Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore

Shan Jiang
Department of Urban Studies and Planning
Massachusetts Institute of Technology, Cambridge, USA
shanjang@mit.edu

Joseph Ferreira, Jr.
Department of Urban Studies and Planning
Massachusetts Institute of Technology, Cambridge, USA
jf@mit.edu

Marta C. González
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology, Cambridge, USA
martag@mit.edu

ABSTRACT
In this study, with Singapore as an example, we demonstrate how we can use mobile phone call detail record (CDR) data, which contains millions of anonymous users, to extract individual mobility networks comparable to the activity-based approach. Such an approach is widely used in the transportation planning practice to develop micro urban simulations of individuals' daily activities and their travel; yet it depends highly on detailed travel survey data to capture individual activity-based behavior. We provide an innovative data mining framework that synthesizes the state-of-the-art techniques in extracting mobility patterns from raw mobile phone CDR data, and design a pipeline that can translate the massive and passive mobile phone records into interpretable spatial human mobility patterns for urban and transportation planning. With growing ubiquitous mobile sensing, and shrinking labor and fiscal resources in the public sector globally, the method presented in this research can be used as a low-cost alternative for transportation and planning agencies to understand the human activity patterns in cities, and provide targeted plans for future sustainable development.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – *Data mining, spatial databases and GIS;* I.2.6 [**Artificial Intelligence**]: Learning – *Knowledge acquisition*

General Terms
Algorithms, Measurement, Experimentation

Keywords
Human mobility networks, CDR data, trajectory data mining, mobility motif detection, Singapore

UrbComp'15, August 10, 2015, Sydney, Australia.

1. INTRODUCTION
To improve urban mobility, accessibility and quality of life, understanding how individuals travel and conduct activities has been the major focus of city planners, geographers and transportation planners for more than half a century [1]–[4]. In the past, this task was mainly accomplished through collecting survey data in small sample sizes and low frequencies (e.g., planning agencies in metropolitan areas usually conduct 1 per cent household travel diary survey once or twice in a decade), due to the huge amount of labor and fiscal resources required in the data collection process. Today, this traditional approach has many challenges. First, more than half of the global population (54 per cent in 2014) lives in urban areas now, and with the evolution of our society and innovation in technology, cities have become more diverse and complex than ever before in the increasingly interconnected world. Second, urban problems in traffic congestion, environmental pollution and degradation, energy consumption, and green house emission become major threats especially for the rapidly growing cities in developing countries. It is projected that additional 2.5 billion urban population will be added by 2050 and 90 per cent of the increase will be concentrated in Asia and Africa [5]. Third, recent financial crisis has brought substantial stress to the public sector in collecting the traditional survey data to understand cities, the economy and the society. For cities in developing countries, governments usually don't have resources in collecting expensive data.

Despite that urban and transportation theories can go beyond data, analytical methods and modeling approaches in the past were developed from and tailored to the traditional type of data, and cannot meet the current challenges. It is increasingly urgent for urban researchers to look for new approaches to address these challenges, especially for cities in the rapid urbanization process. With the rise of the ubiquitous sensing technologies, digital human footprints can be recorded in unprecedentedly massive scale with high frequency and low costs. It brings great opportunity to change the landscape of urban research to a new horizon (e.g., [6]–[8]), and requires innovation to link the new massive and passive data with urban theory and thinking to derive new urban knowledge called the new urban science [9][10].

In this paper, we use Singapore, a city state as an example, to demonstrate the pipeline of parsing, filtering, analyzing and synthesizing mobile phone data to interpret human mobility in a way that can be used and understood in the planning language and

for planning practices. We argue that by properly treating the mobile phone data, planners can derive much richer information for understanding how individuals use different parts of the city and travel, and where to target for future improvement of urban and transportation planning.

The rest of the paper is organized as follows. In Section 2, we present the spatial unit of analysis and data, including mobile phone call detail records (CDR) data and census data for mobility analysis, and household travel survey data for validation purpose. In Section 3, we introduce the data mining pipelines to extract statistically reliable estimates of individual mobility networks from the massive CDR data. In Section 4, we present the spatial patterns of human mobility networks at the transportation planning zone level and interpret the mobility patterns in the urban context. In Section 5, we discuss the planning implications of the finding and future improvement for urban development.

2. STUDY AREA AND DATA

Singapore is a city state— it has landed area of 718.3 square kilometer (sq km), with a range of around 43 kilometer in the west-east direction, and 23 kilometer in the north-south direction. It has a total of 5.18 million population in 2010, of which 3.79 million are Singapore residents (including 3.26 million citizens and 0.53 million permanent residents), and 1.39 million non-residents. It has a mobile penetration rate above 150% (using total population as base), one of the world's highest[1].

2.1 Call Detail Record Data

Call detailed record (CDR) data are generated by the phone service carrier for billing purposes, and record the time and cellular tower information of the phone communication activity (including messages, calls, etc.). As wireless mobile connectivity has changed the way people communicate, work, and play, phone data can be used to derive the spatiotemporal information of anonymous phone users' whereabouts for analysis of their mobility pattern. We use 2 consecutive weeks (i.e., 14 days) of mobile phone CDR data (in March and April of 2011) from one carrier in Singapore to examine the mobility patterns of anonymous individuals in the metropolitan area. The data set of the studied period contains 3.17 million anonymous mobile phone users, and a total of 722.92 million records of phone usages. There are more than five thousand cellular towers in Singapore, with a spacing gap of around 50-meters in the dense downtown area to a few kilometers in the suburban region. In general, the cellular tower network has a very high density covering the whole metropolitan area (see Figure 1).

2.2 Census Data and Geographic Zones

Despite the high mobile phone penetration rate, the CDR data, which were obtained from one carrier for this study, only count for 63% of the total population in Singapore. To get statistically meaningful measurements from the CDR data for the whole population, we need to scale up the phone users—therefore census data with spatial information is useful. Besides the one-category total population, Singapore census data includes population by different sociodemographic groups at the Development Guide Planning zone level (which is the planning district in Singapore). However, population data with geographic zonal information are only available for Singapore residents (not for non-residents). We then scale the phone data at the DGP level to the Singapore residents. We assume that an individual older than 10 years old can possess a mobile phone, and we use the population in this age category for the scaling of phone users.

DGP is a spatial unit of planning area used by the Singapore Urban Redevelopment Authority (URA). There are 56 DGPs in Singapore with sizes ranging from 0.85 sq km, to 66 sq km. 35 of these 56 DGPs have residential population residing in them, and the rest are either industrial zones or reserved land for other purposes (e.g., natural reserve, military reserve, etc.) These DGP zones can be further subdivided into finer-grained spatial units, the transportation analysis zones (MTZs); however, population data at the high spatial resolution (MTZ) level is not available in the census data. Compared with the DGPs, MTZs are closer to the neighborhood level, and are used for transportation planning purposes by the Singapore Land and Transportation Authority (LTA). A total of around 1100 MTZs cover the whole metropolitan area, and their sizes range from 0.015 sq km to 43 sq km. In Figure 1, we illustrate the population density at the DGP level, and demonstrate the spatial relationship of DGP, MTZ and cellular towers. With the CDR data and population data at the DGP level, we will be able to scale the phone users at the tower level to match the total population at the DGP level, and estimate population at the MTZ level, which is useful for transportation planning purposes. We discuss the scaling details in Section 3.

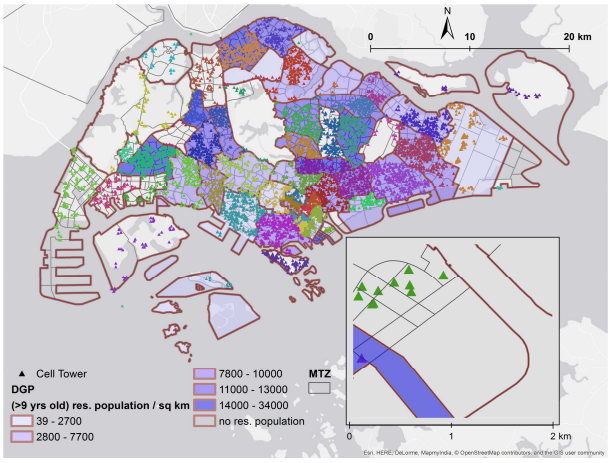

Figure 1 Singapore Development Guide Planning (DGP) zones, Transportation Analysis Zones (MTZ), and cellular tower distribution

2.3 Survey Data

We use the Singapore 2008 Household Interview Travel Survey Data (HITS) collected by the Singapore Land Transport Authority (LTA) for validation purpose to examine and compare the results derived from CDR data. Compared with CDR data, traditional household travel survey data are often in very small sample sizes, and come with expansion factors for the purpose of scaling up survey samples to total population. Survey expansion factors are usually derived from sampling design and individuals' sociodemographic characteristics. Survey data can include detailed information on household, and individual's social demographics and travel records self-reported by survey respondents. The Singapore HITS 2008 data include around 34000 individuals and their 1 day travel (such as departure, and arrival time for each trip, trip purpose, location information of the destination, etc.), besides social demographics of the individual and their household. In this study, we use the time and location information of the survey respondents' travel, and expansion

[1] Reported by the Info-communications Development Authority (IDA) of Singapore. http://www.ida.gov.sg/.

factors provided by the LTA to scale up the sampled individuals to the whole population to compare with our analysis derived from the CDR data.

3. METHODOLOGY

Human daily travels are organized base on activities and anchor locations that are important in their daily life. From this point of view, the activity-based model— the state-of-the-art approach in transportation planning—considers travel demand as the derived needs to conduct activities [11]–[14]. To understand human mobility patterns at the metropolitan level for transportation and city planning purposes, in this study, we synthesize methods in previous research [15]–[17], and develop a pipeline (see Figure 2) of (1) parsing the CDR data to extract stay locations of phone users; (2) detecting phone users' home location; (3) filtering users and select statistically representative samples from the parsed CDR data; (4) identifying the daily mobility networks of the phone users; (5) deriving expansion factors for the filtered phone observations by combining the preprocessed phone data and the census data; scaling up population, trips, and daily motifs; and aggregating them from tower to transportation analysis zones.

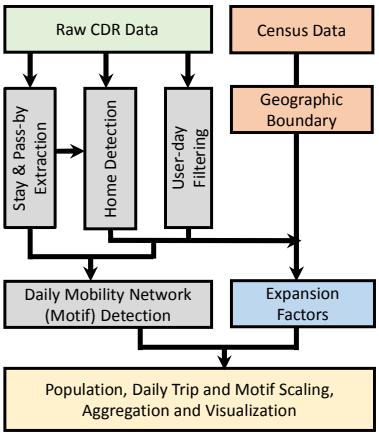

Figure 2 Workflow of population estimation and mobility pattern extraction from the CDR data

3.1 Parsing Trajectory to Extract Stays

Parsing the passive mobile phone CDR data to extract anchor points (the stay locations) of an individual's daily travels and differentiate destinations from pass-by points is the first part in the pipeline that can lead to identifying individual's mobility networks. Zheng [18] presents a comprehensive overview on the research topic of the trajectory data mining. Inspired by the algorithm first introduced by Hariharan and Toyama [19] which was tested previously using triangulated mobile phone data [15], [17], we modified the algorithm for the tower-level CDR trajectories in this study. Other algorithms on stay point detection can also be found in the literature[20].

Three goals need to be achieved in the stay detection process. (1) Eliminating noises of the user's geospatial location records in identifying his/her true location is important in extracting the user's accurate stays. Two types of noises— outliers and signal jumps between towers—exist in the raw CDR data. We applied an outlier detection algorithm (based on the time intervals and distances between consecutive points) to eliminate noises. (2) In the time-stamped location sequence, we need to cluster points that are spatially close (within the threshold of Δd) and temporally adjacent (in the sequence ordering) into a single location, with a collapsed time duration inferred from this clustering (with either one or multiple locations) as the stay. (3) Since we are interested in the unique stay locations that a user visits daily to eventually form his/her mobility network, we need to agglomerate points that are spatially close but not necessarily adjacent in temporal consecutive sequence into one unique location.

- Considering a user i's location recorded in the CDR in sequence as $D_i = (d_i(1), d_i(2), d_i(3), \dots, d_i(n_i))$, by selecting a roaming distance Δd_1 (300 meters in this study) as the threshold, we heuristically cluster spatially close locations within Δd_1 for each point into the medoid (which is the point in a set that minimizes the maximum distance to every other point in the set) and form a new sequence $D'_i = (d'_i(1), d'_i(2), \dots, d'_i(n'_i))$. Where $d_i(k) = (tower(k), t(k))$ for $k = 1, \dots, n_i$, $tower(k)$ and $t(k)$ are the tower id, and time stamp of the user i's k-th observation in the raw CDR data. $d'_i(g) = (cellid(g), \mathrm{t(g)}, \mathrm{dur(g)})$ includes the tower id, starting time of the first observed tower in the cluster, and the duration that the user stayed in the clustering.
- We then further cluster the points in the trajectory sequence set D'_i based on the distance threshold of Δd_2 (300 meters for this study), and keep the points whose duration is greater than the time threshold Δt (10 minutes in this study) as the final stay points $S_i = (s_i(1), s_i(2), \dots, s_i(m_i))$, where $s_i(m) = (cellid(m), \mathrm{t(m)}, \mathrm{dur(m)})$. Using such a method, we eliminate the outliers in the clustering process. We present the input and output of the process in Figure 3 for demonstration.

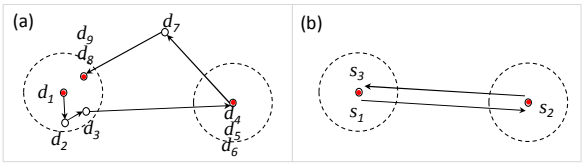

Figure 3 Stay point detection

3.2 Detecting Home

It is important to detect the phone user's home location (tower) for several reasons. First, when we later scale phone users to population, we need to combine the CDR data with census data, in which population are measured by their residential location. Second, in order to form the mobility networks we make an assumption that for each day, an individual always departures from home, and returns to home by the end of the day. Although in the real world there are exceptions, this assumption simplifies the way that we can extract mobility networks from the CDR data and estimates trips in a justifiable way. Since CDR data passively records phone users' spatiotemporal information, it does not always give complete information of an individual's whereabouts. For example, a user may not use her/his phone at home for a particular day, but we can identify her/his home from the entire observation period. Third, for planning purposes, it is important to understand people's mobility patterns from their residential location, as the built environment and land use of individuals' residence influence their travel and activities [21], [22]. Being able to identify human mobility patterns associated with urban space will enable planners to target improvement for future urban development.

- A phone user's home tower is identified as the most frequently communicated tower during nights of weekdays, and the entire days of weekends over the study period (i.e., 14 days in this study).
- We define night time from 7 pm to 7 am. Note that this definition of 'night' is a parameter that can be adjusted in different urban contexts. We select this time period based on the local life and culture in Singapore, where people have a relatively rigid work schedule.

After processing the CDR data, we obtained 2.88 million users whose home towers can be observed in the 2-week data. It counts for 91% of all phone users contained in the data for the study period, and around 56% of the total population. By properly scaling up these users, we can disaggregate population from aggregated planning area level, to the neighborhood transportation analysis zone (MTZ) level with high spatial resolution. We discuss the population scaling method and results in section 3.5.

3.3 Filtering Phone User-Day Samples

As discussed previously, CDR data are passive records of human footprints when their mobile phones communicate with cellular towers. Compared with traditional survey data, the advantage of the CDR data is its longer observation period (in this case, 14 days), and larger sample size (e.g. 63% population in this study); and its disadvantage is that not every phone user necessarily uses her/his phone at every destination that s/he visit, which means CDR data may not always reveal individuals travel as what travel dairy or GPS tracking application data can do. Therefore it is important to carefully select the sampled days when a phone user uses his/her mobile device frequently enough that we can use his/her parsed trajectory derived from the CDR data on that day as a proxy to the travel dairy data. Previous research [16] has found that if a phone user's phone communication activities can exceed certain threshold, the CDR data can be treated equivalently as survey data, and can generate results that are statistically consistent with travel dairy data in terms of individual mobility pattern. In this study we adopt similar rules.

- We only keep a user-day observation, if a user has phone communication activities in at least 8 of the 48 half-hour time slots in a day (24 hours).
- We only keep the observations during the weekday, as mobility patterns can be different on weekdays and weekends.

After the filtering, we keep 6.28 million user-days for 1.55 million users in the 10 weekday days of the two weeks. The total users after filtering are around 49% of the original users contained in the raw CDR data for this time period. We plot the frequency distribution of user-days per user after data filtering in Figure 4 (c). We can see that, after the filtering, for these 1.55 million users, 22% of them only have 1-day observation; 15% have 2-day; 13% have 3-day; 11% have 4-day; 10% have 5-day; 8% has 6-day; 7% has 7-day; and 3% has 10-day observations. On average, each user has around 4 weekday observations. This is a relatively good sampling rate— especially if we can have longer observation period, the number of user-days for each user will be longer, and information on the long term trend could be detected.

We summarize the user-day statistics at the tower level based on phone users' home tower. We see in Figure 4 (e), (b) and (d), the distribution of the average and standard deviation of the number of user-days per user at the tower level. We plot the average number of user-day at the tower level in Figure 4 (a). We can see that in the downtown, the airport, and the west-end of the coastal area, the number of user-day per user is relatively low. Presumably, tourists and foreigners live in these areas. While in residential areas in the suburb, such as Ang Mo Kio, Bukit Batok, Jurong West, Punggol, Sengkang, Tampines, Yishun, and Woodlands, etc. the average numbers of user-days are relatively higher than average.

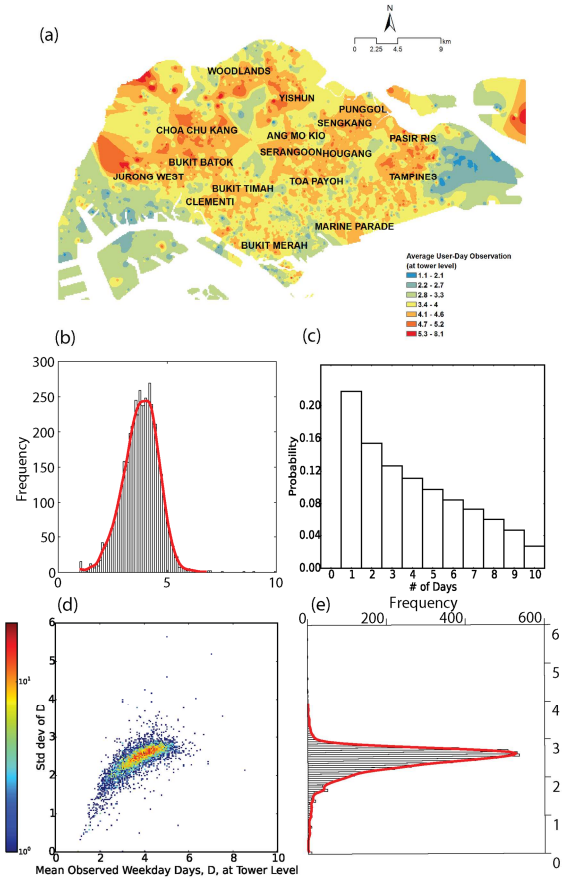

Figure 4 User-day statistics after sample filtering: (a) spatial distribution, frequency distribution of (b) mean number of day- and (e) standard deviation of day-observations at the tower level, (c) frequency distribution of total number of days for each user, and (d) density distribution of (b) and (e).

3.4 Identifying Activity-Based Mobility Patterns

Human daily mobility can be highly structural—organized by a few activities essential to life. It is important to identify individual mobility networks (i.e., daily motifs) for urban and transportation planning purposes. The most contemporary travel demand models developed and employed by regional transportation agencies are the activity-based approach, and self-reported survey are employed in collecting information on individuals' socio-demographics, their daily activity purposes, time, and location. Readers can refer to the comprehensive TRB report [14] for more detailed explanation on activity-based travel demand models used in the transportation planning filed.

Figure 5 (1a-1f) show some examples of individual activity pattern structure (daily tours) commonly seen in the real world and represented in the transportation modeling language. They are categorized as (1a) 1 home-based work tour; (1b) 1 home-based

work tour with a third destination, and work as the primary destination; (1c) 1 home-based work tour with 1 home-based sub-tour; (1d) 2 home-based work tours with a third destination; (1e) 1 home-based work tour with 1 work-based sub-tour; and (1f) 1 home-based work tour with 1 escort tour (to drop off and pick up somebody). These activity pattern structures can go into numerous combinations; however, if we use a more abstract diagram to represent these activity-chains, they can be reduced into daily motifs that can be captured by the massive and passive mobile phone data. Figure 5 (2a-2f) represent the real world activity pattern structure of (1a-1f) in a highly abstract format, which are called *daily motifs*. These daily motifs have been proposed and tested as measures of human mobility in previous studies for Paris, Chicago and Boston [15], [16], and have been found comparable to traditional travel survey in aggregate statistics.

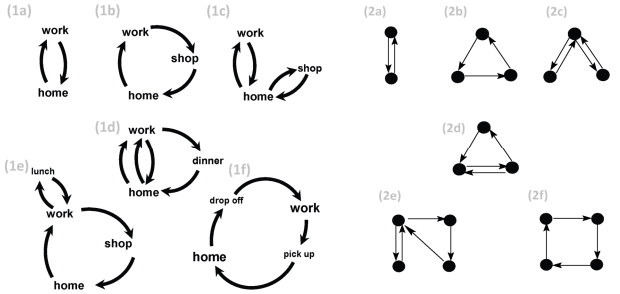

Figure 5 Examples of daily mobility networks (1a-1f) in the real world, (2a-2f) in abstract

Here, we present the algorithm to identify the daily motifs for the filtered phone users whose mobile phone records in a sampled day are frequent enough (explained in section 3.3) for representing their daily activity pattern. Formally speaking, a *motif* is an equivalence class of directed graphs. A directed graph is an ordered pair $D = (V, A)$ where V is the set of nodes (or vertices), and A is a set of ordered pairs of nodes (i.e., directed edges). Without loss of generality, when V is a set of n elements, we take V to be the set $V_n \triangleq \{1,2,\dots,n\}$. Let $\mathfrak{D}_n$ be the set of directed graphs with n nodes, i.e., $\mathfrak{D}_n \triangleq \{D_n | D_n = (V_n, A), A \subseteq V_n \times V_n\}$. For any mapping $f: V_n \to V_n$, it induces a mapping $(f \times f): V_n \times V_n \to V_n \times V_n$ such that $(f \times f)((a,b)) = (f(a), f(b))$. Then an equivalence relation $\sim$ on $\mathfrak{D}_n$ is defined as follows: $D_n = (V_n, A) \sim D'_n = (V_n, A')$ if and only if there exists a bijection $f: V_n \rightarrowtail V_n$ such that $(f \times f)(A) = A'$. Intuitively speaking, $D_n \sim D'_n$ if the "graphs" of D_n and D'_n appear the same when the labelling of nodes is ignored. Thus, a motif $[D_n]$ (i.e., the equivalence class of D_n) can be visualized by D_n.

With the extracted stay locations, including observed and potential stay locations (described in section 3.1), users' identified home location (discussed in section 3.2), we will be able to identify the mobility networks as illustrated in Figure 5 (2a-2f). We find that not all users necessarily have the first stay-point (or potential stay-point) starting from their home tower, or the last stay point (or potential stay-point) ending at their home tower every day in the filtered trajectory, due to the passive nature of the CDR data. In order to form a complete tour for each user, we make the assumption that, if a user travel in a day, s/he starts the first trip from, and ends the last trip at home.

- From the extracted stay points and filtered day(s), we obtain a sequence of destinations on the sampled day *k*, for a given anonymous user *i*, denoted by $S_{ik} = (s_{ik}(1), s_{ik}(2), \dots, s_{ik}(m_{ik}))$.
- We also detected the user *i*'s home h_i.
- We always check if for day *k*, user *i*'s first and last activity locations are h_i. If not, we add home h_i to the stay point sequence, which forms a new sequence $S'_{ik} = (s'_{ik}(1), s'_{ik}(2), \dots, s'_{ik}(m'_{ik}))$.
- We count the total number of trips, l_{ik}, for user *i* on day *k*.
- We count the total number of distinct nodes, n_{ik}, from trajectory S'_{ik}.
- We form a network (directed graph) D_{ik} with n_{ik} nodes: using an $n_{ik} \times n_{ik}$ matrix M_{ik}, we represent the mobility network for user *i* on day *k*. If there is at least one trip from node *o* to node *d*, $M_{ik}(o,d) = 1$; otherwise, $M_{ik}(o,d) = 0$.
- We obtain the motif $[D_{ik}]$, i.e., the equivalence class of D_{ik}.

3.5 Scaling Up Phone Sample to Population

To derive estimates of trips and daily mobility patterns (motifs), and robust spatial patterns for the population in the metropolitan area, and informative for city planning and transportation planning, we need to sale the phone samples (including users and user-days) properly. In this section, we describe the process in generating the expansion factors for the phone observations.

3.5.1 Expansion Factors

Two types of scaling factors are generated, including (1) user expansion factor, and (2) user-day expansion factor. Figure 6 presents the entity relationship diagram useful in generating these expansion factors. After detecting phone users' home tower, we store the users' information in a database table 'user', with their anonymized id number that is unique to each user. Home towers of 2.88 million users were detected. We store the mobility patterns of the filtered users who had frequent phone communication activities (in at least 8 half-hour time slots in a day) in the table 'user_day_motif', in which each row represents a unique user-day observation, with user id, motif id, and number of trip information. 6.28 million user-day records are included in this table for 1.55 million unique users whose home tower information can be found in the 'user' table. The rest of the 1.33 million users, who are in the 'user' table but not in the 'user_day_motif' table, are those whose phone communication activities are not enough such that their trajectories revealed by their phone records can be treated similar to what the self-reported travel diary may reveal.

As exhibited in Figure 6, the 'user' table is linked to the 'cell_tower' table through the cell_tower_id of the users' home_tower_id. By the spatial operation of joining cell towers to the MTZ zones that they fall into, each cell tower is associated with an MTZ. For those towers which are not contained in any MTZs, the nearest MTZ within a 200 meter distance is used to allow for the measurement error in spatial representation of zone and tower location. The 'cell_tower' table is linked to the 'mtz_taz' table (through mtz_id), which is then linked to the 'planning_area' table (through dgp_id), including 56 DGP zones of which 35 contain residential population.

Two sets of expansion factors are considered. One includes the user expansion factors calculated for the 2.88 million users with identifiable home location; the other includes user expansion factors for the 1.55 million users and the user-day expansion factors for the 6.28 million user-day observations (from the 1.55 million users). For the former case, the user expansion factor for every user with home towers in the same planning area is the same. For the latter case, as each user may have observations for more than one day, we normalize the expansion factor by the total number of observation-days for this user. We define the user

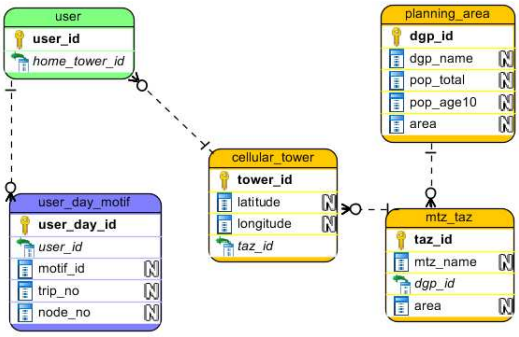

Figure 6 Entity relation diagram of the processed phone user, user-day mobility data from CDR, and geographic spatial units: towers, transportation analysis zones (MTZs), planning area (DGPs) with census population.

expansion factor β_i for user i, and the user-day expansion factor θ_{ik} for user i on day k as follows.

- $\beta_i = P_g / \sum_1^{Q_g} U_{h_g}$, for user i with a home tower h_g. Where P_g is the population (above 10 years old) in DGP zone g reported in the census; U_{h_g} is the total number of users whose home tower is at h_g in DGP zone g, and Q_g is the total number of towers in zone g;
- $\theta_{ik} = \beta_i / K_i$, where K_i is the total number of days when the user i's motifs are identified.

In Figure 7 (a) we present the spatial distribution of the user scaling factors at the tower level across Singapore, color coded by categories in different standard deviation of the scaling factor. We can see that the user scaling factors are higher in the north part of Singapore, meaning that phone user to the population ratio in this region is lower than the city average. However, when looking at Figure 4 we see that for an average phone user in this north region, s/he tends to have relatively more days of observations. In the central region, the user-scaling factors are lower, meaning that the phone user to population ratio in this region is higher than the city average. Figure 7 (b) and (c) present the frequency distribution of user expansion factors (for the 1.55 million phone users) and user-day expansion factors (for the 6.28 million user-day observations), respectively.

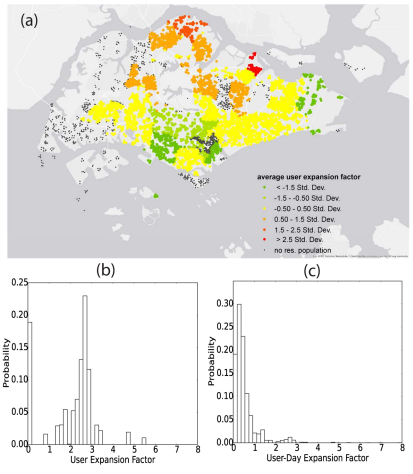

Figure 7 (a) Spatial distribution of user-expansion factors at the tower level, and frequency distribution of (b) user-expansion factors, and (c) user-day expansion factors.

3.5.2 *Population Estimation in High Resolution*

With the estimated scaling factor for each phone user in the study, we can then estimate population with high spatial resolution that are not available in the census statistics but are useful for the transportation and urban planning. Figure 8 (a) presents the estimated residential population density (for people older than 10 years old) in Singapore at the MTZ level (with more than 1100 zones) by using the phone CDR data. To compare the population estimates from users who only have home location detected (denoted as Estimation 1), with the estimates from the filtered user whose mobility patterns can be identified (denoted as Estimation 2), we plot the two population estimates at the tower level, and at the MTZ level in Figure 8 (b) and (c) respectively. We can see that with proper treatment of scaling, by only using the frequent user-observations (albeit it has a smaller sample size), it can provide comparable estimates to that by using all user samples whose home location can be detected without restricting their phone communication frequency observations.

The benefit of using the filtered phone users based on their frequency of phone communication is that we can mine richer mobility patterns from their records, and treat these observations as travel survey in a much larger expanded scale, to inform sound city planning and make wiser decisions on public policies related to urban and transportation development.

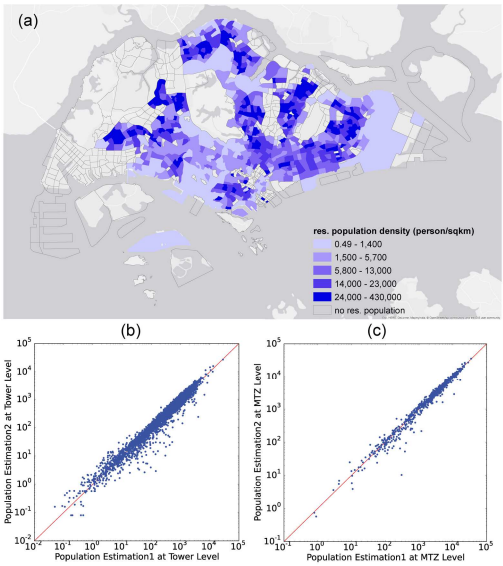

Figure 8 (a) Residential population density estimates at the MTZ level; (b) comparison of population estimation from the sample with all users with detected home (a.k.a. Estimation 1), and filtered users with complete motility networks (a.k.a., Estimation 2) at the tower level, and (c) at the MTZ level.

3.5.3 *Aggregating Daily Trips and Mobility Patterns*

After applying the expansion factors to the user-day observations, we aggregate the daily trips and mobility patterns to Singapore residents (above 10 years old) at the metropolitan level, and inferred human mobility patterns for Singapore. Figure 9 presents the inferred frequency distribution of (a) daily mobility patterns, (b) the number of daily trips, and (c) the number of unique daily

destinations that phone user visited during weekday. In Figure 9 (a) we also compare the inference results obtained from the phone data with the estimates from the HITS travel survey data. From the phone data, We find that on an average weekday, 13.5% Singapore residents stayed at home, 33% visited 2 queue places, 30% 3 places, 14% 4 places, 5.5% 5 places, 2.1% 6 places, and less than 2% visited more than 6 places. These patterns cover around 90% of survey respondents' travel.

When comparing the estimates of daily motifs with survey data, in Figure 9 (a) we find that the motif with 2-nodes reported in the survey is dominant (55%), and much higher than those observed from the phone data. We found the following reasons for explanation. (1) The HITS survey only asks for motorized travel, and excludes non-motorized travel. (2) People tend to under report their secondary destinations to the primary destinations in the survey than the phone could detect. If we refer to travel survey data in other metropolitan areas such as Chicago, Boston, and Paris, we found that the 2-node travel patterns are always less than 40% [15][16].

A separate study [23] that uses smart phone based survey to track and record individual users' travels and activities in Singapore conducted by the Future Mobility Survey (FMS) project confirms similar observation. Even though the FMS project only has 387 individuals who answered both the paper-based survey and validated their travel in the smart-phone app based tracking survey, it is found that people significantly under report their travel in the paper-based survey. 76.3% people in the paper-based survey reported 2 trips, while only 20% were detected and verified that they had 2 trips in the smartphone based survey. Therefore, the mobility patterns inferred from the phone CDR data here can be more reasonable than the travel survey data. This difference will have a significant impact on policy implications for urban and transportation planning.

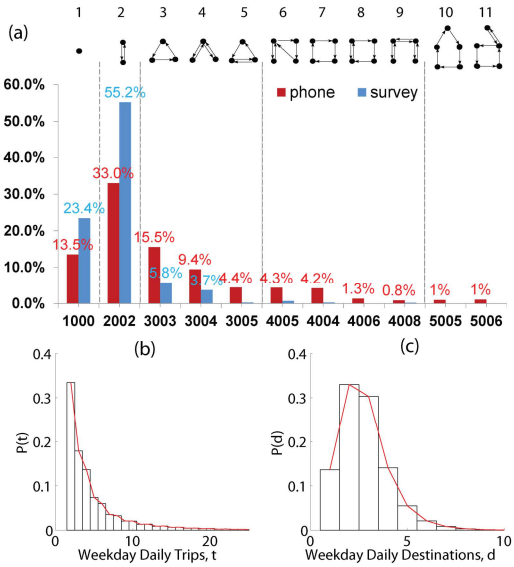

Figure 9 (a) Daily motifs from phone and survey data, and (b) frequency distribution of daily trips, and (c) activity destinations.

4. SPATIAL PATTERNS OF ACTIVITY-BASED HUMAN MOBILITY

Understanding the spatial patterns of human mobility aggregated at their home location is especially important to help planners understand how their neighborhood and/or the built environment may influence their travel. A vast body of literature in planning [24] has tried to test this theory empirically using traditional travel survey data. With the method presented in this study, we can derive new evidences on people's travel behavior in a large scale, and facilitate planners and policy makers to identify areas with great potential for improvement, such as provision of transportation alternatives to reduce motorized travel, improvement of community facilities to targeted population. In this section, we present analytical indicators to measure the spatial distribution of the mobility patterns from the perspective of individuals' home location.

4.1 The 2-node Home-based Tour

Figure 10-b presents the population density of residents with only 2-node motifs. These residents had 1 place besides 'home' that are important in their daily life. It could be workplace for workers, school for students, or a social/recreational place for the non-workers/non students. Since they are the majority of the population (around 33%, see Figure 9-a), the spatial distribution of the group with 2-node motifs is very similar to the spatial distribution of total population (shown in Figure 8-a).

We use the concept of Location Quotient, to compare an area's distribution of residents by mobility type to the distribution of a reference area (i.e. the Singapore metropolitan area). To be more specific, for example, to measure how residents with daily motif m are concentrated in a particular neighborhood (at the MTZ level) compared to the population in the metropolitan area, we use the Location Quotient (LQ_m), detailed as follows.

- $LQ_m = (e_m/e)/(E_m/E)$, where e_m is the residential population with daily motif type m (Figure 8-a) in a zone (e.g., MTZ); e is the total residential population in zone; E_m is residential population with daily motif type m in the metropolitan area (Singapore); E is the total residential population in the metropolitan area.

From Figure 10 (a) and (c), we can see that several suburban neighborhoods have a unique identity with a greater than metropolitan average share of residents whose daily mobility networks are of 2 nodes. These areas include zones in the West region, such as Jurong East, Jurong West, Clemnti, Bukit Batok, Choa Chu Kang; zones in the North-West region, including Woodlands, Sembawang, and Yishun; zones in the North-East region such as An Mo Kio, Sengkang, and Punggol; and zones in Bedok of the East region (see Figure 10-a for zone name references). People who only visit 1 other place besides their home (for work/school as the majority) are overly concentrated in these zones, which are new towns and far away from the city center. It would be potentially important to provide public transportation services with high capacity and frequency during the peak-hour to relieve congestion for residents in these neighborhoods. Meantime it is also important to plan more localized economic and community services in these neighborhoods to help reduce long distance travel for the needs of additional activities.

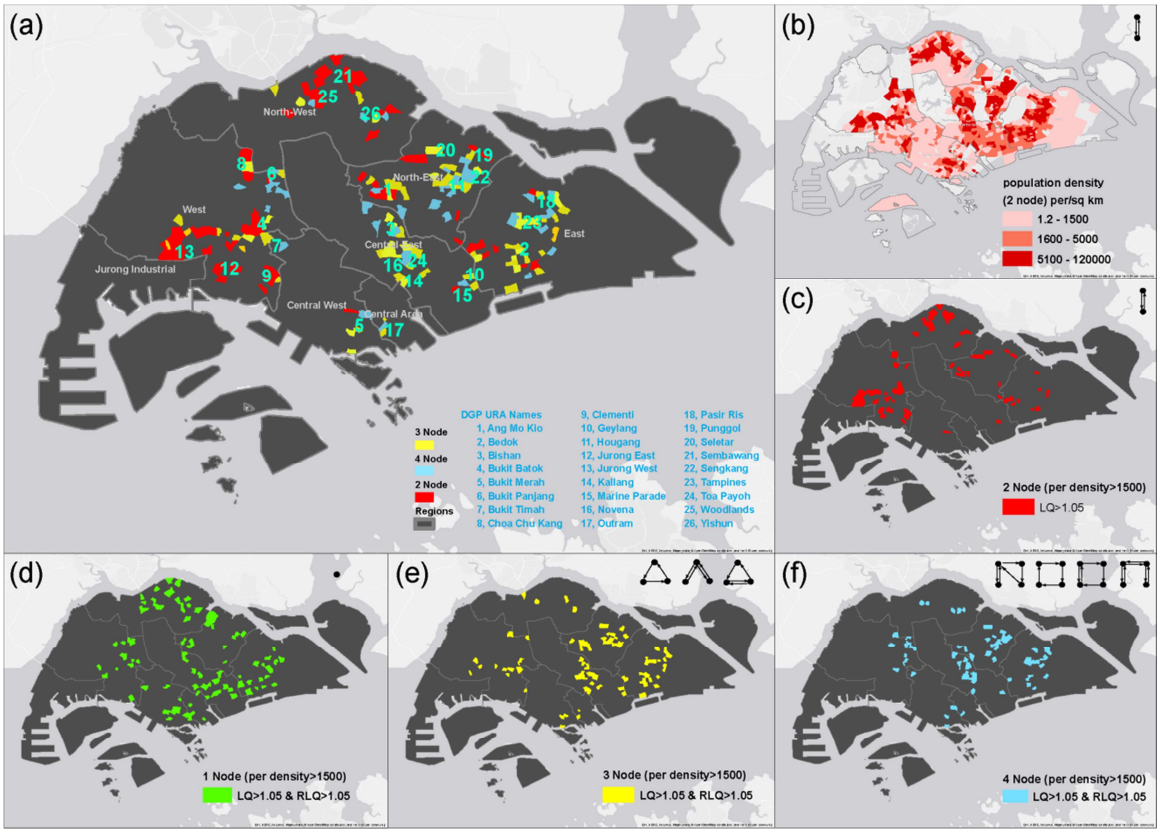

Figure 10 Highly concentrated areas with residents of (c) 2-node, (d) 1-node (stay-at-home), (e) 3-node, (f) 4-node, and (a) overlay of 2-node, 3-node, and 4-node mobility networks; and (b) population density of residents with 2-node motifs. (Note: aggregation based on phone users' home location)

4.2 The 1-node, 3-node and 4-node Motifs

As the 2-node home-based tour pattern (type 2 in Figure 9) is the majority of mobility patterns in the metropolitan area, we also use it as a reference, and compare other types of mobility patterns to the spatial distribution of the 2-node this pattern, and we call it the Relative Location Quotient, $RLQ_d = (LQ_d)/LQ_2$, for motif type d. In Figure 10 (d-f) we highlight MTZs that have high concentration of stay-at-home residents, and residents with 3-node, and 4-node in their daily activity motifs. For the space limitation here, we collapse the spatial distribution of the different motifs with 3-node, and with 4-node into one map for each in Figure 10 (e) and (f). We only show areas with population (for the corresponding type of motif) greater than 500 persons, and population density (for the corresponding type of motif) higher than 1500 person per sq km. We also only show areas that their Location Quotient, and Relative Location Quotient (to 2-node) exceed 1.05 (to allow for 5% error), to illustrate the highly concentrated areas with the corresponding types of daily motifs for residents living in the zones. In Figure 10 (a) we overlay the highly concentrated zones of residents with 2-node, 3-node, and 4-node in different colors, to illustrate the overall patterns of mobility of Singapore residents. We found that weekday stay-at-home residents are highly concentrated in zones in the Central region and along the east coast, and in zones of the North-west region, and some areas in the West region. In the Central region, it contains more mature neighborhoods where senior residents have higher probability to reside.

In Figure 11, we provide maps that zoom into the neighborhoods containing highly concentrated residents with 3- and 4-node motifs demonstrated in Figure 10 (a). We found that these areas include suburbs that are connected by light rail (grey lines in the figure) to the major transit lines, such as Bukit Panjang, and Bukit Bakok in the West region (Figure 11-a), and Sengkang in the North-East region (Figure 11-b), where accessibility are relatively

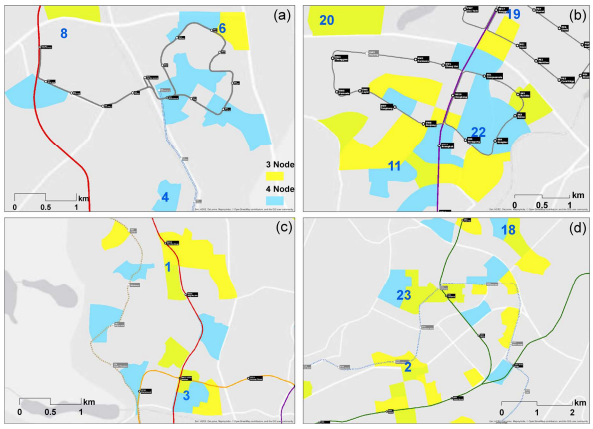

Figure 11 Neighborhoods with high concentration of residents with 3-node and 4-node daily motifs (with zone reference names in Figure 10-a)

low. It also include areas that are currently not served by the transit system, such as Lentor, Mayflower, Bright Hill and Upper Thomason, in Ang Mo Kio and Bishan (Figure 11-c), and neighborhoods in the north part of Bedok and Tempines (Figure 11-d). Neighborhoods that have higher concentration of residents of 3- and 4-node motifs usually have more trips. By providing more convenient public transportation serveries, improved level of services (e.g. for the light rail), and alternative transportation modes can help significantly reduce the total amount of travel. It's worth noting that in some of the areas, the Singapore LTA has provided plans to improve transit network (and some of them are under construction as depicted in grey dashed line in Figure 11-c and d).

5. CONCLUSIONS

Sustainable urban and transportation planning depends greatly on understanding human mobility patterns in the metropolitan area, and in this study by synthesizing state-of-the-art data mining techniques, we present an integrated pipeline that can parse, filter, and scale up the raw passive and massive mobile phone CDR data to extract human mobility patterns for millions of anonymized residents in an metropolitan area, and translate the knowledge into planning-interpretable languages. By understanding the different patterns of human mobility at the neighborhood level, we argue that transportation policies can be more targeted to the local residential neighborhoods which need the most attention of mobility improvement. Such areas are usually not connected directly or conveniently by mass transit system with high level of services. To reduce total number of travel and total mileage of motorized travel, the identified neighborhoods need to be further examined for policy incentives. For example, as city and transportation planners are testing scenarios for future mobility alternatives, the problematic areas identified through the method designed in this study, can be prioritized as testing zones to improve mobility and livability of the city for programs such as autonomous vehicle, and ride sharing.

6. ACKNOWLEDGMENTS

This research was in part funded by the Singapore National Research Foundation (NRF) through the Singapore-MIT Alliance for Research and Technology (SMART) Center for Future Urban Mobility (FM) and by the Center for Complex Engineering Systems (CCES) at KACST. We thank data providers for research collaboration through SMART.

7. REFERENCES

[1] K. Lynch, *The Image of the City*. Cambridge, MA: The MIT Press, 1964.

[2] T. Carlstein, D. Parkes, and N. Thrift, "Human activity and time geography," 1978.

[3] T. Hägerstrand, "What about people in regional science?," *Pap. Reg. Sci.*, vol. 24, no. 1, pp. 7–24, 1970.

[4] S. Jiang, J. Ferreira, and M. C. González, "Clustering daily patterns of human activities in the city," *Data Min. Knowl. Discov.*, vol. 25, no. 3, pp. 478–510, Apr. 2012.

[5] M. Batty, "Cities as Complex Systems: Scaling, Interactions, Networks, Dynamics and Urban Morphologies," in *Encyclopedia of Complexity and Systems Science*, vol. 1, R. Meyers, Ed. Berlin, DE: Springer, 2009, pp. 1041–1071.

[6] "Data for Development (D4D) Challenge." [Online]. Available: http://d4d.orange.com/en/Accueil.

[7] J. Wakefield, "Mobile phone data redraws bus routes in Africa," *BBC*, 01-May-2013.

[8] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1082–1090.

[9] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban Computing," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014.

[10] M. Batty, *The New Science of Cities*. MIT Press, 2013.

[11] M. A. Bradley, J. L. Bowman, and B. Griesenbeck, "Development and application of the SACSIM activity-based model system," in *11th World Conference on Transport Research*, 2007.

[12] J. Hao, M. Hatzopoulou, and E. Miller, "Integrating an Activity-Based Travel Demand Model with Dynamic Traffic Assignment and Emission Models," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2176, no. -1, pp. 1–13, 2010.

[13] A. R. Pinjari and C. R. Bhat, "Activity-based travel demand analysis," *A Handb. Transp. Econ.*, no. 1, pp. 1–36, 2011.

[14] Transportation Research Board, "Activity-Based Travel Demand Models: A Primer," 2015.

[15] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira, E. Frazzoli, and M. C. González, "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*, 2013, pp. 2:1–2:9.

[16] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs.," *J. R. Soc. Interface*, vol. 10, no. 84, p. 20130246, Jul. 2013.

[17] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data," *Transp. Res. Part C Emerg. Technol.*, Mar. 2015.

[18] Y. Zheng, "Trajectory Data Mining: An Overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 29:1–29:41, May 2015.

[19] R. Hariharan and K. Toyama, "Project Lachesis: parsing and modeling location histories," *Geogr. Inf. Sci.*, pp. 106–124, 2004.

[20] C.-C. Hung and W.-C. Peng, "A regression-based approach for mining user movement patterns from random sample data," *Data Knowl. Eng.*, vol. 70, no. 1, pp. 1–20, 2011.

[21] R. Cervero and J. Murakami, "Effects of built environments on vehicle miles traveled: evidence from 370 US urbanized areas," *Environ. Plan. A*, vol. 42, no. 2, pp. 400–418, 2010.

[22] C. Zegras, "The Influence of Land Use on Travel Behavior: Empirical Evidence from Santiago de Chile.," *Transp. Res. Rec.*, vol. 1898, 2004.

[23] C. Carrion, F. Pereira, R. Ball, F. Zhao, Y. Kim, K. Nawarathne, N. Zheng, C. Zegras, and M. Ben-Akiva, "Evaluating fms: A preliminary comparison with a traditional travel survey," 2014.

[24] R. Ewing and R. Cervero, "Travel and the Built Environment -- A Meta-Analysis," *J. Am. Plan. Assoc.*, vol. 76, no. 3, pp. 265–294, 2010.