Data for Development Challenge Senegal

Book of Abstracts: Scientific Papers

At NetMob 2015

Organized by

Universidad Carlos III de Madrid

Sponsored by

RealImpact ANALYTICS

www.d4d.orange.com / Tweeter : @O4Dev
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 4
   Pedro Zufrina
2. [A02] Genesis of millet prices in Senegal: the role of production, markets and their failures 24
   Damien Jacques

Session 2: Energy

3. [E01] Using mobile phone data for rural electrification planning in developing countries 30
   Markus Schläpfer

Session 3: Health

4. [H01] Quantifying the effect of movement associated with holidays on malaria prevalence using cell phone data 39
   Sveta Milusheva
5. [H02] Uncovering the impact of human mobility on schistosomiasis via mobile phone data 71
   Marino Gatto
6. [H11] Mobile data as public health decision enabler: a case study of cardiac and neurological emergencies 97
   Edward Mutafungwa
7. [H12] Modeling Ebola virus diffusion in Senegal using mobile phone datasets and agent-based simulation 103
   Jonathan Leidig

Session 4: National Statistics

8. [N02] Spatial structure and efficiency of commuting in Senegalese cities 110
   Thomas Louail
9. [N03] Construction of socio-demographic indicators with digital breadcrumbs 122
   Timo Schmid
10. [N07] Virtual networks and poverty analysis in Senegal 132
    Neeti Pokhriyal, Wen Dong, Venu Govindaraju

Session 5: Transports / Urbanism

11. [T02] Deviations from the norm: detecting regularities and anomalies in human mobility patterns 144
    Gijs Joost Brouwer
12. [T03] National and regional road network optimization for Senegal using mobile phone data 153
    Erik De Romph
13. [T04] Building workers' travel demand models based on mobile phone data 180
    Liu Feng
14. [T05] Urban road construction and human mobility: evidence from Dakar 200
    Amadou Sy
15. [T06] Travel demand analysis with differentially private releases 214
    David Gundlegard
16. [T10] The municipality and the territory: two scales of understanding the city 228
    Jerome Chenal
17. [T15] Using mobile phone data for spatial planning simulation and optimization technologies 251
    Serigne Gueye
18. [T16] Where do we develop? Discovering regions for urban investment in Senegal 270
    Derek Doran

Session 6: Others

19. [O02] On the anonymizability of mobile traffic datasets 281
    Marco Gramaglia
20. [O05] Data for Development reloaded: visual matrix techniques for the exploration and analysis of massive mobile phone data 289
    Stef Van Den Elzen

A01

Mobility profiles and calendars for food security and livelihoods analysis

Pedro J. Zufiria [1], David Pastor-Escuredo [1], Luis Úbeda-Medina [1], Miguel A. Hernández-Medina [1], Iker Barriales-Valbuena [1], Alfredo J. Morales [1], Wilfred Nkwambi [2], John Quinn [3], Paula Hidalgo-Sanchís [3], Miguel Luengo-Oroz [3]

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.

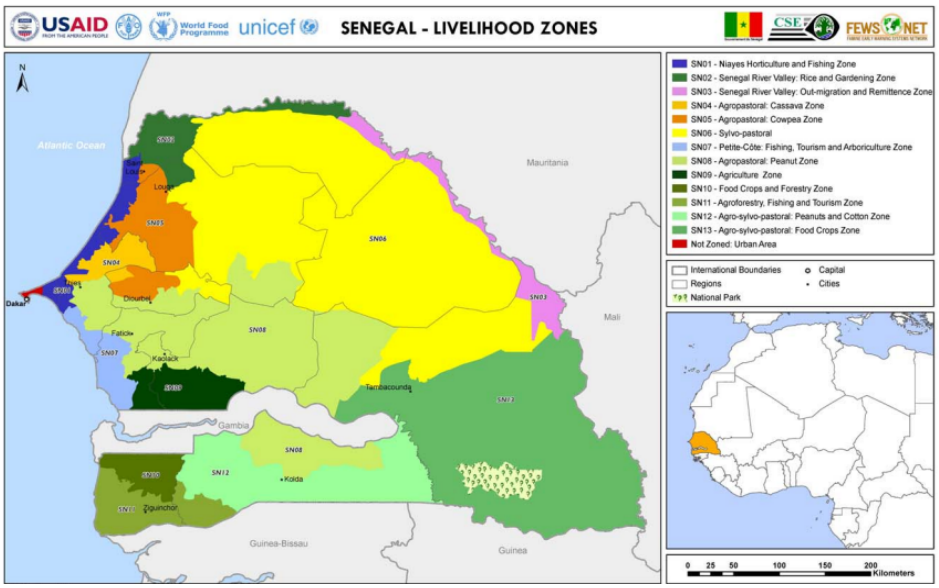


Fig. 1a: Livelihood zones map in Senegal. This map has been used to generate an Arrondissement to Livelihood assignment.

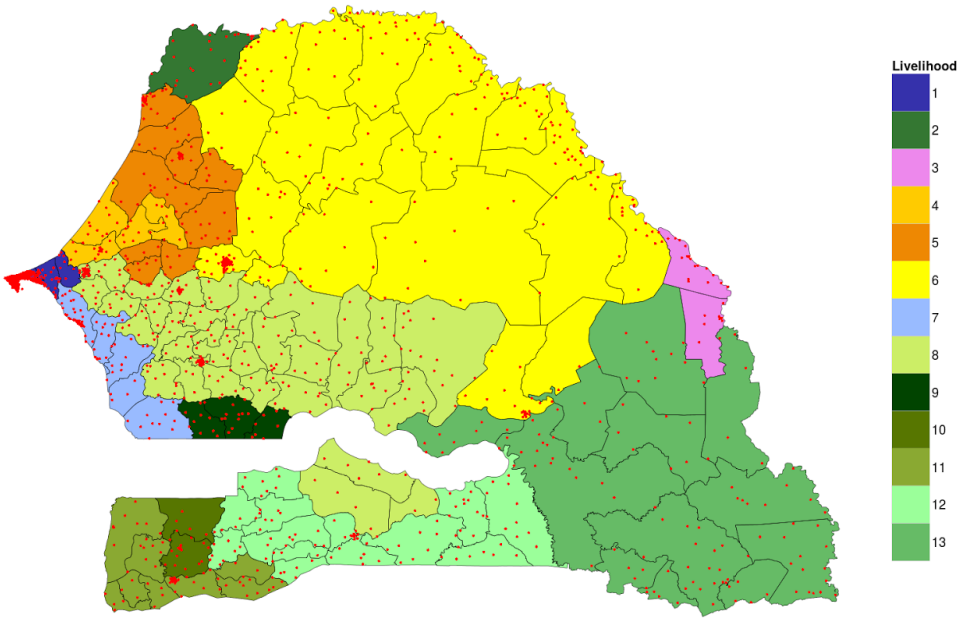


Fig. 1b: Levels of geographical resolution in Senegal based upon D4D datasets. The coverage regions of the antennas (red dots) correspond to Data-set 2 and can be aggregated to the Arrondissement level (line boundaries). In this work, the Arrondissements (Data-set 3) are grouped by Livelihood Zones (colors) 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 Mmin consecutive months in the target LZ.
- The user must have not stayed more than Mmax months in the target LZ.
- The user must have stayed at least Moutmin 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.

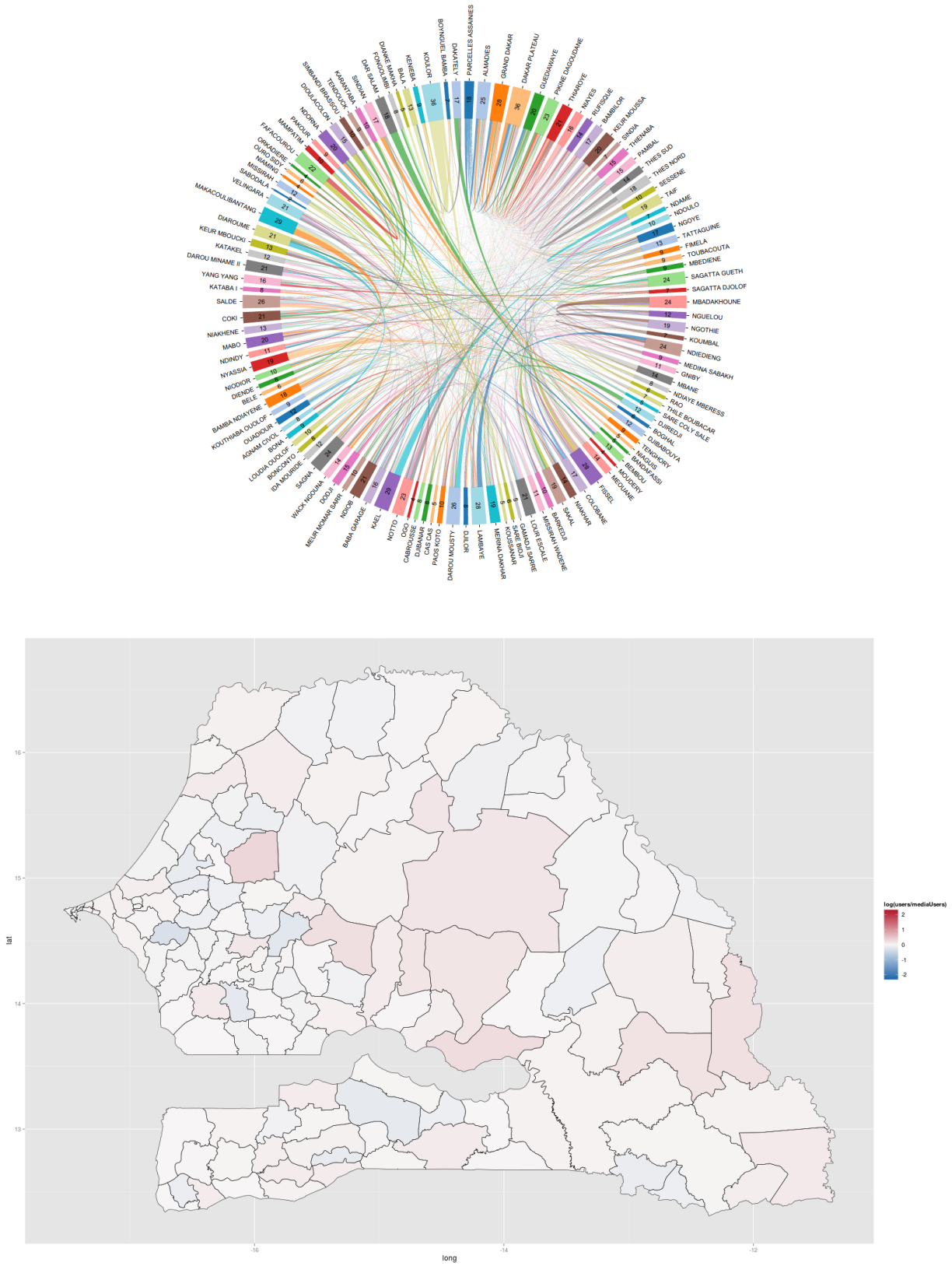


Figure 2. Regular distribution of movements in a normal day of year 2013.

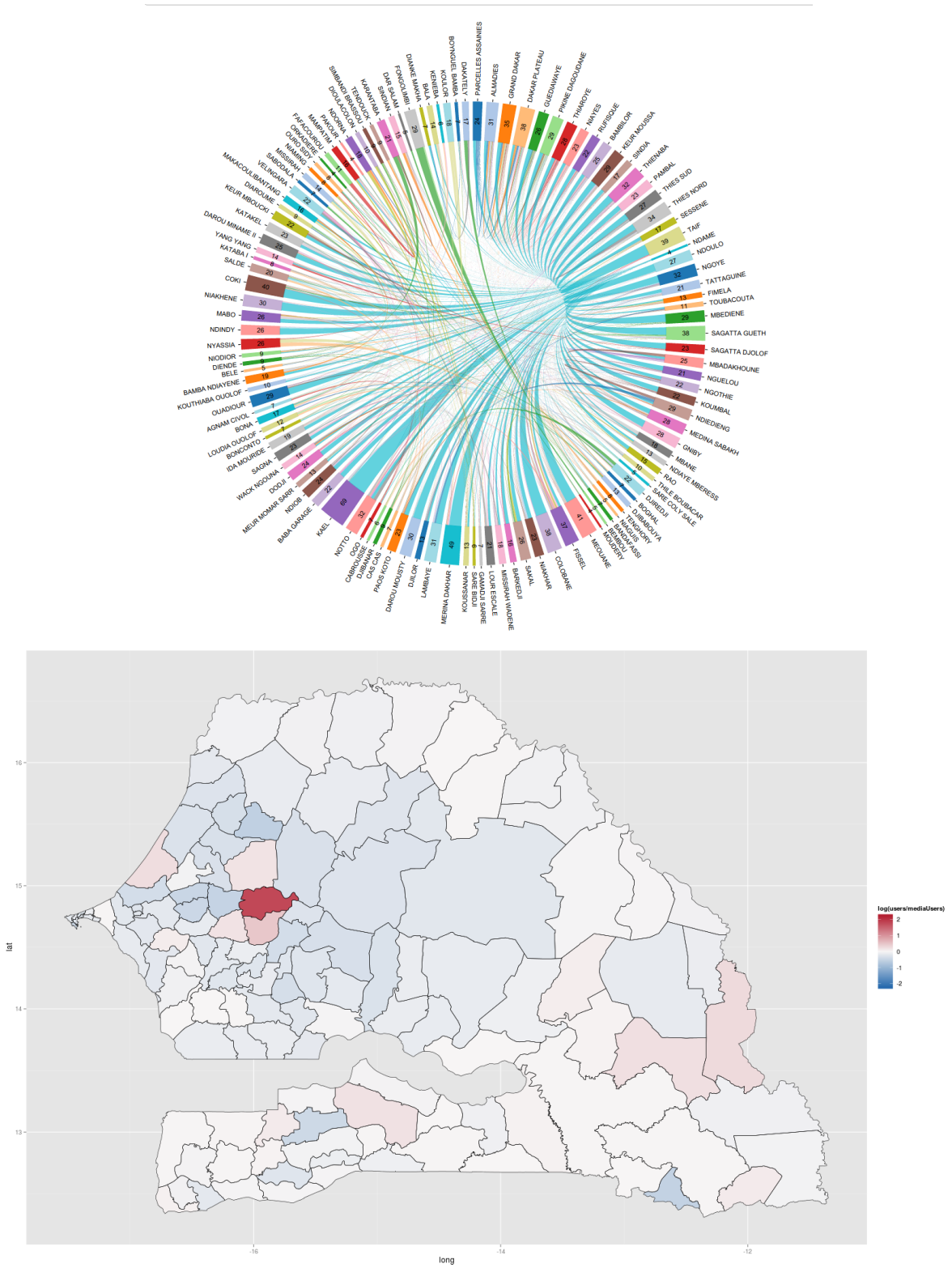


Figure 3. Grand Magal detection on the 21st of December 2013. Both images show how people from almost all Arrondissements move to the city of Touba in the Arrondissement of Ndame.

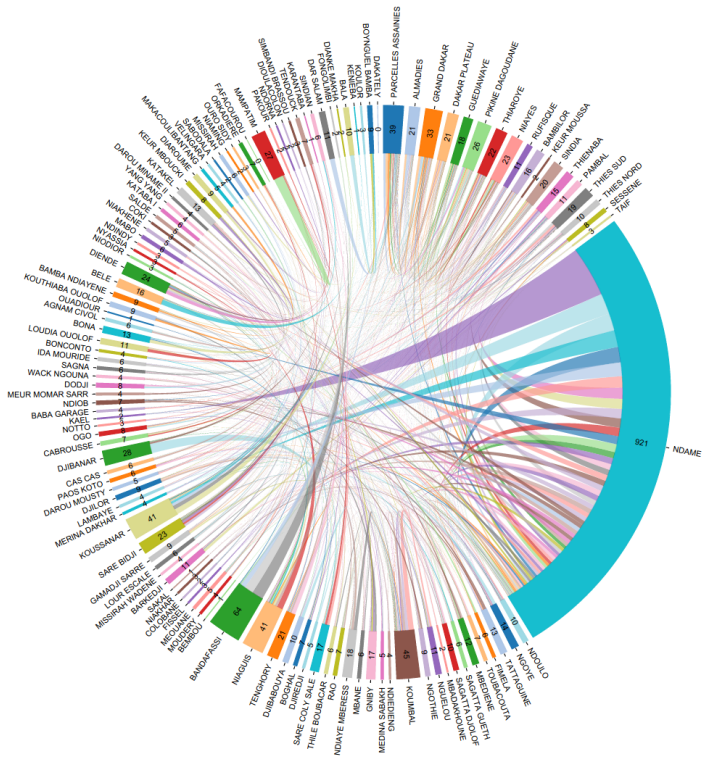


Figure 4. Grand Magal detection on the 23rd of December 2013. The image shows people returning to their home Arrondissements after the celebration of the Grand Magal.

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.

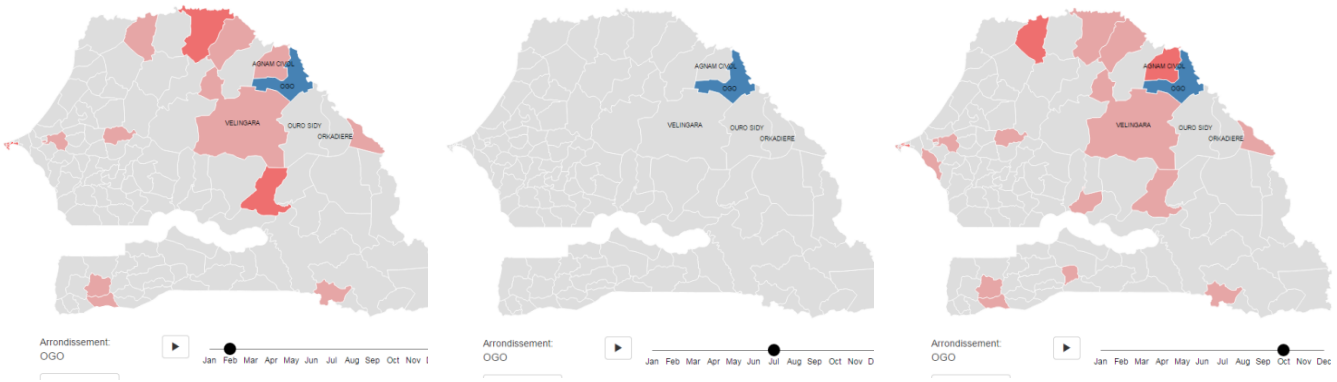


Figure 5: Distribution along Senegal for the population that was completely localized in the Arrondissement of MATAM during July.

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 HAUVs (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.

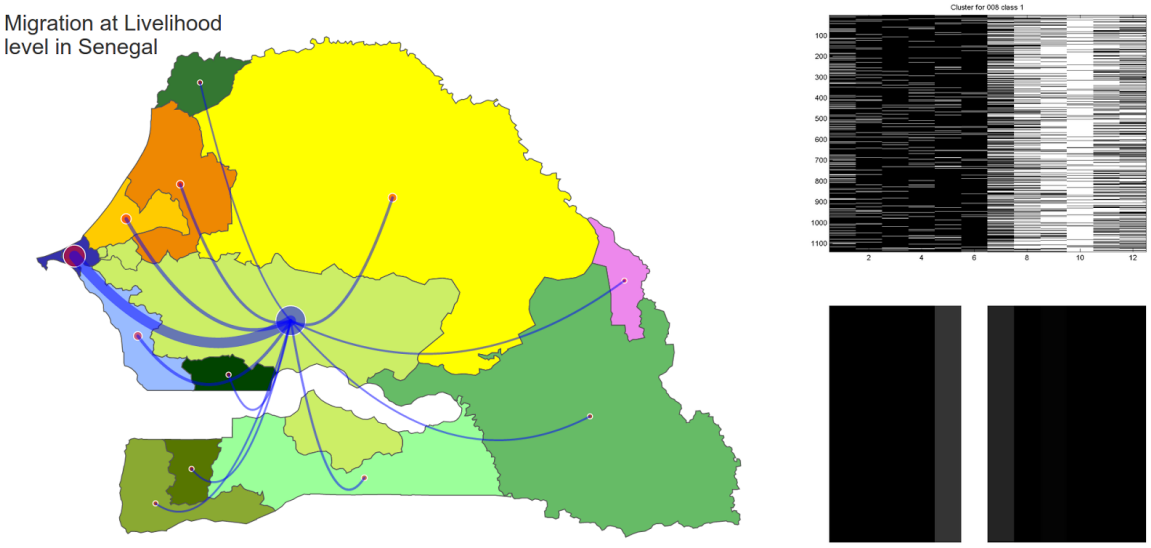


Figure 6. Inter livelihoods flows of people arriving the second half of the year to Livelihood Zone 8.

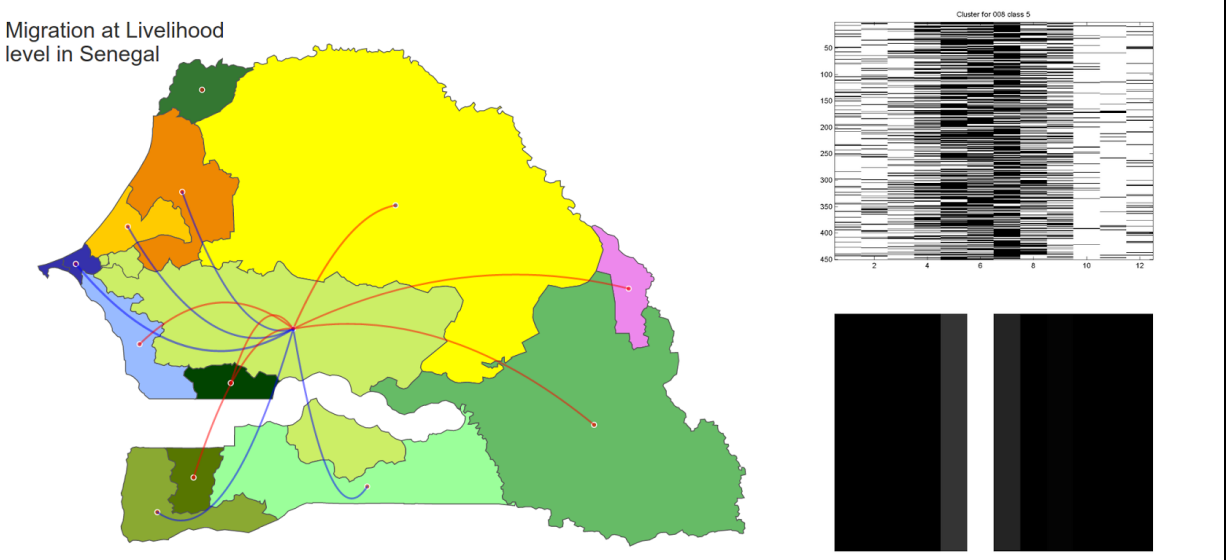


Figure 7. Inter livelihoods flows of people leaving during rainy season from Livelihood Zone 8.

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.

.

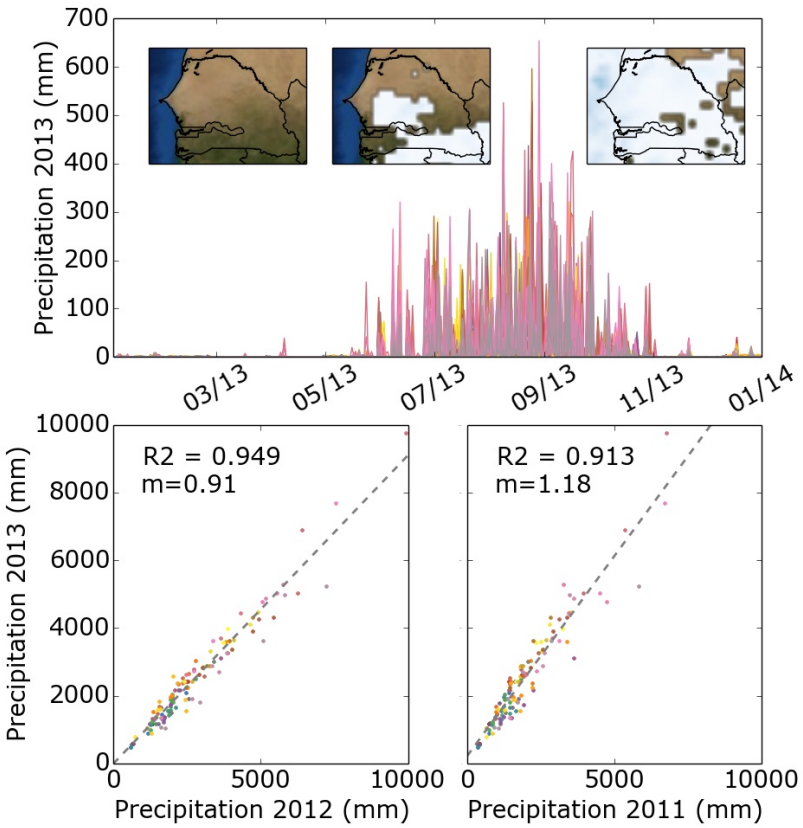


Figure 8. Top: Temporal evolution of precipitation levels by Arrondissement in 2013. Insets: Visualization of rainfalls over the map of Senegal at days: Feb 1st, Jun 1st and Aug 8th (from left to right). Bottom: Correlation between rainfall levels by Arrondissement in 2013 against 2011 and 2012.

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.

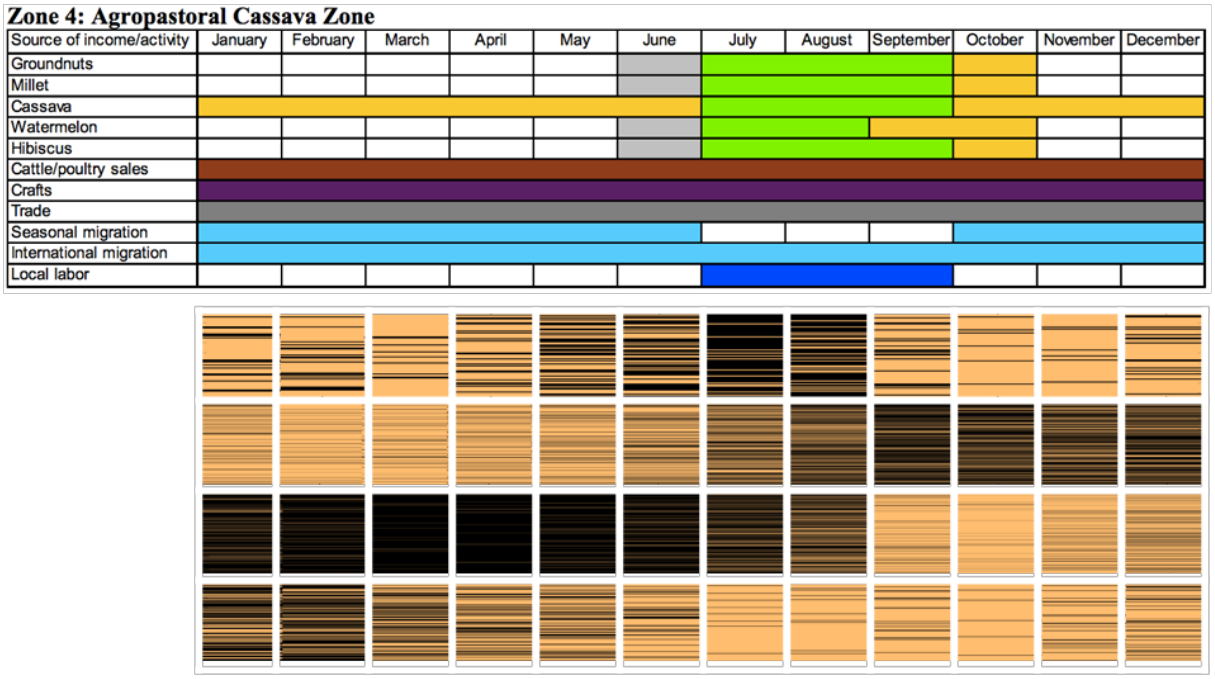


Figure 9. Zone 4 calendar of sources of income and activities against users' mobility profiles. Check [2] for details in the color code of the agricultural maps.

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.

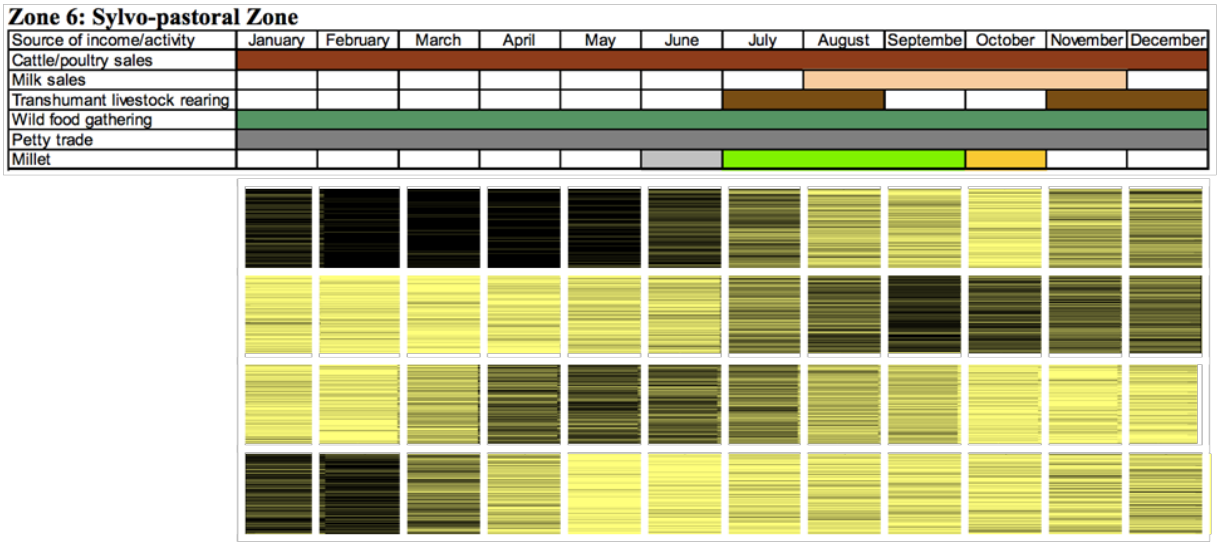


Figure 10. Zone 6 calendar of sources of income and activities against users' mobility profiles.

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.

Zone 8: Agro-pastoral Peanut Zone

Source of income/activity	January	February	March	April	May	June	July	August	Septembe	October	November	December
millet, cowpea, watermelon												
Peanuts												
Maize												
Sorghum												
Cattle/poultry sales												
Baobab gathering												
Jujube gathering												
Salt harvesting												
Market gardening												
Transhumant livestock rearing												
Trade												
Crafts												
Seasonal migration												
International migration												
Local labor												

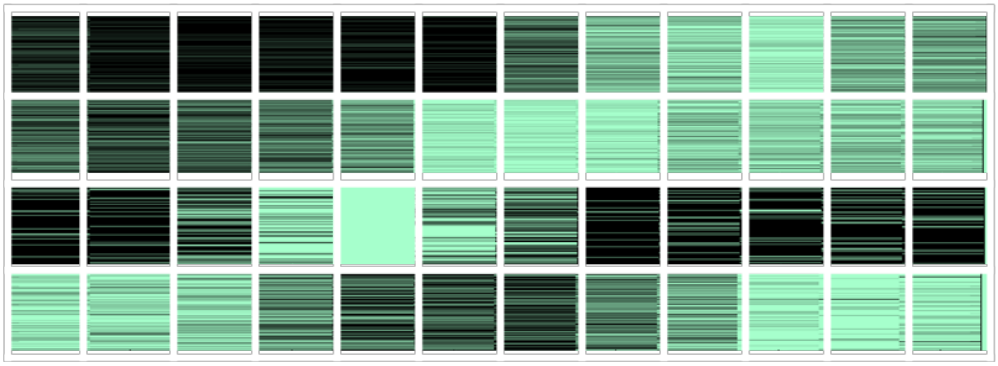


Figure 11. Zone 8 calendar of sources of income and activities against users' mobility profiles.

Zone 13: Agro-sylvo-pastoral Food Crops Zone

Source of income/activity	January	February	March	April	May	June	July	August	Septembe	October	November	December
Maize												
Groundnuts												
Cotton												
Cattle/poultry sales												
Milk production												
Forestry												
Baobab fruit gathering												
Shea nut gathering												
Made fruit gathering												
Gold mining												
Labor migration												

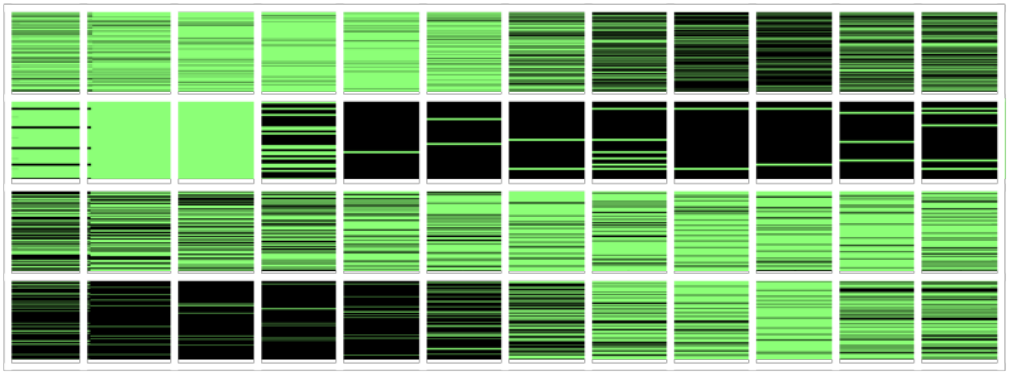


Fig. 12. Zone 13 calendar of sources of income and activities against users' mobility profiles.

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

[1] Y.-A. de Montjoye et al., D4D-Senegal: The Second Mobile Phone Data for Development Challenge.

[2] WFP/FAO/SE-CNSA/FEWS, Comprehensive Food Security and Vulnerability Analysis. Livelihood Zone Descriptions.
www.fews.net/sites/default/files/documents/reports/sn_livelihoodzonedescriptions2011_en.pdf

[3] FEWS.NET, Famine Early Warning Systems Network. http://www.fews.net/ .

[4] Kelly P.M. and Adger W.N., Theory and Practice in assessing Vulnerability to Climate Change and facilitating Adaptation. Climatic Change vol. 47 , no. 4, pp. 325-352 (2000).

[5] Pastor-Escuredo D., García-Morales A. et al. “Flooding from the Lens of Mobile Phone Data”. IEEE GHTC’14.

[6] UNU- EHS: Linking Environmental Change, Migration and Social Vulnerability. Studies of the University: Research, Counsel, Education. No. 12, 2009.

[7] USAID From the American People. http://www.usaid.gov/ .

[8] WFP, World Food Program. http://www.wfp.org/ .

[9] TRMM Project 3B42 daily v7:
http://disc.sci.gsfc.nasa.gov/recipes/?q=datacollection/TRMM_3B42_daily.007/description

[10] Viz1: http://goo.gl/aZ6UJu

[11] Viz2: http://pulselabkampala.ug/d4d-senegal/

[12] Viz3: http://goo.gl/QJKfl4

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).

$$P_i = f(D_i, S_i) \tag{1}$$

with

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

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

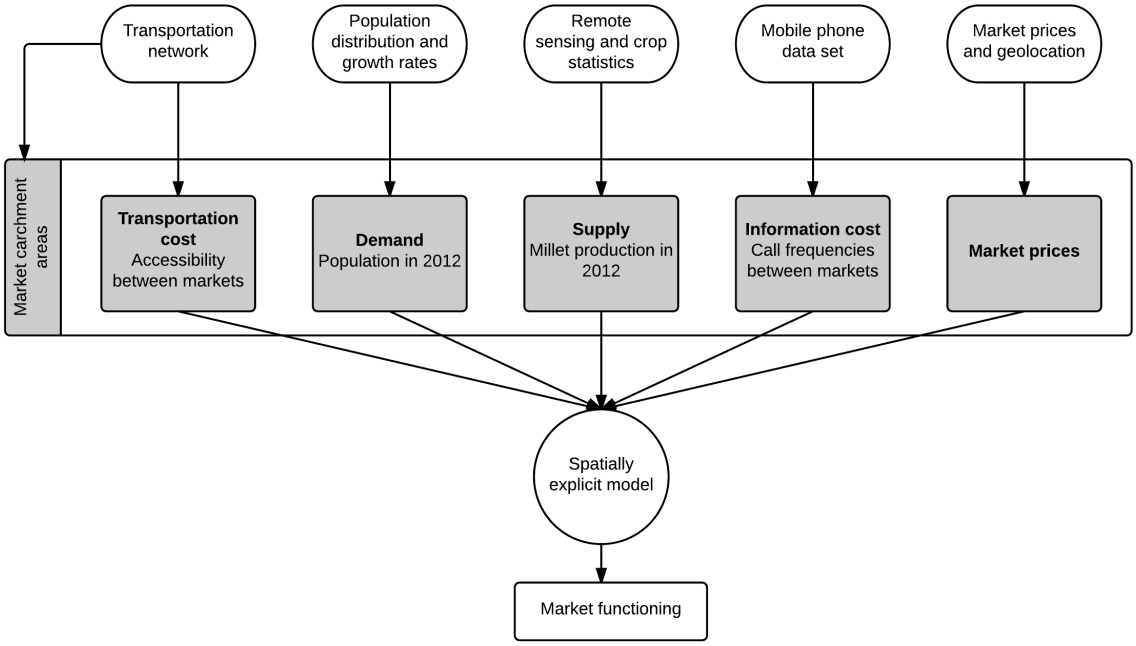


Fig. 1: Overview of the method (erratum: population data used are for 2013)

market (catchment areas) is the area that minimize the journey time (using road network) between each market (see section 2.2).

$$P_i \propto PsPr_i = \frac{Pop_i}{Prod_i + 1} \tag{2}$$

where $PsPr_i$, Pop_i, $Prod_i$ 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):

$$TC_{i,j} = \beta_{i,j} \times \frac{PsPr_i - PsPr_j}{d_{i,j} + 1} \tag{3}$$

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,

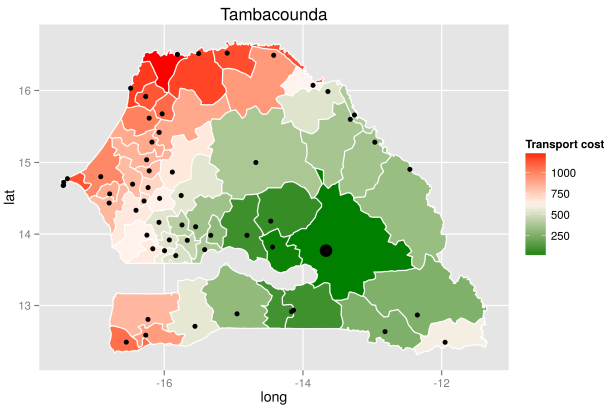


Fig. 2: Distance to Tambacounda market from each geolocated market.

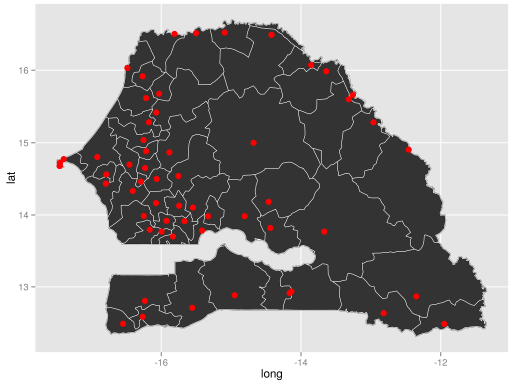


Fig. 3: Market catchment areas derived with an accessibility model using the transportation network

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 vegetations, with

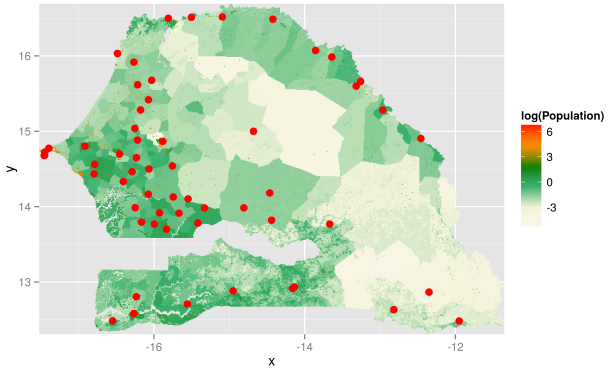


Fig. 4: Estimated distribution of the population for 2013

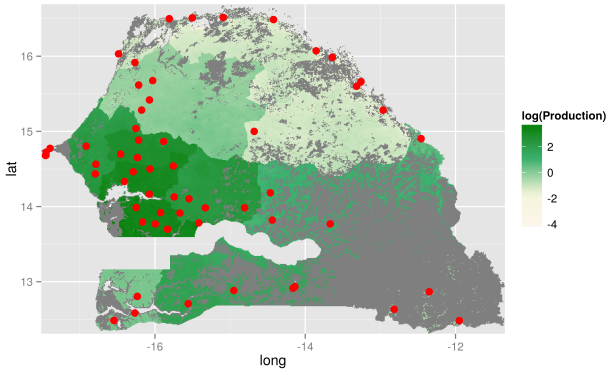


Fig. 5: Estimated Millet production for 2012

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:

$$\text{Prod}_i = \text{CUM}_i^{NDVI} \times \frac{\text{Prod}_{region}}{\text{CUM}_{region}^{NDVI}} \quad (4)$$

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

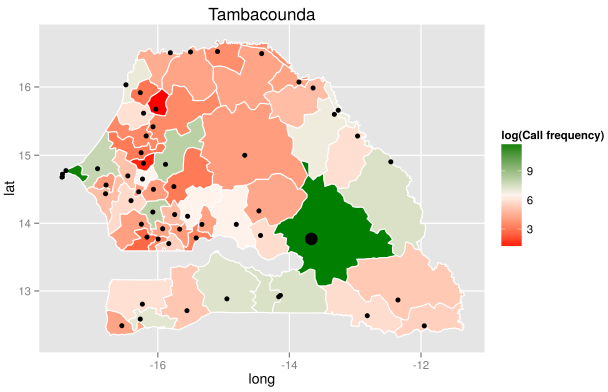


Fig. 6: Incoming calls to Tambacounda

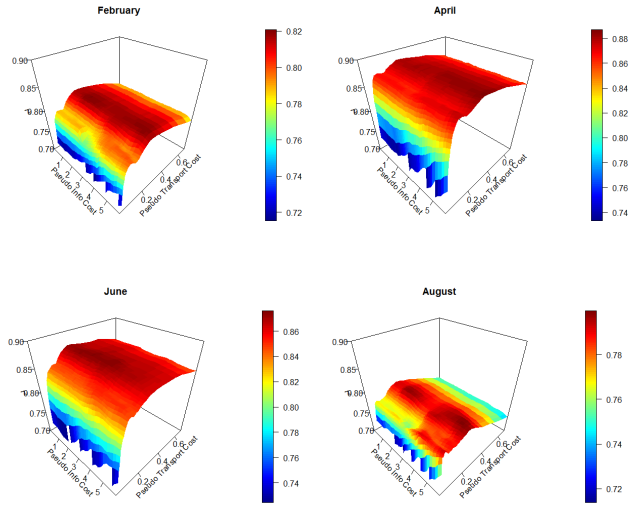


Fig. 7: Correlation between pseudo-prices and actual millet prices

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 β used to estimate the cost of information is defined as:

$$\beta_{i,j} = \begin{cases} 1 \text{ if } \frac{\log(Ncalls_{i,j}) - \min(\log(Ncalls_{i,j}))}{(IC_{i,j} - \min(\log(Ncalls_{i,j})))} > 1 \\ \text{else } \frac{\log(Ncalls_{i,j}) - \min(\log(Ncalls_{i,j}))}{(IC_{i,j} - \min(\log(Ncalls_{i,j})))} \end{cases} \quad (5)$$

where $Ncalls_{i,j}$, $IC_{i,j}$ 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 ($r^2 = 0.26$ for April, $r^2 = 0.23$ 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 between 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 β parameter of information asymmetry, and millet prices in 4 selected months.

Notice the high explanatory power reached by the simulated pseudo-prices for some combinations of transaction costs and beta parameter. Here it is worth nothing that the model is extremely efficient, reaching correlations above 0.8 with the fitting of only two parameters when combining a plethora of data sets in a sound theoretical framework. Correlations increase immediately with the inclusion of even small transportation costs and reach their maximum for pseudo-transportation costs around 0.4 units per kilometer. Clearly, today this is the main source of inefficiencies in millet markets in the country.

On the other hand, due to the widespread coverage of mobile phone network in the country, information asymmetries seem to play a minor, although not negligible, role in millet prices differentials among markets. Interestingly enough, the effect of information asymmetries seems to be more important during the months where the pseudo-prices have lower explanatory power. The reasons for that remain unclear given that the mobile phone data used is limited to 2013, a year with already a good mobile phone coverage throughout the country.

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.

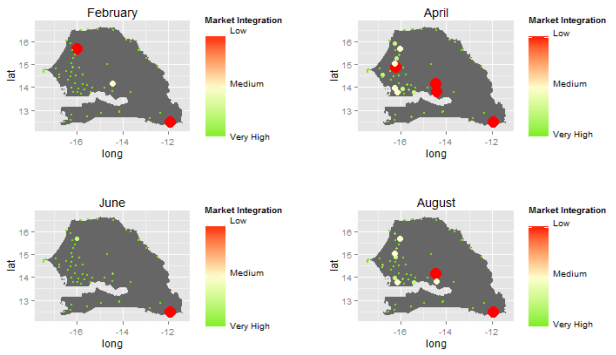


Fig. 8: Market integration defined by mobile phone data (number of markets with a $\beta < 1$ for the best modelling in February, April, June and August)

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

Food and commodity prices data store, 2014. URL `http://foodprices.vam.wfp.org/`.

Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. D4d-senegal: The second mobile phone data for development challenge. *arXiv preprint arXiv:1407.4885*, 2014.

Dong. Millet has many faces. *Global Agricultural Information Network*, 2011.

P. H. Eilers. A perfect smoother. *Analytical chemistry*, 75(14): 3631–3636, 2003.

G. Genovese, C. Vignolles, T. Nègre, G. Passera, et al. A methodology for a combined use of normalised difference vegetation index and corine land cover data for crop yield monitoring and forecasting. a case study on spain. *Agronomie*, 21(1):91–111, 2001.

U. Leonardi. Senegal land cover mapping, 2008.

J. Lewis, J. Rowland, and A. Nadeau. Estimating maize production in kenya using ndvi: some statistical considerations. *International Journal of Remote Sensing*, 19(13):2609–2617, 1998.

C. Linard. Population distribution mapping: the afripop dataset. In *Data from Conflict-Affected Regions: Filling-in the Blanks*, 2011.

K. Manjunath, M. Potdar, and N. Purohit. Large area operational wheat yield model development and validation based on spectral and meteorological data. *International Journal of Remote Sensing*, 23(15):3023–3038, 2002.

N. Quarmby, M. Milnes, T. Hindle, and N. Silleos. The use of multi-temporal ndvi measurements from avhrr data for crop yield estimation and prediction. *International Journal of Remote Sensing*, 14(2):199–210, 1993.

M. S. Rasmussen. Assessment of millet yields and production in northern burkina faso using integrated ndvi from the avhrr. *International Journal of Remote Sensing*, 13(18):3431–3442, 1992.

J. Ren, Z. Chen, Q. Zhou, and H. Tang. Regional yield estimation for winter wheat with modis-ndvi data in shandong, china. *International Journal of Applied Earth Observation and Geoinformation*, 10(4):403–413, 2008.

E01

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 'D4D-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"

E. A. Martinez-Cesena and P. Mancarella are with the University of Manchester, School of Electrical and Electronic Engineering, Electrical Energy and Power Systems Group, Manchester, M13 9PL UK (e-mail: {eduardo.martinezcesena; p.mancarella}@manchester.ac.uk).

M. Ndiaye is with the Ecole supérieure polytechnique de Dakar UCAD, Centre International de Formation et de Recherche en Energie Solaire, 5085 Dakar-Fann, Senegal (e-mail: emamadoulamine.ndiaye@ucad.edu.sn).

M. Schläpfer is with the Santa Fe Institute, Santa Fe, NM 87501 USA (e-mail: schlaepfer@santafe.edu).

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 mobility patterns 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].

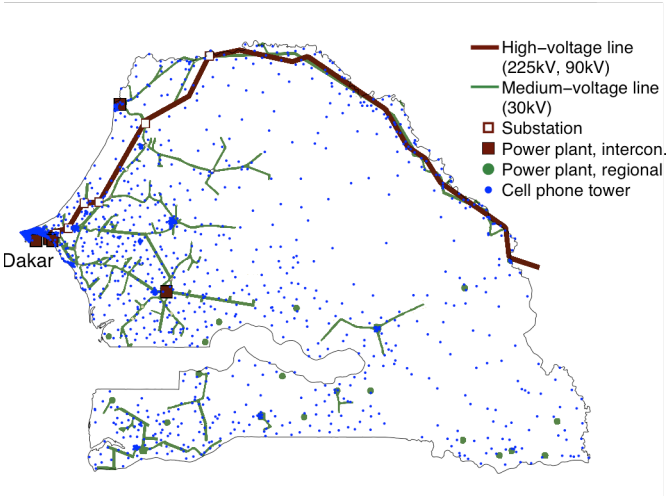


Fig. 1. Existing electricity infrastructure in Senegal and location of the mobile phone towers. The transmission and distribution network as well as the location of the power stations are adopted from [12].

B. Electricity consumption and infrastructure data

For the purpose of this project, the national electric utility in Senegal – the "Société Nationale d'Éléctricité" (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

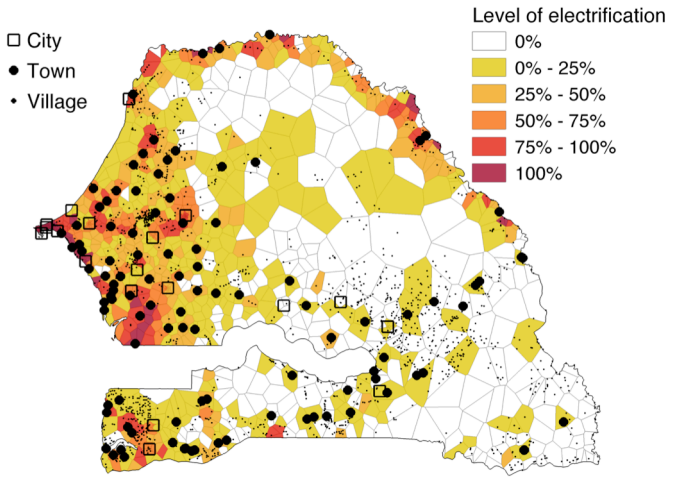


Fig. 2. Level of electrification in the Voronoi polygons defined by the location of the mobile phone towers. The locations of the settlements (cities, towns, villages) are adopted from OpenStreetMap.

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.

TABLE I
EXAMPLE OF THE MAXIMUM AMOUNT OF INSTITUTIONS/SERVICES THAT ARE CONSIDERED FOR AVERAGE VILLAGES OF DIFFERENT POPULATION SIZES IN SENEGAL [15].

	Village size (population)			
	500	1,000	5,000	10,000
Hospitals	1	1	1	2
Schools	1	1	2	3
Markets	1	1	3	13
Public Lighting points	3	6	50	99

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.

Together, this information facilitates the differentiation of the mobile phone data based on the context of the area where the mobile phone towers are located. Therefore, we approximated the reception area of each mobile phone tower by a Voronoi tessellation (i.e., the area corresponding to a given tower comprises all points that are closer to that tower than to any other tower) [1]. As a result, the mobile phone data can be classified based on the level of electrification (or access to education, health services etc.) in each Voronoi polygon, as shown in Fig. 2. This classification is critical to identify the conditions where mobile phone data is a good proxy for energy needs, as will be further discussed below.

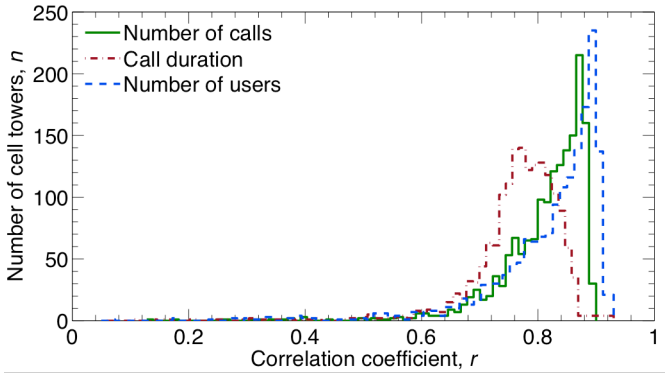


Fig. 3. Histogram of the linear correlation between the electricity load curve and the mobile phone activity within each mobile phone tower area.

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,

$$r_i = \frac{\sum_t \left(A_i(t) - \langle A_i \rangle\right)\left(D(t) - \langle D \rangle\right)}{\sqrt{\sum_t \left(A_i(t) - \langle A_i \rangle\right)^2}\sqrt{\sum_t \left(D(t) - \langle D \rangle\right)^2}} \quad (1)$$

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 $\langle \cdot \rangle$ 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 $\langle r^d \rangle = 0.76$, $\langle r^n \rangle = 0.8$ and $\langle r^u \rangle = 0.81$, 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

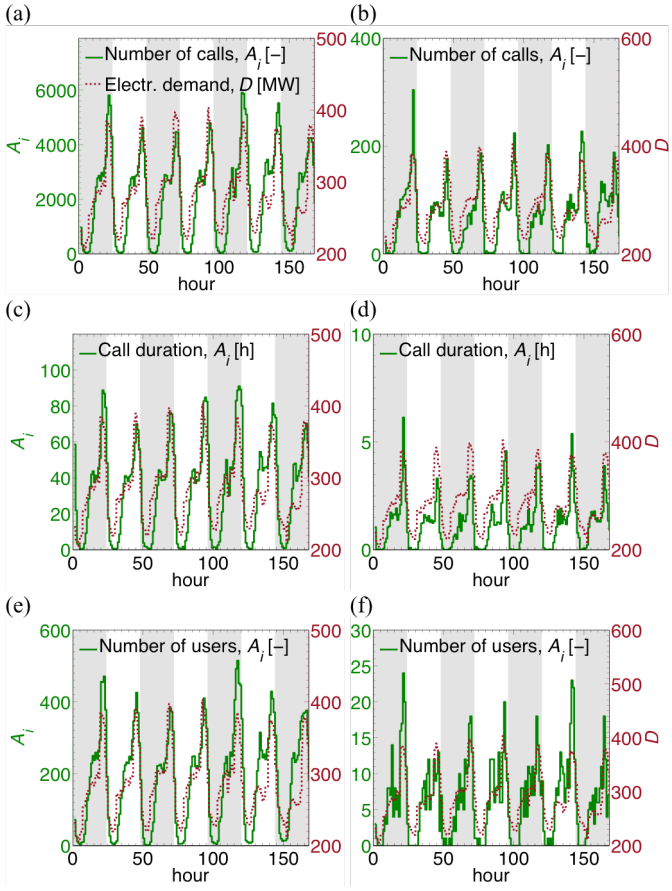


Fig. 4. Comparison of the mobile phone activity (measured as the hourly number of calls, call duration and number of users) with the aggregated electricity load profile (dotted line, in MW) for one week in January 2013 (Monday-Sunday). (a),(c),(e) Mobile phone tower in the city of Dakar. (b),(d),(f) Mobile phone tower in the rural area of Fayil. The shaded areas separate the different days.

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.

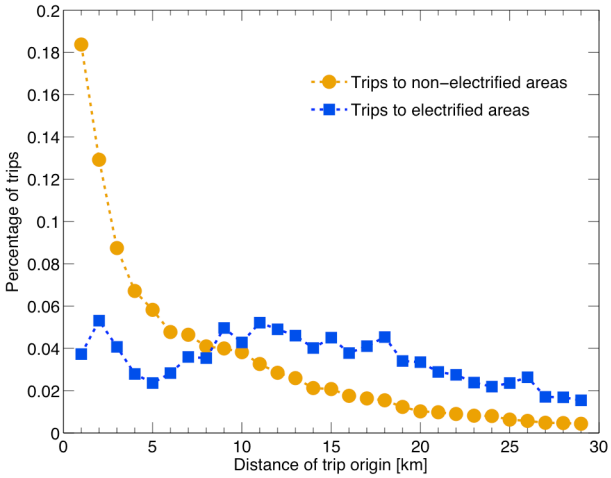


Fig. 5. Normalized histogram of the number of trips to electrified and non-electrified areas, binned by travel distance.

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.

TABLE II
ECONOMIC AND TECHNICAL PARAMETERS FOR THE MV GRID EXTENSION ELECTRIFICATION ALTERNATIVE [15]

MV line	
Cost	8M CFA/km
O&M cost	2% (of the total investment)
Transformer	
Capacity	5 kVA – 100 kVA
Cost	2M CFA – 4M CFA
O&M cost	3%
LV line	
Costs	6M CFA/km
O&M costs	3%
Generation	
Generation costs	83.4 CFA/kWh
Losses	15%

$$NPC_{GE} = I_{GE} + \sum_{i=0}^{T} \frac{C_{GE}\left(D_i + L_{MV,i}\right) + M_{GE}}{\left(1+d\right)^i} \quad (2)$$

$$M_{GE} = M_{MV} + M_{TR} + M_{LV} \quad (3)$$

$$I_{GE} = I_{MV} + I_{TR} + I_{LV} \quad (4)$$

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 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.

TABLE III
ECONOMIC AND TECHNICAL PARAMETERS FOR THE DIESEL ENGINE-BASED MICROGRID ALTERNATIVE [15]

Generation	
Capacity	10kVA – 50kVA
Cost	6.4M CFA – 9.8M CFA
Generation costs	216CFA/kWh
O&M cost	5%
Life time	5 years
LV line	
Costs	6M CFA/km
O&M costs	3%
Losses	5%

$$NPC_{MG} = I_{MG} - S_G + \sum_{i=0}^{T} \frac{C_{MG}\left(D_i + L_{MG,i}\right) + M_{MG}}{\left(1+d\right)^i} \quad (5)$$

$$M_{MG} = M_G + M_{LV} \quad (6)$$

$$I_{MG} = I_{LV} + \sum_{i=1}^{floor\left(\frac{T}{LT_G}\right)} \frac{I_G}{\left(1+d\right)^{i \cdot LT_G}} \quad (7)$$

$$S_G = \frac{I_G}{\left(1+d\right)^{floor\left(T/LT_G\right)\cdot LT_G}} \cdot \frac{T - floor\left(T / LT_G\right)}{LT_G} \quad (8)$$

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 LT_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].

PV Module	
Capacity	20W – 150W
Cost	88k CFA – 660k CFA
O&M costs	1%
Batteries	
Capacity	14Ah – 38Ah
Cost	40k CFA – 70kCFA
O&M cost	1%
Converter	
Costs	28k CFA
O&M costs	1%

$$NPC_{PV} = I_{PV} - S_B + \sum_{i=0}^{T} \frac{M_{PV}}{(1+d)^i} \tag{9}$$

$$M_{PV} = M_{MO} + M_{CV} + M_B \tag{10}$$

$$I_{PV} = I_{MO} + I_{CV} + \sum_{i=0}^{floor\left(\frac{T}{LT_B}\right)} \frac{I_B}{(1+d)^{i \cdot LT_B}} \tag{11}$$

$$S_B = \frac{I_B}{(1+d)^{floor(T/LT_B)\cdot LT_B}} \cdot \frac{T - floor\left(T / LT_B\right)}{LT_B} \tag{12}$$

where I_{PV} is the total investment cost in CFA, S_B is the salvage value of the batteries and LT_B represents the lifetime of the batteries (in years). The parameters M_{PV}, M_{MO}, M_{CV} and M_B denote the annual O&M costs in CFA/year associated with the total investment, modules, converter and batteries, respectively, and I_{MO}, I_{CV} and I_B represent investments in modules, converters and batteries, respectively.

We applied the detailed simulation technique proposed in [18], based on an estimated electricity profile (i.e., a scaled version of the aggregated electricity profile) and the technical characteristics of the converters and batteries, in order to optimize the design (i.e., type and amount of modules and batteries) of the PV system for each residential dwelling and institutional building such as a hospital, school, and so forth. The simulation approach allows the assessment of the lifetime of the batteries, which facilitates the identification of low cost PV designs. This is because it may be convenient to oversize the PV module or battery array if it leads to an increased lifetime for the batteries. The lifetime of the batteries is estimated with the Ah throughput model denoted by (13), (14) and (15), see [18] for a detailed description.

$$LT_B = \frac{AhT}{\sum_{t=1}^{8760} i_{eff,t}} \tag{13}$$

$$AhT = DOD_{rated} \cdot CAP_{rated} \cdot N_{rated} \tag{14}$$

$$i_{eff,t} = \frac{i_{rated}}{CAP(i_t)} \cdot \frac{N_{rated}}{N(DOD_t)} \cdot i_t \tag{15}$$

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 $CAP(i_t)$ (in A) 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).

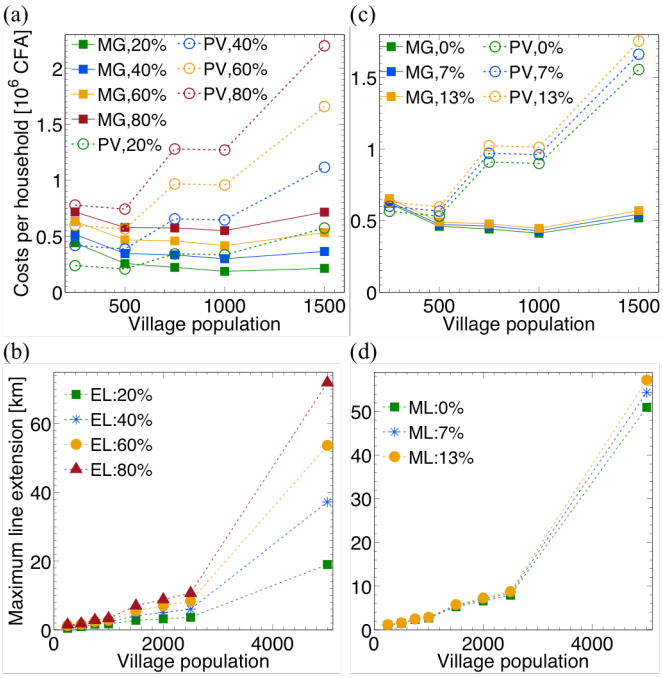


Fig. 6. (a) Costs of PV systems (PV) and diesel-based Microgrids (MG) for different electrification levels. The curves are averages over three migration levels (0%,7%,13%). (b) Maximum MV grid extensions that are economically competitive with PV and MG for different electrification levels. The curves 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

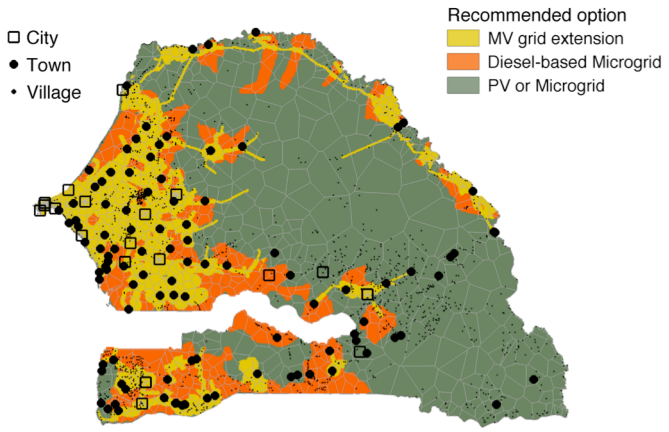


Fig. 7. Electrification recommendations considering 60% electrification levels and 13% migration to electrified villages.

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. Schläpfer thanks Luis Bettencourt, Paul Hines and Seth Blumsack for helpful discussions.

References

[1] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 2008.

[2] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, and A. E. Gaughan, "Dynamic population mapping using mobile phone data," *Proc. Nat. Acad. of Sci. USA*, vol. 111, pp. 15888–15893, 2014.

[3] M. Schläpfer, L. M. A. Bettencourt, S. Grauwin, M. Raschke, R. Claxton, Z. Smoreda, G. B. West, and C. Ratti, "The scaling of human interactions with city size," *J. R. Soc. Interface*, vol. 11, pp. 20130789, 2014.

[4] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *Eur. Phys. J. - Spec. Top*, vol. 214, pp. 481–518, 2012.

[5] J. Keirstead, M. Jennings, and A. Sivakumar, "A review of urban energy system models: Approaches, challenges and opportunities," *Renew. Sust. Energy Rev.*, vol. 16, pp. 3847–3866, 2012.

[6] World Bank, "World Development Indicators." [Online]. Available: databank.worldbank.org/data/home.aspx

[7] International Energy Agency (IEA), "Africa Energy Outlook 2014." [Online]. Available: www.worldenergyoutlook.org

[8] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel, "D4D-Senegal: The second mobile phone data for development challenge." [Online]. Available: arxiv.org/pdf/1407.4885v2.pdf

[9] J. C. Aker and I. M. Mbiti, "Mobile phones and economic development in Africa," *J. Econ. Persp.*, vol. 24, pp. 207–232, 2010.

[10] P. A. Ndiaye, private communication, November 2014.

[11] Senelec. [Online]. Available: www.senelec.sn

[12] Commission de Régulation du Secteur de l'Electricité (CRSE), "Carte de l'ensemble du réseau," [Online]. Available: http://www.crse.sn

[13] A. Sanoh, L. Parshall, O. F. Sarr, S. Kum, and V. Modi, "Local and national electricity planning in Senegal: Scenarios and policies," *Energ. Sust. Dev.* vol. 16, pp. 13–25, 2012.

[14] B. Diuf, R. Pode and R. Osei, "Initiative for 100% rural electrification in developing countries: Case study of Senegal," *Energ. Policy*, vol 59, pp. 926–930, 2013.

[15] Agence Sénégalaise d'Electrification Rurale and the Energy Group, Columbia Earth Institute, "Costing for National Electricity Interventions to Increase Access to Energy, Health Services, and Education." 2007. [Online]. Available: http://modi.mech.columbia.edu

[16] Solar Radiation Data (SoDa). [Online]. Available: http://www.soda-is.com

[17] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira Jr, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example," *Transport. Res. C*, vol. 26, pp. 301–313, 2013.

[18] E. A. Martinez-Cesena, and J. Mutale, "Assessment of demand response value in photovoltaic systems based on real options theory," *IEEE PowerTech 2011,* Trondheim, June 2011.

[19] B. Ould Bilal, M. Ndongo, V. Sambou, P. A. Ndiaye, and C. M. Kebe, "Diurnal characteristics of the wind potential along the North-western coast of Senegal," *Int. J. Phys. Sci.*, vol. 6, pp. 7950–7960, 2011.

H01

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 prevalance 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. Surveyers 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 Yukich 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* 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 Tijanism 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.[1] 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* 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* 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 hypthesis that the majority of short term movements are focused around important holidays.[2] 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.

[2]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:

$$TotalMalaria_{it} = \alpha + \beta R_{it} + \gamma_i + \lambda_t + \epsilon_{it} \quad (1)$$

$$TotalMalaria_{it} = \alpha + \beta V_{it} + \gamma_i + \lambda_t + \epsilon_{it} \quad (2)$$

$$R_{it} = \sum_{j=1}^{73} (R_{jt} * H_j) \quad (3)$$

$$V_{it} = \sum_{j=1}^{73} (V_{jt} * H_j) \quad (4)$$

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, γ_i captures location fixed effects, λ_t captures time fixed effects, ϵ_{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 analagous 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 targe 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 Bakel and Kédougou with R_0s 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.[3] 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

[3]Similarly 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 assuptions 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

Balcan, Duygu et al. (2009). "Multiscale mobility networks and the spatial spreading of infectious diseases". In: *Proceedings of the National Academy of Sciences* 106.51, pp. 21484–21489.

Benyoussef, A et al. (1974). "Health effects of rural-urban migration in developing countries–Senegal". In: *Social Science & Medicine (1967)* 8.5, pp. 243–254.

Blumenstock, Joshua E (2012). "Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda". In: *Information Technology for Development* 18.2, pp. 107–125.

Bulletin de Surveillance Sentinelle du Paludisme No 1-46, 2013 (2013). Tech. rep. PNLP. URL: `http://www.pnlp.sn/Bulletin-Surveillance-Palu/`.

Deshingkar, Priya and Sven Grimm (2004). "Voluntary internal migration: An update". In: *London: Overseas Development Institute* 44.

Doolan, Denise L, Carlota Dobaño, and J Kevin Baird (2009). "Acquired immunity to malaria". In: *Clinical microbiology reviews* 22.1, pp. 13–36.

Enns, E.A. and J.H. Amuasi (2013). *Human mobility and communication patterns in Cote d'Ivoire: A network perspective for malaria control*. Tech. rep. D4D Challenge 1 Book.

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

Gething, Peter W et al. (2011). "A new world malaria map: Plasmodium falciparum endemicity in 2010". In: *Malar J* 10.378, pp. 1475–2875.

Gonzalez, Marta C, Cesar A Hidalgo, and Albert-Laszlo Barabasi (2008). "Understanding individual human mobility patterns". In: *Nature* 453.7196, pp. 779–782.

Huang, Zhuojie, Andrew J Tatem, et al. (2013). "Global malaria connectivity through air travel". In: *Malar J* 12.1, p. 269.

Isaacman, Sibren et al. (2011). "Identifying important places in people's lives from cellular network data". In: *Pervasive Computing*. Springer, pp. 133–151.

Killeen, Gerry F, Amanda Ross, and Thomas Smith (2006). "Infectiousness of malaria-endemic human populations to vectors". In: *The American journal of tropical medicine and hygiene* 75.2 suppl, pp. 38–45.

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.

Lynch, Caroline and Cally Roper (2011). "The transit phase of migration: circulation of malaria and its multidrug-resistant forms in Africa". In: *PLoS medicine* 8.5, e1001040.

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

Osorio, Lyda, Jim Todd, and David J Bradley (2004). "Travel histories as risk factors in the analysis of urban malaria in Colombia". In: *The American journal of tropical medicine and hygiene* 71.4, pp. 380–386.

Pison, Gilles et al. (1993). "Seasonal migration: a risk factor for HIV infection in rural Senegal". In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 6.2, pp. 196–200.

Prothero, R Mansell (1977). "Disease and mobility: a neglected factor in epidemiology". In: *International Journal of Epidemiology* 6.3, pp. 259–267.

Rapport Statistique PNLP Spécial: 2010-2013 (2013). Tech. rep. PNLP. URL: `file:///C:/Users/Sveta/Downloads/Rapport%20annuel%20STATISTIQUE%20PNLP%20Special%202010_2013_VF%20(3).pdf`.

Régulation des Télécommunications et des Postes, Autorité de (2013). *Observatoire de la Telephonie Mobile: Tableau de bord au 31 décembre 2013*. Tech. rep.

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

Sanitaire, Service National de l'Information (2011). "Carte Sanitaire du Senegal: Mise a Jour de 2011". In:

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.

Smith, Thomas and Allan Schapira (2012). "Reproduction numbers in malaria and their implications". In: *Trends in parasitology* 28.1, pp. 3–8.

Stoddard, Steven T et al. (2009). "The role of human movement in the transmission of vector-borne pathogens". In: *PLoS neglected tropical diseases* 3.7, e481.

Tatem, Andrew J, Youliang Qiu, et al. (2009). "The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents". In: *Malar J* 8, p. 287.

Tatem, Andrew J and David L Smith (2010). "International population movements and regional Plasmodium falciparum malaria elimination strategies". In: *Proceedings of the National Academy of Sciences* 107.27, pp. 12222–12227.

Union, International Telecommunication (2013). "World Telecommunication/ICT Indicators Database". URL: `http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx`.

Vazquez-Prokopec, Gonzalo M et al. (2013). "Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment". In: *PloS one* 8.4, e58802.

Wesolowski, Amy et al. (2012). "Quantifying the impact of human mobility on malaria". In: *Science* 338.6104, pp. 267–270.

Williams, Nathalie E et al. (2014). "Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data". In: *arXiv preprint arXiv:1408.5420*.

World Malaria Report 2014 (2014). Tech. rep. World Health Organization.

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.

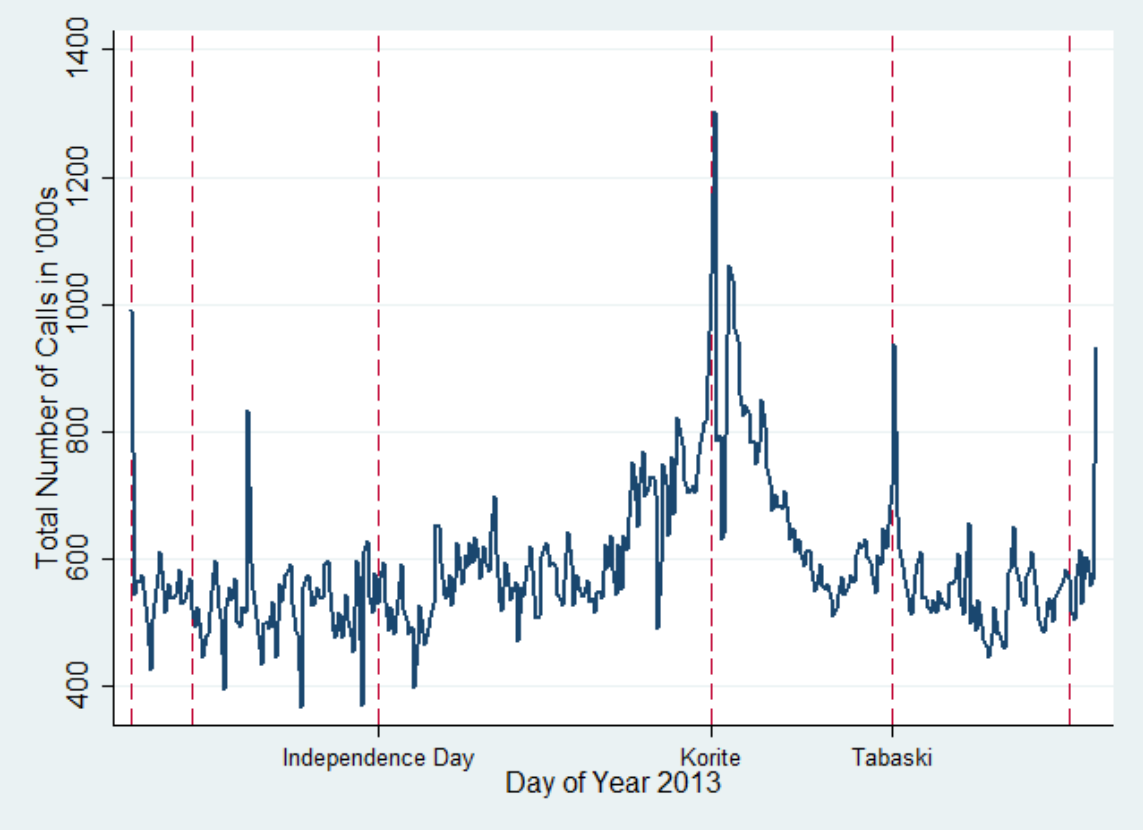
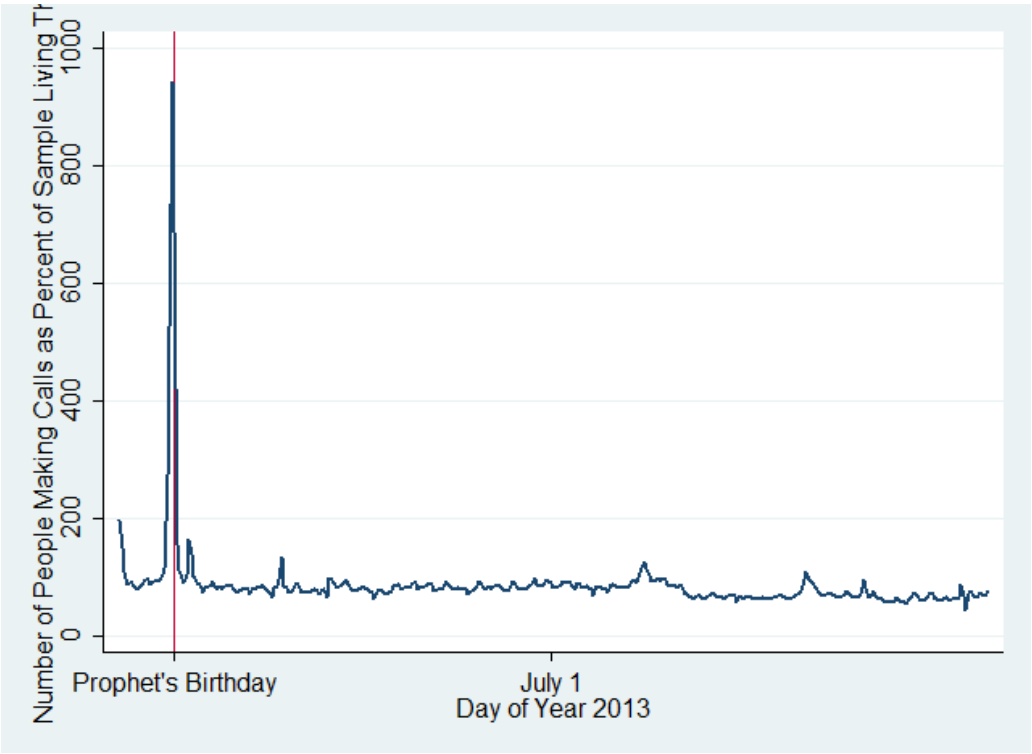
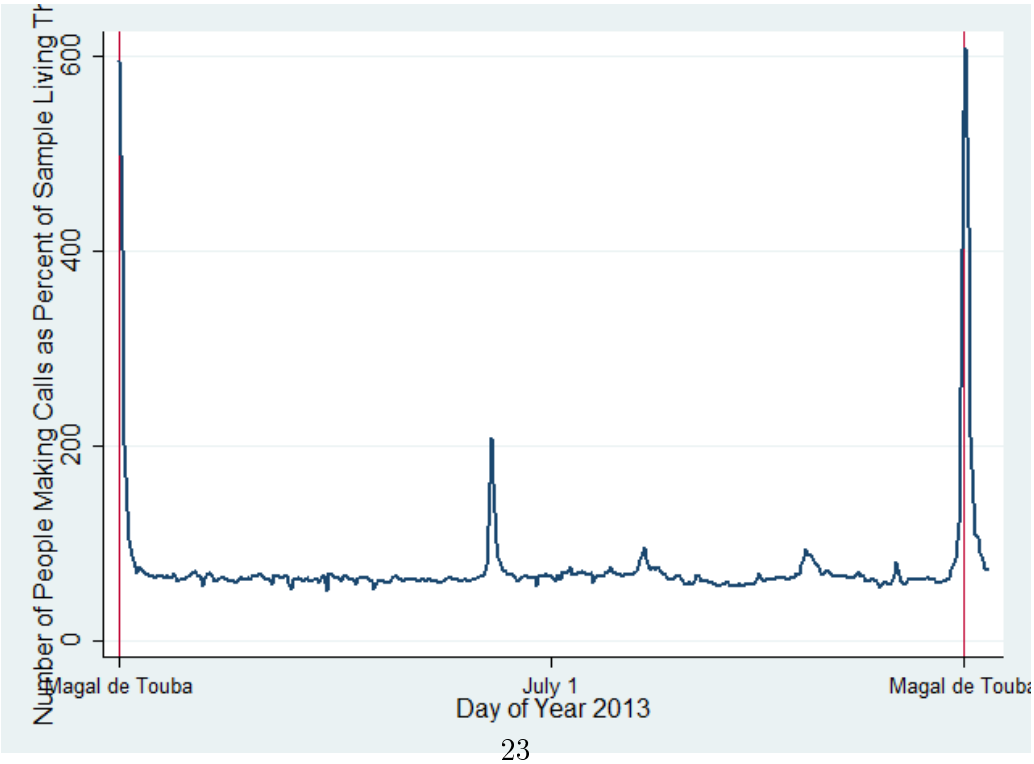


Figure 1: Total Calls Per Day



(a) Pambal



(b) Ndame

Figure 2: People Calling per Day as Percent of Sample Living in the Particular Arrondissement

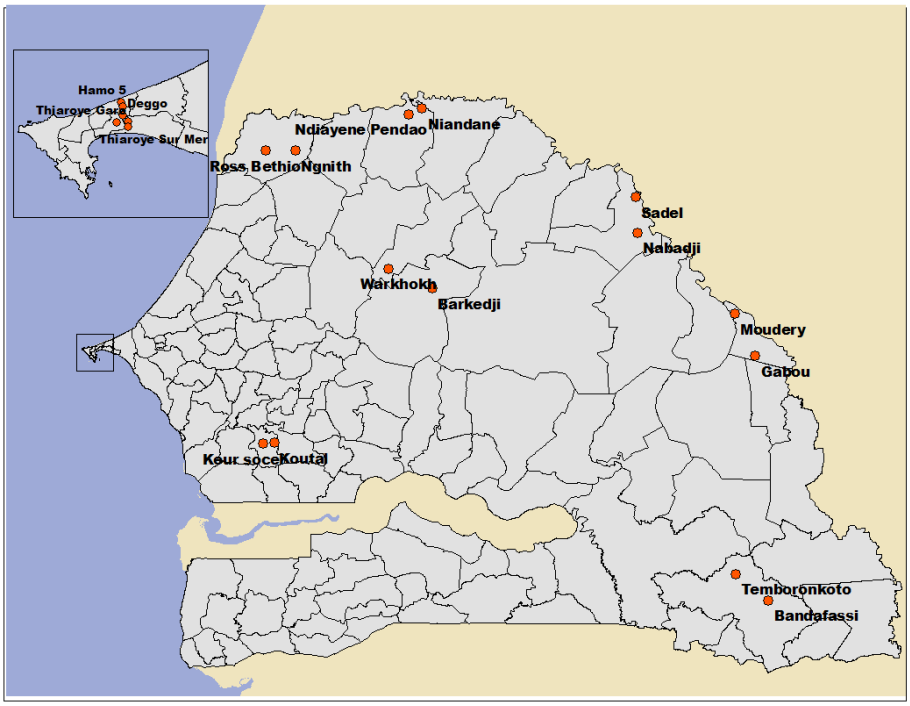


Figure 3: Map of Health Posts with Weekly Malaria Data Collection

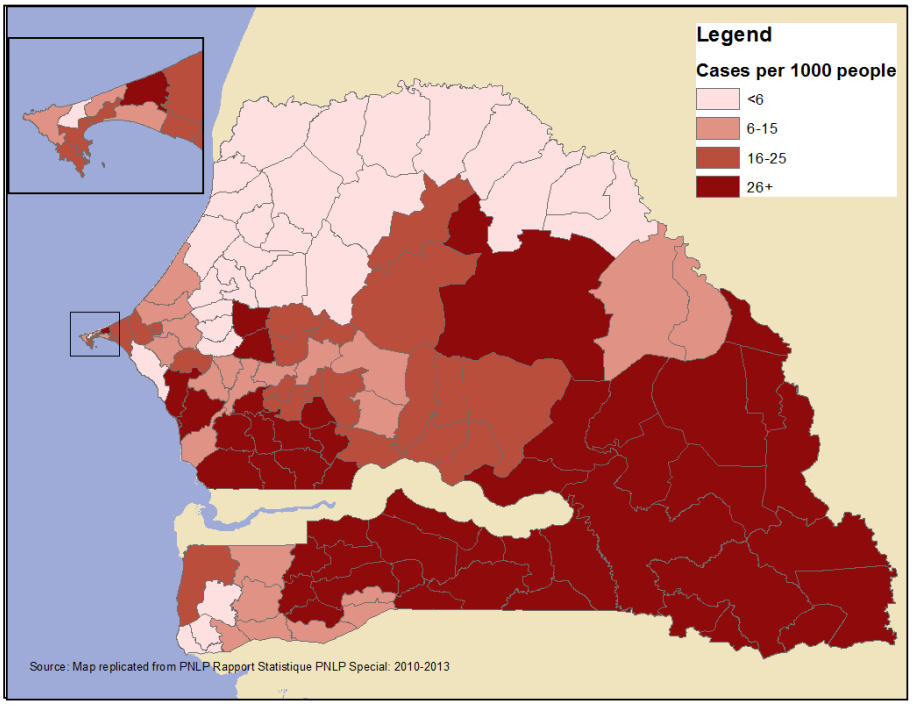


Figure 4: Map of Malaria Prevalence in Senegal in 2013

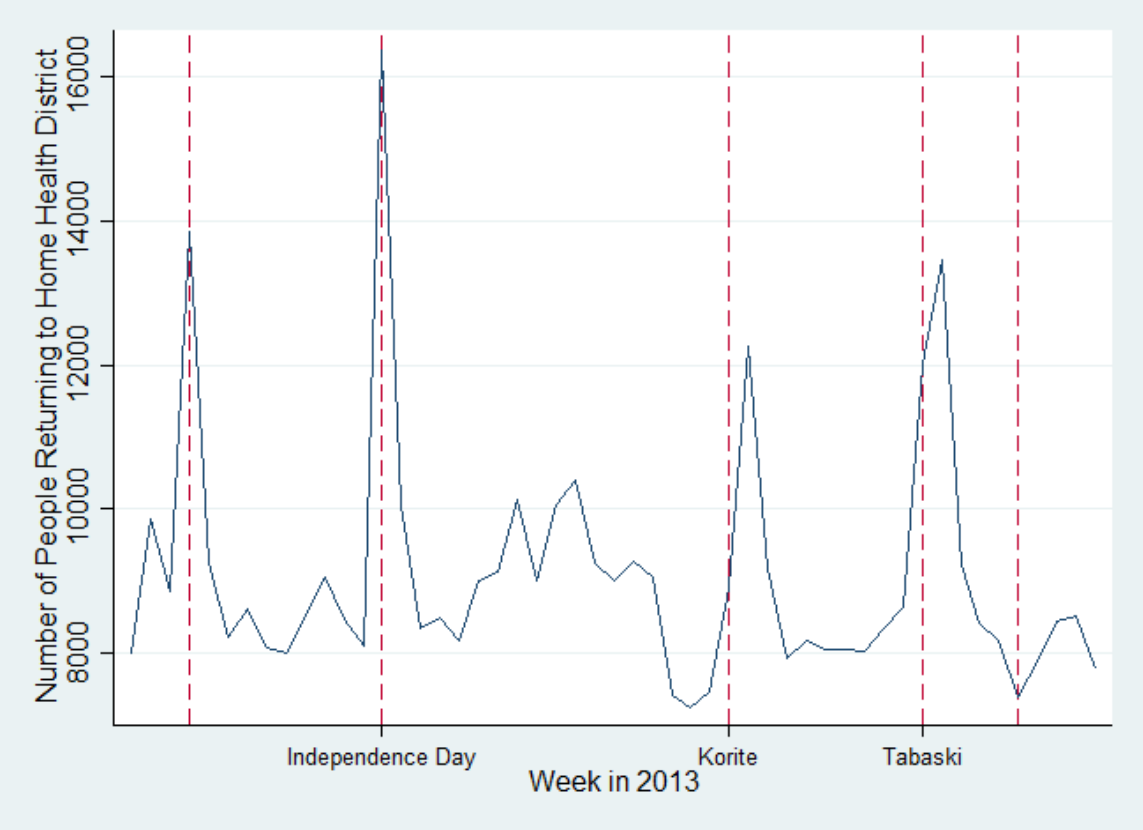


Figure 5: Total Number of People Returning to "Home" Health District Each Week

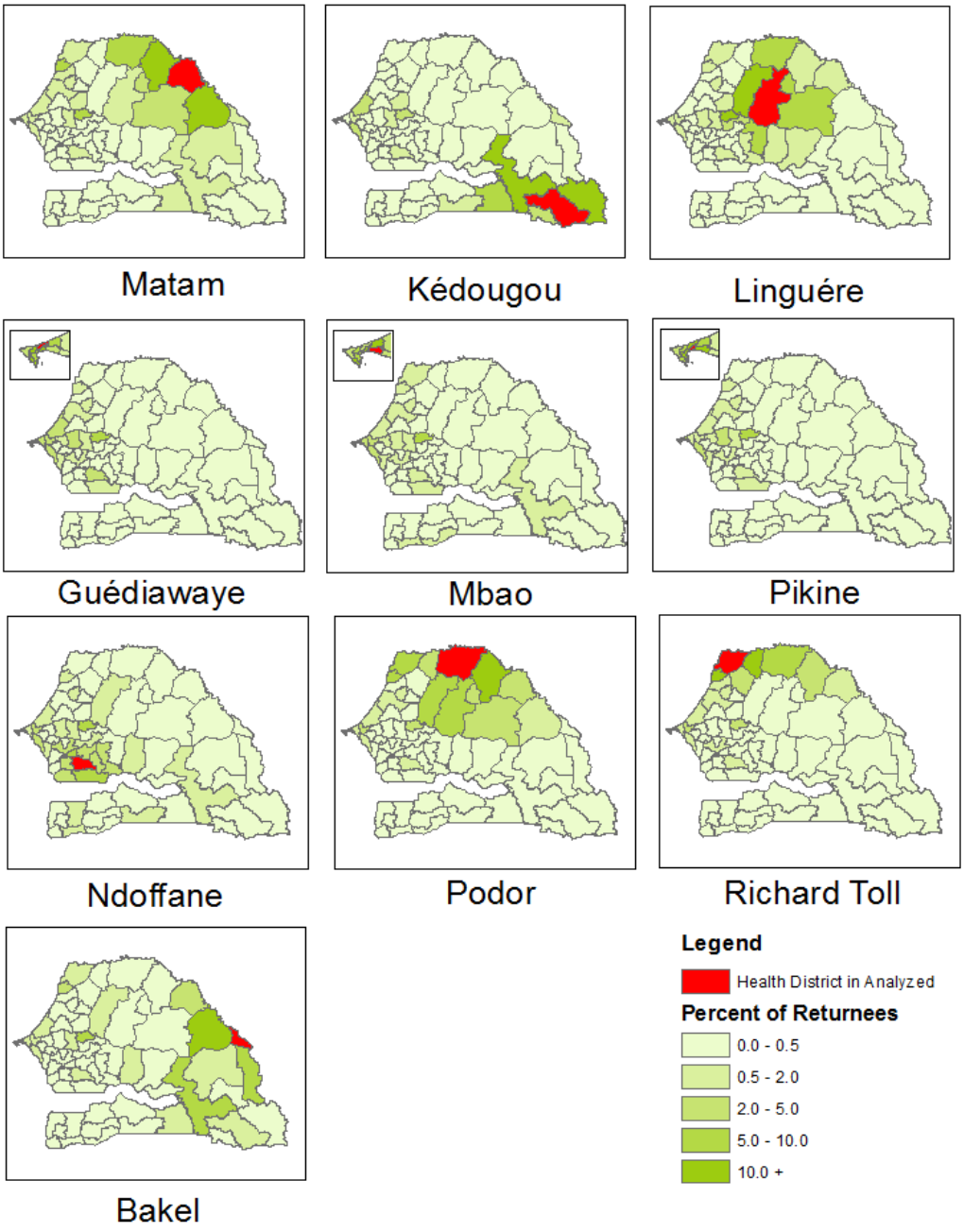


Figure 6: Percent of People Returning from Each Health District to Each Health District Analyzed, 2013

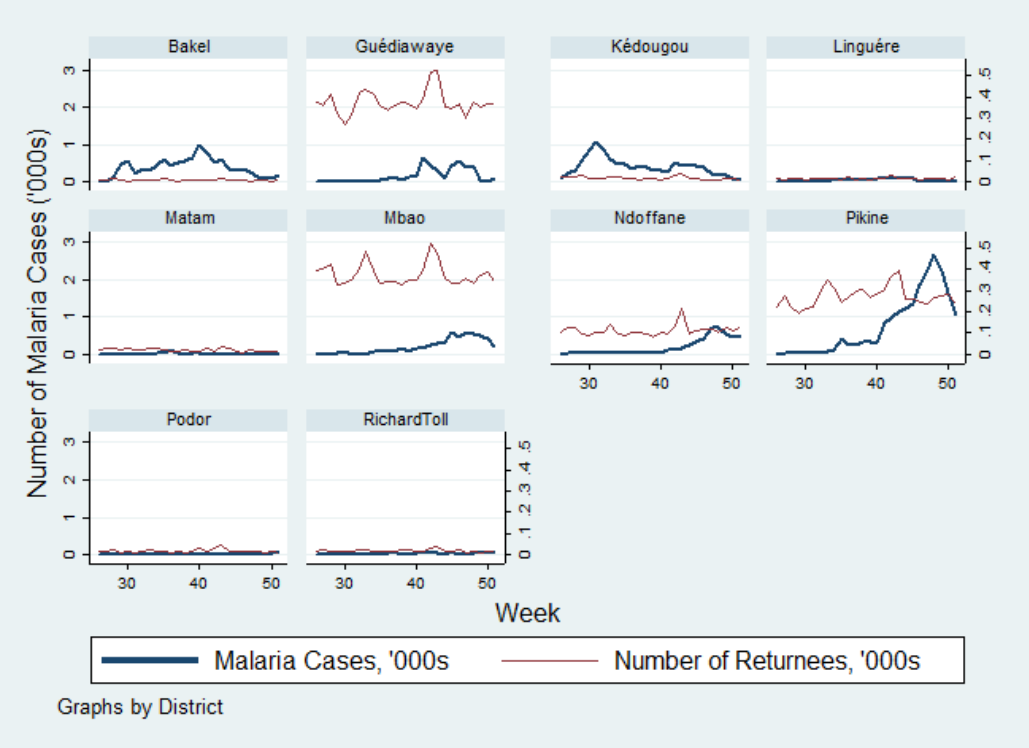


Figure 7: Weekly Cases of Malaria and People Returning Home from High Malaria Places During the Malaria Season

	(1)	(2) Include Low Malaria Return	(3) Include Lagged Malaria	(4) No Pikine	(5) Weeks 25-48
Return, High Malaria	3.128**	3.120**	1.136**	1.624**	3.708***
	(1.231)	(1.224)	(0.487)	(0.698)	(1.269)
Return, Low Malaria		-0.150			
		(0.145)			
Lagged Malaria Cases			0.865***		
			(0.0596)		
Constant	282.4***	284.8***	58.72	380.3***	371.4***
	(73.24)	(73.43)	(39.51)	(55.38)	(71.01)
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
District Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	230	230	230	207	190
R-squared	0.542	0.545	0.905	0.524	0.560
Number of panels	10	10	10	9	10

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

	(1)	(2) Include Low Malaria Return	(3) Include Lagged Malaria	(4) No Pikine	(5) Weeks 25-48
Visit, High Malaria	-0.598	-0.660	0.262	0.410	-0.0790
	(0.916)	(0.917)	(0.347)	(0.727)	(0.863)
Visit, Low Malaria		0.0893			
		(0.0841)			
Lagged Malaria Cases			0.887***		
			(0.0629)		
Constant	303.0***	298.1***	54.76	394.6***	379.0***
	(77.07)	(76.59)	(41.39)	(56.12)	(79.38)
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
District Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	230	230	230	207	190
R-squared	0.506	0.508	0.901	0.500	0.498
Number of panels	10	10	10	9	10

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

	(1)	(2) Include Lagged Malaria	(3) No Pikine	(4) Weeks 25-48
L4.tot_lowreturn	-0.155	-0.0526	-0.0849	-0.142
	(0.148)	(0.0733)	(0.0911)	(0.148)
L.district_cases		0.883***		
		(0.0633)		
Constant	298.5***	59.67	400.1***	380.5***
	(75.89)	(41.57)	(56.15)	(78.64)
Time Fixed Effects	Yes	Yes	Yes	Yes
District Fixed Effects	Yes	Yes	Yes	Yes
Observations	230	230	207	190
R-squared	0.508	0.901	0.500	0.501
Number of panels	10	10	9	10

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

Lags	(1) Returnee Coef.	(2) Lagged Malaria Cases	(3) Constant
No Lag	-0.0656	0.893***	33.90
	(0.149)	(0.0575)	(31.50)
1 Lag	0.0224	0.891***	34.06
	(0.491)	(0.0580)	(31.45)
2 Lags	0.617	0.886***	37.83
	(0.494)	(0.0595)	(32.97)
3 Lags	1.112**	0.876***	85.04*
	(0.478)	(0.0598)	(34.16)
4 Lags	1.136**	0.865***	57.14
	(0.487)	(0.0596)	(35.39)
5 Lags	0.925*	0.850***	5.464
	(0.503)	(0.0613)	(37.64)
6 Lags	0.0728	0.856***	26.69
	(0.527)	(0.0682)	(37.21)
7 Lags	-0.217	0.861***	-2.094
	(0.533)	(0.0728)	(38.40)
8 Lags	0.212	0.847***	49.19
	(0.571)	(0.0783)	(39.60)

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

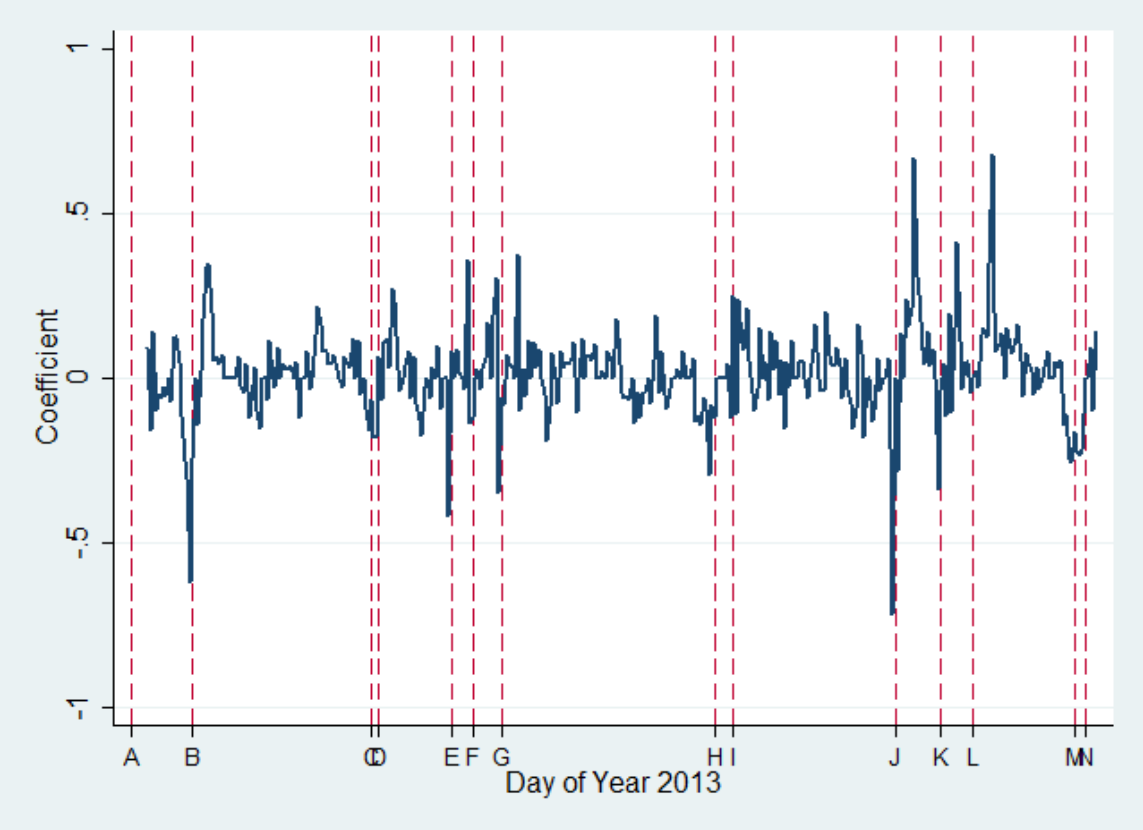


Figure 8: Coefficients from Regressions for Each Day in 2013 of Net Movement Between Arrondissements in Previous 7 days on Long Term Net Migration.

Note: The following holidays are marked on the graph with vertical dash lines-A. New Years Day/Magal de Touba B. The Prophet's Birthday C. Easter Monday D. Independence Day E. Labour Day F. Ascension Day G. Whit Monday H. Koritè I. Assumption Day J. Tabaski K. All Saint's Day L.Tamkharit M. Magal de Touba N. Christmas Day.

H02

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

Lorenzo Mari[1], Renato Casagrandi[1], Manuela Ciddio[1],
Susanne H. Sokolow[2], Giulio De Leo[2], Marino Gatto[1,⋆]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria,
Politecnico di Milano, Italy

[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 (Thétiot-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 (Fenwick, 2012). These figures likely make schistosomiasis one of the most common parasitic diseases (second after malaria) and the deadliest among neglected tropical disesases. 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

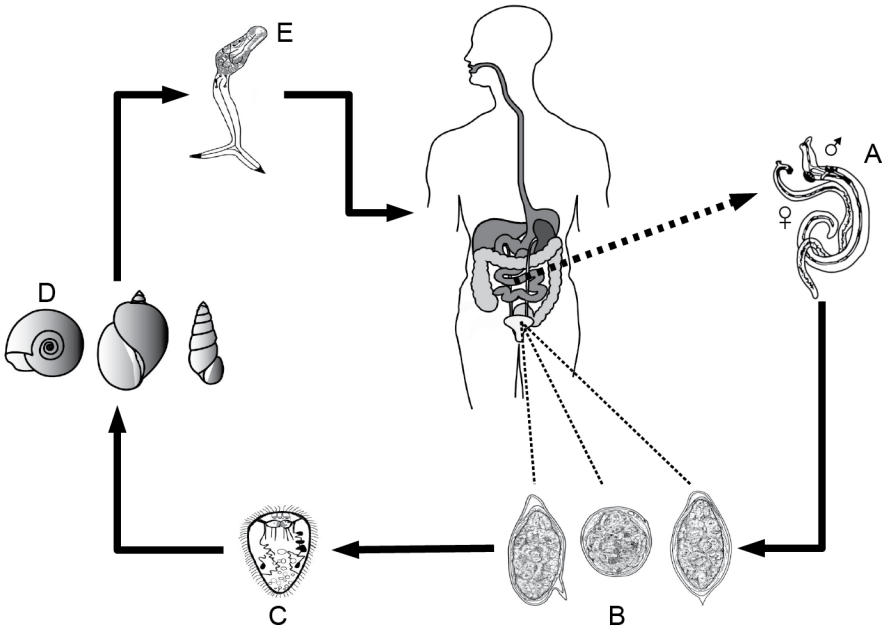


Figure 1: Schistosomiasis transmission cycle. Paired adult worms within human hosts (A) produce eggs (B; left to right: *S. mansoni*, *S. japonicum*, *S. haematobium*) that are shed through faeces or urine and hatch into miracidia (C). Miracidia infect intermediate snail hosts (D; left to right: genus *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; Clennon 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 form 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 (Criscione 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 aquired, 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_i^p 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):

$$\begin{aligned}
\frac{dH_i^0}{dt} &= \mu_H(H_i - H_i^0) - \mathcal{F}_i H_i^0 + \gamma^1 H_i^1 \\
\frac{dH_i^p}{dt} &= \mathcal{F}_i H_i^{p-1} - (\mu_H + \alpha_H^p + \mathcal{F}_i + \gamma^p) H_i^p + \gamma^{p+1} H_i^{p+1} \quad [0 < p < P] \\
\frac{dH_i^P}{dt} &= \mathcal{F}_i H_i^{P-1} - (\mu_H + \alpha_H^P + \gamma^P) H_i^P \\
\frac{dS_i}{dt} &= \mu_S(N_i - S_i) - bM_i S_i \\
\frac{dI_i}{dt} &= bM_i S_i - (\mu_S + \alpha_S) I_i \\
\frac{dC_i}{dt} &= \pi_C I_i - \mu_C C_i \\
\frac{dM_i}{dt} &= \mathcal{G}_i - \mu_M M_i \ ,
\end{aligned} \tag{1}$$

where $\mathcal{F}_i = a \sum_{j=1}^{n} Q_{ij} \theta_j C_j$, $\mathcal{G}_i = \pi_M \delta_i \sum_{j=1}^{n} Q_{ji} \frac{\mathcal{W}_j}{2}$, and $\mathcal{W}_j = \sum_{p=1}^{P} p H_j^p$. 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, μ_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, $\mathcal{F}_i$ is the force of infection for the inhabitants of community i: $\mathbf{Q} = [Q_{ij}]$ 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), θ_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 γ^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 μ_P being the per capita parasite mortality rate. Disease-related mortality in humans is accounted for by the term α_H^p, which describes increasing mortality for increasing parasite burden ($\alpha_H^p = p\alpha_H$, where α_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, μ_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 α_S. As for the dynamics of larval stages, cercariae are shed by infective snails at rate π_C and die at rate μ_C. Similarly, miracidia are shed by infected human hosts and die at rate μ_M; specifically, the total contamination rate for community i is $\mathcal{G}_i$, with π_M being the shedding rate of miracidia by infected humans, δ_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 $\mathcal{W}_j/2$ carried by the residents of human community j). Note that if $\mathbf{Q}$ 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 trasmission 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 Δ. If we also introduce the rescaled state variables

$$h_i^k = \frac{H_i^k}{H_i}\,, \qquad s_i = \frac{S_i}{N_i}\,, \qquad y_i = \frac{I_i}{N_i}\,,$$

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

$$\begin{aligned}
\frac{dh_i^0}{dt} &= \mu_H(1 - h_i^0) - F_i h_i^0 + \gamma^1 h_i^1 \\
\frac{dh_i^k}{dt} &= F_i h_i^{k-1} - (\mu_H + \alpha_H^k + F_i + \gamma^k) h_i^k + \gamma^{k+1} h_i^{k+1} \quad [0 < k < K] \\
\frac{dh_i^K}{dt} &= F_i h_i^{K-1} - (\mu_H + \alpha_H^K + \gamma^K) h_i^K \\
\frac{ds_i}{dt} &= \mu_S(1 - s_i) - G_i s_i \\
\frac{dy_i}{dt} &= G_i s_i - (\mu_S + \alpha_S) y_i\,,
\end{aligned} \tag{2}$$

with

$$F_i = \sum_{j=1}^{n} Q_{ij}\beta_j N_j y_j\,, \qquad \beta_j = \frac{a}{\Delta}\frac{\pi_C}{\mu_C}\theta_j\,, \qquad \gamma^k = \frac{p_k}{\Delta}\mu_P\,, \qquad \alpha_H^k = p_k\alpha_H\,,$$

$$G_i = \chi_i \sum_{j=1}^{n} Q_{ji}W_j\,, \qquad \chi_i = \frac{b}{2}\frac{\pi_M}{\mu_M}\delta_i\,, \qquad W_j = H_j\sum_{k=1}^{K} p_k h_j^k\,.$$

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, 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/`). 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`). 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 $\mathbf{Q} = [Q_{ij}]$ represents the probability that people usually living in community (arrondissement) i come in contact with freshwater in community (arrondissement) j ($j = 1..123$, 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

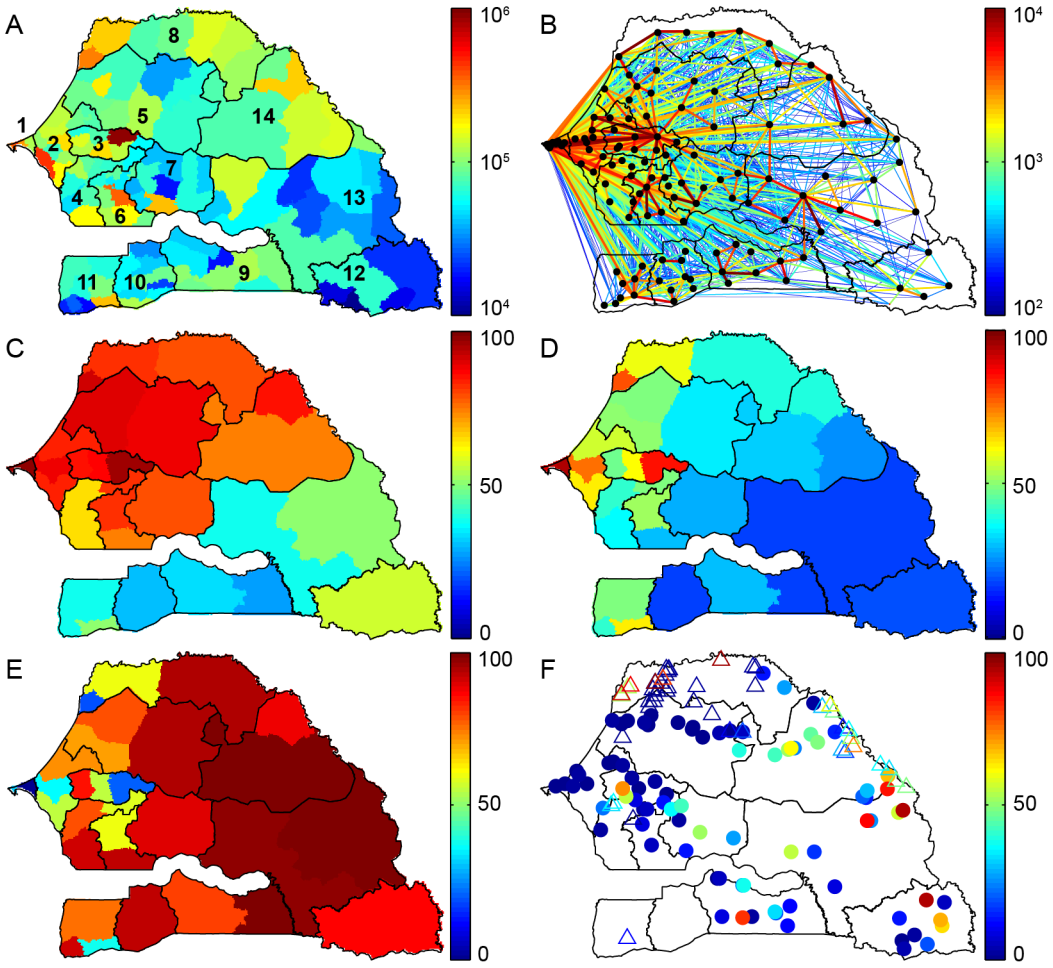


Figure 2: Data for model set-up. A) Population distribution at arrondissement level [number of inhabitants]. Black lines indicate regional borders. Regions are numbered as follows: 1–Dakar, 2–Thiès, 3–Diourbel, 4–Fatick, 5–Louga, 6–Kaolack, 7–Kaffrine, 8–Saint-Louis, 9–Kolda, 10–Sédhiou, 11–Ziguinchor, 12–Kédougou, 13–Tambacounda, 14–Matam. B) Human mobility fluxes [number of people]; the fluxes Φ_{ij} between any two arrondissements (say i and j) are obtained as $\Phi_{ij} = H_i P_{ij}$. Only fluxes ≥ 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 $\mathbf{Q}$ are expected to be roughly proportial 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 β_i and χ_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 `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 lacks access to safe water supplies and improved sanitation, respectively.

Exposure and contamination rates can thus be expressed as

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

respectively, where β_0 and χ_0 are the maximum exposure/contamination rates, ω_i and σ_i represent the fraction of individuals living in arrondissement i with access to water/sanitation, and ϕ_β and ϕ_χ 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 ρ_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

$$\beta_i = \beta_0\rho_i^{\phi_\beta} \qquad \text{and} \qquad \chi_i = \chi_0\rho_i^{\phi_\chi}\,.$$

We retain this latter definition for β_i and χ_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`). 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` 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 ($\mathcal{O}$(100 years)) starting from an initial condition in which human communities are set to be completely uninfected ($h_i^0 = 1$ and $h_i^k = 0$ with $k > 0$ in all arrondissements), while the prevalence of infected snails is tentatively set to be 5% ($s_i = 0.95$ and $y_i = 0.05$ 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 parameters can be reliably estimated from the literature or from epidemiological/demographic records. Specifically, 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 $\mu_H = 4.5 \cdot 10^{-5}$ days^{-1}), snail intermediate hosts (about 1 year according to Feng et al., 2004, hence $\mu_S = 2.7 \cdot 10^{-3}$ days^{-1}) and schistosomes (about 5 years according to Gryseels et al., 2006, hence $\mu_P = 5.5 \cdot 10^{-4}$ 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 β_0 and χ_0, the shape parameter ϕ and the snail density N_0. However, as β_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.

$$RSS = \sum_x (u_x^D - u_x^M)^2 \,,$$

in which u_x^D 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 Kédougou (12), Diourbel (3), Matam (14) and Fatick (4), where the model overestimates (Kédougou 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

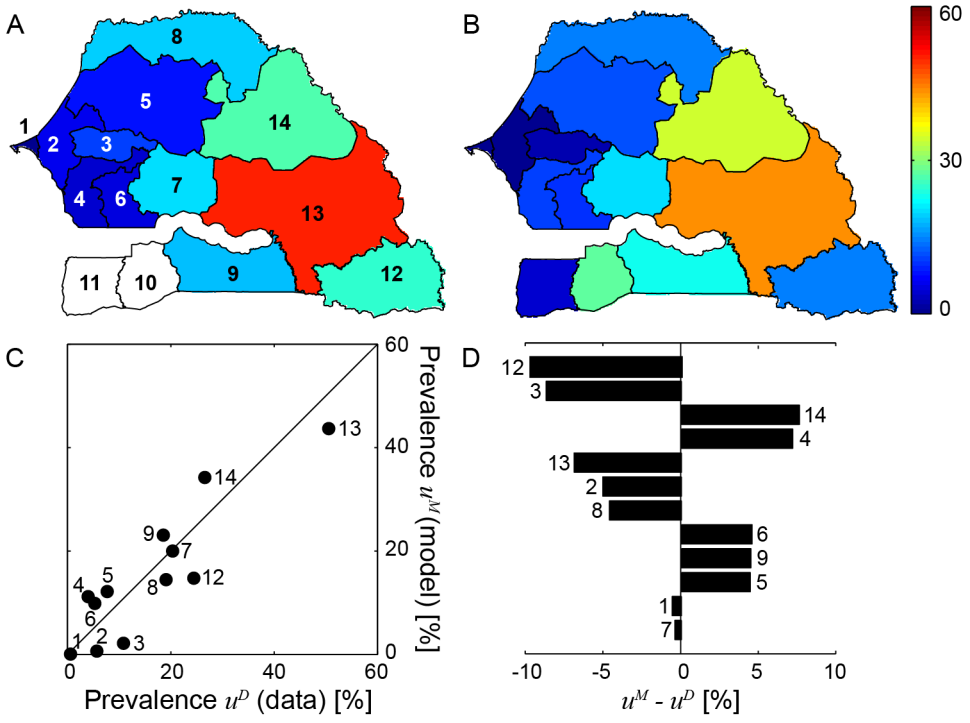


Figure 3: Simulation of model (2) and comparison with epidemiological evidence. A) Regional prevalence of urogenital schistosomiasis [% of infected people] from the 1996 national survey. No data are available for the regions of Sédhiou (10) and Ziguinchor (11). B) Regional schistosomiasis prevalence [% 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.

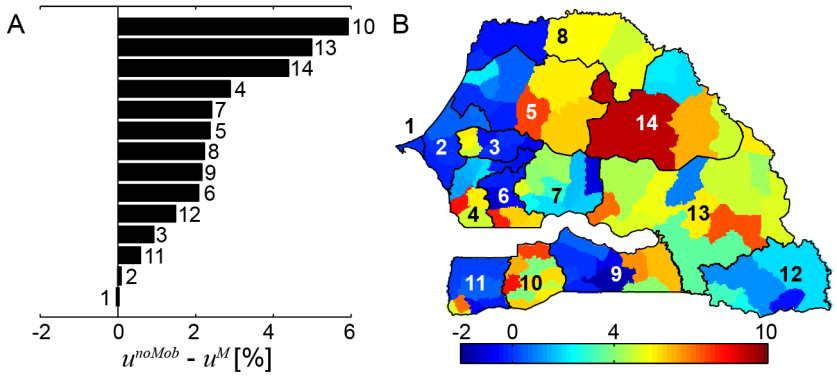


Figure 4: Effects of human mobility on schistosomiasis dynamics. A) Differences in regional disease prevalence obtained without vs. with human mobility (absolute differences sorted in decreasing order). Numbers correspond to administrative regions as shown in panel B. B) Differences [%] in prevalence 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.

Fig. 3, wich accounts for human mobility through matrix **Q** (whose entries are estimated from call detail records), with a simulation in which human mobility is not allowed (i.e. **Q** is set to be 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 ω_i and/or σ_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 ρ_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 τ and η 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 ρ_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 β_0 to β_0' and from χ_0 to χ_0'). Such interventions can be, again, targeted or untargeted. Untargeted interventions can be modeled as

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

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).

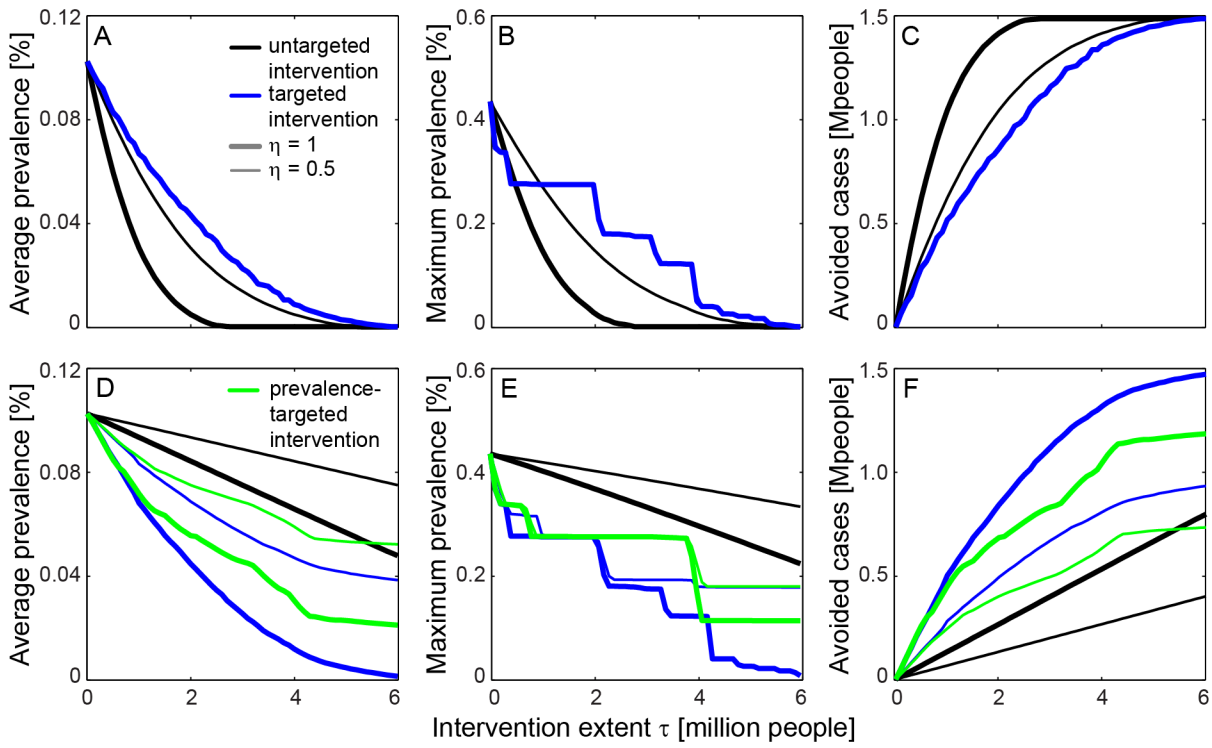


Figure 5: Evaluation of large-scale control strategies. A) Effects of WASH actions aimed at improving community access to safe water and sanitation on population-weighted average schistosomiasis prevalence. B) As in panel A, but for maximum regional schistosomiasis prevalence. C) As in panel A, but for 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. 3). 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

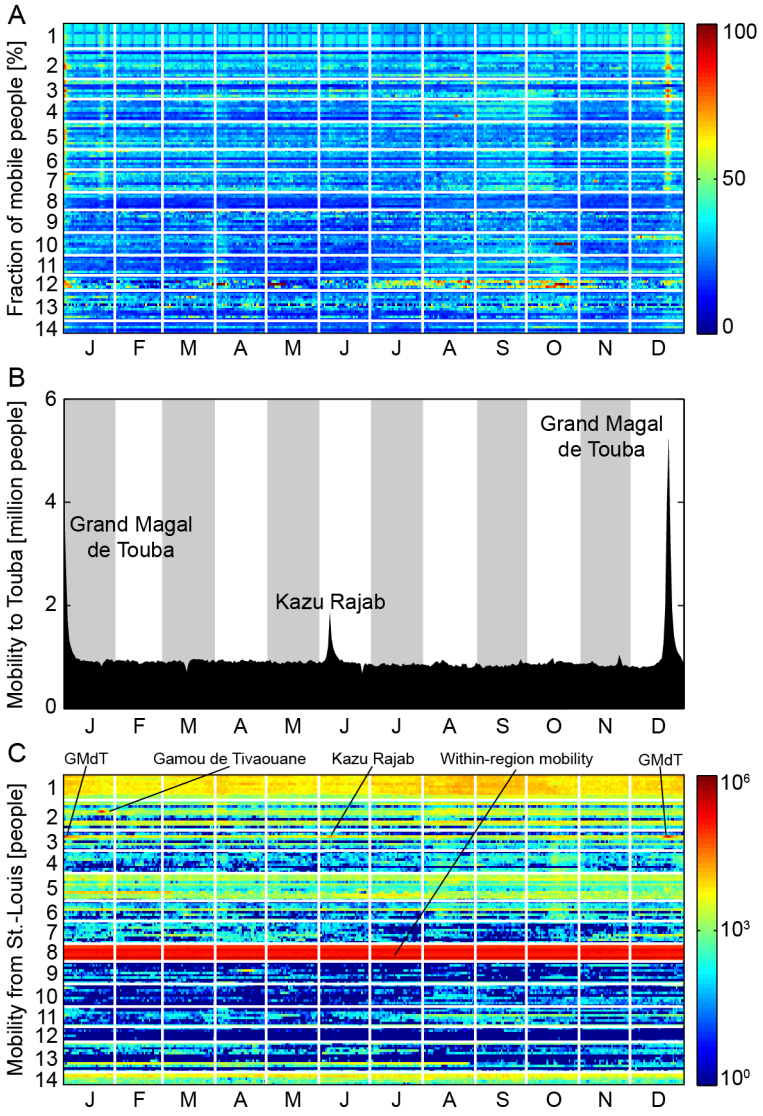


Figure 6: A closer, time-explicit look at human mobility. A) Overall mobility, evaluated as $(1 - Q_{ii})$, $i = 1..123$. Region labels (1..14) are shown for easier visual reference. B) Mobility fluxes to Touba (Ndame arrondissement, Diourbel region), evaluated as $\sum_i \Phi_{i,\mathrm{Ndame}}$, $i = 1..123$. 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 $i \in$ 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 (Thétiot-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

Alkalay, A. S., Rosen, O., Sokolow, S. H., Faye, Y. P. W., Faye, D. S., Aflalo, E. D., Jouanard, N., Zilberg, D., Huttinger, E., and Sag, A. (2014). The prawn *Macrobrachium vollenhovenii* in the Senegal river basin: towards sustainable restocking of all-male populations for biological control of schistosomiasis. *PLoS Neglected Tropical Diseases*, 8:e3060.

Allan, C. and Stankey, G. H. (2009). *Adaptive Environmental Management: a Practitioner's Guide*. Springer/CSIRO Publishing, Collingwood, Australia.

Appleton, C. C., Ngxongo, S. M., Braack, L. E., and le Sueur, D. (1996). *Schistosoma mansoni* in migrants entering South Africa from Moçambique–A threat to public health in north-eastern KwaZulu-Natal? *South African Medical Journal*, 86:350–353.

Bella, H., de C. Marshall, T. F., Omer, A. H. S., and Vaughan, J. P. (1980). Migrant workers and schistosomiasis in the Gezira, Sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 74:36–39.

Belton, V. and Stewart, T. (2002). *Multiple Criteria Decision Analysis: an Integrated Approach*. Kluver Academic Publishers, Dordrecht, the Netherlands.

Briand, V., Watier, L., Le Hesran, J. Y., Garcia, A., and Cot, M. (2005). Co-infection with *Plasmodium falciparum* and *Schistosoma haematobium*: protective effect of schistosomiasis on malaria in Senegalese children? *Americal Journal of Tropical Medicine and Hygiene*, 72:702–707.

Cetron, M. S., Chitsulo, L., Sullivan, J. J., Pilcher, J., Wilson, M., Noh, J., Tsang, V. C., Hightower, A. W., and Addiss, D. G. (1996). Schistosomiasis in Lake Malawi. *Lancet*, 348:1274–1278.

CIA (2014). The World Factbook 2013–14. Technical report, Central Intelligence Agency. Available online at `https://www.cia.gov/library/publications/the-world-factbook/index.html`.

Clennon, J. A., King, C. H., Muchiri, E. M., and Kitron, U. (2007). Hydrological modelling of snail dispersal patterns in Msambweni, Kenya and potential resurgence of *Schistosoma haematobium* transmission. *Parasitology*, 134:683–693.

Colley, D. G., Bustinduy, A. L., Secor, W. E., and King, C. H. (2014). Human schistosomiasis. *Lancet*, 383:2253–2264.

Criscione, C. D., Anderson, J. D., Sudimack, D., Subedi, J., Upadhayay, R. P., Jha, B., Williams, K. D., Williams-Blangero, S., and Anderson, T. J. C. (2010). Landscape genetics reveals focal transmission of a human macroparasite. *PLoS Neglected Tropical Diseases*, 4:e665.

de Montjoye, Y. A., Smoreda, Z., Trinquart, R., Ziemlicki, C., and Blondel, V. D. (2014). D4D-Senegal: the second mobile phone data for development challenge. *Computing Research Repository*, abs/1407.4885.

Feng, Z., Eppert, A., Milner, F. A., and Minchella, D. J. (2004). Estimation of parameters governing the transmission dynamics of schistosomes. *Applied Mathematics Letters*, 17:1105–1112.

Fenwick, A. (2012). The global burden of neglected tropical diseases. *Public Health*, 126:233–236.

Garcia, A. J., Pindolia, D. K., Lopiano, K. K., and Tatem, A. J. (2014). Modeling internal migration flows in sub-Saharan Africa using census microdata. *Migration Studies*, in press.

Gilks, W. R., Richardson, S., and Spiegelharter, D. J. (1995). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York, USA.

González, M. C., Hidalgo, C. A., and Barabási, A. L. (2008). Understanding individual human mobility patterns. *Nature*, 453:479–482.

Grimes, J. E. T., Croll, D., Harrison, W. E., Utzinger, J., Freeman, M. C., and Templeton, M. R. (2014). The relationship between water, sanitation and schistosomiasis: a systematic review and meta-analysis. *PLoS Neglected Tropical Diseases*, 8:e3296.

Gryseels, B., Polman, K., Clerinx, J., and Kestens, L. (2006). Human schistosomiasis. *Lancet*, 368:1106–1118.

Gurarie, D., King, C. H., and Wang, X. (2010). A new approach to modelling schistosomiasis transmission based on stratified worm burden. *Parasitology*, 137:1951–1965.

Gurarie, D. and Seto, E. Y. W. (2009). Connectivity sustains disease transmission in environments with low potential for endemicity: modelling schistosomiasis with hydrologic and social connectivities. *Journal of the Royal Society Interface*, 6:495–508.

Kheir, M. M., Eltoum, I. A., Saad, A. M., Ali, M. M., Baraka, O. Z., and Homeida, M. M. A. (1999). Mortality due to schistosomiasis *mansoni*: a field study in Sudan. *The American Journal of Tropical Medicine and Hygiene*, 60:307–310.

Kloos, H., Correa-Oliveira, R., dos Reis, D. C., Rodrigues, E. W., Monteiro, L. A. S., and Gazzinelli, A. (2010). The role of population movement in the epidemiology and control of schistosomiasis in Brazil: a preliminary typology of population movement. *Memórias do Instituto Oswaldo Cruz*, 105:578–586.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112–147.

Li, Y. S., Raso, G., Zhao, Z. Y., He, Y. K., Ellis, M. K., and McManus, D. P. (2007). Large water management projects and schistosomiasis control, Dongting Lake Region, China. *Emerging Infectious Diseases*, 13:973–979.

Lowe, D., Xi, J., Meng, X., Wu, Z., Qiu, D., and C., S. R. (2005). Transport of *Schistosoma japonicum* cercariae and the feasibility of niclosamide for cercariae control. *Parasitology International*, 54:83–89.

Lozano, R., Naghavi, M., K., F., Lim, S., Shibuya, K., and Aboyans, V. *et al.*. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, 380:2095–2128.

Lu, X., Wetter, E., Bharti, N., Tatem, A. J., and Bengtsson, L. (2013). Approaching the limit of predictability in human mobility. *Scientific Reports*, 3:2923.

Macdonald, G. (1965). The dynamics of helminth infections, with special reference to schistosomes. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 59:489–506.

Mari, L., Bertuzzo, E., Righetto, L., Casagrandi, R., Gatto, M., Rodriguez-Iturbe, I., and Rinaldo, A. (2012). Modelling cholera epidemics: the role of waterways, human mobility and sanitation. *Journal of the Royal Society Interface*, 9:376–388.

Maszle, D. R., Whitehead, P. G., Johnson, R. C., and Spear, R. C. (1998). Hydrological studies of schistosomiasis transport in Sichuan Province, China. *Science of the Total Environment*, 216:193–203.

Ndir, O. (2000). Situation des schistosomoses au Sénégal. In Chippaux, J. P., editor, *La Lutte contre les Schistosomoses en Afrique de l'Ouest*, pages 225–236. IRD Editions.

Ndir, O., Faye, O., Dieng, Y., Dieng, T., and Diallo, S. (1996). National study of bilharzia in Senegal. Technical report, Cheikh Anta Diop University, Dakar, Senegal.

Ogden, S., Gallo, K., Davis, S., McGuire, C., Meyer, E., Addiss, D., and Haddad, D. (2014). WASH and the neglected tropical diseases. A manual for WASH implementers. Senegal. Technical report, WASH NTD. Available online at `http://www.washntds.org/`.

Palchykov, V., Mitrović, M., Jo, H. H., Saramäki, J., and Pan, R. K. (2014). Inferring human mobility using communication patterns. *Scientific Reports*, 4:6174.

Picquet, M., Ernould, J. C., Vercruysse, J., Southgate, V. R., Mbaye, A., Sambou, B., Niang, M., and Rollinson, D. (1996). The epidemiology of human schistosomiasis in the Senegal river basin. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 90:340–346.

Poda, J. N., Sellin, B., Sawadogo, L., and Sanogo, S. (1994). Distribution spatiale des mollusques hôtes intermédiaires potentiels des schistosomes et de leurs biotopes au Burkina Faso. *OCCGE INFO*, 101:12–19.

Poda, J. N., Sondo, B., and Parent, G. (2003). Influence des hydro-aménagements sur la distribution des bilharzioses et de leurs hôtes intermédiaires au Burkina Faso. *Cahiers d'Études et de Recherches Francophones / Santé*, 13:49–53.

Poda, J. N., Traoré, A., and Sondo, B. K. (2004). L'endémie bilharzienne au Burkina Faso. *Société de Pathologie Exotique*, 97:47–52.

Pullan, R. L., Freeman, M. C., Gething, P. W., and Brooker, S. J. (2014). Geographical inequalities in use of improved drinking water supply and sanitation across sub-Saharan Africa: mapping and spatial analysis of cross-sectional survey data. *PLoS Medicine*, 11:e1001626.

Remais, J. (2010). Modelling environmentally-mediated infectious diseases of humans: transmission dynamics of schistosomiasis in China. In Michael, E. and Spear, R., editors, *Modelling Parasite Transmission and Control*, pages 79–98. Springer.

Rollinson, D., Knopp, S., Levitz, S., Stothard, J. R., Tchuem Tchuenté, L. A., Garba, A., Mohammed, K. A., Schur, N., Person, B., Colley, D. G., and Utzinger, J. (2013). Time to set the agenda for schistosomiasis elimination. *Acta Tropica*, 128:423–440.

Schur, N., Hurlimann, E., Garba, A., Traoré, M. S., Ndir, O., Ratard, R. C., Tchuem Tchuenté, L. A., Kristensen, T. K., Utzinger, J., and Vounatsou, P. (2011). Geostatistical model-based estimates of schistosomiasis prevalence among individuals aged ≤ 20 years in West Africa. *PLoS Neglected Tropical Diseases*, 5:e1194.

Simini, F., González, M. C., Maritan, A., and Barabási, A. L. (2012). A universal model for mobility and migration patterns. *Nature*, 484:96–100.

Sokolow, S. H., Lafferty, K. D., and Kuris, A. M. (2014). Regulation of laboratory populations of snails (*Biomphalaria* and *Bulinus* spp.) by river prawns, *Macrobrachium* spp. (Decapoda, Palaemonidae): implications for control of schistosomiasis. *Acta Tropica*, 132:64–74.

Song, C., Koren, T., Wang, P., and Barabási, A. L. (2010a). Modelling the scaling properties of human mobility. *Nature Physics*, 6:818–823.

Song, C., Qu, Z., Blumm, N., and Barabási, A. L. (2010b). Limits of predictability in human mobility. *Science*, 327:1018–1021.

Steinmann, P., Keiser, J., Bos, R., Tanner, M., and Utzinger, J. (2006). Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *The Lancet Infectious Diseases*, 7:411–426.

Stensgaard, A. S., Utzinger, J., Vounatsou, P., Hürlimann, E., Schur, N., Saarnak, C. F. L., Simoonga, C., Mubita, P., Kabatereine, N. B., Tchuem Tchuenté, L. A., Rahbek, C., and Kristensen, T. K. (2013). Large-scale determinants of intestinal schistosomiasis and intermediate host snail distribution across Africa: does climate matter? *Acta Tropica*, 128:378–390.

Talla, I., Kongs, A., Verle, P., Belot, J., Sarr, S., and Coll, A. M. (1990). Outbreak of intestinal schistosomiasis in the Senegal river basin. *Annales de la Societé Belge de Medicine Tropicale*, 70:173–180.

Tatem, A. J., Huang, Z., Narib, C., Kumar, U., Kandula, D., Pindolia, D. K., Smith, D. L., Cohen, J. M., Graupe, B., Uusiku, P., and Lourenço, C. (2014). Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malaria Journal*, 13:52.

Thétiot-Laurent, S. A., Boissier, J., Robert, A., and Meunier, B. (2013). Schistosomiasis chemotherapy. *Angewandte Chemie*, 52:7936–7956.

Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C. M., Blondel, V., Smoreda, Z., González, M. C., and Colizza, V. (2014). On the use of human mobility proxies for modeling epidemics. *PLoS Computational Biology*, 10:e1003716.

Tzanetou, K., Adamis, G., Andipa, E., Zorzos, C., Ntoumas, K., Armenis, K., Kontogeorgos, G., Malamou-Lada, E., and Gargalianos, P. (2007). Urinary tract *Schistosoma haematobium* infection: a case report. *Journal of Travel Medicine*, 14:334–337.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. (2012). Quantifying the impact of human mobility on malaria. *Science*, 6104:267–270.

Wesolowski, A., Stresman, G., Eagle, N., Stevenson, J., Owaga, C., Marube, E., Bousema, T., Drakeley, C., Cox, J., and Buckee, C. O. (2014). Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. *Scientific Reports*, 4:5678.

WHO (2014). Schistosomiasis. Fact sheet n. 115. Technical report, Available online at `http://www.who.int/mediacentre/factsheets/fs115/en/`.

Woolhouse, M. E. J. and Chandiwana, S. K. (1990a). Population biology of the freshwater snail *Bulinus globosus* in the Zimbabwe highveld. *Journal of Applied Ecology*, 27:41–59.

Woolhouse, M. E. J. and Chandiwana, S. K. (1990b). Population dynamics model for *Bulinus globosus*, intermediate host for *Schistosoma haematobium*, in river habitats. *Acta Tropica*, 27:151–160.

H11

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

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

[a] *Faculté des Sciences de Tunis University of Tunis, Tunisa*
[b] *ERIAS INSERM U897,* ISPED, *University of Bordeaux, F-33000, France*
[c] *Virginia Modeling Analysis and Simulation, Old Dominion University, VA, USA*
[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 Infaraction 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.

Takling 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[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 totalizing 196,712 km^2. The country is subdivised 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 busness 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

[1] 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 Decembre 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):

Table 1: Main figures about population in Senegal

Region Number on the map	Name of the region	Number of males	Number of females	Global Population	Area (km²)	Density (/km²)
1	Dakar	1 579 020	1 558 176	3 137 196	547	5735.3
2	Thies	896 572	892 292	1 788 864	6670	268.2
3	Diourbel	716 460	780 995	1 497 455	4824	310.4
4	Kaolack	474 404	486 471	960 875	5357	179.4
5	Saint-Louis	453 315	455 627	908 942	19241	47.2
6	Louga	433 715	440 478	874 193	24889	35.1
7	Fatick	353 716	360 676	714 392	6849	104.3
8	Tambacounda	344 475	336 835	681 310	42364	16.1
9	Kolda	335 018	327 437	662 455	13771	48.1
10	Kaffrine	282 093	284 899	566 992	11262	50.3
11	Matam	276 481	286 058	562 539	29445	19.1
12	Ziguinchor	281 813	267 338	549 151	7352	74.7
13	Sedhiou	229 468	223 526	452 994	7341	61.7
14	Kedougou	78 867	72 490	151 357	16800	9.0
	Total	**6 735 417**	**6 773 298**	**13 508 715**	**196 712**	**68.7**

- 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.

Table 2: Example of voice traffic between antennas

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

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 3: example of dataset 2

user	timestamp	site
1	18/03/2013 21:30	716
1	18/03/2013 21:40	718
1	19/03/2013 20:40	716
1	19/03/2013 20:40	716
1	19/03/2013 20:40	716
1	19/03/2013 20:40	716
1	19/03/2013 21:00	716
1	19/03/2013 21:30	718
1	20/03/2013 09:10	705
1	21/03/2013 13:00	705

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 routinelly 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 tesselation 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 polygonial 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 tessalation generated for the provided 1,666 cell sites in Senegal is shown in Figure 1.

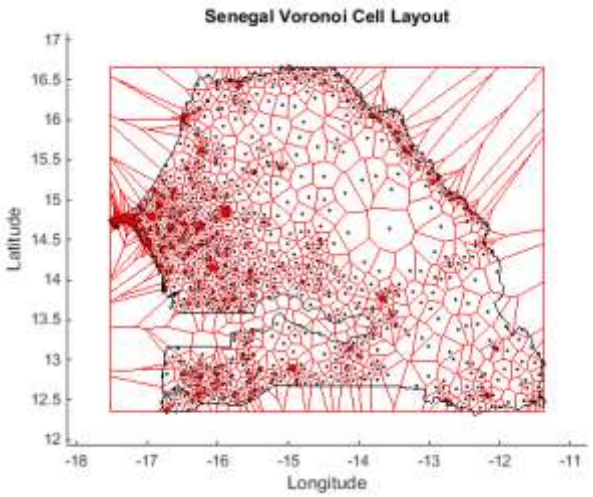


Figure 1 Voronoi cell layout for Senegal based on provided 1,666 site locations

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 agregate 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 ajust 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.

$$\alpha = U_S * 1.2 * {}^{1}/_{Oms}$$

$$\beta = \frac{U_R}{O_R}$$

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 survey[2].
- 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 Sedhiou region and even 509.43 for the Kedougou 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

Site	Region	Unique User/Day
1	Dakar	160,24
1583	Tambacounda	170,05
1405	Saint-Louis	181,04

[2] http://www.gsma.com/

Figure 2 represents distribution of antenna as well as the computed density of the population across the country. Lets assume that N represente 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 1,001<N<10,000, brown for 10001<N<100,000 and black for N > 100,001.

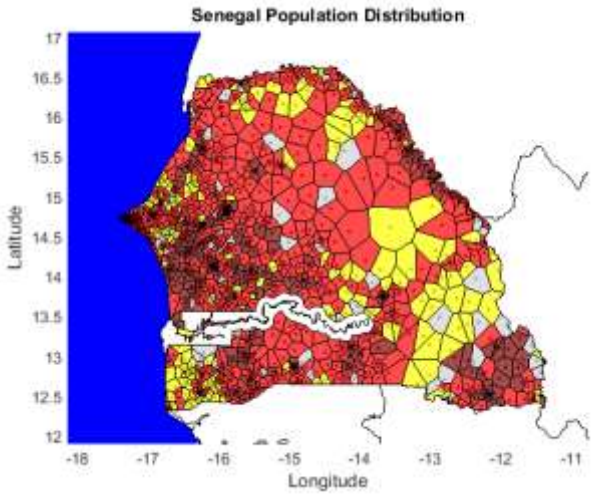


Figure 2: Distribution of Senegal population according to the antennas

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 stokes 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 involes mapping the hospital and populated distribution on to the cellular network layout. A total of 40 hospitals located across the 14 regions of Senegal where 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 where 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.

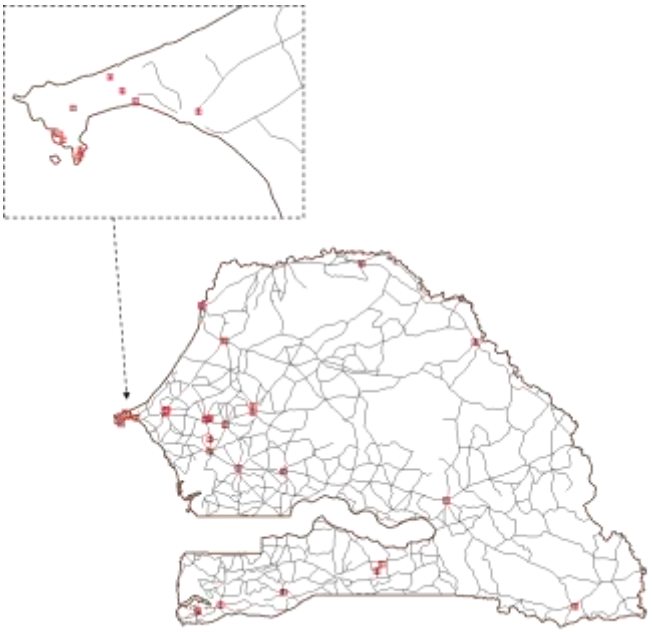


Figure 3: Location of the 40 hospitals considered in the study. They are represented in red symbols while the road network is shown in black lines. The Dakar region is shown inset.

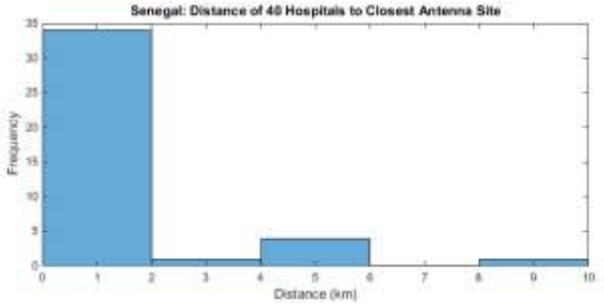


Figure 4 Distance of hospitals to closest antenna sites

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 know 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.

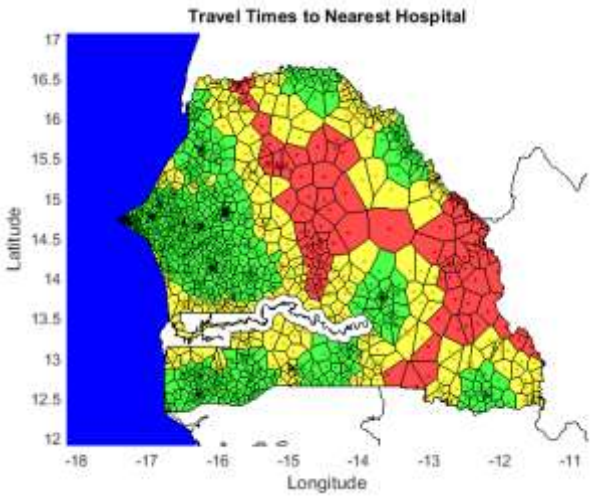


Figure 5 Estimated travel times from different areas to the nearest hospital. (Green: less than 90 minutes, Yellow: 90 to 180 minutes, Red: over 180 minutes).

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.

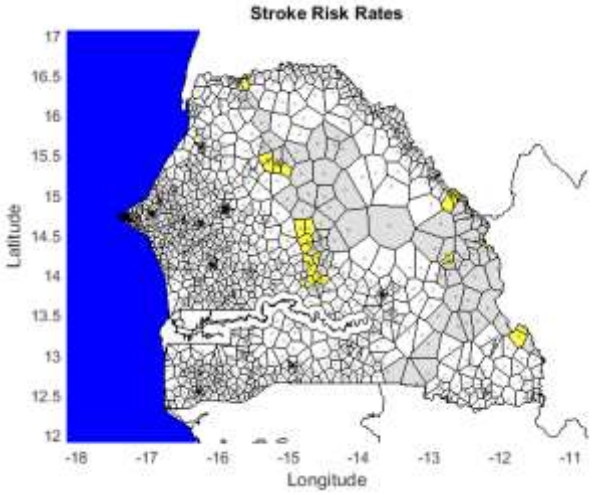


Figure 6. Estimated incident cases of strokes from different areas too far from the nearest hospital to be treated by balloon. (Grey < 5 victims, Yellow 5 to 25).

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 baloon treatment because they cannot reach the hospital in less than 90 minutes (Figure 7).

Discussion

Highlights

Stroke and Myocardial Infarction are two majors issues in Public Health. Several factors contribute to the increasing of the number of the cases annually. This is particulalry true in countries such as Senegal.

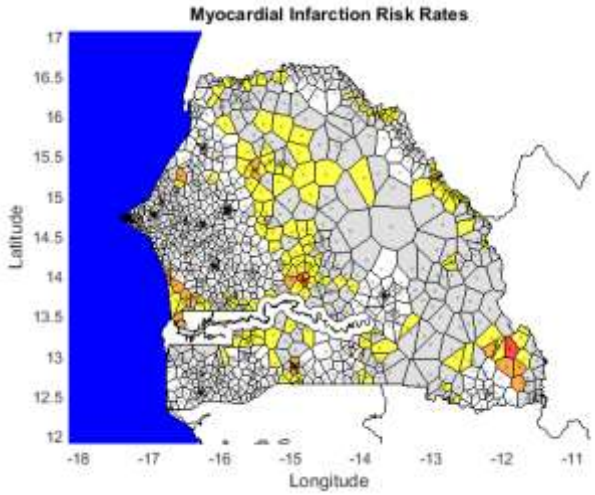


Figure 7. Estimated incident cases of Myocardial Infarction from different areas too far from the nearest hospital to be treated by fibrinolysis. (Grey < 5 victims, Yellow 5 to 25, Orange 26 to 50, Red 51 to 100).

We have shown that by using a considerable amount of recorded mobile data combined with census data it is possible to perform location based estimatimation 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 periode 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 ajust 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 sequelaes (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

- Baert, A.-E. and Seme, D. 2004: Voronoi mobile cellular networks: topological properties, Third International Symposium on/Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks, 5-7 July 2004.
- Boscoe et al. 2013: A Nationwide Comparison of Driving Distance Versus Straight Line Distance to Hospitals, Professional Geographer, 64(2), 2013.
- Dummer, T. J. B. 2008: Health geography: supporting public health policy and planning. CMAJ : Canadian Medical Association Journal, 178(9), 2008.
- FDA, 2013. Paving the Way for Personalized Medicine: FDA's Role in the New Era of Medical Product Development (October 2013)
- Holma H. and Toskala A. 2009: WCDMA for UMTS: HSPA Evolution and LTE, Wiley & Sons Ltd, Chichester, 2009.
- Montjoye et al. 2014 : D4D-Senegal : The Second Mobile Phone Data for Development Challenge. July 2014.
- OSM, 2014: Web Site OpenStreetMap, consulted on December 15, 2014.
- RGPHAE, 2013: Recensement Général de la Population et de l'Habitat
- SD, 2014 : SenDoctor http://sendocteur.com/structureregion.php, consulted on December 16, 2014
- SMD, 2014 : The online Senegal Medical Directory Web Site http://www.annuairemedical-senegal.com, consulted on December 30, 2014
- Shadbolt, N., Hall, W., Berners-Lee, L. "The Semantic Web Revisited", IEEE Intelligent Systems Journal, May/June 2006, 96-101

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

H12

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 know 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 Ebolavirus, family Filoviridae. To date, five different Ebolavirus 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 106 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.**
- **FluTE'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 arrondissements centers and antennas. The antenna-based data set 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).

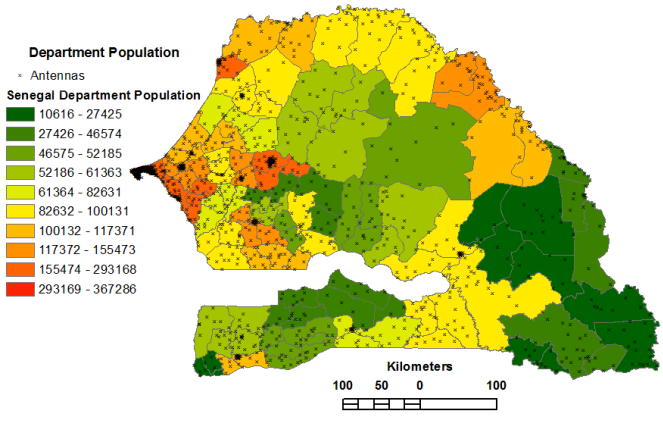


Fig. 1. Population density and sampling locations (antennas). Note: since the population is larger than the random sample for each location (provided by the D4D challenge), the actual population size within each antenna range was projected based on the population of the department (census records) and the relative sample size for each antenna within that department.

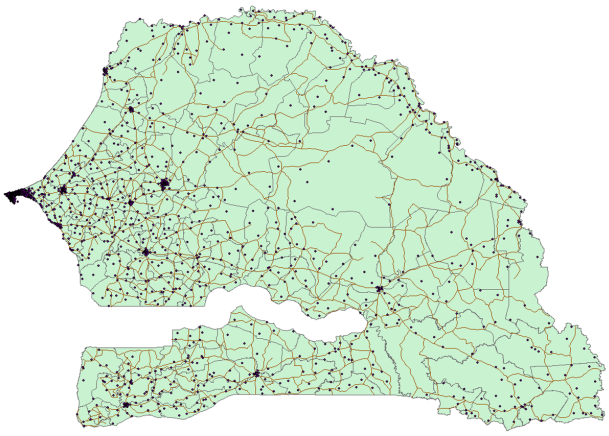


Fig. 2. Geospatial map of the population, roads, and sampling locations (antennas). Movement between the sampling locations provided high-resolution detail of mobility, social contacts, and disease transmission pathways in the population models.

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.

Parameter	Range
Incubation period	2-22 days
Viral Load Trajectories	2 trajectories, one fatal
Case fatality rate	40-70%
Days after death before burial	1-3 days
Simulation length	multiple seasons and years

TABLE I
SOFTWARE MODIFICATIONS REQUIRED TO SIMULATE EBOLA.

Parameter	Range
Secondary infections (R_0)	1.51-2.53
Infectious after death	modified viral load and contacts
Ascertainment delay	3 days after symptomatic
Models	Population, airports, ports, borders, geo-political boundaries, mobility

TABLE II
CONFIGURATION PARAMETERS REQUIRED TO SIMULATE EBOLA.

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 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 boarder 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.

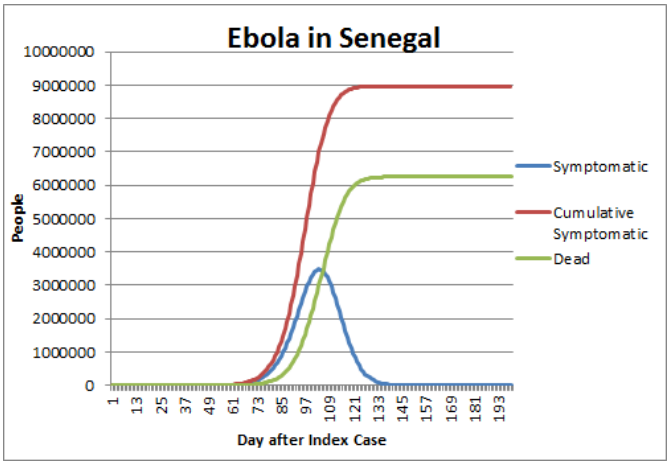


Fig. 3. Simulation results from the D4D-informed model detailing the count of individuals that have been infected by the end of that day (cumulative symptomatic) and currently infected on that day (symptomatic).

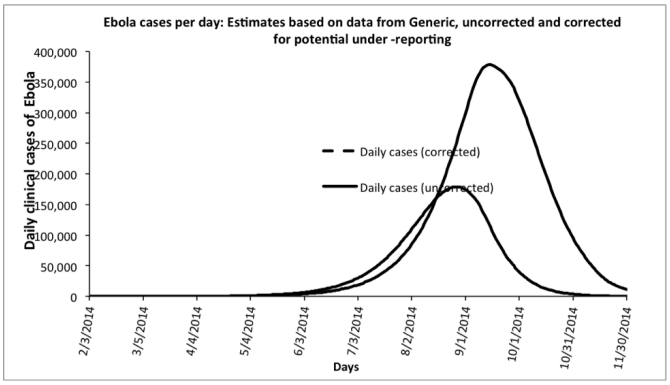


Fig. 4. Simulation results from the CDC model detailing the count of individuals currently infected on a given day.

Number of days infectious	Total cases
10	6,886,748
11	9,370,528
12	10,921,851
13	11,939,709

TABLE III
PREDICTED ATTACK RATES FOR A YEAR-LONG SIMULATION OF EBOLA BY THE CDC MODEL IN A BASELINE SCENARIO (I.E., NO GOVERNMENTAL INTERVENTIONS OR PREVENTATIVE/TREATMENT ACTIONS TAKEN BY INDIVIDUALS).

B. Comparative Mitigation Strategies

Analyzing the results of comparative simulations provides a foundation for isolating and determining the effect of combinations of mitigation strategies. To demonstrate this, simulations were run to determine the impact of a scenario where sick individuals were isolated away from family members within their house in comparison to a baseline scenario without isolation. Figure 5 displays the total infections under each policy. Certain mitigation strategies might delay the onset of the disease (allowing additional time for government intervention), reduce the total number of infections, or prevent the infections from spreading to multiple areas. Figure 6 displays the results from the same simulations based on the number of individuals currently infected on a given day. See Appendix I for a the output of the simulation software aggregating the results per day. The software also provides detailed information on the day each individual is infected, their age, and their location within the country (not included due to size). Other studies might compare the differences in epidemics initially beginning at a major airport in comparison to a border crossings.

VI. DISCUSSION

Anonymous call detail records provide a rich dataset for improving health in developing regions. In this work, a simulation platform was developed to provide predictions of the spread of Ebola throughout Senegal in a variety of potential scenarios. The simulation software is based on a set of models describing Senegal's population, mobility, workforce, travel patterns, and migration. Developing the software required extensive modifications of existing open-source computational epidemiology software. The revised software may be used to conduct comprehensive studies of predicted outcomes regarding public policy, resource allocation, and recommendations for individuals. Further studies that utilize the software are required to analyze potential practices, e.g., the risk associated with opening a currently closed border.

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

[1] C. Althaus. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. In *PLOS Currents: Outbreaks*, September 2014.

[2] A. Ansari. Clinical features and pathobiology of Ebolavirus infection. *J Autoimmun*, 55:1–9, 2014.

[3] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. Marathe. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC '08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, pages 1–12, Piscataway, NJ, USA, 2008. IEEE Press.

[4] K. Bisset, J. Chen, X. Feng, A. Vullikanti, and M. Marathe. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *ICS '09: Proceedings of the 23rd international conference on Supercomputing*, pages 430–439, New York, NY, USA, 2009. ACM.

[5] CDC. Ebola outbreak in West Africa case counts. http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/case-counts.html, Dec. 2014.

[6] CDC. Interim guidance for specimen collection, transport, testing, and submission for patients with suspected infection with Ebola virus disease. http://www.cdc.gov/vhf/ebola/pdf/ebola-lab-guidance.pdf, Dec. 2014.

[7] D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini. FluTE, a Publicly Available Stochastic Influenza Epidemic Simulation Model. *PLoS Comput Biol*, 6(1), 2010.

[8] G. Chowell, N. Hengartner, C. Castillo-Chavez, P. Fenimore, and J. Hyman. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *J Theor Biol*, 229(1):119–126, 2004.

[9] G. Chowell and H. Nishiura. Transmission dynamics and control of Ebola virus disease (EVD): a review. *BMC Medicine*, 12(196), 2014.

[10] S. Dowell, R. Mukunu, T. Ksiazek, A. Khan, P. Rollin, and C. Peters. Transmission of Ebola hemorrhagic fever: a study of risk factors in family members, Kikwit, Democratic Republic of the Congo, 1995. *Journal of Infectious Diseases*, 179:87–91, 1995.

[11] P. Gerland, A. Raftery, H. Sevcikova, and N. Li. World population stabilization unlikely this century. *Science*, 346(6206):234–237, 2014.

[12] D. Grady and S. Fink. Tracing Ebolas breakout to an African 2 year-old. *New York Times*, Aug. 2014.

[13] J. Legrand, R. Grais, P. Boelle, A. Valleron, and A. Flahault. Understanding the dynamics of Ebola epidemics. *Epidemiology and Infection*, 135(4):610–621, 2007.

[14] E. Leroy, B. Kumulungui, X. Pourrut, P. Rouquet, A. Hassanin, P. Yaba, A. Delicat, J. Paweska, J. Gonzalez, and R. Swanepoel. Fruit bats as reservoirs of Ebola virus. *Nature*, 438:575–576, 2005.

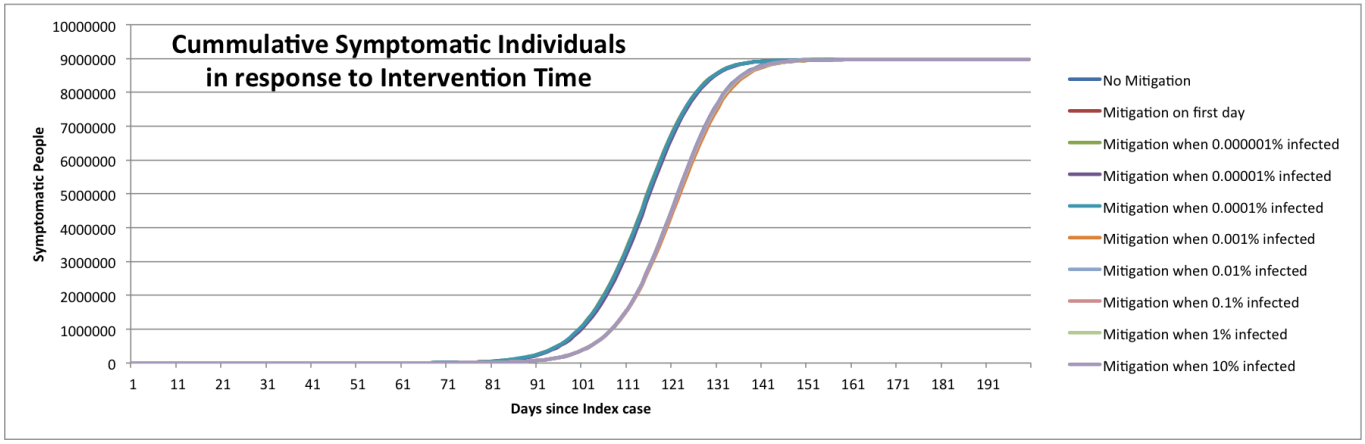


Fig. 5. Cummulative simulation results from the D4D-informed model detailing the count of infected individuals by day.

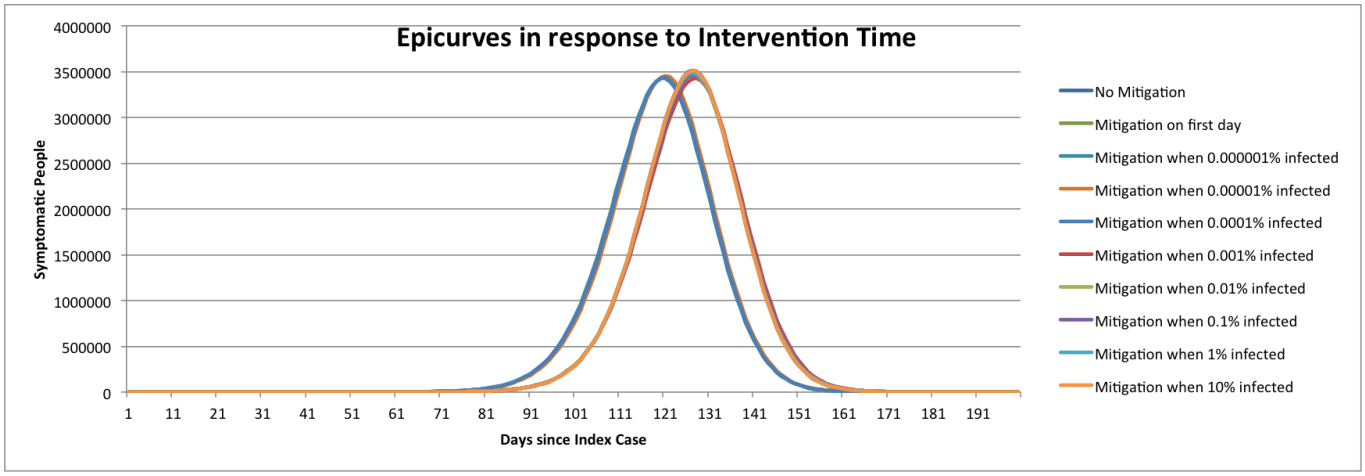


Fig. 6. Daily (current) simulation results from the CDC model detailing the count of infected individuals by day.

[15] **Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 6.3 [Machine-readable database]. *Minneapolis: University of Minnesota Technical Report*, 2014. The authors wish to acknowledge the statistical offices that provided the underlying data making this research possible: National Agency of Statistics and Demography, Senegal.**

[16] **S. Ng and B. Cowling. Association between temperature, humidity and ebolavirus disease outbreaks in Africa, 1976 to 2014. *Euro Surveillance*, 19(35), 2014.**

[17] **P. Rouquet, J. Froment, M. Bermejo, A. Kilbourne, W. Karesh, P. Reed, B. Kumulungui, P. Yaba, A. Dlicat, P. Rollin, and E. Leroy. Wild Animal Mortality Monitoring and Human Ebola Outbreaks, Gabon and Republic of Congo, 2001-2003. *Emerging Infectious Diseases*, 11(2):283–290, 2005.**

[18] **T. Smith, G. F. Killeen, N. Maire, A. Ross, L. Molineaux, F. Tediosi, G. Hutton, J. Utzinger, K. Dietz, and M. Tanner. Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of Plasmodium falciparum malaria: Overview. *Am J Trop Med Hyg*, 75:1–10, 2006.**

[19] **J. Towner, P. Rollin, and D. Bausch. Rapid diagnosis of Ebola hemorrhagic fever by reverse transcriptionPCR in an outbreak setting and assessment of patient viral load as a predictor of outcome. *J Virol*, 78:4330–4341, 2004.**

[20] **M. Washington, C. Atkins, and M. Meltzer. CDC Generic EbolaResponse (ER), Modeling the spread of disease impact and intervention, version 2.5. http://stacks.cdc.gov/view/cdc/24900, Sept. 2014.**

[21] **WHO. Sierra Leone: a traditional healer and a funeral. http://www.who.int/csr/disease/ebola/ebola-6-months/sierra-leone/en, 2014.**

APPENDIX I: SAMPLE SIMULATION RESULT FILE

Model revised from FluTE version: 1.15

Label: senegal-antenna

Population: flute-senegal-antenna

Nodes used: Single processor version

beta: 0.262342

R0: 1.5

Random number generator seed: 4925

Tracts: 1666

Communities: 7125

Families: 1009920

People: 13401030

Essential worker fraction: 0

Travel allowed: 1

School opening days: 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

Seeded 1 infected people

Seeded (n/10000) at airports: 0

Seeded once

Pre-existing immunity level: 1

Pre-existing immunity fraction by age: 0,0,0,0,0,

Baseline VES by age: 0,0,0,0,0,

Prevaccination strategy: none

Reactive vaccination strategy: none

Vaccine priorities : 1,1,1,1,1,1,1,1,1,1,1,1,1,

Response threshhold: 0.001

Response delay: 1

Ascertainment delay: 3

Ascertainment fraction: 1

Reactive strategies deployed on day: 85

School closure policy: none

School closure days: 0

Isolation: 1

Quarantine: 0

Liberal leave: 0

Antiviral policy: none

Vaccination fraction: 0.7

Vaccines initially available: 0

Vaccines deliverable per day: 4294967295

Antivirals available: 4294967295

Antivirals deliverable per day: 4294967295

Vaccines used: 0

Antivirals used: 0

Relative vaccine efficacy by age: 1,1,1,1,1,

Needs boost by age and school: 0,0,0,0,0,0,0,0,

AVEs: 0.3

AVEp: 0.6

AVEi: 0.62

Number symptomatic (daily):0,0,0,0,1,1,1,1,1,0,0,1,1,3,3,4,5,5,6,6,9,12,13,13,16,16,17,15,17,17,17,20,21,22,23,27,33,41,41,43,49,52,58,75,89,104,111,129,152,178,194,231,270,314,364,406,476,550,632,724,809,882,954,1061,1201,1324,1460,1639,1830,2085,2391,2742,3124,3622,4173,4928,5776,6799,7958,9407,11141,13122,15445,18375,21744,25765,30356,35729,42150,49767,58871,69725,82374,97469,115235,136116,160450,188398,221564,259293,303008,353730,411042,477413,551518,635337,728576,831316,945859,1071626,1207814,1354632,1512152,1678007,1850776,2029545,2211417,2393217,2573615,2747027,2909211,3058513,3189650,3299633,3384220,3439486,3464408,3460813,3424968,3358168,3263406,3141800,2996229,2833957,2656481,2466030,2270939,2073221,1876586,1684372,1499856,1324490,1160061,1009183,869965,745194,633129,534112,447922,372810,308096,253522,207233,168935,137338,111019,89424,71815,57632,46148,36850,29484,23515,18807,15018,12013,9605,7763,6247,5029,4076,3333,2718,2186,1821,1518,1264,1047,878,744,633,543,466,381,321,267,218,181,145,120,97,80,70,60,50,42,34,30,29,27,

Cumulative symptomatic (daily):0,0,0,0,1,1,1,1,1,1,1,2,2,4,4,5,6,6,8,8,11,14,15,15,19,19,20,21,23,26,28,32,34,37,40,44,50,59,61,66,74,81,89,110,130,150,164,192,223,256,285,333,388,449,509,572,666,766,876,1004,1131,1247,1371,1541,1744,1950,2168,2430,2732,3080,3490,3962,4515,5186,5922,6886,8008,9325,10803,12647,14836,17355,20294,23943,28170,33220,39042,45883,54041,63697,75145,88839,104871,123859,146249,172618,203372,239099,281336,329949,386400,452041,527143,614012,712485,824705,950750,1092294,1251706,1428761,1624785,1839832,2075430,2330219,2603403,2895343,3204346,3527796,3865366,4212597,4565343,4921377,5276548,5626512,5967604,6295237,6606095,6899516,7170854,7419072,7644425,7845318,8022568,8179048,8313175,8427580,8524576,8606079,8674115,8730453,8776387,8813689,8843979,8868321,8887908,8903671,8916039,8926075,8934048,8940396,8945366,8949334,8952521,8955134,8957140,8958761,8960066,8961059,8961883,8962577,8963116,8963601,8963951,8964254,8964501,8964729,8964913,8965063,8965190,8965284,8965377,8965454,8965531,8965571,8965617,8965650,8965681,8965707,8965731,8965748,8965762,8965780,8965787,8965799,8965805,8965811,8965817,8965820,8965825,8965831,8965832,8965834,8965834,8965834,8965836,8965837,8965839,8965842,8965844,8965844,

Number Dead (daily):0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,2,2,2,2,3,3,3,4,4,6,8,9,9 11,11,12,13,13,15,17,18,19,20,25,28,33,40,44,48,59,66,73,87,101,116,130,148,169,190,218,259,284,322,377,431,492,565,652,746,829,927,1025,1123,1270,1432,1585,1769,1991,2270,2594,2960,3407,3915,4495,5159,5982,6963,8139,9577,11239,13136,15381,18037,21387,24955,29347,34561,40663,47791,56374,66426,78155,91954,108290,127373,149719,175692,206215,241525,282480,329316,383418,445621,516121,596104,686270,787541,900757,1025804,1164533,1316663,1482199,1661619,1853541,2057486,2273759,2500115,2735229,2976763,3222093,3468465,3715252,3958898,4196211,4425092,4642280,4846557,5036780,5211791,5370491,5513049,5639313,5750913,5847154,5929703,6000097,6059338,6109457,6150871,6184758,6212586,6234981,6253185,6268045,6279972,6289371,6296873,6302848,6307680,6311494,6314430,6316849,6318832,6320395,6321613,6322614,6323478,6324135,6324625,6325028,6325368,6325647,6325880,6326097,6326261,6326398,6326503,6326592,6326659,6326734,6326788,6326831,6326861,6326899,6326927,6326950,6326969,6326985,6326999,6327015,6327027,6327035,6327041,6327047,6327050,6327054,6327055,6327057,6327062,6327063,6327064,6327064,

Total symptomatic attack rates by age: 0.669341,0.67008,0.66874,0.668433,0.6686,

Total Fatality rates by age: 0.703276,0.703554,0.707201,0.707406,0.702797,

Total symptomatic attack rate: 0.669041

Total fatality rate 0.705685

Total symptomatic individuals by age: 1336028,2191929,1950177,3256097,231613,

Total dead individuals by age: 939596,1542141,1379168,2303382,162777,

Total symptomatic unvaccinated individuals by age: 1335803,2191583,1949466,3254943,231576,

Total unvaccinated individuals by age: 1996036,3271146,2916194,4871239,346415,

Total individuals by age: 1996036,3271146,2916194,4871239,346415,

N02

Spatial structure and efficiency of commuting in Senegalese cities

Rémi Louf[1], Giulia Carra[1], Hadrien Commenges[2], Jean-Marie Dembele[3], Riccardo Gallotti[1], Maxime Lenormand[4], Thomas Louail[1], Marc Barthelemy[1,5]

[1] *Institut de Physique Théorique, CEA-CNRS (URA 2306), F-91191, Gif-sur-Yvette, France*

[2] *Laboratoire Eau, Environnement et Systèmes Urbains, Ecole des Ponts, F-77455, Marne-la-Vallée, France*

[3] *Université Gaston Berger, UFR SAT, Saint-Louis, BP 234, Senegal*

[4] *IFISC, Instituto de Física Interdisciplinar y Sistemas Complejos (CSIC-UIB), Campus Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain and*

[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 [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 developement; 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_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's ICT devices could revolutionize the quantitative and theoretical understanding of human spatial dynamics, and urban dynamics [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 [6–10]; the universal structure of subway networks and streets networks [11, 12]; the number and the spatial organisation of centers in urban areas [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 peripheric 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 developping 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 developped 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 refered 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), drawed the communication networks upon natural ressource layers and discussed the resulting maps. While providing advices 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 relations of each city with the others. We evaluate their geographical area of influence ('attraction basin') and the degree of 'integration' of their labour and housing markets. We then focus on the intra-urban scale, and provide several measures of the spatial organisation of these cities. We first focus on residential and activity hotspots, and discuss their spatial organisation and the shape of the Senegalese cities. We then analyse the structure of commuting in the cities, and propose an original method to characterize the efficiency of the commuting structure, by evaluating the spatial mismatch between residences and workplaces locations. Finally we discuss the results obtained, and the first hints in terms of planning.

Results

The following analysis is based on origin-destination (OD) matrices extracted from the dataset 2 for the largest Senegalese cities. An OD matrix is a classic object in transport planning and mobility studies, and is a $n \times n$ matrix F_{ij} where n is the number of spatial units that compose the city at the spatial aggregation level considered. The element F_{ij} represents the number of individuals living in location i and commuting to location j where they have their main, regular activity (work or school for most people). The general idea for extracting the OD matrices from individual mobile phone data is to determine for each individual the pair of locations (i.e. antennas) that are the best proxys for their home and main daily activity locations (see Methods for details). In order to reduce the noise as much as possible with the dataset at hand, the results presented here have been obtained by summing the 25 OD matrices of the dataset 2, each covering a 2-weeks period.

Structure of inter-urban commuting

We first study interactions – in terms of commuting – between each of the major cities identified in the data, and between each city and the rest of the country.

For a given city i, we can distinguish three types of commuters:

- The *internal* commuters, who live in i and whose main daily activity is also located in i;
- The *divergent* commuters, who live in i, and whose main daily activity is based outside i;
- The *convergent* commuters, who live outside i, and whose main daily activity is based in i;

In the following we denote by N_i°, $N_i^{\rightarrow}$ and $N_i^{\leftarrow}$ the numbers of internal, divergent and convergent commuters, respectively. The total commuting population C of the urban area (for the summed 25 OD matrices) is then given by the total number of internal commuters plus the total number of divergent commuters

$$C = N^o + N^{\rightarrow} \tag{1}$$

These three types of commuting are represented schematically on Fig. 1.

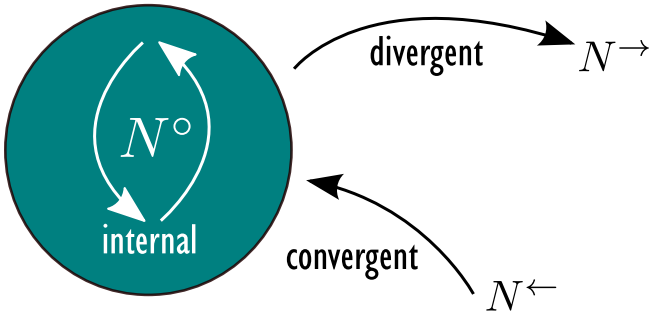


FIG. 1: Schematic representation of the different types of commuters: internal, convergent and divergent commuters.

	C	N°	$N^{\leftarrow}$	$N^{\rightarrow}$	f_0	f_1
Dakar	3,099,116	3,049,579	47,540	49,537	0.03	0.95
Thies	261,050	248,950	10,718	12,100	0.09	0.88
Touba	230,498	218,987	13,980	11,511	0.11	1.21
Mbour	216,418	204,312	11,111	12,106	0.11	0.91
Kaolack	178,365	170,821	8,097	7,544	0.09	1.07
Saint-Louis	162,875	154,910	5,843	7,965	0.08	0.73
Ziguinchor	135,562	130,264	4,255	5,298	0.07	0.81
Tambacounda	65,577	62,900	3,049	2,677	0.09	1.13
Tivaoune	63,198	53,815	7,532	9,383	0.31	0.80
Diourbel	62,345	59,357	3,585	2,988	0.11	1.19
Louga	59,883	56,964	3,649	2,919	0.11	1.25
Kolda	57,895	55,676	2,304	2,219	0.08	1.03

TABLE I: Partition of commuting flows into internal, divergent and convergent commuting flows, for the 12 largest urban areas identified in Senegal (ordered by total commuting population).

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 nightime.
- 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 nightime.
- 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

$$r^{\rightarrow} = \frac{1}{N^{\leftarrow}} \sum_n \ell_n \qquad (2)$$

where ℓ_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 a 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-

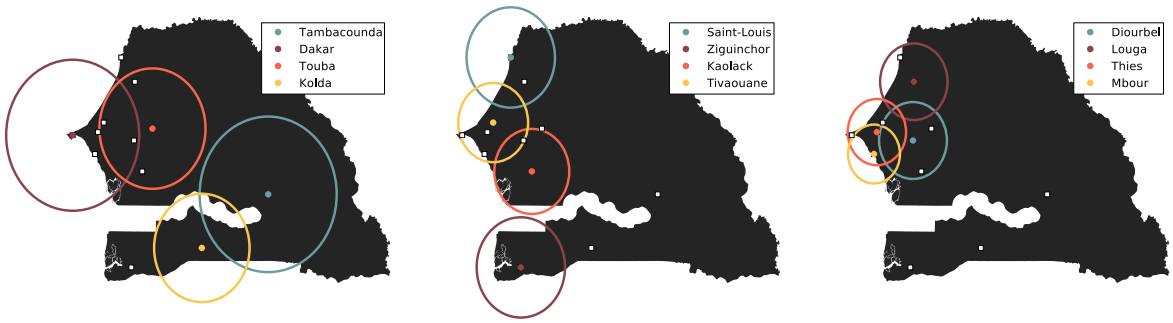


FIG. 2: **Attraction radius for the major Senegalese cities.** We represent the radius of attraction computed using the OD matrices that we extracted from the data. The radius for different cities are shown on three different panels for a matter of clarity.

	$r^{\leftarrow}$ (km)
Dakar	137
Thies	60
Touba	109
Mbour	53
Kaolack	77
Saint-Louis	91
Ziguinchor	91
Tambacounda	141
Tivaouane	72
Diourbel	69
Louga	69
Kolda	98

TABLE II: **Attraction radius** $r^{\leftarrow}$. 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.

ular enough to infer with confidence their home and activity locations (see Methods). To inspect these aspects further, we would need the data provided in dataset 2, but available for a longer period of time for each individual. The dataset 3 could allow to solve this problem, but its spatial resolution is much scarcer, and the small number of individuals per city (Dakar excepted) would raise other difficulties. We thus leave this investigations for further studies and additional datasets. In the following we will focus on what the journey-to-work OD matrices can teach us regarding the intra-urban mobility and the spatial structure of Senegalese cities.

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].

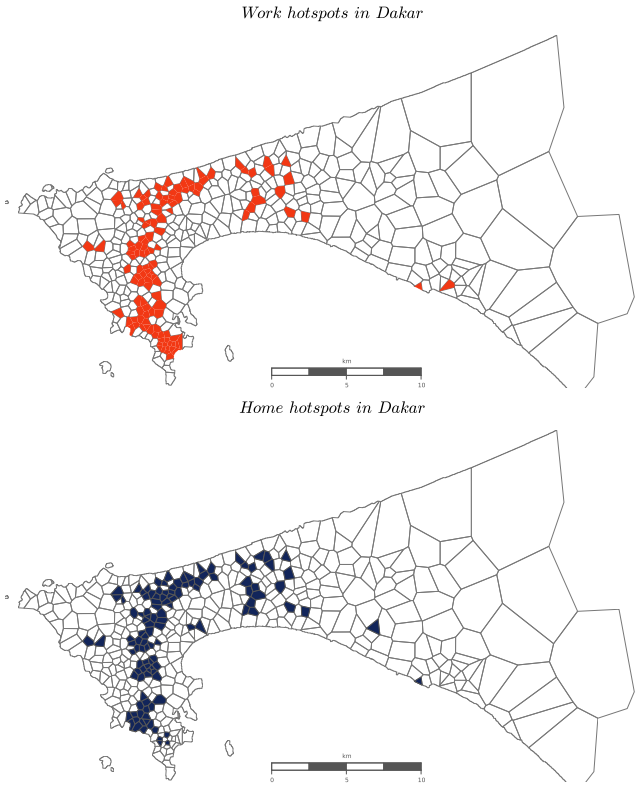


FIG. 3: Location of residential and activity hotspots in Dakar, determined with the method described in [14].

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' hotposts 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' hostpots (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 hopspots 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 hostpots are also home hostpots. 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 influent 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 addind 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

$$O = \frac{L - L_R}{L_O - L_R} \tag{3}$$

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

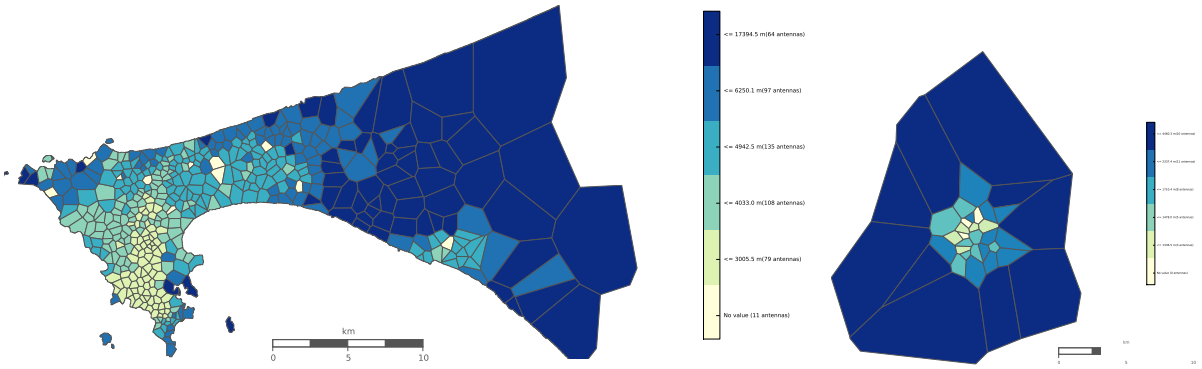


FIG. 4: **Average commuting distance at the antenna level** (Left) Dakar: Average outreach per antenna, i.e. average distance commuted by people living in the area covered by this antenna. When calculating the average we do not take into account the people who are flagged as living and working at the same place. (Right) Same measure for the urban area of Thies. These results are obtained by averaging the OD matrices over the whole year.

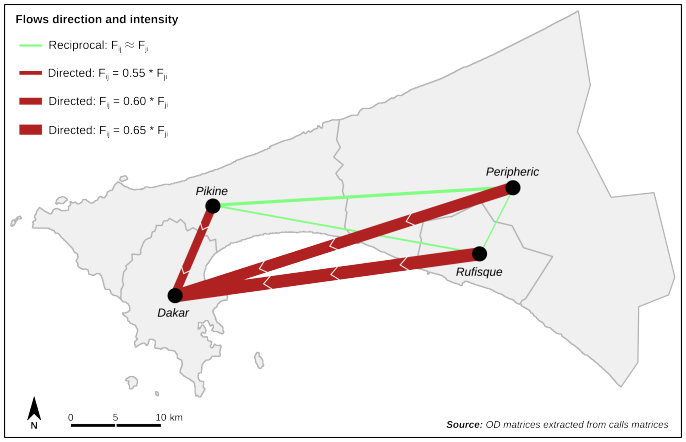


FIG. 5: **Main directed flows of commuters at the municipality scale, in the urban area of Dakar.** As the urban area expanded in direction of the surrounding municipalities, several secondary acivity and employment hotspots appeared in the urban area of Dakar. Still the historical activity centers, located in the municipality of Dakar, attract more commuters that live in the surrounding municipalities, than the opposite.

	L/L_O	L/L_R	O
Dakar	2.09	0.22	0.87
Thies	3.49	0.23	0.82
Touba	3.70	0.18	0.86
Mbour	3.36	0.21	0.84
Kaolack	1.92	0.30	0.83
Saint-Louis	3.84	0.20	0.84
Ziguinchor	3.08	0.29	0.78
Tambacounda	3.90	0.28	0.77
Tivaoune	8.98	0.20	0.82
Diourbel	2.97	0.28	0.79
Louga	4.47	0.31	0.74
Kolda	3.38	0.27	0.79

TABLE III:

	L (km)	L_O (km)	L_R (km)	$\sqrt{A}$ (km)
Dakar	4.6	1.7	21	20

TABLE IV: Comparison of the average commuting length as measured on the data (ℓ), on the optimal OD matrix (ℓ_O), the random OD matrix (ℓ_R), and the typical size of the city ($\sqrt{A}$).

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 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 proove 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 neighbordhood (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

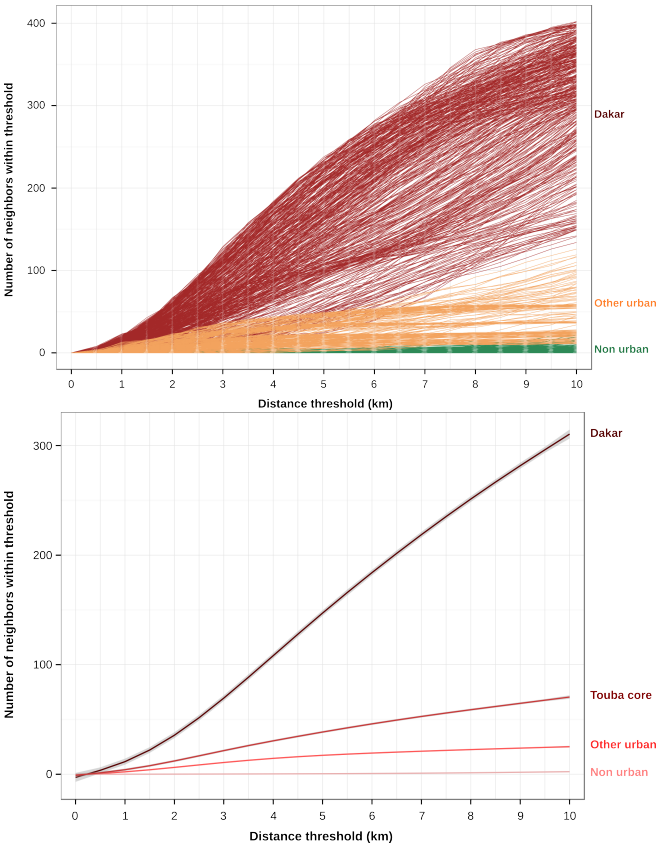


FIG. 6: Illustration of the method used to delimitate cities. For each antenna of the dataset, we count its number of neighbors in a circle of increasing radius (x-axis of the top figure). Each curve represent an antenna. We then apply a hierarchical clustering method on the resulting set of vectors, and represent the average profile of each class on the bottom figure.

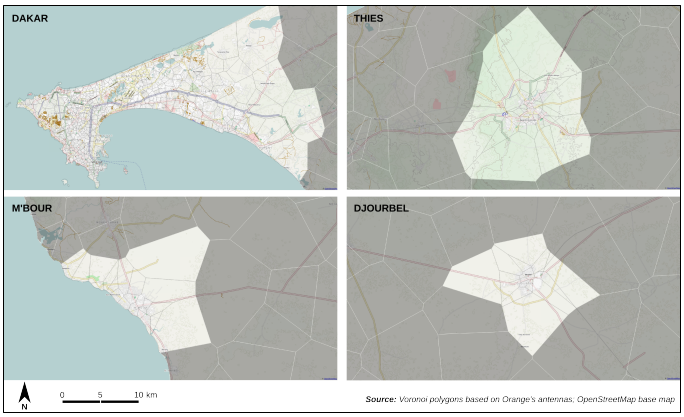


FIG. 7: Comparison of the antennas chosen to delimitate four of the twelve cities selected, following our criterion based on the density of antennas (red), and the corresponding OpenStreetMap layers. We checked each city individually (others not shown here).

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 u, 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 *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 *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 *prop* increases, which is an expected effect. However the proportion of intra-cell flows tends to sharply increase when *prop* increases. In order not to remove users and try to keep the network structure we choosed to fix *prop* to 1/3.

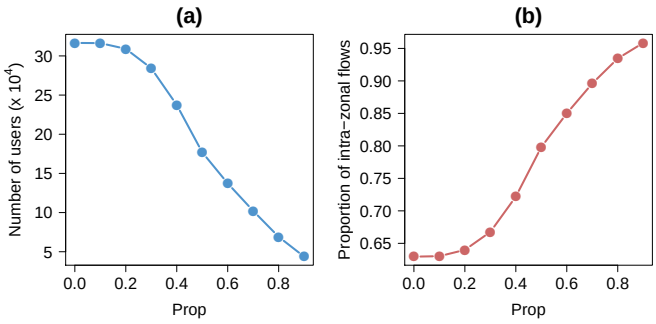


FIG. 8: Sensitivity of (a) the number of individuals selected and (b) the proportion of intra-antenna flows, to the parameter *prop* of the OD matrix extraction method. These tests have been performed on the dataset SET2_P01.

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-detination 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 intepret 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

$$P(1 \to 2 = exp(\frac{L_1 - l_2}{T}) \quad (4)$$

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 τ. We call τ 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.

$$\tilde{A}_\tau(i) = \int_\tau A(i,x)\,\mathrm{d}x \quad (5)$$

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

$$M_\tau(i) = \frac{\tilde{A}_\tau(i)}{\sum_{i\in\mathcal{N}} \tilde{A}_\tau(j)} \quad (6)$$

Depending on the area of study $\mathcal{N}$ and the time-window τ we might be able to catch different phenomena. For instance, setting $\mathcal{N}$ to be a city, τ 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 τ 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

$$DN(i) = \frac{M_{day}(i) - M_{night}(i)}{M_{night}} \quad (7)$$

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 EUNOIA (FP7-DG.Connect-318367). TL is supported by the EU commission through project EUNOIA (FP7-DG.Connect-318367). ML acknowledges funding from the PD/004/2013 project, from the Conselleria de Educacin, 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)

[1] The World Bank, Senegal `http://www.worldbank.org/en/country/senegal` (accessed 22/12/2014).

[2] City population website `http://www.citypopulation.de/Senegal.html`

[3] Senegal Country Assessment Report - MIT `http://web.mit.edu/urbanupgrading/upgrading/case-examples/overview-africa/country-assessments/reports/Senegal-report.html` (accessed 10 december 2014).

[4] Lenormand, M. et al. Cross-checking different sources of mobility information. *PLoS ONE* **9**(8): e105184. doi:10.1371/journal.pone.0105184 (2014).

[5] Ratti, C., Williams, S., Frenchman, D. & Pulselli, R.M. Mobile landscapes: using location data from cell phones for urban analysis. *Environ. Plann. B* **33,** 727–748 (2006).

[6] González M., Hidalgo C. and Barabási A.-L. (2008) Understanding individual human mobility patterns. *Nature* 453, pp 779-782.

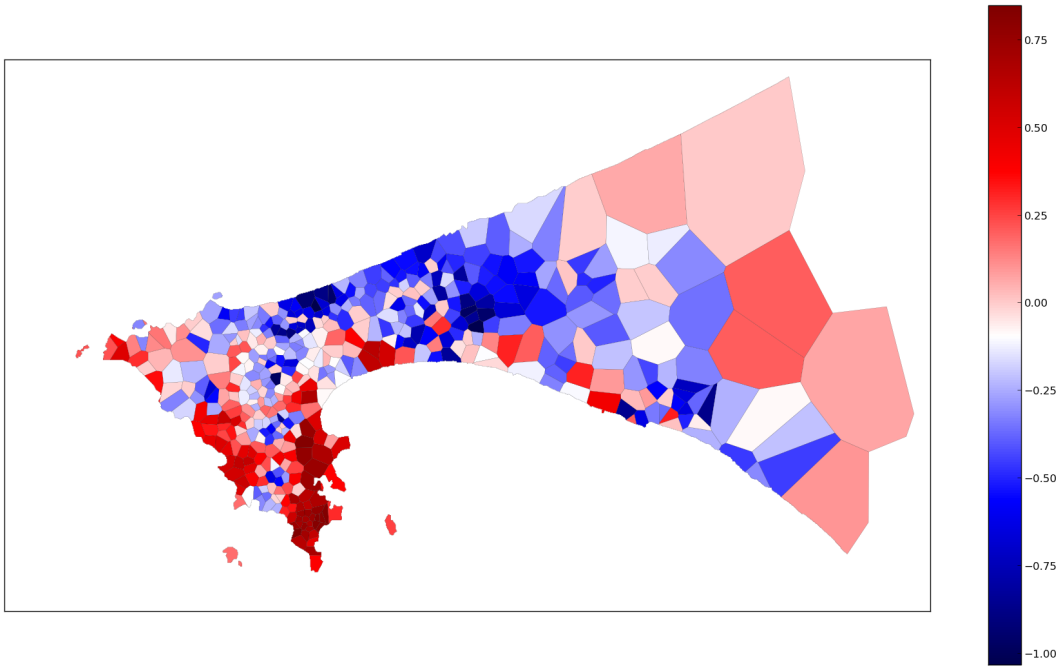


FIG. 9: Relative difference of antenna mass between night and day in Dakar. Each Voronoi cell is colored with respect to the value $\frac{M_{day}(i)-M_{night}(i)}{M_{day}}$ calculated for the corresponding antenna i.

[7] Roth, C., Kang, S.-M., Batty, M. & Barthelemy, M. Structure of urban movements: polycentric activity and entangled hierarchical flows. *Plos ONE* **6,** e15923 (2011).

[8] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. & Mascolo, C. A tale of many cities: universal patterns in human urban mobility. *Plos ONE* **7**:e37027 (2012).

[9] Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z. & González, M.C. Unravelling daily human mobility motifs. *J R Soc Interface* **10**:20130246 (2013).

[10] Louail, T. et al. Uncovering the spatial structure of mobility networks. *Nature Communications*, in press (2014).

[11] Roth C., Kang S.M., Batty M. and Barthelemy M. (2012) A long-time limit of world subway networks. *Journal of the Royal Society Interface*, doi:10.1098/rsif.2012.0259.

[12] Louf R. and Barthelemy M. (2014) A typology of street patterns. *Journal of The Royal Society Interface* 11 (101).

[13] Louf, Rémi, and Marc Barthelemy. "Modeling the polycentric transition of cities." *Physical review letters* **111**.19 (2013): 198702.

[14] Louail, T. et al. From mobile phone data to the spatial structure of cities. *Scientific reports* **4**:5276 (2014).

[15] Expert P., Evans T.S., Blondel V.D. and Lambiotte R. (2011) Uncovering space-independent communities in spatial networks, *Proceedings of the National Academy of Sciences* 108, 7663.

[16] Herrera-Yagüe et al. (2014) The origins of searchability in social networks, online on MIT HumNet's publications page.

[17] Pumain D., Paulus F., Vacchiani C. and Lobo J. (2006) An evolutionary theory for interpreting urban scaling laws, *Cybergeo*, 343, 20 p.

[18] Bettencourt LMA, Lobo J, Helbing D, Kuhnert C and West GB (2007) Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Science*,104:7301–7306.

[19] Louf R. and Barthelemy M. (2014) Scaling: lost in the smog. *Environment and Planning B* 41 (5), 767-769.

[20] Arcaute, E. et al. Constructing cities, deconstructing scaling laws. Journal of the Royal Society Interface, http://dx.doi.org/10.1098/rsif.2014.0745 (2014).

[21] Bertaud, A. & Malpezzi, S. The spatial distribution of population in 48 world cities: implications for economies in transition. *World Bank Report* (2003).

[22] Bretagnolle, A., Pumain, D. & Vacchiani-Marcuzzo, C. [The organisation of urban systems] in *Complexity perspective in innovation and social change*, [Lane, D., Pumain, D., van der Leeuw, S.E. and West, G. (eds)][197–220], Springer (2009).

[23] Grataloup C., Fumey, G. and Boucheron, P. L'atlas Global, *Editions les Arènes* (2014).

[24] Simini F., González M., Maritan A. and Barabási A.-L. A universal model for mobility and migration patterns. *Nature* **484**, pp 96–100 (2012).

[25] Kung K.S., Sobolevsky, S. & Ratti, C. Exploring universal patterns in human home/work commuting from mobile phone data. *Plos ONE* **9**(6) (2014).

[26] Berlingerio M. et al. AllAboard: a system for eexploring urban mobility and optimizing public transport using cellphone data. *D4D 2013 Challenge book* (2013).

[27] Mamei M. and Ferrari L. Daily commuting in Ivory Coast: development opportunities. *D4D 2013 Challenge book* (2013).

[28] Andris C. and Bettencourt, L. Development, Information and Social Connectivity in Côte d'Ivoire. SFI Working paper (2013).

[29] Wakita, K., Kawasaki, R. and Takami, M. Observation of Human Dynamics in Cote d'Ivoire Through D4D Call Detail Records, *AAAI Technical Report WS-13-04 When the City Meets the Citizen 2*, pp 29–32 (2013).

[30] Eubank, S. et al. Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180-184 (2004)

[31] Balcan, D. et al. Multiscale mobility networks and the spatial spreading of infectious diseases. *PNAS* **106**, 21484–21489 (2009).

[32] Dalziel, B.D., Pourbohloul, B. & Ellner, S.P. Human mobility patterns predict divergent epidemic dynamics among cities. *Proc R Soc B* **280**:20130763 (2013).

[33] Bretagnolle, A., Paulus, F. & Pumain, D. Time and space scales for measuring urban growth. *Cybergeo* 219, `http://cybergeo.revues.org/3790` (2002).

[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).

[35] Mézard, M., Parisi, G. and Virasoro, M. A. Spin glass theory and beyond. World Scientific (1987).

[36] Jiang, S. et al. A Review of Urban Computing for Mobile Phone Traces:Current Methods, Challenges and Opportunities. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, doi:10.1145/2505821.2505828 (2013).

[37] Wu, L., Zhi, Y., Sui, Z. & Liu, Y. Intra-urban human mobility and activity transition: evidence from social media check-in data.*Plos ONE* **9**(5)e97010 (2014).

[38] Zhong, C., Arisona, S.M., Huang, X., Batty, M. & Schmitt, G. Detecting the dynamics of urban structure through spatial network analysis. *Int J Geo Inf Sc* **28** (2014).

[39] Deville, Pierre, et al. "Dynamic population mapping using mobile phone data." *Proceedings of the National Academy of Sciences* **111**(45) : 15888-15893 (2014).

[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

N03

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-grain 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 departements 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 Senegalese 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 interactions on more than 75% of the days, but less than 1000 interactions a week are included. While three different datasets are provided, antenna-to-antenna traffic of all 1666 antennae, fine-grained mobility patterns of frequently changing sub-populations and coarse-grained mobility patterns for a selected sub-population, we concentrate on the first dataset in our analysis. Mobility patterns require sophisticated 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 representative sample of 7,902 households covering 77,269 individuals 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: Locations of survey centres and antennae

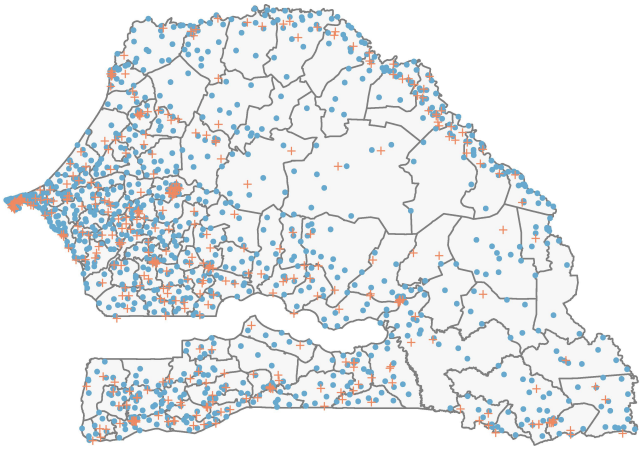


Figure 1 shows the GPS locations of the DHS cluster centres (orange crosses) next to the D4D antennae (blue points).

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 assumption of representativeness of the survey. Second, sample 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

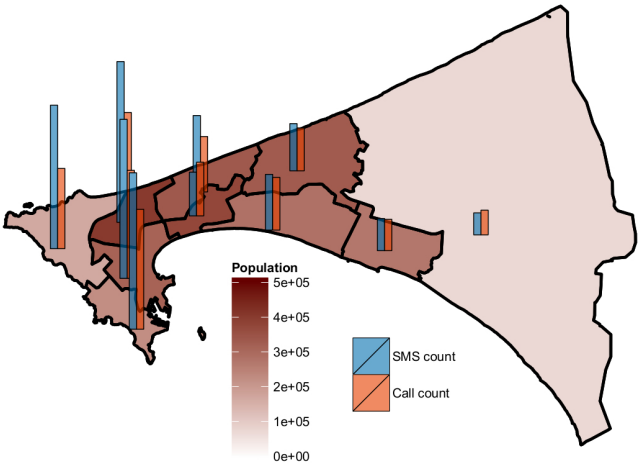
Min	1st Qu.	Median	Mean	3rd Qu.	Max.
81	163	197	197.6	230.5	409

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:

$$E(Y|X) = \hat{\beta}_{0r} + X_1'\hat{\beta}_1 + ... + X_n'\hat{\beta}_n + ... + X_i'\hat{\beta}_{(n+j)r}$$

$$\forall\, r \in R \wedge \forall\, i \in 1...n \wedge \forall\, j \in 1...m$$

R consists of the 14 regions of the Senegal and $\hat{\beta}_{0r}$ contains an intercept for each of these regions. The term $X_i'\hat{\beta}_{(n+j)r}$ 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

the 123 administrative areas of the Senegal, we use the coefficients from our fitted models to predict socio-demographic indicators for the 1666 towers. The 2.5 % and the 97.5 % quantiles of the predictions are winsorized to account for outliers. Standard cross-validation approaches such as non-exhaustive k-fold cross-validation would assess the predictive power in the same (spatial) dimension. In order to assess inter-spatial accuracy, we average the obtained tower-level predictions for every arrondissement, respectively, and then compare them to the arrondissement values from the DHS survey (benchmarks) using a root mean squared error approach. We call this measure 'Inter-Spatial RMSE' (IS-RMSE). Sampling uncertainty deriving from the mobile phone dataset is not taken into account, since a mobile penetration rate of close to 100% [12] and a client base of around 9 million users [11] (of approx. 13.5 million inhabitants [14]) suggest the assumption of being negligible.

3.3 Robustness

So far, we have used annual aggregates for estimation. Calculations based on such aggregates suffer, however, from two disadvantages. First, data preparation involves high computing capacities, which might not always be available. Second, sub-annual volatility of the target variables is not captured. Sub-annual availability of data appears to be less relevant for indicators such as literacy or educational attainment. Frequent food security or population estimates, however, especially 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 individually, 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 interpolation 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 hexagons 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, timely estimates of socio-demographic variables from antenna data only is summarized in the recipe below.

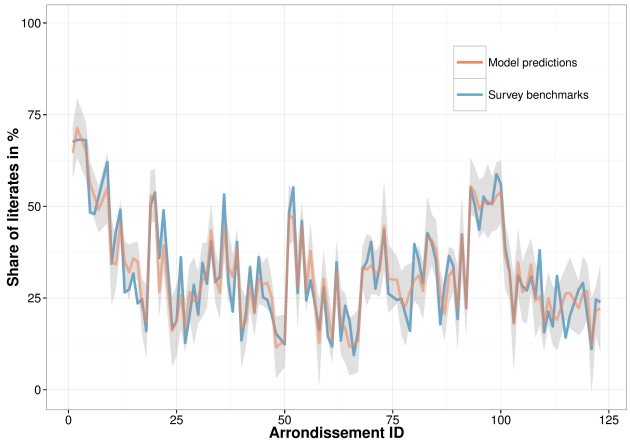
3.5 The Recipe

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 level 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 predictions with the 95% confidence intervals (CI) under normality. The blue line represents the share of literates calculated 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 variation with the confidence interval (grey band) staying close to the estimated values, signifying a robust fit.

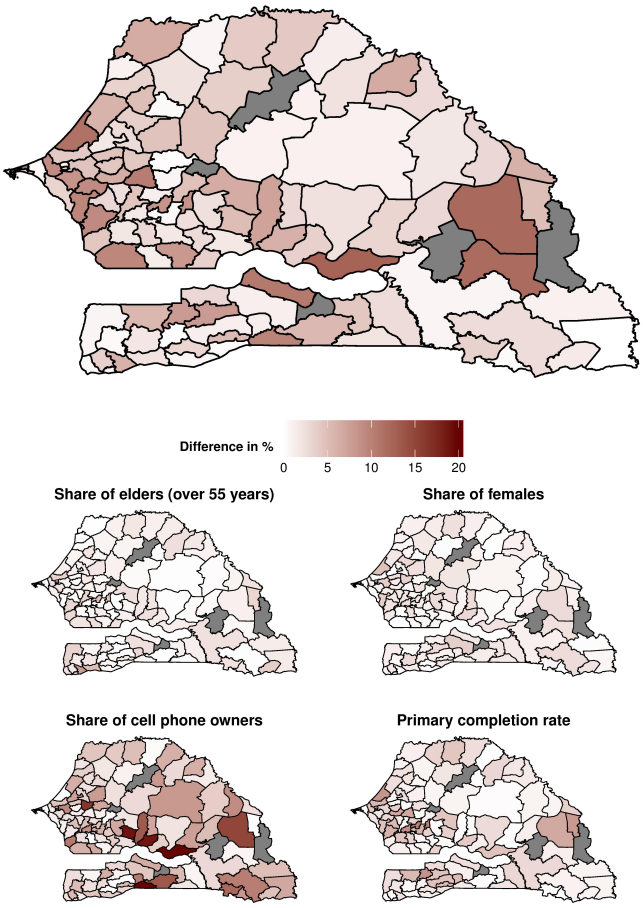
Table 2 presents different performance measures for selected socio-demographic indicators on the arrondissement level. The adjusted R^2 shows the share of total variation captured from the respective model, thereby penalizing for the number of variables used for explanation. The root mean squared error (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 mobile phone data only. While this already demonstrates the value of mobile phone data for national statistics, this value 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 perform as well in the first place but shows more inter-spatial robustness.

Table 2: Basket precision

Variable of interest	Adj. R^2	RMSE	IS-RMSE
Literacy rate	0.9647	0.0502	0.0841
Poverty rate	0.8697	0.1334	0.1587
Primary completion rate	0.9177	0.0338	0.0456
Share of Minors	0.9966	0.0255	0.0343
Share of Elders	0.9625	0.0174	0.0177
Share of Women	0.9984	0.0162	0.0398
Share of Wolofs	0.9526	0.0605	0.0624
Share of Poular	0.9722	0.0421	0.0518
Share of Serer	0.9256	0.0248	0.0377
Safe water access rate	0.9244	0.1524	0.3092
Electricity access rate	0.8677	0.1483	0.1800
Share of cell phone owners	0.9910	0.0629	0.0801

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.

Figure 4: Prediction deviances on arrondissement level (literacy rate as main plot)



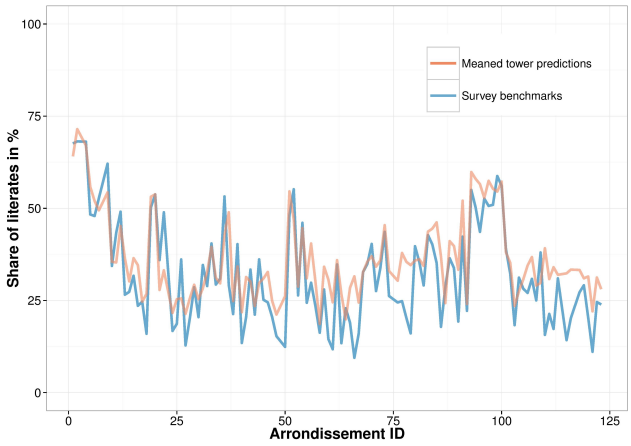
Dark areas point to larger differences between survey values and estimates. For the variables displayed, these differences never rise above 15%. Accurate in-sample fits, however, do not prove the out-of-sample performance of the models.

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.

$$\widehat{\textbf{literacy\_rate}} = region - 0.0002 * \frac{call\_volume}{outgoing\_sms} - 0.4019 * sms\_ratio + 0.0833 * call\_ratio - 0.1096 * sms\_entropy + 0.1055 * call\_entropy - 0.0216 * mean\_sms\_distance + 0.1240 * \frac{calls\_to\_Dakar}{outgoing\_calls} + region * mean\_sms\_distance + region * \frac{calls\_to\_Dakar}{outgoing\_calls}$$

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: Literacy in Dakar

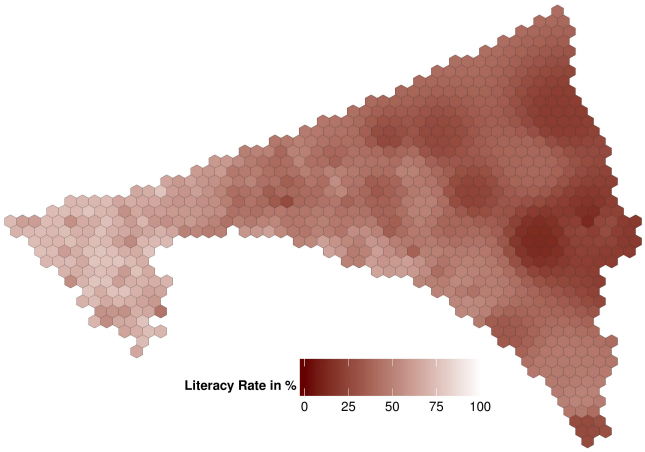


Figure 7: Literacy in Senegal (without Dakar)

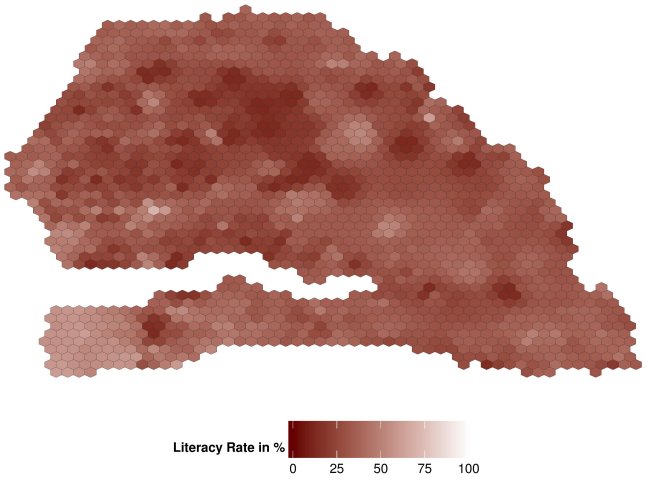


Figure 6 shows the expected image. Literacy peaks in central Dakar and fades away with increasing distance to the city centre. In Figure 8, pockets of illiteracy are clearly visible. Overall, rural Senegal exhibits higher illiteracy. The coastal Casamance, Touba and other major cities in Senegal can 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)

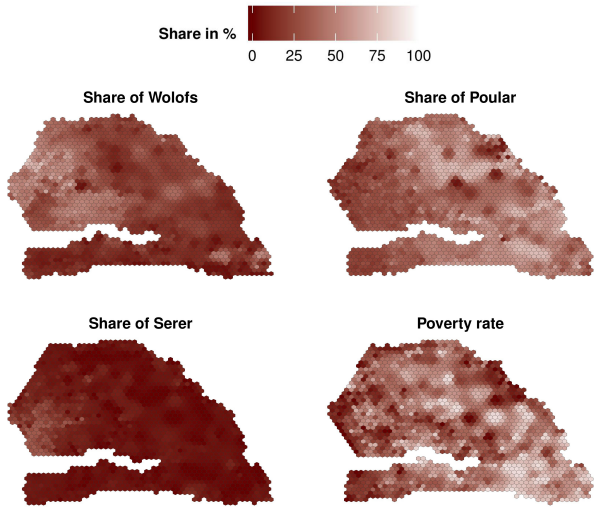


Figure 9 shows the distribution of the three major ethnic groups in the Senegal: Wolof, Poular and Serer. The fourth heat map indicates the population share of the poorest fifth of the country. Poor areas of the Senegal are dominated by the Poular. This also holds true for areas in otherwise Wolof dominated regions.

Figure 9: Age, Education, Gender and Cell Phone Ownership in Senegal (without Dakar)

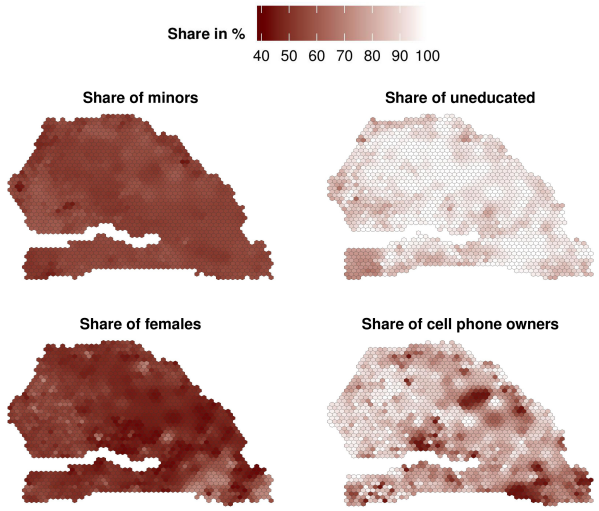


Figure 10 presents the shares of under 18 years olds, of people without completed primary education, of women and of people owning a cell phone. The share of women varies stronger than one would assume. Areas with a high share of people without school education often exhibit low cell phone penetration 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 subpopulations 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

[1]Agence Nationale de la Statistique et de la Démographie, S. and MEASURE DHS, U. 2012. *Demographic and health survey senegal 2010-11.*

[2]Bengtsson, L. et al. 2011. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti. *PLoS medicine*. 8, 8 (2011), e1001083.

[3]Berlingerio, M. et al. 2013. AllAboard: A system for exploring urban mobility and optimizing public transport using cellphone data. *Machine learning and knowledge discovery in databases*. Springer. 663–666.

[4]Deville, P. et al. 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*. 111, 45 (2014), 15888–15893.

[5]Economist, T. 2014. Off the map. *The Economist; November 15th*. (2014).

[6]Estadística, D.A.N. de 2014. Use of satellite images for agricultural statistics. (2014).

[7]Frias-Martinez, E. et al. 2011. An agent-based model of epidemic spread using human mobility and social network information. *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)* (2011), 57–64.

[8]Landefeld, S. 2014. Uses of big data for official statistics: Privacy, incentives, statistical challenges, and other issues. *Discussion Paper for International Conference on Big Data for Official Statistics*. (2014).

[9]Lazer, D. et al. 2009. Life in the network: The coming age of computational social science. *Science (New York, NY)*. 323, 5915 (2009), 721.

[10]Lima, A. et al. 2013. Exploiting cellular data for disease containment and information campaigns strategies in country-wide epidemics. *arXiv preprint arXiv:1306.4534*. (2013).

[11]Montjoye, Y.-A. de et al. 2014. D4D-senegal: The second mobile phone data for development challenge. *arXiv preprint arXiv:1407.4885*. (2014).

[12]Régulation des Télécommunications et des Postes, A. de 2011. *Rapport d'activité*.

[13]Smith, C. et al. 2013. Ubiquitous sensing for mapping poverty in developing countries. *Paper submitted to the Orange D4D Challenge*. (2013).

[14]Statistique et de la Démographie, A.N. de la 2013. *Recensement général de la population et de l'Habitat, de l'Agriculture et de l'Elevage*.

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 $\frac{\#\ of\ outgoing\ events}{\#\ of\ incoming\ events}$.

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 departement 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).

$$I(t_i) = \sum_j \mathbf{I}_{E(t_i,t_j)}\ \forall\ j\ \in\ 1, ..., 1666$$

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 departement.

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 multiply 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:

$$\begin{aligned}\hat{\textbf{poverty\_rate}} &= region + \\ &0.0002 * \frac{call\_volume}{outgoing\_sms} + 1.2633 * sms\_ratio + \\ &0.5342 * call\_ratio + 0.2964 * sms\_entropy - \\ &0.2931 * call\_entropy + 0.0037 * mean\_call\_distance + \\ &0.0010 * distance\_to\_dakar + region * call\_ratio\end{aligned}$$

$$\begin{aligned}\hat{\textbf{share\_of\_minors}} &= region - \\ &0.0042 * \frac{outgoing\_sms}{outgoing\_calls} + 0.0001 * \frac{call\_volume}{outgoing\_sms} - \\ &0.0200 * call\_entropy - 0.2273 * \frac{calls\_to\_dakar}{outgoing\_calls} + \\ &region * \frac{calls\_to\_dakar}{outgoing\_calls} + region * \frac{outgoing\_sms}{outgoing\_calls}\end{aligned}$$

$$\begin{aligned}\hat{\textbf{share\_of\_elders}} &= region + \\ &0.0095 * \frac{outgoing\_sms}{outgoing\_calls} + 0.0702 * sms\_ratio - \\ &0.0364 * call\_ratio - 0.0122 * sms\_entropy + \\ &0.0117 * call\_entropy + 0.0001 * mean\_sms\_distance + \\ &0.0974 * \frac{calls\_to\_dakar}{outgoing\_calls}\end{aligned}$$

$$\begin{aligned}\hat{\textbf{share\_of\_females}} &= region + \\ &0.0231 * \frac{outgoing\_sms}{outgoing\_calls} + 0.0001 * \frac{call\_volume}{outgoing\_sms} - \\ &0.0954 * sms\_ratio + 0.0002 * sms\_isolation - \\ &0.0011 * call\_isolation - 0.0335 * sms\_entropy + \\ &0.0194 * call\_entropy - 0.0005 * distance\_to\_dakar + \\ &region * distance\_to\_dakar + region * call\_entropy\end{aligned}$$

$$\textbf{share\_of\_wolof} = region - 0.00004 * \frac{call\_volume}{outgoing\_sms} - 0.2679 * sms\_ratio + 0.3978 * call\_ratio - 0.6220 * call\_entropy - 0.0004 * mean\_sms\_distance + 0.0012 * mean\_call\_distance - 0.2462 * \frac{calls\_to\_dakar}{outgoing\_calls} - 0.0264 * distance\_to\_dakar + region * distance\_to\_dakar + region * call\_entropy$$

$$\textbf{safe\_water\_access\_rate} = region - 0.0001 * \frac{call\_volume}{outgoing\_sms} - 1.5330 * sms\_ratio + 0.5433 * call\_ratio - 0.0015 * sms\_isolation + 0.0082 * call\_isolation - 0.1634 * sms\_entropy + 0.2727 * call\_entropy - 0.0005 * mean\_sms\_distance - 0.2738 * \frac{calls\_to\_dakar}{outgoing\_calls} - 0.0119 * distance\_to\_dakar + region * distance\_to\_dakar$$

$$\textbf{share\_of\_poular} = region - 0.1686 * call\_ratio + 0.1885 * sms\_entropy + 0.0026 * mean\_call\_distance + 0.1125 * \frac{calls\_to\_dakar}{outgoing\_calls} + 0.0143 * distance\_to\_dakar + region * distance\_to\_dakar + region * \frac{calls\_to\_dakar}{outgoing\_calls} + region * sms\_entropy$$

$$\textbf{share\_of\_cell\_phone\_owners} = region - 0.0995 * \frac{outgoing\_sms}{outgoing\_calls} - 0.0001 * \frac{call\_volume}{outgoing\_sms} - 0.8713 * sms\_ratio - 0.0069 * mean\_sms\_distance - 0.0029 * mean\_call\_distance + 0.1043 * \frac{calls\_to\_dakar}{outgoing\_calls} + 0.0004 * distance\_to\_dakar + region * mean\_sms\_distance + region * mean\_call\_distance$$

$$\textbf{share\_of\_serer} = region - 0.0220 * \frac{outgoing\_sms}{outgoing\_calls} - 0.00002 * \frac{call\_volume}{outgoing\_sms} - 0.0379 * sms\_ratio + 0.0002 * sms\_isolation - 0.0003 * call\_isolation - 0.0125 * sms\_entropy - 0.0168 * mean\_sms\_distance + 0.0093 * mean\_call\_distance + 0.1657 * \frac{calls\_to\_dakar}{outgoing\_calls} + 0.0050 * distance\_to\_dakar + region * distance\_to\_dakar + region * mean\_sms\_distance + region * mean\_call\_distance$$

$$\textbf{primary\_completion\_rate} = region + 0.0797 * \frac{outgoing\_sms}{outgoing\_calls} - 0.4642 * sms\_ratio + 6.6716 * call\_ratio - 0.0793 * sms\_entropy + 0.0914 * call\_entropy - 0.0015 * mean\_call\_distance - 0.2498 * \frac{calls\_to\_dakar}{outgoing\_calls} + 0.0125 * distance\_to\_dakar + region * call\_ratio + region * distance\_to\_dakar$$

$$\textbf{electricity\_access\_rate} = region - 2.0363 * sms\_ratio + 10.3692 * call\_ratio - 0.4880 * sms\_entropy + 0.4278 * call\_entropy - 0.0044 * mean\_call\_distance - 0.9215 * \frac{calls\_to\_dakar}{outgoing\_calls} + region * call\_ratio$$

N07

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%)s, 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 Banks $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-theoritic 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

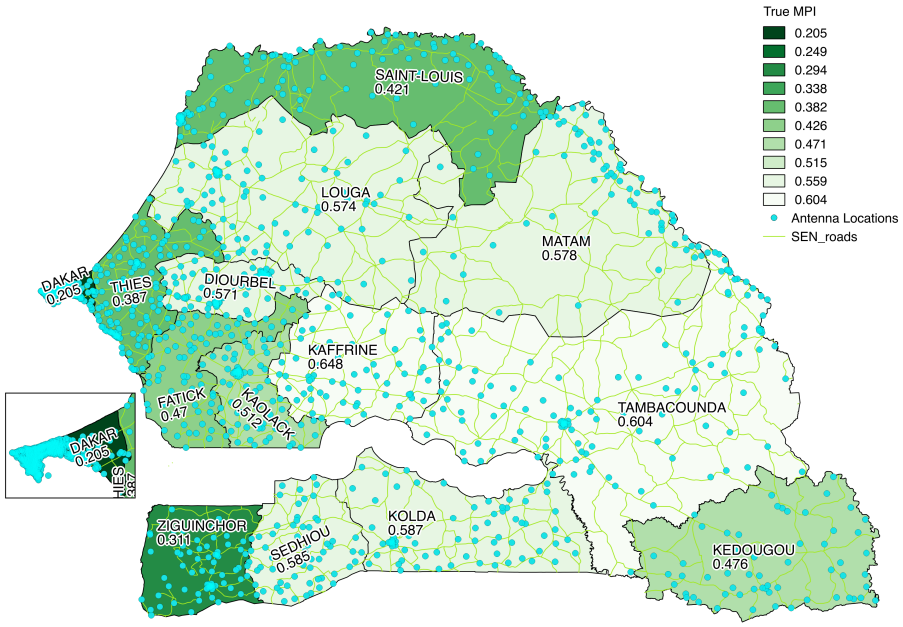


Figure 1: Overlay of Multidimensional Poverty Index (MPI) and cell phone site locations for 14 regions of Senegal.

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 re quiring 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 124×124 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 14×14 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.

4.2.1 Normalization of Information Flow Matrix While previous researchers [11] have used the information flow matrix M^r to characterize the virtual network, we first normalize the matrix to account for the disparity in the population across regions, especially the strong influence of Dakar owing to its relatively high population share (22.9%). We construct a normalized information matrix $\hat{M}^r$ as follows:

$$\widehat{M}^r_{ij} = \frac{M^r_{ij} * d_{ij}}{n_i n_j} \tag{4.1}$$

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

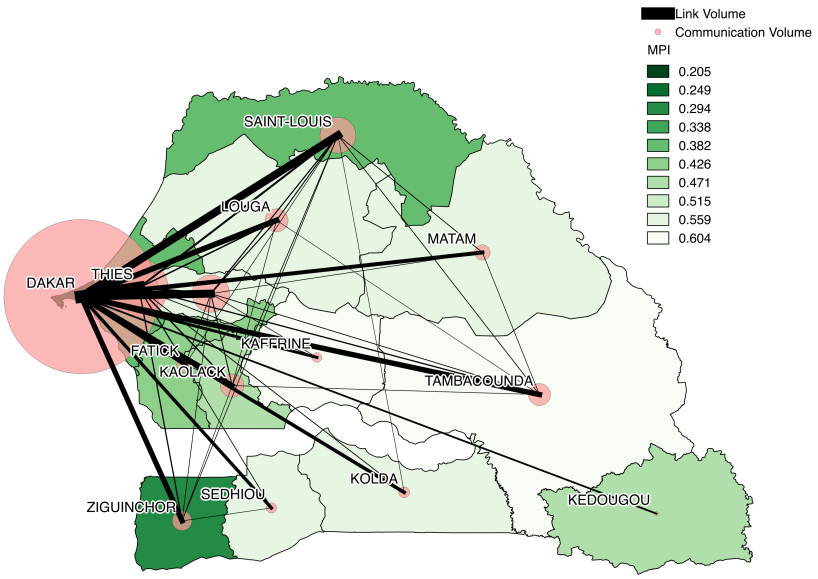


Figure 2: Virtual network for Senegal at Region-level with MPI (Multi-dimensional Poverty) as an overlay. Thickness of links indicate the volume of calls and texts exchanged between a pair of regions. Size of the circle at each region indicates the total number of incoming and outgoing calls and texts 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] ($numcalls \propto \frac{p_i p_j}{d_{ij}^{\alpha}}$). However, instead of using the population of the two regions (p_i and p_j), we use the number of sites to get a estimate of the population *with* mobile phones. We set the exponent α 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 region 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 ($\widehat{M}^r$) 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 ($= \sum_{i \neq j} \widehat{M}^r_{ij}$ 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 $\widehat{M}^r$,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:

$$x_i = \frac{1}{\lambda} \sum_j \widehat{M}^r_{ij} x_j \tag{4.2}$$

where λ is some constant. In matrix notation, this can be written as $\lambda \mathbf{x} = \widehat{M}^r \mathbf{x}$, such that $\mathbf{x}$ is the eigenvector of the matrix $\widehat{M}^r$ 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

Measure	H		A		MPI	
	corr	p-value	*corr*	p-value	*corr*	p-value
PageRank	-0.87	6×10^{-5}	-0.81	0.0004	-0.82	0.0003
Eigenvalue Centrality	-0.83	0.0002	-0.80	0.0005	-0.79	0.0007
Gravity Residual	-0.81	0.0003	-0.76	0.0015	-0.79	0.0007
Introversion	0.82	0.0002	0.70	0.0040	0.79	0.0006
Activity (*Normalized*)	-0.81	0.0008	-0.76	0.0003	-0.79	0.0015
Activity (*Raw*)	-0.80	0.0006	-0.68	0.0075	-0.71	0.0040

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.

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 $\widehat{M}^r$ 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:

$$Residual_i = \sum_j \widehat{M}^r_{ij} \tag{4.3}$$

  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:

$$Introversion_i = \frac{M^r_{ii}}{\sum_j M^r_{ij}} \tag{4.4}$$

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 or headcount ratio (H) of poverty, i.e., the proportion of the population that is multidimensionally poor and the average intensity (A) 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:

$$\tilde{H}_i = -708.32 \times PageRank_i + 131.94 \tag{4.5}$$

$$\tilde{A}_i = -346.66 \times PageRank_i + 84.58 \tag{4.6}$$

Finally, we combine the two estimates to predict the MPI as:

$$\widetilde{MPI}_i = \frac{\tilde{H}_i}{100} \times \frac{\tilde{A}_i}{100} \tag{4.7}$$

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.

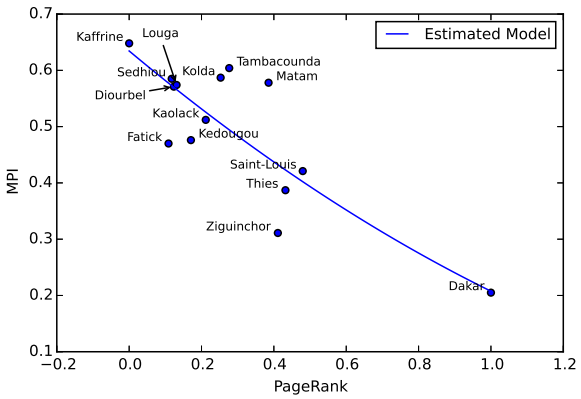


Figure 4: Estimated model for predicting MPI using the Page Rank feature.

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 $\widehat{M}^a$.

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.

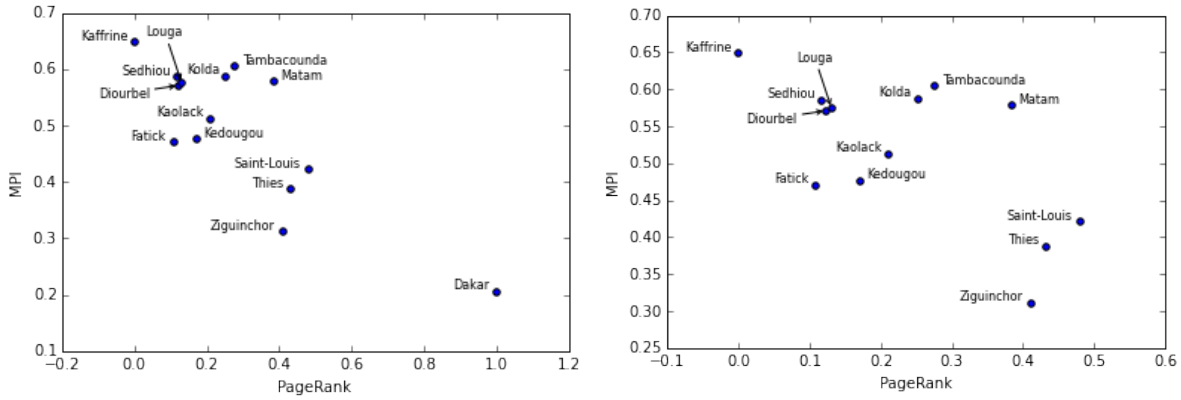
5.2.1 Localization of Users For each user we consider the calls made between 8 PM and 12 PM on each of the 365 days of the year. We measure the following quantities:

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).



(a) With Dakar

(b) Without Dakar

Figure 3: Illustrating influence of Dakar on the relationship between MPI and Page Rank for regions.

Measure	H		A		MPI	
	corr	p-value	*corr*	p-value	*corr*	p-value
PageRank with Dakar	-0.87	6×10^{-5}	-0.81	0.0004	-0.82	0.0003
PageRank without Dakar	-0.68	0.01	-0.65	0.016	-0.64	0.018

Table 2: Pearson's r Correlation of H, A and MPI with Pagerank of the regions considering Dakar and NOT considering Dakar.

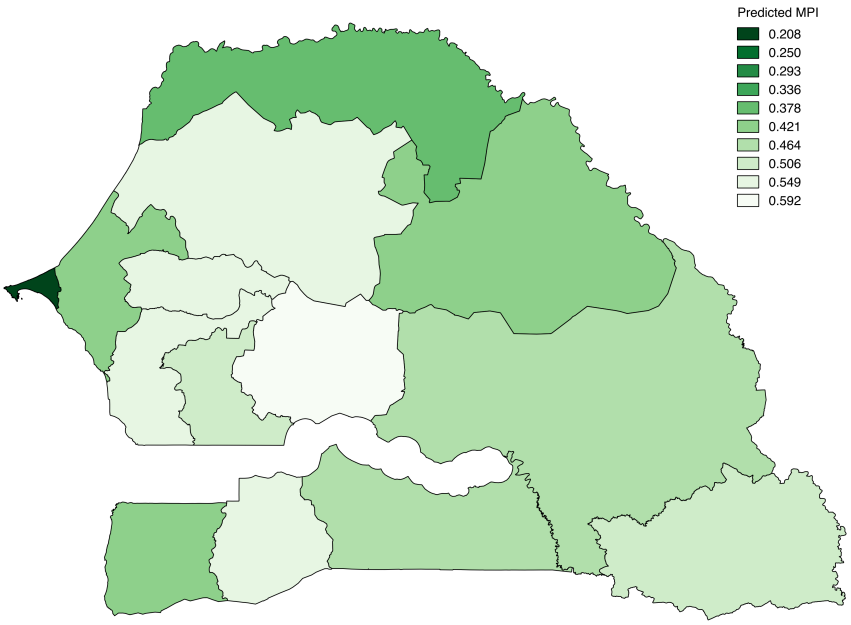


Figure 5: Predicted region level Poverty map using the virtual network.

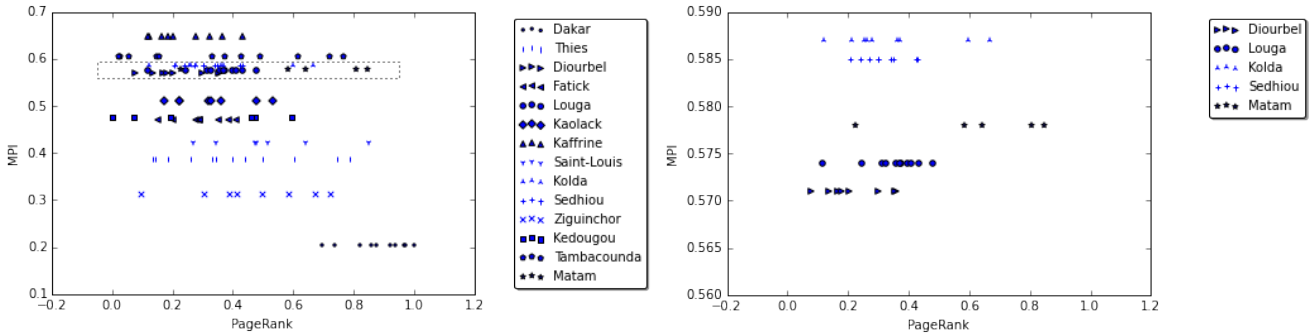


Figure 6: Visual depiction of what happens when MPI is calculated at a region level. All arrondissements within each region are assigned the same poverty, but they have varying centrality measures, signifying importance! Thus, need to generate finer povery maps for targeted eradication of poverty. The dashed box in the top panel is expanded in the bottom panel.

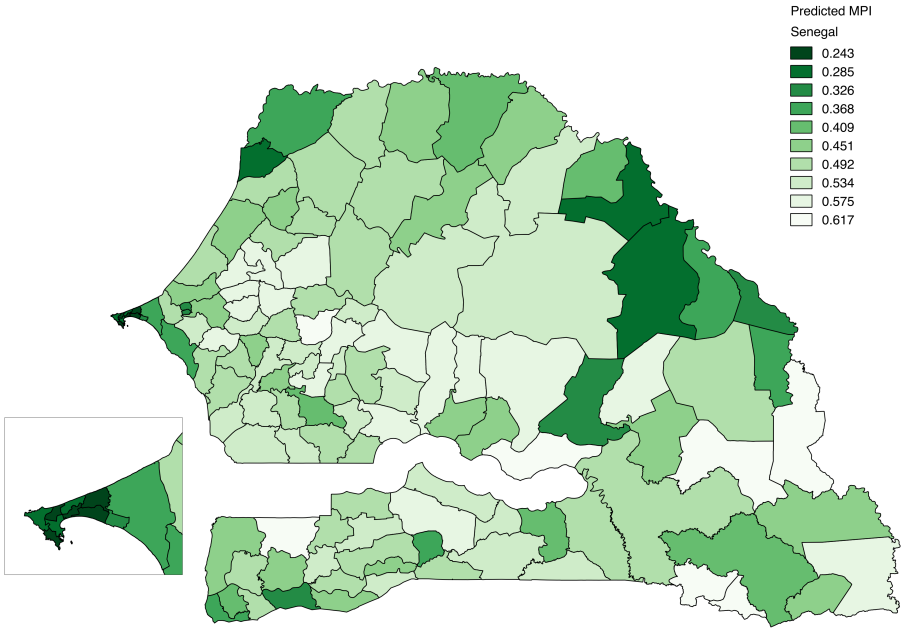


Figure 7: Predicted Arrondissement level Poverty map using the virtual network.

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.

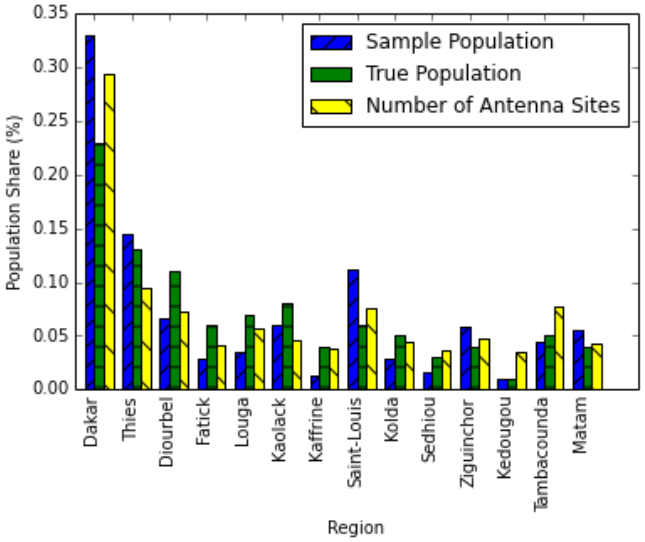


Figure 8: Comparing the region-level population share for the indicator sample and the true population.

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:

$$\tilde{H}_i = -302.65 \times PIC_i + 119.35 \quad (5.8)$$

$$\tilde{A}_i = -151.53 \times PIC_i + 78.84 \quad (5.9)$$

The MPI is calculated by multiplying the two estimates

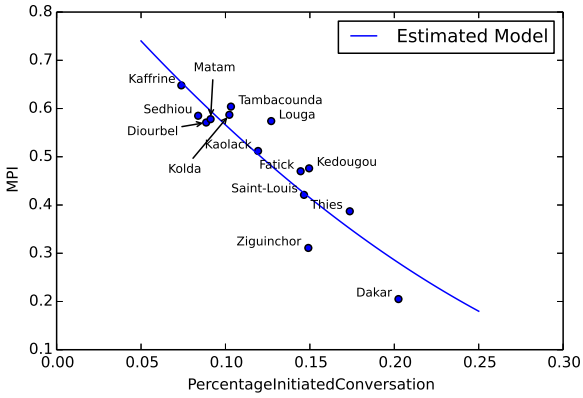


Figure 9: Estimated model for predicting MPI using the Page Rank feature.

(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

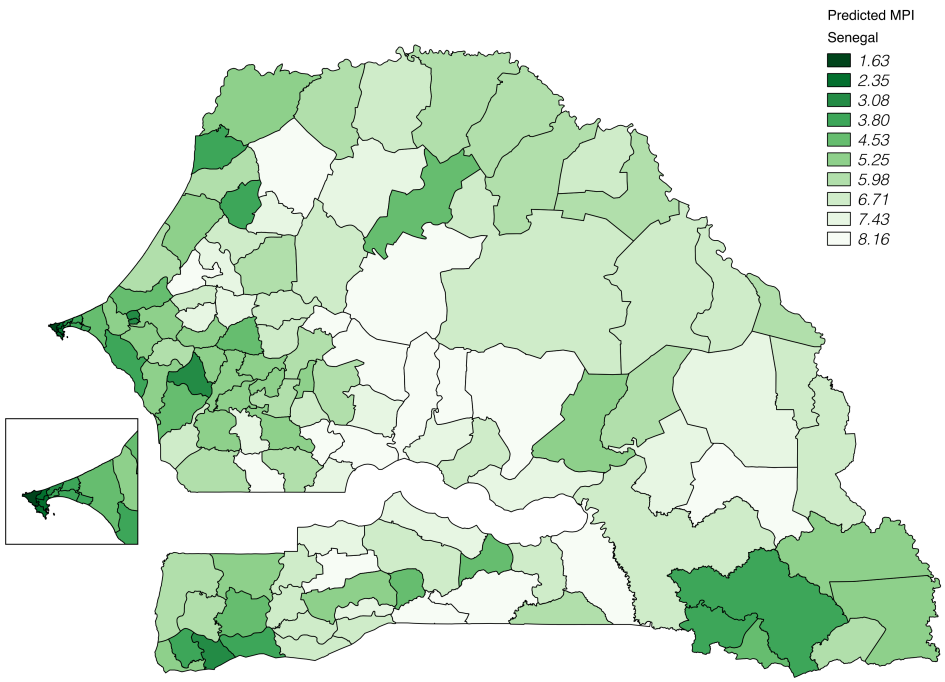


Figure 10: Predicted arrondissement-level map of MPI for Senegal using Behavioral Indicators.

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

[1] `http://www.itu.int/en/ITU-D/Statistics/Pages/intlcoop/mdg/default.aspx`.

[2] `http://data.worldbank.org/indicator/IT.CEL.SETS.P2`.

[3] `http://mdgs.un.org/unsd/mdg/Metadata.aspx?IndicatorId=2`.

[4] `https://www.imf.org/external/pubs/ft/dp/2013/afr1304.pdf`.

[5] F. Calabrese, F. C. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti. The geography of taste: Analyzing cell-phone mobility and social events. In *Proceedings of the 8th International Conference on Pervasive Computing*, Pervasive'10, pages 22–37, Berlin, Heidelberg, 2010. Springer-Verlag.

[6] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[7] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, May 2010.

[8] V. Frías-Martínez, V. Soto, J. Virseda, and E. Frías-Martínez. Can cell phone traces measure social development? In *NetMob*, 2013.

[9] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, (07):L07003+, July 2009.

[10] M. Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[11] C. Smith, A. Mashhadi, and L. Capra. Ubiquitous sensing for mapping poverty in developing countries. In *3rd International Conference on the Analysis of Mobile Phone Datasets (NetMob 2013) - Data for Development (D4D) challenge.*, 2013.

[12] V. Soto, V. Frías-Martínez, J. Virseda, and E. Frías-Martínez. Prediction of socioeconomic levels using cell phone records. In *User Modeling, Adaption and Personalization - 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings*, pages 377–388, 2011.

T02

Deviations from the norm: Detecting regularities and anomalies in human mobility patterns

Gijs Joost Brouwer[1], and Foster Provost[2]

[1] Integral Ad Science, New York, USA

[2] Stern School of Business, New York University, New York, USA

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[1,2], socio-political instabilities like mass protests[3] 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 response 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[4]. Described as quasi-democratic, Senegal is stable compared to other countries within its region[5]. 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[6]. It also experiences heavy urbanization, leading to extreme poverty in these urban areas[7]. In addition, the literacy rate in 2011 was estimated to be 52.1%, making Senegal the 9th most illiterate country in the world[8]. 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 Senegali owned a cellphone, 13% of which were internet-enabled smart phones[9]. 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[10]. 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[11-13], prevent epidemics[14-16] as well as insights into the effects of weather[17] and conflict[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.

Methods

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 ~7.4 million users[19]. In the current study, we used one out of the three data sets provided by Orange. This fine-grained date 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.

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*x*n* square matrix (*n* = number of unique towers visited by the user), each $T_{i,j}$ 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[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[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 *n*x*m* 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.

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[22,23].

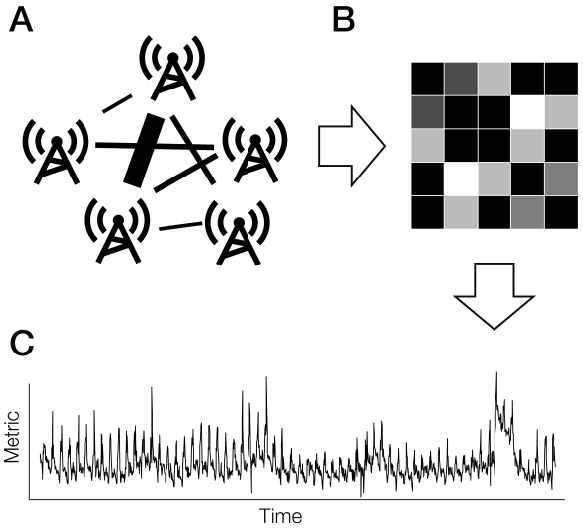


Figure 1. Steps in the computation of transition probability, distance and surprise. (A) For each user, we collected the number of times they visited each available cell phone tower. From this we compute the mode tower (the tower most frequently visited) for this user. In addition, we computed the measure of surprise of seeing the user at a specific tower, which we defined as (1 - frequency of visiting the tower) divided by the 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 - frequency of a cellular event occurring at this tower) divided by the maximum frequency observed across towers observed for a particular user. 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 *n*x*m* 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 permutated 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 *n*x*m* 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 *n*x*20* 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

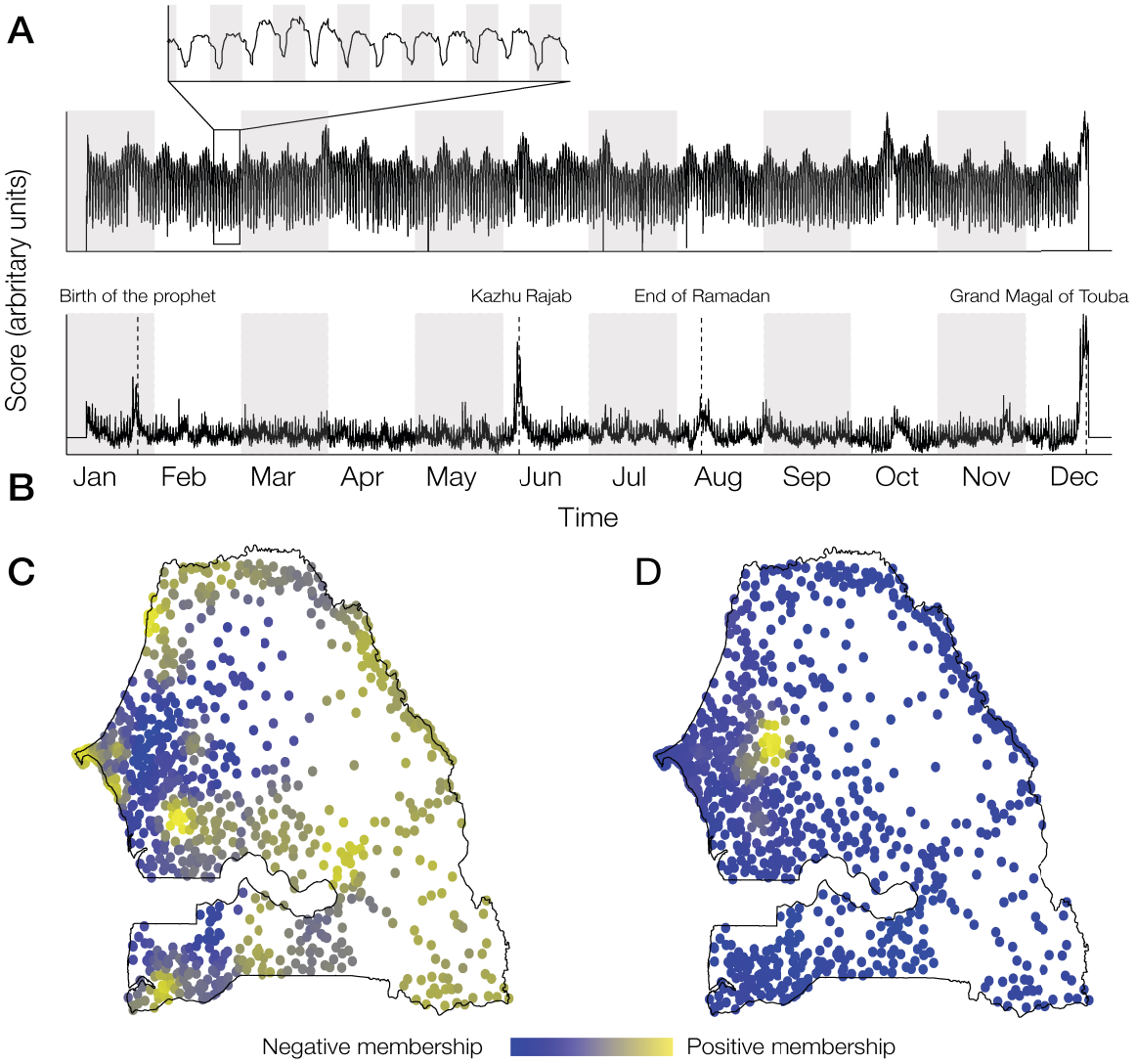


Figure 2: Mobility patterns, part 1. (A) The main source of variance in the mobility data was explained by the regular daily patterns of commuting back and forth between the home location and work. Inset shows the regular patterns observed over a few days. **(B)** The second source of variation revealed four major religious observances: the Grand Magal of Touba, the end of Ramadan, Kazhu Rajab and the birth of the prophet Mohammed. **(C)** Cell phone tower membership to the regular daily commuting pattern was high for urbanized regions, low for more rural regions. **(D)** 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 $\mathbf{D}_{ij}$ 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[24-26], here we computed them using the algorithm described in detail by Arenas et. al.[27]. 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, we computed the probability of a certain user transitioning between two particular towers, the distance between the destination tower and the most commonly visited cell phone tower, and the surprise of seeing the user at a particular tower. We used principal component analysis (PCA) to reveal patterns of human mobility shared across cell phone towers. After reducing the dimensionality of the original *n*x*m* matrix (where *n* = number of time points and *m* the number of towers with sufficient amount of traffic), we visualized the components that explain the largest amount of variance. We did this as a proof of principle; to show that the components derived from the mobility patterns were indeed meaningful. We found that the three metrics resulted in very similar component patterns

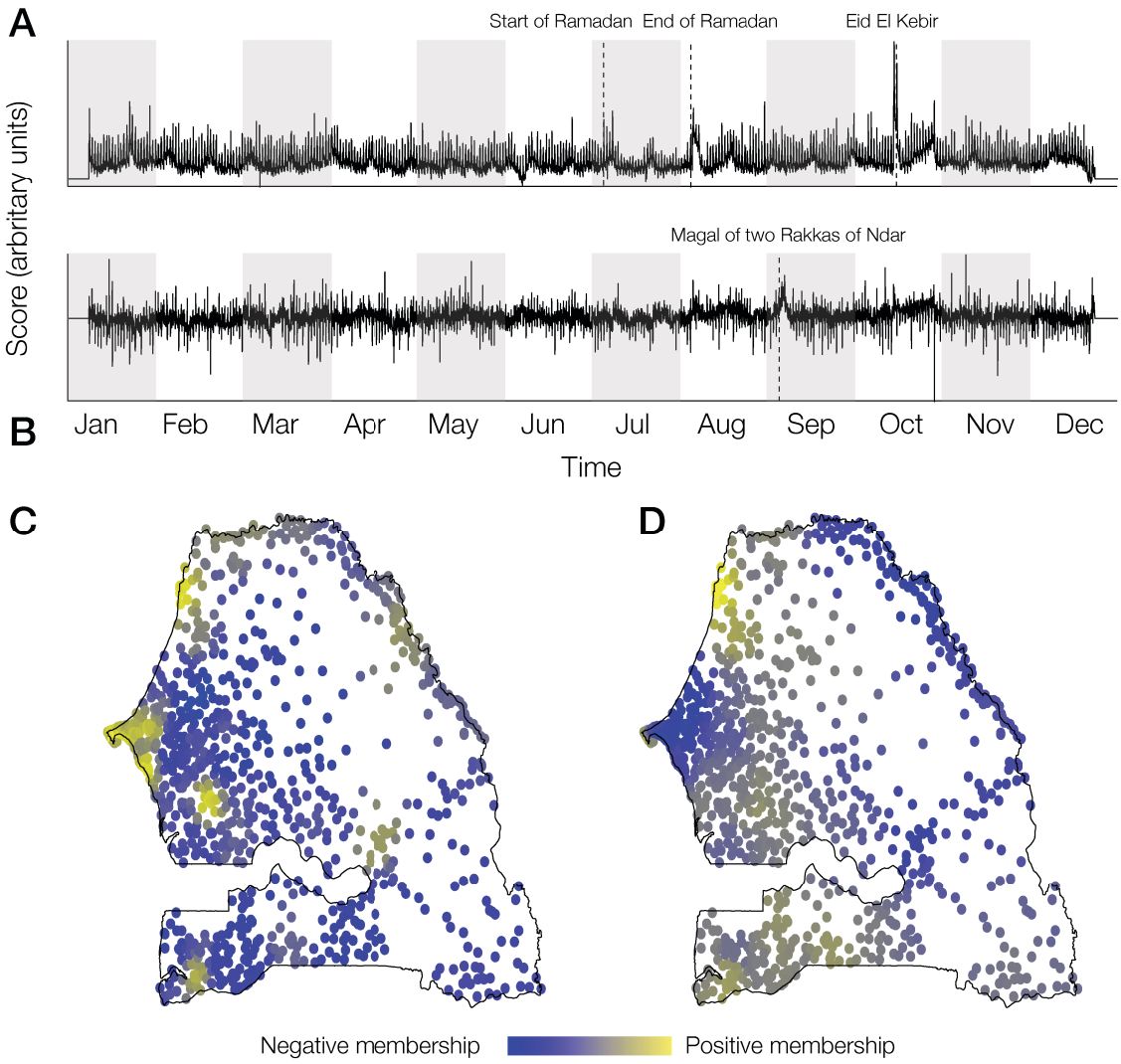


Figure 3: Mobility patterns, part 2. (A) The third source of variance in the mobility data was explained by the religious holiday of Eid El Kebir (feast of sacrifice) and a reduction in the average distance metric within the month of Ramadan. **(B)** The fourth source of variation revealed a singular 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 i 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

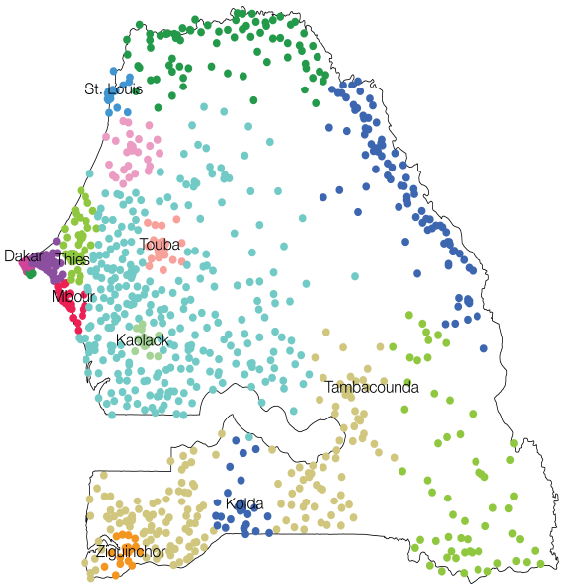


Figure 4. Component-based communities. Community analysis of the weights placed on the 20 first components revealed a total of 13 distinct communities. These communities were geographically separated, even though no direct information about distance was used to detect the communities. Each dot represents 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 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 drop, 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 by. 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.

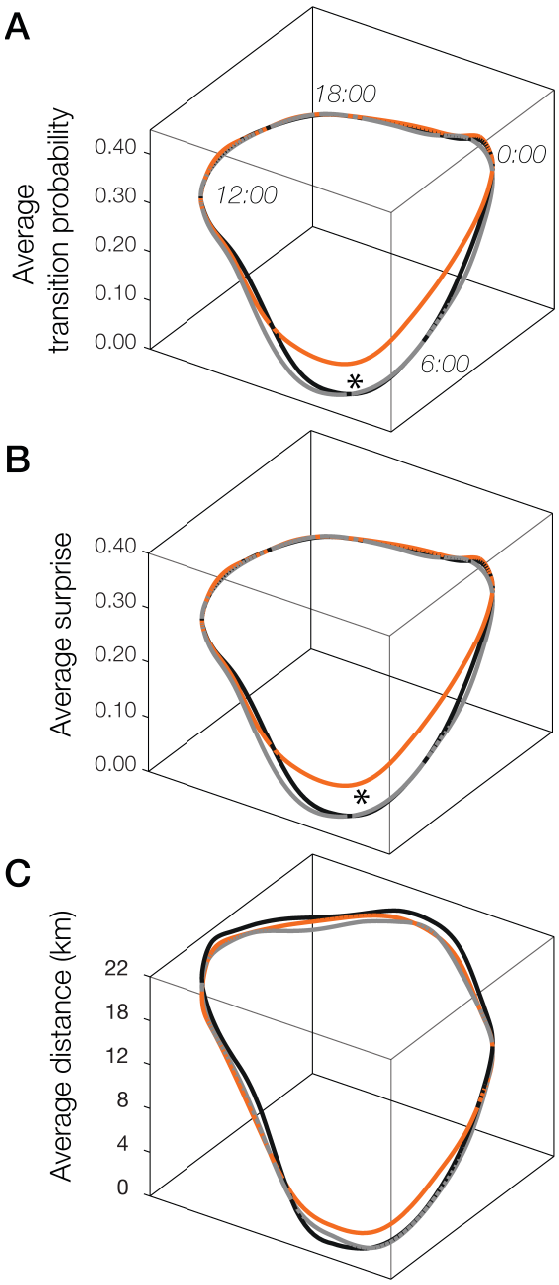


Figure 5. Daily activity patterns prior (black), after (gray) and during the month of Ramadan (red). Each line represents a circular trace of a full day of measurements, starting at midnight, and moving clockwise. **(A)** During Ramadan, nightly transition probabilities were significantly higher than 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 their 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 Ramada. 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[21,28]. 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[1-3,29-32], 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 Cote d'Ivoire. For now, it has also has 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.
2. Bagrow JP, Wang DS, Barabasi AL (2011) Collective response of human populations to large-scale mergencies. Plos One 6:e17680
3. Lu X, Bengtsson L, and Holme P (2012). Predictability of population displacement after the 201 Haiti earthquake. Proc. Nat. Aca. Sci. 109:11576-11581
4. Stewart F (2002) Root causes of violent conflict in developing countries. BMJ 324-342-345
5. https://freedomhouse.org/report/freedom-world/2013/senegal#.VH3GlWTF9yM
6. Piel G (1997) The urbanization of poverty worldwide. Challenge 40:58-68
7. Ndiaye M (2007) Senegal agricultural sutuation country report 2007. USDA Foreign Agricultural service
8. Unesco Institute for Statistics. http://www.uis.unesco.org/Pages/default.aspx
9. http://www.pewglobal.org/category/publications/survey-reports/2013/

10. de Montjoye YA, Smoreda Z, Trinquart R, Ziemlicki C, and Blondel VD (2014). D4D-Senegal: The second mobile phone dta for development challenge. http://www.d4d.orange.com/en/home
11. Paraskevopoulos P, Dinh TC, Dashdorj Z, Palpanas T, Serafini L (2013) Identification and characterization of human behavior patterns from mobile phone data. Orange D4D Mobile Phone Data for Development
12. Angelakis V, Gundlegard D, Rajna B, Rydergren C, Vrotsou K, Carlsson R, Forgeat J, Hu TH, Liu EL, Moritz S, Zhao S, and Zheng Y. (2013) Mobility modeling for transport efficiency - analysis of travel characeteristics based on mobile phone data. Orange D4D Mobile Phone Data for Development
13. McInerney J, Rogers A, and Jennings NR (2013) Crowdsourcing physical delivery using the existing routine mobility of a local population.
14. Enns EA, and Amuasi JH (2013) Human mobility and communication patterns in Cote d'Ivoire: a network perspective for malaria control
15. Saravanan M, Karthikeyan P, Aarthi A, Kiruthika M, and Suganya S (2013) Exploring community structure to understand diesase spread and control using mobile call detail records. Orange D4D Mobile Phone Data for Development
16. Wesolowski A, and Buckee CO (2013) Are gravity models appropriate for estimating the spatial spread of malaria?
17. Craveiro JP, Ramos FMV, Kanjo E and El Mawass N (2013) Towards an early warning system: the effect of weather on mobile phone usage. Orange D4D Mobile Phone Data for Development
18. Linardi S, Kalyanaraman S, and Berger D (2013) Does conflict affact human mobility and cellphone usage? Evidence from Cote d'Ivoire. Orange D4D Mobile Phone Data for Development
19. https://www.telegeography.com/products/commsupdate/articles/2013/12/17/artp-senegal-mobile-market-tops-12-721m-at-end-september-as-expresso-leapfrogs-tigo/
20. Murphy KP (2012) Machine Learning: a probabilistic perspective. MIT, Massachusetts.
21. Eagle N and Pentland AS (2009) Eigenbehaviors: identifying structure in routine. Behav Ecol Sociobiol 63:1057-1066
22. Song CM, Qu ZH, Blumm N and Barabasi AL (2010) Limits of preditacbility in human mobility. Science 327:1018-1021
23. Gonzalez MC, Hidalgo CA and Barabasi AL (2008) Understanding individual human mobility patterns. Nature 453:779-782
24. Expert P, Evans TS, Blondel VD, and Lambiotte R (2011) Uncovering space-independent communities in spatial networks.
25. Blondel VD, Guillaume JL, Lambiotte R, and Lefebvre E (2008) Fast unfolding of communites in large networks. J. Stat. Mech. P10008.
26. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E74:036104
27. Arenas A, Fernandez A, and Gomez S (2008). Analysis of the structure of complex networks at different resolution levels. N. J. Phys. 10:053039
28. Simini F, Gonzales MC, Maritan A, and Barabasi AL (2012) A universal model for mobility and migration patterns. Nature 484:96-100.
29. Kavanaugh A, Yang S, Sheetz S, Li LT, and Fox E (2011) Between a Rock and a Cell Phone: Social Media Use during Mass Protests in Iran, Tunisia and Egypt. Technical Report TR-11-10, Computer Science, Virginia Tech.
30. Ryan, Yasmine. How Tunisia's revolution began. January 26, 2011. AlJazeera.net.
31. Sheetz S, Kavanaugh A, Quek F, Kim BJ, and Liu, SC (2010) Expectation of connecteness and cell phone use in crises. Int. J. Erm Man.
32. Vieweg, S., Palen, L., Liu, S., B., Hughes, A. L., & Sutton, J. (2008, May 2008). Collective Intelligence in disaster: Examination of the phenomenon in the aftermath of the 2007 Virginia Tech shooting. Paper presented at the Information Systems for Crisis Response and Management, Washington, DC.

T03

National and Regional Road Network Optimization for Senegal Using Mobile Phone Data

Yihong Wang[1], Gonçalo Homem de Almeida Correia[1], and Erik de Romph[*1,2]

[1]Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands
[2]DAT.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

*Corresponding author. E-mail: E.deRomph@tudelft.nl

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 occured 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:

$$T_{ij} = K_0 \frac{P_i P_j}{d_{ij}^2} \quad [1]$$

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 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):

$$T'_{ij} = K_1 {P_i}^a {P_j}^b f(c_{ij}) \quad [2]$$

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):

$$f(c_{ij}) = e^{-\beta c_{ij}} \quad [3]$$

$$f(c_{ij}) = {c_{ij}}^{-n} \quad [4]$$

$$f(c_{ij}) = {c_{ij}}^{n} e^{-\beta c_{ij}} \quad [5]$$

where, β 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):

$$I_{ij} = K_2 \frac{P_i P_j}{d_{ij}^2} \quad [6]$$

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:

$$\frac{T_{ij}}{I_{ij}} = \frac{K_0}{K_2} \quad [7]$$

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

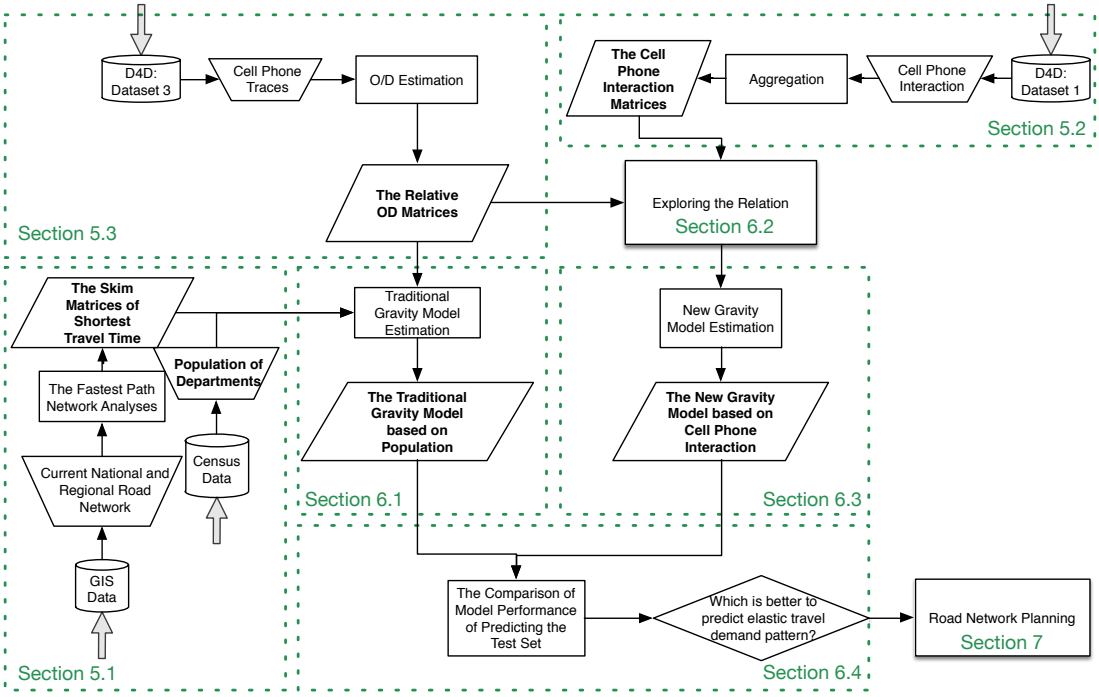


Figure 1: The Flowchart of Methodology

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

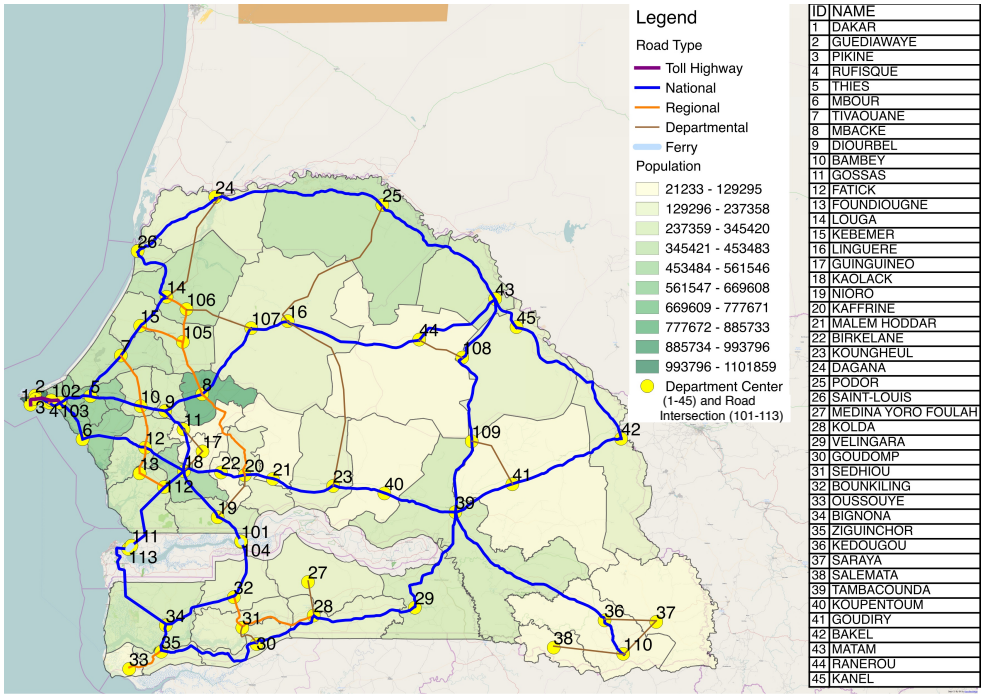


Figure 2: Cleaned Road Network and Population Distribution of Senegal

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 rth 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}, $t(D_{u(r-1)}, D_{ur})$ 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 rth 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.

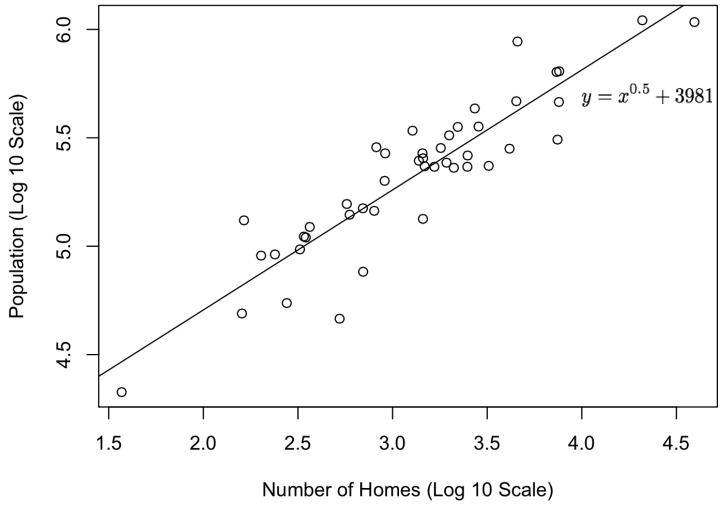


Figure 3: The Linear Regression between Population per Department and the Number of Sampled Users' Homes per Department

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-Diamniadio 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.

OD Pair	Average Estimated Number of Undirected Trips between OD Pairs Made by Sampled Users per Month	
	Before the opening of new highway section (From January to July)	After the opening of new highway section (From August to December)
1-4	54563	57932
1-5	15023	16447
1-6	14287	13980
2-4	6219	6587
2-5	1937	2186
2-6	1444	9209
3-4	87416	92972
3-5	10806	11006
3-6	8981	8452

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-Diamniadio 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].

$$T_{ij} = K_1 {P_i}^a {P_j}^b t_{ij}^{-n} \quad [8]$$

$$T_{ij} = K_1 {P_i}^a {P_j}^b e^{-\beta t_{ij}} \quad [9]$$

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. β 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:

$$T_{ij} = (1.27e-09) \times {P_i}^{1.07067} \times {P_j}^{1.08714} \times t_{ij}^{-2.53015} \quad [10]$$

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

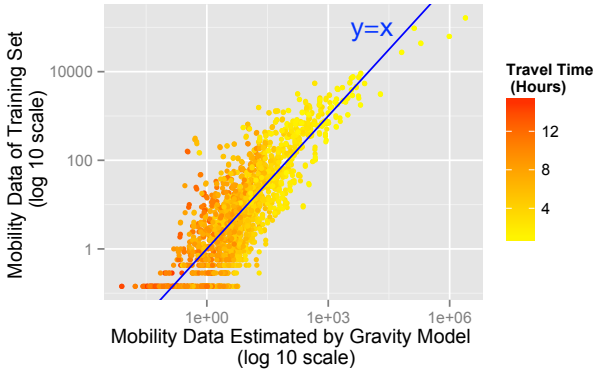


Figure 4: The Fitness of Traditional Gravity Model Based on Population with the Power Parameter for Cost Function

Parameter	Estimate	Std. Error	t value
$\log_{10} K_1$	-8.89705	0.34927	-25.47
a	1.07067	0.04305	24.87
b	1.08714	0.04305	25.25
n	2.53015	0.05034	50.26
Adjusted R-Squared	0.7299		

Table 2: The Estimated Values of Parameters and Adjusted R-Squared

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.

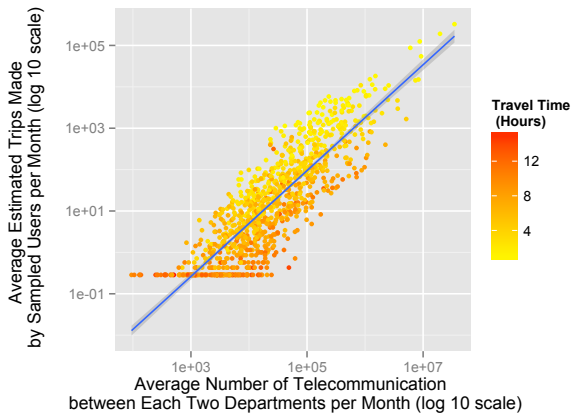


Figure 5: The Relation between Cell Phone Interaction Data and the Estimated Mobility Data before August 1st, 2013

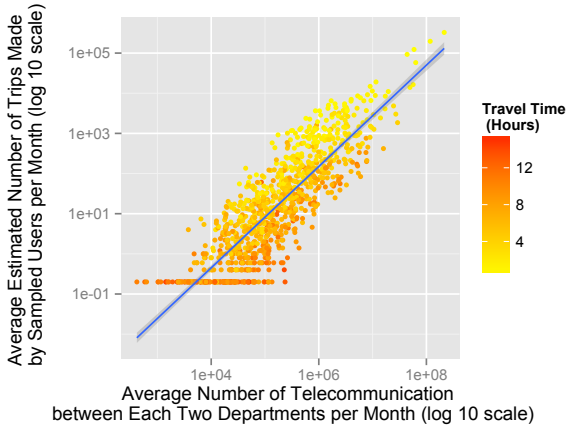


Figure 6: The Relation between Cell Phone Interaction Data and the Estimated Mobility Data after August 1st, 2013

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:

$$T_{ij} = K_3 {I_{ij}}^{\alpha} {t_{ij}}^{-m} \quad [11]$$

$$T_{ij} = K_3 {I_{ij}}^{\alpha} e^{-\theta t_{ij}} \quad [12]$$

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. α is the parameter for I_{ij}. Since we have observed a close-to-linear relationship between mobility data and interaction data, we assume α 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. θ 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.

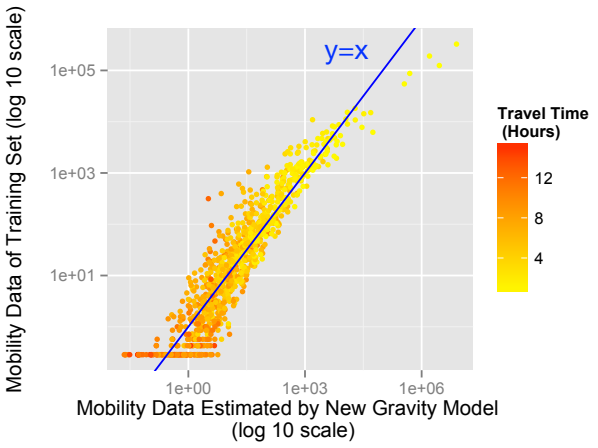


Figure 7: The Fitness of New Gravity Model Based on Cell Phone Interaction with the Power Parameter for Cost Function

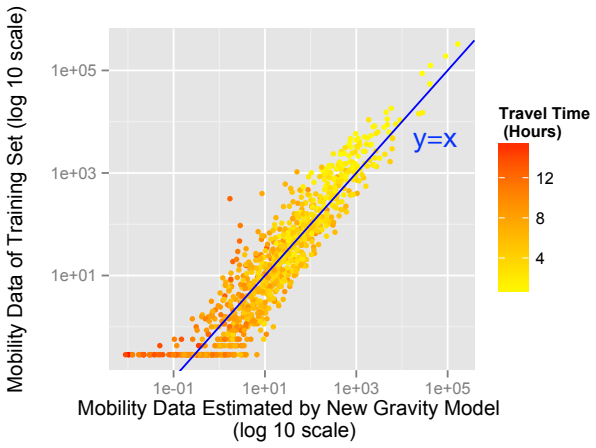


Figure 8: The Fitness of New Gravity Model Based on Cell Phone Interaction with the Exponential Parameter for Cost Function

Parameter	Estimate	Std. Error	t value
$\log_{10} K_3$	-1.8021	0.11102	-16.23
α	0.94989	0.01976	48.16
m	1.72147	0.05384	31.98
Adjusted R-Squared	0.8724		

Table 3: The Estimated Values of Parameters and Adjusted R-Squared of New Gravity Model with Power Parameter

Parameter	Estimate	Std. Error	t value
lnK_3	-5.31188	0.25115	-21.15
α	1.00108	0.02041	49.05
θ	0.34970	0.01236	28.3
Adjusted R-Squared	0.8566		

Table 4: The Estimated Values of Parameters and Adjusted R-Squared of New Gravity Model with Exponential Parameter

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:

$$T_{ij} = 0.00493 \times {I_{ij}}^{1.00108} \times e^{-0.3497 \times t_{ij}} \quad [13]$$

The value of α, 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.

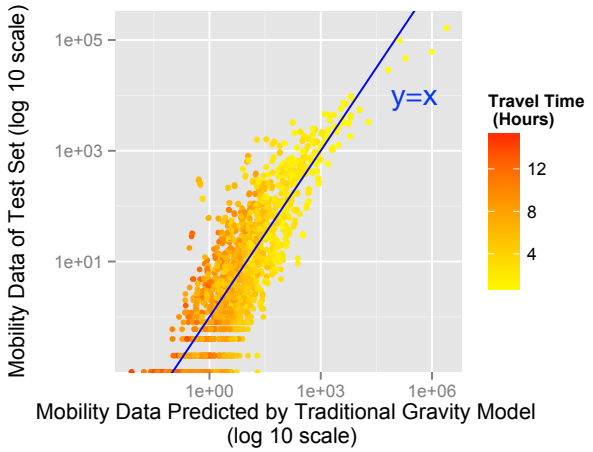


Figure 9: Model Performance of the Estimated Traditional Gravity Model Based on Population (RMSE: 157229.3)

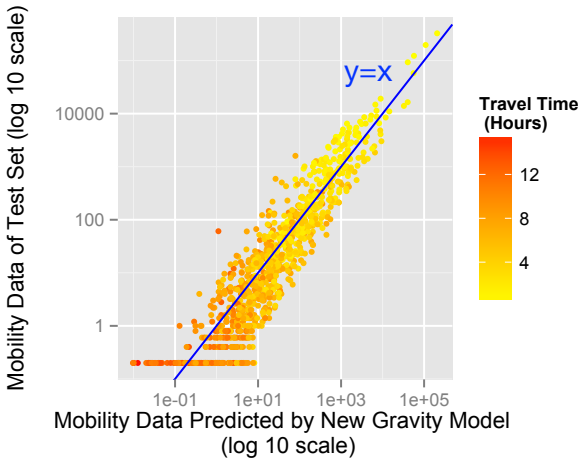


Figure 10: Model Performance of the Estimated New Gravity Model Based on Cell Phone Interaction (RMSE: 5590.407)

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-Diamniadio 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

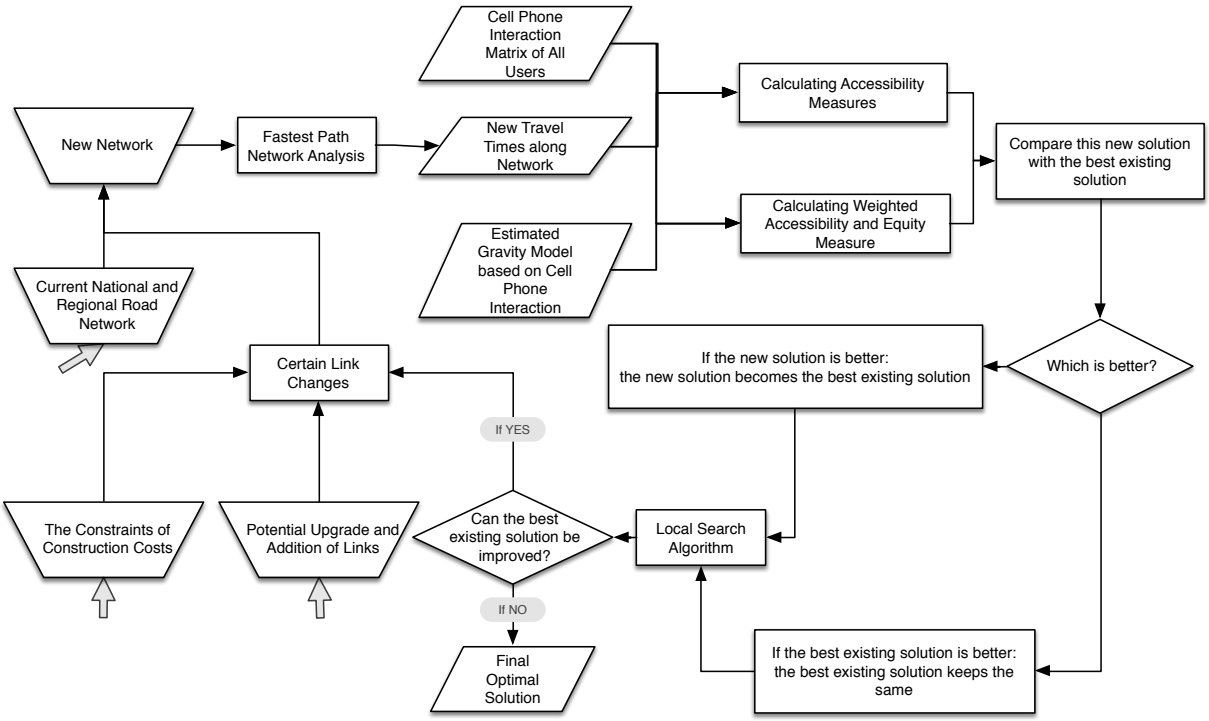


Figure 11: The Flowchart of Methodology Regarding Road Network Planning

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

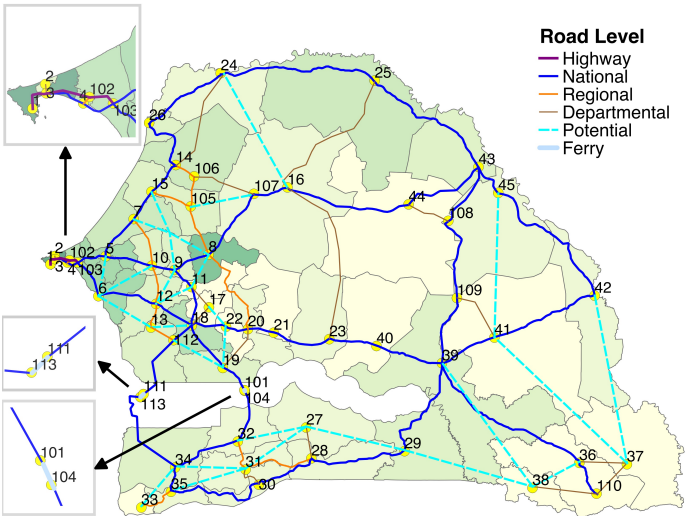


Figure 12: Potential Links to Be Added and Existing Links to Be Upgraded

	Upgraded Level	Potential	Departmental	Regional	National	Highway	Bridge
Existing Level	Average Speed	0 km/h	30 km/h	45 km/h	60 km/h	80 km/h	60 km/h
Potential	0 km/h	–	–	1.2	1.6	4	–
Departmental	30 km/h	–	–	0.2	0.6	3	–
Regional	45 km/h	–	–	–	0.4	2.8	–
National	60 km/h	–	–	–	–	2.4	–
Ferry	0-2 km/h	–	–	–	–	–	8

Table 5: Design Characteristic of Different Road Levels and Relative Unit Costs for Road Construction and Upgrading

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:

$$Z = \sum_{i \in N} A_i \times \frac{P_i}{P} \quad \text{and} \quad A_i = \sum_{j \in N \setminus i} P_j \times f(c_{ij}) \qquad [14]$$

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

based on the new gravity model is adapted as follows:

$$Z = \sum_{i \in N} A_i \times \frac{1}{2I} \quad \text{and} \quad A_i = \sum_{j \in N \setminus i} I_{ij} \times f(c_{ij}) \tag{15}$$

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]:

$$f(c_{ij}) = e^{-0.3497 c_{ij}} \quad \text{and} \quad c_{ij} = t_{ij} \tag{16}$$

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):

$$E = \sum_{i \in N_{low}} P_i \times A_i \quad \text{and} \quad A_i = \sum_{j \in N \setminus i} P_j \times f(c_{ij}) \tag{17}$$

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:

$$E = \sum_{i \in N_{low}} A_i \quad \text{and} \quad A_i = \sum_{j \in N \setminus i} I_{ij} \times f(c_{ij}) = \sum_{j \in N \setminus i} I_{ij} \times e^{-0.3497 t_{ij}} \tag{18}$$

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:

$$maxV = w_Z \times \frac{Z(y) - Z_0}{Z_B - Z_0} + w_E \times \frac{E(y) - E_0}{E_B - E_0} \tag{19}$$

subject to:

$$Z(y) = \sum_{i \in N} \sum_{j \in N \setminus i} I_{ij} \times e^{-0.3497 \times t_{ij}(y)} \times \frac{1}{2I}, \quad \forall i, j \in N \quad (i \neq j) \tag{20}$$

$$E(y) = \sum_{i \in N_{low}} \sum_{j \in N \setminus i} I_{ij} \times e^{-0.3497 \times t_{ij}(y)}, \quad \forall i, j \in N \quad (i \neq j) \tag{21}$$

$$\sum_{m \in M_l} y_{lm} = 1, \quad \forall l \in L \tag{22}$$

$$\sum_{m \in M_l} e_{lm} \times y_{lm} \leq b \tag{23}$$

$$T_{ij} \geq 0, \quad \forall i, j \in N, \quad l \in L, \quad y_{lm} \in \{0, 1\}, \quad l \in L, \quad m \in M \tag{24}$$

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 $y\{y_{lm}\}$; $y\{y_{lm}\}$ = 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.

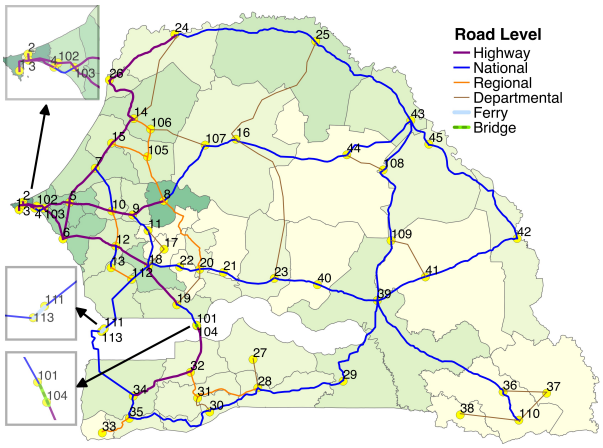


Figure 13: New Network of the Optimal Solution to Achieve the Single Efficiency Objective

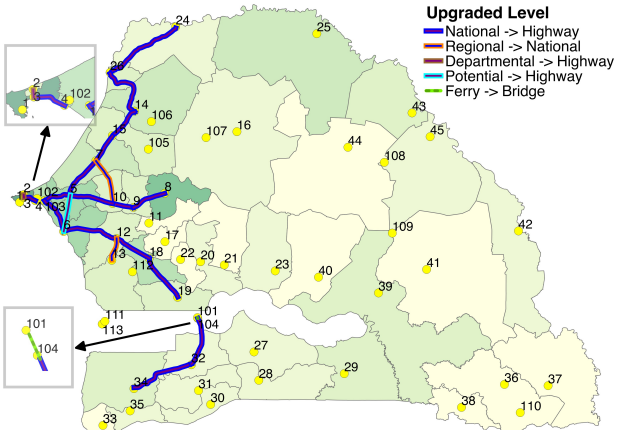


Figure 14: All Upgraded Links of the Optimal Solution to Achieve the Single Efficiency Objective

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-Diamniadio 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. 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.

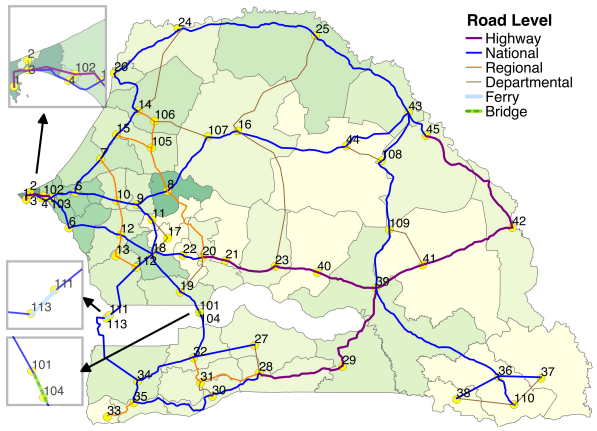


Figure 15: New Network of the Optimal Solution to Achieve the Single Equity Objective

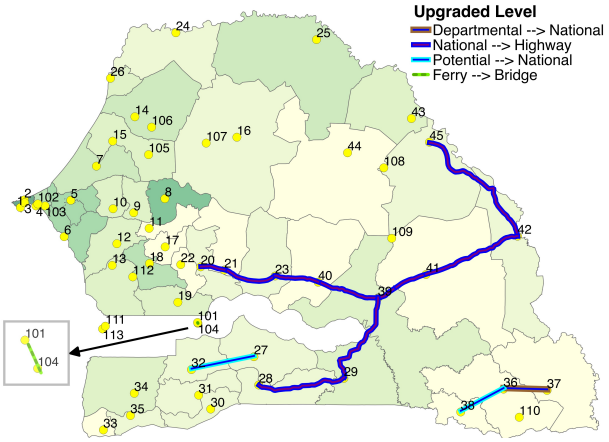


Figure 16: All Upgraded Links of the Optimal Solution to Achieve the Single Equity Objective

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.

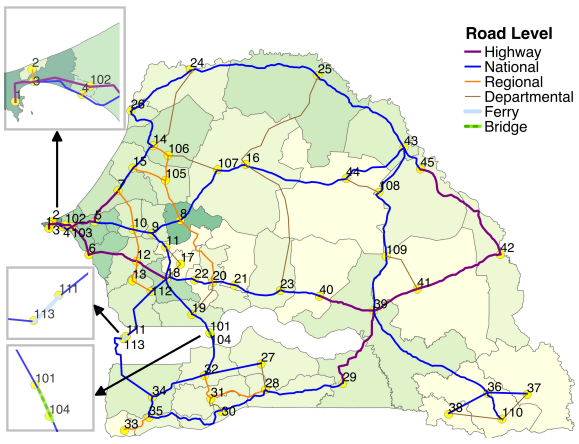


Figure 17: New Network of the Optimal Solution to Achieve both Efficiency and Equity Objective

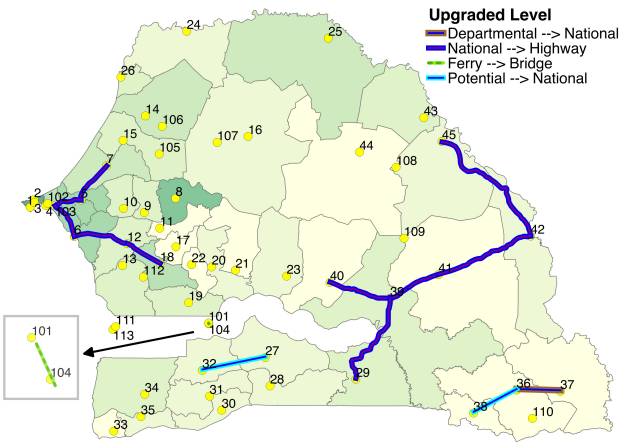


Figure 18: All Upgraded Links of the Optimal Solution to Achieve Both Efficiency and Equity Objective

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%.

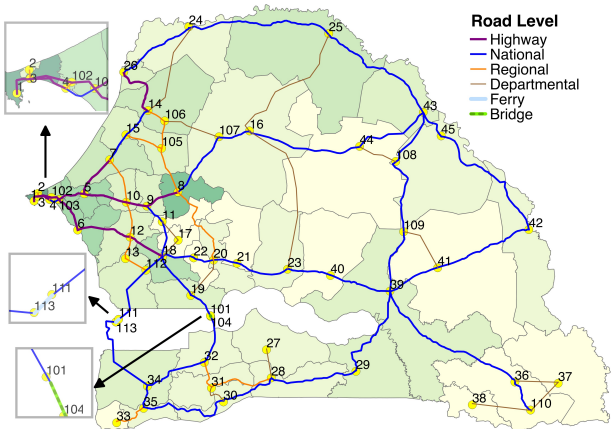


Figure 19: New Network of the Optimal Solution to Achieve Efficiency Objective Given 1/2 Budget

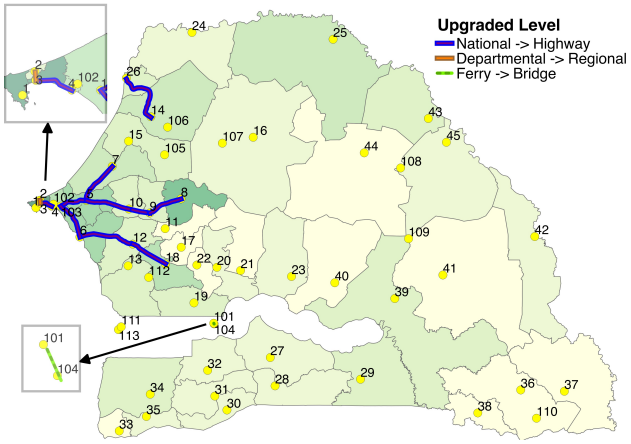


Figure 20: All Upgraded Links of the Optimal Solution to Achieve Efficiency Objective Given 1/2 Budget

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.

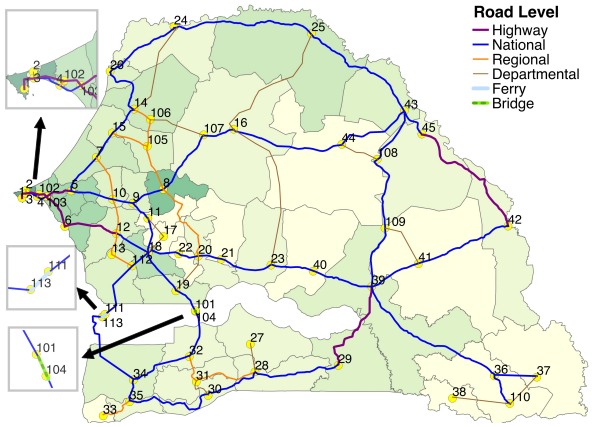


Figure 21: New Network of the Optimal Solution to Achieve Both Efficiency and Equity Objective Given 1/2 Budget

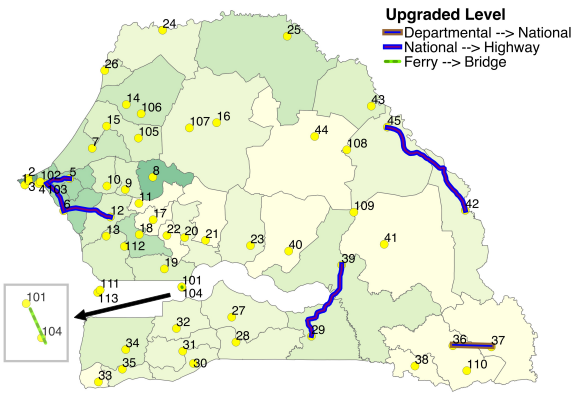


Figure 22: All Upgraded Links of the Optimal Solution to Achieve Both Efficiency and Equity Objective Given 1/2 Budget

Solution	Measure	Budget	
		50%	100%
For the Single Efficiency Objective	Z (Efficiency)	4.644%	6.548%
For Both Efficiency and Equity Objective	Z (Efficiency)	2.158%	3.537%
	E (Equity)	8.988%	18.341%

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 patters 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

ADBG (2014). *Executive Summary of the Environmental and Social Impact Assessment, Dakar Toll Road Project - Phase 2 Diamniadio-AIBD Section*. [Online; accessed 22-December-2014]. URL: http://www.afdb.org/fileadmin/uploads/afdb/Documents/Environmental-and-Social-Assessments/Senegal_-_Dakar_Toll_Road_-_Phase_2_-_Diamniadio-Aibd_Section_-_ESIA_Executive_Summary.pdf.

ANSD (2013). In:

Caceres, N, JP Wideberg, and FG Benitez (2007). "Deriving origin destination data from a mobile phone network." In: *Intelligent Transport Systems, IET* 1.1, pp. 15–26.

Calabrese, Francesco et al. (2011a). "Estimating origin-destination flows using mobile phone location data." In: *Pervasive Computing, IEEE*.

Calabrese, Francesco et al. (2011b). "Interplay between telecommunications and face-to-face interactions: A study using mobile phone data." In: *PloS one* 6.7, e20814.

Calabrese, Francesco et al. (2011c). "The connected states of america: Quantifying social radii of influence." In: *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. IEEE, pp. 223–230.

Cantarella, Giulio Erberto and Antonino Vitetta (2006). "The multi-criteria road network design problem in an urban area." In: *Transportation* 33.6, pp. 567–588.

Csáji, Balázs Cs et al. (2013). "Exploring the mobility of mobile phone users." In: *Physica A: Statistical Mechanics and its Applications* 392.6, pp. 1459–1473.

Demissie, Merkebe Getachew, Gonçalo Homem de Almeida Correia, and Carlos Bento (2013). "Exploring cellular network handover information for urban mobility analysis." In: *Journal of Transport Geography* 31, pp. 164–170.

Dios Ortuzar, Juan de, Luis G Willumsen, et al. (1994). *Modelling transport*. Wiley.

Eiffage (2013). *Dakr-Diamniadio Motorway*. [Online; accessed 18-December-2014]. URL: http://www.eiffage.com/en/home/le-groupe-1/our-projects/projets-acheves/autoroute-de-lavenir-au-senegal.html.

Eto, David (2012). "Senegal's mobile penetration passes 88 percent." In: URL: http://www.itnewsafrica.com/2012/10/senegals-mobile-penetration-passes-88-percent/.

Hoteit, Sahar et al. (2014). "Estimating human trajectories and hotspots through mobile phone data." In: *Computer Networks* 64, pp. 296–307.

Hsiao, Ming-Hsiung (2007). "Predicting Aggregate Travel Demands of Japan under the Impact of Telecommunications." In: *Proceedings of the Eastern Asia Society for Transportation Studies*. Vol. 2007. 0. Eastern Asia Society for Transportation Studies, pp. 90–90.

Kamargianni, M and A Polydoropoulou (2013). "Does Social Networking Substitute for or 20 Stimulate Teenagers' Travel? Findings from a Latent Class Model." In: *3rd* 21.

Krings, Gautier et al. (2009). "Urban gravity: a model for inter-city telecommunication flows." In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.07, p. L07003.

McNally, Michael G (2008). "The four step model." In: *Center for Activity Systems Analysis*.

Michalewicz, Zbigniew and David B Fogel (2004). *How to solve it: modern heuristics*. Springer.

Mokhtarian, Patricia L (2002). "Telecommunications and travel: The case for complementarity." In: *Journal of Industrial Ecology* 6.2, pp. 43–57.

Montjoye, Yves-Alexandre de et al. (2014). "D4D-Senegal: The Second Mobile Phone Data for Development Challenge." In: *arXiv preprint arXiv:1407.4885*.

Nanni, Mirco et al. (2014). "Transportation Planning Based on GSM Traces: A Case Study on Ivory Coast." In: *Citizen in Sensor Networks*. Springer, pp. 15–25.

Nobis, Claudia and Barbara Lenz (2009). "Communication and mobility behaviour–a trend and panel analysis of the correlation between mobile phone use and mobility." In: *Journal of Transport Geography* 17.2, pp. 93–103.

Plaut, Pnina O (1997). "Transportation-communications relationships in industry." In: *Transportation Research Part A: Policy and Practice* 31.6, pp. 419–429.

Ratti, Carlo et al. (2006). "Mobile landscapes: using location data from cell phones for urban analysis." In: *Environment and Planning b Planning and Design* 33.5, p. 727.

Santos, Bruno, António Antunes, and Eric Miller (2009). "Multiobjective approach to long-term interurban multilevel road network planning." In: *Journal of transportation engineering* 135.9, pp. 640–649.

Santos, Bruno, António Antunes, and Eric J Miller (2008). "Integrating equity objectives in a road network design model." In: *Transportation Research Record: Journal of the Transportation Research Board* 2089.1, pp. 35–42.

Santos, Bruno Filipe Lopes (2009). "Road Network Planning With Efficiency, Equity, and Robustness Objectives." In:

"Senegal Road Network Information" (2013). In: URL: http://dlca.logcluster.org/display/public/DLCA/2.3+Senegal+Road+Assessment.

Snelder, Maaike et al. (2007). "Optimal Redesign of Dutch Road Network." In: *Transportation Research Record: Journal of the Transportation Research Board* 2029.1, pp. 72–79.

White, Joanna and Ivan Wells (2002). "Extracting origin destination information from mobile phone data." In: *Road Transport Information and Control, 2002. Eleventh International Conference on (Conf. Publ. No. 486)*. IET, pp. 30–34.

Wikipedia (2013). *Trans-Gambia Highway — Wikipedia, The Free Encyclopedia*. [Online; accessed 22-December-2014]. URL: http://en.wikipedia.org/w/index.php?title=Trans-Gambia_Highway&oldid=562547018.

World Bank (2004). "Performance and impact indicators for transport in Senegal." In: URL: http://www.worldbank.org/transport/transportresults/regions/africa/senegal-road-redi_final.pdf.

Yang, Hai, H. Bell, and Michael G (1998). "Models and algorithms for road network design: a review and some new developments." In: *Transport Reviews* 18.3, pp. 257–278.

T04

Building workers' travel demand models based on mobile phone data

Feng Liu[a], Davy Janssens[b], JianXun Cui[c], Geert Wets[b]

[a,b] Transportation Research Institute (IMOB), Hasselt University, Wetenschapspark 5, bus 6, B-3590, Diepenbeek, Belgium

[c] Department of transport engineering, Harbin Institute of Technology (HIT), 1500, Harbin, China

[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.

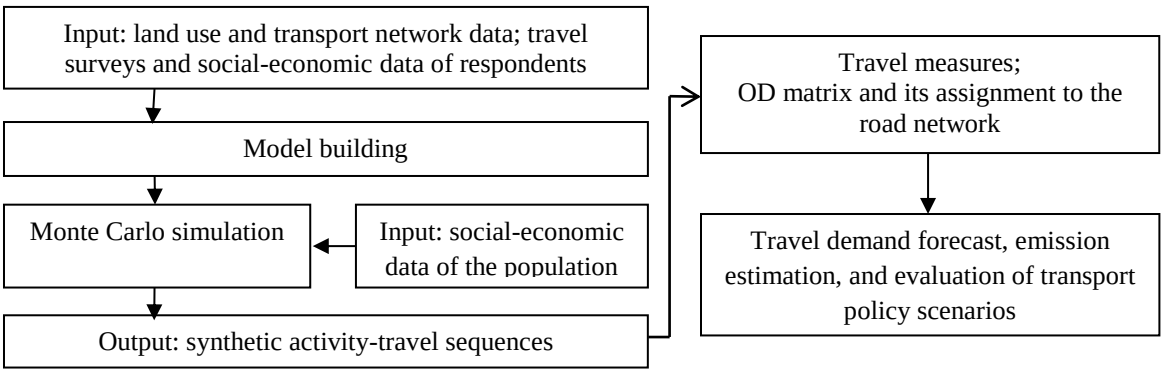


Fig. 1. The entire process of an activity-based transportation model

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 1. The typical call data of an individual

Time	11:57:00	13:40:00	16:59:00	17:43:00	21:28:00
Cell_id	751	749	177	751	751

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.

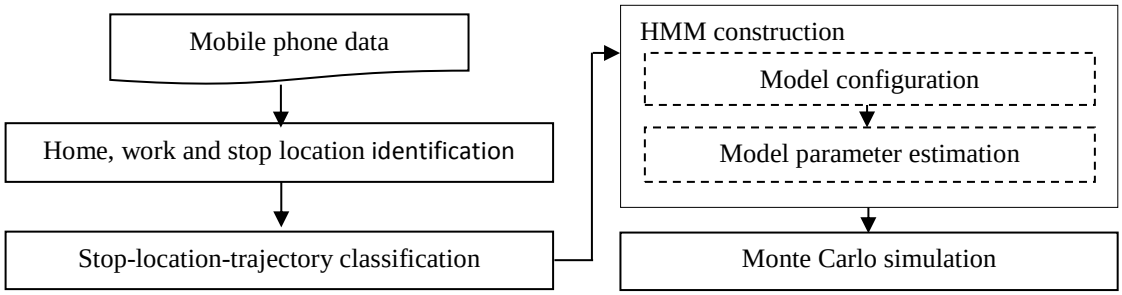


Fig. 2. The overall structure of the methodology

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 -> l_2 -> ... -> 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 $T(l_i,1)$, $T(l_i,2)$, ..., $T(l_i,k_i)$, 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 $l_1(T(l_1,1),T(l_1,2),...,T(l_1,k_1))$ -> ... -> $l_n(T(l_n,1),T(l_n,2),...,T(l_n,k_n))$, simplified as $l_1(T(1),T(2),...T(k_1))$ -> ... -> $l_n(T(1),T(2),...,T(k_n))$. 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) ->* l_2*(17:51) ->* l_3*(17:56,19:41) ->* l_4*(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) ->* l_2*(22:00) ->* l_3*(22:02) ->* l_4*(22:05) ->* 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 in 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 $T_{call\text{-}location\text{-}duration}$, 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 *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. $T(l_{i+1},1) - T(l_{i-1},k_{i-1})$, is computed. If this time period is longer than a threshold value, defined as $T_{maximum\text{-}time\text{-}boundary}$, l_i is perceived as a stop visit. (*iii*) When l_i is in the first or last position of a trajectory and the call-location-duration is shorter than $T_{call\text{-}location\text{-}duration}$, 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 $T_{call\text{-}location\text{-}duration}$ and $T_{maximum\text{-}time\text{-}boundary}$ respectively, as set up in our experiment described in Section 4, the obtained stop-location-trajectories for *user265* and *user72* are l_1 -> l_3 -> l_4 and l_1 -> 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. HWHWH), or 3 (or more)-home-based-work-tours (e.g. HWHWHWH), 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. HOHOH), 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

distance-based-tour-class-distribution, which 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 m_1, m_2 and m_3 respectively, and the other stop locations corresponding to each part of the sequences, defined as $m_{1,1}$, $m_{1,2}$, $m_{2,1}$, $m_{2,2}$ and $m_{3,1}$, 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 $p_{emit}(x|state)$.

At each of the states, maximum 3 possible transition probabilities πs 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 $m_{3,1}$, i.e. $\pi(m_{3,1}|m_3)$, when a trip is made in the evening after coming back from work. (iii) Transitions from state $m_{k,1}$*(k=1, 2, 3)* to state m_k, i.e. $\pi(m_k|m_{k,1})$, 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. $\pi(m_{k,1}|m_{k,1})$, 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 *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 $\pi(End \mid m_3)$.

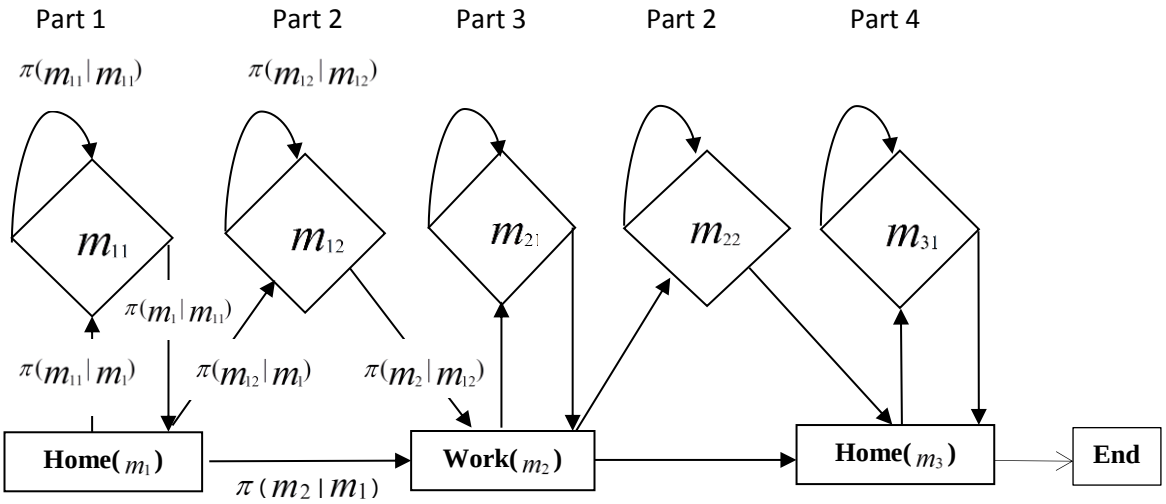


Fig. 3. The HMM for a home-based-work-tour

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 πs 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. $p_{emit}(x|state)$. 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.

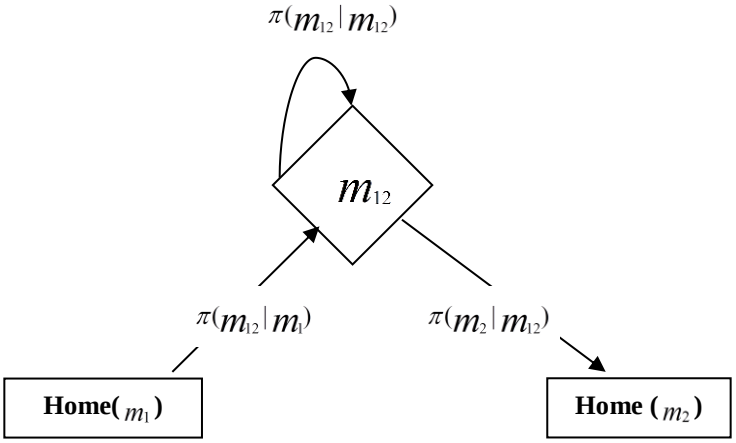


Fig. 4. The HMM for a home-based-non-work-tour

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 πs and $p_{emit}(x|state)$ 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 $E(x|state)$ as the frequencies of letter x at state $state$, respectively. The estimators for the parameters are given by the following formula.

$$\pi(r \mid \mathrm{q}) = \frac{A(r \mid \mathrm{q})}{\sum_r A(r \mid \mathrm{q})}, \; p_{emit}(x \mid \mathrm{state}) = \frac{E(x \mid \mathrm{state})}{\sum_{x'} E(x' \mid \mathrm{state})}$$

Where, $x, x' \in \{set\ of\ all\ letter\ types\ at\ the\ state\}$.

In the parameter estimation process, a *pseudocount* is set, which is a small value added to $A(r|q)$ or $E(x|state)$ 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 $p_{emit}(x|\ m_{11})$ 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. $s=Hxx$), based on $\pi(m_1|m_{11})$ and $\pi(m_{11}|m_{11})$ 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 *End* state of the model is realized based on $\pi(End | m_3)$.

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.

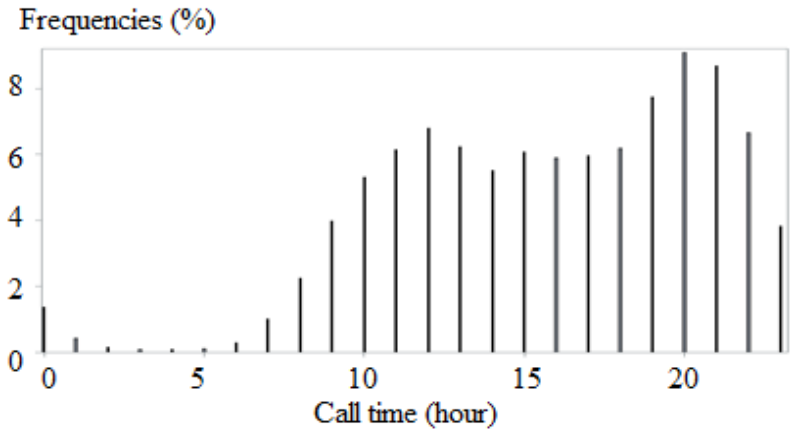


Fig. 5. The distribution of the time of calls

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. $T_{call\text{-}location\text{-}duration}$ and $T_{maximum\text{-}time\text{-}boundary}$

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 $T_{call\text{-}location\text{-}duration}$ defines the minimum time interval at a location, above which the location is considered as a possible stop. The other parameter $T_{maximum\text{-}time\text{-}boundary}$ 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, $T_{call\text{-}location\text{-}duration}$ and $T_{maximum\text{-}time\text{-}boundary}$ 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.

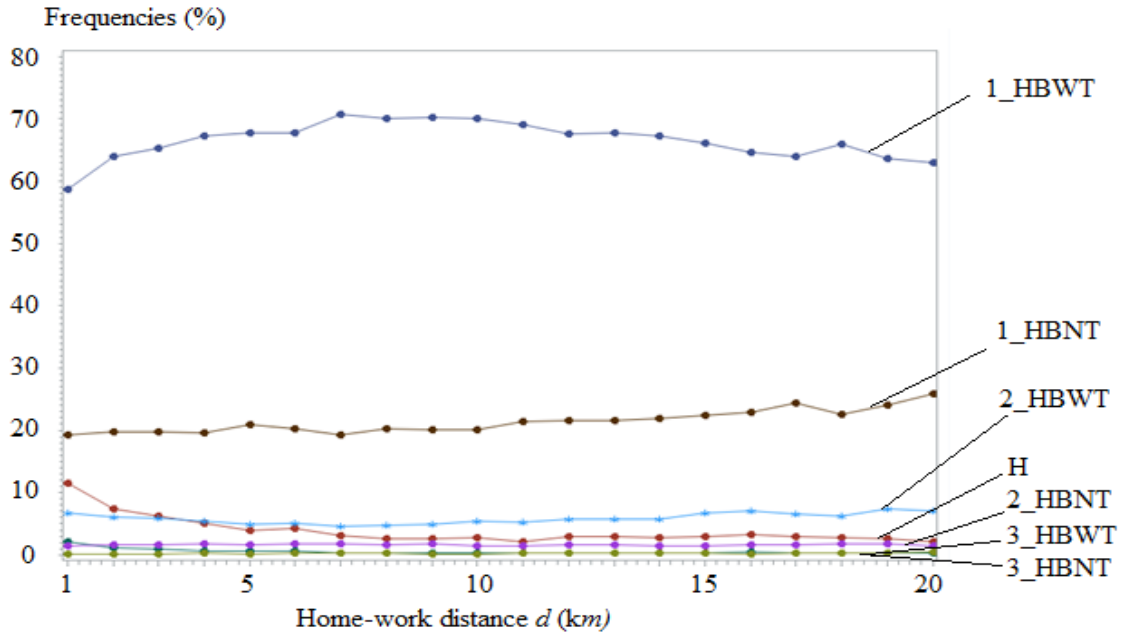


Fig. 6. The distribution of sequence frequencies in each class over home-work distances

Based on the observation from Fig. 6, we thus divide *d* into 4 intervals including *≤2km, 2-6km, 6-11km,* and *>11km*. 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. 11km, 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>11km* 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 (%)

Distance(d)	1_HBWT	2_HBWT	3_HBWT	1_HBNT	2_HBNT	3_HBNT	H	Total
≤2	61.00	9.76	1.70	19.42	1.57	0.15	6.41	100
2-6	66.69	5.14	0.77	20.05	1.68	0.21	5.47	100
6-11	70.05	2.69	0.31	20.15	1.60	0.25	4.95	100
>11	58.80	1.61	0.20	28.62	2.46	0.45	7.87	100
Average	63.05	5.29	0.84	22.31	1.86	0.26	6.39	100

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 *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 $\pi(m_{k,2}|m_k)$, $\pi(m_{k,2}|m_{k,2})$ and $\pi(m_{k+1}|m_{k,2})$ are not expected, they are thus represented with *'Null'*.

Table 3. Transition probabilities of the HMM derived from the 1_HBWT cluster

Locations	$\pi(m_{k,1} \mid m_k)$	$\pi(m_{k,2} \mid m_k)$	$\pi(m_{k+1} \mid m_k)$	$\pi(m_{k,1} \mid m_{k,1})$	$\pi(m_k \mid m_{k,1})$	$\pi(m_{k,2} \mid m_{k,2})$	$\pi(m_{k+1} \mid m_{k,2})$
Start home (m_1)	0.02	0.29	0.72	0.02	0.02	0.38	0.62
Work (m_2)	0.18	0.31	0.51	0.24	0.76	0.41	0.59
End home (m_3)	0.05	Null	0.95	0.24	0.76	Null	Null

Regarding the emission probabilities $p_{emit}(x|state)$, 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 $p_{emit}(x|state)=1$ 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 (%)

Types	HWH	HWOH	HOWH	HWOWH	HWOOH	HOWOH	HOOWH
Validation	41.38	11.38	7.97	4.60	3.53	3.26	2.21
Simulated	35.09	12.29	8.39	4.79	4.92	2.99	3.13
Types	HWOWOH	HWOOOH	HWHOH	HOWOOH	HOWOWH	HOOWOH	
Validation	1.83	1.34	1.33	1.21	1.12	1.01	
Simulated	1.75	2.01	1.42	1.24	1.12	1.18	

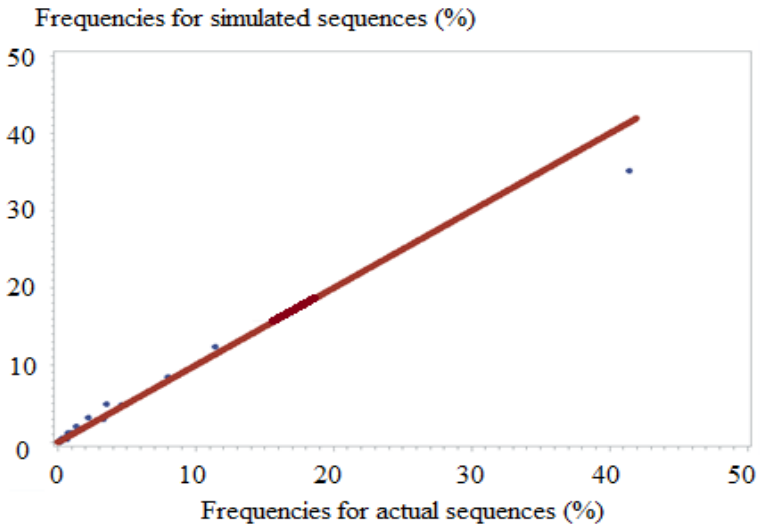


Fig. 7. The correlation of sequence frequencies for each type between actual phone location sequences and simulated ones

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 $T_{call\text{-}location\text{-}duration}$ and $T_{maximum\text{-}time\text{-}boundary}$ are used. $T_{call\text{-}location\text{-}duration}$ defines the maximum time duration needed to traverse a single cell area; while $T_{maximum\text{-}time\text{-}boundary}$ 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 $p_{emit}(x|state)$ 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. $p_{emit}(x|state)$. 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

Angelakis, V., Gundlegård, D., Rajna, B., Rydergren, C., Vrotsou, K., Carlsson, R., Forgeat, J., Hu, T. H., Liu, E. L., Moritz, S., Zhao, S., & Zheng, Y. T. (2013). Mobility Modeling for Transport Efficiency - Analysis of Travel Characteristics Based on Mobile Phone Data. Third International Conference on the Analysis of Mobile Phone Datasets. NetMob, Special session on the D4D challenge, MIT, May 1-3, 2013.

Arentze, T. A., & Timmermans, H. J. P. (2004). A learning-based transportation oriented simulation system. Transportation Research Part B: Methodological, 38(7), 613-633.

Asakura, Y., & Hato, E. (2006). Tracking individual travel behavior using mobile phones: recent technological development. Paper presented at 11th International Conference on Travel Behaviour Research, Kyoto.

Bayir, M. A., Demirbas, M., & Eagle, N. (2009). Discovering spatiotemporal mobility profiles of cellphone users. World of Wireless, Mobile and Multimedia Networks & Workshops, WOWMOM, IEEE, 1-9.

Becker, R., Cáceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A tale of one city: Using cellular network data for urban planning. IEEE Pervasive Computing, 10(4), 18–26.
Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T., & Timmermans, H. J. P. (2010). Implementation Framework and Development Trajectory of Feathers Activity-Based Simulation Platform. Transportation Research Board: Journal of the Transportation Research Board, 2175, 111-119.
Berlingerio, M., Calabrese, F., Lorenzo, G. D., Nair, R., Pinelli, F., & Sbodio, M. L. (2013). AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. Third International Conference on the Analysis of Mobile Phone Datasets. NetMob, Special session on the D4D challenge, MIT, May 1-3, 2013.
Bhat, C. R., & Koppelman, F. S. (1999). A Retrospective and Prospective Survey of Time-Use Research. Transportation, 26(2), 119-139.
Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing, 10(4), 36-44.
Cools, M., Moons, E., Bellemans, T., Janssens, D., & Wets, G. (2009). Surveying activity-travel behavior in Flanders: Assessing the impact of the survey design. Proceedings of the BIVEC-GIBET Transport Research Day, Part II, VUBPress, Brussels, 370, 727-741.
Davidson, W., Donnelly, R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J., Castiglione, J., & Picado, R. (2007). Synthesis of first practices and operational research approaches in activity-based travel demand modeling. Transportation Research Part A: Policy and Practice, 41(5), 464–488.
Delafontaine, M., Versichele, M., Neutens, T., & Van de Weghe, N. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. Applied Geography, 34, 659–668.
de Montjoye, Y., Smoreda, Z., Trinquart, R., Ziemlicki, C., & Blondel, V. D. (2014). D4D-Senegal: The Second Mobile Phone Data for Development Challenge.
Do, T. M. T., & Gatica-Pereza, D. (2013). Where and what: Using smartphones to predict next locations and applications in daily life. Pervasive and Mobile Computing. http://dx.doi.org/10.1016/j.pmcj.2013.03.006
Durbin, R., Eddy, S.R., Krogh, A., & Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.
Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., & Punta, M. (2014). Pfam: the protein families database. Nucleic Acids Res. 42(D1): D222–D230.
García-Díez, S., Fouss, F., Shimbo, M., & Saerens, M. (2011). A sum-over-paths extension of edit distances accounting for all sequence alignments. Pattern Recognition, 44(6), 1172–1182.
González, M. C., Hidalgo, C. A., & Barabási, A. L. (2008). Understanding individual human mobility patterns. Nature, 453, 779-782.
Gühnemann, A., Schäfer, R. P., &Thiessenhusen, K. U. (2004). "Monitoring Traffic and Emissions by Floating Car Data." Institute of Transport Studies, Working Paper, Issue ITS-WP-04-07.
Hartgen, D. T. (2013). Hubris or humility? Accuracy issues for the next 50 years of travel demand modeling. Transportation, 40(6), 1133-1157.
Janssens, D., Giannotti, F., Nanni, M., Pedreschi, D., & Rinzivillo, S., (2012). Data Science for Simulating the Era of Electric Vehicles. KI - Künstliche Intelligenz, 26(3), 275-278.
Joh, C. H., Ettema, D., & Timmermans, H. J. P. (2008). Improved Motif Identification of Activity Sequences: Application to Interactive Computer Experiment Data. Transportation research record: Journal of the Transportation Research Board, 2054, 93-101.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K., & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235(5):1501-1531.
Liu, C. C., & Chen, J. C. H. (2013). Using Q methodology to explore user's value types on mobile phone service websites. Expert Systems with Applications, 40(13), 5276–5283.
Liu, F., Janssens, D, Cui, J.X. , Wang, Y.P., Wets, G, & Cools, M. (2014). Building a validation measure for activity-based transportation models based on mobile phone da. Expert Systems with Applications. 41(14), 6174–6189.
Liu, F., Janssens, D., Wets, G., & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. Expert Systems with Applications, 40(8), 3299–3311.
Monares, A., Ochoa, S. F., Pino, J. A., Herskovic, V., Rodriguez-Covili, J., & Neyem, A. (2013). Mobile computing in urban emergency situations: Improving the support to firefighters in the field. Expert Systems with Applications, 38(2), 1255–1267.
Roorda, M. J., Miller, E. J. & Habib, K. M. N. (2008). Validation of TASHA: A 24-H Activity Scheduling Microsimulation Model. Transportation Research Part A: Policy and Practice, 42(2), 360-375.
Saneinejad, S., & Roorda, M. J. (2009). Application of sequence alignment methods in clustering and analysis of routine weekly activity schedules. Journal Transportation Letters: The International Journal of Transportation Research, 1(3), 197-211.
Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling Daily Human Mobility Motifs. Journal of The Royal Society Interface, 10(84),
Shan, J., Viña-Arias, L., Ferreira, J., Zegras, C., & González, M. C. (2011). Calling for Validation, Demonstrating the use of mobile phone data to validate integrated land use transportation models. In Proceedings 7VCT 2011.
Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in human mobility. Science, 327(5968), 1018-1021.
Spissu, E., Pinjari, A. R., Bhat, C. R., Pendyala, R. M., & Axhausen, K. W. (2009). An analysis of weekly out-of-home discretionary activity participation and time-use behavior. Transportation, 36(5), 483-510.
Steenbruggen, J., Borzacchiello, M. T., Nijkamp, P., & Scholten, H. (2013). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. GeoJournal, 78(2), 223-243.
Wegener, M. (2013). The Future of Mobility in Cities: Challenges for Urban Modelling. Transport Policy, 29, 275–282.
Wilson, C. (2008). Activity patterns in space and time: calculating representative Hagerstrand trajectories. Transportation, 35(4), 485-499.

T05

Urban Road Construction and Human mobility: Evidence from Dakar, Senegal. *

Thiemo Fetzer

December 29, 2014

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

*London School of Economics and STICERD, Houghton Street, WCA2 2AE London.

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 users, 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.

3 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 Δd_{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 Δd_{AB}. For every cell A, we compute the weighted average $\Delta d_A = \sum_{b \in B} \omega_{Ab} \Delta d_{Ab}$. ω_{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.

$$X_{AB} = G(M_A^{\beta_1} M_B^{\beta_2} / D_{AB}^{\beta_3}) \tag{1}$$

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 α_{AB} as well as general time trends δ_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 γ_t. These measure the effect of the reduction in travel time Δd_{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 γ_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 γ_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:

$$X_{ABt} = \alpha_{AB} + \delta_t + Post_t \times \Delta d_{AB} + \epsilon_{ABt} \tag{3}$$

Where $Post_t = 1$ 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:

$$X_{ABt} = \alpha_{AB} + \delta_t + \sum_{a \in A} Post_t \times \gamma_a + \epsilon_{ABt} \tag{4}$$

This provides a set of coefficients γ_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 γ_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 γ_t with an indicator that is equal to 1 in for the period following August 1st, 2013. We replace the measure ΔD_{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%.

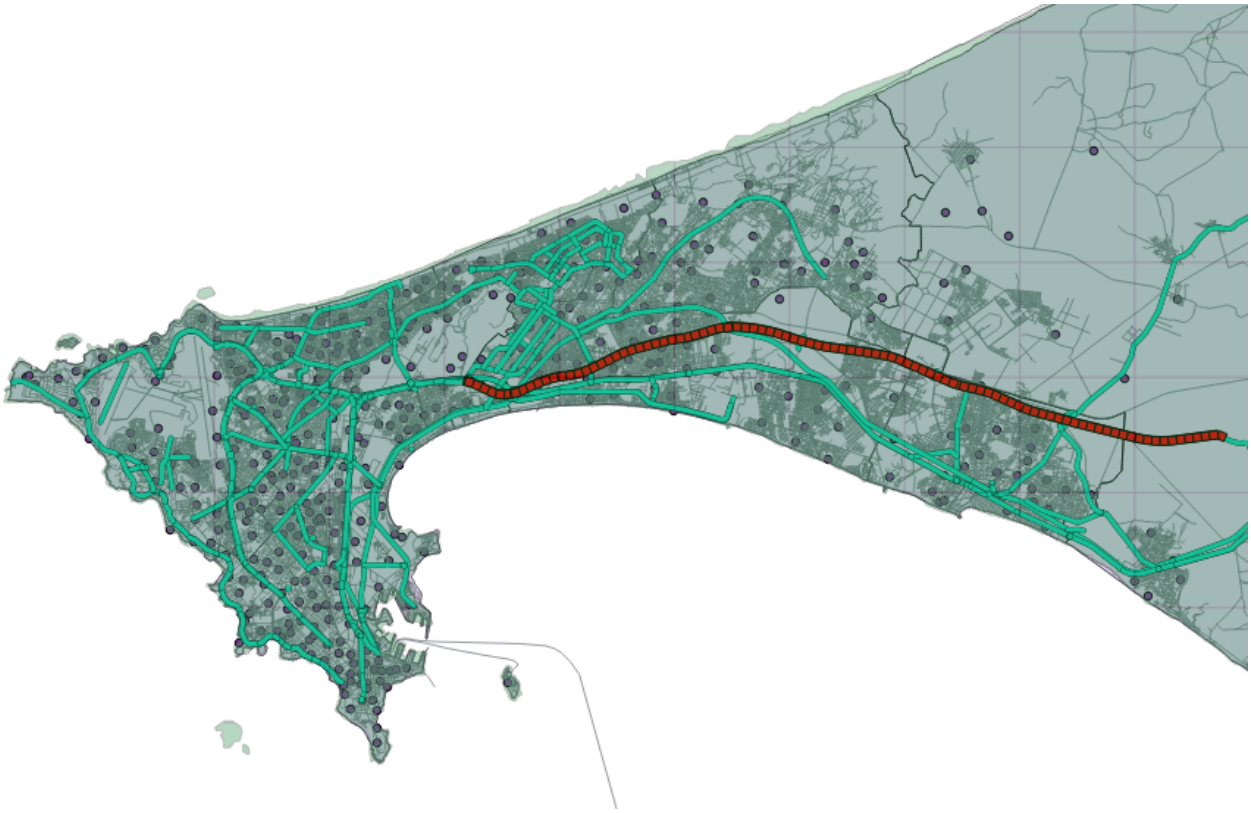


Figure 1: Toll Road Opened in August 2013 marked in red; other main roads are marked in. 0.03 x 0.03 degree grid cells are outlined. Data from Open Street Map.

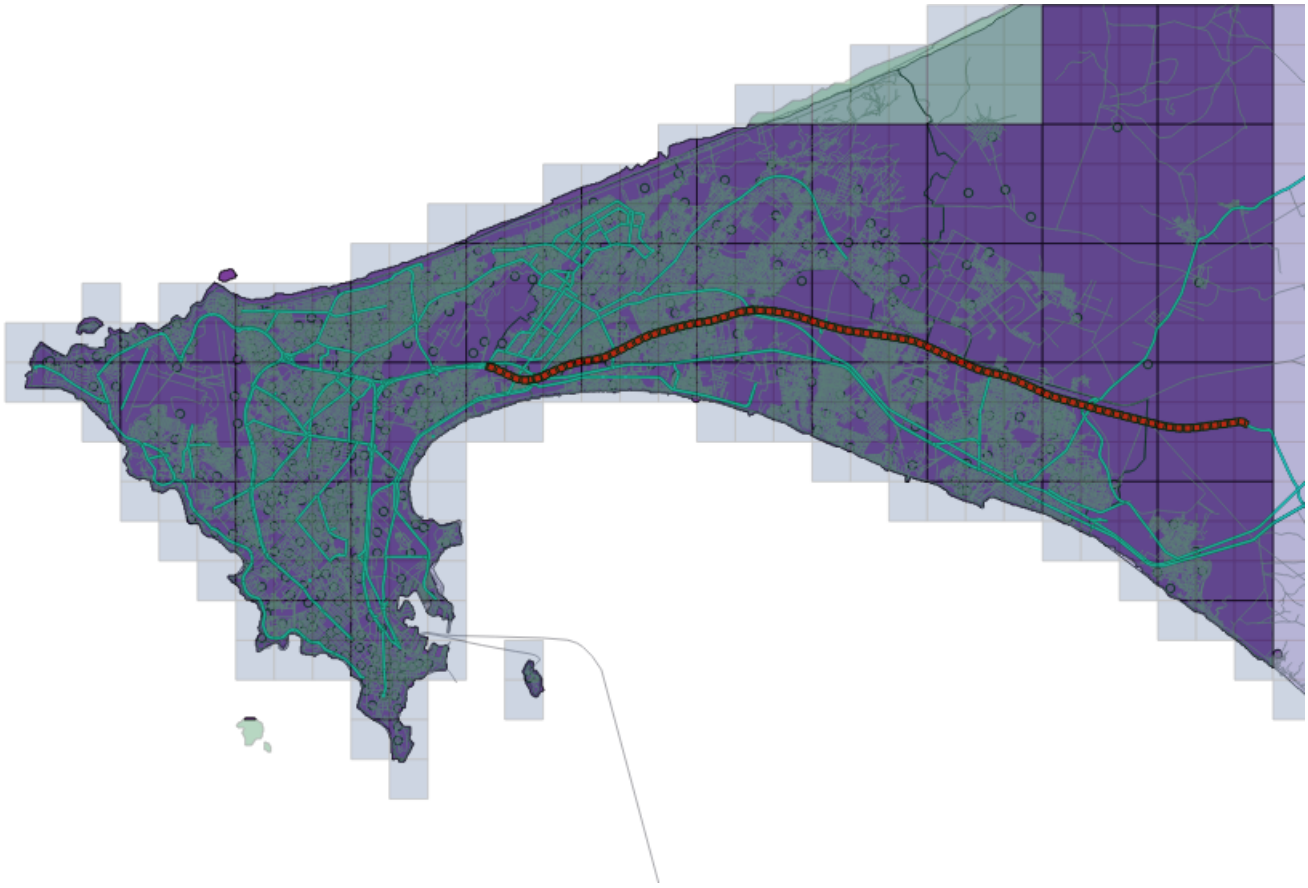


Figure 2: Study Area - Grid Cells included in the Analysis of Cross Cell Intra Day Mobility Over Time.

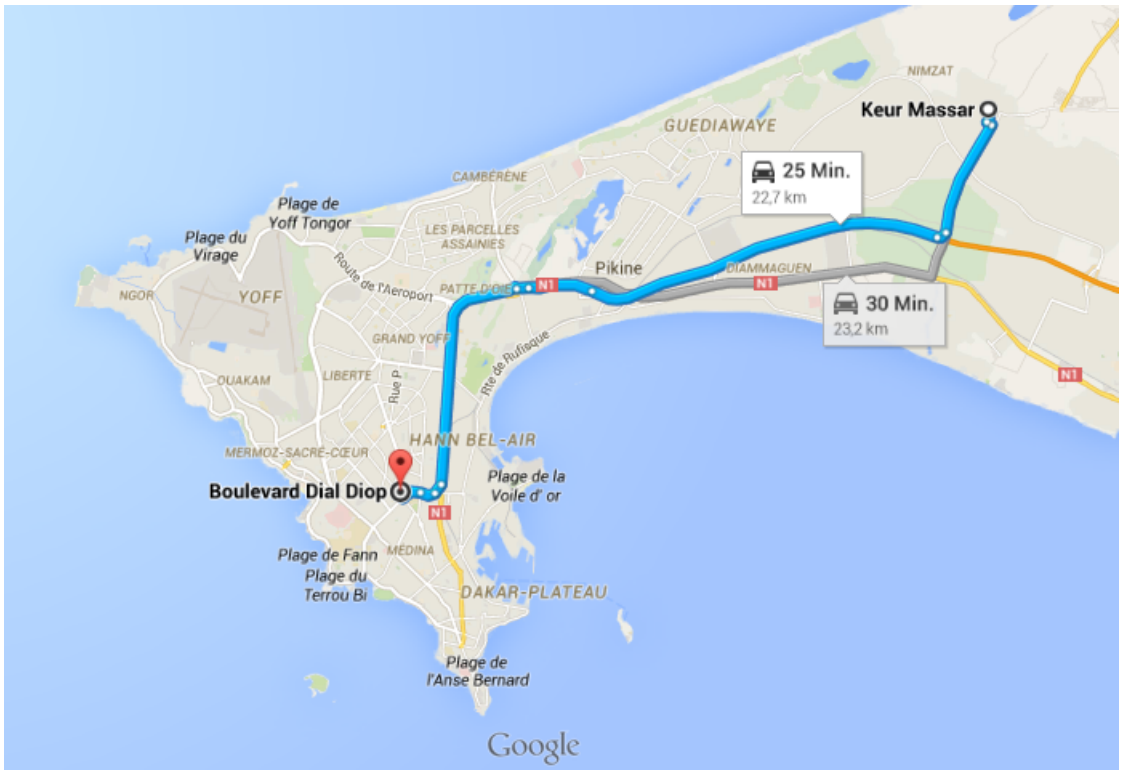


Figure 3: Example Computation of Travel Time Reduction due to Toll Road Access. The route using the toll road is highlighted in blue, while the alternative route is highlighted in light grey. Data from Google Maps.

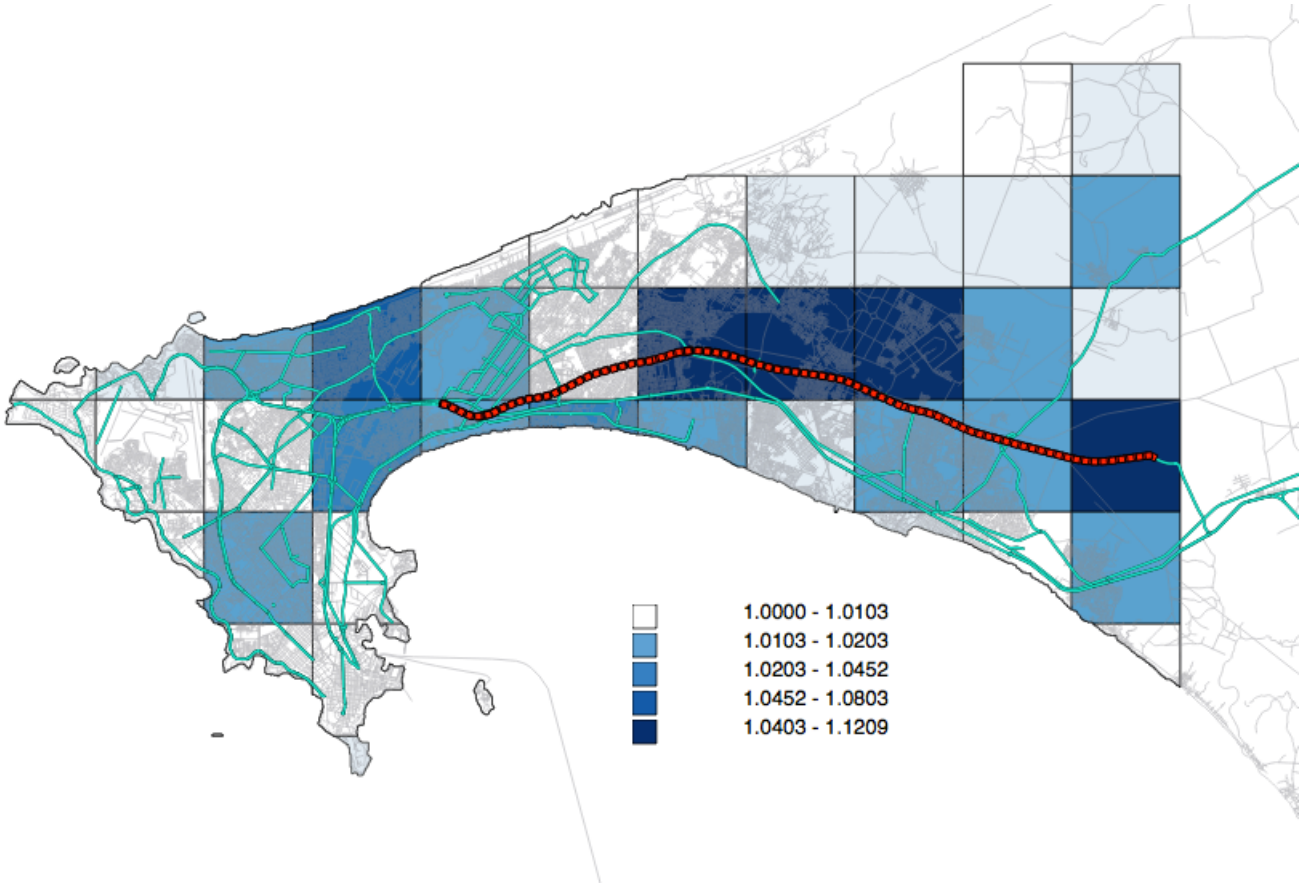


Figure 4: Spatial Distribution of Travel Time Reductions. Travel time reductions are computed as weighted averages, weighted by the total number of people who commute across grid cells before the opening of the Toll Road.

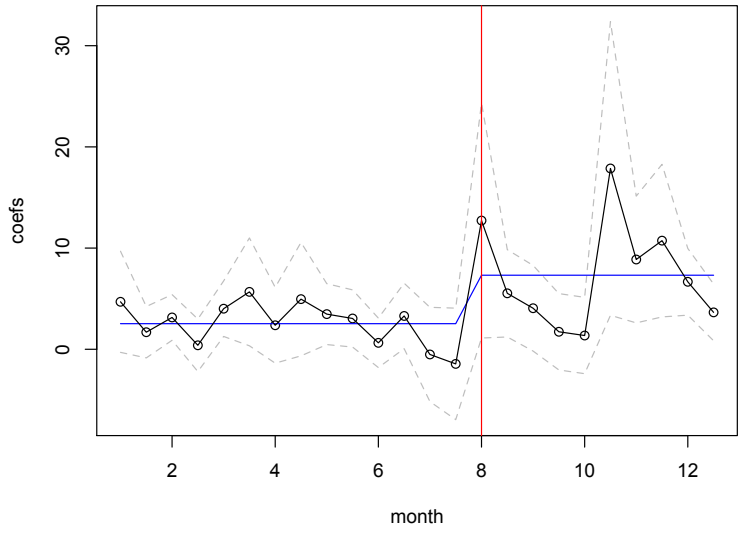


Figure 5: Estimated Effect of Travel Time Reduction on Inter Cell Mobility Over Time. Dashed lines are 90% confidence intervals; the blue line corresponds to the estimated structural break in the series of estimated coefficients. The vertical line corresponds to the official opening of the toll road.

Table 1: Increase in Commuting Activity following Opening of Toll Road

	Including zeroes		Excluding zeroes	
	(1)	(2)	(3)	(4)
Post x Δd_{AB}	4.958**	8.466**	8.736**	14.359**
	(2.004)	(3.489)	(3.587)	(5.936)
Mean of DV	29.1	29.1	66.4	66.4
Time FE	.	.	Yes	.
Pair FE	Yes	Yes	Yes	Yes
Time x Origin/ Destination FE	No	Yes	.	Yes
Clusters	44	44	43	43
Observations	47300	47300	20742	20742
R-squared	.993	.994	.993	.994

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 *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

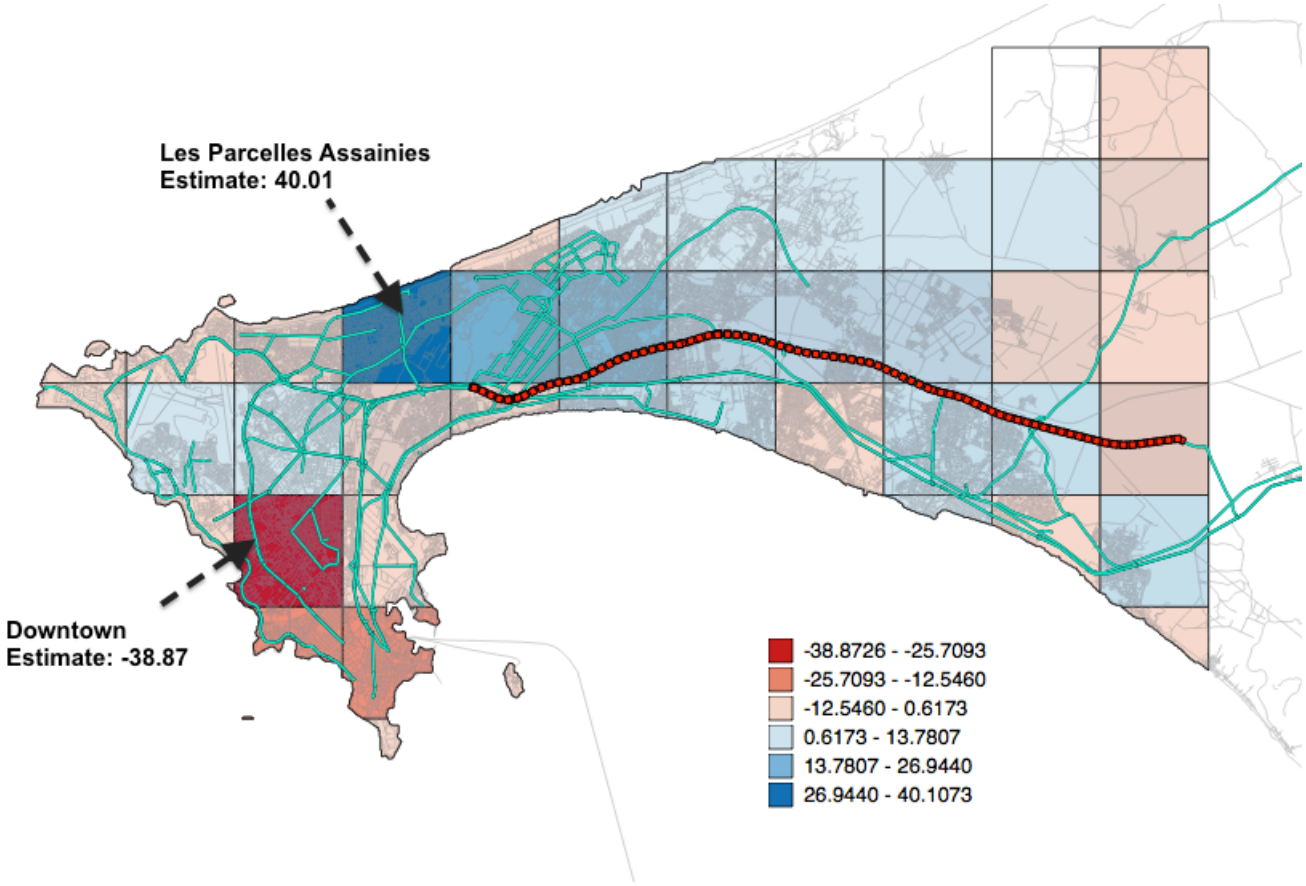


Figure 6: Estimated Effect of Toll Road Introduction on Inter Cell Mobility By Non-Day Time Cells. The incidence of increased inter cell mobility is observed in locations along the route.

References

Atkin, D. and D. Donaldson (2014). Who ' s Getting Globalized ? The Size and Implications of Intranational Trade Costs . (February).

Clark, X., D. Dollar, and A. Micco (2004, December). Port efficiency, maritime transport costs, and bilateral trade. *Journal of Development Economics 75*(2), 417–450.

Donaldson, D. (2013). Railroads of the Raj: Estimating the Impact of Transportation Infrastructure. *forthcoming, American Economic Review*.

Ji, Y. (2011). Understanding Human Mobility Patterns Through Mobile Phone Records : A cross-cultural Study.

Simini, F., M. C. González, A. Maritan, and A.-L. Barabási (2012, April). A universal model for mobility and migration patterns. *Nature 484*(7392), 96–100.

Tinbergen, J. (1962). Shaping the World Economy; Suggestions for an International Economic Policy.

Yuan, Y. and M. Raubal (2013). Extracting dynamic urban mobility patterns from mobile phone data.

T06

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 the 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 discard 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 underlooked: 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

* Corresponding author: David Gundlegård, david.gundlegard@liu.se, Linköping University, 601 74 Norrköping, Sweden.

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.

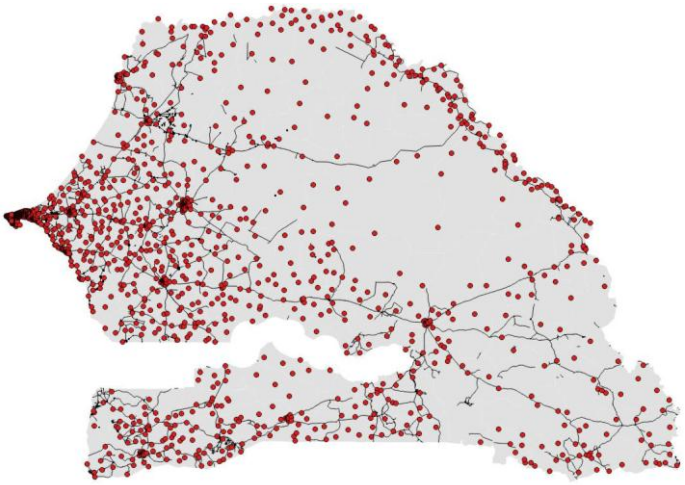


Figure 1: Antenna distribution and road network from Open Street Map for Senegal.

B. *Mobile phone data for mobility analysis*

The mobile phone data for Senegal originate from the mobile operator Orange (de Montjoye et al. 2014) and consists of Call Detail Records (CDR) of phone calls and text exchanges (SMSes) between customers in Senegal. The data is collected between January 1, 2013 and December 31, 2013. The data used in this paper consists of 1666 antenna ID and locations and mobility data on a rolling 2-week basis for a year for about 300,000 randomly sampled users.

The mobility data consists of timestamps and antenna IDs. The positions of the calls are identified according to the connected antenna. The position of each antenna is given as the longitude and latitude, slightly blurred due to sensitive information. The time stamps are truncated to 10 minute intervals. The data accuracy is as described in de Montjoye et al. (2014) and the data looks sound according to our initial plots and analysis, except minor anomalies with “jumps” between the two relatively distant antennas, e.g. between antenna 1057 and 604.

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 Gundlegård 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 approached 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 differentially 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.

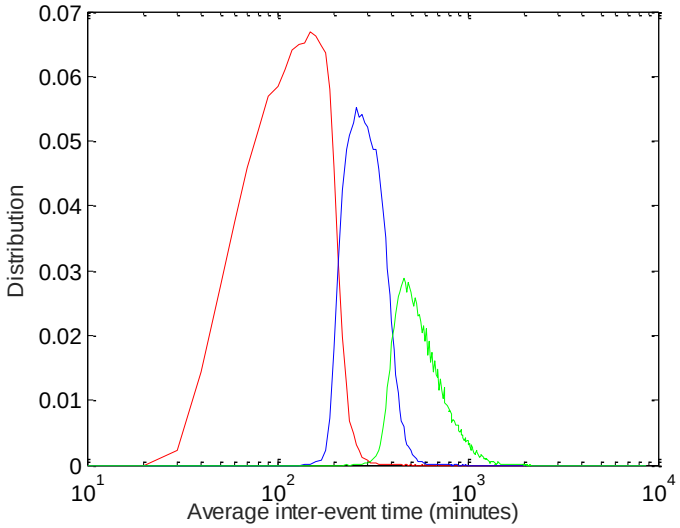


Figure 2: Distribution of average inter-event times for three quantiles.

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.

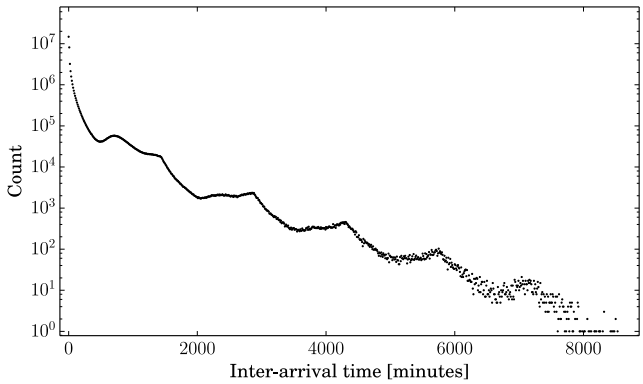


Figure 3: Counts of time duration between recorded sites.

A standard problem when analyzing cellular network data based on user activities, e.g. SMS and phone calls, is the time bias in the samples; typically users have a tendency to make fewer phone calls early in the morning. Figure 4 shows the total number of events over the day and it can be seen that there is much more phone activity late in the evening compared to early in the morning, whether this is a true reflection of the human activity in Senegal or not, is not clear to the authors at this time. If this is a bias in the data set, it is important to take into account when scaling up results from the data set.

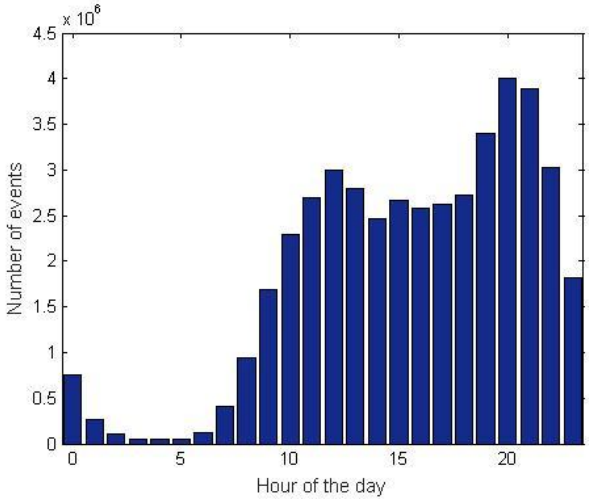


Figure 4: CDR data events per hour for the first two week period of data.

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 POI:s, 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.

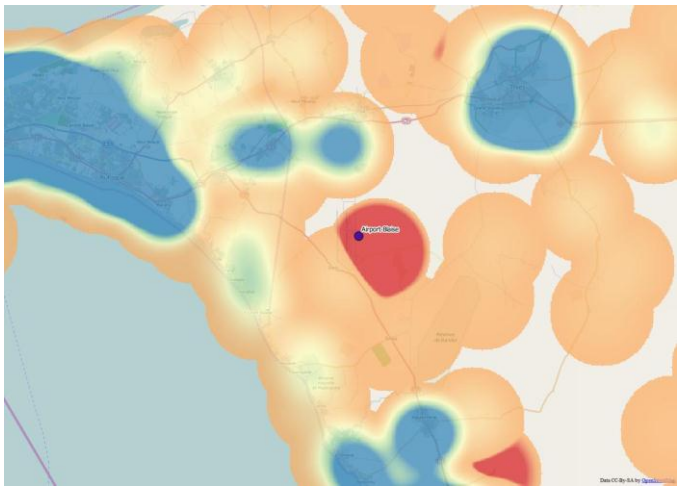


Figure 5: Heat map of difference of number of home locations and number of work locations. Blue indicates potential residential areas and red indicates areas with large daytime activity. The red area in the middle of the figure is an airport.

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_{max}. Furthermore, a threshold value, d_{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.

```
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
```

Algorithm 1a: Main function for the trip generation.

```
detect_trip_end ()
if(p_udk == workbase or p_udk == homebase)
   destination = p_udk
else
   if(p_ud(k+1) exists)
      if(p_udk == p_ud(k-1))
         destination = p_udk
      end
   else
      if(d(p_udk,homebase) < d_max)
         destination = homebase
      else
         destination = p_udk
      end
   end
end
```

Algorithm 1b: Function for detecting trip end.

```
detect_trip_start()
if (trip_set empty)
   if(p_udk != homebase and d(p_udk,homebase) < d_max
and d(p_udk,homebase) > d_min)
      trip_active = true
      origin = homebase
   end
   if(p_udk != homebase and d(p_udk,homebase) > d_max)
      trip_active = true
      origin = p_udk
   end
   if(p_udk == workbase and d(p_udk,homebase) > d_max)
      trip_active = true
      origin = homebase
      destination = workbase
   end
else
   if(p_udk != previous_trip_start(trip_set) and
d(previous_trip_start(trip_set), p_udk) > d_min)
      origin = previous_trip_start(trip_set)
   end
end
```

Algorithm 1c: Function for detecting trip start.

Figure 6 shows an example of generated trips for a specific user for a specific day. Blue circles are antennas in the trajectory for this user and day, the home location is marked by H, the work POI is marked by W and an additional location A. The location A is identified by two consecutive calls referring to the same antenna, here taken as an indication of an activity at this location. The trips generated in this case are 1) from H to W, 2) from W to H, 3) from H to A and 4) from A to H. Note that the fourth trip (A to H) is generated even if the trajectory does not end in H for the specific day. This corresponds to the generation of an activity profile, HWHAH, as discussed in Liu et. al. (2013).

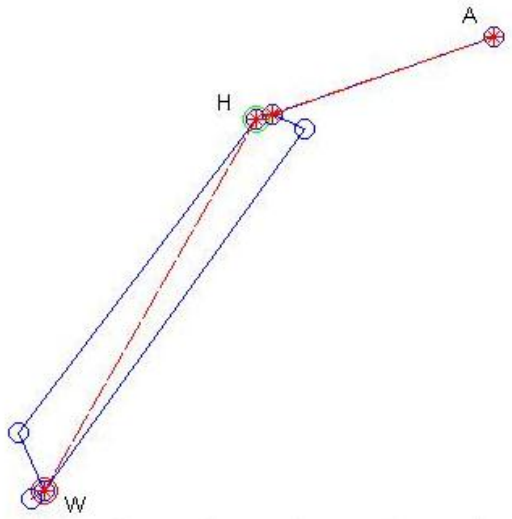


Figure 6: Example of trips generated for one user one day. Blue lines are trajectories and red lines are the generated origin-destination trips.

The number of trips generated using this trip definition is approximately 0.7 trips per day and user, with d_{min} set to 3 km and d_{max} 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).

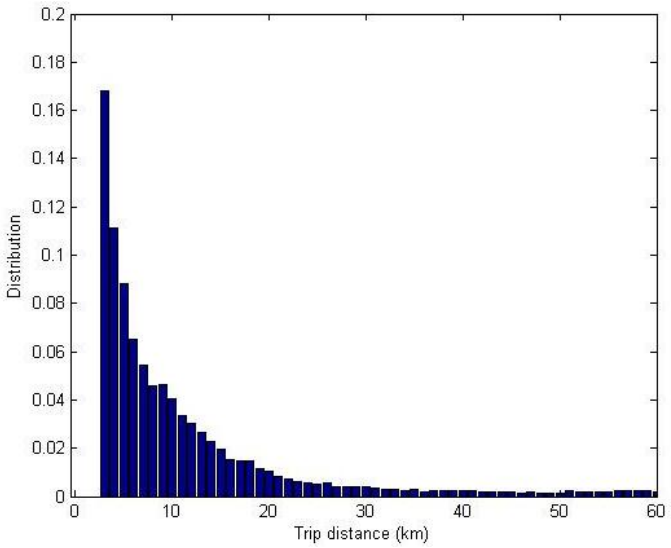


Figure 7: Trip distance distribution for the generated trips.

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.

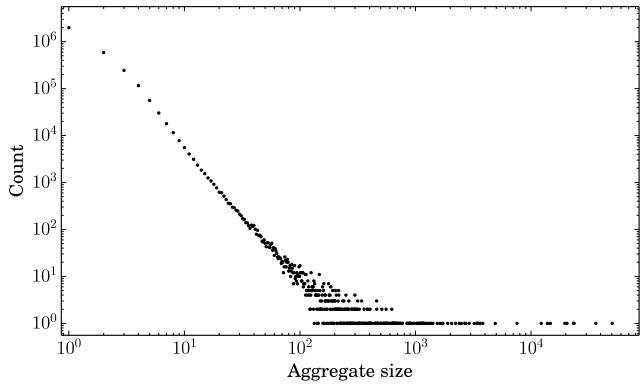


Figure 8: Counts of aggregate sizes as found by clustering algorithm.

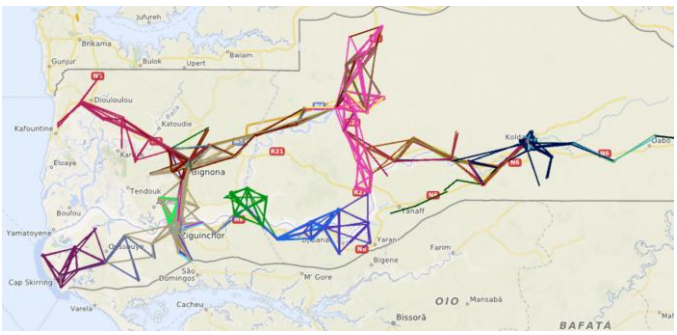


Figure 9: Example of sequence aggregates illustrated as site-to-site transitions color coded by aggregate.

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 γ 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.

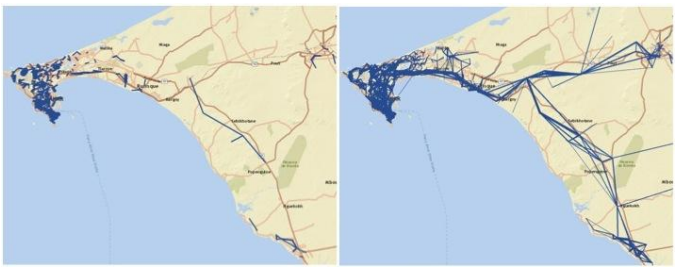


Figure 10: Sample result for the capital area for the algorithm Seqwog (left) and MG-FSM (right).

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.

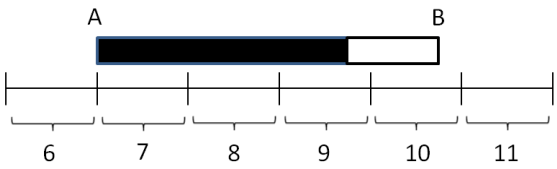


Figure 11: Assignment of trip weight to the time sliced origin destination matrix

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.

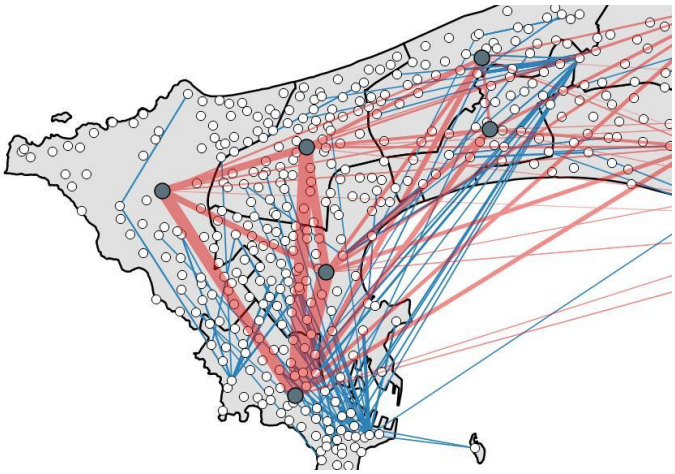


Figure 12: OD demand on the level of arrondissement for the first two week time period filtered on large OD demands for the city of Dakar in red and antenna level demand in blue.

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.

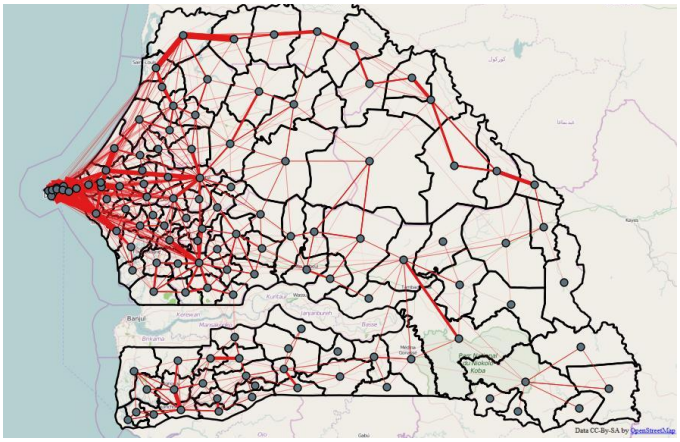


Figure 13: OD demand on the level of arrondissement for the first two week time period filtered on large OD demands.

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.

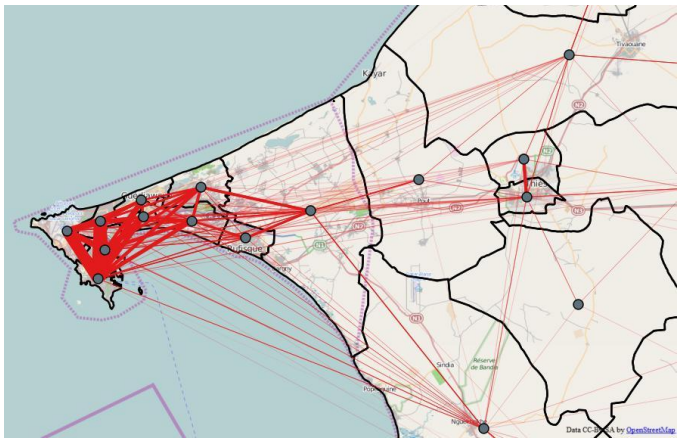


Figure 14: OD demand on the level of arrondissement for the first two week time period filtered on large OD demands for the Dakar area.

In order to get a potentially higher temporal resolution for trips, we have further analyzed trips that has 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 par. 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.

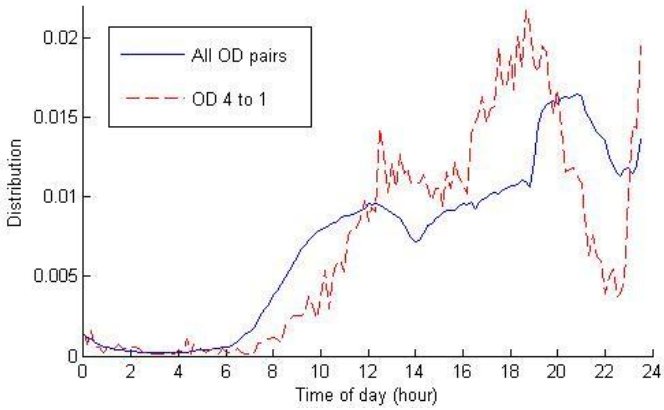


Figure 15: Temporal distribution of trips for OD pair from arrondissement 4 to 1, based on trips with an estimated average speed larger than 10 km/h. The red line spike at midnight is most likely due to a combination of antenna oscillations and trip definition.

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).

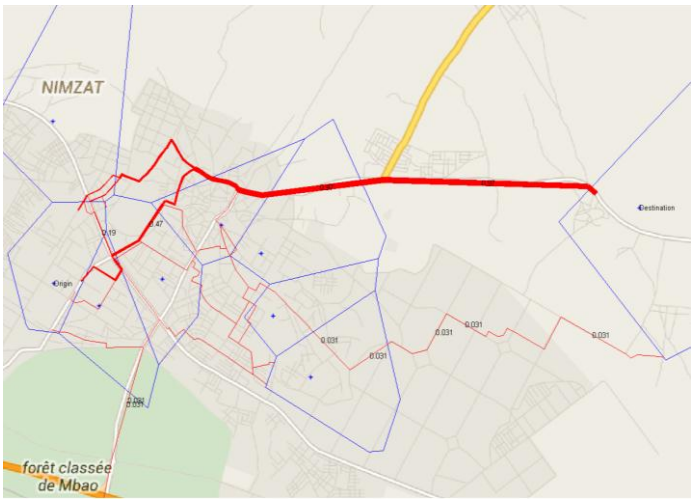


Figure 16: Route choice distribution for an example antenna pair, based on assigned demand according to probability of route choice, calculated using sequences with frequent sampling of antennas.

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.

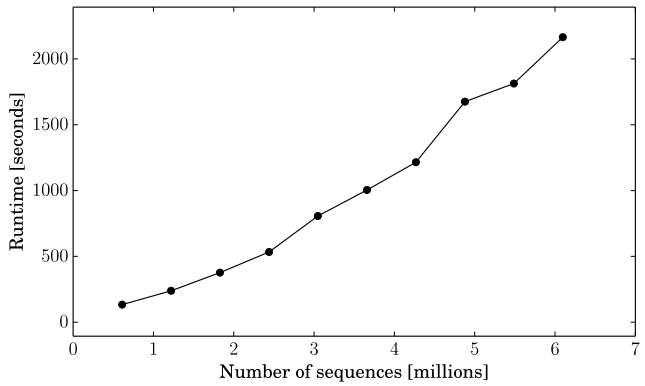


Figure 17: Runtime of clustering algorithm for different number of sequences on an Apple MacBook Pro with a 2.8 GHz Intel Core i7 processor and 16 GB of RAM.

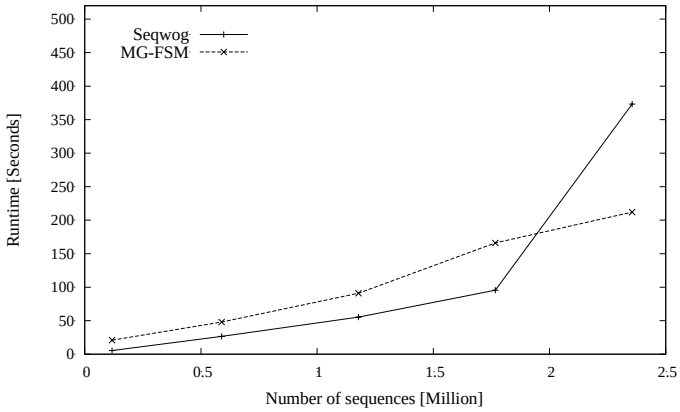


Figure 18: Runtime of frequent sequence mining algorithms Seqwog and MG-FSM for different number of sequences.

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 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 that ranges of Epsilon converge. In this figure only best and converged results are shown.

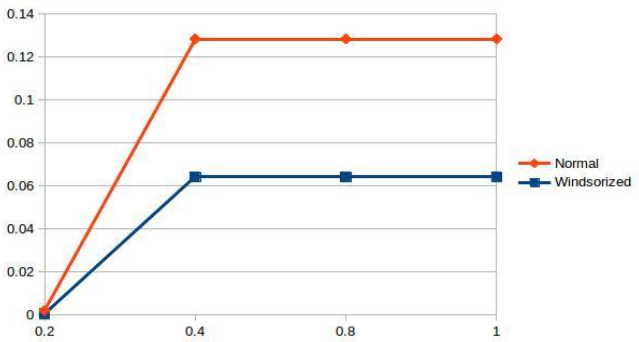


Figure 19: Convergence of privacy budget estimations for a site sequence cluster input, Epsilon values taken from range of [0,1].

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.

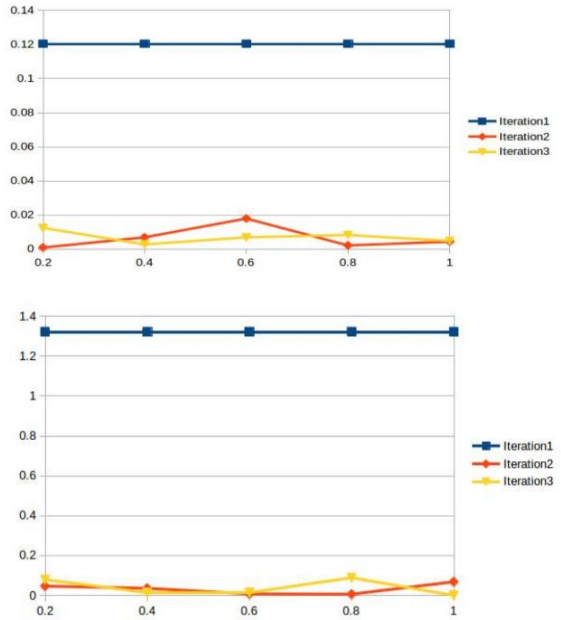


Figure 20: Convergence of privacy budgets for two frequent sequence mining route aggregation techniques. Each aggregate set were analyzed using an Epsilon from range of [0.2,1]. Estimator was run for three iterations.

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.

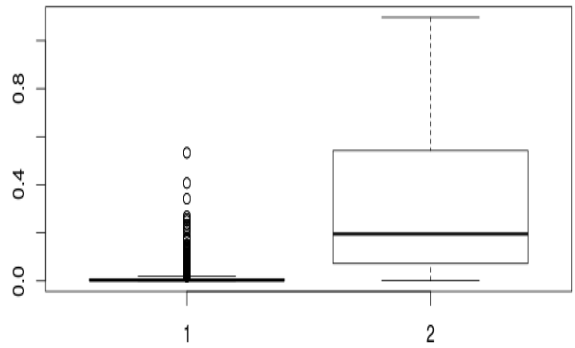


Figure 21: Comparative analysis of privacy budget ranges for longitudinal data units.

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 γ) 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.

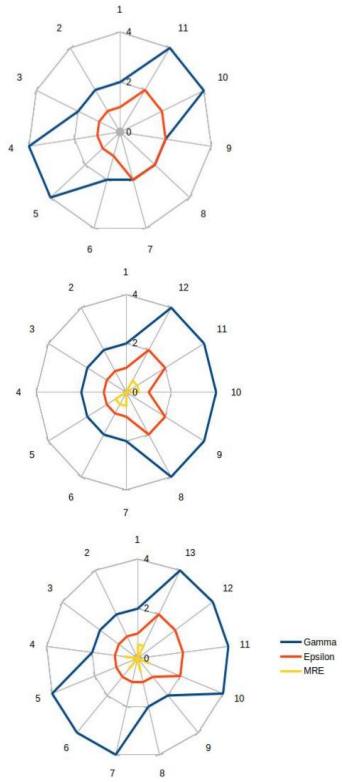


Figure 22: Output range estimations for statistical release of three experimental site sequence clusters. Gamma values are taken from the range of [0,4] and Epsilon values are taken from the range of [1,2].

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

G. Acs, C. Castelluccia, "A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14, 2014, pp. 1679–1688.

V. Angelakis, D. Gundlegård, B. Rajna, C. Rydergren, K. Vrotsou, R. Carlsson, J. Forgeat, T.H. Hu, E.L. Liu, S. Moritz, S. Zhao, Y. Zheng, "Mobility Modeling for transport efficiency: Analysis of travel characteristics based on mobile phone data," in V.D. Blondel, N. de Cordes, A. Decuyper, P. Deville, J. Raguenez, S. Zbigniew (eds), Mobile Phone Data for Development, Analysis of mobile phone datasets for the development of Ivory Coast, viewed 29 December 2014, http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4Dbook.pdf, 2013, pp. 412-422.

N. Andrienko, G. Andrienko, "Spatial Generalization and Aggregation of Massive Movement Data," IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 2, 2011, pp. 205–219.

N. Andrienko and G. Andrienko, "Visual analytics of movement: an overview of methods, tools, and procedures," Information Visualization, vol. 12, no. 1, 2012, pp. 3–24.

S. Bekhor, M. E. Ben-Akiva, M. S. Ramming, "Evaluation of choice set generation algorithms for route choice models," Annals of Operations Research, volume 144, 2006, pp. 235-247.

M. Berlingerio, F. Calebrese, G. Di Lorenzo, R. Nair, F. Pinelli, M-L Sbodio, "AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data," in V.D. Blondel, N. de Cordes, A. Decuyper, P. Deville, J. Raguenez and S. Zbigniew (eds), Mobile Phone Data for Development, Analysis of mobile phone datasets for the development of Ivory Coast, viewed 29 December 2014, http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4Dbook. pdf, pp. 397-411.

V.D. Blondel, N. de Cordes, A. Decuyper, P. Deville, J. Raguenez, S. Zbigniew, "Mobile Phone Data for Development, Analysis of mobile phone datasets for the development of Ivory Coast (D4D Book)", viewed 29 December 2014, 2013, http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf

C. Borgelt. "Keeping things simple: Finding frequent item sets by recursive elimination," International workshop on open source data mining: Frequent pattern mining implementations (OSDM), 2005, pp. 66–70.

A.Z. Broder, M. Charikar, A.M. Frieze, M. Mitzenmacher "Min-wise independent permutations," Journal of Computer and System Sciences 60, 1998, pp. 327–336.

N. Cáceres, L.M. Romero, F.G. Benitez and J.M. del Castillo, "Traffic Flow Estimation Models Using Cellular Phone Data," IEEE Transactions on Intelligent Transportation Systems, vol. 13, no. 3, 2012, pp. 1430-1441.

F. Calabrese, G. Di Lorenzo, L. Liu, C. Ratti, "Estimating origin-destination flows using mobile phone location data," IEEE Pervasive Computing, vol. 10, no. 4, 2011, pp. 36-44.

F. Calabrese, M. Diao, G. Di Lorenzo, J.Ferreira Jr, C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example", Transportation Research Part C, 2013, pp. 301-313.

B.C. Csáji, V.A. Traag, A. Browet, J.-C. Delvenne, E. Huens, P. Van Dooren, V.D. Blondel, Z. Smoreda, "Exploring the mobility of mobile phone users," Physica A: StatisticalMechanics and its Applications, vol. 392, no. 6, 2012, pp. 1459-1473.

M. Dash, H.L. Nguyen, C. Hong, G.E. Yapm, M. N. Nguyen, X. Li, S.P. Krishnaswamy, J. Decraene, S. Antonatos, Y. Wang, D. T. Anh ; A. Shi-Nash, "Home and work place predication for urban planning using mobile network data," proceedings from Mobile Data Management (MDM), 2014 IEEE 15th International Conference on Mobile Data Management, 2014, pp. 37-42.

C. Dwork, “Differential Privacy,” in Automata, Languages and Programming, vol. 4052, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Springer Berlin Heidelberg, 2006, pp. 1–12.

L. Fan and L. Xiong, “Real-time aggregate monitoring with differential privacy,” Proc. 21st ACM Int. Conf. Inf. Knowl. Manag. - CIKM ’12, 2012, pp. 2169-2173.

P. Fiadino, D. Valerio, F. Ricciato, K.A. Hummel, ”Steps towards the extraction of vehicular mobility patterns from 3G signaling data,” Traffic Monitoring and Analysis - 4th International Workshop, TMA 2012, Proceedings Lecture Notes in Computer Science, 2012, 7189 LNCS, 2012, pp.66-80.

B. C. M. Fung, K. Wang, R. Chen, P. S. Yu, “Privacy-preserving data publishing,” ACM Comput. Surv., vol. 42, no. 4, 2010, pp. 1–53.

D. Gundlegård and J.M Karlsson, “Generating Road Traffic Information from Cellular Networks - New Possibilities in UMTS”. In: ITS Telecommunications, 2006, pp. 1128-1133.

M.C. González, C.A. Hidalgo, A-L. Barabasi, “Understanding individual human mobility patterns,” Nature, vol 453, 2008, pp. 479-482.

O. Görnerup, “Scalable Mining of Common Routes in Mobile Communication Network Traffic Data”, Proceedings of the 10th International Conference on Pervasive Computing, 2012, pp. 99-106.

S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, “Identifying Important Places in People’s Lives from Cellular Network Data”. Proceedings of the 9th international conference on Pervasive computing, 2011, pp 133-151.

S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, “Human mobility modeling at metropolitan scales,” Communications of the ACM, volume 56, issue 1, 2013, pp. 74-82.

S. Isaacman, R. Becker, R. Cáceres, J. Rowland, A. Varshavsky, S. Kobourov, “A tale of two cities. HotMobile’, The 11th Workshop on Mobile Computing Systems and Applications, 2010, pp. 19-24.

Md. S. Iqbal, C.F. Choudhury, P. Wang, M.C. González, “Development of origin-destination matrices using mobile phone call data,” Transportation Research Part C, 2014, vol 40, pp. 63-74.

F. D. McSherry, “Privacy integrated queries: an extensible platform for privacy-preserving data analysis,” SIGMOD Conference, 2009, pp. 19–30.

I. Miliaraki, K. Berberich, R. Gemulla and S. Zoupanos. “Mind the gap: Large-scale frequent sequence mining”. ACM SIGMOD International Conference on Management of Data (SIGMOD), 2013.

D. J. Mir, S. Isaacman, R. Caceres, M. Martonosi, and R. N. Wright, “DP-WHERE: Differentially private modeling of human mobility,” 2013 IEEE Int. Conf. Big Data, 2013, pp. 580–588.

P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler, “GUPT: privacy preserving data analysis made easy,” in Proceedings of the 2012 international conference on Management of Data - SIGMOD ’12, 2012, pp. 349-360.

Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the Crowd: The privacy bounds of human mobility,” Sci. Rep., vol. 3, 2013.

Y-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, V. D. Blondel, “D4D-Senegal: The Second Mobile Phone Data for Development Challenge”, 2014, http://arxiv.org/pdf/1407.4885v2.

C.H. Mooney and J.F. Roddick. “Sequential pattern mining - approaches and algorithms,” ACM Computing Surveys (CSUR), vol. 45, no. 2, 2013.

D. Naboulsi, M. Fiore, R. Stanica, “Human Mobility Flows in the City of Abidjan”, in V.D. Blondel, N. de Cordes, A. Decuyper, P. Deville, J. Raguenez, S. Zbigniew (eds), Mobile Phone Data for Development, Analysis of mobile phone datasets for the development of Ivory Coast, viewed 29 December 2014, http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4Dbook.pdf, 2013, pp. 594-601.

B. Rajna, “Mobility analysis with mobile phone data”, Master Thesis, Linköping University, LiU-ITN-TEK-A-14/003-SE, 2014.

C. Ratti, D. Frenchman, R.M. Pulselli, S. Williams, ”Mobile landscapes: Using location data from cell phones for urban analysis”, Environment and planning B: Planning and design, vol. 33, no. 5, 2006, pp. 727-748.

M.R. Vieira, V. Frías-Martínez, N. Oliver, E. Frías-Martínez, “Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics”, Proceedings – SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust, 2010, pp. 241-248.

H. Wang, F. Calabrese, G. Di Lorenzo, C. Ratti, “Transportation mode inference from anonymized and aggregated mobile phone call detail records’, IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2010, pp. 318-323.

P. Wang, T. Hunter, A.M. Bayen, K. Schechtner, M.C. González, “Understanding Road Usage Patterns in Urban Areas,” 2012, Scientific reports, http://dx.doi.org/10.1038/srep01001

A.G. Wilson, A statistical theory of spatial distribution models, Transportation Research, Volume 1, Issue 3, 1967, pp. 253-269.

M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker, and I. Stoica. "Spark: cluster computing with working sets," In Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, HotCloud’10, 2010.

T10

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Groupe 8
PredicSis

The municipality and the territory: two scales of understanding the city

Project Summury

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.

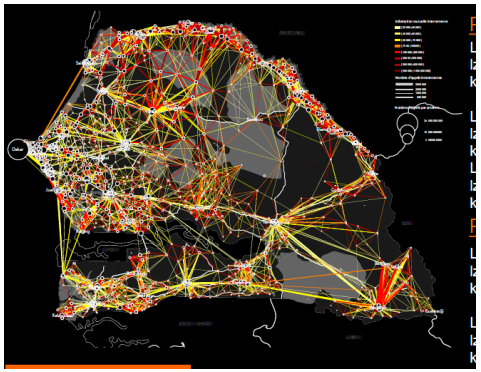
2. Dakar unbalance the network, in terms of land use that would argue for a greater balance between Dakar and the "desert" of Senegal.

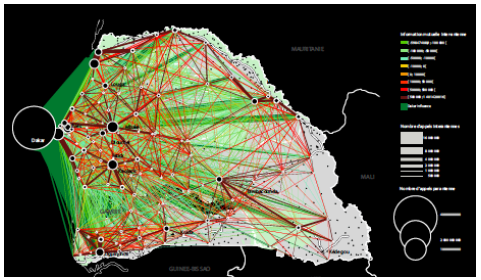
3. The important role of natural barriers, like the Gambia, but also the areas of "fossil valleys".

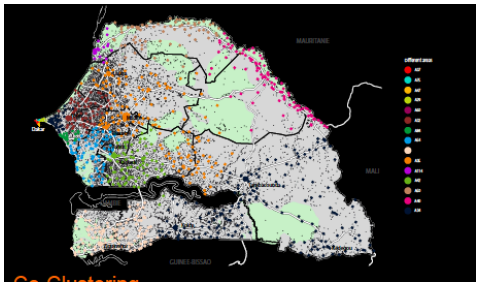
4. Calls are close to relatives, from one city to another with the exception of Dakar.

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.

Question

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.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

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.

Results

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.

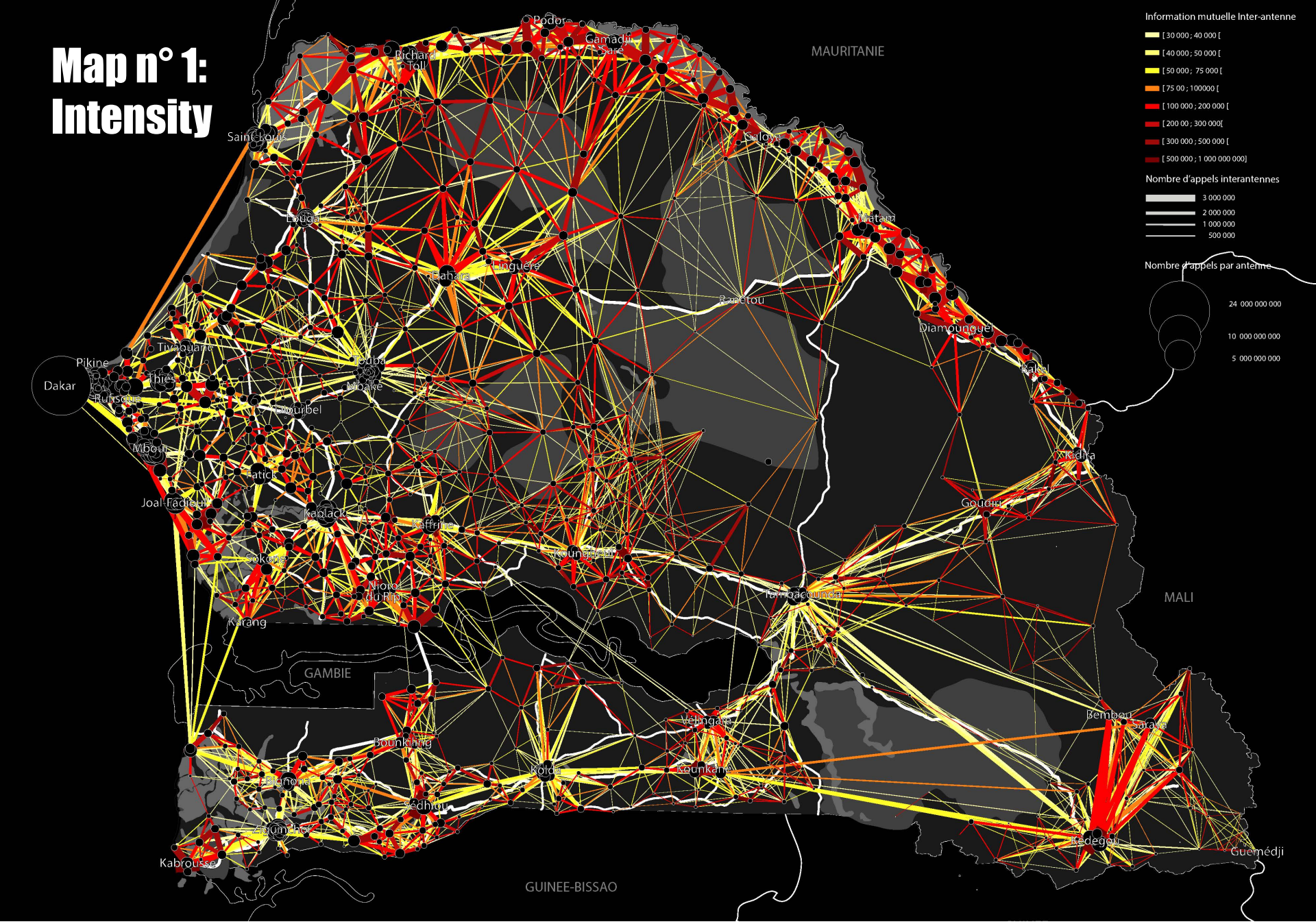
6. The network of cities of the Senegal River is remarkable and it is an entity which takes place along the river and not just at the delta. We can assume an even larger network with the cities of Mauritania on the other side of the river that are not on the map.

7. The Casamance is not connected with the rest of the country and develops its own network between Lower and Upper Casamance.

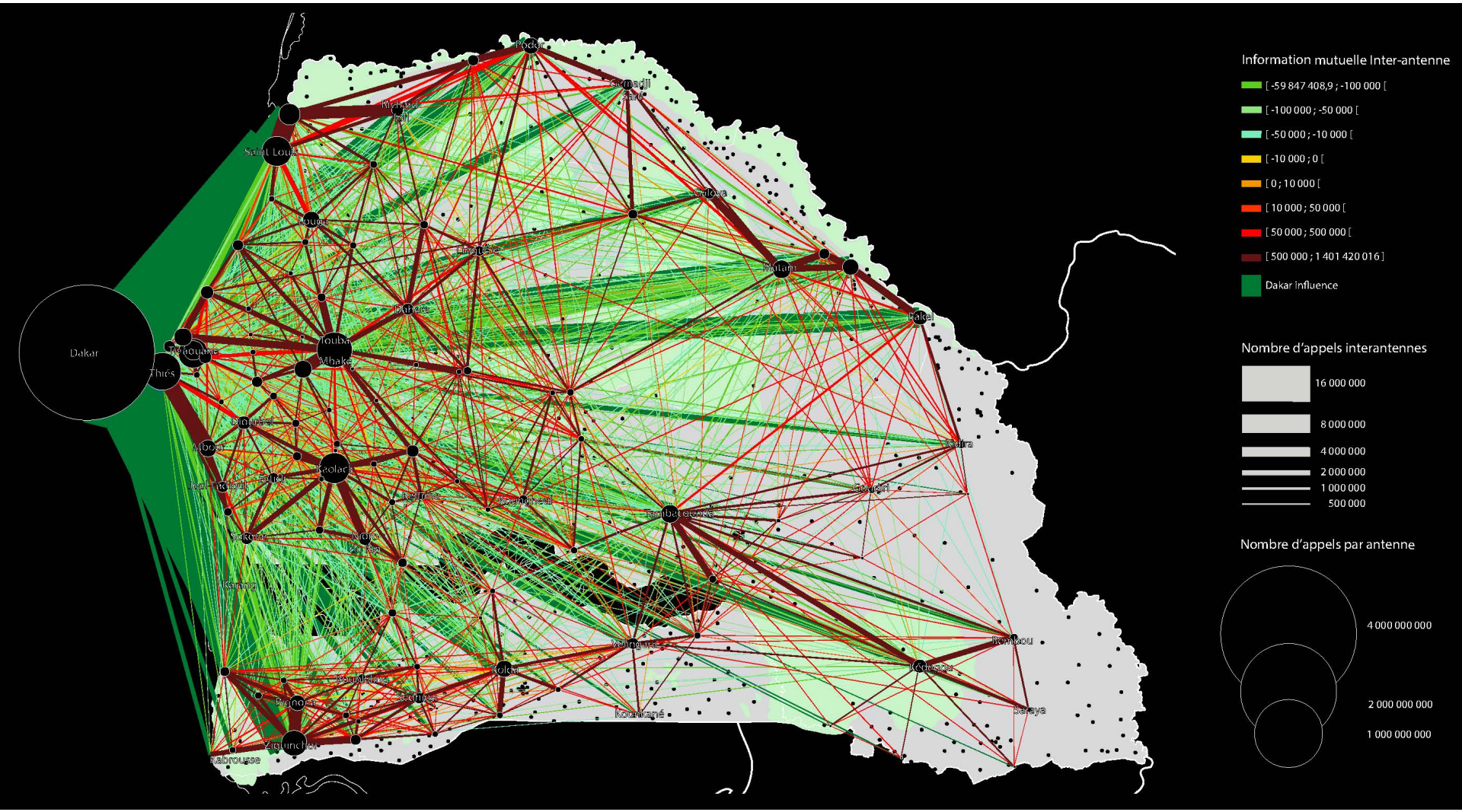
8. Tambacounda is quite independent and acts as a regional centre with strong connections with the River, with High Casamance and Kedougou close to the border.

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

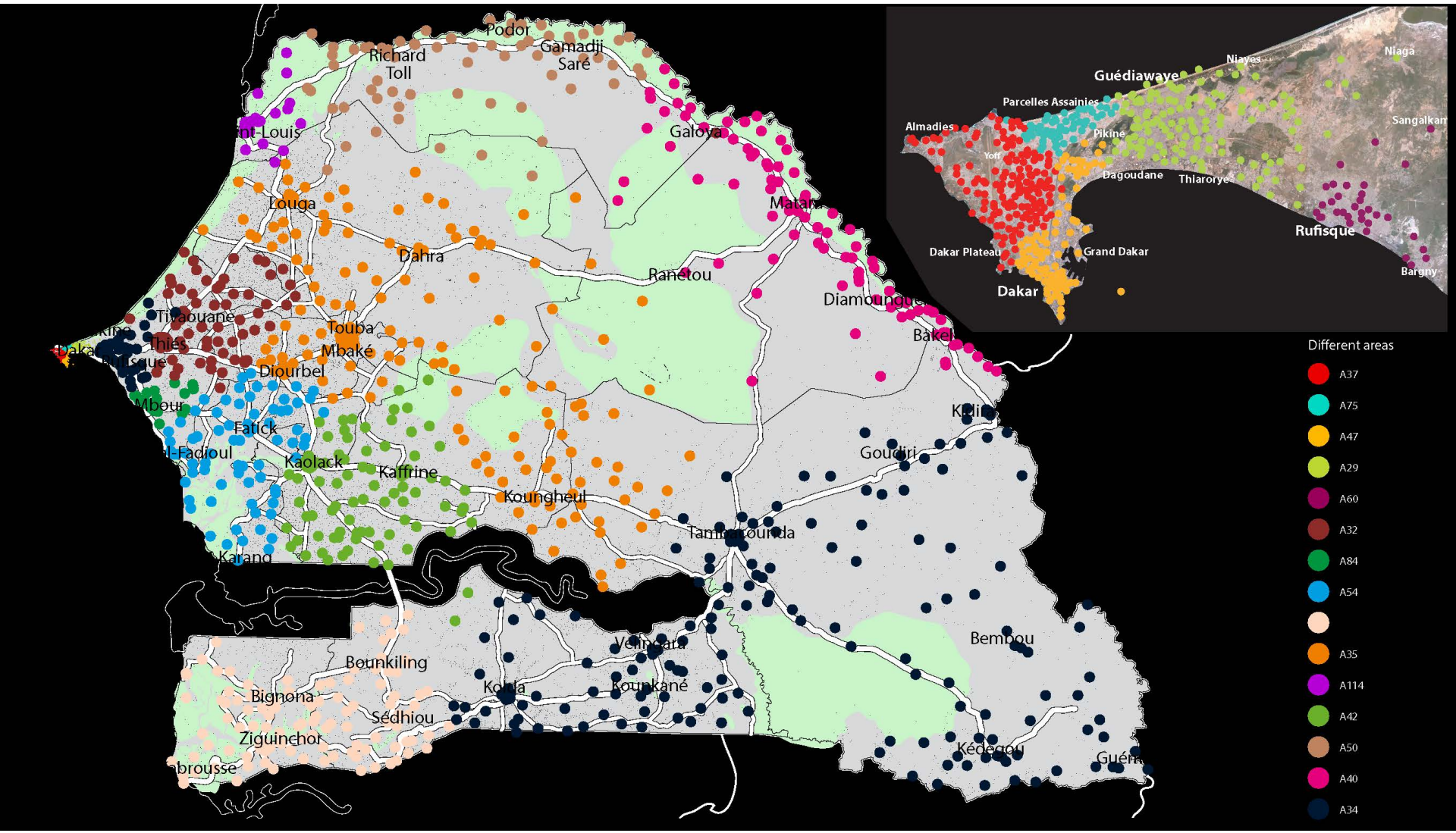
Map n° 1: Intensity



Map n° 2: Flow inter-antennas



Map n° 3: Co-clustering



Results

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 a 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.

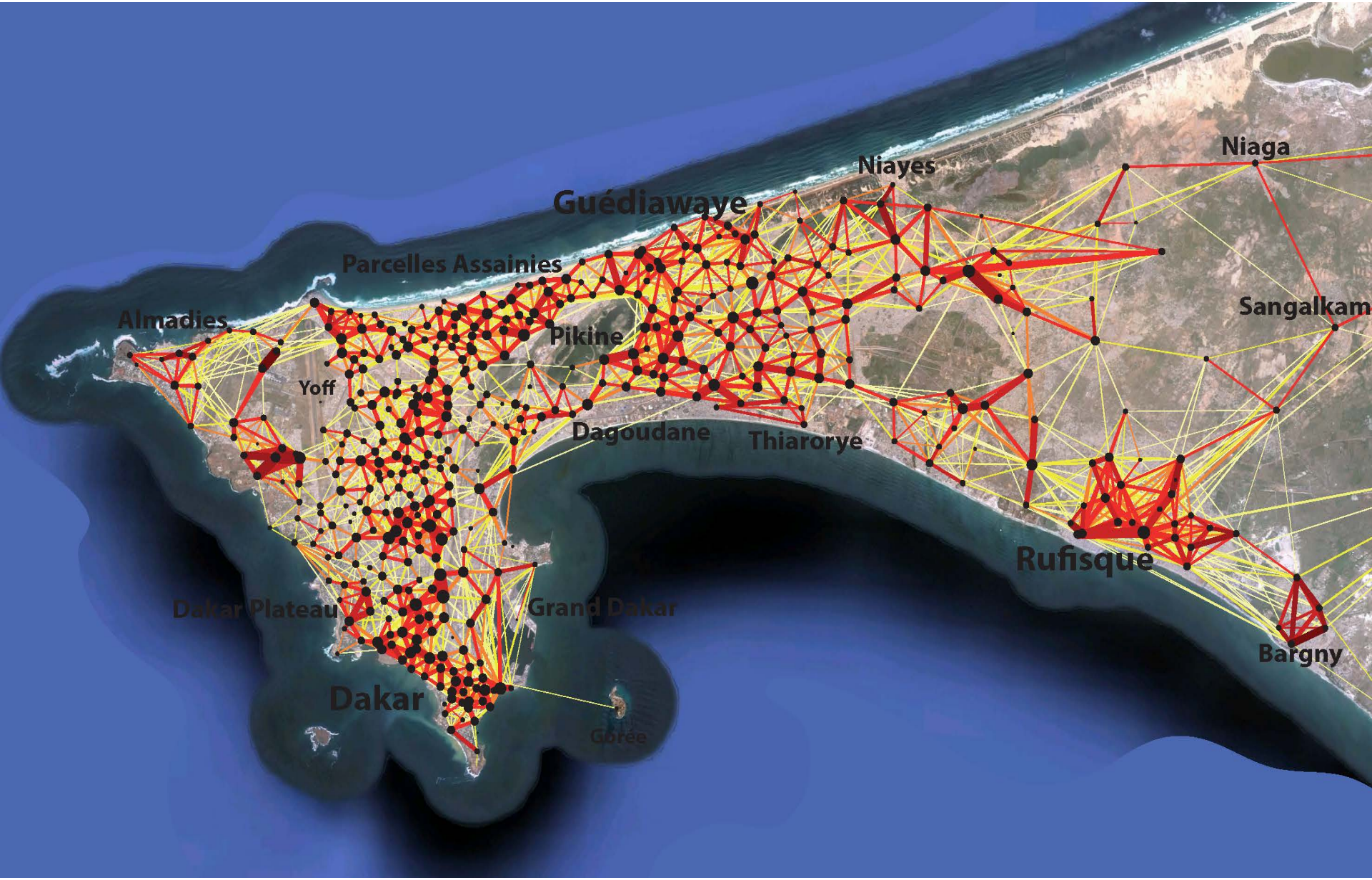
3. Develop regional balance

4. Anchor Dakar in a larger territory

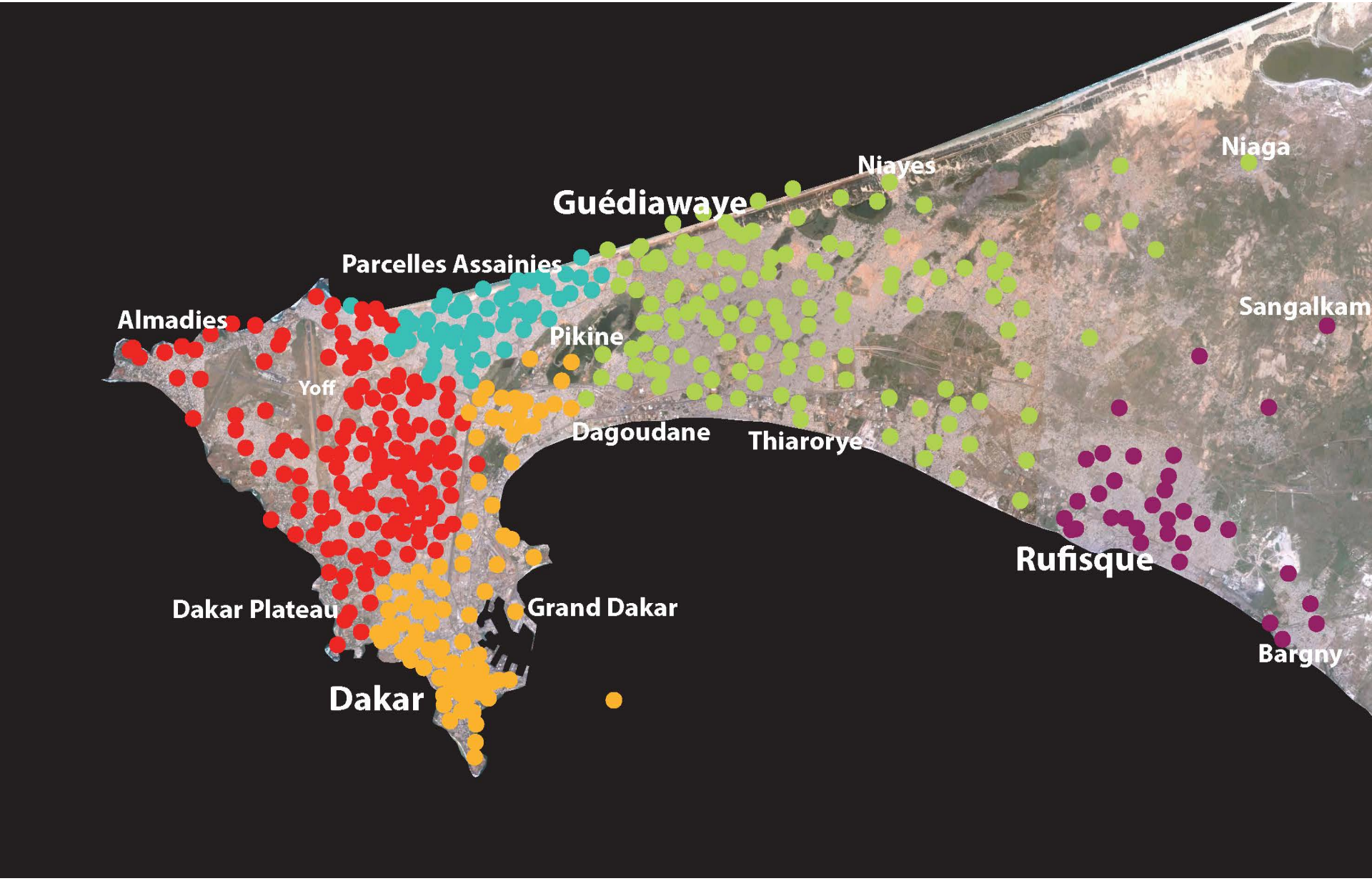
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Focus on Dakar

Map Dakar n°1: Intensity



Map Dakar n°2: Co-clustering



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;

Team

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-Besset, Hadrien Chicault, Cedric Thao
www.predicsis.com

T15

Using mobile phone data for Spatial Planning simulation and Optimization Technologies (SPOT)

Serigne Gueye[1], Babacar M. Ndiaye[3], Didier Josselin[3], Michael Poss[5], Roger M. Faye[2], Philippe Michelon[1], Cyrille Genre-Grandpierre[3], and Francesco Ciari[4]

[1] Laboratoire d'Informatique d'Avignon (LIA), Université d'Avignon, France
{serigne.gueye,philippe.michelon}@univ-avignon.fr
[2] Laboratoire de Traitement de l'Information (LTI),
ESP- Cheikh Anta Diop University, Dakar, Sénégal
roger.faye@ucad.edu.sn
[3] Laboratoire de Mathématiques de la Décision et d'Analyse Numérique (LMDAN)
FASEG-University of Cheikh Anta Diop, Sénégal
babacarm.ndiaye@ucad.edu.sn
[4] Institute for Transport Planning and Systems (IVT), Zurich, Switzerland
ciari@ivt.baug.ethz.ch
[5] UMR Etude des Structures, des Processus d'Adaptation et des Changements de l'Espace (ESPACE), Avignon, France
{didier.josselin,cyrille.genre-grandpierre}@univ-avignon.fr
[6] Heudyasic, UTC, France. mjposs@gmail.com

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 6 [7] 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 futur 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. Nevetheless, 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 concernes 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 garantee 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.

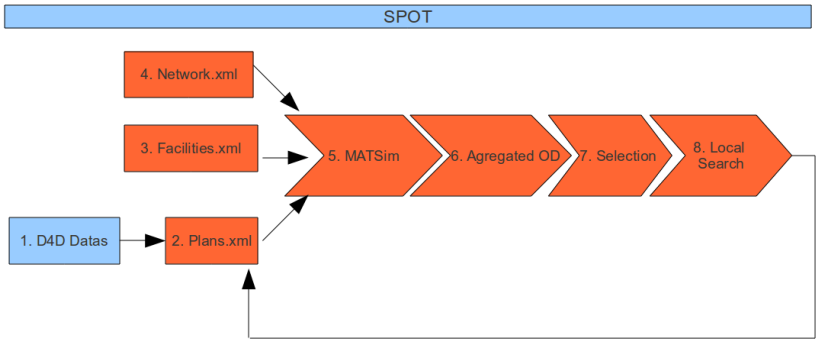


Fig. 1. SPOT

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 poplate the file network.xml, and the amenities are used for the file facility.xml.

Most of time, the type of amenities in the list was not correctly annoted. We processed a semi-automatic assignment using specific requests in the QGIS data base. 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 informations 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

Ville	Arrondissement	Population	Pourcentage (%)
Dakar	Plateau	215.343	8,71
	Grand-Dakar	253.434	10,25
	Almadies	121.006	4,90
	Parcelles Assainies	237617	9,61
Pikine	Thiaroye	239.053	9,67
	Dagoudane	461.648	18,68
	Niayes	209.859	8,49
Guédiawaye		435.350	17,61
Rufisque		160.860	6,51
Bargny		41.220	1,67
Sébikotane		19.400	0,78
Zone rurale		76.940	3,11
TOTAL		2.471.730	

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 SENEGAL-ARR.csv of the challenge data.

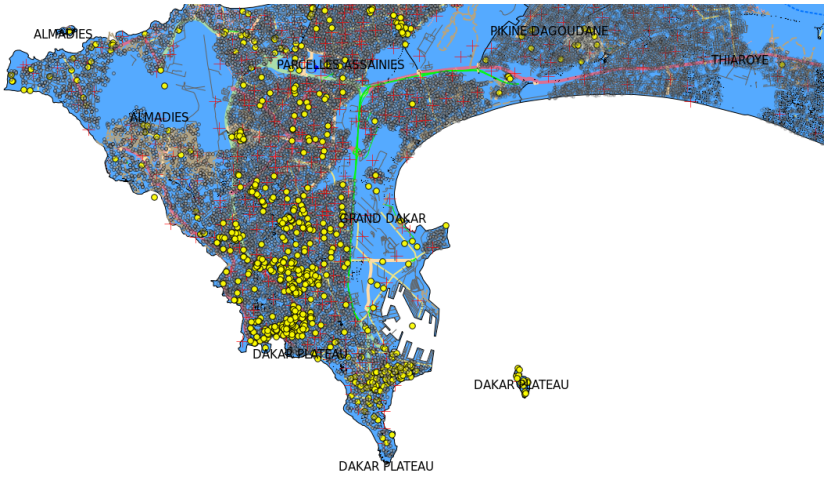


Fig. 2. Homes (grey points), facilities (yellow points), antennas (crosses), and district locations

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/

3 Generating MATSim plan files from mobile phone data

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" >
.     < act type = "home" facility = "1202" x = " − 17.446145" y = "14.7382253" end_time = "11 : 10 : 00"/ >
.     < leg mode = "car" >
.     < /leg >
.     < act type = "shop" facility = "476" x = " − 17.388872" y = "14.7697" end_time = "15 : 30 : 00"/ >
.     < leg mode = "car" >
.     < /leg >
.     < act type = "home" facility = "1202" x = " − 17.446145" y = "14.7382253" end_time = "23 : 59 : 59"/ >
.   < /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 $15h30:00$ 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:M0:00): 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

1,2013-01-08 19:30:00,323
1,2013-01-08 21:00:00,323
1,2013-01-09 11:00:00,323
1,2013-01-09 14:50:00,323
1,2013-01-09 15:10:00,318
1,2013-01-09 15:10:00,318
1,2013-01-09 15:20:00,318
1,2013-01-09 20:50:00,323
1,2013-01-09 21:40:00,323
1,2013-01-09 21:40:00,318
1,2013-01-09 21:50:00,318

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 agregated 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:

2,2013-01-10 19:30:00,408
2,2013-01-10 19:30:00,416
2,2013-01-10 19:30:00,419

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 $p^j = (c^j, s^j, e^j, a^j, l^j, m^j)$ 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 $T^j =< p_1^j, p_2^j, ..., p_n^j >$.*

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, l^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 it geographical location assigned to the attributes l^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 additionaly 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 combinatory 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 **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 **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 $t^j = (v^j, w^j)$ 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_i^j 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_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 :

$$d(t^j, t^k) = \sum_{i=1}^{144} \chi(v_i^j, v_i^j) + \sum_{i=1}^{144} \chi(w_i^j, w_i^j)$$

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 π 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 π 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 maraichers”: 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, whith 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 Assignement 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.

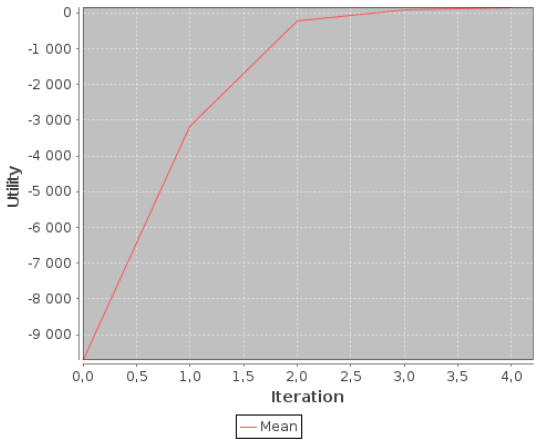


Fig. 3. Utilities Mean through simulation iterations $SET2\_P01.CSV$

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 explaination above. We should intuitively expect that while utility increases, travel time decreases. However, some agents may realize more important activities, explaining this counter-intuive 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

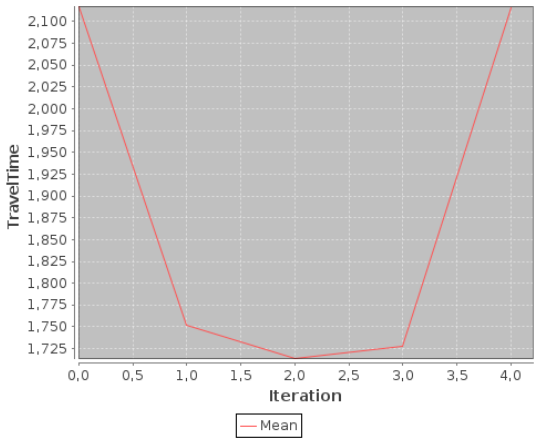


Fig. 4. Travel Times Mean $SET2\_P01.CSV$

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 [12]. This movie is available in the dropbox link given in footnote [13]. 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 $SET2\_P05.CSV$ 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

[12] http://http://www.senozon.com/
[13] https://www.dropbox.com/s/syxgkn9w7w69q12/SPOT-SENOZON.mov?dl=0

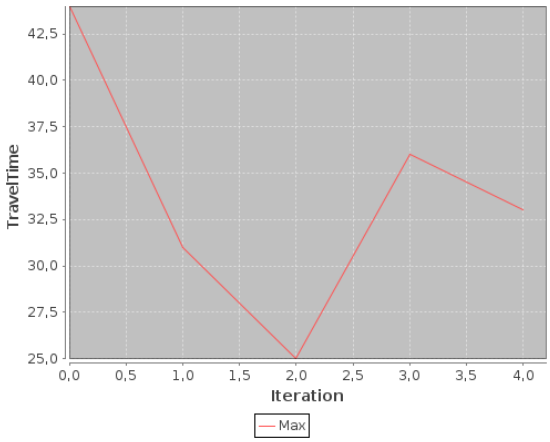


Fig. 5. Travel Times Max $SET2\_P01.CSV$

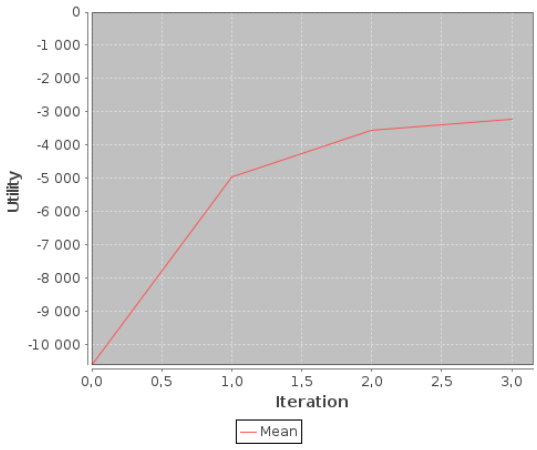


Fig. 6. Utilities Mean for $SET2\_P05.CSV$

iteration by iteration.

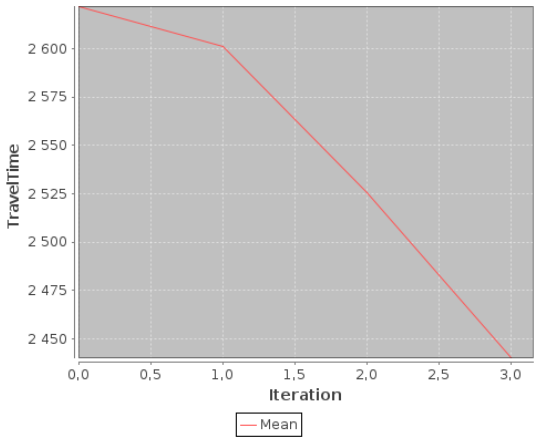


Fig. 7. Travel Times Mean $SET2\_P05.CSV$

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

1. M. Balàc, F. Ciari, C. Genre-Grandpierre, F. Voituret, S. Gueye, and P. Michelon. Decoupling accessibility and automobile mobility in urban areas. In *Transport Research Arena, Paris*, 2014.
2. M. Balmer, M. Rieser, K. Meister, N. Lefebvre D. Charypar, and K. Nagel. Matsim-t: Architecture and simulation times. *Multi-Agent Systems for Traffic and Transportation Engineering*, pages 57–78, 2009.
3. V.D. Blondel, N. De Cordes, A. Decuyper, P. Deville, J. Raguenez, and Z. Smoreda. Mobile phone data for development : Analysis of the mobile phone datas for the development of ivory coast. Technical report, 2013.
4. R-E. Burkard, E. Cela, P-M. Pardalos, and L. Pitsoulis. *The quadratic assignment problem*, pages 241–238. Kluwer Academic Publishers, Dordrecht, 1998.
5. CETUD. Enquête sur la mobilité, le transport et les services urbains à dakar (emtsu) - 2000. Technical report, 2000.
6. MMTTA CETUD, GMAT Ltée. Projet de plan de déplacements urbains pour l'agglomération de dakar, (ppdud) horizon 2025. Technical report, 2007.
7. D. Charypar and K. Nagel. Generating complete all-day activity plans with genetic algorithms. *Transportation*, 32(4):369–397, 2005.
8. M. S. Daskin. *Network and Discrete Location: Models, Algorithms and Applications*. John Wiley and Sons, Inc., New York, 1995.
9. J.H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, Michigan; re-issued by MIT Press, 1975.
10. H. Horni, D.M. Scott, M. Balmer, and K.W. Axhausen. Location choice for shopping and leisure activities implemented in the activity-based multi-agent transport simulation matsim. In *8th Swiss Transport Research Conference*, 2008.
11. T-C. Koopmans and M-J. Beckmann. Assignment problems and the location of economic activities. *Econometrica*, 25:53–76, 1957.
12. B.M. Ndiaye, M.A.M.T. Baldé, and S. Gueye. Rapport technique - activitÃ© 1 : Optimisation des réseaux de transport : Analyse des flots et localisation dâactivité (ortrans). Technical report, 2012.
13. M. Steinbach P-N. Tan and V. Kumar. *Introduction to Data Mining*. 2005.
14. P-M. Pardalos, F. Rendl, and H. Wolkowicz. The quadratic assignment problem: A survey and recent developments. In *In Proceedings of the DIMACS Workshop on Quadratic Assignment Problems, volume 16 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 1–42. American Mathematical Society, 1994.

T16

Where do we Develop? Discovering Regions for Urban Investment in Senegal

Derek Doran
Dept. of Computer Science & Engineering
Kno.e.sis Research Center
Wright State University
derek.doran@wright.edu

Andrew Fox
Dept. of Industrial Engineering and
Management Science
Northwestern University
andrewfox2014@u.northwestern.edu

Veena Mendiratta
Bell Laboratories
Alcatel-Lucent
veena.mendiratta@alcatel-lucent.com

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 numbers 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 long-standing 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

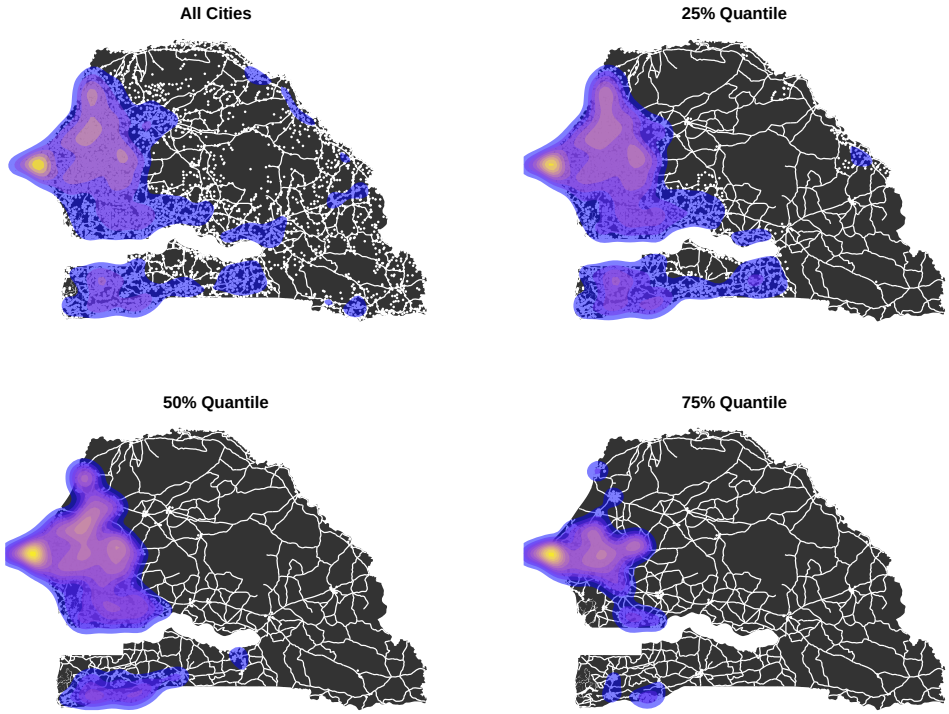


Figure 1: The affect of urbanization in Senegal. Each plot filters out cities whose population density falls below the stated quartile. The 25% largest cities are mostly concentrated in a band east of the capital city of Dakar.

vices 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. *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 *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-

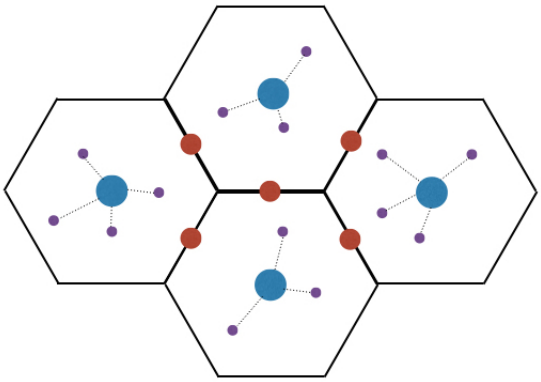


Figure 2: Idealistic spatial heirarchy of Central Places (blue), Low Places (Purple), and Middle Places (red). Hexagons correspond to the region Central Places influence by providing low- and high-level outputs. Low places rely on the Central Place to thrive. Middle places are necessarily self-sustaining due to their distance between Central Places.

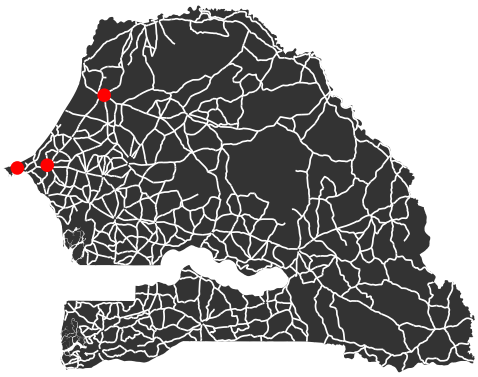


Figure 4: Locations of Dakar, Louga, and Thies. Dakar is 178km away from Louga and 77km from Thies. Louga is 114km away from Thies.

tures relevant to Central Place and Flow theory for use in our analysis. Section 4 discusses the results of our model and the most promising places it identifies for investment. Conclusions and future work are presented in Section 5.

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

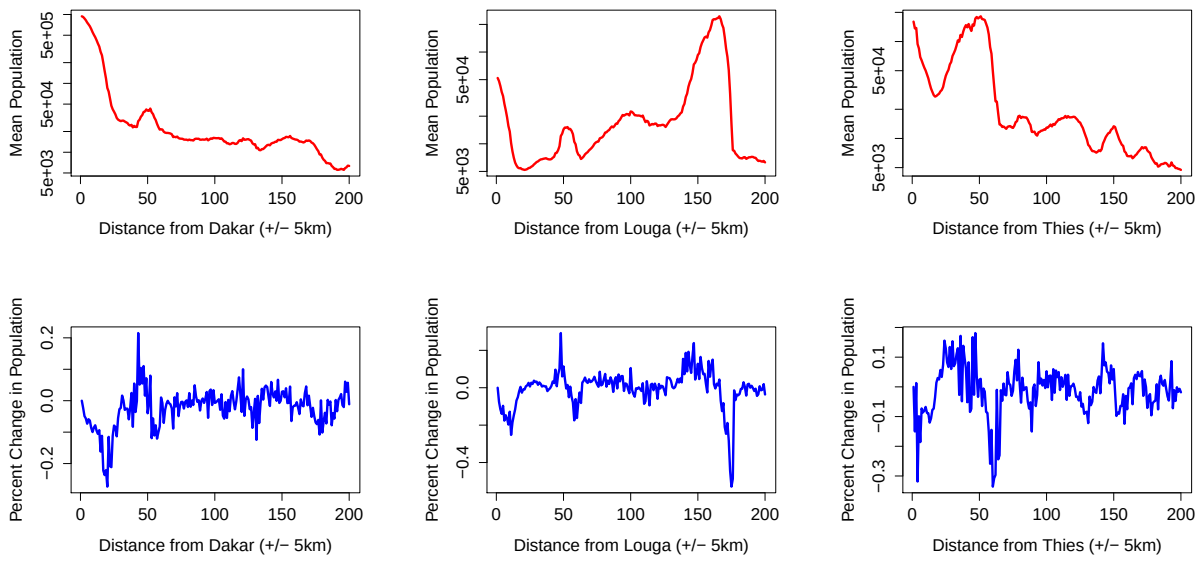


Figure 3: Comparing changes in city populations with distance from large urban centers

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 form 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 ouputs that drive their economy, but their further development hinges on the free exchange of ideas and integration

[3] http://www.fallingrain.com/world/index.html

of 'foreign' commerce (CFT) [34, 33]. Agent-based simulations further explain the interlocking relationship between CPT and CFT for Central Place development [21]. We hypothesize 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 integrate 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 economic and infrastructure investments.

To evaluate the degree to which CFT holds across Senegal, 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. Figure 5 plots the distribution of the total duration of all conversations 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 individual places. Towards this end, we only consider communication 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 communication 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 activity as described above. To evaluate the popularity of a calling 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 considers the popularity p_i of calling tower i to be proportionate to the popularity of the towers it communicates with. It is defined by:

$$p_i = \alpha \sum_j \mathbf{A}_{ij} \frac{p_j}{g_j} + (1-\alpha)\frac{1}{N}$$

where $\mathbf{A}$ is a matrix with $\mathbf{A}_{i,j}$ given as the cumulative length of all conversations between towers i and j, k_j is the degree 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 Figure 6, we compare the location of the 10 most populated cities in Senegal in the Global Gazetteer against the location and PageRank centrality of calling towers (larger vertices 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 centralities is skewed, many call towers with high PageRank lying

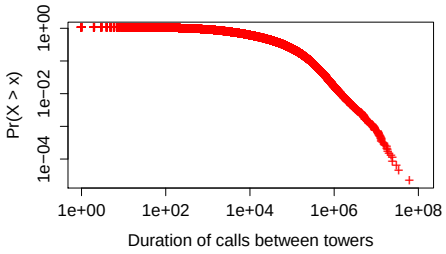


Figure 5: Distribution of total calling times across towers

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 information with Central Places. According to CFT, such exchanges 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 unsupervised methods for identifying areas in the country most likely to correspond to Middle Places. We intentionally decided 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 theoretically 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 because: (i) high-resolution data expressing the calls made between villages, towns, and cities are unavailable; (ii) government 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 Middle 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, according 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:

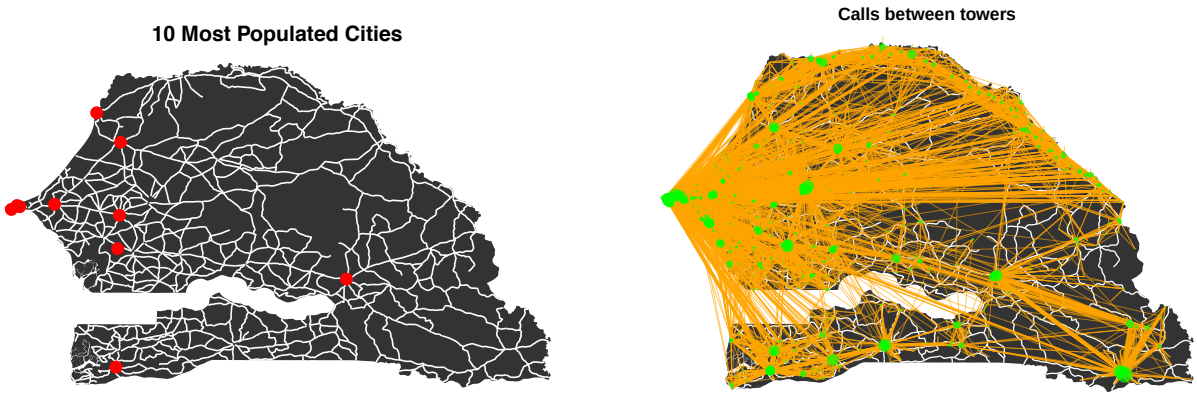


Figure 6: Spatial comparison of the most populated cities in Senegal and Pagerank centrality of calling towers

- **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 features 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 across the country 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

Distance Traveled by X% of Calls	Correlation with Self-Sufficiency	Variance of Distance Traveled
50% (median)	-0.58	454
60%	-0.37	854
70%	-0.17	3,581
80%	0.10	10,872

Table 1: Correlation between distance of calls and self-sufficiency features

Solution	Variables	BIC	Pseudo-F
Best	Self-Sufficiency, Partnership, Betweenness Distance ($X = 60\%$)	-1,288	41.6
Alt. A	Self-Sufficiency, Partnership, Betweenness, Distance ($X = 50\%$)	-1,297	46.2
Alt. B	Self-Sufficiency, Partnership, Betweenness, Demand-Weighted Dist.	-1,342	31.0
Alt. C	Self-Sufficiency, Partnership, PageRank Centrality, Demand-Weighted Dist.	-1,316	38.4

Table 2: Clustering solutions with different variable settings

of mixture models and number of clusters [12].

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 Θ, $d = |\Theta|$, 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 Psudeo-F scores correspond to 'tighter' clusterings where intra-cluster distances between vectors is small and inter-cluster distances are high.

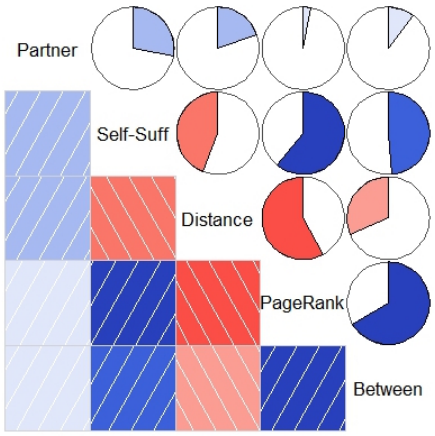


Figure 8: Correlations between transformed and standardized features. PageRank shows high ($|r| > 0.5$) correlations with call distance and self-sufficiency, while betweenness is only moderately correlated with Self-Sufficiency.

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

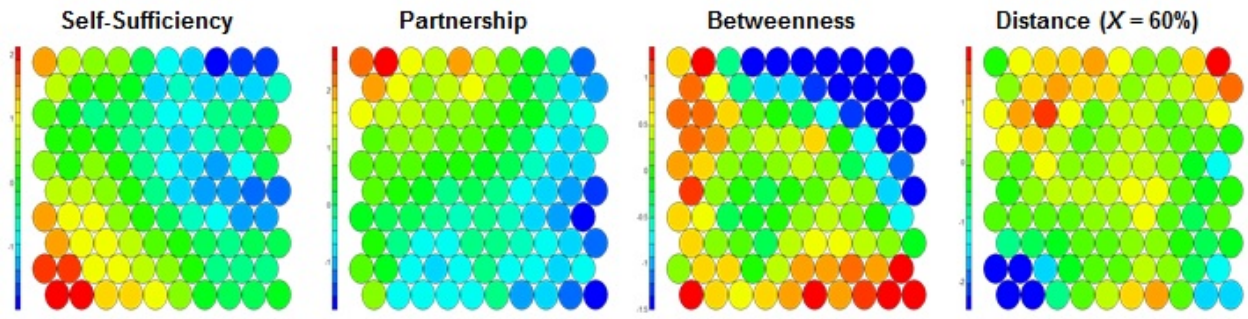


Figure 7: Feature values for best clustering solution

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 Psudeo-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 are 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.

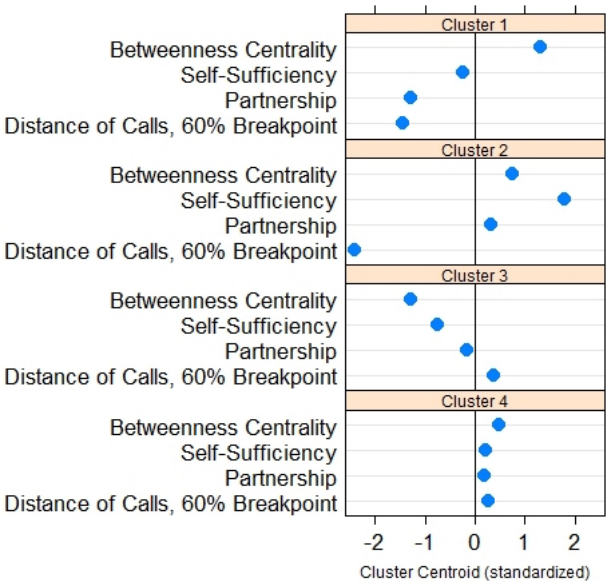


Figure 9: Dot plot of cluster centroids

4 Cluster results and discussion

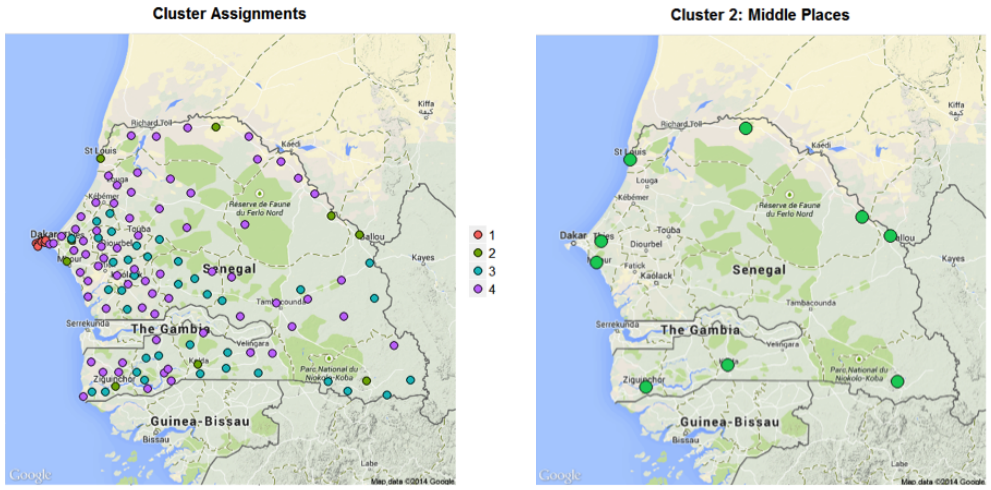
Figure 9 uses a dot plot to present the centroid positions of the four clusters in the best FMM solution. We subjectively map these values to being relatively LOW, MODERATE, or HIGH for each cluster in Table 3. We label each of the four clusters as:

- **Cluster 1: Dakar and its Suburbs**. Cluster 1 identifies 8 arrondissements that, as visualized in Figure 10(a), contain Dakar and its suburbs. These arrondisements 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

Cluster Description	Self-Sufficiency	Partnership	Betweenness	Distance	Cluster Size
Central Places: Dakar and Suburbs	MODERATE	LOW	HIGH	LOW	8
Middle Places: Emerging Opportunities	HIGH	MODERATE	HIGH	LOW	9
Low Places: Villages Between	LOW	MODERATE	LOW	MODERATE	37
Middle-Low Places: Common Towns	MODERATE	MODERATE	MODERATE	MODERATE	69

Table 3: Cluster labels and features



(a) (b)

Figure 10: Cluster assignments (a) and Middle Places (b)

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[4]. It is also a major producer of peanuts and fertilizer that are among the country's top exports, and host reserves of important metals[5]. 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

[5] http://www.britannica.com/EBchecked/topic/592085/Thies

of growth[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[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 [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[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

[1] B. J. Berry and W. L. Garrison. A note on central place theory and the range of a good. *Economic Geography*, pages 304–311, 1958.

[2] B. J. Berry and W. L. Garrison. Recent developments of central place theory. *Papers in Regional Science*, 4(1):107–120, 1958.

[3] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137*, 2012.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[5] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[6] D. J. Casley and D. A. Lury. Data collection in developing countries, 1987.

[7] W. Christaller. *Central places in southern Germany*. Prentice-Hall, 1966.

[8] D. Christian. *Maps of Time: An Introduction to Big History, With a New Preface*, volume 2. Univ of California Press, 2011.

[9] M. J. Daniels. Central place theory and sport tourism impacts. *Annals of Tourism Research*, 34(2):332–347, 2007.

[10] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. D4d-senegal: The second mobile phone data for development challenge. *arXiv preprint arXiv:1407.4885*, 2014.

[11] D. Doran, V. Mendiratta, C. Phadke, and H. Uzunalioglu. The importance of outlier relationships in mobile call graphs. In *Proc. of Intl. Conference on Machine Learning and Applications*, volume 2, pages 24–29. IEEE, 2012.

[12] C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306, 1999.

[13] J. Friedmann. The world city hypothesis. *Development and change*, 17(1):69–83, 1986.

[14] A. Gilbert and J. Gugler. *Cities poverty and development: Urbanization in the third world*. New York NY/Oxford England Oxford University Press, 1982.

[15] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*. Springer, 2009.

[16] V. Henderson. Urbanization in developing countries. *The World Bank Research Observer*, 17(1):89–112, 2002.

[17] P. M. Hohenberg and L. H. Lees. *The making of urban Europe, 1000-1994*. Harvard University Press, 1995.

[18] W.-T. Hsu, T. J. Holmes, and F. Morgan. Optimal city hierarchy: A dynamic programming approach to central place theory. In *Meeting Papers from Society for Economic Dynamics*, 2009.

[6] http://http://en.wikipedia.org/wiki/Saint-Louis,_Senegal

[7] http://en.wikipedia.org/wiki/M'Bour

[8] http://www.britannica.com/EBchecked/topic/657131/Ziguinchor

[9] http://en.wikipedia.org/wiki/Ziguinchor

[19] K. Ikeda, K. Murota, T. Akamatsu, T. Kono, Y. Takayama, G. Sobhaninejad, and A. Shibasaki. Self-organizing hexagons in economic agglomeration: core-periphery models and central place theory. Technical report, Technical Report METR 2010–28. Department of Mathematical Informatics, University of Tokyo, 2010.

[20] J. Jacobs. *The death and life of great American cities*. Random House LLC, 1961.

[21] D. Knitter. Central places and the environment, 2013.

[22] L. K. Lim, F. Acito, and A. Rusetski. Development of archetypes of international marketing strategy. *Journal of International business studies*, 37(4):499–524, 2006.

[23] E. Linden. The exploding cities of the developing world. *Foreign Affairs*, pages 52–65, 1996.

[24] K. Lynch. *Good city form*. MIT press, 1984.

[25] J. Malone, K. McGarry, S. Wermter, and C. Bowerman. Data mining using rule extraction from kohonen self-organising maps. *Neural Computing & Applications*, 15(1):9–17, 2006.

[26] E. Malthouse. *Segmentation and lifetime value models using SAS*. SAS Institute, 2013.

[27] P. McCann and F. van Oort. Theories of agglomeration and regional economic growth: a historical review. *Handbook of regional growth and development theories*, pages 19–32, 2009.

[28] G. F. Mulligan. Agglomeration and central place theory: a review of the literature. *International Regional Science Review*, 9(1):1–42, 1984.

[29] G. F. Mulligan, M. D. Partridge, and J. I. Carruthers. Central place theory and its reemergence in regional science. *The Annals of Regional Science*, 48(2):405–431, 2012.

[30] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.-L. Barabasi. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States*, 104:7332–7336, 2007.

[31] A. I. Saichev, Y. Malevergne, and D. Sornette. *Theory of Zipf's law and beyond*, volume 632. Springer, 2009.

[32] S. Sassen. *The global city*. Princeton University Press Princeton, NJ, 1991.

[33] E. W. Soja. Cities and states in geohistory. *Theory and Society*, 39(3-4):361–376, 2010.

[34] P. J. Taylor. Extraordinary cities: Early city-ness and the origins of agriculture and states. *International Journal of Urban and Regional Research*, 36(3):415–447, 2012.

[35] P. J. Taylor, M. Hoyler, and R. Verbruggen. External urban relational process: introducing central flow theory to complement central place theory. *Urban Studies*, 47(13):2803–2818, 2010.

[36] R. Wehrens and L. M. Buydens. Self-and super-organizing maps in r: the kohonen package. *Journal of Statistical Software*, 21(5):1–19, 2007.

O02

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 anonimizability 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 communicaiton 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 sensible 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 timespans 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.

Pseudo-id	Gender	Age	ZIP	Degree	Income	...
00013701	Male	21	77005	Bachelor	13,000	...
08936402	Male	37	77065	Master's	90,000	...
42330327	Female	60	89123	High School	46,000	...
...	...	...	...	...	...	...

Table 2: Mobile traffic database format.

Pseudo-id	Spatiotemporal samples (fingerprint)						
a	c_1,8	c_2,14	c_3,17				
b	c_4,8	c_5,15	c_6,15	...	c_{13},15	c_{14},16	c_{15},17
c	c_{16},7	c_{17},20					
...	...						

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 reindetifiable, 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 reidentifiable 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*

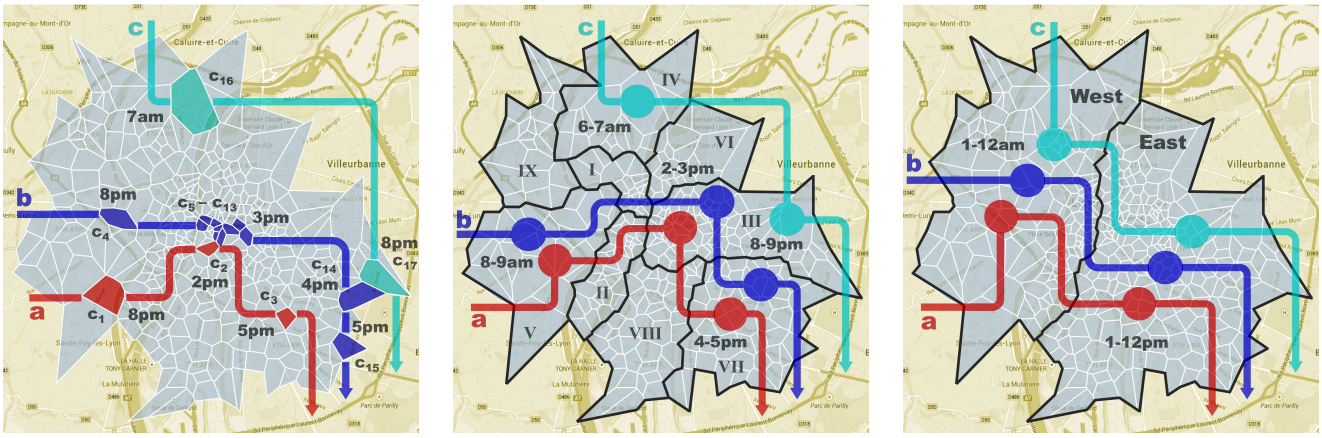
Traditional micro-data databases are structured into matrices where each row maps to one individual, and each column to an *attribute*. An example is provided in Tab.1. Individuals are associated to one *identifier*, i.e., a value that uniquely pinpoints the user across datasets (e.g., his complete name, 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[1] 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

[1]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.



(a) Initial (b) Aggregated (c) More aggregated

Figure 1: Example of mobile traffic fingerprints of three subscribers. (a) Initial dataset granularity: user locations are represented 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 a 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 metropolitan 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

$$d_{ab}(i,j) = w_s \delta_s \left(s_i^a, s_j^b\right) + w_t \delta_t \left(t_i^a, t_j^b\right). \quad (1)$$

Here, δ_s and δ_t are functions that determine the distance along the spatial and temporal dimensions, respectively. The former thus operates 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 weigth 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 δ_s and δ_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

$$\delta_s \left(s_i^a, s_j^b\right) = \begin{cases} \dfrac{dist\left(s_i^a, s_j^b\right)}{\delta_s^{max}} & \text{if } dist\left(s_i^a, s_j^b\right) \leq \delta_s^{max} \\ 1 & \text{otherwise,} \end{cases} \quad (2)$$

and

$$\delta_t \left(t_i^a, t_j^b\right) = \begin{cases} \dfrac{|t_i^a - t_j^b|}{\delta_t^{max}} & \text{if } |t_i^a - t_j^b| \leq \delta_t^{max} \\ 1 & \text{otherwise.} \end{cases} \quad (3)$$

In (2), $dist\left(s_i^a, s_j^b\right) = |s_i^a.x - s_j^b.x| + |s_i^a.y - s_j^b.y|$ 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

$$\Delta_{ab} = \begin{cases} \dfrac{1}{n_a} \displaystyle\sum_{h=1}^{n_a} \min_{k=1,\ldots,n_b} d_{ab}(h,k) & \text{if } n_a \geq n_b \\ \dfrac{1}{n_b} \displaystyle\sum_{h=1}^{n_b} \min_{k=1,\ldots,n_a} d_{ab}(k,h) & \text{otherwise.} \end{cases} \quad (4)$$

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 Δ_{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

$$\Delta_a^k = \frac{1}{k-1} \sum_{b \in \mathbb{N}_a^{k-1}} \Delta_{ab}, \quad (5)$$

where $\mathbb{N}_a^{k-1}$ 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 the 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, δ_s^{max} and δ_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 anoymizability in (5) can be intended as a dissimilarity measure, and employed in legacy definitions used to understand micro-data database sparsity, e.g., `(ε,δ)-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

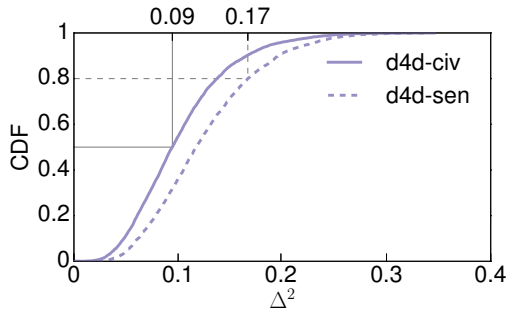
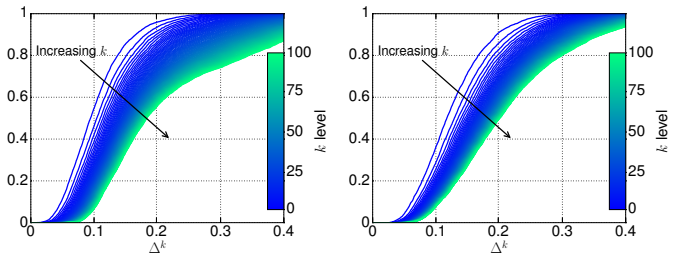


Figure 2: CDF of the anonymizability measure, under the 2-anonymity criterion, in the `d4d-civ` and `d4d-sen` mobile traffic datasets.



(a) d4d-civ (b) d4d-sen

Figure 3: CDF of the anonymizability measure, for varying k of the k-anonymity criterion, in the `d4d-civ` and `d4d-sen` mobile traffic datasets.

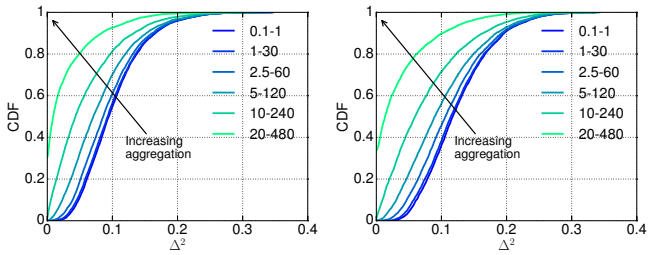
that no single mobile subscriber is 2-anonymous in either of the original datasets, which confirms previous findings on the elevate re-identifiability of mobile traffic datasets [4, 5]. However, the probability mass gathered in both cases in the 0.1-0.2 range, i.e., it is quite close to the origin. This 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

[2]The spatial information maps to the antenna location in `d4d-civ`, and to a random point within the voronoi cell associated to the antenna in `d4d-sen`.

[3]At 100-m spatial granularity, each square cell contains at most one antenna or voronoi location from the original dataset. In other words, this discretization does not implies any spatial aggregation.



(a) d4d-civ (b) d4d-sen

Figure 4: CDF of the anonymizability measure, under the 2-anonymity criterion and for varying spatiotemporal aggregation levels, in the `d4d-civ` and `d4d-sen` mobile traffic datasets. The legend reports the level of spatial (in kilometers) and temporal (in minutes) aggregation each curve refers to.

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 $\mathbb{N}_a^{k-1}$ of k-1 other subscribers whose fingerprints are the closest to that of a, according to (5). Then, we disaggregate all the fingerprint distances Δ_{ab} between a and the users $b \in \mathbb{N}_a^{k-1}$ 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 $\mathbb{S}_a^k = \{w_s\delta_s\}$ and $\mathbb{T}_a^k = \{w_t\delta_t\}$. 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.

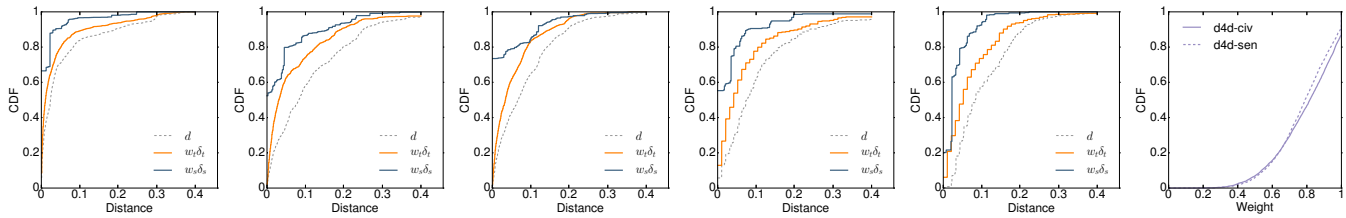
Examples of the spatial and temporal distance distributions we obtain in the case of 2-anonymity are shown in Fig. 5a-5e. Each plot refers to one random user in the `d4d-civ` or `d4d-sen` dataset, and portrays the CDF of the spatial ($w_s\delta_s$) and temporal ($w_t\delta_t$) component distance, as 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., $\sum_{\mathbb{T}_a^k} w_t\delta_t / \sum_{\mathbb{S}_a^k} w_s\delta_s$, 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*

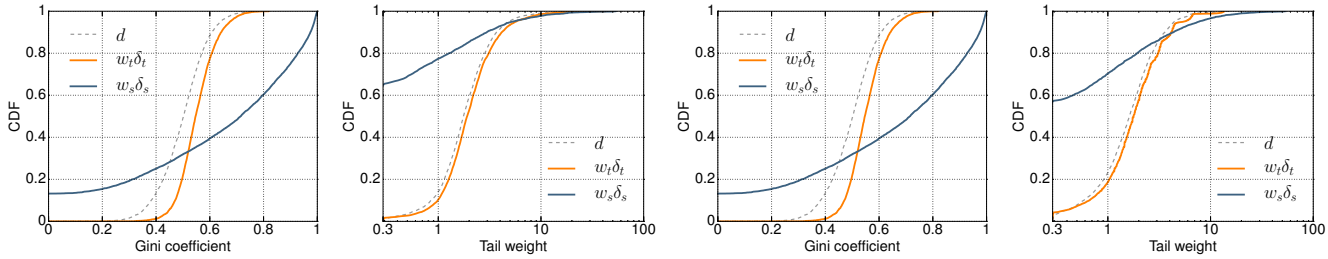
Not only temporal components weight much more than spatial ones in the fingerprint distance, but they also seem to show longer tails in Fig. 5a-5e. Longer tails imply the presence of more samples with a large distance: this, in turn, significantly increases the level of aggregation needed to achieve k-anyonimity, 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 $\mathbb{S} = \{s_i\}$,



(a) d4d-civ, id 370 (b) d4d-civ, id 224 (c) d4d-civ, id 3175 (d) d4d-sen, id 658 (e) d4d-sen, id 2130 (f) Time weight

Figure 5: (a)-(e) CDF of the sample distance, and of its spatial and temporal components, under the 2-anonymity criterion, for five random mobile users in the `d4d-civ` and `d4d-sen` mobile traffic datasets. (f) Contribution of the temporal components to the total sample distance, expressed as the ratio between the sums of temporal component distances and spatial component distances.



(a) d4d-civ, Gini (b) d4d-civ, Tail weight (c) d4d-sen, Gini (d) d4d-sen, Tail weight

Figure 6: (a,c) CDF of the Gini coefficient computed on the sample distance distributions of all users in the `d4d-civ` and `d4d-sen` datasets, for the 2-anonymity criterion. (b,d) CDF of the Tail weight index computed on the sample distance distributions of all users in the `d4d-civ` and `d4d-sen` datasets, for the 2-anonymity criterion.

the coefficient is computed as

$$G(\mathbb{S}) = 1 - \frac{2\sum_{i=1}^{N} i s_i + \sum_{i=1}^{N} s_i}{N \sum_{i=1}^{N} s_i}, \qquad (6)$$

where N is the cardinality of $\mathbb{S}$. We compute the Gini coefficient on the sets $\mathbb{S}_a^k$ and $\mathbb{T}_a^k$, for all users a.

The second metric is the Tail weight index [10], which quantifies the weight of the tail of a distribution with empirical CDF F as

$$T_F = \frac{F^{-1}(0.99) - F^{-1}(0.5)}{F^{-1}(0.75) - F^{-1}(0.5)} \frac{\Phi^{-1}(0.75) - \Phi^{-1}(0.5)}{\Phi^{-1}(0.99) - \Phi^{-1}(0.5)}. \qquad (7)$$

In the expression above, $F^{-1}(\cdot)$ is the inverse function of the empirical CDF and $\Phi^{-1}(\cdot)$ is the inverse function of a standard normal CDF. We compute again the Tail weight index on the distributions obtained from both $\mathbb{S}_a^k$ and $\mathbb{T}_a^k$, for all a.

Fig. 6 shows the results returned by the two metrics in the `d4d-civ` or `d4d-sen` datasets. No significant differences emerge among the two mobile traffic datasets. In both cases, the Gini coefficient, in Fig. 6c and Fig. 6a, has, for all mobile user fingerprints (d), high values around 0.5 that denote significant dispersion around the mean. However, two opposite behaviors are observed for the spatial ($w_s\delta_s$) and temporal ($w_t\delta_t$) components. The former show cases where no dispersion at all is recorded (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 im-

[4] As a reference, an exponential distribution with mean equal to 1 has a Tail weight index of 1.6, 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

[1] L Sweeney, *k-anonymity: A model for protecting privacy*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.

[2] P. Ohm, *Broken promises of privacy: Responding to the surprising failure of anonymization*. UCLA L. Rev. 57 (2009): 1701.

[3] A. Narayanan V. Shmatikov, *Robust de-anonymization of large sparse datasets*. Security and Privacy, 2008. SP 2008. IEEE Symposium on. IEEE, 2008.

[4] H. Zang and J. Bolot *Anonymization of location data does not work: A large-scale measurement study*. Proceedings of the 17th annual international conference on Mobile computing and networking. ACM, 2011.

[5] Y. de Montjoye, C.A. Hidalgo, M. Verleysen and V. Blondel, *Unique in the Crowd: The privacy bounds of human mobility*. Nature Sci. Rep. (2013).

[6] A. Machanavajjhala, D. Kifer, J. Gehrke abd M. Venkitasubramaniam, *l-diversity: Privacy beyond k-anonymity*. ACM Transactions on Knowledge Discovery from Data (TKDD) 1.1 (2007): 3.

[7] N. Li, T. Li and S. Venkatasubramanian, *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity*. ICDE. Vol. 7. 2007.

[8] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, L. Murphy, *A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners*. Transactions on Data Privacy, 7:337370, 2014.

[9] E. Krause, *Taxicab geometry*, The Mathematics Teacher, Vol. 66, No. 8 (1973), pp. 695-706.

[10] D. Hoaglin, F. Mosteller and J.W. Tukey *Understanding robust and exploratory data analysis*. Vol. 3. New York: Wiley, 1983.

O05

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

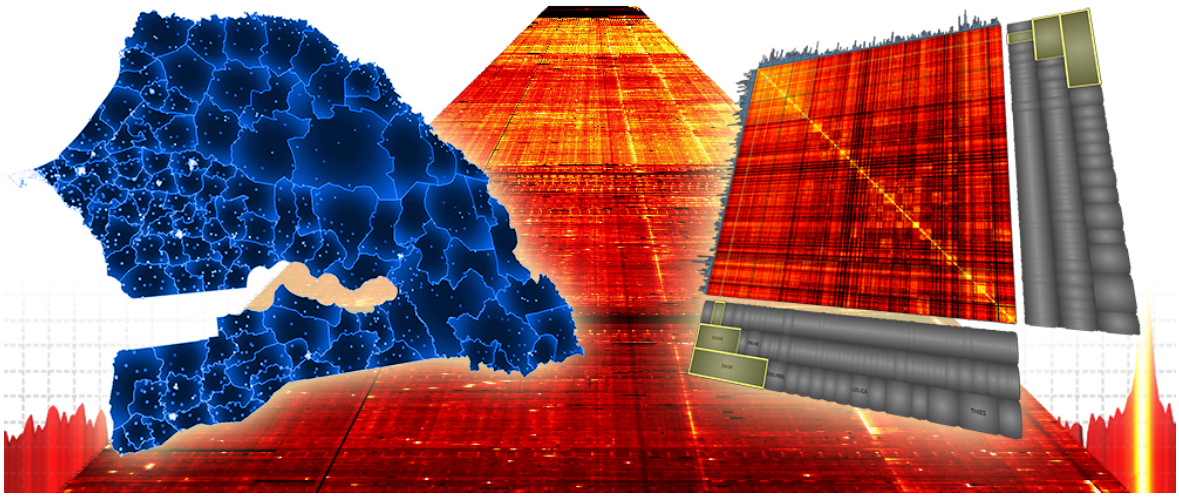


Fig. 1. The visual matrix is a central element in our visual analytics multiple coordinated view approach for the exploration and analysis of massive mobile phone data. On the left, geographical visualization of Senegal divided into 123 arrondissements that contain a total of 1666 cell towers, shown as white dots (correlating with population density). On the right, our scalable and flexible visual matrix showing tower-to-tower call intensity using a heatmap colormap. The visual matrix is enriched with interactive configurable icicle plot and histogram visualizations for enhanced exploration and analysis in the sensemaking process.

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 at 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.

- *Stef van den Elzen, Martijn van Dortmont are with Eindhoven University of Technology and SynerScope B.V. E-mail: {s.j.v.d.elzen, m.a.m.m.v.dortmont}@tue.nl.*
- *Jorik Blaas, Danny Holten, Willem van Hage, Jan-Kees Buenen are with SynerScope B.V. E-mail: {jorik.blaas, danny.holten, willem.van.hage, jan-kees.buenen}@synerscope.com.*
- *Jarke J. van Wijk is with Eindhoven University of Technology. E-mail: j.j.v.wijk@tue.nl.*
- *Robert Spousta, Simone Sala, Alison Kuzmickas, Steve Chan are with Sensemaking Fellowship. E-mail: {spousta, salas, akuzmic}@mit.edu, stevechan@post.harvard.edu*

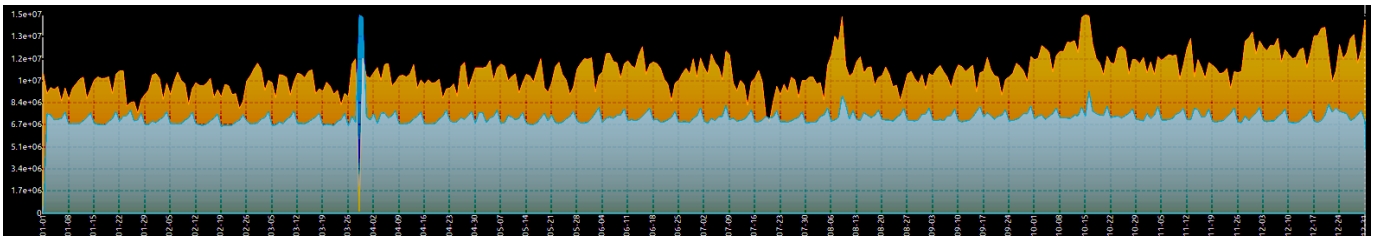


Fig. 2. Number of calls (orange) with the computed change in call behavior overlaid (blue). The change in calls is computed based on the extended Jaccard index (as defined in previous work [47]), rather than taking the absolute difference of the number of calls for two points in time. For the number of calls, heightened communication is present at several religious events (Ramadan, Feast of Sacrifice) and New Years Eve. We also see a significant drop in communication at March 29. For the change in call behavior we also see these events as spikes, as well as a regular weekly pattern. However, other than these, no significant spikes are present.

In this paper we describe our approach to contribute to the development and welfare of Senegal by providing visual analytics tools and techniques for the exploration and analysis of massive mobile phone data. Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces [45]. The general belief is that there is a mutual benefit in novel combinations of visualization and automated methods such as statistics, data mining and machine learning. In our approach we combine different visualization techniques in a multiple coordinated view setup tightly coupled with automated clustering and sorting methods to gain insight in massive mobile phone data. We use a visual matrix as a central element in our solution. The visual analytics methods are implemented in a prototype and applied to the provided data to enable and support users in the discovery of global and local patterns, outliers, trends, counter-trends, periodicity and anomalies. We provide a smooth user experience by adhering to design principles and requirements introduced in earlier work [47]. The insights gained in the exploration and analysis process can be used for better policy decision making.

In Section 2 we describe related work. Next, we present and discuss our visual analytics method in Section 3 and explain why we choose the visual matrix. Results and example use cases are discussed in Section 4. Finally, conclusions and extensions for future work are provided in Section 5.

2 Related Work

First, visual analytics approaches for the exploration and analysis of (massive) mobile phone data are discussed. Literature on mobile phone data analysis using methods without interaction and visualization but using only purely automated methods is plenteous, for example, [9, 19, 27, 46].

Visual analytics approaches for the extraction of spatio-temporal urban mobility information from mobile network traffic is explored by Kwan and Lee [28]. Similar visual analytics systems focusing on social aspects and historical place extraction from mobile data are explored by Andrienko et al. [6, 7] and Sagl et al. [35].

The D4D book [11] containing all submissions of the previous challenge that were selected provides a rich source on different approaches for the analysis of mobile phone data, ranging from purely automated methods to purely visual methods and any combination in between.

Next, to provide context, we discuss the related work that is closest to features that are implemented in our current prototype: we started the data exploration using our successful prototype solution of the previous D4D challenge [47]. However, we quickly found that focusing on event detection via change in call behavior was not fruitful, see Figure 2: we did not see the clear local changes in call behavior like we did in the previous challenge. This provided the first insight; Senegal was not a country in turmoil in contrast to Ivory Coast. Therefore we decided to focus more on global patterns and consequently needed a better overview visualization showing all patterns we are interested in: normal behavior, outliers, anomalies, periodicity, trends, and counter-trends. These patterns are described in more detail in the next section. 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.

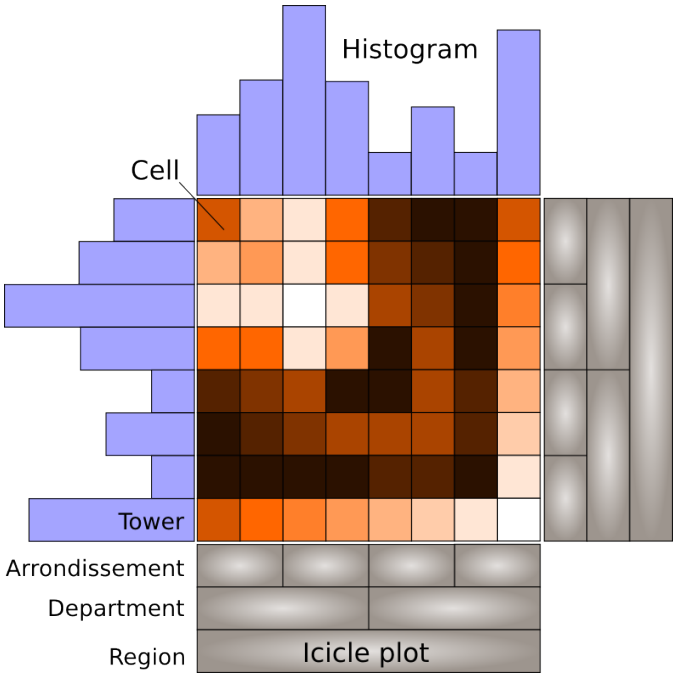


Fig. 3. Schematic representation of visual matrix with all different configurable elements; icicle plots at right and bottom for hierarchical grouping of the towers based on geography, histograms at top and left for aggregated information about the values in the cells, and cells are colored with a user 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-

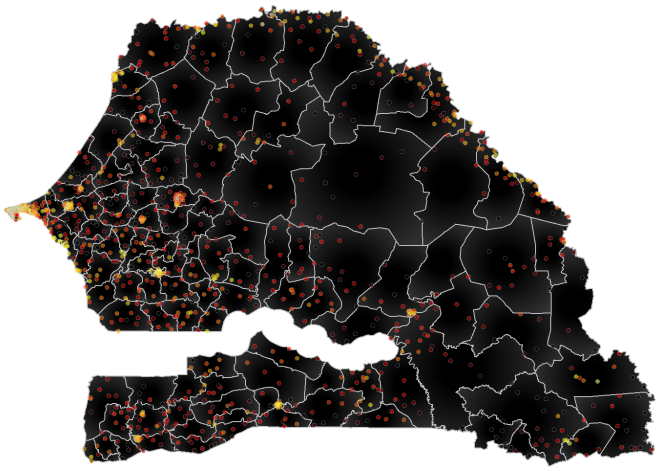


Fig. 4. Geographical View showing overall call activity. Each cell tower is represented by a dot; dots are colored according to a user selected measure. Here we use a heatmap colormap ranging from black (few calls) to red, yellow, white (many calls). Note that in urban areas, i.e., Dakar, Mbacké, Kaolack, Saint-Louis, and Ziguinchor communication is higher (many yellow/white dots) compared to rural areas.

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, showing the relative outgoing and incoming activity 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 dendogram 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).

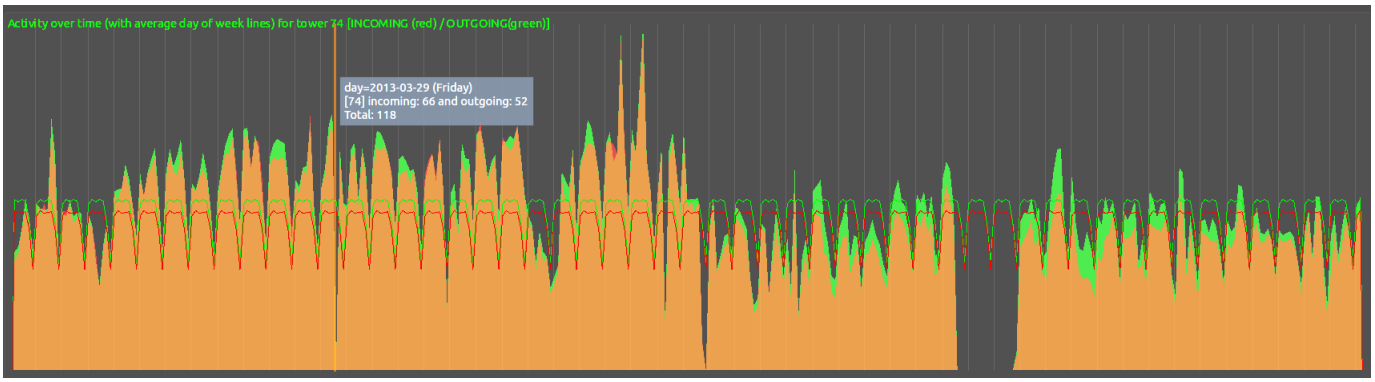
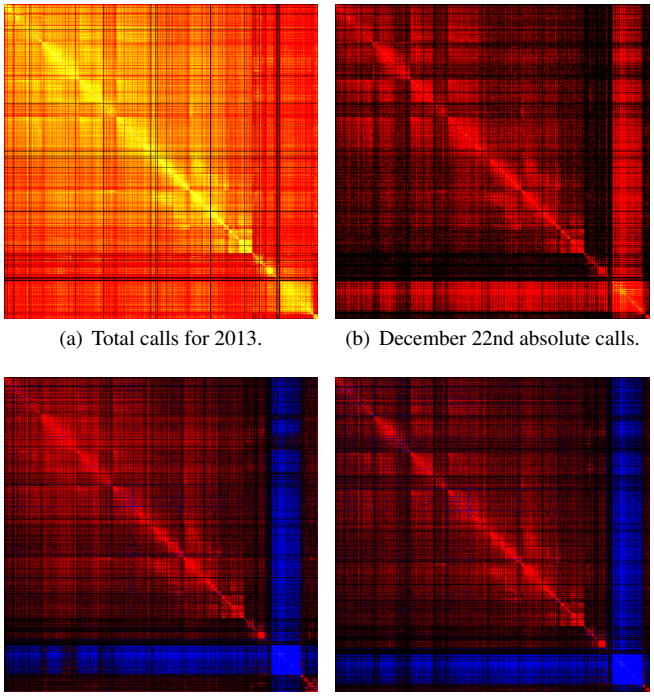


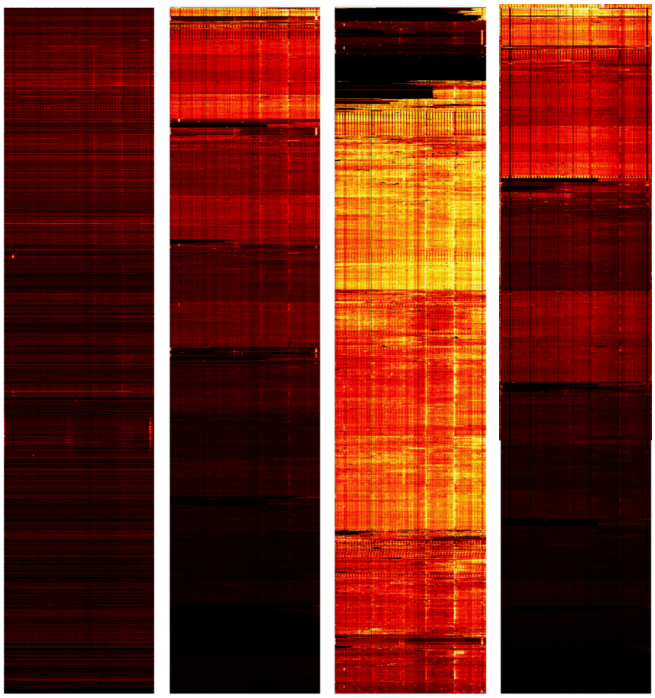
Fig. 5. Incoming and outgoing activity for a tower for the entire year as red and green area charts respectively. Overlapping areas will show up as yellow. Also visible are the average day of week activity as a green (outgoing) and red (incoming) line showing a clear week-weekend pattern. There are a number of spikes with heightened communication and also a gap with no communication at all.



(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.

Fig. 6. Zoomed section of matrix of Touba and Dakar using various display modes. (a) absolute number of calls in Touba and Dakar for 2013 using a black-red-yellow-white colormap. (b) absolute number of calls on December 22. (c) Difference in number of calls on December 22 (Grand Magal of Touba) compared to average using a red (below average) to blue (above average) colormap. (d) Difference in number of calls on December 22 compared to average number of calls on Mondays using a similar red-blue colormap.



(a) Default ordering. (b) Clustering. (c) Row. (d) Column.

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.

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.

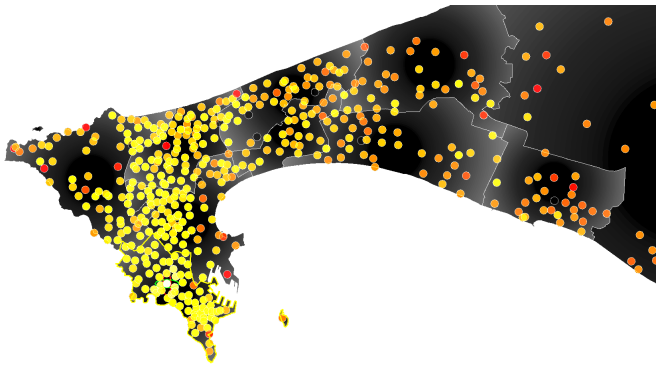


Fig. 8. Tower density in the *Dakar* region is particularly high; 489 cell towers of the total 1666 are located here.

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.

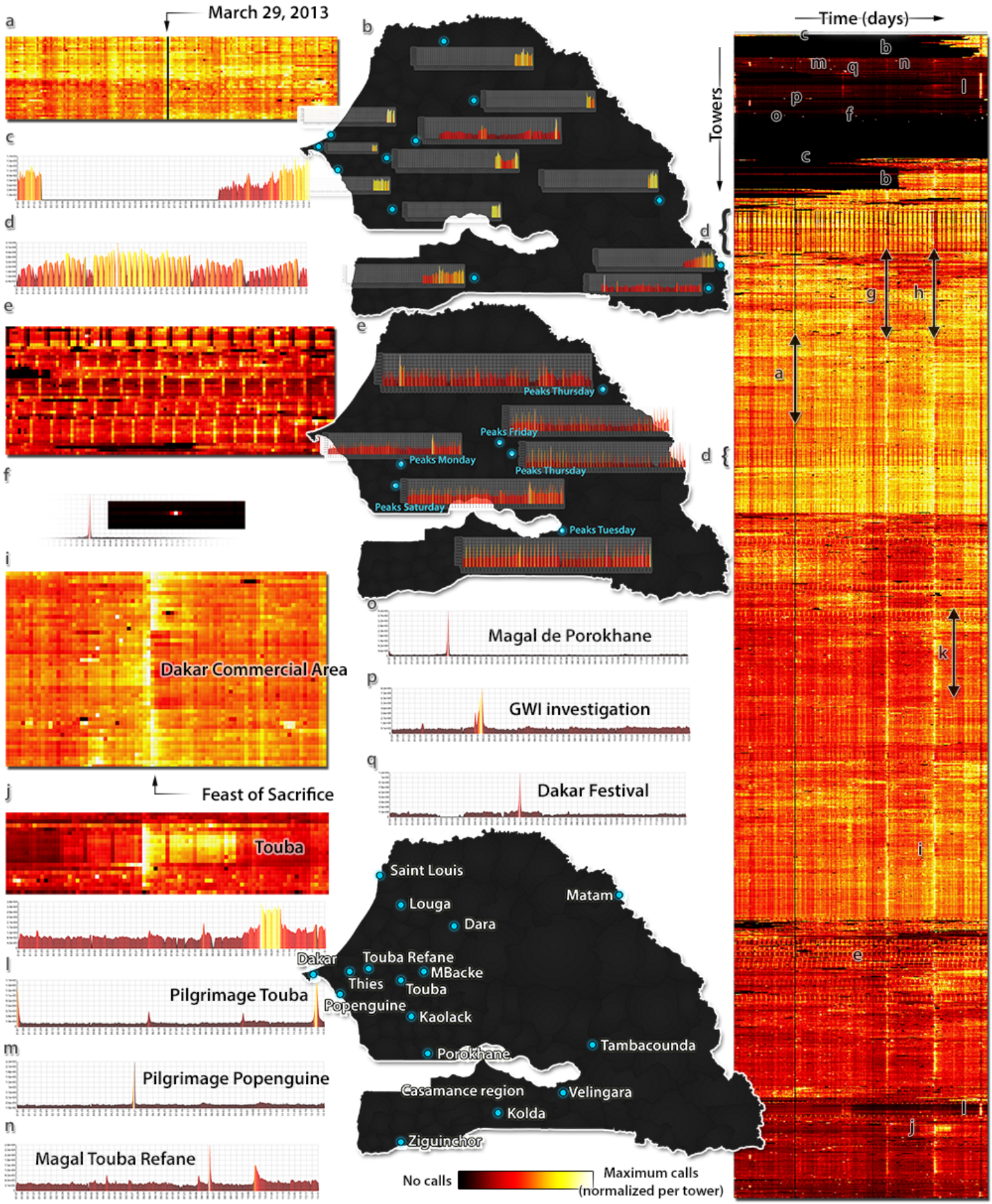


Fig. 9. Temporal matrix (right) with annotated events (left). A more detailed description is given in Sections 4.2, 4.3, 4.4 and 4.5.

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]. For the same towers, we see increased call intensity on December 25th when Christmas is celebrated, strengthening our hypothesis as we do not see this strong spike for other towers.

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*), contrasted 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 *Nioro du 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 *Senegal 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 Assainies* 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.

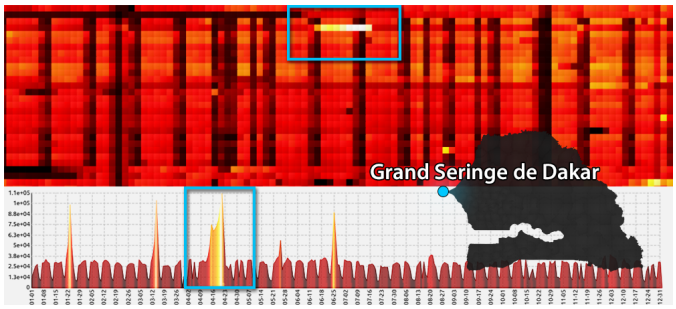


Fig. 10. Death of the old and reassignment of the new religious leader the *Grand Serigne de Dakar*. (top) pattern stands out due to breaking the week-weekend pattern; (bottom) more detailed line graph showing call intensity for the selected cell tower for the entire year. (right) linked geographical visualization that shows the location of the tower providing context.

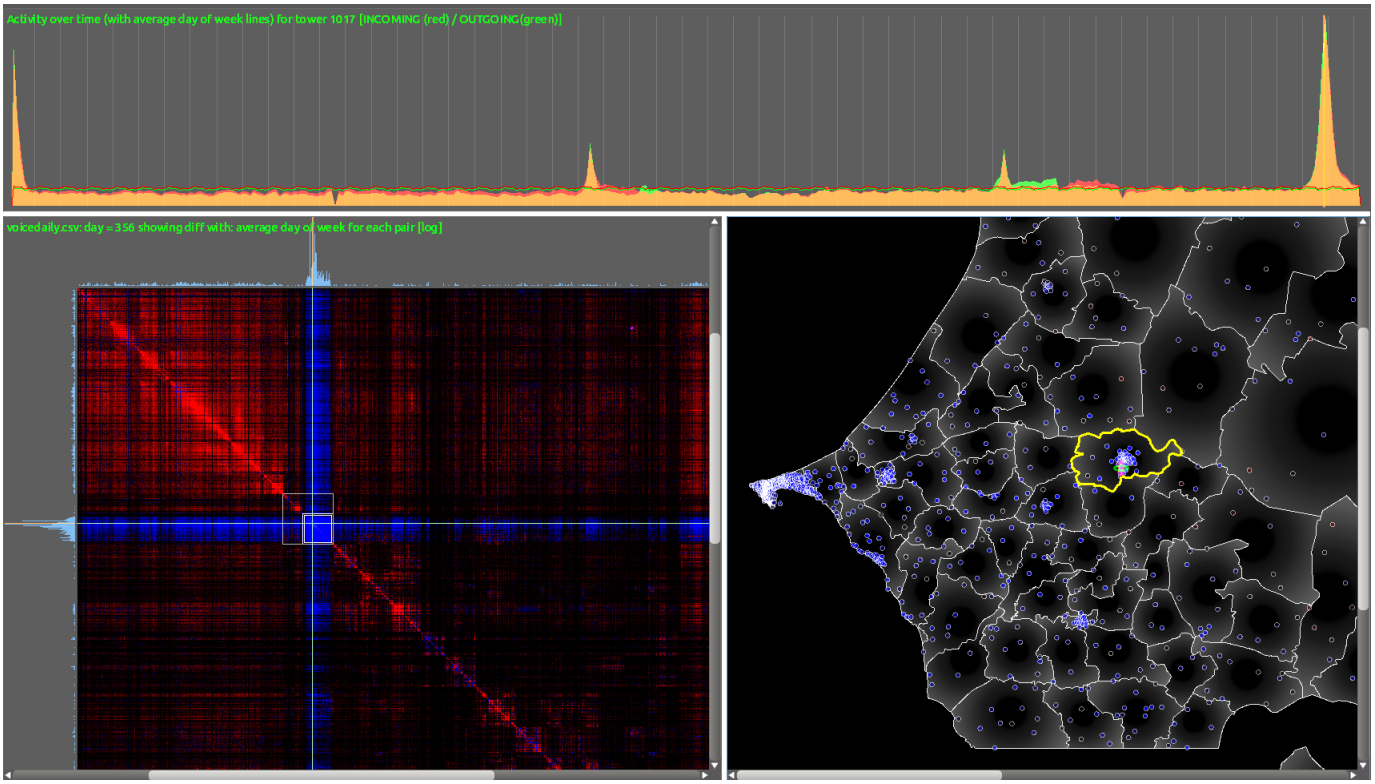


Fig. 11. Combination of linked views showing heightened activity in *Touba* around the 22nd of December compared to the average day of the week using a red (below average) to blue (above average) colormap. (bottom left) tower to tower matrix visualization; (bottom right) geographical visualization showing 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.

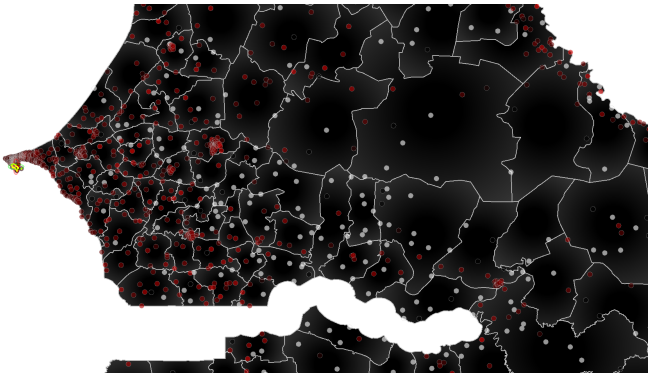
One interesting observation we have made is that the ratio between voice and text messages seems to differ for communication to rural and urban areas. Communication to and from rural areas seems to be dominated by voice traffic, while traffic in urban areas seems more balanced, see Figure 12. We suspect 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

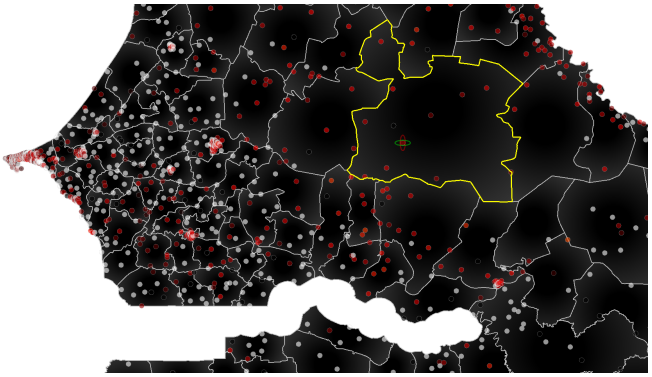
We developed a highly interactive prototype system for the exploration of massive mobile phone data in the context of the D4D challenge. As a central element and starting point of the exploration process we implemented a scalable and flexible visual matrix.

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



(a) Traffic from a typical urban tower located in *Dakar*. Here we use a black (balanced voice and text) to red, yellow, white (predominantly voice traffic) colormap. Traffic between cities is mostly dark, meaning relatively balanced, while traffic to rural areas shows up as bright white dots, meaning a high preference for voice communication.



(b) Traffic from a typical rural tower in the *Matam* region. Here most traffic, to both urban and rural areas, is white, suggesting a strong preferences for voice communications.

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.

year. Also week-weekend patterns for the identification of commercial areas stand out in the matrix as well as a global decrease of traffic on one particular day. In addition, the found patterns could be used to identify Islamic or Christian areas, due to the difference of celebrated holidays. We also found correlations with the harvesting season and weather conditions influencing call intensity such as thunderstorms. Finally, several towers that are clearly visible in the matrix show unusual behavior in the form of significant increased call intensity for a particular day. Whether this correlates with some event or the observations are anomalous data is still an open question.

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.

References

[1] Centre for Research on the Epidemiology of Disasters, EM-DAT, The International Disaster Database. http://www.emdat.be/database, 2013. [Online; accessed 2014-12-18].

[2] Senegal official public holidays. http://senegal.officialpublicholidays.com/archive-day, 2013. [Online; accessed 2014-12-19].

[3] Weather Underground. Historical Weather. http://www.wunderground.com/history/, 2013. [Online; accessed 2014-12-18].

[4] CIA World Fact Book, Senegal. https://www.cia.gov/library/publications/the-world-factbook/geos/sg.html, 2014. [Online; accessed 2014-12-20].

[5] United States Afica Command. http://www.africom.mil/, 2014. [Online; accessed 2014-12-28].

[6] G. Adrienko, N. Adrienko, M. Mladenov, M. Mock, and C. Politz. Identifying place histories from activity traces with an eye to parameter impact. *Visualization and Computer Graphics, IEEE Transactions on*, 18(5):675 –688, may 2012.

[7] G. Andrienko, N. Andrienko, M. Mladenov, M. Mock, and C. Politz. Discovering bits of place histories from people's activity traces. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 59 –66, oct. 2010.

[8] T. Barlow and P. Neville. A Comparison of 2-D Visualizations of Hierarchies. In *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*, pages 131–141, Washington, DC, USA, 2001. IEEE Computer Society.

[9] R. Baruah and P. Angelov. Evolving social network analysis: A case study on mobile phone data. In *Evolving and Adaptive Intelligent Systems (EAIS), 2012 IEEE Conference on*, pages 114 –120, may 2012.

[10] J. Bertin. *Smiologie graphique*. Mouton, Paris, 1967.

[11] V. D. Blondel, N. de Cordes, A. Decuyper, P. Deville, J. Raguenez, and Z. Smoreda. Mobile Phone Data for Development (Selected contributions to the D4D challenge. In *Proceedings of the 3rd International Conference on the Analysis of Mobile Phone Datasets (NetMob 2013)*, Cambridge, MA, May 2013.

[12] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: the D4D challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.

[13] D. Borland and R. M. T. II. Rainbow Color Map (Still) Considered Harmful. *IEEE Computer Graphics and Applications*, 27(2):14–17, 2007.

[14] W. Brinton. *Graphic Methods for Presenting Facts*. New York: The Engineering Magazine Company, 1914.

[15] A. Buja, J. A. McDonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. In *Proceedings of the 2nd Conference on Visualization*, pages 156–163, Los Alamitos, CA, USA, 1991. IEEE Computer Society Press.

[16] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction*

to Algorithms. The MIT Press, 2nd edition, 2001.

[17] Y. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. D4d-senegal: The second mobile phone data for development challenge. *CoRR*, abs/1407.4885, 2014.

[18] A. de Presse Senegalaise. 6eme édition du Festival culturel, artistique et sportif pour enfants les 8 et 9 juin. http://www.sen360.com/actualite/6eme-edition-du-festival-cultural-et-sportifs-pour-enfants-les-8-et-9-uin-42522.html, 2013. [Online; accessed 2014-12-20].

[19] N. Eagle and A. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63:1057–1066, May 2009.

[20] Food and Aggriculture Organization of the United Nations. Senegal Crop Calendar. http://www.fao.org/giews/countrybrief/country.jsp?code=SEN, 2014. [Online; accessed 2014-12-20].

[21] M. Ghoniem, J.-D. Fekete, and P. Castagliola. On the readability of graphs using node-link and matrix-based representations: A controlled experiment and statistical analysis. *Information Visualization*, 4(2):114–135, July 2005.

[22] I. Hathie, L. Diouf, M. Diouf, and M. Kama. Recherche-action sur les moyens de subsistance des agriculteurs et les options d'intervention de Global Water Initiative. http://www.inter-reseaux.org/IMG/pdf/GWI_Rapport_final_Senegal_draft_atelier_GWI.pdf, 2013. [Online; accessed 2014-12-21].

[23] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall Advanced Reference Series. Prentice Hall PTR, 1988.

[24] S. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[25] D. Keim. Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):1–8, Jan 2002.

[26] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, June 1938.

[27] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.

[28] M.-p. Kwan and J. Lee. Geovisualization of Human Activity Patterns Using 3D GIS : A Time-Geographic Approach In Michael F . Goodchild and Donald G . Janelle . Eds . 2003 . *Spatially integrated social science*, 27:48–66, 2003.

[29] Lepays-Senegal. TIVAOUANE : La Ziarra Generale 2013 sera célébrée ce Dimanche 10 Mars. http://www.lepays-senegal.com/sn/index.php/politique/item/1605-tivaouane-la-ziarra-generale-2013-sera-celebree-ce-dimanche-10-mars, 2013. [Online; accessed 2014-12-20].

[30] I. Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining*, 3(2):70–91, 2010.

[31] T. Loua. *Atlas statistique de la population de Paris*. J. Dejey & cie, 1873.

[32] L. Lugo. *Tolerance and Tension: Islam and Christianity in Sub-Saharan Africa*. Pew Research Center, 2010.

[33] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[34] K. Pearson. Notes on the History of Correlation. *Biometrika*, 13(1):25–45, Oct. 1920.

[35] G. Sagl, M. Loidl, and E. Beinat. A Visual Analytics Approach for Extracting Spatio-Temporal Urban Mobility Information from Mobile Network Traffic. *ISPRS International Journal of Geo-Information*, 1(3):256–271, 2012.

[36] SAP. Sap hana. http://saphana.com, 2013. [Online; accessed 2014-12-18].

[37] Senegalese News Agency. Matam: l'érosion fluviale suscite l'inquiétude dans certains villages. http://www.aps.sn/newsedit/spip.php?article111560, 2013. [Online; accessed 2014-12-20].

[38] SeneWeb. Derniere minute : Abdoulaye Makhtar Diop, nouveau Grand Serigne de Dakar. http://www.seneweb.com/news/Societe/derniere-minute-abdoulaye-makhtar-diop-nouveau-grand-seringe-de-dakar_n_93554.html, 2013. [Online; accessed 2014-12-20].

[39] SeneWeb. Grand magal de touba en décembre 2013 : Démarrage des préparatifs. http://www.seneweb.com/news/Societe/grand-magal-de-touba-en-decembre-2013-demarrage-des-preparatifs_n_105487.html, 2013. [Online; accessed 2014-12-15].

[40] SeneWeb. MAGAL DE POROKHANE CE JEUDI : Sokhna Diarra Bousso, la voisine de Dieu. http://www.seneweb.com/news/Societe/magal-de-porokhane-ce-jeudi-sokhna-diarra-bousso-la-voisine-de-dieu_n_90650.html, 2013. [Online; accessed 2014-12-15].

[41] SeneWeb. Seneweb popenguine 2013. http://www.seneweb.com/news/Popenguine-2013, 2013. [Online; accessed 2014-12-18].

[42] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–346, Washington, DC, USA, 1996. IEEE Computer Society.

[43] C. Spearman. The proof and measurement of association between two things. By C. Spearman, 1904. *The American journal of psychology*, 100(3-4):441–471, 1987.

[44] Sud Online. Bambey, Magal de Touba Refane: Le bitumage de la route, la préoccupation du Comité d'organisation. http://www.sudonline.sn/index.php/articles-images/vignettes/articles-images/crop/NOURISSONS_w300_h280/le-bitumage-de-la-route-la-preoccupation-du-comite-d-organisation_a_15187.html, 2013. [Online; accessed 2014-12-20].

[45] J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005.

[46] V. Traag, A. Browet, F. Calabrese, and F. Morlot. Social event detection in massive mobile phone data using probabilistic location inference. In *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*, pages 625 –628, oct. 2011.

[47] S. van den Elzen, J. Blaas, D. Holten, J.-K. Buenen, J. J. van Wijk, R. Spousta, A. Miao, S. Sala, and S. Chan. Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics approach. In *Proceedings of the 3rd International Conference on the Analysis of Mobile Phone Datasets (NetMob 2013)*, Cambridge, MA, May 2013.

[48] Wikipedia. August 15 — Holidays and Observances, 2004. [Online; accessed 2014-12-20].

[49] L. Wilkinson and M. Friendly. The History of the Cluster Heat Map. *The American Statistician*, 63(2):179184, 2009.