can be attributed to the fact that the least squares parameter $a_0$, which is the only dynamic aspect of that model across different data aggregation periods, varied only slightly between the two months. Thus in the context of that metric, the actual underlying topology, and not just its barcode, underwent very little change between July and August. This suggests that for the particular choice of subset and aggregation periods made here, the GCM metric is not a sensitive probe for changes in structure of the data. While our initial hope was that there might be some qualitatively obvious change in the structure of the 1st dimensional homology between July and August appearing readily in at least one of the barcodes, their appearance indicates that detection of any significant change will necessitate sensitive analysis.

In order to further analyze whether the topological structure of the data encoded in these metrics underwent a significant change between July and August of 2013, we consider geometrical realizations of the most persistent 1 dimensional generators in each barcode. This has the double benefit of providing more information than is present in the barcode plots, which will aid in the potential recognition of a topology change, as well as giving insight into what such a change indicates in terms of the specific affects on local call activity near Dakar. In Figures 7, 8, 9 representative cycles of the three generators with the highest relative dominance, or “the most persistent” generators, are overlayed for the two aggregation periods on a map with the locations of the actual cell tower locations. The clustering of the towers in these pictures was made with respect to an arbitrary
filtration parameter cutoff value merely to plot a majority of the towers, and is irrelevant to the location of these representative cycles.

In Figure 9, one can see that all three of the most persistent generators can be represented by cycles in or near Dakar, and furthermore that these representative cycles remain virtually unchanged between July and August. This further validates the claim that the GMC metric and the underlying simplicial complex to which it gives rise remain unchanged between July and August, and provides more evidence than the barcode alone. For the ICD and DPC metrics, one can likewise see in Figures 7 and 8 that the most persistent generators can be represented by cycles which enclose the nearby local regions in July and August. Thus even considering the geographical representatives of the generators of the first homology group, it appears that none of our metrics were able to capture a significant topology change due to the introduction of a section of the Dakar Motorway. We conclude from these results that there is no detectable change in the topological structure of the call duration data with respect to the models here considered. This null result suggests that either the introduction of the Motorway had no real affect on call traffic, and social activity coupled to this traffic, or if there was an effect that its detection would require a different measure of call activity and distance between towers.

Further Analysis of ICD Metric

As we saw in the geographical metric example previously, using the entire point cloud can often obscure the
underlying topology within a higher dimensional space. In order to determine the underlying topology and reduce noise, we may instead choose a representative subset of the points. From our prior analysis, the ICD metric led to the highest relative dominance of generators for the $H_1$ homology group; to explore the data further and attempt to understand if there is a true latent topological structure, we choose 350 of the 500 towers via the sequential maxmin methodology and then further reduce this by using a $k = 1$ core subset in which we choose 200 of the densest towers. The goal is to reduce the noise sufficiently that any hidden structure will become evident.

After choosing these subsets of the original 500 towers, the barcodes for the 0th and 1st dimensional homology for the Vietoris-Rips complexes are displayed in Figure 10. Again, as in the case of the full point cloud, a topological change between period $T_1$ and period $T_2$ is not clearly evident, however, the relative dominance of the top three $H_1$ generators decreased as seen in Table II, potentially hinting at evidence of a topology change, however, without further analysis such as methodologically sweeping parameters, it still does not represent a detectable change in the underlying topology.

These three most persistent $H_1$ generators are mapped for both time periods, $T_1$ and $T_2$, in Figure 12. As an additional check to the existence of a topology change between times $T_1$ and $T_2$, we also construct the Lazy Witness complex for each choice in metric space, and parameter we scan over. See Figure 11 for an example of the Barcode plot we produce from a choice in parameters where: $k = 100$ (neighbourhood parameter), $P = 40$ (percentage of points from the original metric space that are used to produce the new metric space), $\nu = 0$ (used in the construction of $LW(Z, L, \epsilon, \nu)$), the Landmark Set $L$ is produced randomly. Similarly, given our choices in parameters, we see no statistically significant topology change for the Lazy Witness complex.

**Communities via 0th Homology and Comparison with Modularity**

Using Vietoris-Rips to construct simplices, which is equivalent to single-linkage clustering, we may identify connected components at any desired filtration value. With our three metrics, we compare the clustering results to that of another approach based on modularity. Modularity $Q$ is a quality index for decompositions of a
FIG. 11: Lazy Witness Barcodes for the simplicial complexes created from the ICD metric for the aggregation periods $T_1$ (with $\max(\delta_R) = 0.02$) and $T_2$ (with $\max(\delta_R) = 0.08$) respectively. $\nu = 0$, $P = 40$, $k = 100$; $L \subset Z$ produced randomly.

FIG. 12: Representatives of the three most persistent generators of $H_1$ for the ICD metric for aggregation periods $T_1$ and $T_2$, on a map with the tower locations. Here we have taken a 200 tower subset via sequential maxmin and core subsetting.

network into communities, which measures the fraction of edges that fall within the given communities minus the expected fraction if those edges were distributed at random, and has a value between -1 and 1. For a particular weighted graph $G = (V,E)$, with edge weights $A_{ij}$, the modularity of a decomposition into communities $V = \bigcup_i C_i$, $C_i \cap C_j = \emptyset \ \forall i \neq j$, may be defined as

$$Q(C) = \sum_{i,j} \left( \frac{A_{ij}}{A} - \frac{k_i k_j}{A^2} \right) \delta(c_i,c_j) \quad (2)$$

where $A = \sum_{i,j} A_{ij}$, $k_i = \sum_j A_{ij}$, and $c_i$ is the community label of vertex $i$. The optimal community decomposition for a given weighted graph is defined to be that which maximizes the modularity $Q$.

In practice, for any reasonably sized network, it is not computationally practical to compute an exactly optimal decomposition into communities. We therefore make use of the hierarchical algorithm, and accompanying software, introduced by Blondel et. al. [11]. We find that it is very effective in finding near optimal community decompositions for mobile phone networks of this sort [12].

Employing this algorithm, we compute communities for two weighted mobile phone networks, one weighted by the total volume of voice communication between each pair of towers in July, and the other weighted by the total voice volume in August. The results are depicted in Figures 13 and 14. We note in particular that, as above, the Dakar motorway does not have any appreciable effect on the modularity communities.

To compare these with communities identified as generators of $H_0$, we show in Figures 15 and 16 a clustering resulting from the ICD and DPC metrics. We note that, interestingly enough, the zeroth homology generators arising from the ICD metric are not terribly dissimilar from the communities detected by the modularity optimizing algorithm. This suggests that the ICD metric may be capturing similar information as the edge weights which go into the modularity optimization, and that therefore the higher dimensional homology generators arising from that metric may be exploring richer structures present in the graph weighted by call volume, than is available from the modularity analysis alone.

CONCLUSIONS

We chose both to construct a reduced metric space from core subsetting, and to analyze the towers most likely to be affected by the introduction of the Dakar motorway. By comparing the barcode plots and the relative dominances for the Rips complex for the months
FIG. 13: Map of the ten communities within Senegal (above) and within Dakar (below) detected via modularity, using total voice communication volume in July. Communities with only a single tower are ignored.

of July and August, we determine no statistically significant changes in the homology from month to month, given these choices. We also detect no significant topological change in the Lazy Witness complex construction given these same choices.

In both the Vietoris-Rips and Lazy Witness constructions, we scan over various parameters: the choice in metric space (either ICD, GMC, or DPC); the percentage of points that are used in the core subset (“P”); the density neighbourhood parameter (“k”); and the choice in the determination of the Landmark Set $L$ (either random or maxin).

Thus, the null result suggests that either the introduction of the Motorway had no real affect on call traffic, and social activity coupled to this traffic, or if there was an effect then its detection would require a different measure of call activity and distance between towers.

In a future analysis, large scale averaging relative dominance over multiple trials will improve any claim to the statistical relevance of a topological feature. In the current analysis, we do no large scale averaging, and instead average over just a few trials.

With regard to modularity-based community detection, we find that, for the ICD metric, the $H_0$ genera-

FIG. 14: Map of the ten communities within Senegal (above) and within Dakar (below) detected via modularity, using total voice communication volume in August. Communities with only a single tower are ignored.

FIG. 15: Map of the nine largest detected communities within Senegal and within Dakar via the most persistent $H_0$ groups found with the ICD metric.

tors provide a qualitatively similar decomposition. Given that constructing simplices via Vietoris-Rips and building the clusters is mathematically equivalent to single-linkage clustering, a rudimentary approach, it is promising that the results can mimic that of the modularity communities.

Acknowledgements We would like to thank the developers of javaPlex, and specifically to Henry Adams for
FIG. 16: Map of the nine largest detected communities within Senegal and within Dakar via the most persistent $H_0$ groups found with the DPC metric.

all the insight he provided into the world of persistent homology.

REFERENCES

Analyzing Internal Migration Patterns using Aggregated Cell Phone Data

Tina Yu

1 Introduction

Internal migration, both permanent and seasonal, is very common in most developing countries. For example, Aker et al [4] observed that over 45 percent of households in Niger had at least one seasonal migrant while Skeldon [8] noted that roughly 30 percent of Indians reported permanently living in a place other than the place of survey enumeration. Internal migration within a country can have a profound impact on regional labour markets [3], can affect levels of urban and rural inequality [2], and also can provide vectors for disease transmission [1]. Although governments of developing countries seek to understand the causes of internal migration to regulate population movement for social and economical improvement, the internal migration information used are either sparse or unreliable [5]. In this study, we are interested in the internal migration patterns in Senegal during the rainy seasons. Using the cell phone data collected during the flooding of August in 2013, we modeled the distribution of the phone calls made in the country. The distribution patterns may be perceived as the internal migration trend emerged to adapt to the rainy seasons.

The report is organized as follows. We first explain our approaches in Section 2 and then present the data in Section 3. In Section 4, we describe the computer experiments and in Section 5 we discuss our results. Finally, Section 6 concludes the report and outlines our future works.

2 Assumptions and Approaches

The commonly used methods to study internal migration are government censuses and household surveys. However, due to the infrequency of the survey conducted and the bias of the surveyed samples, the results do not reflect the internal migration well. With the popularity of the cell phones in the developing countries\(^1\), cell phone data are reasonable choices to study population migration. The D4D challenge consortium has made such data available for our study. The data set contains aggregated antenna-to-antenna traffic information for 1666 antenna on an hourly basis from January 1 to December 31 of 2013. They are in the following format:

<table>
<thead>
<tr>
<th>timestamp</th>
<th>outgoing_site_id</th>
<th>incoming_site_id</th>
<th>number_of_calls</th>
<th>total_call_duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-04-01 00</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>138</td>
</tr>
<tr>
<td>2013-04-01 00</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>136</td>
</tr>
</tbody>
</table>

Assume the number of outgoing phone calls made from a particular antenna site represents the number of people living around the geological location of that antenna, the distribution of the outgoing phone calls made from each antenna can be used as a proxy of the population distribution. Hence, by tracing the daily changes of the distribution, we can get an idea about population migration. Moreover, in addition to tracing the migration patterns, we are also interested in capturing that pattern in a model to be used for future prediction of population migration. Hence the project consists of two parts:

- Processing the data and present a sequence of daily phone calls distributions to trace the population migration.
- Using the processed time series data to train a model that makes one-time step prediction of the phone call distribution at each antenna site.

The data preparation process is explained in the following Section.

\(^1\)There is a roughly 68 percent penetration in the developing world (need a citation).
3 Data Preparation

As the first attempt using aggregated cell phone data to model internal migration patterns, we only used data from August of 2013 in this study. This is because there was a flooding took place during that period of time and we are interested in modeling the migration patterns during rainy seasons.

We examined the data set and found that among the 1666 sites, 106 of them were down for the entire month, where both the number of outgoing and incoming calls were zero. Additionally, there were a number of antenna sites which had zero number of outgoing and incoming calls during part of the month. Although it is possible that the antenna went down due to the flooding, it is also possible that no phone calls were made from that antenna during those days. Since we do not have information about the actual well-being of those antenna during that period of time, we only removed the antenna that were down for the entire month from the data set. This reduced the number of sites to 1560 which we calculated their daily outgoing phone call distributions. The results will be presented and discussed in Section 5.

To train a model that captures the migration patterns of the daily phone call distributions, we have to prepare the data further. First, to prevent model overfitting to the training data, we split the 1560 31-day time series into training and testing data, where training data are used to build the model while the testing data are used to evaluate the performance of the model. Ideally, the two data sets should have similar characteristics so that a model trained under one data set would work well on the other data set. But how do we measure the similarity between two 31-day time series from two antenna sites?

In this work, we used $R^2$, which is the square of the Pearson product moment correlation coefficient, to measure the similarity of two time series $x$ and $y$:

$$R^2 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 - \sum (y - \bar{y})^2}}$$

where $\bar{x}$ is the mean of the time-series $x$ and $\bar{y}$ is the mean of the time-series $y$. $R^2$ has values between 0 and 1. The higher the $R^2$ is, the more similar the $x$ and $y$ are.

Once the $R^2$ is calculated for each pair of the 1560 time series, we used a $R^2$ threshold as the selection criterion. We first selected the time series which have the largest number of ($R^2 >$ threshold) and added it into the training set. Next, those sites whose $R^2$ with the selected training site is greater than the threshold are added to the testing set. We repeated this process to select training and testing data until no time series have $R^2 >$ threshold in the original data set.

We experimented with different $R^2$ threshold and the splitting results are given in the following table:

<table>
<thead>
<tr>
<th>$R^2$ threshold</th>
<th>Training Data</th>
<th>Testing Data</th>
<th>Unselected</th>
<th>Down</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>104</td>
<td>393</td>
<td>1063</td>
<td>106</td>
<td>1666</td>
</tr>
<tr>
<td>0.8</td>
<td>180</td>
<td>580</td>
<td>800</td>
<td>106</td>
<td>1666</td>
</tr>
<tr>
<td>0.7</td>
<td>236</td>
<td>813</td>
<td>511</td>
<td>106</td>
<td>1666</td>
</tr>
<tr>
<td>0.6</td>
<td>256</td>
<td>1031</td>
<td>273</td>
<td>106</td>
<td>1666</td>
</tr>
<tr>
<td>0.5</td>
<td>233</td>
<td>1218</td>
<td>109</td>
<td>106</td>
<td>1666</td>
</tr>
<tr>
<td>0.45</td>
<td>209</td>
<td>1295</td>
<td>56</td>
<td>106</td>
<td>1666</td>
</tr>
</tbody>
</table>

We decided to use $R^2 = 0.45$ as the threshold to split the data so most of the data would be selected for either training or testing. We added the unselected time series (which has $R^2 < 0.45$) to the training set. Hence the number of training time series is 265 and the number of testing time series is 1295.

The second step of preparation the data for model training is to decide which information (features) to use to train the prediction model. We decided to use information from the previous 3 days ($t - 1$, $t - 2$, $t - 3$) to make one time step ($t$) prediction:

• $\text{outvol}$: Outgoing call volume
• **outlen**: Outgoing call length
• **invol**: Incoming call volume
• **inlen**: Incoming call length

With such, the data are prepared such that each row consists of 13 columns: 4 features from each of the 3 days and the target call volume. Each site is a 30-day time-series, which is made into a 28 data entries (the first 3 days can not be used). The number of entries in the training set is 7,420 and the number of entries in the test set is 36,260.

Once the data are properly prepared, we can proceed with modeling training which is explained in the following Section.

### 4 Modele Training

The machine learning method we used is Genetic Programming (GP)[7], which applies Genetic Algorithms (GA)[6] to perform models training based on the given data. The model in this case is an if-then-else rule that uses information from the previous 3 time steps (days) to make one-time step (day) prediction.

The rules can contain the following operators: `and`, `or`, `not`, `+`, `-`, `*`, `if-then-else`. The objective function is to minimize mean error:

\[
\min f = \frac{\sum_{i=1}^{7420} \text{abs}(\text{outvol}_i - \text{predicted}_i)}{7420}
\]

where **outvol**<sub>i</sub> is the correct outgoing call volume of data **i** and **predicted**<sub>i</sub> is that predicted by the prediction model. We made one GP run and the trained best model is evaluated on the testing data.

We present the model and its performance in the following Section.

### 5 Results and Analysis

We first present the outgoing call volume distribution of the 1560 sites on 4 representative days (August 4, 18, 25 and 31) in Figures 1, 2, 3, 4. As shown, most calls were made from the city of Dakar, indicating its population density is higher than the rest of the country. There are also more calls made from the Northern border to Mauritania. Additionally, there is an increased number of phone calls made around the border region. One possible interpretation is that people migrated to the other country to avoid the flooding. We will need to verify that with the actual record.

Aside from that, there is no significant difference among the call volumes distributions on the 4 respective days. This suggests that the migration might be an independent random walk without any pattern, even under the flooding. We therefore build a random walk model that takes the average of the time-series as the predicted call volume at each site. Its performances on training and testing data are given in Table 3 to be compared with the GP trained model.

The GP trained model used information from the previous 3 days to predict the outgoing call volume distribution on the current day(t):

```plaintext
if (invol(t-1)==outvol(t-1))
    temp=invol(t-3)*2
else
    temp = invol(t-2);
if( temp < (outvol(t-2)-outvol(t-1)-invol(t-2)) )
    outvol(t)=(outvol(t-3)-(invol(t-2)-invol(t-1)) × outvol(t-3)*35.26
else
    outvol(t) = outvol(t-1);
```
The model can be roughly interpreted as follow: checking the call volume information from the three previous days using some criteria. If the criteria are satisfied, use previous 3 day’s call volume distribution to predict this day’s call volume distribution. Otherwise, simply use the previous day’s call volume as this day’s call volume prediction. Note that the call length information is not used for prediction, indicating it does not have impact on the predicted call volume.

The performance of the model on training and testing data are given in Table 3. Moreover, the simple one-time-step model, which uses the previous day’s call volume distribution to predict the current day’s outgoing call volume distribution, is also evaluated:

\[ \text{outvol}(t) = \text{outvol}(t-1); \]

Table 3 shows that the one-time-step model does not do better than the random walk model, which suggests that there is no dependency of the call volume distribution time series. However, when using information from the 3 previous days, the GP trained model gives the smallest mean errors on both the training and testing data. This indicates that there is long-term dependency in the time series. It is possible that even better model can be trained when information of longer term is provided. We will test that theory in our future work.

The trained model gives similar mean errors for both training and testing data, indicating the model is robust for unseen data. This indicates that the data splitting scheme using \( R^2 \) is appropriate for this
Table 3: Mean error of the Three Different Models.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Training Data</th>
<th>Testing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Walk</td>
<td>0.8069847864940344</td>
<td>0.9059290074068214</td>
</tr>
<tr>
<td>One-Time-Step</td>
<td>0.8778886</td>
<td>0.96313363</td>
</tr>
<tr>
<td>Three-Time-Step</td>
<td>0.7388</td>
<td>0.8226</td>
</tr>
</tbody>
</table>

data set.

6 Conclusions and Future Works

We have developed a methodology to investigate the internal migration patterns using the cell phone data. We have tested the method on the raining season data and the preliminary results are encouraging. First, we found that the devised method builds a model that can be interpreted as a one-time-step model with extra conditions. Second, the model gives better performance than the random walk model and the one-time-step model, indicating that there is longer term dependency of the call volume distribution time series. Third, we identify that there is increase in call volume close to the Northern border region during the raining season. We will verify its accuracy and investigate its cause in our future work.

We are also interested in those antenna sites that have zero incoming and outgoing phone calls during part of the month. What has caused such behavior? Is it flooding related? This will also be investigated in our future work.

References

1. Introduction

Social connections between individuals play an important role in numerous aspects of life in Sénégal, including affiliation with the Sufi brotherhoods, the economic well-being of the country, public health, and the ability to respond to natural disasters, to name a few. In the previous Data for Development (D4D) Challenge, focused on Côte d’Ivoire, the organizers provided four sets of data. Three were analogous to the three sets provided by the D4D Sénégal Challenge [1]:

Set 1, which details the mobile communications between network antennae,

Set 2, which provides mobility data for individual users, at the spatial resolution of individual antennae, during two week time periods (fortnights), and

Set 3, which details mobility information for individual users for the entire time of the challenge, at the level of arrondissements.

Additionally a fourth data set was provided [2]:

Set 4, which detailed the social connections of individual users, including information on friends of friends.

As far as we are aware, not a single presentation at the D4D Challenge day at NetMob 2013, nor even during the rest of NetMob 2013, covered analysis of data from Set 4. Perhaps for this reason, no analog of Set 4 was provided for the D4D Sénégal Challenge.

We feel that it is unfortunate that this sort of direct social network information was excluded from both D4D challenges, whether due to technical glitches or privacy concerns, and therefore seek to reproduce similar information, for Sénégal, from the data provided in Sets 1, 2 and 3. This analysis is also relevant to the question of mobile phone data anonymization, and highlights challenges involved in any attempts to generate useful synthetic data.

2. Combining the mobility data

Data sets 2 and 3 provide mobility data, for different time periods, at different resolutions, each with a different set of user IDs. Set 3 provides data for 146352 users over the entire
year of 2013, while Set 2 provides data for approximately 300,000 users during each of 25 non-overlapping fortnights during the year. By way of comparison, the D4D network provider, Sonatel, had 7.362 million subscribers in Sénégal in 2013 [3]. Neglecting details such as individual subscribers being excluded from selection in the mobility datasets (due to insufficient activity, for example) the probability that a given subscriber will appear in the Set 3 mobility data is roughly $146.352/7.362 \times 10^6 \approx 0.01988$, i.e., 2%. Thus, each of the 300,000 users in a given fortnight of Set 2 has about a 2% chance of appearing in Set 3, assuming these samples were taken uniformly at random and independently. From this one can easily estimate that the number of such users is approximately $300,000 \times 0.02 = 6000$. One can continue, asking about the probability that a single user from Set 3 appears in at least two specific fortnights of Set 2, which, again assuming uniform, independent sampling, should be about $(6000/150,000)^2 = 0.0016$. We can therefore expect approximately $0.0016 \times 150,000 = 240$ such users.

To identify the users who appear in both a fortnightly data set and the annual data set, we compared the list of calls of each fortnightly user, each cell tower ID being replaced by the arrondissement in which that tower lies, with the list of calls of each annual user, restricted to that fortnight. In no case did a fortnightly user match more than one annual user, supporting the reasonable inference that call lists are good “fingerprints”. The resulting numbers of matches are listed in Table 1. Notice that the average number per fortnight is about twice what we estimated above, suggesting that the total number of users who satisfy the sufficient activity criteria is only about half of Sonatel’s subscribers, namely about 3.75 million.

As expected, some Set 3 users match with users in multiple fortnights. In fact, of all the Set 3 users, only 29,823 do not match at least once, 37,251 match in 1 fortnight, 40,807 match in 2, 28,930 match in 3, 14,848 match in 4, 5,803 match in 5, 1,941 match in 6, 474 match in 7, 109 match in 8, 13 match in 9, and 1 matches in 10 fortnights (i.e., for this user the Set 2 data provides 20 weeks of call data at the cell tower level!). As we noted in the previous section, such a Poisson distribution should exist in any realistic synthetic data, which was the original reason we computed these matches.

3. Deducing relationships between users

It is useful to deduce identifications among the user IDs of a given subscriber in the various datasets, but what we are really aiming for is information about the relationships among the subscribers. Can we also glean some information about this, from the data provided? It turns out that the answer is “yes”—some of the ‘Set 4 information’, about networks of individuals, is buried in the data of Sets 1-3. Here we present an approach to begin to extract this information.

The central idea comes from the notion of a user saturating a antenna hour. We say that a user saturates an antenna hour if the total number of records for a user ID from Set 2, for a given hour during the year, at a given antenna, matches the total number of sms messages handled by that antenna during that hour. (We restrict to sms communications
<table>
<thead>
<tr>
<th>dates</th>
<th>no. users</th>
<th>no. annual users</th>
<th>no. matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 Jan–20 Jan</td>
<td>319 508</td>
<td>145 199</td>
<td>12 407</td>
</tr>
<tr>
<td>21 Jan–3 Feb</td>
<td>319 543</td>
<td>145 923</td>
<td>12 394</td>
</tr>
<tr>
<td>4 Feb–17 Feb</td>
<td>319 204</td>
<td>146 021</td>
<td>12 362</td>
</tr>
<tr>
<td>18 Feb–3 Mar</td>
<td>319 528</td>
<td>146 031</td>
<td>12 616</td>
</tr>
<tr>
<td>4 Mar–17 Mar</td>
<td>319 466</td>
<td>146 088</td>
<td>12 210</td>
</tr>
<tr>
<td>18 Mar–31 Mar</td>
<td>319 488</td>
<td>146 080</td>
<td>12 470</td>
</tr>
<tr>
<td>1 Apr–14 Apr</td>
<td>319 507</td>
<td>146 097</td>
<td>12 306</td>
</tr>
<tr>
<td>15 Apr–28 Apr</td>
<td>319 366</td>
<td>146 040</td>
<td>12 616</td>
</tr>
<tr>
<td>29 Apr–12 May</td>
<td>319 389</td>
<td>146 050</td>
<td>12 294</td>
</tr>
<tr>
<td>13 May–26 May</td>
<td>319 398</td>
<td>146 076</td>
<td>12 549</td>
</tr>
<tr>
<td>27 May–9 Jun</td>
<td>319 500</td>
<td>146 068</td>
<td>12 206</td>
</tr>
<tr>
<td>10 Jun–23 Jun</td>
<td>319 473</td>
<td>146 057</td>
<td>12 712</td>
</tr>
<tr>
<td>24 Jun–7 Jul</td>
<td>319 471</td>
<td>146 017</td>
<td>12 460</td>
</tr>
<tr>
<td>8 Jul–21 Jul</td>
<td>319 374</td>
<td>145 977</td>
<td>12 539</td>
</tr>
<tr>
<td>22 Jul–4 Aug</td>
<td>319 511</td>
<td>145 975</td>
<td>12 554</td>
</tr>
<tr>
<td>5 Aug–18 Aug</td>
<td>319 422</td>
<td>146 016</td>
<td>12 462</td>
</tr>
<tr>
<td>19 Aug–1 Sep</td>
<td>319 453</td>
<td>145 978</td>
<td>12 587</td>
</tr>
<tr>
<td>2 Sep–15 Sep</td>
<td>319 395</td>
<td>145 972</td>
<td>12 730</td>
</tr>
<tr>
<td>16 Sep–29 Sep</td>
<td>319 422</td>
<td>145 942</td>
<td>12 865</td>
</tr>
<tr>
<td>30 Sep–13 Oct</td>
<td>319 478</td>
<td>145 781</td>
<td>12 452</td>
</tr>
<tr>
<td>14 Oct–27 Oct</td>
<td>319 339</td>
<td>145 844</td>
<td>12 196</td>
</tr>
<tr>
<td>28 Oct–10 Nov</td>
<td>319 456</td>
<td>145 964</td>
<td>12 266</td>
</tr>
<tr>
<td>11 Nov–24 Nov</td>
<td>319 456</td>
<td>145 938</td>
<td>12 089</td>
</tr>
<tr>
<td>25 Nov–8 Dec</td>
<td>319 467</td>
<td>145 953</td>
<td>11 936</td>
</tr>
<tr>
<td>9 Dec–22 Dec</td>
<td>319 510</td>
<td>145 731</td>
<td>11 747</td>
</tr>
</tbody>
</table>

**Table 1.** Numbers of matching users for each fortnight.

for convenience. A more thorough analysis would make use of the voice communications as well.)

To search for such saturated antenna-hours, we begin by computing the total sms usage of each antenna during each hour of the year, from Set 1. Equipped with such a database, we
can then sum the number of records in the data of Set 2 for a given userid and antenna-hour, and ask if that user saturates that antenna-hour. In this way we can build up a database of saturated antenna-hours, along with the responsible user IDs.

Equipped with a such database of saturated antenna-hours, with the associated user IDs, we can now look through Set 1, to identify entries for which both ends of the communication involve a saturated antenna. Each such entry indicates a direct sms communication between two individual user IDs. From these connections, we generate a directed graph of a social network of Sonatel users in Sénégal, where the direction indicates who sent an sms message to whom. We eliminate user IDs for which we have no connection information, and also pairs of user IDs which only communicate with each other. The resulting social networks are depicted in Figures 1 and 2, each visualized using a different algorithm within the Graphviz package [4].

![Figure 1. Annual social network, visualized by Graphviz's twopi.](image)

4. Extensions

We’ve shown that, since they are not synthetic, i.e., they are compiled from the same set of call detail records (CDRs), the data in Sets 1, 2 and 3 of the D4D Sénégal Challenge can be combined to give a ‘glimpse’ of the social network among Sonatel’s users. To deduce
additional relationships we would first also use the non-sms calls in Set 1, as noted above. Furthermore, we could relax “saturated hours” to “almost saturated hours”, and infer the existence of additional relationships, albeit with probabilities smaller than 1. For the kind of structural analysis of social networks in which we are also interested [5], this is more relevant than knowing exactly who is communicating with whom, and has the advantage of minimizing any possible de-anonymization of the users.

Figure 2. Annual social network, visualized by Graphviz’s circo.
References


