2013

Detecting Mobility Patterns in Mobile Phone Data from the Ivory Coast

Matthew Dixon
University of San Francisco, mfdixon@usfca.edu

Spencer P. Aiello

Funmi Fapohunda

William Goldstein

Detecting Mobility Patterns in Mobile Phone Data from the Ivory Coast

Matthew F. Dixon, Spencer P. Aiello, Funmi Fapohunda, William Goldstein

Graduate Program in Analytics
University of San Francisco
2130, Fulton Street, San Francisco, CA 94117

This paper investigates the Data for Development (D4D) challenge [3], an open challenge set by the French mobile phone company, Orange, who have provided anonymized records of their customers in the Ivory Coast. This data spans a 5 month (150 day) horizon spread across 4 different sets containing antenna-to-antenna traffic, trace data for 50,000 customers at varying spatial resolution, and social graphs for 5,000 customers. By leveraging cloud-based and open-source analytics infrastructure to (1) merge the D4D datasets with Geographic Information System (GIS) data and (2) apply data mining algorithms, this paper presents a number of techniques for detecting mobility patterns of Orange customers in the Ivory Coast.

By applying a k-medoid clustering algorithm to the antenna locations and their average distance to nearby antennas, we show how the high spatial resolution mobile phone dataset reveals a number of daily mobility patterns and properties, including trends in week-day versus weekend, public holiday mobility behavior, and distributional properties of daily trip distances across each cluster. With a view towards providing tools to assist with transport infrastructure planning, we combine the high spatial resolution D4D dataset with GIS data for transport infrastructure and demonstrate an approach for detecting whether a mobile phone user is traveling on a segment of transport infrastructure. This work culminates in a preliminary cloud-based GIS tool for visualizing mobility traces.

1 Introduction

The Ivory Coast is a developing West African country renound for their cocoa and coffee. It faces disease outbreaks at its West, North, and coastal borders, suffers drought caused by the seasonal Harmattan winds, and is divided north-south by ethnic and religious tensions, occasionally encountering out-bursts of civil war [5]. The Ivory Coast is an important subject of study for economists attempting to quantify reactions to Harmattan winds and civil war outburst, for epidemiologists' models of human migration to guide the dispursing of resources for disease eradication, and for sociologists considering all aspects of social policy.

The emergence of vast quantities of communication and movement data generated by mobile phone customers has the potential to positively impact urban planning, enconomic development, and public health decisions [4]. While the validity of such data to answer broad questions of human behavior remains dubious [2], it is possible to harvest meaningful insights from mobile data, even if it only spans a 5-month horizon [1].

One promising area of application in particular is the use of high spatial resolution mobile phone data to study mobility patterns for the ultimate purpose of mitigating congestion in urban roads, urban planning, traffic prediction and the study of complex networks. Such studies are conventionally based on primitive travel surveys and typically fail to support transport planners with information needed to design future road networks able to withstand modern mobility demand [9]. In fact, across the world, our understanding of the origins and destinations of commuters through a particular segment of road or railline remains for the most part poorly understood and unquantified.

Recently Wang, Hunter, Bayen, Schechtner Gonzlez [9] presented an approach for understanding road usage patterns in urban areas through the integration of mobile phone data and GIS data. This approach revealed hidden patterns in road usage in the San Francisco Bay and Boston areas and demonstrated a basis for more informed and cost effective transport planning and congestion mitigation. This study used three-week-long mobile phone billing records generated by 360,000 San Francisco Bay Area users (6.56% of the population, from one carrier) and 680,000 Boston Area users (19.35% of the population, from several carriers) respectively [9]. Given the high density of service towers (there are 892 antenna service areas in the San Francisco Bay Area), the authors are able to perform a detailed study of road usage patterns.

Mobility patterns Motivated by the study of Wang et al. [9], this paper sets out to characterize and quantify the daily

distances travelled by mobile phone users. In order to do this, we begin in Sections 2 and 3 with a description of the data from which it becomes apparent that the spatial resolution of the antennas is not only much lower than in the study by Wang et al. [9], but is considerably more varied depending on proximity to large cities and its region. Variable resolution challenges the interpretation of mobility distance studies and in Section 4 we turn to k-medoid clustering to partition the antennas into artificial regions in which antennas density are more uniformly distributed. By partitioning the antennas into a small number of clusters, we proceed to characterize the distribution and time history of daily trip distances and identify trends in mobility behavior and anomalies which are unencountered for by religious events or national holidays.

Route detection An additional challenge to interpretating the mobility patterns is that commuters clearly do not follow the shortest path through a sequence of antennas, but travel on road and railroads whose routes may be significantly different to the mobility trace. Section 5 demonstrates a simple methodology for detecting whether a user has travelled on a segment of transport infrastructure. This methodology is based on integrating the D4D datasets with Geographic Information System (GIS) data and open-source software infrastructure. Using this methodology, we show the full mobility patterns of users who use the railroad on a particular day and reveal their source and end destinations. Such insight provides a basis for better informed transportation planning, including targeted strategies to mitigate congestion.

Visualization tool An effective and low-cost mechanism for transferring analytics research to application domains such as transport planning, is to provide an open source web-based prototype visualization tool which enables the user to study individual user trajectories and observe how they interact with the transport infrastructure. Section 6 describes the infrastructure used to create this tool and provides a URL which the reader can use to access this prototype visualization tool. The site is password protected and reader should contact the corresponding author for login credentials.

2 Data Description

The data, spanning December 2011 to April 2012, provided by Orange has been released to research teams in order to examine developmental questions in new ways [3].

There are four sets of data provided for the D4D project:
SET1: Antenna-to-antenna, number of calls as well as the duration of calls between any pair of antennas aggregated hour by hour. The antenna positions are given by longitude/latitude pairs.
SET2: Individual Trajectories, High Spatial Resolution Data movement trajectories for 50,000 users. The sub-prefectures are given by longitude/latitude pairs.
SET3: Individual Trajectories, Low Spatial Resolution Data movement trajectories for 50,000 users over the entire observation period at lower spatial resolution (phone calls aggregated by prefecture, rather than antenna position). The sub-prefectures are given by longitude/latitude pairs.
SET4: Communication Subgraphs, communication subgraphs for 50,000 randomly selected individuals

In this paper we use the first two datasets for detecting daily mobility patterns. We additionally use country and administrative subdivision outlines provided by GDAM [7] (an open geospatial datasource), and road and railroad vectors from the Digital Chart of the World, both of which can be accessed through the open GIS software and data content provider DIVA-GIS [6].

3 Route Visualization

In order to gain some initial insight into the mobility patterns of users, it is useful to visualize their mobility traces over the course of a day. Appendix A provides the details of how individual mobility traces are aggregated from the call data. Figure 1 shows the antennas (black dots), the major cities (black circles) with a population of at least 1M and the mobility trace of all users (red translucent lines) travelling on the 6th of December 2011, the second day of the D4D dataset. The department boundaries are also shown on the map of the Ivory Coast. Figure 2 shows the Ivory Coast road network and by comparing this figure with Figure 1, we are able to explain many of the features in the mobility traces. Hubs of mobility activity are observed to coincide with major city locations and the most intense mobility activity is in the Abidjan area in the south of the Ivory Coast. The rest of the southern half of the country has a fairly even distribution, while the western and northern areas of the country lag behind in terms of mobile capacity. This distribution of antenna intensity is further observed to be commensurate with the population density map shown in Figure 3.

4 Daily Trip Distances

There are many challenges in using call data in SET2 to study daily mobility patterns. Aside from the obvious fact that the 50k mobile phone users included in the dataset represent a small fraction of the estimated 20M people (source: World Bank, 2011) who reside in the Ivory Coast, there are more technical challenges. There are only 1231 antennas listed in the SET2 dataset to cover a country with an area of 124,500 sq miles. Furthermore, these antennas are non-uniformly located, thus introducing variation in the minimum threshold distance that a user needs to travel in order for that trip to be detected. In Abidjan, for example, the antennas density is higher and thus there is higher fidelity in trip detection. In contrast, antennas are sparsely located over the north of the Ivory Coast and shorter trips may not be detected.

In order to study the distribution and time series of daily mobility distances using all available call data recorded in SET2 , we use a clustering algorithm to separate the antennas into a small number of distinct clusters. The purpose of creating clusters is to partially address the above concern regarding the variation in minimum threshold distance for trip

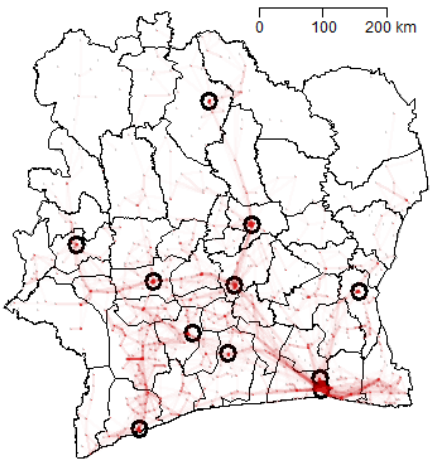

Fig. 1. A map of the Ivory coast showing the department borders, major cities with a population in excess of 1M (black circles) , antenna locations (black dots) and the mobility traces of all users on the 6th of December, 2011 (translucent red lines).

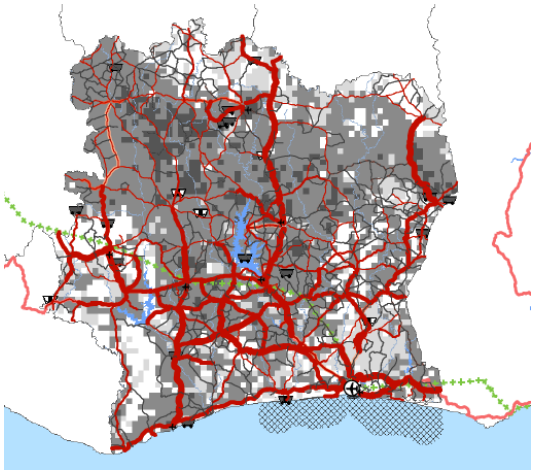

Fig. 2. The Ivory Coast Road Network. Source: AICD Interactive Infrastructure Atlas for The Ivory Coast downloadable from http://www.infrastructureafrica.org/aicd/system/files/civ_new_ALL.pdf

detection which is apparent by viewing the antenna locations in Figure 1. This direction is predicated on the notion that antenna service areas are approximately equal in radius and that the impact of terrain topology on coverage can be considered secondary. We introduce an antenna distance dispersion measure over an area A on the map. For each antenna location $p_i \in A$, we define the set of other antenna locations P_i which are in the neighborhood of p_i as

$$P_i := \{p_j \ : \ \text{dist}(p_i, p_j) \leq C_0, p_i \neq p_j\}.$$

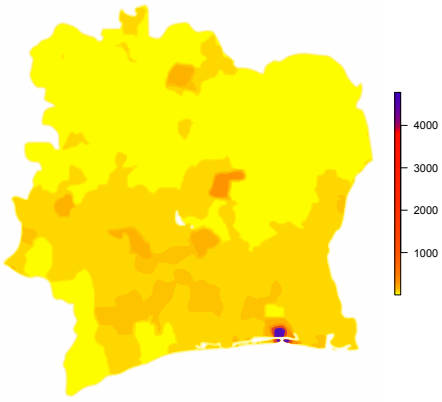

Fig. 3. A map of the population density (per sq. km) over the Ivory Coast.

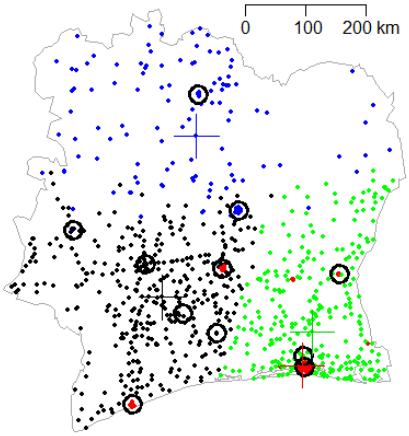

Fig. 4. This map show the clusters of antenna locations. The centroids of each cluster are shown by large colored crosses. Cities are shown with black circles. Key: Cluster 1 (black), Cluster 2 (red), Cluster 3 (green) and Cluster 4 (blue).

Defining the average distance between p_i and all other points $p_j \in P_i$

$$d_i = \frac{1}{|P_i|} \sum_{p_j \in P_i} dist(p_i, p_j),$$

and our antenna distance dispersion measure as the standard

deviation of the average antenna distances over the area A

$$\sigma^2(D) = \frac{1}{|D|-1} \sum_{d_i \in D} (d_i - \mu(D))^2,$$

where D denotes the set of d_i corresponding to all points p_i and $\mu(D)$ is the mean of D.

Clusters are formed using the Partitioning Around Medoids (PAM) algorithm [8] applied to the standarized antenna longitude and latitude co-ordinates, **and** the standardized average distances between antennas D. C_0 is chosen to be 25km which is found by trial and error to yield a sufficiently high number of antennas with at least one other nearby antenna, while still resolving urban areas of dense antenna locations. A is chosen to cover the entire region of the Ivory Coast. This choice of features for clustering is based on a competing desire to preseve the contiguous geographic regions on the map but also partition the dataset by the average distance between antennas.

PAM is a type of k-medoids algorithm which attempts to minimize the distance between points labeled to be in one of k clusters and a point designated as the center of each cluster. k was set arbitrarily to 4 so that the granularity of the mobility study is coarser than a regional study but sufficiently granular to reduce dispersion in the average distance between antennas. Table 1 shows the details of the cluster properties - the number of points in the cluster (size) and the first two moments of the average distance between antennas over the cluster. The color codes listed in the table correspond to the colors of labelled antennas shown in Figure 4. The bottom row shows the moments of the average distance between antennas over the entire dataset, without clustering.

We note in particular that Cluster 2 (shown in red) represents some of the most significant dense urban areas, including Abidjan, and is characterized by more dense and uniformly distributed antenna locations. Clusters 1 and 3 have similar density and uniformity characteristics, with the average of antenna distances being higher than the national average and the level of dispersion being lower. Hence, Clusters 1-3 are observed to exhibit more uniform antenna locations. Cluster 4, representing the north region, on the other hand exhibits a smaller set of antennas and a high level of dispersion. For this reason, we approach the measurement of daily commute distances of users associated with Cluster 4 with more caution.

Data preparation The daily distances travelled by users can be associated with each cluster by the location of the antenna which they are most frequently closest to over the two-week period. This of course is not a reliable indicator of where a user resides, but in most cases it serves as a starting pointing for approximating a user's central place of calling activity and hence from in which most trip distance estimates are made. In the proceeding analysis below, it is important to note that we first applied two filters: (1) exclusion of users who do not travel on any day over two week period and (2) exclusion of daily distances for a user if at any point during the day the nearest antenna is listed as '-1'. The motivation for the first filter is to focus on the mobile cohort of customers in each cluster and the effect is to reduce the number of users included in the study by approximately 30%. The second filter removes daily distances which may be flawed and the effect is to remove a further 10% of all daily distances per user logged.

cluster	color	size	μ	σ
1	black	380	16.15	2.85
2	red	384	9.06	1.59
3	green	320	16.35	2.85
4	blue	147	15.22	5.54
Orig.		1231	13.84	4.41

Table 1. The characteristics of each cluster are shown for comparison with the original dataset (bottom row).

Times series of user mobility The top left graph in Figures 5 and 6 shows the historical time series of the average of the daily commute distances over each cluster between December the 5th, 2011 and February the 15th, 2012, and between February the 16th and April the 22nd, 2012 respectively. The splitting of the time series into two components is purely to improve the readibility of the graphs. The start and end of weekends are shown in the time series plot as two vertical gray lines. The bottom axis shows the monthly, weekly and daily markers, and the top axis shows the periods representing each two-week period in the dataset (every other marker indicates the start of each two-week period).

Weekends On first glance, it would seem apparent that the average daily distance travelled by users is generally lower at the weekend. However, this is a misleading interpretation of the mobility traces. We see in the top right graph in Figures 5 and 6 the corresponding ratio of 'undetected commuters' on any given day to the cluster cohort size. Undetected commuters are any combination of the following: users who either do not travel, are confined to the service area of one antenna, or make less than two calls on that day. We observe that the least sparse cluster (Cluster 2), representing dense urban areas, generally exhibits the lowest ratio of undetected commuters on any given day. In any two week period, we further observe that the Cluster 2 ratio of undetected commuters generates fluctuates the least between the period up until the 3rd weekend in March, after which point, ratios fluctuate considerably.

Although the user cohort is constant over any two-week period, the average number of calls that a user makes in a day (shown in the bottom left graph of Figures 5 and 6) varies over time. We observe a general trend of call volumes per user being lower at weekends, although this is less pronounced than in the ratio of undetected commuters. In each cluster, average call volumes are found to be correlated with average daily distances, $\rho = (0.42, 0.76, 0.67, 0.58)$.

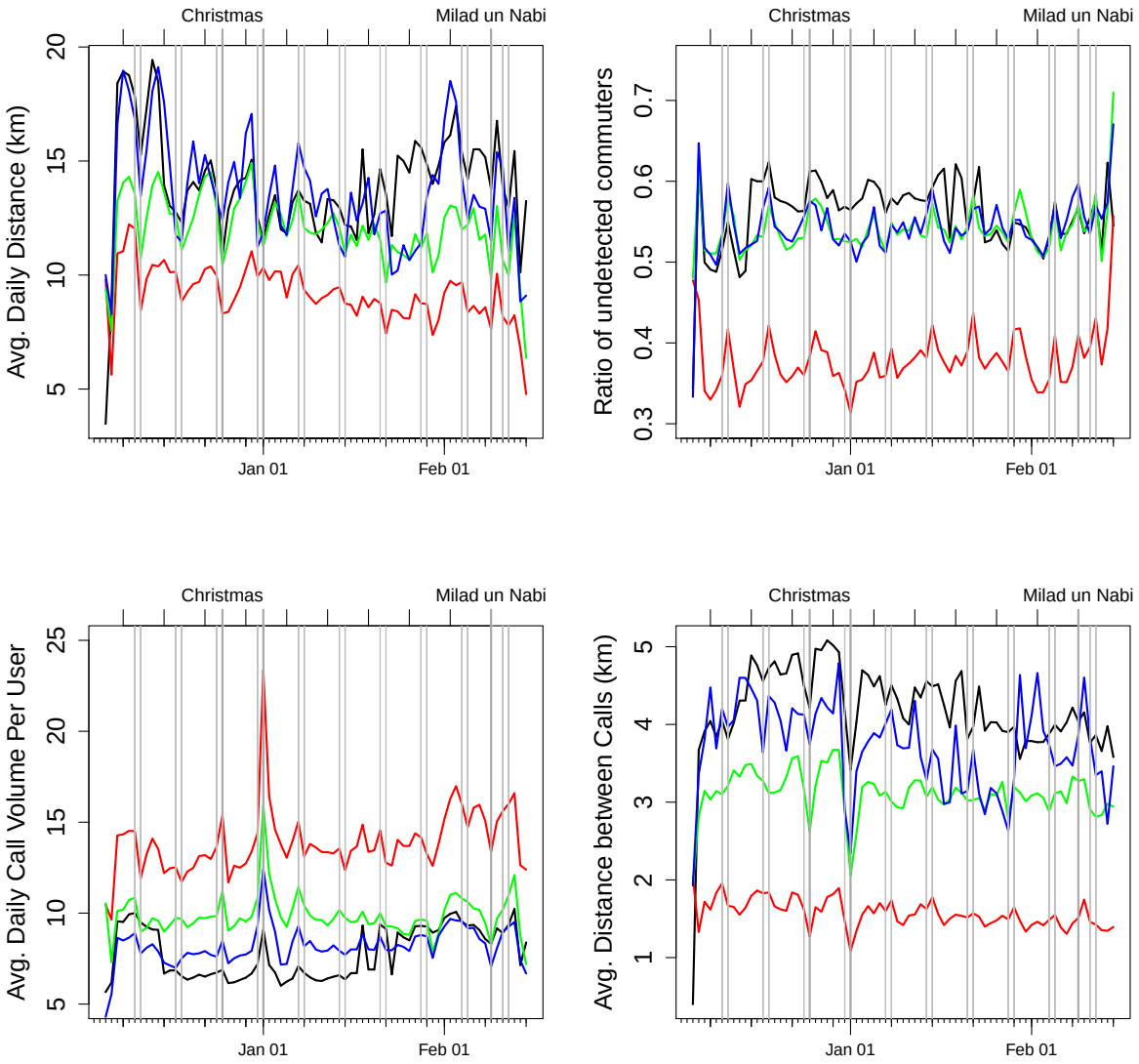

Fig. 5. This figure shows the historical time series of the daily average distance (top left), the ratio of undetected commuters (top right), the average daily call volume per user (bottom left) and the average distance between calls (bottom right) over the period of December the 5th, 2011 and February the 15th, 2012. The start and end of weekends are shown in the time series plot as two vertical gray lines. The bottom axis shows the monthly, weekly and daily markers, and the top axis shows the elapsed weeks since the date of the first call in the entire dataset (every other marker indicates the start of each two-week period). Key: Cluster 1 (black), Cluster 2 (red), Cluster 3 (green) and Cluster 4 (blue).

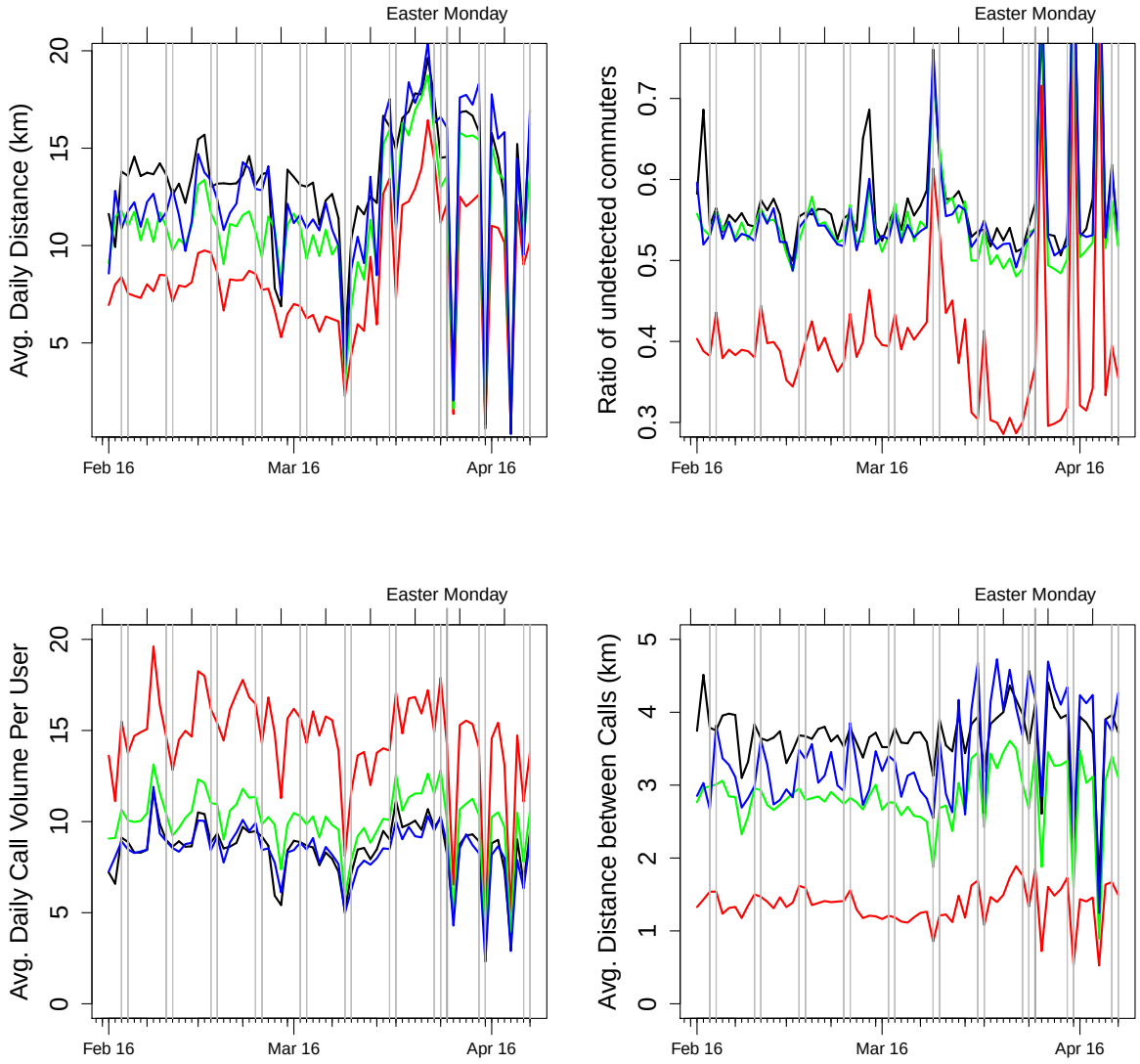

Fig. 6. This figure shows the historical time series of the daily average distance (top left), the ratio of undetected commuters (top right), the average daily call volume per user (bottom left) and the average distance between calls (bottom right) over the period of February the 16th, 2012 through to April the 22nd, 2012. The start and end of weekends are shown in the time series plot as two vertical gray lines. The bottom axis shows the monthly, weekly and daily markers, and the top axis shows the elapsed weeks since the date of the first call in the entire dataset (every other marker indicates the start of each two-week period). Key: Cluster 1 (black), Cluster 2 (red), Cluster 3 (green) and Cluster 4 (blue).

In contrast, average call volumes are found to be negatively correlated with the ratio of undetected commuters with $\rho = (-0.71, -0.69, -0.77, -0.65)$ for each cluster.

We further note in passing that users in Cluster 2 (dense urban areas) make more calls per day, and one might speculate that this is a function of better mobile phone network coverage perhaps. The challenge that we thus encounter in interpreting whether users, in general, actually travel less at weekends is obscured by the drop in call volumes at weekends and a corresponding increase in the ratio of undetected commuters. In other words, if users actually travel less at weekends, as the average distance time series suggests, then we would need to look to another measure in order to confirm this.

In the bottom right graph of Figures 5 and 6), we normalize the daily distance travelled with the number of calls that the user makes to yield a measure of the average distance travelled between calls. We posit that this normalized distance is a preferable indicator of daily trip distances since it attempts to counteract the effect of call volume variation on the distance estimates. This normalized distance is much lower in Cluster 2 as we would expect for a dense urban environment. There is some indication of increased average distance travelled between calls at the weekends in Cluster 2 but the effect is less pronounced compared to the average daily distances. To reiterate, the suspected causation is that the drop in average distances travelled on weekend days is due to the higher number of undetected commuters because call volumes are lower.

Holidays We do observe a notable drop in the distance per call on Christmas and New Year's day in each cluster. We note a surge in call volume on New Year's day, likely causing a drop in the ratio of undetected commuters and a misleading increase in the detected average distance travelled. We further observe a mid-week drop in average call volume across all clusters on February the 9th which coincides with 'Milad un Nabi' - the Suni celebration of the birth of Prophet Mohammed.

In the months of December 2011 and January 2012, there is evidence of regular peaked ratios at weekends (especially in Cluster 2), with the exception of the weekend before Christmas, where the peak is observed on the day after Christmas. There is also a marked drop in the Cluster 2 ratio on January 1st which is a national holiday. Other noteworthy holidays include Easter Monday when the average call volume is observed to drop in the call volume but with no significant change in the average distance between calls. We speculate that the remaining very large drops in the average volume, such as on February the 15th and April the 10th are caused by nationwide power outages since the call volume drops significantly over all clusters.

Distribution of daily distances Figure 7 shows the histogram of daily distances travelled by users over the entire twenty week period, partitioned by cluster. The histograms exclude users who are not detected as travelling on a given day, so that the distribution represents only non-zero daily commute distances. The variance of each distribution is significantly larger than the mean, an effect referred to as 'overdispersion'. Because of this property, the distribution can not be accurately described by a Poisson distribution and we fit instead a negative binomial distribution to the daily distances travelled by each user. The distribution can be expressed as a mixture of Poisson distributions with the mean distributed as a gamma distribution with scale parameter $(1 - prob)/prob$ and shape parameter *size*. The fitted negative binomial distribution together with the parameters *size* and *prob* are shown in each plot and observed to vary between cluster. Consistent with the increased antenna density, we observe that Cluster 2 exhibits a higher proportion of daily distances in the semi-open interval $(0, 10]km$ and thus the fitted density function shows a sharper decay rate with increasing distance than for the other clusters.

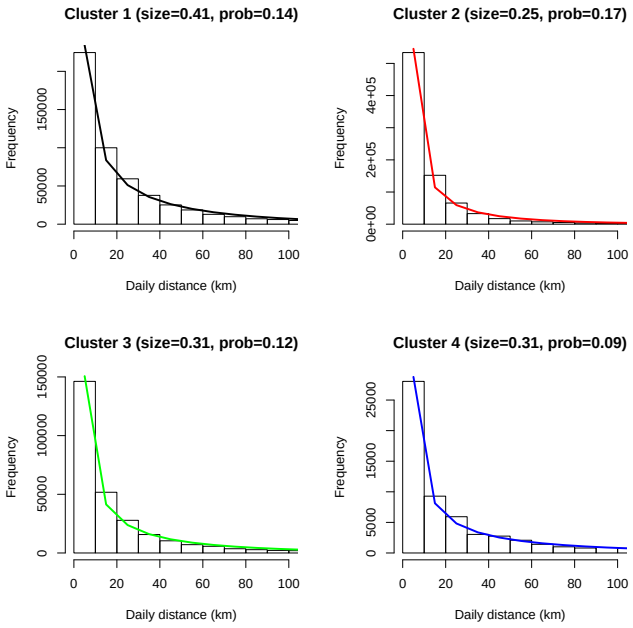

Fig. 7. Histograms of daily distances together with the fitted negative binomial distributions of daily commute distances over the entire twenty week period, partitioned by cluster.

5 Route Segment Detection

This section partially addresses the problem of how to detect that a mobile phone user is travelling on a particular segment of transport infrastructure such as a railroad or road based on the call data provided in SET2. By detecting that a mobile phone user is travelling on a particular segment of transport infrastructure, it is then possible to more accurately view their mobility traces, estimate their daily commute distances and determine the origins and destinations of their trips. Such insight may be useful in infrastructure extension projects such as new sections of railroads and new roads.

The proposed algorithm for detecting whether a user's trajectory interacts with a route segment in a particular interval consists of the following two steps:

Step 1: Associate antennas with route segments Identify the set of antennas whose minimum Euclidean distance is within a threshold of the target route segment.

Assume that a route segment S_j is a curve in a two dimensional plane. Let P_j denote the set of antenna locations $p_i \in \mathbb{R}^2$ whose minimum Euclidean distance with the j^{th} route segment S_j is within C_0 units of distance so that

$$P_j := \{p_i \ : \ \text{dist}(p_i, S_j) \leq C_0\}.$$

Denote the corresponding set of antenna ids as $A_j := (a_1, a_2, \ldots, a_m)$, where m denotes the cardinality of the set P_j. Define the point of intersection with the shortest path from p_i and the curve S_j as x_{ij}, which for ease of exposition, we shall assume is unique. Adopting a convention for determining which end of the curve is the start and end point based on its latitude and longitude, we may denote the normalized distance along the curve x_{ij} lies from the starting end point as $d_{ij} \in [0,1]$. Further denote the corresponding set of m normalized distances as D_j.

Step 2: Detect trip interaction with route segment Denote the chronologically ordered sequence of antennas id which a user calls on a given day as $A' := (a'_0, a'_1, \ldots, a'_N)$ with corresponding times $T := (t_0, t_1, \ldots, t_N)$

A user is identified as travelling on a route segment S_k at any period within the time interval $\tau := [t_1, t_2]$ if there exists at least one pair of antennas (a'_i, a'_j) where $a'_i \neq a'_j$, $t_i, t_j \in \tau$, $t_i < t_j$, $a'_i, a'_j \in A_k$, $|d_{i'k} - d_{j'k}| > \varepsilon$ and i', j' denote the local indices of a'_i, a'_j in A_k.

This algorithm is parameterized by two constants C_0 and $\varepsilon << C_0$. We demonstrate this algorithm by detecting which users travel on a segment of the main railroad which runs north-south from Adbijan. Figure 8 shows the detection of antennas within $C_0 = 10$km of the rail line. The `ST_Distance` function provided in postGIS is used to determine the set of normalized distances D_j.

We then apply Step 2 to detect which users' trajectories interact with the railroad at any point over the day using the minimum normalized threshold distance $\varepsilon = 0.1$. The sign of $d_{i'k} - d_{j'k}$, where $|d_{i'k} - d_{j'k}| > \varepsilon$, determines which direction the user is travelling. Figures 9 and 10 show the entire trip of all users who are travelling northbound or southbound on the railroad respectively.

6 Web-based Individual Trajectory Visualization Tool

In this section, we outline a visualization tool which has been developed for the purpose of enabling the community to interact with SET2 and visualize the mobility traces in combination with GIS data. The user is able specify date ranges, user IDs, and toggle topographical and infrastructural elements. Our tool dynamically loads the D4D data in real time, and allows granular playback of call events, exposing commute and travel patterns.

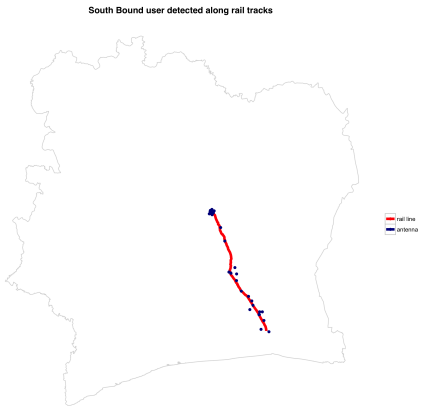

Fig. 8. This figure shows the antennas which have been detected as being located less than 10km from the illustrated railroad segment.

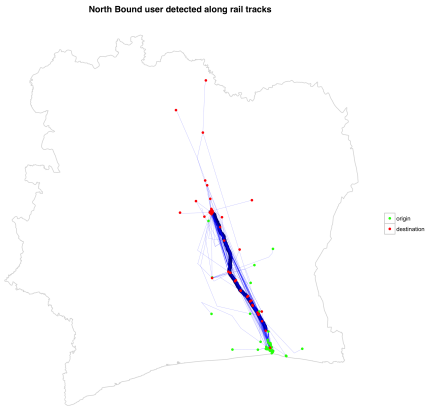

Fig. 9. This figure shows the mobility traces of all users who travel north on the segment of the railroad on a particular day.

There are three ways in which we've made the data more accessible for user interaction. First, we allow for several static toggles that display topographic and infrastructural information for roads, railways, water areas, rivers and antenna locations. Second, we enable the tool user to choose between a more narrow study of particular individual customer trajectories or a more exploratory study of customer samples. By inputting a user ID and Week ID, a user's mobility trace for the 2 week period is plotted on the map. For more open-ended data exploration, we've included the ability to selectively plot all data for a given day or randomly plot customer data for a particular or random date. Lastly, we allow for trajectories to be explored by cluster using the k-medoids algorithm described in Section 4. For a chosen cluster and two-week time period, we plot all of the customer trajectories associated with the cluster and allow for toggling of trajectories by cluster. The tool can be accessed at http://67.202.19.246/proj/finw.html by requesting login credentials from the corresponding author. Further details of the analytics infrastructure are included in Section B.

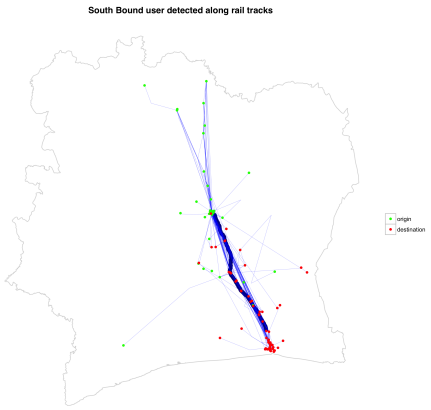

Fig. 10. This figure shows the mobility traces of all users who travel south on the segment of the railroad on a particular day.

7 Conclusion

This paper has identified a number of challenges in detecting mobility patterns from individual trajectories defined by antenna locations across the Ivory Coast. One such challenge is the highly variable antenna density which somewhat opaquely impairs the estimate of daily commute distances. Our approach partially addresses this issue by using a k-medoids algorithm to attempt to cluster the antennas into a small number of artificial geographic regions in which the antennas are more evenly distributed. By further associating user cohorts with each cluster, we are able to study the temporal and distributional characteristics of daily mobility traces per cohort.

We find within each cluster that average daily call volumes are anti-correlated with the number of detected commuters and, in turn, correlated with the average daily distance measured by mobility traces. Because of a suspected undesirable causational effect between call volume and distance measured, we measure the average distance travelled between calls and find evidence of changes in mobility patterns around national holidays and religious events. We also find possible evidence of power-outages through very large decreases in daily call volumes which yield spurious mobility traces on these days.

Another central challenge is how to detect whether a mobile phone user is travelling on a particular segment of transport infrastructure such as a road or railroad. Detection of routes travelled along transport infrastructure not only leads to improved trip distance estimates but ultimately serves to inform transport planners about the origin and destination of users who use such a segment. We demonstrate a simple methodology for transport segment detection which uses postGIS and postgres applied to the D4D datasets. We further introduce a preliminary cloud-based GIS tool for visualizing user trajectories and it is the subject of future work to enable the visualization tool to fully automate route detection.

Acknowledgements

We are grateful to Professor Terence Parr for setting up the Amazon EC2 account and Deron Aucoin for his contributions to the earlier stages of this project.

A Data Manipulation in Postgres

Data Aggregation: Data was aggregated over SET1 and SET 2 to provide daily summaries per user and per antenna respectively. An example of how the daily trip distances for each user were aggregated is provided below.

Distance Travelled Calculation: A daily user aggregate view ("mobility trace") of SET2 was created. This view has a column with the aggregated list of time and antenna position aggregated for each day.

Example

UserID: 6836

Connection Date: $2011-12-16$

Aggregated Antenna List: $06:51-619, 06:55-596, 08:45-596, 13:21-619, 16:54-619$

Computed Distance: $26.175km$

B Analytics infrastructure

Using an EC2 instance in the Amazon cloud, we built a server-side data analytics stack consisting of Apache 2.4.3, postGres 9.1 and PHP 5.4.11. A combination of html and d3.js is used to create a web-based GIS visualization tool, which is run and tested in the Chrome browser (version 24.0.1312.57 m).

The visualization tool provides user-selected parameters to a php script, which connects to and queries the postGres database. The query returns rows consisting of user IDs, the connection date-time, and the user's longitude-latitude pair. This php script packages the query results into an array to be exported for processing in javascript.

Figures and animations are created in D3js (data-driven documents), where geographic coordinates are projected into SVG data. D3 requires the data to be in geoJSON, which is a JSON-style formatting that allows for Point features (an array with a longitude-latitude pair) and LineString features (an array of longitude-latitude pairs) among others. We reformat the raw longiture-latitude pairs for each user ID into the geoJSON format and combine all of the users together into a single feature collection.

Within the javascript, we iterate through the user IDs and create a feature collection for each date requested by the user. We store each user ID in a javascript object, and we hash the user ID for later reference. Each feature collection consists of a feature with either a LineString or Point Geometry type, as specified by the geoJSON format (see http://www.geojson.org/geojson-spec.html). For each new user ID we encounter, we create the geoJSON point feature, add it to the feature collection and store its position in the feature collection in the hashmap. If we encounter the same user ID again, we access the feature collection using the ID as the

hash key, convert the point feature to a LineString feature and add the new longitude-latitude pairs in the LineString array.

References

[1] L. Akoglu and C. Faloutsos, *Event detection in time series of mobile communication graphs*, in Proc. of Army Science Conference, 2010, pp. 1–8.

[2] D. Berry, *The computational turn: Thinking about the digital humanities*, Culture Machine (12), 2011, pp. 1–22.

[3] V. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda and C. Ziemlicki, *Data for Development: The D4D Orange Challenge on Mobile Phone Data*, 2012, pp. 1–10.

[4] J.E. Blumenstock, D. Gillick and N. Eagle. *Whose Calling? Demographics of Mobile Phone Use in Rwanda*, Association for the Advancement of Artificial Intelligence, 2010, pp. 1–2.

[5] A.L. Dabalen and S. Paul, *Estimating the Causal Effects of Conflict on Education in the Ivory Coast*, Policy Research Working Paper, 2012, pp. 1–31.

[6] Diva-GIS, *http://www.diva-gis.org/gdata*

[7] Global Administration Areas, *http://www.gadm.org/*.

[8] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd ed., 2006.

[9] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner and M.C. Gonzalez, *Understanding Road Usage Patterns in Urban Areas*, Sci. Rep. (2), No. 1001, 2012.