Profiling presence patterns and segmenting user locations from cell phone data

Yan Leng
MIT Media Lab
Cambridge, MA, USA
yleng@mit.edu

Haris Koutsopoulos
Northeastern University
Boston, MA, USA
h.koutsopoulos@neu.edu

Jinhua Zhao
MIT
Cambridge, MA, USA
jinhua@mit.edu

ABSTRACT

The dynamic monitoring of commuting flows is crucial for improving transit systems in fast-developing cities around the world. However, existing methodology to infer commutingOriginations and destinations have to either rely on large-scale survey data, which is inherently expensive to implement, or on Call Detail Records but based on ad-hoc heuristic assignment rules based on the frequency of appearance at given locations. In this paper, we propose a novel method to accurately infer the point of origin and destinations of commuting flows based on individual’s spatial-temporal patterns inferred from Call Detail Records. Our project significantly improves the accuracies upon the heuristic assignment rules popularly adopted in the literature. Starting with the historical data of geo-temporal travel pattern for a panel of individuals, we create, for each person-location, a vector of probability distribution capturing the likelihood that the person will appear in that location for a given the time of day. Stacked in this way, the matrix of historical geo-temporal data enables us to apply Eigen-decomposition and use unsupervised machine learning techniques to extract commonalities across locations for the different group of travelers, which ultimately allows us to make inferences and create labels, such as home and work, on specific locations. Testing the methodology on real-world data with known location labels shows that our method identifies home and workplaces with significant accuracy, improving upon the most commonly used methods in the literature by 79% and 34%, respectively. Most importantly, our methodology does not bear any significant computation burden and is easily scalable and easily expanded to other real-world data with historical tracking.

KEYWORDS

Human mobility; Home-Workplace; Eigenlocations; K-means clustering;

1 INTRODUCTION

Understanding human mobility and urban dynamics is the basis for researchers, transit operators, urban planners and location-based service providers in better understanding transportation demand, planning services and implementing urban policies. Traditionally, the mainstream data source in understanding travel demand is the household travel survey, which provides abundant transportation records, such as socio-demographic information, travel time, trip purposes, and travel mode. While it contains detailed travel logs and personal information, it is labor-intensive and costly to obtain in many aspects [29]. For example, the time interval between consecutive surveys conducted in developing countries is around 5 to 10 years, making it impossible to keep pace with the rapid urban development [20]. The rise of ubiquitous digital data collection infrastructures, embedded in urban areas, leads to a dramatic increase in monitoring urban dynamics and human mobility in an unprecedented wide-scale and finer granularity [3, 13, 22]. Various urban intelligent infrastructures, such as cell towers, Wi-Fi hotspots, and blue-tooth beacons, have exploded in the building of "smart cities" and making pervasive computing possible [9, 21]. Digital devices and sensors, such as mobile phones and chips in credit cards, have "intruded" into and monitored various aspects of our lives, such as mobility, health, financial transactions. Overloaded with the amounts and varieties of data sources, we need statistical and machine learning techniques to reveal salient behavioral features, inferring and generating hypothesis about behavior patterns, to optimize transportation systems, support planning decisions, and dynamically monitor regional delineations or land use classifications [4]. Call Detail Records (CDR), an opportunistic, large-scale, and longitudinal data source, generate many new possibilities in travel behaviors at both individual and aggregate level. It is the most widely-penetrated data source to be used as a proxy for travel behaviors. We have witnessed some efforts in both academia and industry in using CDR in transportation to solve critical problems. Alexander (2015)estimate Origin-Destination matrix and segment trips based on the inferred trips purposes, including home-based work, home-based other and non-home-based [2]. Kung (2014) compared commuting patterns, including travel time and distances, across different parts of the world at both country and city level [19]. Phithakkitnukoon (2010) analyzed the correlations in activity patterns for people who work in the same land use categories, which can further be used to estimate the most probable activities associated with the particular regions of the city [25]. Diao (2015) and Calabrese (2014) extracted the embedded travel activity information and infer activity patterns by merging mobile phone data and travel diary surveys. Identifying home and workplace locations...
are the building blocks for not only research mentioned above but also other real-world applications [6,9]. From an application perspective, understanding the spread of individual daily activities, detecting emerging residential and commercial regions, help policymakers better monitor urban dynamics and spatial-temporal distribution of the population. Knowing home and workplaces enable service providers to deliver better location-based services, such as targeting advertisements. Meanwhile, some studies couple the information extracted from CDR with census or large-scale survey as the ground truth to de-bias and up-scale mobile phone users to the whole population [16].

Along with the exciting opportunities and extensive applications, challenges are unavoidable. There are uncertainties, complexities and biases in the data collection as well as the human behavior itself. Despite the importance of accurately inferring home and workplace locations at the individual and aggregate level, existing methods are not accurate and flexible enough in inferring home and workplaces mostly with simplistic assumptions.

Fortunately, it has been shown by several existing studies that human behaviors, especially mobility behaviors, are highly regular and predictable [12,23,27]. The presence of such regular behavioral patterns has enormous practical and policy implications. In the field of transportation engineering, for example, it is a notoriously challenging task to infer the point of commuting origins (home) and, destinations (workplaces) to dynamically monitor the commuting flows, especially in fast-developing cities.

This paper aims to propose a methodology to extract regular behavioral patterns at locations and infer home/workplaces in urban spaces based on Call Detail Records by mapping physical CDR coordinates to user locations with enriched interpretations. Existing methodologies solve this problem by either relying on labor-intensive and untimely survey data or using ad-hoc heuristic assignment rules based on the frequency of appearances at given locations from CDR. In this research, we apply Eigen-decomposition and unsupervised machine learning tools to large-scale mobility data in a populated city in China to extract the commonalities in behavioral patterns across locations for different groups of travelers. In particular, we solve this problem by answering three questions: 1) what are the behavioral patterns at the user locations based on the longitudinal observations, 2) are there any common behavioral structures across the population, 3) add contextual information to user locations by labeling home and workplace locations.

The contributions of our work are four folds:

(1) We introduce a method to infer home and workplaces on CDR with proved better accuracy. This method is readily adaptable and tractable to other data with mobility or behavioral tracking.

(2) We propose a feature, Normalized Hourly Presence, to extract behavioral characteristics from CDR-based user locations which can uncover the shared behavioral patterns across the population using Eigen-decomposition.

(3) Testing based on real-world collected data shows that our methodology is remarkably successful at accurately inferring location labels such as home and workplaces, improving upon the most commonly used methods by 79% and 34% respectively.

(4) We apply the method to the CDR data in a populated city in China, which proved its feasibility and scalability in revealing the behavioral patterns and labeling home/workplaces in real-world settings.

We organize the remainder of the paper as follows. The next section introduces related works on inferring home and work locations based on CDR. Then, we describe the conceptual framework, the definition and calculation of Normalized Hourly Presence and eigenlocation, and the clustering techniques. After that, we test the proposed method on MIT reality mining data and compare it with the state-of-art approach in the literature. Moreover, we implement the proposed method on real-world CDR, collected in a populated city in China, to validate the feasibility, practicability, and scalability.

2 RELATED WORKS

With the increasing penetration and rising popularity of mobile phone and mobile communication, passive mobile phone location data become a possible source of the geographical data source to locate individuals and detect significant locations [1]. There is much research in inferring home and workplaces using CDR. Most of the studies focus on using intuitive features to capture the behavioral patterns at home or workplaces.

Number of presences. The most widely-accepted methods in inferring home and workplaces assume that home and workplace locations are two locations where people visit the most frequently. To quantify this assumption, they aggregate the number of presences at user locations. Calabrese (2011), Phithakkitnukoon (2010) and Jiang (2013) identifies home and workplaces as the ones with the most presences during home and work hours [5,17,25]. The simple assumption - the amount of data records is proportional to the stay length - is problematic given the characteristics of the CDR data. For example, people may use landlines at home or workplaces, the cost of which is low. Besides, not all users have at least one detectable home and workplaces from CDR if people do not use phones at home or they have multiple home and workplaces. However, this algorithm restricted the number of home and workplace to be one.

Dwelling time. Some studies use dwelling time, which measures the staying time at one location, as the feature to segment home and workplaces. Kung (2014) claimed that dwelling time is the longest at home and workplace during specific period [19]. Similarly, Sun (2014) set a dwelling time threshold for user locations to be home and workplaces [28]. They calculate dwelling time at a specific user location as the time difference between the first appearance at the user location and first appearance at another user location. However, this is unreliable due to the event-driven characteristic of CDR data. Moreover, the boundary thresholds to separate home and work hours are arbitrary.

Distance from home. Distance is used to capture the characteristics of the workplace. According to Alexander (2015) assumed that home is the most frequent location during home hours (before 8 am and after 7 pm) as most other research [2]. Meanwhile, a workplace is assumed to be the user location with the maximum distance from home. The assumption is that for a given frequency of visits,
longer trips are more likely to be commuting trips than shorter trips. This assumption, hanging on life-experiences and some empirical evidence from some old studies, needs further validation on commuting patterns worldwide and update-to-date. Besides, the characteristics of trips from travel surveys with high spatial granularities, are different from the un-triangulated CDR observations which have a spatial resolution of 0.5–2 kilometers.

Apart from proposing features to infer home and workplaces in an unsupervised fashion, some methods use small-scale experiments to calibrate parameters and calculate thresholds for segmentation. Ahas (2010) proposed mean and the standard deviation of the earliest call to differentiate home and workplace, which should be among the most frequent two user locations [1]. The justification for this study is that people either call early in the morning or late at night at home. For example, if the mean starting time of the earliest call is later than 17:00 or the standard deviation of the time of the first call in a day is higher than 0.175, they label this user location as home. They compute the thresholds from 14 individuals by tracking them for two months using CDR data. Though data collection is in real-world, this small-scale experiment is unrepresentative and cannot be generalized to other cities.

In conclusion, most of the methods proposed in the literature are problematic in ignoring the behavioral gap between the actual presence patterns and observed presence patterns from CDR. Using some arbitrary time boundaries and thresholds to differentiate home/ workplace/ third place from other user locations certainly generates bias. Moreover, the thresholds calculated from small-scale experiments are not transferable to other geographical regions due to the cultural differences and many reasons.

3 METHODOLOGY

In this section, we describe a behavioral method to map CDR-based user locations to home, workplace and third place based on longitudinal coordinates. Two basic terms used throughout this paper are user location and presence. We define a user location as the weighted centroid of a cluster of cell towers that approximates the exact locations of a user. We define the presence as the appearance of a user at a user location. Both terms are individually based.

3.1 Data description

CDR data records cell phone users’ traces with timestamps and approximate locations of cell phone users whenever they initiate phone calls, send/receive SMS or browse the web. It is event-driven and therefore does not cover the full picture of places that people have visited. The raw CDR data include encrypted user ID, timestamp, Location Area Code (LAC), cell tower ID and event type. It needs to be connected to another cell tower database to approximate the coordinates of the users with a range of 0.5–2km, which is low in spatial resolutions.

3.2 Conceptual framework

We now describe the conceptual framework of the behavioral method, as shown in Figure 1. For each user, we observe a sequence of coordinates with timestamps, representing the digital footprints of the user across the observational period. Each record, which we refer to as “presence” in this paper, can be associated with an activity type, which is the trip purpose. Note that there is no one-to-one correspondence between activities and user locations, meaning that people perform a particular activity at different user locations and people perform several activities at the same user location. Hence, the third layer, which we call activity layer, is observed from the passive-positioning and semantic-poor CDR data. People conduct daily routine activities in a limited number of user locations, which we roughly segment into home/workplace, as the anchor points, and third place. Presence patterns at user locations, shown at the bottom layer of the framework, are one of the most elementary aspects of human mobility. The objective of the behavioral method is to skip the activity layer and identify home/workplace from user locations based on the observed longitudinal presences. CDR-based home and workplace are different from the traditional concept of home and workplace. CDR-based home is the user location with home-like normalized presences. Similarly, CDR-based workplace is the user location with weekday and weekend work-like normalized presences.

3.3 Normalized hourly presence

We now introduce the proposed features to capture when, how often and how long do people appear at a particular user location. This feature extracts not only the location-based presence frequency but also the temporal presence variations at user locations on weekdays and weekends from the longitudinal records. Therefore, to characterize the temporal presence patterns of individuals at user locations and condense the time series data in a structured way, we propose a new feature: Normalized Hourly Presence (NHP). We sum the number of presences in each hour of a weekday and a weekend across the observational period due to the sparsity of the data. Weekdays and weekends are aggregated separately due to the different schedules, i.e., peoples schedules are different on weekdays when they need to go to work and weekends when they have more spare time. We normalize the hourly presences to the percentage of presences concerning the total number of phone connections of the individuals. The normalization captures not only the frequencies of visits but also the various call rates. Equation (1) shows the calculation of NHP, which most programming language can efficiently process, and we use PostgreSQL. In essence, we characterize each user location by a vector with 48 NHPs.

\[ \bar{P}_{L,h} = \frac{p_{I,h}}{\sum_{i=1}^{L} p_{I,h}}, \]

where \( p_{I,h} \) and \( P_{L,h} \) are the NHPs and absolute hourly presence for individual \( i \) during hour \( h \) at user location \( l \). \( L \in [1, 48] \), representing 24 hours on weekdays and weekends. \( L_{i,h} \) represents the unique set of user locations for individual \( i \) during hour \( h \).

A made example is show in Figure 2. User A presented at 5 places for 1, 0, 5, 1 and 0 times respectively during 6:00 – 7:00 on the weekday across the observational period (\( T_A^5 = 5 \)). His normalized presence at place 3 during this time period can be calculated as in equation (2).

\[ \bar{P}_{A,6} = \frac{p_{A,6}}{\sum_{l=1}^{48} p_{A,6}} = \frac{5}{5+0+5+1+0} = \frac{5}{7} \]
3.4 Eigenlocations

Due to the highly regularize and generalizability of human mobility, we reveal and extract the common behavioral patterns and daily routines at user locations across from the large-scale noisy user location dataset. This step is critical in enabling the understanding whether there exist common behaviors across the population, and if so, what are the presence patterns at user locations.

Principal Component Analysis (PCA) has been used to extract underlying structures from large-scale behavioral datasets according to [7, 11, 26]. The resulting, Eigenlocation, is used to describe the common presence patterns across the urban-wide population. Consequently, to capture the common presence patterns across the urban-wide population, we apply PCA based on the assumption that human’s presence patterns at user locations with similar functions are similar across the sample population. Each eigenvector, named as eigenlocation, represents a typical presence pattern by explaining a portion of the behavioral presences variances. We rank the eigenlocations by the explained variances, which is mostly the associated eigenvalue. Projecting original presence vectors onto eigenlocations reveal common behavioral structures and reduce noisy and random behaviors. Representing presence patterns with few eigenlocations indicate the generalizability of human behaviors.

The entire user locations set can be represented by \( \Gamma_1, \Gamma_2, \Gamma_3, \ldots, \Gamma_U \) for a total number of \( U \) user locations. \( \Gamma_U \) is a 48 dimension vector characterized by the 48 NHPs. The average presence pattern of the user location is \( \Psi = \frac{1}{U} \sum_{u=1}^{U} \varphi_u \). For normalization, \( \varphi_i = \Gamma_i - \Psi \) is the deviation of a user location from the mean presence patterns. Matrix \( A \) equals \( [\varphi_1, \varphi_3, \ldots, \varphi_M] \). The calculations are shown in equation (3) and equation (4).

\[
C = \frac{1}{48} \sum_{h=1}^{48} \varphi_h \varphi_h^T = AA^T \quad (3)
\]

\[
V'CV = \Lambda \quad (4)
\]

where \( \Lambda = \text{diag}\{\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_{48}\} \) are the eigenvalues and \( V = [v_1, v_2, v_3, \ldots, v_{48}] \) is an orthogonal matrix where the \( j_{th} \) column \( v_j \) is the eigenvector correspondence to \( \lambda_j \).

3.5 Clustering

We use clustering techniques to achieve the goal of extracting home and workplaces from the collection of unlabeled user locations characterized by normalized hourly presences in an unsupervised way. Clustering analysis is a standard explanatory tool to discover structures and grouping similar objects. It helps us to find patterns in a collection of unlabeled samples by organizing items that are similar in some way.

The first task of clustering is to define the similarity measure. The similarity measure captures the closeness between presence patterns, more specifically, weekday and weekend normalized presences. We use Euclidean distance, the sum of squares of differences between normalized hourly presences with each hour equally weighted to measure similarity. The observed presence patterns reveal the function of the location to the individuals. In other words, the cluster of user locations are the locations that have similar meanings to users, such as home and workplaces.
We integrate two clustering techniques in our framework: hard clustering and soft clustering, represented by K-means clustering and Fuzzy C-means Clustering (FCM). K-means clustering assigns a single cluster label to each user location. FCM, on the other hand, attributes a vector of membership of belonging to each user location cluster (home, workplace, third place in our case). The method assigns each user location to the cluster with the largest "membership", which we interpret as the confidence of belonging to that cluster. There exist uncertainties in inferring the home and workplaces due to the complexities of human behaviors and the characteristics of the data. FCM, though not as interpretable as K-means, captures the confidence of the inference results. On the other hand, this technique allows the trade-off between certainties and inference rate for different application purposes.

3.6 Algorithm
To summarize, the algorithm of our behavioral method works in the following way. The inputs are the presences, and the analysis unit is user locations, which is characterized by NHPs within 24 hours of weekdays and weekends. Eigen-decomposition is performed on the user locations to extract the underlying presence structures at the user locations. The output from this step is the eigenlocations, each representing a common presence structure. The projections onto the eigenlocations can reconstruct the presence patterns and rule out redundant and noisy patterns at user locations. K-means clustering and Fuzzy C-means clustering are used to cluster and segment user locations into the home, workplaces and third places.

4 EVALUATION AND APPLICATIONS
In this section, we apply our method on two datasets for three purposes:

(1) Test and evaluate our method by accuracy, inference rate, and flexibility.
(2) Reveal the feasibility of using the proposed method in extracting common behavioral structures and revealing interesting and interpretable patterns.
(3) Illustrate the practicality and scalability in real-world settings of our method in inferring home, workplaces and third places.

We first test and evaluate our method on MIT Reality Mining data by Human Dynamics Group at MIT Media Lab directed by Professor Pentland [10]. Professor Pentland’s team collected the data on more than 100 individuals by tracking them for more one year started in 2004. We compare our method with the state-of-art method, named as the "Most Frequent Appearance" method, as described later in this section, and show the improved accuracy and flexibility of our method. Last, we implement the method on the real-world data collected in a crowded city in China to demonstrate its feasibility and scalability real-world data set.

4.1 Data

4.1.1 Small-scale experiment. To show the accuracy of our method, we use a small-scale experimental data with labeled ground truth, the MIT Reality Mining data [10]. The Reality Mining project was conducted from 2004 to 2005 at the MIT Media Laboratory. The Reality Mining study followed more than 100 subjects (including students and faculty), 73 of which are usable. The researchers track the subjects by mobile phones pre-installed with software to record data about call logs, cell tower IDs, and phone status (idling or charging). The locations individuals reported include home, workplace, third place and no-signal.

4.1.2 Real-world data. The large-scale CDR data we used covers a two-month period in a populated and fast-developing city in China. One of the three mobile carriers in China provided the data. We use a sample of 100,000 mobile phone users and 217,753 user locations as a case study to test the scalability and the feasibility of our method.

We preprocess this data which are noisier than MIT Reality Mining data. We preprocess the CDR in the populated city in China in the following way. User locations are clustered in geographic coordinates using the algorithm developed by Isaacman (2011) [14]. In this method, it first ranks the cell towers according to the total number of days that they are connected. The next step is to cluster cell towers according to Hartigan’ leader algorithm with a spatial threshold of 1 km. This algorithm starts from the first cell tower in the sorted list as the center of the cluster. The subsequent cell tower is checked to see if they fall within the radius of 1 km. If it does, it groups the cell tower into the existing cluster. Otherwise, it becomes a new cluster centroid. The algorithm completes when every cell tower belongs to a cluster. Readers interested in the detailed implementation of the method are encouraged to refer to Isaacman (2011) [14].

4.2 Results analysis and comparisons
In this section, we compare the proposed method with the most widely-used method in the literature, named as the Most Frequent Appearance method on MIT Reality Mining data. As stated in the literature review section, the Most Frequent Appearance method makes the simplistic assumption that the user location with the most presences during the home time (00:00 - 08:00 and 19:00 - 24:00) are home and daytime (09:00 - 18:00) are workplaces respectively [5, 17, 25]. Two metrics are used for performance comparisons: accuracy and inference rate. We use the percentage of the correct inferences as for the accuracy measure. We compute inference rate as the percentage of inferred home, workplace and third places. We show the results comparisons in Table 1. There exists a trade-off between the accuracy and inference rate, meaning that the larger number of home/workplaces inferred, the more likely they are incorrectly predicted. In our opinion, CDR data is a large-scale dataset, and we, therefore, should prioritize accuracy more than inference rates.

The Most Frequent Appearance method identifies one home and one workplace for every subject. The accuracies are 53% and 62% respectively. K-means clustering and FCM infer 56%/58% home and 82%/84% workplaces with 90%/88% and 75%/74% accuracy. Though the inference rates are relatively low compared to the Most Frequent Appearance method, the accuracy improved considerably.

FCM has the flexibility in compromising accuracy and inference rate for different application purposes. To improve inference rate, relabeling third place whose membership is less than the median as either home or workplace, the method can identify 78% and 100% home and workplaces respectively, and the accuracies are 91% and
We first analyze the pattern of eigenlocations, demonstrating that it is possible to use the computational time to show scalability and efficiency of the method.

By interpreting the pattern, we find that the first eight eigenlocations are intuitively interpretable. With the objective of segmenting user locations, eight is then determined to be the optimal number of eigenlocations. This number is plausible since the non-interpretable patterns are more likely to be noises. The eight eigenlocations can explain 56% of the behavioral variances of user locations in total. A linear combination of the eigenlocations reconstructs the presence patterns at user locations.

4.3.2 Optimal number of cluster. In the inference of home and workplace, prior knowledge and hypothesis are there should exist four clusters, including one workplace, one third place and two home locations, one for normal-schedule workers and one for non-workers or short-distance commuters. To test the validity and stability of the optimal number of cluster, we bootstrap Davies-Bouldin (DB) index for the different number of clusters. DB index measures the scatter within the cluster and separation between clusters by the distances between each observation and its most similar ones [8, 15, 18, 24]. Accordingly, the lower the DB index, the better the cluster configuration. We show the results from bootstrapping the DB indexes in Figure 4. The y-axis and x-axis show the DB index and the cluster size respectively. Each boxplot corresponds to the distribution of DB index for one cluster size. From the figure, we can see that there is an increase in DB index when the cluster number increases from 4 to 5. The decrease in DB indexes is small when the number of clusters exceeds 5. These observations indicate that the optimum number of cluster is four, which confirms our hypothesis.

4.3.3 Clustering results. Figure 5 shows the clustering results from K-means clustering and FCM. The x-axis represents 24 hours on weekdays and 24 hours on weekends. Y-axis represents median NHP for each hour. Each line corresponds to the 48 median NHPs for one of the four clusters, including two types of home locations (red and blue), one workplace cluster (green) and third place cluster (purple).

Overall, we can see that the results from the two clustering techniques are similar. The resulting performances are in line with the actual home and workplace patterns in our daily lives. The red line represents the home cluster for non-commuters or commuters who work near home. The percentages of presences at these locations are quite high throughout the week from 7:00 to 24:00. Note that for those who work near home, it is difficult to differentiate home and workplaces due to the low spatial resolution of CDR. The blue line

![Figure 3: Top three eigenlocations](image)

![Figure 4: Bootstrapping DB index per number of clusters](image)
represents the home for regular commuters, who present at these user locations early in the morning and late at night. The presence frequencies are high before 7:00 possibly due to the automatic data fetching by some mobile Apps. The green line represents the work cluster. Commuters present at these locations more frequently during 9:00 - 20:00 on weekdays. The purple line represents a type of location clustering with infrequent and irregular presences. We also show the home and workplace distributions in the urban area in Figure 6. On the maps, we show that the population and workplace density distributions of the Traffic Analysis Zones (TAZs). We scale up the 100,000 random sample to the whole population with the values in logarithmic transformation for better visualization results. The color scales are positioned underneath the figures. From the map, we can see that home and workplace locations distributed the densest in the city center which is in line with the reality. The density of workplaces in the eastern urban area is higher than that of the home density where there is a high-tech district with many new employment opportunities created.

We acknowledge that there are some limitations produced by this assumption. The group of mobile phone users with flexible work locations, such as shippers and drivers, or workers with irregular work schedules, such as night shifters, is undetectable or prone to be misidentified. The mistake is because this group of population performs unusual pattern, which is a small portion of the population and is hard to observe from the eigenlocation reconstruction. Another limitation is that phone usage patterns can also influence the inference result. For example, if individuals use landlines instead of mobile phones at home, it is hard to estimate a home or workplace for these individuals due to the unobservable presence patterns solely based on CDR data.

4.3.4 Uncertainty in behavioral inference. The confidence of inference results can be learned from membership via FCM, which is the confidence of belonging to the group. The larger the membership, the more confident the results are. We show the membership distributions for each cluster in Figure 7. The x-axis denotes the membership, and the y-axis represents the count of user locations in each membership range. Lighter color indicates a more substantial number of user locations. The median memberships are 0.56, 0.50 and 0.94 respectively for home, workplace and third place. We can see that third place has the highest confidence due to the small observable presences at these locations. The confidence for labeling
workplaces is the lowest since people are more active during the
daytime, making the inference more difficult.

The most significant advantage of applying FCM in this setting is
the flexibility in trading-off between accuracy and inference
rate using the membership. If we increase the accepted confidence
level by setting an accepted membership threshold, we will infer
less home/workplaces. We only label home and workplaces with
membership higher-than-threshold as home/workplace. On the
other hand, if we want to improve inference rate, we can reduce
the accepted threshold to accommodate more inferred home and
workplaces.

4.3.5 Computation time. Computational complexity is an
important consideration for practical application. The computational
complexity of K-means clustering is $O(ndt)$ and that of FCM is
$O(ncdt^2)$. $n$ is the number of observations, which is the total num-
er of user locations for all sampled mobile phone users. $d$ is
the number of features, which is 48. $c$ is the pre-specified number of
clusters. $t$ is the number of iterations until convergence. We test
the method on 1,000,000 mobile phone users with 2,177,530 user
locations; each has two-month presences. The running time for
PCA is 15 seconds. The running time for K-means clustering is
approximately 6.2 seconds, and that for FCM clustering is 120.2 sec-
onds. While K-means clustering outperforms FCM in computation
time, they are both efficient, practical and scalable in practice.

5 CONCLUSIONS
The wide-penetrated mobile phone data provides longitudinal records
for tracking human mobility and urban dynamics. The low spatial
resolution and sparse sampling characteristics make this promising
dataset challenging to apply in transportation and urban planning
fields. Home and workplace, origins and destinations of commuting
and other trips, are the most crucial user locations and are the
foundations of many transportation research. However, the existing
literature, making simple, intuitive but biased assumptions, are
problematic in inferring home and workplace due to the consider-
able discrepancy between observed presence patterns from mobile
phone data and actual presence patterns at user locations.

In this paper, we propose a novel behavioral method to accur-
ately infer the point of origins and destinations of commuting
flows based on individual’s spatial-temporal patterns inferred from
Call Detail Records. Our method significantly improves in accuracy
upon the heuristic assignment rules popularly adopted in the litera-
ture. Starting with the historical data of geo-temporal travel pattern
for a panel of individuals, we create, for each person-location, a
vector of probability distribution capturing the likelihood that the
person appears in that location for a given the time of day. Stacked
in this way, the matrix of historical geo-temporal data enables us to
apply Eigen-decomposition and use unsupervised machine learning
techniques to extract commonalities across locations for a different
group of travelers, which ultimately allows us to make inferences
and create labels, such as home and work, on specific locations.

Testing the methodology on real-world data with known location
labels show that our method identifies home and workplaces with
significant accuracy, improving upon the most commonly used
methods in the literature by 79% and 34%, respectively. Most impor-
tantly, our methodology does not bear any significant computation
burden and is easily scalable with real-world Call Detail Records
data.

For future research, the proposed method can be extended to
cluster not only home and workplace locations but also other user
locations, such as late night locations or weekend locations. The
number of clusters can also be increased to recognize more types of
user locations for different applications purposes as future research.
The method is also useful to combine with other geographical
data sources, such as land use data, Point of Interests, to infer trip
purposes and activity types. Also, it can be further expanded to
estimate commuting characteristics, such as commuting distances,
departure and arrival times. With more data available, this method
can be extended to other data with longitudinal behavioral trackings
in inferring home and workplaces, such as online-social networks
check-ins (Flickr, Twitter), and bank transactions.

REFERENCES
[1] Reis Ahas, Sirri Silm, Olle Jarv, Erki Salouver, and Margus Tira. 2010. Using
mobile positioning data to model locations meaningful to users of mobile phones.
Journal of Urban Technology 17, 1 (2010), 3–27.
[2] Lauren Alexander, Shan Jiang, Mikael Muruga, and Marta C González. 2015. Origin–
destination trips by purpose and time of day inferred from mobile phone data.
Transportation Research Part C: Emerging Technologies (2015).
[3] Vincent D Blondel, Adeline Decuyper, and Gauthier Krings. 2015. A survey of
results on mobile phone datasets analysis. EPJ Data Science 4, 1 (2015), 1–55.
[4] Iva Bojie, Emanuele Massaro, Alexander Bely, Stanislav Sobolevsky, and Carlo
Ratti. 2015. Choosing the right home location definition method for the given
dataset. CoRR abs/1510.03715 (2015). http://arxiv.org/abs/1510.03715
[5] Francesco Calabrese, Massimo Colonna, Piero Lovisolo, Dario Parata, and Carlo
Ratti. 2011. Real-time urban monitoring using cell phones: A case study in Rome.
Intelligent Transportation Systems, IEEE Transactions on 12, 1 (2011), 141–151.
[6] Francesco Calabrese, Laura Ferrari, and Vincent D. Blondel. 2014. Urban Sensing
Using Mobile Phone Network Data: A Survey of Research. ACM Comput. Surv.
47, 2, Article 25 (Nov. 2014), 20 pages. https://doi.org/10.1145/2655691
[7] Francesco Calabrese, Jonathan Reades, and Carlo Ratti. 2010. Eigenplaces: seg-
menting space through digital signatures. Pervasive Computing, IEEE 9, 1 (2010),
78–84.
[8] David L Davies and Donald W Bouldin. 1979. A cluster separation index.
Pattern Analysis and Machine Intelligence, IEEE Transactions on 2 (1979), 224–227.
[9] Mi Diao, Yi Zhu, Joseph Ferreira, and Carlo Ratti. 2015. Inferring individual
daily activities from mobile phone traces: A Boston example. Environment and
Planning B: Planning and Design (2015). https://doi.org/10.1068/b36581
[10] Nathan Eagle and Alex Pentland. 2006. Reality mining: sensing complex social
systems. Personal and ubiquitous computing 10, 4 (2006), 255–268.
[11] Nathan Eagle and Alex Sandy Pentland. 2009. Eigenbehaviors: Identifying struc-
ture in routine. Behavioral Ecology and Sociobiology 63, 7 (2009), 1057–1066.
[12] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779.

[13] Sebastian Graswin, Stanislav Sobolevsky, Simon Moritz, István Göder, and Carlo Ratti. 2015. Towards a comparative science of cities: using mobile traffic records in new york, london, and hong kong. In *Computational approaches for urban environments*. Springer, 363–387.

[14] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. Identifying important places in peoplesAzEs lives from cellular network data. In *Pervasive computing*. Springer, 133–151.

[15] Anil K Jain and JV Moreau. 1987. Bootstrap technique in cluster analysis. *Pattern Recognition* 20, 5 (1987), 547–568.

[16] Shan Jiang, Joseph Ferreira Jr, and Marta C González. [n. d.]. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore.

[17] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM, 2.

[18] M Kathleen Kerr and Gary A Churchill. 2001. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences* 98, 16 (2001), 8961–8965.

[19] Kevin S Kong, Karl Greco, Stanislav Sobolevsky, and Carlo Ratti. 2014. Exploring universal patterns in human home-work commuting from mobile phone data. (2014).

[20] Yan Leng et al. 2016. Urban computing using call detail records: mobility pattern mining, next-location prediction and location recommendation. Ph.D. Dissertation. Massachusetts Institute of Technology.

[21] Yan Leng, Alejandro Nortiega, Alex‘Sandy’ Penland, Ira Winder, Nina Lutz, and Luis Alonso. 2016. Analysis of Tourism Dynamics and Special Events through Mobile Phone Metadata. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM, 2.

[22] Yan Leng, Larry Rudolph, Jinhua Zhao, and Haris N Koutsopoulos. 2017. Synergistic data-driven travel demand management based on phone records. (2017).

[23] Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. 2013. Approaching the limit of predictability in human mobility. *Scientific reports* 3 (2013), srep02923.

[24] JE McKenna. 2003. An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. *Environmental Modelling & Software* 18, 3 (2003), 205–220.

[25] Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. 2010. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding*. Springer, 14–25.

[26] Jonathan Reades, Francesco Calabrese, and Carlo Ratti. 2009. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design* 36, 5 (2009), 824–836.

[27] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. 2013. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface* 10, 84 (2013), 20130246.

[28] Lishan Sun, Liya Yao, Shouwei Wang, Jing Qiao, and Jian Rong. 2014. Property analysis on travel intensity of land use patterns. In *Mathematical Problems in Engineering* 2014 (2014).

[29] Jameson L Toole, Serdar Colak, Fahad Alhasoun, Alexandre Evsukoff, and Marta C Gonzalez. 2014. The path most travelled: mining road usage patterns from massive call data. *arXiv preprint arXiv:1403.0636* (2014).