Studying Intercity Travels and Traffic Using Cellular Network Data

Wei Wu, Eng Yeow Cheu, Yuzhang Feng
Duy Ngan Le, Ghim Eng Yap, Xiaoli Li

Data Analytics Department
Institute for Infocomm Research, A*STAR, Singapore

Abstract—We use anonymized mobile phone CDR (call detail record) dataset provided by Orange D4D and crowdsourced OpenStreetMap data to study people’s intercity travels and traffic speed on intercity highways in Ivory Coast. Knowledge about people’s intercity travel behaviour and speed on highways would help government officers such as transportation strategic planners make informed development decisions. It is also very valuable to end users who would like to understand the traffic situation to avoid traffic jams or slow speed time. We design algorithms and implement a system that discovers people’s intercity trips and extracts intercity travel trajectories suitable for estimating traffic speed on intercity highways. The data and results are visualized using a visualization engine we developed for this project. Besides describing our ideas and system, we in this paper also present our analysis results of intercity travels and highway traffic.

I. INTRODUCTION

Mobile phones are increasingly a part of our life as we move around and stay in contact with one another. Data of mobile phone calls, messages, and Internet access (CDR data) can directly convey the approximate location of each phone user at a specific time, and allow for effective and automated collation of movement data for an entire population over large area [11], [14], [10]. Such data can be used to study social or economic behaviours and make high-impact contributions towards predicting epidemics [13], [15], detecting crisis [2], and urban planning [3], among others.

Taking on the opportunity provided by Orange “Data for Development” (D4D) challenge [1], we propose to use anonymized mobile phone (CDR) data collected in Ivory Coast and publicly-available geographical map data to study two important and interrelated problems:

- People’s intercity travels. For example, what cities people travel from and travel to, and how frequently they travel.
- Health of intercity highway networks. For example, how is the highway traffic speed at different hours of a typical weekday.

We believe CDR data can be used to study the above two problems, although CDR data of a mobile user can be temporally sparse and spatially coarse. People normally use their mobile phones before they depart from their current city and after they arrive at new city. For example they will contact their family to report safety. They also call or SMS their friends during their intercity journey when they are bored or excited. These phone usage before, in, and after their intercity trip produce data for our study.

The technical and data challenges that we address in our study include the following. How to find out whether a user is in a city? How to discover that a user travels from one city to another? How to get the intercity transportation infrastructure data of Ivory Coast? What kind of intercity trip trajectories can be used to estimate highway traffic speed? How to estimate the highway traffic speed with such intercity travel trajectories? How to effectively visualize data and results?

Fig. 1: Example illustrating our study approach.
We design and implement a system that solves all the above challenges. Figure 1 illustrates our solution. We apply clustering algorithm on antenna position data (red and green points) to automatically discover antennas in the major cities (green points). From users’ CDR data we associate users with cities and extract intercity trips (e.g., a user travels from Abidjan to Yamoussoukro). By looking at the detailed CDR records during the trip, we discover and select the intercity trajectories (cyan polyline) along highways (orange lines). We then use our carefully designed algorithm to select suitable intercity trajectories to estimate the traffic speed on highways. City information and highway road networks are extracted from OpenStreetMap data [6].

The results are visually communicated to the decision makers using a visualization engine we develop for this project. Figure 2 shows two representative examples of our analysis and mining results. Figure 2a illustrates the volumes of intercity travels from Abidjan to other major cities. Lines in different color are trips to different cities. Figure 2b is a visualization of the average highway speed in the 3PM-6PM interval. Red segments are the places where traffic is very slow.

Our system reveals some very interesting facts about people’s intercity travel behavior and the state of highway travel in Ivory Coast. For example, a lot of people travel from Abidjan to Yamoussoukro in the middle of April 2012. As another example, our temporal analysis of the CDR data consistently reveals that the slowest commute speeds occur daily between 9p.m. to 6a.m. on most intercity highways over the five months of data. We are intrigued by this and conducted some information search online, which revealed the possible explanation: inadequate lighting make driving conditions hazardous [5], [8].

Such insights from data analytics demonstrate the immense potential of using the readily available mobile CDR dataset to reliably find out something that is typically much more costly to implement, such as transport network analysis using traditional measurement-intensive techniques like travel survey. Our solution is much faster and cheaper as it requires only anonymized mobile phone data and open source map data.

Our system can be used to make recommendation to the Ivory Coast government on the locations where the most benefits will be felt if they have to prioritize their infrastructure improvement works. Possible actions that can then be taken include short/medium term works like lane expansions and addition of slip roads, and longer-term investments such as commissioning the constructions of a new highway linking two major cities. The decisions shall be supported by scientific, data-driven observations from our fully-automated system and will be based on key considerations including intercity travel volume and highway speed limitation.

The ability to recommend specific segments of each highway for immediate improvement possibilities is highly impactful for a country with limited resources and growing economy that depends a lot on the efficacy of the transportation network for the workforce and the businesses.

The rest of this report is organized as follows: Section II describes the details of our proposed solution. In Section III we present the main results that demonstrate the usefulness of our proposed system. Section IV introduces our supplementary materials that demonstrate our system, including videos and other graphical visuals. Finally, Section V concludes this report.

II. SOLUTION

The Orange “Data for Development” (D4D) datasets consist of anonymized Call Detail Records (CDR) of phone calls and Short Message Service (SMS) exchanges of 500,000 of Orange’s customers in Ivory Coast between December 1, 2011 and April 28, 2012. For this study, we use the second set of data (SET2) which contains the coordinates of antenna positions and high resolution trajectories of 50,000 randomly sampled individuals in every two-week periods for the entire observation period.

Figure 3 depicts the broad overview steps involved in extracting intercity trips and estimating the speed profile of the highway segment using CDR data, antenna location data, and geographical map data.
Together with the location of the city extracted from the geographical map data of Ivory Coast from OpenStreetMap (OSM) (see section II-A) as the starting seed location, we design a clustering algorithm that autonomously determines the subset of cellular antennas within each city (see section II-B). After we have determined the set of cellular antennas in every city, we extract all communication records of individuals that have CDR data originated from a cellular antenna in a city and terminated at a cellular antenna in another city. These communication records are used to estimate the inter-city trips made by individuals within each two-week time period (see section II-C). Using the cellular antenna position associated with these inter-city communication records and highway geospatial information from OSM, we determine whether a trajectory is along highways and decide the trajectory can be used to estimate highway traffic speed. For those selected trajectories, we map their individual communication records to the nearest highway segment. We then estimate the speed profile of the interconnecting highway segments using the estimated distance which individuals travelled during the time between two communication records.

A. Map Data and Processing

OpenStreetMap (OSM) is a free worldwide map created and maintained by the public. Worldwide geographical data from OSM is available under the Open Data Commons Open Database License (ODbL) [6]. OSM map data is structured with basic data primitives: node, way, and relation. A node contains geospatial information of a single location. A road, stream or railway is defined by way data primitive by using an ordered interconnection of between 2 and 2000 nodes. A relation data primitive defines the relationships between other data primitives.

For this study, we only study the transport infrastructure interconnecting the following major cities of Ivory Coast[7]. They are, Abidjan, Bouaké, Daloa, Yamoussoukro, Korhogo, San Pédro, and Divo. The city information, including name and a coordinate (probably city center), is described by node data primitives in Ivory Coast geographic data from OpenStreetMap. However, the OpenStreetMap data does not contain city boundary information. As a result, we need to devise a way to identify antennas in each city.

The geospatial information of Ivory Coast highways is described by way data primitives in OSM map data. Each highway is described by an interconnection of geospatial points. A highway segment is a section of the highway connecting two way points. A total of 356770 segments were extracted from the Ivory Coast geographic data from OpenSteetMap.

B. Antennas-in-City Discovery

The objective of Antenna-in-City discovery is to automatically detect the set of antennas in each city using the cellular antenna data across the whole Ivory Coast and city information extracted from OpenStreetMap data.

We define the antenna-in-city discovery problem as follows. Given a set of antennas \( A = \{a_1, ..., a_n\} \), and a set of points each representing a location in one of the major cities, the task is to identify the set of antenna in each city without knowing the city boundaries.

We propose a novel clustering algorithm for antenna-in-city discovery. To perform clustering, we first represent each antennas \( a_i, i = 1, 2, ..., n \) into a feature vector \( V(a_i) = (X(a_i), Y(a_i), Volume(a_i)) \), where \( (X(a_i), Y(a_i)) \) represents \( a_i \)'s geographical location coordinate and \( Volume(a_i) \) denotes \( a_i \)'s average call volume.

Intuitively, population density in cities is higher than that in non-city areas, therefore the antennas in cities are geographically close to each other and densely distributed whereas the distances between antennas outside cities are relatively far away from each other. At the same time, people in cities normally will call more (due to business) than people in non-city areas, so antenna in cities would have higher call volumes. As such, our feature representation takes both geographical location and average call volume information into consideration.

We modify a density based clustering algorithm [12] to group the antennas into a number of clusters/cities.
To initialize clustering algorithm, we also select a seed data point for each city/cluster. In particular, we have searched OpenStreetMap using each city name as a query word to get a point which represents an internal position of the city. We then expand each individual seed nodes by including its nearby antennas to form city clusters. Particularly, for each seed, if its neighborhood contains more than a certain number of neighbors, then we expand all its direct neighbors into corresponding clusters. Similarly, for each of the newly added neighbors, we continue adding their neighbors if their neighborhoods are also located at the high density regions and exhibit high call volumes. This recursive clustering process stops until no more neighbors can be added into the clusters. Finally, we output all the clusters where each of them corresponds to a city.

Note our proposed density based algorithm can detect arbitrarily shaped and sized clusters as long as the antenna data points within each clusters are densely connected. This is especially important for the city boundary detection, as in our case cities can have totally different shapes and/or with different sizes.

We have compared our automatically discovered antenna clusters with the city boundaries approximated through Google Maps. Figure 4 shows the antenna clusters and city boundaries. Our clustering results match very well with the sets of antenna in city boundaries, indicating that our clustering algorithm is very effective for antennas-in-city discovery.

There are a few advantages of using an algorithm to automatically discover antenna-in-city over relying on manually collected city boundaries data. First, the process is automated. Second, the algorithmic solution can scale to many cities. Third, our solution adapts with data and therefore adapts with the development (e.g., expansion) of cities.

C. Intercity Trip Trajectory Extraction

After discovering antennas for each city, we associate antennas in city with the city they belong to. The CDR records are now also attached with city tags. A mobile user made an intercity trip if two of his/her CDR records are associated with different cities. The trajectory of the
intercity trip is the sequence of CDR records from one city to another city.

For convenience, we define a city as the set of antenna in that city.

Definition 1 (city): Given a set of antennas \( A = \{a_1, \ldots, a_n\} \) and a location in city \( i \), a city \( c_i \) is defined as the cluster of antenna our clustering algorithm outputs for city \( i \).

In other words, the city \( c_i \) is defined as a set of \( m \) cellular antennas \( \{a_{i1}, a_{i2}, \ldots, a_{im}\} \) which is a subset of the whole antenna set \( A \).

Definition 2 (CDR trace): CDR trace \( ct_i \) of user \( i \) is the sequence of his/her CDR records \( ct_i = \{(a_{i1}, t_{i1}) \ldots (a_{in}, t_{in})\} \) where \( (a_{ij}, t_{ij}) \) is a CDR record. \( a_{ij} \) is an antenna and \( t_{ij} \) is a timestamp.

Given a user’s CDR trace, the user made a intercity trip if there exist two records in his/her trace whose antenna belong to different cities.

Definition 3 (Intercity Trip Trajectory): Given a CDR trace \( ct_i \) of user \( i \), an inter-city trajectory \( tr \) of the user \( i \) is defined as the shortest subsequence of \( ct_i \) such that the first CDR record and the last CDR record are in different cities. \( tr = \{(a_{ip}, t_{ip}) \ldots (a_{iq}, t_{iq}) | a_{ip} \in c_1, a_{iq} \in c_2\} \).

Note that a user can have several intercity trip trajectories extracted from his/her CDR trace. Here “shortest” means that the number of records in that trajectory can not be smaller. More precisely, it means that the first CDR record of the trajectory is the last record of the user in the departure city and the last record of the trajectory is the first record of the user in the arrival city.

D. Trajectory Along Highway

Because one of our objectives is to study the highway transportation infrastructure amongst cities of Ivory Coast, we need to select intercity trajectories that use the highway. For this purpose we need a way of identifying whether an intercity trajectory is along highways.

Our basic idea is to find out the set of antennas that will serve the users on highways. If during a trip the user made all his/her calls through these antennas then with high probability the user was travelling along highways.

We apply the Voronoi Diagram technique [9] to find out the antenna along highways. Given a space and \( n \) points, called seeds, in the space, a Voronoi diagram is a way of dividing the space into \( n \) regions, called Voronoi cells, such that every cell covers exactly one seed and all points in the Voronoi cell are closer to the seed in this cell than to any other seeds in any other cell.

Definition 4 (Highway Antennas): Given a set of cellular antennas \( A = \{a_1, \ldots, a_n\} \), a function \( V(a_i) \) which gives the Voronoi cell of a given antenna \( a_i \), and a set of highway segments \( S = \{s_1, \ldots, s_m\} \), the set of highway antennas is defined as \( A' = \{a \in A | \exists s : S, V(a) \cap s \neq \emptyset\} \).

Basically a highway antenna is an antenna whose Voronoi cell spatially overlaps with at least one highway segment. Here we slightly misuse the symbol \( \cap \) and \( \emptyset \).

If all the CDR records in a trajectory are associated with highway antennas, we say this trajectory is a highway trajectory.

Definition 5 (Highway Trajectory): Given a set of highway antennas \( A' \) and an intercity trajectory \( tr = \{(a_{ip}, t_{ip}) \ldots (a_{iq}, t_{iq})\} \), \( tr \) is a highway trajectory if and only if \( \forall (a,t) \in tr, a \in A' \).

Please refer to Figure 1 to see an example of highway trajectory.

E. Further Selection of Highway Trajectories

Not all highway trajectories are used to derive highway speed. We further select the trajectories by looking at trip’s “observed” duration and the number of CDR records available during the trip.

The “observed” duration of a trip is the time from the traveller’s last CDR record in departure city to the first CDR record in arrival city. If the “observed” duration is too long, for example more than 12 hours, the traveller might have breaks or even stopover during his/her trip, and therefore speed estimation based on such trips would be very erroneous.

CDR records during a trip are in fact the time and location samples of the traveller’s true trajectory. If there is very few CDR records during a trip, we won’t know have enough information to find out which highway the traveller had took and again the speed estimation can be very wrong.

For these two reasons, we further restrict the trajectories used in highway speed study to the highway trajectories with reasonable “observed” travel duration and at least a few CDR records.

F. Derive Highway Speed Profiles

From the previous steps, we obtain a set of intercity trajectories which are close to the highway and fulfill our other selection criteria. Now we use them to build speed profiles for different segments of the highway.

We project the trajectories onto the highway segments (using map-matching) and estimate the speed the user was travelling (assuming the user was travelling with
constant speed between two consecutive CDR records). Then for each of the highway segments passed by the user, we use the estimated travel speed as observed traffic speed at the time the user passed that segment.

Note that a highway segment will have many speed observations because many users’ highway trajectories are mapped to that highway segment. These speed observations are timestamped. Although the speed estimation for each particular trajectory may not be very accurate (due to limited antenna spatial resolution), the statistical measures (e.g., the median) of these speed observations could be quite telling.

**Definition 6 (Highway Segment Speed Profile):** For each highway segment $s_i$, a time-dependent speed profile $P_i$ is defined as $P_i = \{(v_{i,j}, t_{i,j}) \mid j = 1 \ldots n_i\}$ where $n_i$ is the total number of trips recorded for the highway segment $s_i$, $v_{i,j}$ and $t_{i,j}$ are the speed and time stamp of the $j^{th}$ trip over this highway segment respectively.

Recall that each highway trajectory is a set of tuples $tr = \{(a_1, t_1), \ldots, (a_n, t_n)\}$ where $a_i$ is the antenna through which the call is made and $t_i$ is the time of call. We perform the trajectory projection by pairs of consecutive calls. We define a function $N(a)$ which returns the nearest way point of a given antenna $a$. We also use a function $SP(p_1, p_2)$ which implements a standard Shortest Path algorithm and returns a set of highway segments for two given way points $p_1$ and $p_2$. Lastly we implement a function $D(p_1, p_2)$ which returns the road network distance of two given way points $p_1$ and $p_2$.

Now for each pairs of consecutive call records $(a_{i_1}, t_{i_1})$ and $(a_{i_2}, t_{i_2})$ in a highway trajectory, we compute the speed $v = \frac{t_{i_2} - t_{i_1}}{D(N(a_{i_1}), N(a_{i_2}))}$. Then we update the speed profile of all highway segments in $SP(N(a_{i_1}), N(a_{i_2}))$ by appending a tuple $(v, t^*)$ where $t^*$ the estimated time by assuming that the user travels at constant speed from way point $N(a_{i_1})$ to way point $N(a_{i_2})$.

### III. Results

#### A. City, Antennas-in-city, and Road Networks

We focus on the seven major cities in Ivory Coast. Figure 5 depicts our results of antennas-in-city discovery and road network extraction. Antennas are shown with red and green points. Antennas in cities are shown with green points. They are the dense clusters discovered by our program. Yellow lines are the highway networks extracted from the OpenStreetMap road data.

![Figure 5: City names (black), antennas (red and green), antennas in city (green), and highways (orange).](image)

Note that if authentic data of city boundaries and highway road network are available (for example from Ivory Coast government), they can be easily incorporated into our system to improve data quality. For example, authentic city boundary data can be used to find out the accurate sets of antenna-in-city.

#### B. Intercity Trips

The CDR data set provided by Orange D4D contains 10 samples where each sample contains the records of randomly selected 50,000 users in 2 weeks. The time spans of the 10 samples do not overlap. The whole time span is 20 weeks: last four weeks of 2011, and the first sixteen weeks of 2012. Totally we have CDR records of 500,000 users collected in 20 weeks (2011-12-05 to 2012-04-22) where data of each user are collected in only two weeks.

With this dataset, our system discovers that

- 15,800 users had intercity trips, and
- they made 28,860 intercity trips.

That means that about 3.2 percent (15800/5000000) of the sampled population make intercity trips during two weeks, and the average number of trips made by an intercity traveller is about 2. Note that travelling from city A to B and then travelling back from city B to A are counted as two trips.

Figure 6 shows the distribution of the number of intercity trips made by the intercity travellers. Most of the travellers made one or two intercity trips within two
weeks, and the number of travellers who made one trip is little more than the number of travellers who made two trips. It is reasonable because the time span of each data sample is only two weeks, thus the data do not capture both the “go” and “return” trips for all travellers. From the chart, we also see that there are some users who travels a lot. For example, there are users who made more than six trips in two weeks.

Fig. 6: Distribution of the number of intercity trips made by users.

Fig. 7: The number of intercity trips in each month.

Figure 7 and Figure 8 show the time series characteristics of the intercity trips by month and week respectively. The number of trips observed in different months are quite similar. Considering that the dataset includes 27 days of December and 22 days of April, then it seems that people travel a little more in December and April than in the other months. The time series by week is more fluctuant, as shown in Figure 8. The sudden drop in the 12th week of 2012 and the sudden increase in the 14 week of 2012 are interesting.

1) Origins and Destinations: Figures 9 is a visualization of the intercity trips. Different colors are used to represent trips between different cities.

Table II lists the top ten most travelled city pairs and their shares of intercity trips. We see that the shares of trips from city A to B and from B to A are quite similar. Figures 10a and 10b illustrate the percentage of intercity trips from and to the major cities respectively. For example, of all the inter-city trips, 32% are from the city of Abidjan and 31% are to the city of Abidjan. Again, it can be observed that the percentage of trips from a city is approximately equal to that of trips to the city. It is

<table>
<thead>
<tr>
<th>City Pair</th>
<th>Share of Outbound</th>
<th>Share of Inbound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abidjan - Yamoussoukro</td>
<td>0.137</td>
<td>0.131</td>
</tr>
<tr>
<td>Yamoussoukro - Abidjan</td>
<td>0.090</td>
<td>0.077</td>
</tr>
<tr>
<td>Abidjan - Divo</td>
<td>0.064</td>
<td>0.063</td>
</tr>
<tr>
<td>Abidjan - Bouaké</td>
<td>0.052</td>
<td>0.046</td>
</tr>
<tr>
<td>Bouaké - Korhogo</td>
<td>0.039</td>
<td>0.033</td>
</tr>
</tbody>
</table>

TABLE I: Most travelled city pairs
also clear that Abidjan, Yamoussoukro, and Bouaké are the most busiest cities in Ivory Coast.

Some intercity trips are quite short, e.g., 2 and 3 hours only. They probably are flight trips.
- There are also many very long trips, e.g., above 12 hours. They are probably due to the following reasons. The user did not make call right before he/she leaves. The user did not make call right after he/she arrives. The user makes stopover during trip.

When we use intercity trip trajectory data to study highway speed, we filter out those very short and very long trips.

3) Calls Made in Travel: Figure [13] shows the histogram of the numbers of CDR records of intercity trips. We observe that many people do not make many calls during their trips. This could because because making calls during trips is expensive.

Fortunately, we also see that about 40 percent trips have more than have five or more CDR records during their trips.
In our highway traffic study we only use highway trajectories with more than five CDR records.

![Graph showing the distribution of intercity trips based on the number of calls made in trips.](image)

Fig. 13: The distribution of intercity trips based on the number of calls made in trips.

C. Highway Speed

As described in Section II, we use intercity trip trajectories along highway that fulfill our selection criteria to study the traffic speed on highway. Our system selected about 4,000 intercity trip trajectories to derive highway segment speed profiles.

Figure 14 shows the average speed on highway by hour of day and day of week. We see that it exhibits clear patterns. Speed varies in the range from 20 km/h to 50 km/h. Speed in night is much slower than speed in daytime. In particular, speed between 9pm and 6am is very slow. A possible explanation (we found through Internet search) is that inadequate lighting make driving conditions hazardous [5], [8].

We also see that speed in weekend nights is faster than that in weekday nights. It is probably because people travel less on weekends and therefore there is less traffic on weekends.

Figure 15a-15h show the estimated median travelling speed at different segments of Ivory Coast major highways at different time intervals on every Thursday for the entire observation period. Segments that are coloured in red suggest that commuters are travelling at slow speed. Segments in yellow indicate commuters are travelling at moderate speed. Those segments in green suggest commuters are travelling at a faster speed. Segments in gray have no speed data available. Figure 15i shows the colour bar for the speed.

From these figures, we see that lesser number of segments are coloured during night time than during day time. In addition, we see that driving at night is understandably slower. This two findings are also consistently observed in other days of the week. We postulate that increase risk to personal security and hazardous driving condition during night time than during the day may be the two reasons for such findings [5], [8].

We also see that the highway segments directly connected to the cities normally have slower traffic than the other segments. It is reasonable since such segments would experience more traffic volumes. However, it is also possible due to the security checkpoints at the boundaries of the cities.

IV. VISUALIZATION

a short description of our visualization video

V. CONCLUSION

contribution

- study intercity travels with mobile phone data and OpenStreetMap data.
- study the intercity transport infrastructure with mobile phone data and OpenStreetMap data.
- a framework for the above studies.

The proposed method to using mobile call detail records in estimating the health of highway road network is limited by the following factors.

- Use of minor roads We have simply assumed commuters are travelling along the major highways. It is likely that commuters travelled along other minor roads that interconnect the cities that are under our study. However, we feel that majority of commuters would likely use the faster major highways for intercity trip. Thus, speed estimation of major highway segments will still be useful for the majority of the commuters.
- No communication during inter-city travel The method relies on communication records to estimate the travelling speed of commuters. However if commuters do not communicate during their travels, it will be less accurate in predicting the highway that commuters actually travelled since there can be many roads linking two cities.
- Delayed communication upon departure/arrival The proposed method relies on the time a user communicate. If a user did not communicate right before departure or arrival, speed of highway segments will be underestimated.
- Overlapping cellular antenna coverage Due to overlapping coverage of adjacent cellular antennas, distance travelled by a commuter will not be
captured even though he/she might have travelled some distance and communicated more than one time. This is because proposed method relies on the coordinates of connecting cellular antennas to estimate the speed of nearby highway.

todo: facing all above limitations what can be done to minimize inaccuracy

Nevertheless, the relative travel volume and relative speed on various highway segments will still represent the reality well.

REFERENCES


Fig. 15: Highway median speed plotted with different colours over different time intervals of every Thursday. Segment in red, yellow, and green respectively means the corresponding speed of vehicles at the segment is slow, moderate, and fast respectively. Segments in gray colour indicate no reliable data is available.