Inferring Commuting Flows using CDR Data: A Case Study of Lisbon, Portugal

Thanisorn Jundee
Dept of Computer Engineering
Faculty of Engineering
Chiang Mai University, Thailand
thanisorn_ju@cmu.ac.th

Chanadda Kunyadoi
Dept of Computer Engineering
Faculty of Engineering
Chiang Mai University, Thailand
sureerat_kunyadoi@cmu.ac.th

Anya Apavatjrut*
Dept of Computer Engineering
Faculty of Engineering
Chiang Mai University, Thailand
anya@eng.cmu.ac.th

Santi Phithakkitnukoon*
Dept of Computer Engineering,
and Excellence Center in
Infrastructure Technology and
Transportation Engineering
(ExCITE), Faculty of Engineering
Chiang Mai University, Thailand
santi@eng.cmu.ac.th

Zbigniew Smoreda
Sociology and Economics of
Networks and Services
Department
Orange Labs, France
zbigniew.smoreda@orange.com

*Corresponding author

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5966-5/18/10…$15.00
https://doi.org/10.1145/3267305.3274159

Abstract
Commuting is recurring travel between home and workplace, which accounts for most trips made daily. Understanding commuting patterns and flows is therefore essential for city and transport system design and planning. Traditionally, commuting flow information was collected using surveys and interviews, which are expensive and time-consuming. This paper introduces a way to extract commuting flow and route choice information from analyzing mobile phone communication logs. We present two new methods for inferring individual commuting route choice, which collectively constitutes city-level commuting flows. Both morning and evening flows are inferred and visualized. We believe that our methods and results are useful and contributing to both theory and practice, especially in the interdisciplinary field of urban computing and city science.

Author Keywords
Commuting flows; CDR data; mobile phone data; route choice.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

Introduction

Today's mobile phone is not just a communication device anymore. It has evolved significantly over the past few years with its additional advanced sensing technologies and useful features for handheld use, which makes it a dispensable part of our everyday lives. With its high penetration rate, a mobile phone is being carried by almost everyone these days. When connecting to the cellular network for voice, short message (SMS), or data services, communication logs are collected by the telecom service providers for billing purposes, in forms of the Call Detail Records (CDR) where each record contains a timestamp, corresponding communication activity (e.g., voice, SMS, or data), and location of the connected cellular tower. To use the service, the mobile phone thus needs to connect to the cellular network via the nearest cellular tower. Therefore, each time when the user connects for the cellular service, the user's communication and location information are recorded. Collectively, CDRs constitute a longitudinal behavioral data that can be analyzed methodically to reveal and understand various aspects of human behavior at different aggregate levels both in time and space.

Mobile phone data has a great advantage over the traditional human behavioral datasets that are mostly collected through surveys and interviews, which could be inaccurate, limited, expensive, and time-consuming. With the location traces of individuals, the CDR data can be used to advance research in human mobility, which is important for understanding transport behavior that requires a massive amount of data to truly explain or model each phenomenon with interdependent properties. Several studies benefited from the use of CDRs in human mobility research have yielded interesting findings. For instance, Song et al. [5] found that human mobility is highly predictable, showing an upper bound of 93% predictability that significantly reveals regularity in human movement. Phithakkitunukoon et al. [3] further show that human mobility is greatly influenced by social networks, as they found that 80% of the places that we visit are within just 20 km from a person we know, and we are 15% more likely to be traveling near our weak ties than strong ties. Not only the destinations that we travel to, but how we travel there is also influenced by our social networks as Phithakkitnukoon et al. [4] show that strong ties are more important to determine if driving is the person's transport mode choice, whereas weak ties are more important to determine if public transit is the person's choice. Understanding human mobility has a useful implication in transport system design and planning. Demissie et al. [1] show that CDRs can be used to infer travel demands that facilitate public transport network design, especially for developing countries where traditional travel surveys are costly and infeasible.

CDR-derived mobility pattern is shown to be a reasonable alternative – arguably is perhaps a better option because the results are not biased by the subjectivity of the surveyed participants' perception. Commuting (traveling between one's place of residence and place of work) is the most frequent and common trip made by a typical person, which collectively makes up the profound mobility flow patterns that often define the core mobility characteristic of the area – generally, morning flows (residence to workplace) and evening flows (workplace to residence). Commuting flows are therefore important for transport design and planning. Extending from our previous study [4], from the point

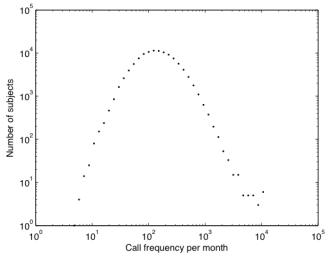

Figure 1: Histogram of call frequency.

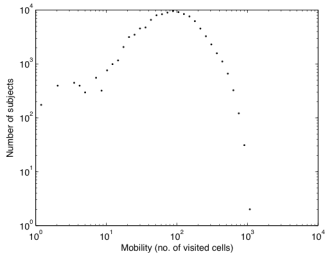

Figure 2: Histogram of mobility.

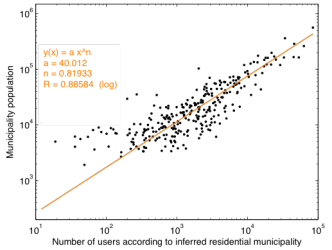

Figure 3: Correlation between the CDR-based population density and the census's.

of view of inferring individual commuting trips based on CDR data, in this paper we introduce two methods to determine routes used in both directions, which consequently makes up an individual commuting pattern. In addition, this paper also presents our developed visualization tool that graphically shows the preprocessed CDR information, individual inferred commuting route choices, and the city-level commuting flows.

Data Description and Processing

The CDR data used in this study was a set of 435,701,811 communication logs over one-year period (April 2006 to March 2017) from 1.3 million mobile phone users (1,318,905) in Portugal. The data accounted for approximately 13% of the population and was collected for billing purposes from all 308 municipalities of Portugal by one of the European telecom operators. To safeguard personal privacy, individual phone numbers were anonymized by the operator before leaving their storage facilities and were identified with a security ID (hash code). The CDR comprised the voice call information: timestamp, caller's ID, callee's ID, call duration, caller's connected cellular tower ID, and callee's connected cellular tower ID. The dataset did not contain information relating to text messages (SMS) or data usage (Internet). The location of the mobile phone user was recorded as the nearest connected cellular tower location when the users made or received a call. The dataset provided us with mobility characteristics of the mobile phone users over an extensive temporal window of observation. There were over 6500 cell tower locations in total, and each on average serves an area of 14 km^2, which reduces to 0.13 km^2 in urban areas such as Lisbon and Porto.

Based on our data, on average, a user makes 173 connections monthly (approx. 8 connected calls daily) to the cellular network – i.e., how frequent the mobile user location is recorded (becomes known). The users connected to the cellular network using on average 98 different cell towers throughout the year. The histograms of the call frequency and mobility (number of visited cell towers) are shown in Figs. 1 and 2.

As our study focuses on the commuting trips, so we first needed to identify a most probable place of residence (home) and work. We adopt the approach in [3] that infers the user's home location proximity based on the location of the most frequently used cell towers during nighttime (10PM – 7AM), and workplace based on the location of the most frequently used cell towers during business hours (9AM – 5PM). With this approach, our estimated home locations are comparable with the census population density with the correlation value of 0.89, as shown in Fig. 3.

In this preliminary research work, we focused our study to only analyzing commuting flows within the city Lisbon. To obtain the call history within the area of Lisbon, we filtered the entire CDR data with the following constraints:

- Each record must be an incoming or outgoing call established to and from Lisbon.
- Each record must be during weekdays (Monday - Friday). This is to filter only for weekday mobility data as most commuting trips take place on a weekday.

- Each user must be connected to the cellular network at least five times each month. This is to ensure fine-grained mobility traces.
- Each user must have at least 100 total connections during the morning commuting hours (7AM – 11AM) and 100 connections during the evening commuting hours (3PM – 7PM). This is to ensure an efficient amount of CDRs for our commuting trip analysis.

This data filtering process yielded a total of 6,813 mobile users for our study. As a technical note, since we were dealing with a big data here, we used *Google BigQuery*[1] as a tool for data processing and *Google Cloud Storage*[2] for quick and easy data retrieval and export.

Commuting Route Inference

We had inferred home and workplace locations for each user in Lisbon. Our goal was to identify the route that was likely to be used for commuting between home and workplace. Following [4] and [6], we used the *Directions API*[3] of the *Google Maps Platform* to first generate route choices for each user given the inferred home and workplace locations. This simplifies our problem into choosing the most probable route used among the suggested route choices by using the user's mobile phone usage history (i.e., locations and frequency) as a clue. An example is shown in Fig. 4 where there are three commuting route choices. The user's home and workplace are marked with 'A' and 'B', respectively. Call history consists of morning-hours connectivity (7AM – 11AM) and evening-hours connectivity (3PM – 7PM), which are marked with red and blue circles, respectively. The size of the circle corresponds to the amount of connections at the location.

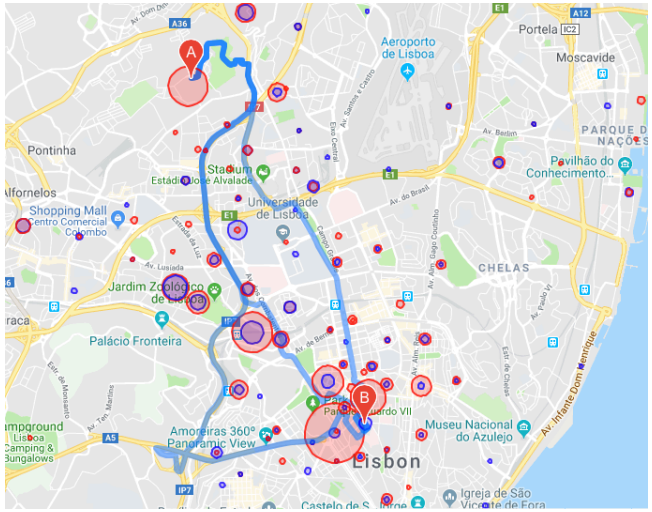

Figure 4: An example of commuting route choices (3 routes) of a mobile phone user, along with cellular network usage history (morning connectivity is in red, evening connectivity is in blue). Home is marked with 'A' and workplace is marked with 'B'.

To infer the commuting route, we proposed two methods:

Method A: Minimum Distance

The idea of this Method A is to select the route with the minimum distance to locations of the visited or used cell towers by the user. This method lies on the possibility of the user using his/her mobile phone to

[1] https://cloud.google.com/bigquery/

[2] https://cloud.google.com/storage/

[3] https://developers.google.com/maps/documentation/directions

connect to the cellular network from different locations along his/her commuting route throughout the year.

Basically, this method calculates an average distance of each route choice and selects the one with the minimum value. Supposed that X_k is a set of the geographical coordinates (latitude, longitude) or waypoints obtained from the Google Directions API of the route k, i.e.,

$$X_k = \{(x_1^{lat}, x_1^{lon}), (x_2^{lat}, x_2^{lon}), (x_3^{lat}, x_3^{lon}), \dots, (x_n^{lat}, x_n^{lon})\}, \quad (1)$$

where (x_v^{lat}, x_v^{lon}) is a latitude-longitude pair of the waypoint v and the total number of waypoints is n. Let Y be a set of visited cell tower location coordinates, i.e.,

$$Y = \{(y_1^{lat}, y_1^{lon}), (y_2^{lat}, y_2^{lon}), (y_3^{lat}, y_3^{lon}), \dots, (y_m^{lat}, y_m^{lon})\}, \quad (2)$$

where (y_u^{lat}, y_u^{lon}) is a latitude-longitude pair of the cell tower location u and the total number of previously visited cell towers is m. The route to be chosen (k) is the one that minimizes the average Euclidean distance (D) i.e.,

$$\arg\min_{k \in \{1,2,\dots,n\}} D = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \sqrt{\left(x_i^{lat} - y_j^{lat}\right)^2 + \left(x_i^{lon} - y_j^{lon}\right)^2} \quad (3)$$

This method however has a drawback. The waypoints obtained from the Google Directions API do not equally spread along the route. A straight line would be represented with two waypoints at one of the ends. Curving part of the route would consist of more waypoints than a straighter part of the route, as shown in in Fig. 5. Therefore, when calculating D the curving portion of the route dominates the result. An example is shown in Fig. 6 where each red line represents the Euclidean distance from a waypoint along the route to one visited cell tower. Denser lines can be observed near curving part of the route. For this reason, the result from the method A favors the route with visited cell towers closer to the curving part of the route.

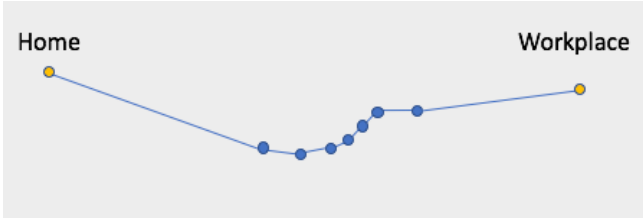

Figure 5: An example of waypoints that are denser at curving part of the route.

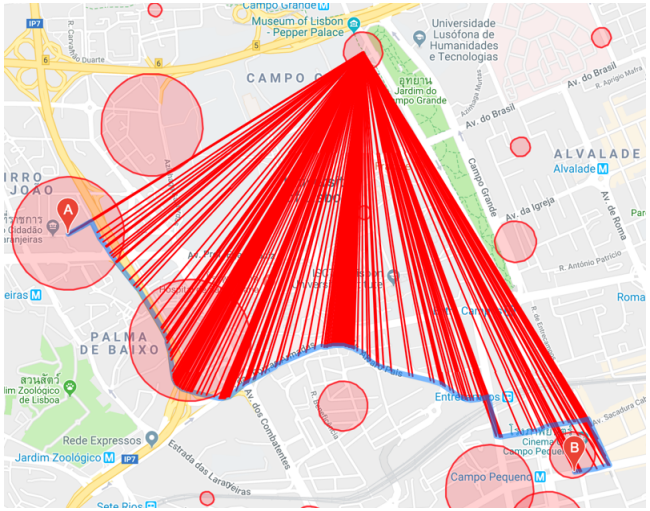

Figure 6: An example of calculating a distance from each cell tower location to waypoints along the route. Denser lines (distance measures) can be observed near curving part of the route.

Method B: Maximum Overlap

To resolve the issue of inconsistent waypoints distribution along the route, we proposed another method that interpolates and extrapolates the waypoints to normalize the spacing between them. We do so by simply using grids as reference and create a new data point to represent a waypoint at the centroid of the grid within which the route passes by. An example is shown in Fig. 7 where red dots are the new data points created by this interpolation/extrapolation process to replace the original waypoints represented by blue dots.

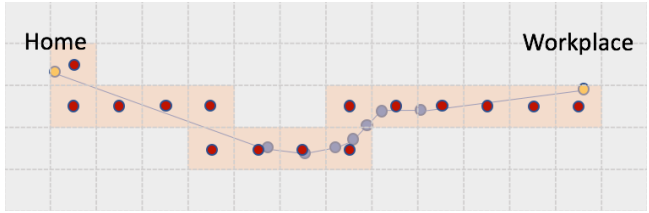

Figure 7: An example of our interpolation/extrapolation process of waypoints along the route by using grids. Blue dots are the original waypoints and red dots are the result of the interpolation/extrapolation process.

As opposed to the minimum distance to cell towers calculation, we created the coverage area of each used cell tower, so the route that maximizes the overlap area with the used cell tower coverages is to be selected. The average coverage distance of all used cell towers by the user was used as the coverage area for each cell tower for this calculation. Mathematically, suppose that W_k is a set of processed waypoints of route k, i.e.,

$$W_k = \{(w_1^{lat}, w_1^{lon}), (w_2^{lat}, w_2^{lon}), (w_3^{lat}, w_3^{lon}), \dots, (w_p^{lat}, w_p^{lon})\} \quad (4)$$

where (w_v^{lat}, w_v^{lon}) is a latitude-longitude pair of the processed waypoint v and the total number of waypoints is p. The route to be chosen (k) among n alternative routes is the one that maximizes the overlap cardinality ($\bar{\bar{O}}$) or the set elements of O, i.e.,

$$\arg\max_{k \in \{1,2,\dots,n\}} \bar{\bar{O}} = |\{W_k | \exists\, v : (w_v^{lat}, w_v^{lon}) \in C\}|, \quad (5)$$

where O is an overlap, which is defined as a set of all waypoints that are located inside the coverage area of cell towers (C), i.e., $O = \{W_k | \exists\, v : (w_v^{lat}, w_v^{lon}) \in C\}$. An example of using this Method B in choosing the commuting route (with grid size of 10 meters) is shown in Fig. 8 where each red marker is the element in set O or the waypoint located inside the cell towers' coverage area. The number of red mark thus is the overlap cardinality ($\bar{\bar{O}}$), which is to be maximized.

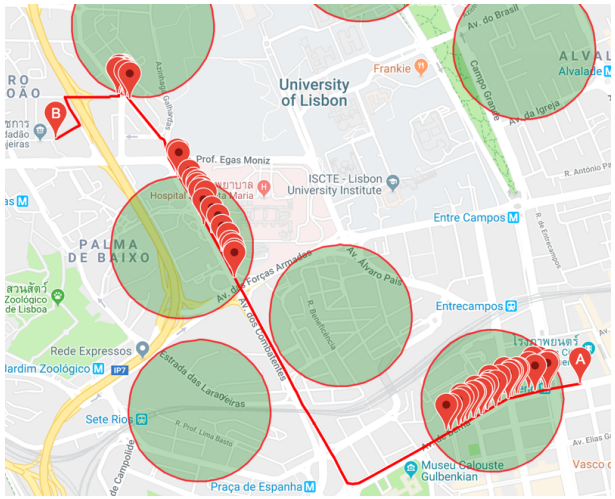

Figure 8: An example of using Method B for choosing the commuting route that maximizes the overlap cardinality.

Visualization Tool and Commuting Flows

In this study, we've developed an online visualization tool for our analysis. We built the tool with *Google Firebase*[4] as a database server and *Google Maps API*[5] to display map, shapes, and markers. The tool is available at http://myweb.cmu.ac.th/thanisorn_ju/index.html. A snapshot of the online tool is shown in Fig. 9. The tool allows the user to select to view an individual user's CDR-based information such as morning (Day) and evening (Night) connectivity with associated locations, inferred home and workplace locations (marked with A and B, respectively), route choices, selected route based on methods A and B, and the overall commuting traffic flows.

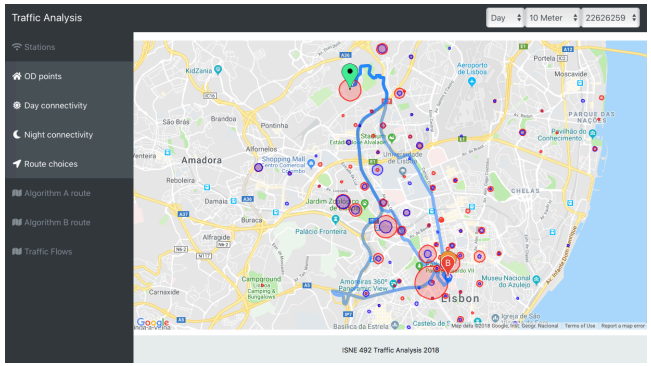

Figure 9: A snapshot of our online visualization tool.

In most of the cases, the methods A and B yield the same route choices. However, there are a few cases where their selections differ.

For example, in Fig. 10, the tool shows the selected morning commuting route (trip from home to workplace) based on the Methods A and B, which is shown with blue and red lines respectively. Each method selects a different morning route for this user. In Fig. 11, with the same user, the selected evening commuting route (trip from workplace to home) is the same one from both methods. Interestingly, this evening route appears to be different from the morning route. This case seems to support the concept of asymmetry in travel behavior [2], which states that people change their route approximately 15% of the time, even in their commuting trip.

Each of individual commuting trips collectively makes up the commuting traffic flows, which are very important to understand for traffic design and planning. In Fig. 12, our visualization tool shows all selected morning commuting routes of all 6,813 mobile users, which illustrates the overall morning commuting flows in the city of Lisbon, based on Method A. Likewise, Fig. 13 shows the evening commuting flows based on Method A. Similarly, the morning commuting flows and the evening commuting flows based on the Method B are shown in Figs. 14 and 15, respectively. Slightly different commuting flows in the morning and evening can be observed. This information is a useful insight that can benefit a range of stakeholders.

[4] https://firebase.google.com

[5] https://developers.google.com/maps/documentation/

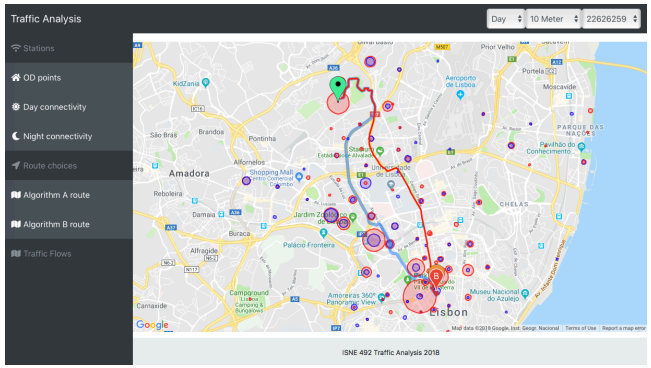

Figure 10: An example of the tool showing the selected morning commuting route by Methods A and B, with blue and red lines respectively.

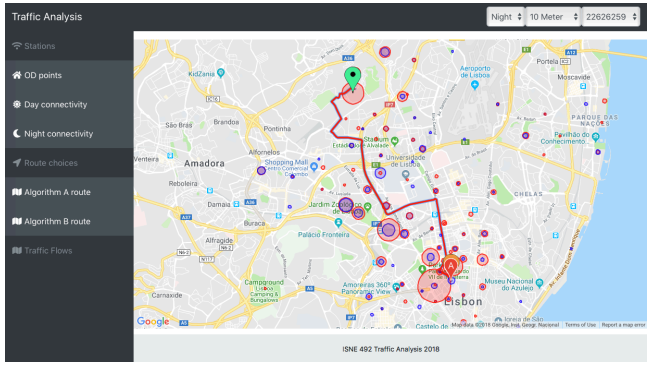

Figure 11: An example of the tool showing the selected evening commuting route by Methods A and B, with blue and red lines respectively. Same route is selected by both methods.

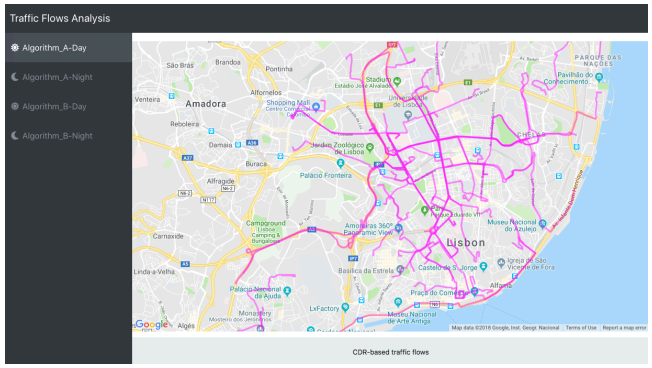

Figure 12: Morning commuting flows (home to workplace) in Lisbon, based on Method A.

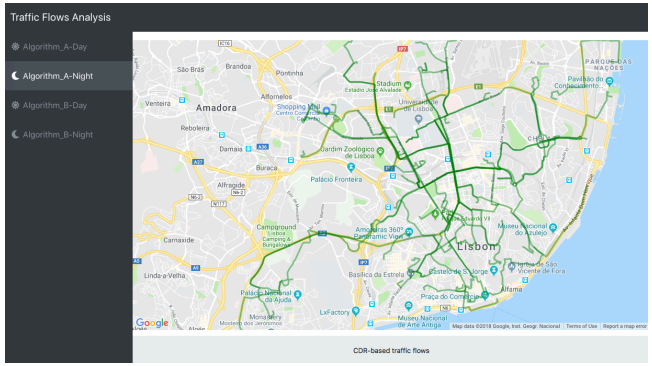

Figure 13: Evening commuting flows (workplace to home) in Lisbon, based on Method A.

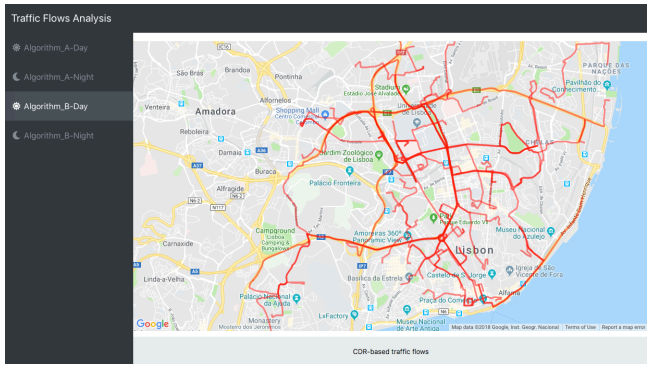

Figure 14: Morning commuting flows (home to workplace) in Lisbon, based on Method B.

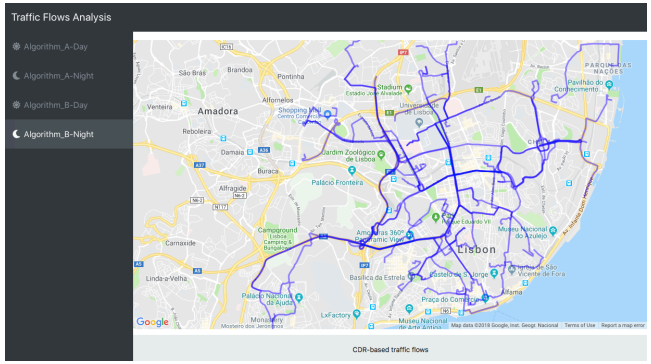

Figure 15: Evening commuting flows (workplace to home) in Lisbon, based on Method B.

Although the result obtained in this study cannot be realistically validated against the ground truth information, we still believe that the result is valid to a certain extent. The result and its methodology can be used to support decision making, design, and planning of urban space.

Conclusion

This study turns typical mobile phone communication logs collected from billing purposes into useful insight about the routes people use to commute from home to workplace in the morning as well as in the evening when people travel back home from their workplaces. Transport engineers and researchers typically spend a huge budget to obtain such information regarding route choices of commuters. Here, we introduce methods to extract the commuting route information from CDR data. We introduced two methods to infer individual commuting route based on the mobile phone communication history. One method aims at minimizing the distance between the route and the historic locations where communications were made. The other method aims at maximizing overlap area between interpolated/extrapolated waypoints along the route and the location proximity where communications were made. We've also developed an online visualization tool for our analysis to view intermediate (i.e., individual commuting route) as well as final results (i.e., commuting flows).

Nonetheless, there are some limitations to our current study, which will be addressed in our future work that include incorporating call frequency into our inference model and integrating both Methods A and B. We will continue with this investigation, which we believe is useful and contributing to both theory and practice in interdisciplinary domains such as urban computing, intelligent transportation systems, and city science.

References

1. M.G. Demissie, S. Phithakkitnukoon, T. Sukhvibul, F. Antunes, R. Gomes, and C. Bento. 2016. Inferring Passenger Travel Demand to Improve

Urban Mobility in Developing Countries Using Cell Phone Data: A Case Study of Senegal. *IEEE Transactions on Intelligent Transportation Systems* 17, 9.

2. Nick Malleson, Anthony Vanky, Behrooz Hashemian, et al. 2018. The characteristics of asymmetric pedestrian behavior: A preliminary study using passive smartphone location data. *Transactions in GIS* 22, 2: 616–634.
3. S. Phithakkitnukoon, Z. Smoreda, and P. Olivier. 2012. Socio-geography of human mobility: A study using longitudinal mobile phone data. *PLoS ONE* 7, 6.
4. S. Phithakkitnukoon, T. Sukhvibul, M. Demissie, Z. Smoreda, J. Natwichai, and C. Bento. 2017. Inferring social influence in transport mode choice using mobile phone data. *EPJ Data Science* 6, 1.
5. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert László Barabási. 2010. Limits of predictability in human mobility. *Science*.
6. Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*.