https://doi.org/10.1016/j.compenvurbsys.2018.04.001

Human Mobility and Socioeconomic Status: Analysis of Singapore and Boston

Yang Xu*[1,2], Alexander Belyi[2,3], Iva Bojic[2], and Carlo Ratti[4]

[1]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[2]Senseable City Laboratory, FM IRG, SMART Centre, 1 Create Way, Singapore
[3]Faculty of Applied Mathematics and Computer Science, Belarusian State University, 4 Nezavisimosti Ave., 220030 Minsk, Belarus
[4]Senseable City Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

*Electronic address: yang.ls.xu@polyu.edu.hk

Abstract

Recently, some studies have shown that human movement patterns are strongly associated with regional socioeconomic indicators such as per capita income and poverty rate. These studies, however, are limited in numbers and they have not reached a consensus on what indicators or how effectively they can possibly be used to reflect the socioeconomic characteristics of the underlying populations. In this study, we propose an analytical framework — by coupling large scale mobile phone and urban socioeconomic datasets — to better understand human mobility patterns and their relationships with travelers' socioeconomic status (SES). Six mobility indicators, which include radius of gyration, number of activity locations, activity entropy, travel diversity, k-radius of gyration, and unicity, are derived to quantify important aspects of mobile phone users' mobility characteristics. A data fusion approach is proposed to approximate, at an aggregate level, the SES of mobile phone users. Using Singapore and Boston as case studies, we compare the statistical properties of the six mobility indicators in the two cities and analyze how they vary across socioeconomic classes. The results provide a multifaceted view of the relationships between mobility and SES. Specifically, it is found that phone user groups that are generally richer tend to travel shorter in Singapore but longer in Boston. One of the potential reasons, as suggested by our analysis, is that the rich neighborhoods in the two cities are respectively central and peripheral. For three other mobility indicators that reflect the diversity of individual travel and activity patterns (i.e., number of activity locations, activity entropy, and travel diversity), we find that for both cities, phone users across different socioeconomic classes exhibit very similar characteristics. This indicates that wealth level, at least in Singapore and Boston, is not a factor that restricts how people travel around in the city. In sum, our comparative analysis suggests that the relationship between mobility and SES could vary among cities, and such relationship is influenced by the spatial arrangement of housing, employment opportunities, and human activities.

Key words: mobile phone data, human mobility, socioeconomic characteristics, data fusion

1 Introduction

The last decade has witnessed an explosive growth of scientific research that characterizes and models how people move around in space and time. The interdisciplinary field — broadly conceived as *human mobility analysis* — has attracted researchers across various backgrounds to tackle questions in epidemiology [1], sociology [2] and urban planning [3], among others. With rapid developments of information and location-aware technologies, researchers nowadays have access to large datasets of different types (e. g., mobile phone records, social media data, public transit records). This allows for acquisition of new knowledge about important aspects of human mobility patterns [4, 5, 6].

Despite the numerous insights uncovered by recent human mobility research, there have been limited studies — especially the ones leveraging new and emerging data sources — that analyze the relationships between movement patterns and socioeconomic characteristics of the travelers. This is partially due to a lack of multimodal data that could reveal both travel behavior and socioeconomic status (SES) of the same population. An improved understanding of the relationship between mobility and SES is very important for many scientific domains and real-world applications, especially the ones that call for human-centered approaches. For example, knowing how travel patterns vary across social classes could help decision makers to control spread of infectious diseases more effectively by targeting the right population groups [7], or improve the performance of transportation systems by providing customized mobility solutions to travelers [8]. It can also shed light on many societal issues such as spatial inequality and social stratification [9, 10].

Recently, some studies have shown that human mobility patterns are strongly associated with regional socioeconomic indicators such as per capita income and poverty rate [11, 12, 13]. However, these studies are limited in numbers and they have not reached a consensus on what indicators or how effectively they can possibly be used to reflect the socioeconomic characteristics of the underlying populations. Hence, this research proposes an analytical framework — by coupling large scale mobile phone and urban socioeconomic datasets — to better understand human mobility patterns and their relationships with travelers' socioeconomic status. Using Singapore and Boston as case studies, this work aims to answer one important research question: How do people belonging to different social classes move around in a city, and whether they use urban spaces in different ways?

By analyzing large scale mobile phone data in Singapore and Boston, we introduce six indicators — which are (1) radius of gyration, (2) number of activity locations, (3) activity entropy, (4) travel diversity, (5) k-radius of gyration, and (6) unicity — to quantify important aspects of phone users' mobility characteristics. Among these indicators, radius of gyration and the entropy-based measures (e.g., activity entropy and travel diversity) have been widely used in existing studies to quantify two salient dimensions of human mobility patterns [4, 5, 14, 11], namely, the *spatial dispersion* and *predictability* of individual movements. K-radius of gyration and unicity are two measures that were proposed more recently to quantify individual movements among the most frequented locations [15] and the uniqueness of an individual's activity patterns relative to others [6]. These six mobility indicators, which have gained considerable attentions in human mobility research, can either be derived from raw mobile phone data or meaningful location sequences extracted from mobile phone users' trajectories. They capture a comprehensive picture of phone users' travel behavior, such as the spatial extent of activity space (radius of gyration and k-radius of gyration), the regularity of daily activities (number of activity locations and activity entropy), the diversity of movements among important activity locations (travel diversity), and the re-identifiability of mobility traces (unicity).

By further incorporating several socioeconomic datasets — (1) the sale price of residential properties and household interview travel survey in Singapore, and (2) per capita income estimated at census tract level in Boston — we propose a data fusion approach to approximate, at an aggregate level, the socioeconomic status (SES) of mobile phone users. We then compare the statistical properties of the six mobility indicators in the two cities, and analyze their relationships with the phone users' SES. The comparative analysis reveals the socioeconomic dimensions of human mobility, and suggests whether there exist universal patterns across the cities.

The remainder of this article is organized as follows. Section 2 provides an overview of related work of this research. Section 3 introduces the study areas as well as the mobile phone and socioe-

conomic datasets. In section 4, we introduce how the mobility indicators are derived and the data fusion approach for approximating phone users' SES. We then present analysis results in section 5. Finally, in section 6, we conclude our findings and discuss future research directions.

2 Literature Review

2.1 Dimensions of human mobility

Human mobility analysis is an interdisciplinary field that aims to understand the intrinsic properties of human movements as well as the mechanisms behind the observed patterns. The concept of human mobility is broad in a sense that it encompasses various dimensions of human travel at both individual and group levels. The conceptualization and representation of human mobility also vary depending on the contexts of studies and backgrounds of researchers. One important concept that is widely used in geographical and urban studies is *activity space*. Namely, it denotes the daily environment that an individual is using for his or her activities [16]. It is usually conceptualized as the set of locations that a particular person has visited as well as his/her travels among those locations [17]. Previous studies have employed various activity space measures, such as standard deviational ellipse [18, 19], confidence ellipse and minimum spanning trees [17, 20], and space-time prisms [21, 22], to better understand people's travel and daily activity patterns. The *activity space* measures mainly focus on quantifying a person's mobility patterns from three perspectives: (1) the spatial extent of daily activities, (2) one's frequented activity locations (i.e., activity "anchor" points), and (3) movements between those locations [17]. They collectively form a geographic representation of individual human mobility, and have been widely used to study household travel behavior [23, 24] and individual accessibility to urban facilities [25, 26].

Recent advancements in information and location-aware technologies have produced many new datasets (e.g., mobile phone records and social media data) that capture the whereabouts of people in space and time. These new datasets have empowered researchers from a wide range of fields, such as computer science, statistical physics, and transportation engineering, to characterize and model individual mobility for large populations. Using a six-month cellphone trajectories of 100,000 users, Gonzalez et al. found that individual travel distance (i.e., displacement) can be approximated by a truncated power-law and that people tend to return to a few highly frequented locations [4]. By analyzing a three-month mobile phone trajectories of 50,000 users, Song et al. found that human travel patterns are highly predictable and there is a remarkable lack of variability (in predictability) across the population [5]. In another research [14], which was also based on mobile phone data, the authors developed a microscopic model (*exploration and preferential return*) that is able to reproduce many intrinsic properties (e.g., jump size, visitation probability) of human travel behavior. By applying eigendecomposition to the *MIT Reality Mining* dataset, researchers were able to reconstruct and predict an individual's travel behavior with a high accuracy based on the principle components of his or her activity diary [27]. Several important indicators — such as radius of gyration and entropy-based measures — have been used in these studies to capture the spatial dispersion and regularities of human mobility, respectively [4, 5, 14]. These studies mark a new wave of scientific efforts to uncover the hidden mechanisms that govern individual movements.

Another strand of research focuses more on analyzing collective human behavior and space-time structures of cities. Topics include, but are not limited to, visual analytics of cellular usage [28, 29], community detection in urban population flow [30, 31], and quantification of urban spatial structures [32, 33]. Some studies have also taken advantage of big urban datasets to compare human mobility patterns across cities. For instance, by analyzing mobile phone data in three major US metropolitan areas (Los Angeles, San Francisco and New York), researchers observed notable differences in people's travel ranges among the three cities [34]. Similarly, based on three mobility indicators (daily activity range, number of activity anchor points, and frequency of movements) extracted from large-scale mobile phone data, the authors in [35] performed a comparative analysis of human travel patterns in two metropolitan regions in China (Shenzhen and Shanghai). These studies highlighted the unique properties of human movements in each city that are potentially shaped by the urban spatial structure and socio-demographic characteristics.

In transportation planning and modeling, human mobility is often linked to concepts such as activity locations, origin-destination (OD) matrices, individual trip making, and commuting patterns. Studies in this field mainly focus on identifying and modeling human travels among important activity locations (e.g., home, work place, shops and restaurants). For example, some recent studies have shown that mobile Call Detail Records (CDRs) have the potential of complementing or even substituting traditional surveys in addressing questions such as OD estimation [3, 36] and travel demand modeling [37, 38]. Although the approaches used for studying human mobility vary across disciplines (e.g., geography, urban planning, transportation engineering, and network science), there seem to be a lot of overlapping interests, such as quantifying the spatial extent of individual activity space, understanding individual travels among important activity "anchor" points, and uncovering the inherent regularities in human movements.

2.2 Human mobility and sociodemographic characteristics

Socioeconomic status (SES) and demographic characteristics are important factors that shape individual travel behavior. Before information and communication technologies (ICTs) proliferated, travel surveys had been used as the most reliable data source for assessing and comparing human travel behaviors across social classes and demographic tiers [39, 40, 41, 42, 43]. In these studies, the differences in people's gender, race or ethnicity were found to be correlated with their daily activities and movement patterns. In particular, some studies have observed notable differences in the travel-activity patterns of men and women [39, 40]. They suggested that "women encounter higher levels of daytime fixity constraint" [39, p. 370], while working men "frequented recreation places and workplaces more often than did the women" [40, p. 298]. On the other hand, it is found that socioeconomic status is related to an individual's daily travel patterns. As suggested by [41], an individual's travel frequency is positively correlated with employment status, and income has a positive impact on the spatial dispersion of destinations visited. Another interesting finding from the same study is that education level, which also describes a person's SES, is negatively associated with travel range. The study observed an intertwined relationship between SES and individual travel behavior. Despite the great contributions of these studies, one issue is that collecting travel surveys is usually costly and time-consuming. The difficulties in data collection limited the scope of the studies, which usually focused on investigating a small size of participants during a short-period of time. Such difficulties also pose additional challenges to performing comparative analysis across areas with different social-cultural characteristics (i.e., inter-city comparisons).

In recent years, mobile phone data have become a new data source for studying the social aspects of human mobility [44]. For example, by analyzing mobile phone trajectories of two linguistic groups in Tallinn, Estonia, the authors found that ethnicity has a significant influence on the activity spaces of individuals [45, 46]. Using a CDR dataset collected in Shenzhen, China [47], the authors proposed a home-based approach to analyze how people's daily activities take place around their home locations, and the results revealed a 'north-south' contrast of human activity space that is in general agreement with the socioeconomic divide in the city. In these studies, sociodemographic characteristics of phone users are implicitly considered or approximated using certain variables (e.g., linguistic background). However, there is still a remarkable lack of research that would reveal the relationship between movement patterns and socioeconomic characteristics of the travelers. This is partially due to the difficulty of coupling large individual tracking datasets with SES. One approach usually adopted is to associate individual home location to census tract, where aggregate characteristics of SES (e.g., mean/median household income) are available [48].

The associations between travel behavior and SES revealed by previous studies also spurred a collection of research that aimed at predicting socioeconomic levels based on human mobility datasets. For example, some recent studies found that mobile phone data can be used to predict individual SES [49] and regional socioeconomic characteristics such as poverty and wealth levels [50, 51, 52]. In these studies, mobility indicators of phone users are used along with other variables, such as calling patterns and social network structures. The prediction methods (e.g., regression, SVM, and random forest) as well as the social-cultural characteristics of areas studied also varied. In other words, whether there exists any universal relationship between mobility and SES remains

to be better understood.

3 Study Area and Datasets

3.1 Mobile phone data

In this research, two mobile phone datasets collected in Singapore and Boston Metropolitan Area are used to derive phone users' mobility indicators. The Singapore dataset covers 4.4 million cellphone users during a period of 50 days in 2011. Each mobile phone record tracks the unique ID of the phone user, the communication type (call/SMS), as well as the date, time and the user's location when the phone communication started. The location of each record was reported as the latitude/longitude of the phone user's connected cellphone tower. There are about 5,000 cellphone towers that are densely distributed across the whole Singapore, and the average nearest distance between them is about 100 meters. On the other hand, Boston dataset contains location estimations of about one million anonymous mobile phone users. It was collected by AirSage (http://www.airsage.com) during four months in 2009. In this dataset, location information is generated each time a phone user engages in calling, messaging, or web browsing activities (we refer to all these activities as 'calls' in the following text). In contrast to the Singapore dataset, which explicitly provides phone users' locations at the cellphone tower level, the Boston dataset reports location estimations obtained through triangulation technology. The uncertainty range of the estimations has a mean of 320 meters and a median of 200 meters. More detailed information about this dataset can be found in [53].

To control the issue of data sparsity and in order to take into account that the two datasets were passively generated during certain types of mobile phone activities, this research focuses on the cellphone users, who were active (i.e., have at least one record in the dataset) at least half of the days during the given data collection period. Specifically, the subset of Singapore dataset consists of phone users with at least 25 active days of phone usage, as compared to those of 60 active days for the Boston dataset. This allows us to mitigate the data sparsity issue by filtering individuals: (1) who are short-term subscribers and/or (2) who have inactive phone usage during the study period. For example, the average number of active hours (i.e., number of one hour time slots with at least one call) per active day increased from 4 to 6 in both datasets, and average number of calls per active day increased from 17 to 25 in Singapore and from 31 to 47 in Boston. At the same time, average time between calls on active days also increased from 63 to 73 and from 27 to 31 minutes for the Singapore and Boston datasets respectively, indicating that phone calls of short-term subscribers are more concentrated in time, i.e., bursty. The resulting datasets after this filtering consist of 2.1 and 0.5 million phone users in Singapore and Boston, respectively. Detailed plots with distributions of time-related characteristics of the data could be found in Appendix A.

3.2 Socioeconomic data

In this research, two different types of socioeconomic data are used to reflect the SES of the mobile phone users in Singapore and Boston. For the case of Singapore, we use a housing price dataset acquired from a private company (https://www.99.co), which is one of the two largest companies in the country that provide map-based search for a comprehensive coverage of housing properties[1]. The dataset used in this research includes information of thousands of residential properties across the country collected between 2011 and 2012. Each record corresponds to a unique housing property in geographic space, with information such as its property type (i.e., condo, landed, or HDB[2]), the geographic coordinates (i.e., latitude and longitude), and the total sale price of one housing unit.

The housing price dataset is used as a proxy for mobile phone users' SES. Basically, we hypothesize that people who live in areas with a higher average housing price tend to be richer in general. To support the usage of this dataset, we incorporate another dataset — the Household Interview Travel Survey (HITS) — collected by the Singapore Land Transport Authority (LTA) in 2012. The

[1]The other one is Property Guru (https://www.propertyguru.com.sg/).

[2]HDB, which is short for Housing Development Board, is a type of residential housing property that is publicly governed and developed in Singapore. The HDB flats were built primarily to provide affordable housing.

Singapore 2012 HITS collects 1-day travel diary of 35,715 individuals (sampling rate of about 1%) along with other socio-demographic attributes — such as monthly income — self-reported by the respondents. To understand the correlation between the housing price and monthly income recorded by HITS, we extract all the individuals in the HITS data who reported their income (12,111 in total). We then aggregate these individuals — based on the postal code of their residencies — by planning areas (as shown in Figure 1A), and calculate the average value of monthly income for individuals in each planning area. Similarly, we compute the average sale price of housing units for each planning area and explore the relationship between the two variables. As illustrated in Figure 1B, the average housing price matches relatively well with the income level at the corresponding planning area except for 3 outliers (i.e., Novena, Sungei Kadut, and Southern Islands). We think this is partially caused by the sampling bias. For example, we find that only two individuals are sampled from the Southern Islands, a planning area where several luxury housing communities locate. Both of them reported a monthly income of 500 SGD. Figure 1C shows the relationship between the two variables after filtering these three outliers. The Pearson's correlation is 0.88, which suggests that housing price is a strong indicator of the residents' SES.

Note that we use housing price data instead of HITS because only a limited number of individuals are sampled from HITS, which provide sporadic observations about people's SES. The housing price data, however, enable us to capture the heterogeneity of mobile phone users' SES at a finer spatial granularity. Specifically, the mobile phone users' SES can be approximated at the level of cellphone tower service areas, which have a much finer spatial resolution than the planning areas. Details on how mobile phone users and socioeconomic variables are associated will be described in section 4.4.

For the case of Boston, we use per capita income estimated at the census tract level (Figure 1D) as a proxy of phone users' SES. The data, which is included in the 2010 American Community Survey (ACS), is publicly available and can be downloaded through the American FactFinder (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml) provided by the United States Census Bureau. Specifically, we define the geographic type as census tract, and then choose the dataset as 2010 ACS 5-year estimates. The dataset includes one table that provides per capita income estimated for the past 12 months.

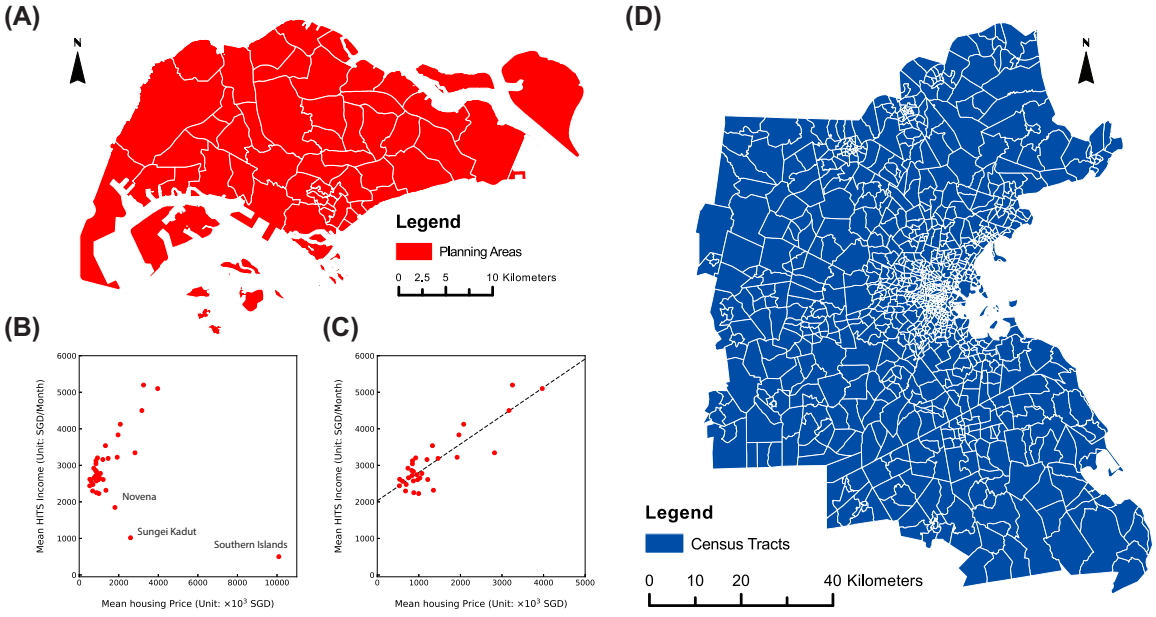

Figure 1: (A) The planning areas of Singapore; (B) The relationship between average housing price and average monthly income computed at the planning area level; (C) The correlation between the two variables after removing three outliers (Pearson's $r = 0.88$); (D) The census tracts in the metropolitan Boston area.

4 Methodology

In this section, we introduce the methods for deriving mobility indicators as well as how we approximate phone users' socioeconomic status (SES). Figure 2 illustrates the overall research design. First, we derive a collection of indicators to quantify important aspects of phone users' mobility characteristics. The first indicator, radius of gyration, is derived from the raw mobile phone data to describe the typical range of user's activity territory. The indicator has been widely used in previous studies [4, 5] to quantify the spatial dispersion of a phone user's daily activities. When calculating this indicator, previous studies often consider cellphone towers visited by a phone user as independent locations, and the total number of records generated at each tower indicates its importance to the user. Since radius of gyration is calculated from the raw data, we refer to it as a *low level mobility indicator* (LMI) in this research.

However, there are several issues with LMI. One issue is that an individual's reported locations depend on the cellphone tower he/she is connected to. And this connection could switch between different cellphone towers due to signal jump or load balancing [54]. That means mobile phone observations might not reflect a phone user's real locations. Also, the intensity of a phone user's communication activities at a particular location is only one way of measuring the location's importance. Other properties, such as duration of stay, are also essential to the understanding of phone users' travel behavior. To obtain a behaviorally more realistic measure of individual mobility, we apply a trajectory segmentation method to process raw data of each user into a meaningful location sequence, from which a set of *high-level mobility indicators* (HMIs) is derived. The methodologies for deriving HMIs and what properties of individual mobility each HMI could capture will be introduced in section 4.3.

Next, we associate mobile phone users — based on their estimated home locations — to spatial units (i.e., cellphone tower service areas for Singapore and census tracts for Boston) where socioeconomic variables are available. The socioeconomic variables are used to reflect, at an aggregate level, the SES of mobile phone users. The mobility indicators and SES of mobile phone users are then analyzed to better understand how people belonging to different social classes move around in the two cities.

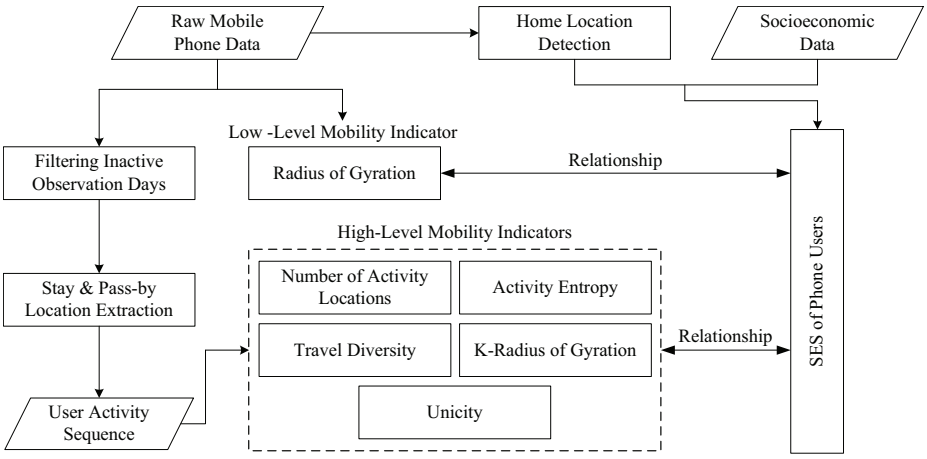

Figure 2: In this research, we derive a collection of indicators to quantify important aspects of phone users' mobility characteristics. We then examine the relationship between mobility characteristics and SES of mobile phone users.

4.1 Radius of gyration as a low level indicator (LMI)

Given a cellphone user's observations as a list of tuples $\{(l_1, t_1), (l_2, t_2), ..., (l_n, t_n)\}$, where l_i and t_i denote the location and the time of the i^{th} observation, the radius of gyration, R_g, is defined as follows:

$$R_g = \sqrt{\frac{\sum_{i=1}^{n} \left(\overrightarrow{l_i} - \overrightarrow{l_c}\right)^2}{n}} \quad (1)$$

Here $\overrightarrow{l_i}$ denotes the location vector (i.e., x, y coordinates) of l_i, and $\overrightarrow{l_c} = \sum \overrightarrow{l_i}/n$ refers to the center of mass. R_g can be used to measure the spatial dispersion of a phone user's daily activities. A large value of R_g usually indicates a large activity space, while a small value suggests that the phone user's daily activities mainly concentrated in a small geographic area.

4.2 Trajectory segmentation and stay location extraction

The purpose of trajectory segmentation is to obtain a behaviorally more realistic representation of a phone user's movements over time. Due to the sparsity of the mobile phone data used in this research, it is rather challenging to extract meaningful activity sequences — such as location and duration of stays — for phone users who have very few records in a day. Hence, it is necessary to filter users and observation days with few mobile phone records. In this research, we adopt the methods and workflow from [3, 37] to perform the trajectory segmentation. First, by dividing a day into 24 one-hour time windows, we calculate, for each phone user, the number of active time windows (e.g., ones with call observations) for each day. We refer to the days where the total number of active time windows is equal or greater than a threshold (set to 8 in this research) as *active observation days* [37]. For each phone user, only active observation days are used for the trajectory segmentation.

Let $\{(l'_1, t'_1), (l'_2, t'_2), ..., (l'_{n'}, t'_{n'})\}$ denote a phone user's records on the active observation days. We compare each record with its subsequent observation, and we merge them into a segment if they are within a roaming distance Δd_1. We then calculate the medoid (or mean center)[3] of the segment, which is then compared with the next observation. We iteratively add an observation into the segment if its location and the medoid (or mean center) is within Δd_1. Otherwise, a new segment is created.

Iterating through $\{(l'_1, t'_1), (l'_2, t'_2), ..., (l'_{n'}, t'_{n'})\}$ based on the described above process results in a sequence $\{(m_1, t''_1, dur_1), (m_2, t''_2, dur_2), ..., (m_{n''}, t''_{n''}, dur_{n''})\}$, where m_i, t''_i, and dur_i denote the medoid (or mean center), starting time, and the stay duration of the i^{th} segment, respectively. Then, we further cluster the medoids or mean centers (m_i) with $dur_i > 0$ that are within a roaming distance Δd_2. The purpose is to merge the stay segments that are close to each other in geographic space. Finally, we keep the segments which duration is greater than a threshold Δt. The final location sequence for a phone user is denoted as $\{(s_1, \bar{t}_1, \overline{dur}_1), (s_2, \bar{t}_2, \overline{dur}_2), ..., (s_{\bar{n}}, \bar{t}_{\bar{n}}, \overline{dur}_{\bar{n}})\}$.

Considering the average spacing gap of cellphone towers in Singapore (i.e., 100 meters) and the median range of location uncertainty in the Boston dataset (i.e., 200 meters), we set both Δd_1 and Δd_2 to 300 meters, and Δt was set to 10 minutes[4]. Figure 3 shows an example of the cellphone towers that were visited by a randomly selected user in the Singapore dataset during the whole data collection period (384 towers in total, see Figure 3A) as well as on the active observation days (366 towers in total, see Figure 3B). By performing the trajectory segmentation and keeping the stay segments with duration above Δt, we successfully extracted 35 stay locations for this user (see Figure 3C).

4.3 High-level mobility indicators (*HMIs*)

Given cellphone user's stay segments $\{(s_1, \bar{t}_1, \overline{dur}_1), (s_2, \bar{t}_2, \overline{dur}_2), ..., (s_{\bar{n}}, \bar{t}_{\bar{n}}, \overline{dur}_{\bar{n}})\}$, we derive the following five indicators:

- A: Total number of activity locations

[3]Since the locations in the Singapore dataset are reported explicitly at the cellphone tower level, we use medoid to represent the location of the stay segment. However, as the location information in the Boston dataset is generated based on triangulation technology, in this case we use mean center.

[4]Refer to [3, 37] for further information on the method.

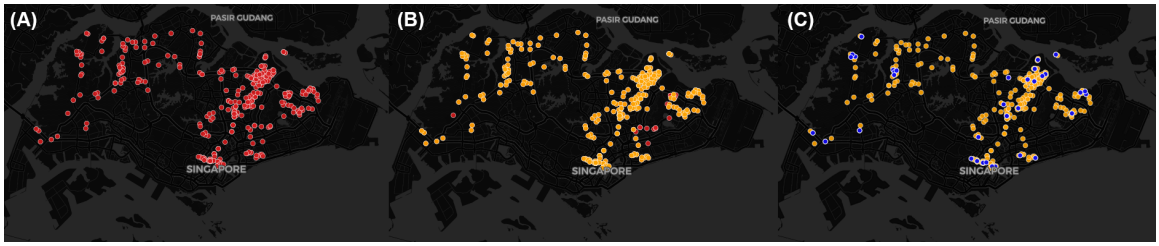

Figure 3: (A) Cellphone towers (red) visited by a randomly selected user during the data collection period; (B) Towers (orange) that are visited by the same user on active observation days; (C) The extracted stay locations (blue) on active observation days.

- H_1: Activity entropy
- H_2: Travel diversity
- $R_g^{(k)}$: K-Radius of gyration
- U: Unicity

The first high-level indicator, A, measures the total number of activity locations visited by a phone user during the data collection period:

$$A = |set(s_1, s_2, ..., s_{\bar{n}})| \tag{2}$$

A large value of A indicates that phone users' activities are distributed across a variety of locations during the study period.

Given a vector $\{p_1, p_2, ..., p_A\}$, where $p_i = \frac{\sum_{s_j=s_i} \overline{dur}_j}{\sum_{j=1}^{\bar{n}} \overline{dur}_j}$ denotes the proportion of duration of stay at location s_i, the activity entropy is calculated as:

$$H_1 = -\sum_{i=1}^{A} p_i \log(p_i) \tag{3}$$

Note that $\sum p_i = 1$. Since the locations in $set(s_1, s_2, ..., s_{\bar{n}})$ are weighted by the (observed) duration of stay rather than the number of mobile phone records (e.g., calls/SMS), the measure is less sensitive to phone users' calling behavior and the 'bursty' nature of human communication activities [55]. From a spatial point of view, a large value of H_1 suggests that the diversity of a phone user's daily activities is high.

The travel diversity, H_2, measures the regularity of a phone user's movements among his/her activity locations. Given a phone user's stay segments $\{(s_1, \bar{t}_1, \overline{dur}_1), (s_2, \bar{t}_2, \overline{dur}_2), ..., (s_{\bar{n}}, \bar{t}_{\bar{n}}, \overline{dur}_{\bar{n}})\}$, we define origin-destination trips as movements between consecutive stay locations. Let E denote all the possible origin-destination pairs (without considering direction) extracted from this phone user's activity locations (i.e., $set(s_1, s_2, ..., s_{\bar{n}})$), the travel diversity is measured as:

$$H_2 = -\sum_{i \in E} p_i' \log(p_i') \tag{4}$$

where p_i' is the probability of observing a movement between the i^{th} origin-destination pair. Note that $\sum p_i' = 1$. A large value of H_2 indicates that a phone user's trips distribute across a variety of origins and destinations.

K-radius of gyration, $R_g^{(k)}$, is a radius of gyration calculated using only k the most visited places. It was proposed by Pappalardo et al. in [15] to measure to what extent the most important locations determine a user's radius of gyration. A precise definition of 'the most visited place' could vary

depending on a context. We define it as a place where a user spent the longest time, i.e., a place with the longest total duration of stay. So in our case, a formulae for $R_g^{(k)}$ takes the form:

$$R_g^{(k)} = \sqrt{\frac{\sum_{i=1}^{k} \left(\overline{dur}_i \left(\overrightarrow{s_i} - \overrightarrow{s_c} \right)^2 \right)}{\sum_{i=1}^{k} \overline{dur}_i}} \tag{5}$$

where s_i, $i = \overline{1,k}$, are k locations with the longest total duration of stay, and $s_c = \frac{\sum_{i=1}^{k} \left(\overline{dur}_i \overrightarrow{s_i} \right)}{\sum_{i=1}^{k} \overline{dur}}$ is the center of mass of these k locations. By using only stay locations from aggregated trajectory we try to avoid cases when some of the most visited locations are actually one location and apparent split is caused by switching between cell towers. For example, we expect home and work to be the two places where most users spend the longest time. Then, 2-radius of gyration measures user's activity space between these two places. If it is small compared to total radius of gyration then those two places play a less important role in person's mobility habits. And if it is on par with the overall radius of gyration then those two places completely explain person's traveling behavior. Using k-radius of gyration we can divide users into two categories: returners and explorers. This notion was also originally proposed by Pappalardo et al. in [15]. As authors suggested in their study, k-returners are those for whom $R_g^{(k)} \geq R_g/2$ and k-explorers are those for whom $R_g^{(k)} < R_g/2$. Intuitively, k-returners are those who tend to spend most of the time between k the most important locations, while k-explorers are those who's activity space cannot be well described by only k top locations.

Our unicity measure, inspired by [6], estimates the number of top locations needed to uniquely identify a particular person. The fewer points needed, the more unique the person is, meaning that it is easier to re-identify him/her using outside information about top locations he/she visited. Since collected locations in the Boston and Singapore datasets were not at the same spatial resolution (i.e., in Singapore they were on a cellphone tower level, while Boston dataset reported locations that were determined using a triangulation method), we decide to perform the analysis at the level of grid cells. Specifically, considering the uncertainty range of mobile positioning in the two cities (100 meters for Singapore and 200 meters for Boston), we divide the study areas into regular grid cells using two different spatial resolutions (500m and 1km). 1km grid is used to examine how a coarser spatial resolution affects the unicity results. We then map each user's stay locations onto the grid cells and evaluate his/her re-identifiability.

Given a phone user's stay segments as $\{(s_1, \bar{t}_1, \overline{dur}_1), (s_2, \bar{t}_2, \overline{dur}_2), ..., (s_{\bar{n}}, \bar{t}_{\bar{n}}, \overline{dur}_{\bar{n}})\}$, we map them onto the grid cells $\{g_1, g_2, ..., g_z\}$ and calculate the total duration that the user stayed at each grid cell (i.e., $\forall i$ where $s_i \in g_z \rightarrow \sum \overline{dur}_i$). After this step, we order all the grid cells visited by the phone user in descending order of stay duration. The top l locations are defined as the first l locations in this ordered list. After deriving these locations for all the users, we examine the re-identifiability of the users by analyzing the uniqueness of their top l locations. In particular, we introduce the concept of unicity, U_l, as the percentage of mobile phone users who can be uniquely identified using the top l locations:

$$U_l = \frac{number\ of\ phone\ users\ with\ unique\ top\ l\ locations}{total\ number\ of\ phone\ users\ in\ a\ city} \tag{6}$$

It is important to note that we are building a set of top l visited locations, which means that the order does not matter. For example, if for a given user the first most visited location is g_5 followed by g_2 and for another user his/her top two locations are g_2 and g_5, in the case when $l = 2$, those two users would not be considered as unique ones. Finally, we explore the relationship between U_l and SES of phone users. The research question we pose here is: Is a lower socioeconomic status a limiting factor for people to be more unique? Meaning if richer people are more privileged to visit more different places and consequently be more unique.

4.4 Home location detection and association with socioeconomic variables

To establish a link between mobile phone users and SES, one approach usually adopted in existing studies is to associate individuals — based on their residential locations — to spatial units where socioeconomic variables are available. In this research, we use a similar approach. Specifically, we estimate each phone user's home location, which is then associated with a value derived from the corresponding socioeconomic variable (e.g., average housing price or per capita income).

There have been many studies which discuss how home locations can be inferred from mobile phone data [56, 57, 58, 47, 44]. For the Singapore dataset, we estimate each individual's home location as the most used cellphone tower before 06:00 and after 19:00. We then generate Voronoi polygons based on the spatial distribution of the cellphone towers to approximate their service areas. For each cellphone tower service area, we then extract all the housing properties (i.e., condo, landed or HDB) that fall inside. And then we calculate the average value of unit sale price of all these properties, which is then used to represent the housing price level at the corresponding cellphone tower service area. On average, we have 24.5 housing properties in each cellphone tower service area to calculate the average sale price (median is 12, 25th and 75th percentiles are 6 and 23, respectively; readers could refer to Figure B.3 in the appendix for detailed information). Finally, each phone user is associated with a housing price value based on the service area of his/her home tower.

To ensure that our approach captures the heterogeneity of the housing price in Singapore, we calculate the standard deviation of unit sale price within each cellphone tower service area (i.e., within-cell std), and compare that with the overall standard deviation (i.e., overall std). As illustrated in Figure 4A, a large proportion of cellphone tower service areas have a small ratio between within-cell std and overall std. The median value is 0.16. That means the cellphone tower service areas (i.e., Voronoi polygons), to a large extent, capture the heterogeneity of the residential housing price in Singapore. Figure 4B shows the percentage of mobile phone users that are associated with different housing price values. The distribution is highly skewed to the right, indicating that a limited proportion of people live in expensive neighborhoods.

Since the mobile phone locations in the Boston dataset are not provided at the cellphone tower level, we use a different home location estimation method. Specifically, given a phone user's stay locations as $set(s_1, s_2, ..., s_{\bar{n}})$, we calculate, for each location, the total number of mobile phone records before 06:00 and after 19:00 during the entire data collection period. The location with the highest frequency (i.e., the most used location) is estimated as the phone user's home. Similar to what we perform on the Singapore dataset, each phone user is associated with the per capita income value of his/her home census tract. Figure 4C shows the histogram of phone users that are associated with different income values, which generally follows a symmetric distribution.

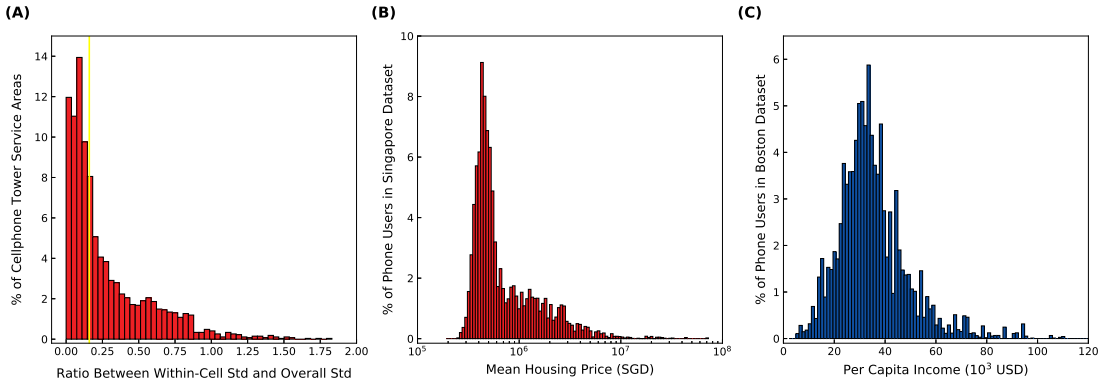

Figure 4: (A) Histogram of the ratio between within-cell standard deviation and overall deviation of housing price at the level of cellphone tower service areas; (B) Percentage of phone users that are associated with a particular housing price value in Singapore (C) Percentage of phone users that are associated with a particular per capita income value in Boston.

5 Analysis Results

In this section, we present our findings — by comparing Singapore and Boston — on the properties of the derived mobility indicators (LMI & HMIs) and their relationships with phone users' SES.

5.1 Radius of gyration and user class definition

We first examine the relationships between radius of gyration (R_g) and SES of phone users in the two cities. Since mobile phone users are associated with different housing price or income values, to distinguish their SES, we use a social stratum model [10] to group them into different classes. Note that these classes are only used to reflect group-level characteristics, and they do not represent individual socioeconomic status. Specifically, for each city, we sort mobile phone users in ascending order based on their associated socioeconomic values v. By calculating the cumulative sum of v, the model groups mobile phone users into q classes such that the sum of socioeconomic values in each user class is the same (i.e., equal to $(\sum v)/q$). Compared to quantile classification from which each class gets an equal size, the social stratum model partitions individuals into classes with decreasing sizes, such that the differences of SES among classes can be distinguished more effectively (i.e., richer groups have smaller sizes).

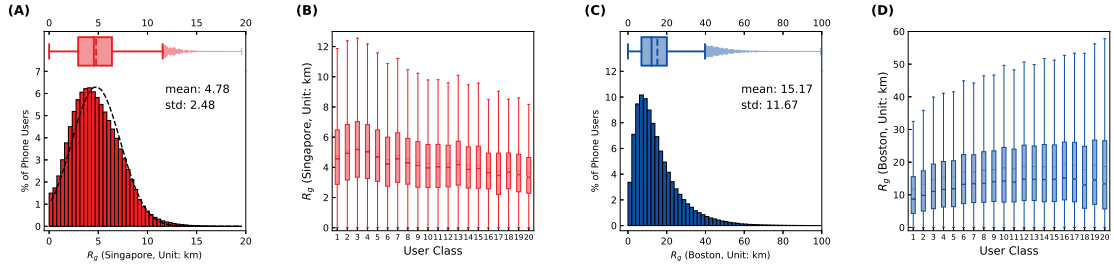

Figure 5: (A-B) The distribution of radius of gyration (R_g) for the Singapore dataset and its relationship with user classes (box plots for 20 classes; whiskers indicate 1.5 interquartile range, mean is shown as dotted line, notch around median shows its standard error); (C-D) The distribution of R_g for the Boston dataset and its relationship with user classes.

Figure 5 illustrates the relationships between R_g and SES of mobile phone users in the two cities. Here we use $q = 20$ as an example. As illustrated in Figure 5A, the R_g of phone users in Singapore generally follows a normal distribution. By investigating their relationship with SES, as shown in Figure 5B, we find that user groups living in richer areas tend to have a smaller R_g on average, with Spearman correlation coefficient between housing price and radius of gyration of -0.10 (p-value less than 10^{-6}). By examining the geographic patterns of housing prices in Singapore, it is found that many expensive and luxury residential communities locate in planning areas such as Bukit Timah, Tanglin, River Valley, and Marine Parade (see Figure 6A). These areas sit relatively close to the *downtown core*, the central business district (CBD) of Singapore. People who live in these planning areas tend to have better access to the various job opportunities provided by the CBD. Moreover, many retail and entertainment hubs (e.g., Orchard Road and Sentosa Island) are also highly accessible from these planning areas. That means people who live in these rich areas were able to perform different types of daily activities (e.g., working and recreational) within a short travel range. Furthermore, by computing phone users' radius of gyration on weekdays and weekends separately (i.e., $R_{g(weekdays)}$ and $R_{g(weekends)}$), we find that their relationships with SES are consistent with Figure 5B, indicating a shorter travel range for richer groups on both weekdays and weekends (see Appendix C).

For the case of Boston, the distribution of phone users' R_g is highly skewed to the right (Figure 5C). It starts to resemble closer power-law distributions observed in other studies [4, 15] where much larger county-level coverage areas were analyzed. The reason for this is that some phone users tend to travel very far during the data collection period.

The relationship between R_g and SES, as illustrated in Figure 5D, is different from that of Singapore. In general, mobile phone users who live in poorer areas tend to travel shorter, but the difference between the middle and upper user classes is relatively small. Spearman correlation coefficient between income and radius of gyration equals to 0.17 (p-value less than 10^{-6}). One possible explanation, as suggested in some studies [59, 60], is that higher socioeconomic status would lead to longer travel distances, which allows people to access better housing and/or job opportunities (see Figure 6B). Note that we also compute $R_{g(weekdays)}$ and $R_{g(weekends)}$ for the Boston dataset, and their relationships with SES are consistent with the result in Figure 5D (see Appendix C).

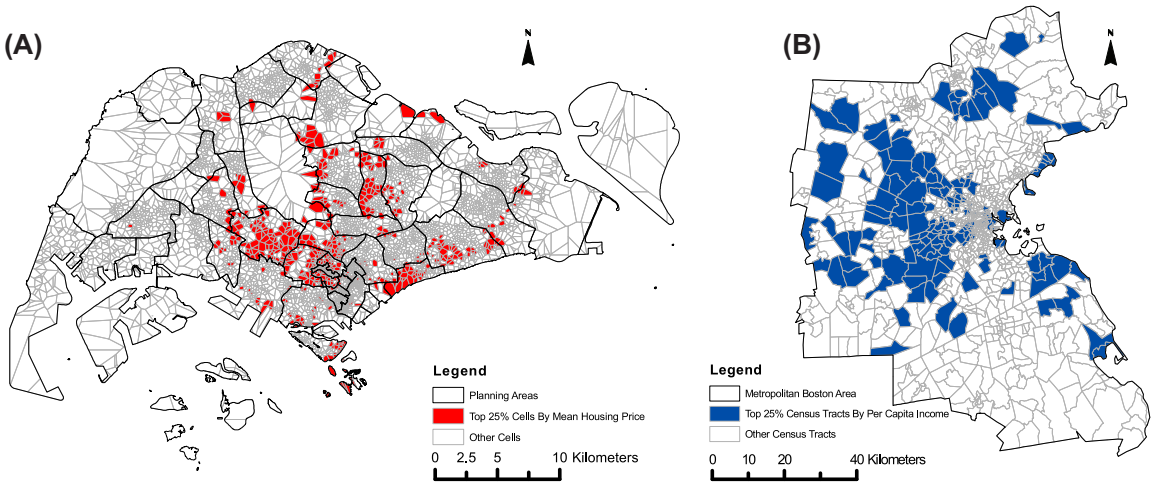

Figure 6: (A) Top 25 percent of the cellphone tower service areas by average housing price in Singapore; (B) Top 25 percent of the census tracts by per capita income in Boston.

5.2 Activity locations, entropy, and travel diversity

Figure 7 shows the relationships between SES and three high-level mobility indicators (HMIs), which are the total number of activity locations (A), activity entropy (H_1), and travel diversity (H_2). These three indicators describe — from different perspectives — the regularity of phone users' daily travel and activity patterns.

The results reveal — in both cities — a lack of variability across SES for all three indicators. For the first indicator A, we find that both cities have a right skewed distribution (Figure 7A and Figure 7C), suggesting that a large proportion of individuals tend to use a small set of locations for their daily activities. The mean and standard deviation of A for Singapore are 14.24 and 10.44, as compared to 21.90 and 15.87 for Boston. As illustrated in Figure 7B and Figure 7D, the average value of A seems not to vary across user classes, despite that the average value of A for the lower and upper classes (e.g., 1, 18, 19 and 20) in Boston are slightly lower than the overall mean, with Spearman correlation coefficient of 0.05 for Singapore and -0.06 for Boston.

For activity entropy (H_1), our analysis produces similar distributions for both cities, with the average value of 1.10 for Singapore (Figure 7E) and 1.12 for Boston (Figure 7G). Although phone users in Boston have more activity locations (A) on average, we can see that the activity diversity of phone users in the two cities are highly comparable. One potential reason is that for most of the people, their daily activities mainly concentrate at a few locations (e.g., home and work location). The regularity of human activities at these locations would have a notable impact on the activity diversity of phone users. We next examine the relationships between H_2 and SES. Again, the results suggest that users across different socioeconomic classes exhibit similar levels of travel diversity (Figure 7J and Figure 7L). This suggests that the wealth level of people, as least in Singapore and Boston, is not a limiting factor that affects how they travel around in the city.

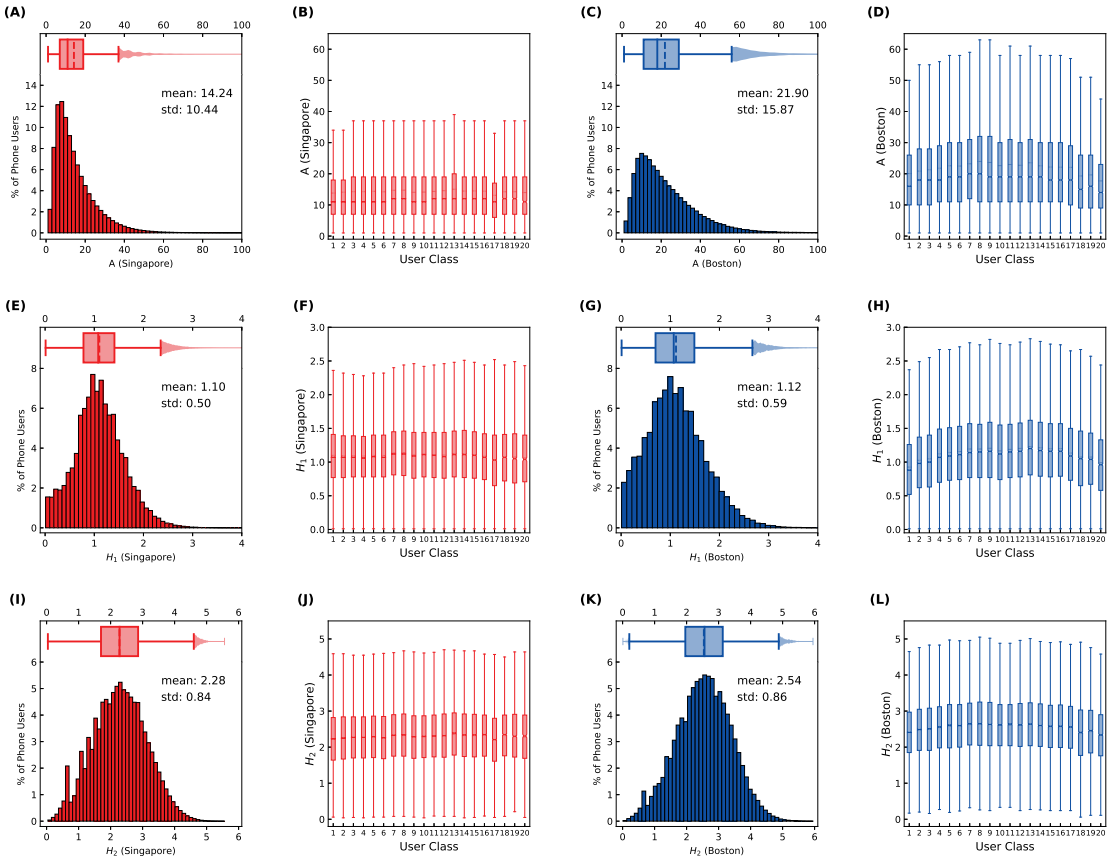

Figure 7: (A-B) The distribution of the number of activity locations (A) for the Singapore dataset and its relationship with user classes; (C-D) The distribution of A for the Boston dataset and its relationship with user classes; (E-F) The distribution of activity entropy (H_1) for the Singapore dataset and its relationship with user classes; (G-H) The distribution of H_1 for the Boston dataset and their relationship with user classes; (I-J) The distribution of travel diversity (H_2) for the Singapore dataset and its relationship with user classes; (K-L) The distribution of H_2 for the Boston dataset and its relationship with user classes.

5.3 Properties of returners and explorers (k-radius of gyration)

In this part we consider only users for whom radius of gyration based on stay locations is greater than zero, i.e., users for whom we found at least two stay locations. This filtering left us with about 280,000 users in Boston area and about 1.25 million users in Singapore. We also use the split of the whole population into $q = 9$ socioeconomic classes following the logic of [10] that there are 3 major classes (namely poor, rich and middle class) each of which could be further split into 3 subclasses.

We calculate and compare k-radius of gyration with overall radius of gyration (for $k = 2, 4, 8$). We have found that in both cities there are more returners than explorers even for $k = 2$. Figure 8 shows how 2-radius of gyration relates to overall gyradius (see Appendix B for plots for $k = 4, 8$). In agreement with [15], there is a split between users for whom k-radius is almost equal to overall radius of gyration versus those for whom k-radius is much smaller than overall. Interestingly, in our case 2-returners already represent more than 70% of all users (70.35% of returners vs 29.65% of explorers in Singapore and 72.49% of returners vs 27.51% of explorers in Boston). This means that for more than 70% of the users top two locations (presumably home and work for majority of the population) already explain more than a half of overall activity space.

In Figure 9 we show how the ratio of returners changes from one class to another overall and

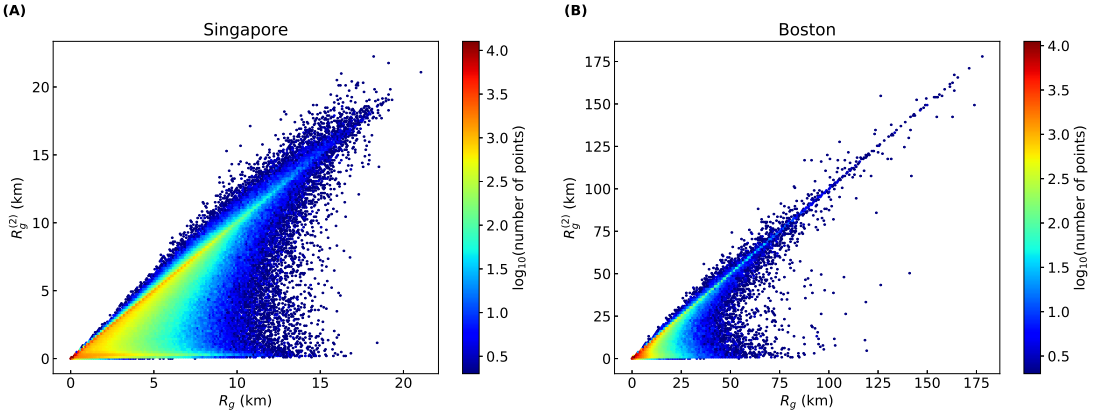

Figure 8: 2-radiuses of gyration versus overall gyradiuses in Singapore and Boston.

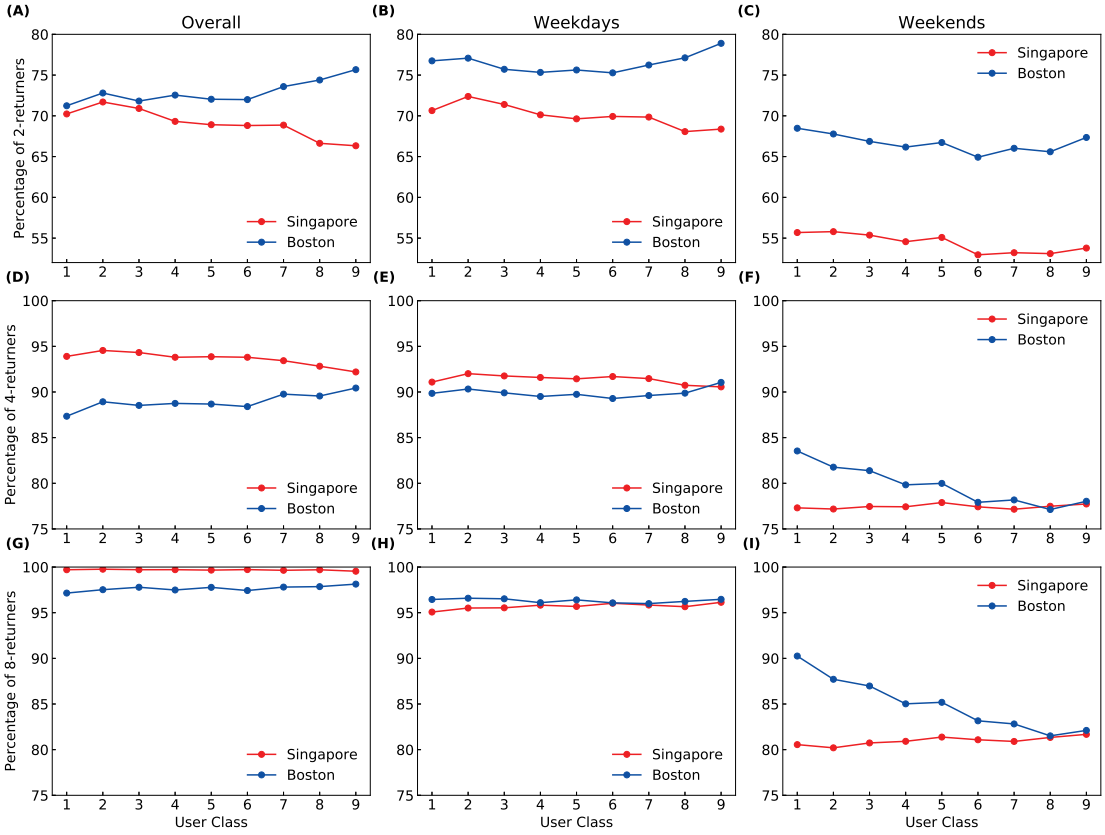

Figure 9: A percentage of returners in each class overall and on weekdays and weekends separately for $k = 2, 4, 8$.

separately on weekdays and weekends. These plots show that for $k = 2$ there are more returners in Boston in each class. And this stays the same on weekends for all k but changes in overall patterns for $k = 4$ and $k = 8$. This means that top two locations better describe users' overall activity space in Boston, but four top locations already better explain total gyradius in Singapore. This shows that people in Singapore are more likely to travel beyond the top two locations for daily activities. Higher prevalence of 4-returners and especially 8-returners in Singapore could be explained by very

restricted area of the city-state that limits both people's movements and coverage of our data.

We can see smaller number of returners on weekends than on weekdays (and respectively higher number of explorers), i. e., the most popular locations play lesser role in people's travel distance on weekends than on weekdays. Figure 9 also shows that percentage of returners grows when people become richer in Boston and shrinks in Singapore in overall patterns for k equal 2 and 4 while trend is opposite for $k = 4, 8$ on weekends. The later could be explained as some richer people could occasionally travel quite far on weekends and this was captured by the Boston dataset. The former though deserve a closer look.

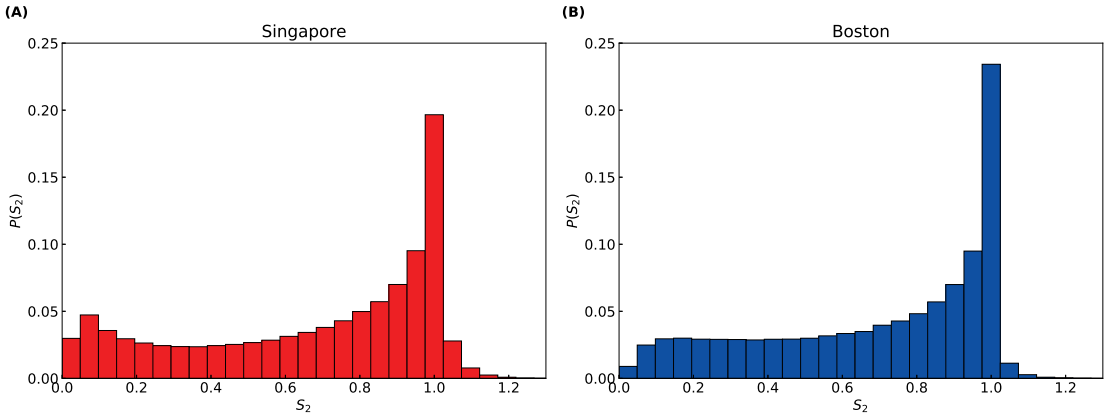

Figure 10: A distribution of the ratio $S_2 = R_g^{(2)}/R_g$. A peak close to the right indicates a high number of strong 2-returners.

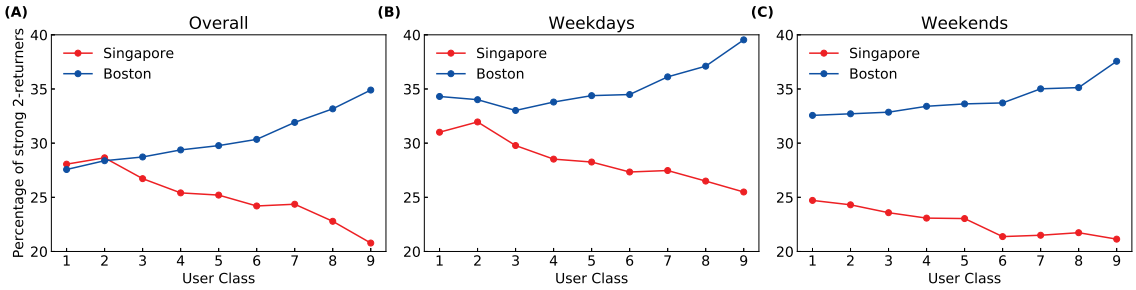

Figure 11: A percentage of strong returners in each class overall and on weekdays and weekends separately showing opposite trends in Boston and Singapore.

In Figure 10 we plot a distribution of ratio $S_2 = R_g^{(2)}/R_g$. The closer this value to 1, the better two the most popular locations describe person's overall activity space. We will call *strong returners* those for whom 2-radius and overall radius of gyration are very close or, more precisely, when $0.9 \leq R_g^{(2)}/R_g \leq 1.1$. A high bar around $S_2 = 1.0$ in Figure 10 indicates that there are a lot of people for whom $R_g^{(2)}$ is almost equal to R_g. In Figure 11 we show proportions of strong 2-returners in each class. We can see here a clear trend: in Boston the percentage of strong returners increases with class while in Singapore it increases a little from the poorest to the second class but then decreases when people become richer. Plots in Appendix B show similar picture for k equal 4 and 8. This could be caused by the same reasons that cause overall radius of gyration to increase in Boston and decrease in Singapore with socio-economic class. Namely, wealthier people can afford living closer to the city center and their work location in Singapore. So their 2-radius of gyration tends to become smaller, while other trips, including leisure and recreational trips, could require traveling longer distances. Conversely, in Boston, higher social classes choose better housing options outside

the city and travel longer distances to work. This makes their 2-radius of gyration relatively large. In the same time, other destinations mostly fall within the home-work circle and do not increase overall gyradius.

5.4 Unicity

Our analysis on unicity consists of two parts. First, we distinguish two different levels of spatial aggregation, i.e., the 500m grid and 1km grid levels. Figure 12 shows the percentage of users that can be uniquely identified (shown on y-axis) when considering from 2 to 10 activity locations (shown on x-axis). According to the previous results published in [6], 95% of the users can be uniquely identified when taking into an account 4 spatiotemporal points. We got similar results when looking at top four locations of 500m grid (94.33% and 99.53% for Singapore and Boston, respectively), but this percentage can drop down to 75.08% when aggregating on 1km grid level in Singapore. However, after adding the knowledge about just one more location a user frequently visited, more than 90% of all users can be uniquely identified no matter which spatial aggregation is used. These results show that most of the users have unique traces in a sense that a few top locations they visited distinguish them from the rest of the users.

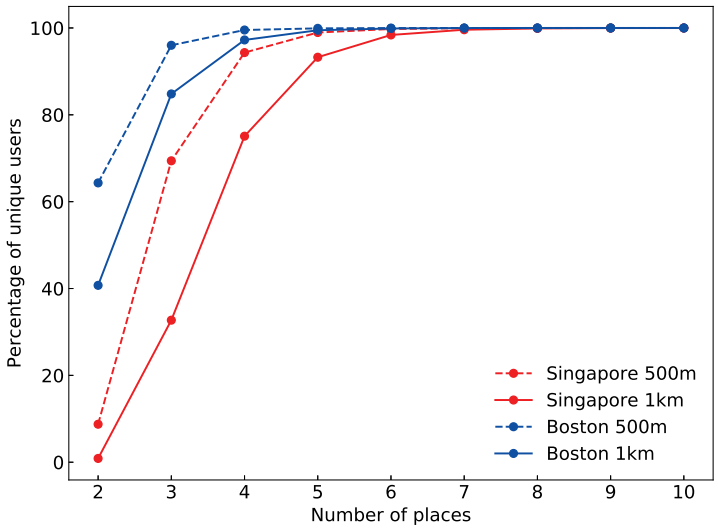

Figure 12: The percentage of all unique users, who were left after filtering out inactive users, both in the Boston and Singapore datasets, aggregating on 500m and 1km grid levels.

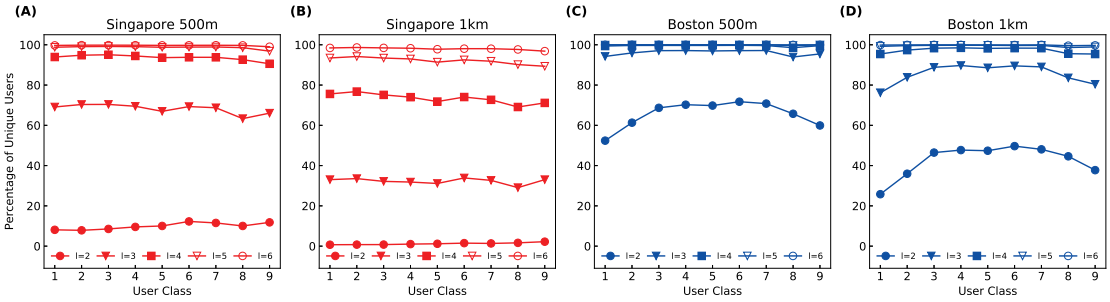

Figure 13: The percentage of unique users in each of nine socioeconomic classes, who were left after filtering out inactive users, both in the Boston and Singapore datasets aggregating on 500m and 1km grid levels.

We next explore the relationships between unicity and SES in the two cities. For each city, we divide users into nine classes using the social stratum model and for each value of the number of top locations l, we compute the percentage of users that have been uniquely identified in each class. Figure 13 shows the results for both cities. For Singapore, as shown in Figure 13A and Figure 13B, the unicity value seems to be independent of SES under both spatial resolutions, no matter what value of l (from 2 to 6) is chosen. The results suggest that although people who belong to different social classes might prefer to visit different types of places in the city, such difference does not cause a particular class (e.g., rich people) to be more identifiable than others. For Boston, when l equals 2 or 3, it is found that the percentages of unique users in the lower (e.g., 1^{st}) and upper (e.g., 9^{th}) classes are lower. That means the users in the poorer or richer classes in the city are less likely to be uniquely identified when considering only the top few activity locations (e.g., home and work).

6 Discussion and Conclusion

By coupling large scale mobile phone and urban socioeconomic datasets, this study introduces an analytical framework to better understand human mobility patterns and their relationships with travelers' socioeconomic status (SES). Six mobility indicators, which include (1) radius of gyration, (2) number of activity locations, (3) activity entropy, (4) travel diversity, (5) k-radius of gyration, and (6) unicity, are derived to quantify important aspects of mobile phone users' mobility characteristics. We then propose a data fusion approach to approximating, at an aggregate level, the SES of mobile phone users. Using Singapore and Boston as case studies, we compare the statistical properties of the six mobility indicators in the two cities and analyze how they vary across socioeconomic classes.

The analysis results provide a multifaceted view of the relationships between mobility and SES. By first examining radius of gyration, a measure that quantifies the spatial dispersion of individual daily activities, we find that phone user groups that are generally richer tend to travel shorter in Singapore but longer in Boston. The contradictory findings in the two cities suggest a complicated relationship between travel distance and wealth level. Such contradictions have also been discussed in previous studies. For example, some studies suggest a positive correlation between economic status and travel distance (e.g., commuting distance) [59, 60] while some others observe an opposite trend mostly in smaller US cities [61] and Europe [62]. These differences have deep roots in spatial arrangement of cities, which have been studied intensively during the last century. Works by Alonso [63], Mills [64] and Muth [65] developed a bid rent theory that explains concentric zones model proposed earlier [66]. These models assume the monocentric and isotropic city. They could explain pretty well the spatial form of cities in the United States, where traditionally wealthier people live in suburbs and more disadvantaged occupy city centers. Although this arrangement is changing and recent studies argue that the members of the advantaged class tend to take up central locations [67], our results still show that richer people travel further to access amenities in Boston. At the same time, there are studies arguing that these classical models, developed and shown to be working for North America, do not work for Europe and Latin America, where their predicted distributions of social classes are not observed [68, 69]. One possible reason for this is that not all cities can be explained by the same model. Singapore is a polycentric city and there are models that better describe such cities [70, 71]. Other attempts to explain observed phenomena include models involving public transportation [72] or population density and segregation [73]. An amenity-based theory [74] ties location by income to city's idiosyncratic characteristics. Its authors use Paris as an example of a city where center is populated by higher social classes in contrary to Detroit. They argue that the urban amenities of some city centers are so attractive that the wealthy want to stay. This is disputed by the authors of "Why Do The Poor Live In Cities?" [72], who present Paris as an exception from the rule stated in the title of their paper. They explain this exception by good public transportation connecting suburbs with the city. The same is true for Singapore, which has a very good public transportation system that connects all parts of the island very well. It is also a very special case of city-island-state strictly constrained in land, so it cannot allow for vast and sparse suburbs, but there are areas with expensive landed houses relatively close to the center.

By investigating the relationships between SES and three other mobility indicators — the number

of activity locations, activity entropy, and travel diversity — we find that for both cities, phone users across different socioeconomic classes exhibit very similar characteristics. The results suggest that wealth level, at least in Singapore and Boston, is not a factor that restricts how people travel around in the city. Note that a recent study based on mobile phone data collected in France observes a positive correlation between travel diversity and per capita income [11]. The finding is different from our study and is worth discussing. First, the two studies are conducted at different scales (i.e., national scale vs. city scale), which make the results to some extent incomparable. Second, it is important to mention that both Singapore and Boston are highly-developed cities with efficient public transportation systems. This enables people to travel among destinations conveniently even without relying on automobiles. It is thus meaningful to repeat our analysis in other cities with underdeveloped public transportation or where certain societal issues (e.g., poverty) are pronounced. This would yield a more comprehensive view of the mobility gap among socioeconomic classes in different types of cities.

The comparison between radius of gyration (R_g) and k-radius of gyration ($R_g^{(k)}$) enables us to better understand to what extent a phone user's top activity locations describe his/her overall activity space. According to the analysis results, for both cities, 2-returners (defined as phone users with $R_g^{(2)} \geq R_g/2$) already represent more than 70% of the phone users, which indicates that the top two activity locations to a large extent capture the overall activity space of the majority of the populations. By gradually increasing the value of k, it is found that a higher number of activity locations (e.g., $k = 4$ and $k = 8$) better explains the overall activity space of phone users in Singapore than in Boston. That means people in Singapore are more likely to travel beyond the top two locations (e.g., home and work place) for their daily activities. By further examining the relationship between SES and percentage of strong returners (defined as phone users with $0.9 \leq R_g^{(2)}/R_g \leq 1.1$), we find that in Boston the percentage increases with socioeconomic class while in Singapore it generally decreases as people become richer. That means compared to Singapore, the home-work circle of richer people in Boston tend to be more inclusive of other types of individual activities.

Finally, the unicity test explores to what extent phone users can be uniquely identified based on the top l activity locations visited. The results suggest that in Singapore, the percentage of phone users that can be uniquely identified are very similar across different socioeconomic classes, no matter what value of l is chosen. However, in Boston, the percentages of unique users in the lower and upper classes are lower when l equals 2 or 3. This suggests that the users in the poorer or richer classes in the city are less likely to be uniquely identified when considering only the top few activity locations. While it is difficult to find out the reasons at this moment, we think it is worthwhile to further examine, in our future research, whether social segregation plays a critical role in this. For example, if both the rich and the poor tend to share the top few activity locations with similar others, this would cause both social groups to be less unique in some sense.

Our comparative analysis shows that the relationship between mobility and SES could vary among cities. It also indicates that certain mobility indicators (e.g., travel diversity and the number of visited locations), which have been used in previous studies to predict socioeconomic development [50, 12], might not be effective in particular types of cities. This is also why mobility characteristics are usually combined with other sociality indicators (e.g., number of phone calls/social contacts) in the prediction algorithms.

We want to point out a few limitations of this research. Although the socioeconomic status of mobile phone users is approximated at an aggregate level, the approximations are not perfect and there are still some uncertainties. The method we use for inferring home locations cannot provide perfect results, although they were shown to correlate very well with the official statistics [44]. In Singapore, we associated phone users with the average housing price at their home cells (i.e., Voronoi polygons). For some areas, especially where the densities of cellphone towers are low, uncertainties could be introduced when residential properties from different price levels are highly mixed. In Boston, we used per capita income at census tract level provided by the American Community Survey. The income data based on surveys can also introduce uncertainties because of issues such as self-report errors [75]. Moreover, since only one monetary variable is used in each city, we were unable to capture other sociodemographic characteristics such as phone users' education status,

family size, and household ownership. These are all important dimensions of SES that can be considered in future studies. Another point is related to the potential bias of the mobile phone datasets. Although the datasets capture a large proportion of residents in the two cities, certain demographic tiers, for example, the elders who tend to use mobile phones less frequently, might be underrepresented. Lastly, the high-level mobility indicators were based on location sequences derived from the trajectory segmentation, which improves the quality of estimates of the stay locations compared to estimates using the raw data. But those indicators still capture a partial view of individual daily mobility patterns due to the sparsity of the mobile phone datasets used in this research. In the future, this issue can be overcome by incorporating mobility datasets with finer temporal granularity. Nevertheless, this research contributes to the broad field of human mobility analysis by introducing a framework that integrates large scale mobile phone and socioeconomic datasets. The framework can be applied to other cities to better understand human travel behaviour and activity patterns as well as their links with socioeconomic development.

Acknowledgement

We would like to thank Daniel Kondor for his help with preprocessing the Boston datasets. We also thank 99.co for providing the housing price data in Singapore.

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its CREATE programme, Singapore–MIT Alliance for Research and Technology (SMART) Future Urban Mobility (FM) IRG. The authors thank MIT SMART Program, Accenture China, Allianz, American Air Liquide, Emirates Integrated Telecommunications Company, ENEL Foundation, Ericsson, Kuwait-MIT Center for Natural Resources and the Environment, Liberty Mutual Institute, Regional Municipality of Wood Buffalo, Volkswagen Electronics Research Lab, UBER, and all the members of the MIT Senseable City Lab Consortium for supporting this research. We also acknowledge the support of research project "Big Data as A Testbed for Urban Theories", funded by the Hong Kong Polytechnic University Start-Up Grant under project 1-BE0J.

Appendices

A Phone users' calling activity patterns.

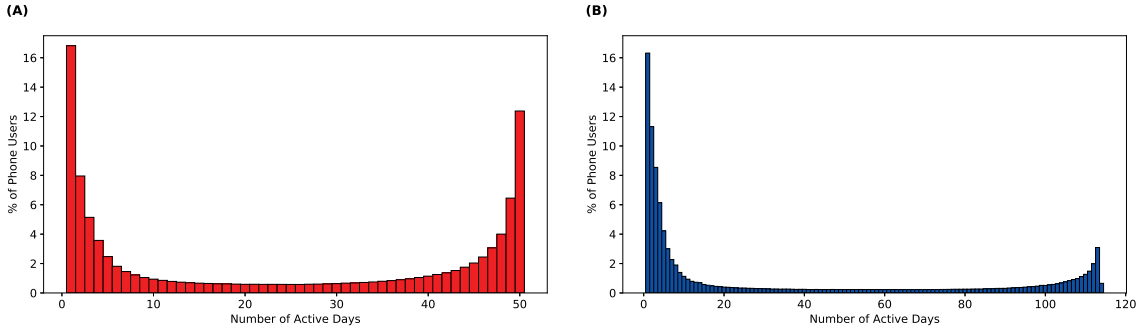

Figure A.1: The distribution of users depending on number of days when they had recorded activity for the Singapore (A) and Boston (B) datasets.

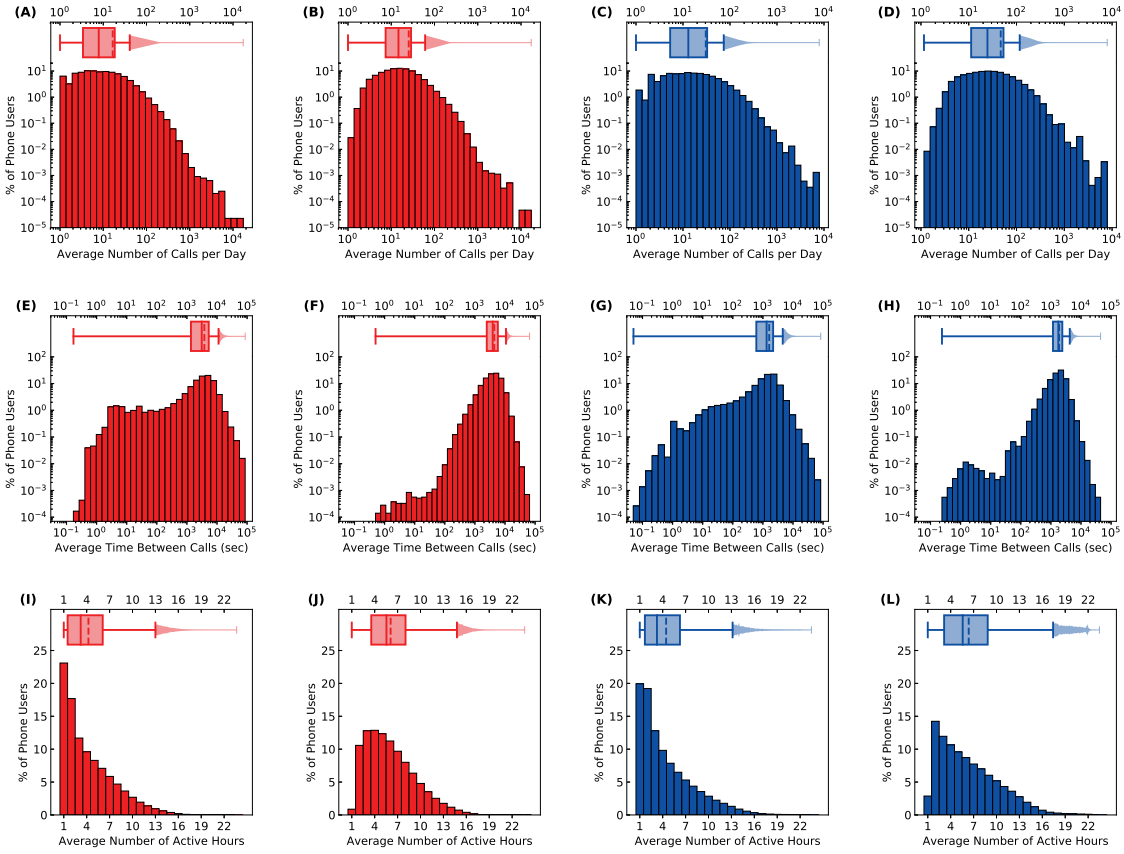

Figure A.2: The distribution of the proportion of users depending on the average (A-D) number of calls per day, (E-H) time between calls within each active day, (I-L) number of active hours per day in Singapore (red) and Boston (blue) before (first and third column) and after (second and forth column) filtering inactive users.

B Number of housing properties that fall within each cellphone tower service area.

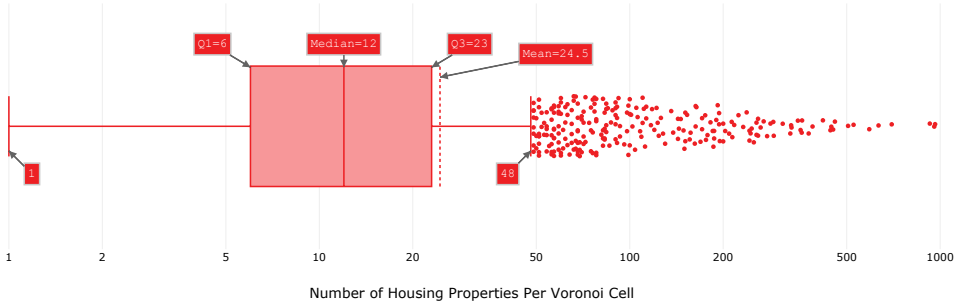

Figure B.3: Distribution of the number of housing properties within each cellphone tower service area.

C Relationships between phone users' SES and radius of gyration computed on weekdays/weekends.

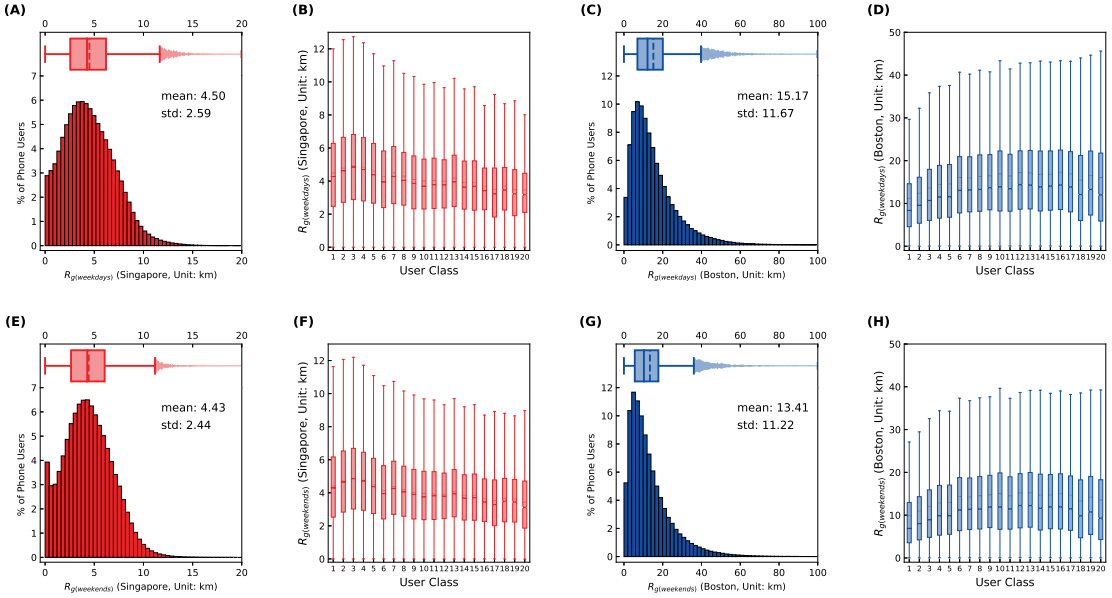

Figure C.4: (A-D) The histogram of $R_{g(weekdays)}$ in the Singapore (red) and Boston (blue) datasets and their relationship with user classes (box plots for 20 classes; whiskers indicate 1.5 interquartile range, mean is shown as dotted line, notch around median shows its standard error); (E-H) The same for $R_{g(weekends)}$.

D K-Radius of gyration for higher values of k.

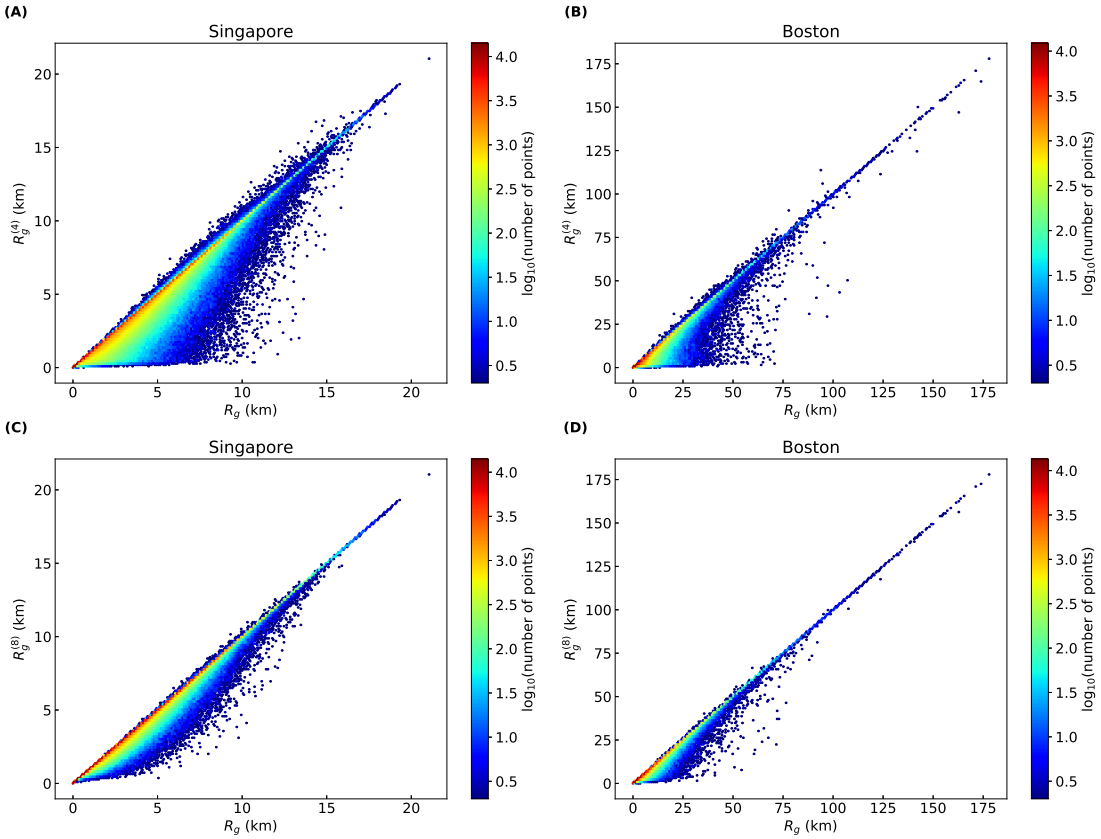

Figure D.5: 4- and 8-radiuses of gyration versus overall gyradiuses in Singapore and Boston.

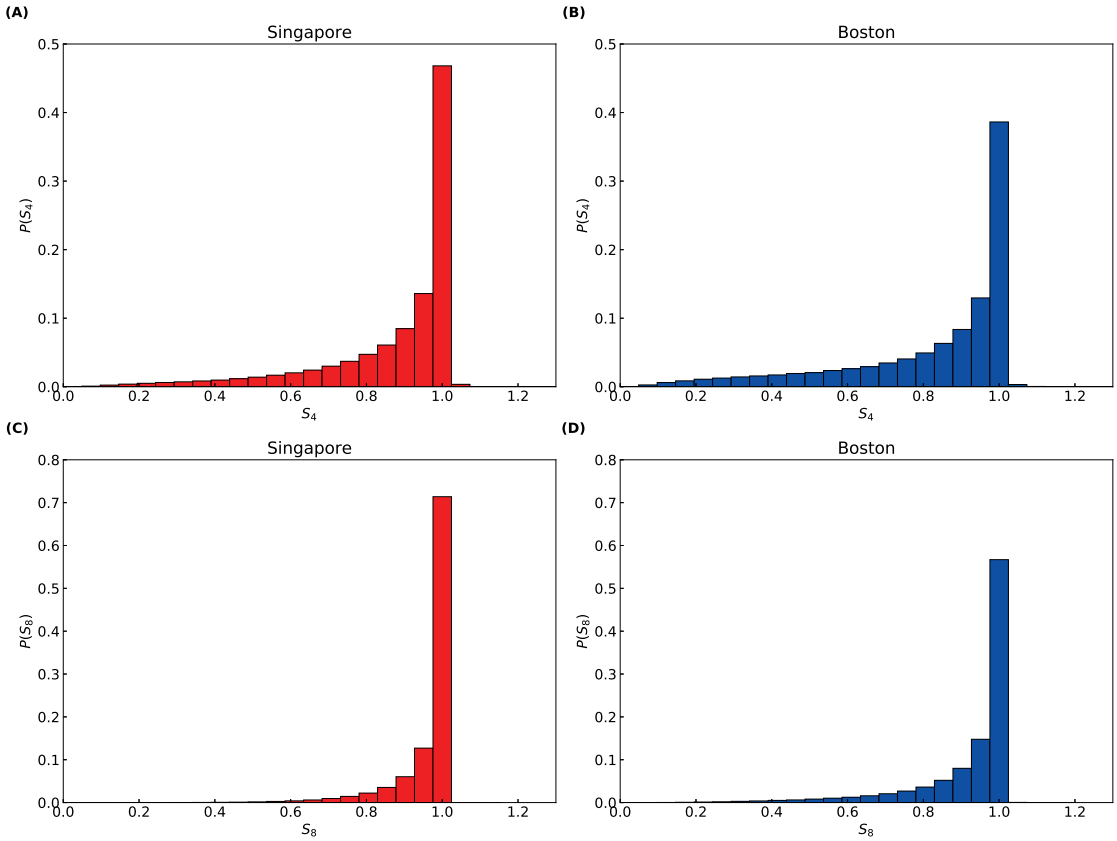

Figure D.6: A distribution of the ratio $S_k = R_g^{(k)}/R_g$ for $k = 4$ and $k = 8$. A peak close to the right indicates a high number of strong k-returners.

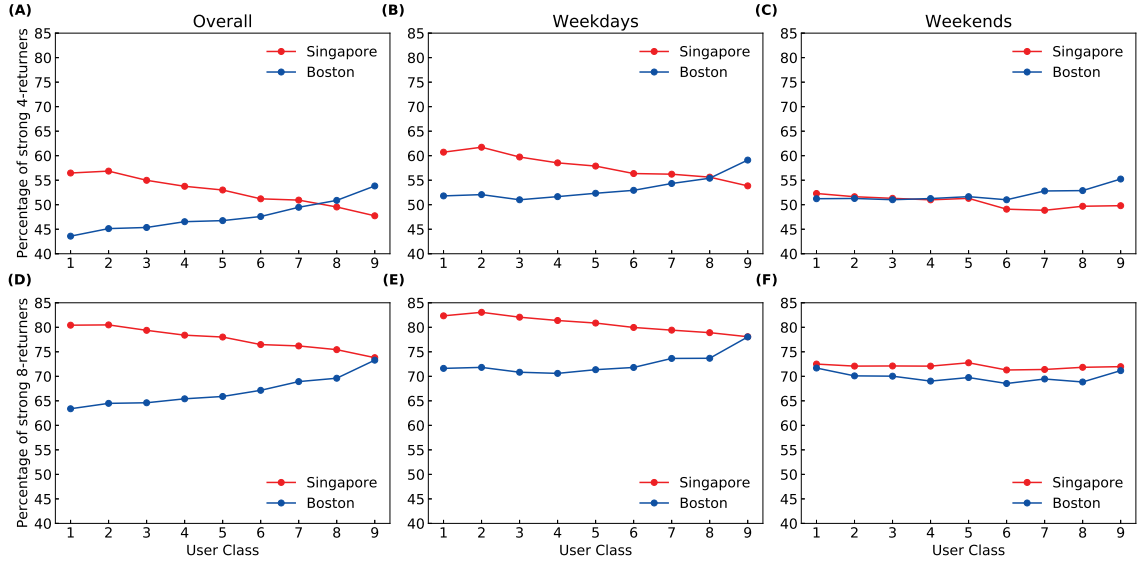

Figure D.7: A percentage of strong k-returners ($k = 4, 8$) in each class overall and on weekdays and weekends separately showing opposite trends in Boston and Singapore.

References

[1] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. Von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti," *PLoS Med*, vol. 8, no. 8, p. e1001083, 2011.

[2] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Life in the network: the coming age of computational social science," *Science (New York, NY)*, vol. 323, no. 5915, p. 721, 2009.

[3] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data," *Transportation research part c: emerging technologies*, vol. 58, pp. 240–250, 2015.

[4] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[5] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[6] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 2013.

[7] F. Finger, T. Genolet, L. Mari, G. C. de Magny, N. M. Manga, A. Rinaldo, and E. Bertuzzo, "Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks," *Proceedings of the National Academy of Sciences*, p. 201522305, 2016.

[8] R. Alsnih and D. A. Hensher, "The mobility and accessibility expectations of seniors in an aging population," *Transportation Research Part A: Policy and Practice*, vol. 37, no. 10, pp. 903–916, 2003.

[9] F. Echenique and R. G. Fryer, "A measure of segregation based on social interactions," *The Quarterly Journal of Economics*, vol. 122, no. 2, pp. 441–485, 2007.

[10] Y. Leo, E. Fleury, J. I. Alvarez-Hamelin, C. Sarraute, and M. Karsai, "Socioeconomic correlations and stratification in social-communication networks," *Journal of The Royal Society Interface*, vol. 13, no. 125, p. 20160598, 2016.

[11] L. Pappalardo, D. Pedreschi, Z. Smoreda, and F. Giannotti, "Using big data to study the link between human mobility and socio-economic development," in *Big Data (Big Data), 2015 IEEE International Conference on*, pp. 871–878, IEEE, 2015.

[12] A. Almaatouq, F. Prieto-Castrillo, and A. Pentland, "Mobile communication signatures of unemployment," in *International Conference on Social Informatics*, pp. 407–418, Springer, 2016.

[13] V. Frias-Martinez, C. Soguero-Ruiz, E. Frias-Martinez, and M. Josephidou, "Forecasting socioeconomic trends with cell phone records," in *Proceedings of the 3rd ACM Symposium on Computing for Development*, p. 15, ACM, 2013.

[14] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

[15] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási, "Returners and explorers dichotomy in human mobility," *Nature communications*, vol. 6, 2015.

[16] R. G. Golledge and R. J. Stimson, *Spatial behavior: A geographic perspective*. Guilford Press, 1997.

[17] S. Schönfelder and K. W. Axhausen, "Activity spaces: measures of social exclusion?," *Transport policy*, vol. 10, no. 4, pp. 273–286, 2003.

[18] Y. Zehavi, *The" UMOT" project*. US Department of Transportation, 1981.

[19] D. W. Lefever, "Measuring geographic concentration by means of the standard deviational ellipse," *American Journal of Sociology*, vol. 32, no. 1, pp. 88–94, 1926.

[20] S. Schönfelder and K. Axhausen, "Structure and innovation of human activity spaces," *Arbeitsberichte Verkehrs-und Raumplanung*, vol. 258, pp. 1–40, 2004.

[21] H.-M. Kim and M.-P. Kwan, "Space-time accessibility measures: A geocomputational algorithm with a focus on the feasible opportunity set and possible activity duration," *Journal of Geographical Systems*, vol. 5, no. 1, pp. 71–91, 2003.

[22] H. J. Miller, "A measurement theory for time geography," *Geographical analysis*, vol. 37, no. 1, pp. 17–45, 2005.

[23] M. Dijst, "Two-earner families and their action spaces: A case study of two dutch communities," *GeoJournal*, vol. 48, no. 3, pp. 195–206, 1999.

[24] T. H. Newsome, W. A. Walcott, and P. D. Smith, "Urban activity spaces: Illustrations and application of a conceptual model for integrating the time and space dimensions," *Transportation*, vol. 25, no. 4, pp. 357–377, 1998.

[25] M.-P. Kwan, A. T. Murray, M. E. O'Kelly, and M. Tiefelsdorf, "Recent advances in accessibility research: Representation, methodology and applications," *Journal of Geographical Systems*, vol. 5, no. 1, pp. 129–138, 2003.

[26] J. E. Sherman, J. Spencer, J. S. Preisser, W. M. Gesler, and T. A. Arcury, "A suite of methods for representing activity space in a healthcare accessibility study," *International Journal of Health Geographics*, vol. 4, no. 1, p. 24, 2005.

[27] N. Eagle and A. S. Pentland, "Eigenbehaviors: Identifying structure in routine," *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, pp. 1057–1066, 2009.

[28] C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams, "Mobile landscapes: using location data from cell phones for urban analysis," *Environment and Planning B: Planning and Design*, vol. 33, no. 5, pp. 727–748, 2006.

[29] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, 2011.

[30] G. D. Nelson and A. Rae, "An economic geography of the united states: from commutes to megaregions," *PloS one*, vol. 11, no. 11, p. e0166083, 2016.

[31] A. Belyi, I. Bojic, S. Sobolevsky, I. Sitko, B. Hawelka, L. Rudikova, A. Kurbatski, and C. Ratti, "Global multi-layer network of human mobility," *International Journal of Geographical Information Science*, vol. 31, no. 7, pp. 1381–1402, 2017.

[32] C. Roth, S. M. Kang, M. Batty, and M. Barthélemy, "Structure of urban movements: polycentric activity and entangled hierarchical flows," *PloS one*, vol. 6, no. 1, p. e15923, 2011.

[33] T. Louail, M. Lenormand, O. G. Cantú, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, "From mobile phone data to the spatial structure of cities," *Scientific Reports*, vol. 4, no. 5276, 2014.

[34] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Communications of the ACM*, vol. 56, no. 1, pp. 74–82, 2013.

[35] Y. Xu, S.-L. Shaw, Z. Zhao, L. Yin, F. Lu, J. Chen, Z. Fang, and Q. Li, "Another tale of two cities: Understanding human activity space using actively tracked cellphone location data," *Annals of the American Association of Geographers*, vol. 106, no. 2, pp. 489–502, 2016.

[36] M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González, "Development of origin–destination matrices using mobile phone call data," *Transportation Research Part C: Emerging Technologies*, vol. 40, pp. 63–74, 2014.

[37] S. Jiang, J. Ferreira, and M. C. Gonzales, "Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore," *IEEE Transactions on Big Data*, 2016.

[38] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González, "The timegeo modeling framework for urban motility without travel surveys," *Proceedings of the National Academy of Sciences*, p. 201524261, 2016.

[39] M.-P. Kwan, "Gender, the home-work link, and space-time patterns of nonemployment activities," *Economic geography*, vol. 75, no. 4, pp. 370–394, 1999.

[40] S. Hanson and P. Hanson, "Gender and urban activity patterns in Uppsala, Sweden," *Geographical Review*, pp. 291–299, 1980.

[41] S. Hanson and P. Hanson, "The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics," *Economic geography*, vol. 57, no. 4, pp. 332–347, 1981.

[42] S. Hanson, "The determinants of daily travel-activity patterns: relative location and sociodemographic factors," *Urban Geography*, vol. 3, no. 3, pp. 179–202, 1982.

[43] N. Limtanakool, M. Dijst, and T. Schwanen, "The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium-and longer-distance trips," *Journal of transport geography*, vol. 14, no. 5, pp. 327–341, 2006.

[44] Y. Xu, A. Belyi, I. Bojic, and C. Ratti, "How friends share urban space: An exploratory spatiotemporal analysis using mobile phone data," *Transactions in GIS*, vol. 21, no. 3, pp. 468–487, 2017.

[45] O. Järv, K. Müürisepp, R. Ahas, B. Derudder, and F. Witlox, "Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia," *Urban Studies*, vol. 52, no. 14, pp. 2680–2698, 2015.

[46] S. Silm and R. Ahas, "Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data," *Annals of the Association of American Geographers*, vol. 104, no. 3, pp. 542–559, 2014.

[47] Y. Xu, S.-L. Shaw, Z. Zhao, L. Yin, Z. Fang, and Q. Li, "Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach," *Transportation*, vol. 42, no. 4, pp. 625–646, 2015.

[48] Q. Huang and D. W. Wong, "Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us?," *International Journal of Geographical Information Science*, vol. 30, no. 9, pp. 1873–1898, 2016.

[49] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[50] C. Smith-Clarke, A. Mashhadi, and L. Capra, "Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 511–520, ACM, 2014.

[51] S. Šćepanović, I. Mishkovski, P. Hui, J. K. Nurminen, and A. Ylä-Jääski, "Mobile phone call data as a regional socio-economic proxy indicator," *PloS one*, vol. 10, no. 4, p. e0124160, 2015.

[52] V. Soto, V. Frias-Martinez, J. Virseda, and E. Frias-Martinez, "Prediction of socioeconomic levels using cell phone records," in *International Conference on User Modeling, Adaptation, and Personalization*, pp. 377–388, Springer, 2011.

[53] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example," *Transportation research part C: emerging technologies*, vol. 26, pp. 301–313, 2013.

[54] B. C. Csáji, A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel, "Exploring the mobility of mobile phone users," *Physica A: statistical mechanics and its applications*, vol. 392, no. 6, pp. 1459–1473, 2013.

[55] A.-L. Barabási, *Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades*. Penguin, 2010.

[56] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *Journal of Urban Technology*, vol. 17, no. 1, pp. 3–27, 2010.

[57] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *International Conference on Pervasive Computing*, pp. 133–151, Springer, 2011.

[58] I. Bojic, E. Massaro, A. Belyi, S. Sobolevsky, and C. Ratti, "Choosing the right home location definition method for the given dataset," in *Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings* (T.-Y. Liu, C. N. Scollon, and W. Zhu, eds.), pp. 194–208, Cham: Springer International Publishing, 2015.

[59] J. O. Wheeler, "Occupational status and work-trips: A minimum distance approach," *Social Forces*, vol. 45, no. 4, pp. 508–515, 1967.

[60] D. L. Poston Jr, "Socioeconomic status and work-residence separation in metropolitan america," *Pacific Sociological Review*, vol. 15, no. 3, pp. 367–380, 1972.

[61] T. A. Maraffa and S. R. Brooker-Gross, "Aspects of the journey-to-work within a small city laborshed," *Urban Geography*, vol. 5, no. 2, pp. 178–186, 1984.

[62] A. Aguiléra, S. Wenglenski, and L. Proulhac, "Employment suburbanisation, reverse commuting and travel behaviour by residents of the central city in the paris metropolitan area," *Transportation Research Part A: Policy and Practice*, vol. 43, no. 7, pp. 685 – 691, 2009.

[63] W. Alonso *et al.*, *Location and land use*. Harvard University Press Cambridge, MA, 1964.

[64] E. S. Mills, "An aggregative model of resource allocation in a metropolitan area," *The American Economic Review*, vol. 57, no. 2, pp. 197–210, 1967.

[65] R. F. Muth, *Cities and Housing*. University of Chicago Press, Chicago, 1969.

[66] R. E. Park and E. W. Burgess, *The city*. University of Chicago Press, 1925.

[67] R. Florida and P. Adler, "The patchwork metropolis: The morphology of the divided postindustrial city," *Journal of Urban Affairs*, pp. 1–16, 2017.

[68] P. M. Hohenberg and L. H. Lees, *The making of urban Europe, 1000–1994*. Harvard University Press, 1995.

[69] G. K. Ingram and A. Carroll, "The spatial structure of latin american cities," *Journal of urban Economics*, vol. 9, no. 2, pp. 257–273, 1981.

[70] C. D. Harris and E. L. Ullman, "The nature of cities," *The Annals of the American Academy of Political and Social Science*, vol. 242, no. 1, pp. 7–17, 1945.

[71] R. Louf and M. Barthelemy, "Modeling the polycentric transition of cities," *Physical review letters*, vol. 111, no. 19, p. 198702, 2013.

[72] E. L. Glaeser, M. E. Kahn, and J. Rappaport, "Why do the poor live in cities? the role of public transportation," *Journal of urban Economics*, vol. 63, no. 1, pp. 1–24, 2008.

[73] R. Louf and M. Barthelemy, "Patterns of residential segregation," *PloS one*, vol. 11, no. 6, p. e0157476, 2016.

[74] J. K. Brueckner, J.-F. Thisse, and Y. Zenou, "Why is central paris rich and downtown detroit poor?: An amenity-based theory," *European economic review*, vol. 43, no. 1, pp. 91–107, 1999.

[75] P. Eckerstorfer, J. Halak, J. Kapeller, B. Schütz, F. Springholz, and R. Wildauer, "Correcting for the missing rich: An application to wealth survey data," *Review of Income and Wealth*, vol. 62, no. 4, pp. 605–627, 2016.