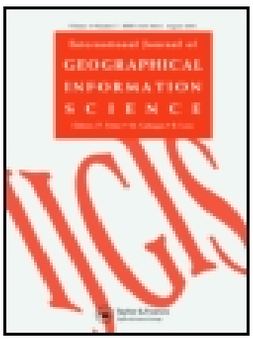

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study

Yihong Yuan & Martin Raubal

To cite this article: Yihong Yuan & Martin Raubal (2016): Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study, International Journal of Geographical Information Science

To link to this article: http://dx.doi.org/10.1080/13658816.2016.1143555

Published online: 12 Feb 2016.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2016
http://dx.doi.org/10.1080/13658816.2016.1143555

Taylor & Francis
Taylor & Francis Group

Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study

Yihong Yuan[a] and Martin Raubal[b]

[a]Department of Geography, Texas State University, San Marcos, TX, USA; [b]Institute of Cartography and Geoinformation, ETH Zurich, Zurich, Switzerland

ABSTRACT

Travel activities are embodied as people's needs to be physically present at certain locations. The development of Information and Communication Technologies (ICTs, such as mobile phones) has introduced new data sources for modeling human activities. Based on the scattered spatiotemporal points provided in mobile phone datasets, it is feasible to study the patterns (e.g., the scale, shape, and regularity) of human activities. In this paper, we propose methods for analyzing the distribution of human activity space from both individual and urban perspectives based on mobile phone data. The Weibull distribution is utilized to model three predefined measurements of activity space (radius, shape index, and entropy). The correlation between demographic factors (age and gender) and the usage of urban space is also tested to reveal underlying patterns. The results of this research will enhance the understanding of human activities in different urban systems and demographic groups, as well as providing novel methods to expand the important and widely applicable area of geographic knowledge discovery in the age of instant access.

ARTICLE HISTORY
Received 9 September 2014
Accepted 31 December 2015

KEYWORDS
Activity space; human mobility; mobile phones; Weibull distribution; data mining; Big (geo)data

1. Introduction

Modeling the basic laws of human mobility has become an important research question in various fields such as physics, transportation, and geographic information science (GIS). Researchers have employed different models to quantify the distribution of human mobility indicators; however, most existing models (e.g., random walk and its numerous derivatives) have concentrated on the displacement (i.e., step size) and direction distribution of human activities in abstract models; therefore, they are not directly applicable to quantify the distribution of other activity indicators, such as the size and shape of human activity space, in a real geographic environment.

The measurement of activity space is a crucial topic when studying both the spatial distribution of individual behavior and the aggregated activity patterns of urban systems. In urban geography, activity space is often defined as the local areas within which people travel during their daily activities (Mazey 1981). Previous research focused on measuring the size, geometry, and inherent structure of human activity space (e.g., the

CONTACT Yihong Yuan yuan@txstate.edu

randomness of activity patterns), as well as the reasons why activity space forms (Golledge and Stimson 1997). In practice, investigating the quantitative properties of human activities often involves model fitting. An appropriate mathematical model provides insights for many application areas, ranging from building a smart system in urban planning and geography to a deeper understanding of the basic laws of human activity in physics (González *et al*. 2008, Song *et al*. 2010). However, the modeling of the distribution of activity space is still an ongoing process. In particular, the study conducted by Kang *et al*. (2012) investigated the patterns of human activity from an urban morphology perspective, where they calculated the radius of gyration (ROG) to represent the scale of user activity space, and correlated the ROG values with the size and shape of corresponding cities. They further developed the spatial heterogeneity constrained Levy flight (SHCLF) model to simulate intra-urban human motion. Their results provide valuable insights regarding the distribution patterns of human mobility and confirm the relations between urban morphology and human mobility patterns; however, this study can be extended from two perspectives: (1) other aspects of human activity (such as the shape and regularity) should also be explored to reveal more detailed patterns; and (2) investigating the activity heterogeneity of different demographic groups provides valuable input for urban planners regarding the usage of activity space in different population groups.

In the big data era, the datasets used in quantitative analyses evolved from limited small data (e.g., socioeconomic survey data/human participant experiments) (Pendyala *et al*. 1991, Harvey and Taylor 2000) to larger datasets contributed by the development of information and communication technologies (ICTs) such as mobile phones. These datasets usually cover a larger study area, therefore providing more comprehensive evidence when studying the whole urban system (Yuan and Raubal 2012). Based on a large call detailed record (CDR) dataset from China, this paper concentrates on the distribution of human mobility from both computational (i.e., model-fitting) and geographical perspectives (i.e., spatial variations). We employ three indicators to represent different aspects (scale, shape, and randomness) of activity behavior: (1) radius, (2) shape index (SI, defined as 1-eccentricity, indicating the degree that an activity space deviates from a straight line), and (3) entropy (indicating the degree of randomness; see Section 3.2.1 for detailed definitions). The first two measure the basic descriptive characteristics of individual activity space, whereas the third one depicts the internal structure of activity space by measuring the regularity of individual trajectories. The objective of this study is to develop a deeper understanding of how individual activity spaces are distributed from two perspectives: (1) From the *methodological perspective*, we explore the possibility of utilizing a flexible and unified distribution (the Weibull distribution) to model all three measurements and perform a comparison; (2) From the *empirical perspective*, individual attributes in large-scale CDR datasets are often missing due to privacy issues; the age and gender data utilized in this research provide valuable information to investigate how activities correlate with the built environment and individual attributes, in particular, exploring age and gender differences in activity spaces. As illustrated, this work can also be considered as an extension of the work conducted by Kang *et al*. (2012) from both methodological and empirical perspectives. The interpretations of distributions provide informative input regarding the determining characteristics of an urban system, as well as generating valuable resources for city

planners to understand urban mobility patterns and travel demand (Fiore *et al*. 2014, Yue *et al*. 2014). It also provides an opportunity to complement traditional mobility models such as random walk (Rhee *et al*. 2011).

The remainder of this paper is organized as follows: Section 2 describes related work in the areas of activity space modeling, mobile phone data analysis, human mobility, and the Weibull distribution and its applications. Section 3 introduces the fundamental research design, including the dataset description and the methodology. Section 4 presents the data analyses and discusses various aspects of the output in detail. We conclude this research and present directions for future work in Section 5.

2. Related work

2.1. *Studies on human activity space*

Among all the activity-based research, the measurement of activity space is an important topic when studying the spatial distribution of individual behavior. Activity space is defined as the local areas within which people travel during their daily activities (Mazey 1981). There are several related concepts in this field, such as the *action space*, defined as the collection of all urban locations about which an individual has subjective utility or preference with (Horton and Reynolds 1971), the *awareness space*, which refers to the places a household had knowledge of before searching for a new neighborhood (Brown and Moore 1970), or *space–time prism*, defined as the set of points that can be reached by an individual given a maximum possible speed from a starting point and an ending point in space–time (Hägerstrand 1970). All these concepts facilitated the studies of human activity space from both the qualitative and quantitative aspects. The former includes social studies related to the definition, nature, and causes of human activities, such as population segmentation and neighborhood assessment (Talen 1999, Knox and McCarthy 2012, Silm and Ahas 2014). The latter one focuses on analyzing activity space from a more computational perspective (e.g., by defining statistical and/or computational models). Generally, the quantitative measurement of activity space depicts its basic characteristics, such as size and shape. For instance, Schönfelder and Axhausen (2002) introduced the concept of intensity estimation to measure the probability of areas visited by a certain person. They used confidence ellipses to approximate a travel probability field for individual travelers. In González *et al*. (2008) and Song *et al*. (2010), activity space is calculated based on the rotation of user trajectories, and the results provided fundamental contributions to understand the basic laws of human mobility (i.e., human activities are predictable). Salingaros (1998) formalized human activities as nodes (e.g., home, work) and edges, and connected them within a network to model the information exchange in an urban system. In addition, other studies also emphasized the reasons of how activity space forms. For instance, as argued by Golledge and Stimson (1997), there are three determinants of activity space for a given individual: (1) home location; (2) regularly visited activity locations (points of interest – POIs) such as work, grocery stores, gym, cinemas, etc.; (3) travel between and around POIs such as the duration of movements between the regularly visited places. The combination of these three factors can be used to describe the development of activity space, as well as studying the causes and effects of human daily activities. Due to privacy issues, it is very

challenging to acquire detailed daily activity information for a large number of the population. As such, researchers have investigated the potential for utilizing various georeferenced datasets like social networking data (Mennis and Mason 2011) and mobile phone data (González *et al*. 2008) to approximate an individual's activity space. The activity space discussed in this research is approximated based on the locations of connected cell towers. Previous studies have demonstrated the feasibility of using such data to model general characteristics of activity space and urban-scale patterns (González *et al*. 2008).

2.2. *Modeling human activity and urban patterns from mobile phone data*

Based on the scattered spatiotemporal points provided in typical mobile phone datasets, it is possible to identify user trajectories through interpolation methods, as well as studying the aggregated patterns of urban systems. In the urban planning field, the development of ICTs has allowed for a new paradigm: smart cities, which concentrate on employing ICTs to achieve sustainable economic development, a higher quality of life, and a wiser management of natural and social resources (Caragliu *et al*. 2009, Miller 2009). This is best exhibited by the analysis of real-time cities by Ratti *et al*. (2007) and the study of behavior analysis and spatiotemporal data mining by Gao *et al*. (2013). The former focused on analyzing aggregated data from cell phones to better understand urban dynamics in real time, while the latter extracted community structures and provided a quantitative framework to identify clusters and interaction patterns.

Undoubtedly, these technologies are a major step forward in identifying and characterizing the clusters, dynamics, and morphology of urban systems. As discussed in Section 1, researchers have focused on the following three aspects when studying the dynamics of urban mobility and the development of regional planning based on mobile phone data, including but not limited to:

- *Urban planning and morphology*: The spatiotemporal characteristics of an urban system can be viewed as a generalization of individual behavior; therefore, mobile phone data also provide new insights into the analysis of the mobility patterns in urban systems. Researchers believe that urban structure has a strong impact on urban-scale mobility patterns, indicating that different areas inside a city are associated with different inhabitants' motion patterns (Gordon *et al*. 1989). Phithakkitnukoon *et al*. (2010) further elaborated this research by discussing the correlation between various urban units and mobility patterns.
- *Urban clusters and spread*: The issue of hotspot clustering patterns has been addressed in numerous studies. For example, researchers have identified the potential of employing mobile phone data in recognizing the clusters of crime activities (Chainey *et al*. 2008, Traunmueller *et al*. 2014). Similar studies have also been conducted in the transportation field to detect traffic congestion (Herrera *et al*. 2010). In those analyses, hotspots are often defined as areas with 'unusually high occurrence of point incidents', and the point observations are sometimes transformed into area measurement, in which hotspots are defined as areas with a high quantity or intensity for a specific attribute (Lu 2000). These density-based analytics methods are usually developed to discover clusters in which the density

of data exceeds a threshold, and to understand the overall trend of point density in the study area (i.e., first-order effects).

- *Urban rhythms*: Although the extraction of aggregated patterns (i.e., hotspots and clusters) offers valuable input for maintaining the sustainability of urban mobility, it fails to provide sufficient information for understanding the 'rhythm' of an urban system (i.e., analyzing these patterns with respect to time). The temporal dimension is considered an important factor for most social activities, therefore understanding the dynamics of the mobility patterns is essential for the management and planning of urban facilities and services. The objective of urban rhythm research is to go a step beyond the aggregation of individual mobility (Schönfelder 2006, Yuan and Raubal 2012, Hasan *et al*. 2013). For instance, by analyzing the time series of mobility aggregation in different urban areas, Yuan and Raubal (2012) discussed outlier time series associated with certain urban districts. In the real-time Rome project (http://senseable.mit.edu/realtimerome/) conducted by the MIT SENSEable City Lab, researchers also studied the temporal pattern of people gathering during special events.

As indicated in Section 1, our research can be considered as an empirical study of the first category. We explore how the central tendency and/or the dispersion of human activities are distributed and how this distribution relates to the characteristics of different cities and demographic groups. The results also provide computational support for real-time city analysis in the age of instant access.

2.3. *The distribution of human mobility*

Borrel *et al*. (2006) summarized existing models, such as the basic random walk model and its many derivatives, which describe basic laws of human activities. However, researchers have distinguished between the two concepts: position and location. The former refers to coordinates majorly captured by positioning technology, whereas the latter refers to both the position and a place where the location belongs (e.g., a city, a town, or a street) (Warf 2010, Groves 2013). There is a slight difference between studying mobility patterns from a 'physics perspective' and a 'geography perspective': In physics, individual trajectories are mostly structured and located in an abstract mathematical coordinate system (not necessarily longitude and latitude); in geography, however, individual trajectories are usually georeferenced within geographic/projected coordinate systems and geocoded to a certain place (e.g., a specific city). The research conducted by Liu *et al*. (2012) found that the direction distribution of human mobility should not be modeled as evenly distributed as in Lévy flight models (Rhee *et al*. 2011), since human activities are highly constrained by the morphology of an urban system. A recent study by Jiang and Yin (2013) also demonstrates how quantitative measures may help to reveal the internal structure of an urban system by introducing a new indicator (*Ht* index). Additionally, in social and cognitive sciences, researchers have investigated the complex nature of activity space with regards to social context, personal relations, and emotions (Cheyne and Efran 1972, Mason and Korpela 2009).

Other research has investigated various types of distributions to explain the complex nature of human activities, including but not limited to:

- *Power law distribution*: Researchers utilized power law to model the distribution of steps in classic random models (Borrel *et al*. 2006, Rhee *et al*. 2011).
- *Exponential law distribution*: Studies have shown that despite the fact that power law has been utilized widely in theoretical studies, people's intra-urban travel in general follows the exponential law, which is often restricted by the built environment in real-world applications (Candia *et al*. 2008, Kang *et al*. 2012).
- *Lognormal distribution*: Azevedo *et al*. (2009) stated that the direction angle variation and the pause time follow a lognormal distribution. Jiang and Jia (2011) also explored the possibility of applying various heavy-tailed distributions such as lognormal distribution to model human movement patterns.

Although various distributions are applicable to human activity modeling, this research utilizes the Weibull distribution to model the three measurements radius, shape index (SI), and entropy for the following reasons:

- *The flexibility of fitting into different shapes of curves*: Mathematically it is more applicable for comparison if the three indicators can be fitted using the same type of distribution. Previous studies have demonstrated the capability of Weibull distribution to adapt into various curve shapes such as bell-shape curves and exponential curves (Weibull 1951).
- *The possibility of interpreting the model parameters from a spatial perspective*: Traditionally, Weibull distributions are most often used to analyze reliability and survival *temporally*, but they are not widely utilized for modeling the *spatial* dimension of human activities (Rinne 2008). One innovative contribution of this research is to extend the interpretation of Weibull parameters to the spatial dimension, i.e., from 'survival in time' to 'accessibility in space', which will be discussed in detail in Sections 2.4 and 3.2.

2.4. *Weibull distribution and its applications*

As mentioned in Section 2.3, the combination of simplicity and flexibility in the shape of the Weibull distribution makes it an effective model for various applications such as industrial engineering (Weibull 1951). The probability density function (PDF) of the Weibull distribution is defined as

$$f(x;\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^{k}} \quad (x \geq 0), \tag{1}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution. With $k = 1$, the Weibull distribution turns into an exponential distribution (see, the Appendix for parameter estimation).

The most common application of the Weibull distribution is to model the life span of an industrial or natural system. For instance, in reliability analysis, the shape parameter k of the Weibull distribution is considered as an indicator of how the failure rate of a system is proportional to a power of time (Rinne 2008). Here failure

rate is defined as the frequency with which an item or system fails per unit of time (Leitch 1995, Neubeck 2004).

- $0 < k < 1$ indicates that the failure rate decreases over time, i.e., unreliable items failing early and the failure rate decreasing over time.
- $k = 1$ indicates that the failure rate does not change over time, i.e., failure can happen randomly at any point during the life span of a system.
- $k > 1$ indicates that the failure rate increases with time. This happens if the system or the item aged out during the process.

Increasing the value of the scale parameter λ has the effect of stretching out the PDF to the right side and decreasing its height (as well as the 'peak' of the PDF), since the area under the curve should be constant (Figure 1).

Although traditionally the failure rate is defined from the temporal perspective, in various fields such as physics and geography, the interaction between temporal and spatial dimensions has provided theoretical and methodological support to more comprehensively interpret human behavior (Ott and Swiaczny 2001). Researchers have also explored the application of Weibull models in a variety of application fields related to space and movement, such as the modeling of wind speed in physical geography (Morgan *et al*. 2011) and precipitation data in hydrology (Singh 1987). However, this method has not been fully addressed in the human mobility field. In this research, we attempt to extend the indication of the Weibull distribution to the spatial perspective to provide an informative interpretation for activity space distributions.

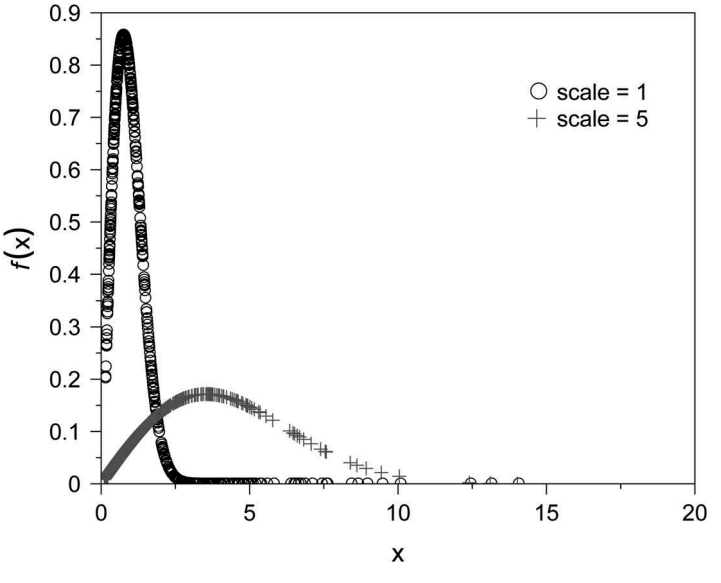

Figure 1. Two Weibull distributions with the same shape parameter ($k = 2$) and different scale parameters ($\lambda = 1$ and $\lambda = 5$).

- The indication of shape parameter *k:* *k* indicates failure rate in the traditional Weibull distribution. Similar to the definition of failure rate, we define 'failure' as 'a certain individual failed to reach a higher level of activeness for a certain mobility measurement (i.e., in this paper the three measurements are radius, SI, and entropy respectively). For radius, 'failed at 2 km' indicates that the user is willing to maintain a daily activity radius of 2 km (without expanding to a broader region). When calculating the aggregated patterns, this can be considered as an indicator of spatial constraints at different scales for a given city or a given population groups, i.e., for radius,
  - $0 < k < 1$ indicates that the failure rate decreases when the radius increases, i.e., inactive individuals failing within a small radius and the failure rate decreasing over space (the distribution is heavy-tailed and people with a large activity space are more persistent and willing to expand their activity space).
  - $k = 1$ indicates that the failure rate does not change across space.
  - $k > 1$ indicates that the failure rate increases when the radius increases. This happens when the urban system has more restrictions on large-scale activities (e.g., people with a large activity space are less persistent or unwilling to expand their activities due to poor public transportation in suburban areas).
- The indication of scale parameter λ: In the Weibull distribution, the scale parameter controls how much the distribution stretches to the right, which can be considered as a measurement of both central tendency and dispersion, therefore it is more comprehensive than using mode/mean/median values or variances separately. For instance, when k is fixed, a population group with a larger λ value for radius distribution has a larger mode value, and the radius range is more dispersed.

In general, the scale parameter λ determines the magnitude of central tendency and the variance of the given variable, while k controls how the variance is distributed along with the value of a variable (generally, a smaller k indicates a heavier-tail distribution and more outliers). Hence, the model parameters provide a feasible method to interpret the mathematical characteristics of human activities at an aggregated level. It is worth noting that each fitted distribution has certain characteristics and affordances – things that it clearly represents versus things that are difficult to examine based on a certain mathematical formulation. Similar to any other distribution fitting, the power of the Weibull distribution is restrained by the limited number of fitted parameters (i.e., λ and k). Therefore, it may not be able to reflect the detailed spatial configuration of a study area. Instead, it represents the absolute magnitude of a mobility variable (e.g., movement radius) and how the variance of this variable is allocated along with the value, as well as providing a subtle indication of how and why the spatial configuration of a study area may result in such λ and k values. Table 1 provides example questions that may or may not be answered based on the Weibull fitting.

Table 1. Weibull distribution – example questions.

Condition: City *X* has a smaller *k* value and the same *λ* value as city *Y* for movement radius		
Questions: Weibull fitting can answer	*Questions: Weibull fitting may be able to answer (with additional urban planning data)*	*Questions: Weibull fitting cannot answer*
Does this mean city *X* has more tail users (users with a relatively large movement radius)?	Does this mean city *X* facilitate long commuting or maybe the local industry requires more people to travel further?	Which direction do users travel the most in city *X*?
Does this mean in city *X* residents with a relatively small movement radius are more likely to expand their activity space compared to long – commuters?		Which area(s) of city *X* are the most clustered?

3. Research design

3.1. *Dataset*

The analysis in this research utilizes a dataset from China, covering around 4.3 million people from 10 densely populated cities in province *H* located in northeast China (Figure 2, due to a signed agreement with our data provider, descriptive statistics of the involved cities can be mentioned; however, the official city names must be removed). The dataset includes mobile phone connection records (both incoming and outgoing calls) for a time span of 9 days (from 21 July 2007 to 29 July 2007). The data include the time, duration, and approximate coordinates of mobile phone connections, as well as the age and gender attributes of a majority of users.[1] Table 2 indicates the

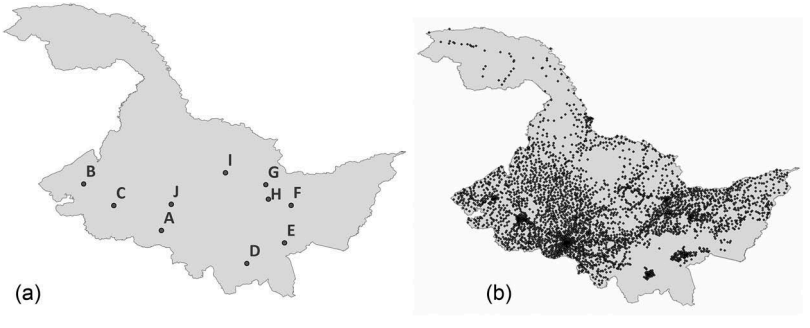

Figure 2. The spatial distributions of: (a) ten cities; (b) the mobile phone towers in Province *H*.

Table 2. Metadata of the 10 cities.

City	Urban area (km^2)	Built-up area (km^2)	Average annual income (10^4 Chinese Yuan)	Urban area population (10^6)	No. of users in dataset (10^6)	No. of records in dataset (10)	Percent of IDs with age and gender (%)
A	7086	336	2.16	4.75	1.70	90.45	84.64
B	4365	135	1.76	1.42	0.41	19.52	85.63
C	5107	169	2.97	1.28	0.66	26.49	96.23
D	1351	64	1.69	0.80	0.34	14.67	96.26
E	2300	79	1.73	0.91	0.27	13.34	95.31
F	1074	60	1.47	0.83	0.30	19.21	94.47
G	4551	43	1.86	0.68	0.19	9.26	98.33
H	1760	62	1.89	0.50	0.19	9.94	94.53
I	19567	159	0.97	0.81	0.09	2.91	95.75
J	2759	28	1.34	0.89	0.13	5.50	86.55

basic information of the 10 cities in the dataset, including size, population of the built-up area, average income[2] and number of users/records in the city.

For each user, the coordinates of the nearest mobile phone tower are recorded both when the user makes and receives a phone call, resulting in a positional data accuracy of about 300–500 m in city centers.

It worth noting that the spatiotemporal data points recorded in mobile phone datasets are neither accurate nor precise. First, the accuracy of positioning data often depends on the density of mobile phone towers in the study area. In this study, the estimation accuracy is higher for cities with a better coverage density of cell towers. Second, the positioning data cannot represent the accurate movement trajectories of each user, since the connected towers are recorded only when a phone-call connection has been established, and during the time span when fewer phone calls occur (e.g., when a user sleeps or drives), the positioning information is not collected. Third, the precision of spatial information varies for different datasets, e.g., a record such as '126.51551E, 45.15153 N' is more precise than '126.52E, 45.15 N'.

Additionally, In CDRs one user may attempt to establish a large number of call connections in a very short time span (e.g., a salesman may make 10 phone calls within 30 min in the office; however, it is not sensible to assume that the salesman visits the office 10 times within 30 min) (Yuan and Raubal 2014). Hence, we eliminated the repeated phone calls made by the same user based on the following rules: For any two consecutive points p_i and p_{i+1} in a given trajectory, if p_i and p_{i+1} are located within the cell of the same mobile phone tower, and the time difference $t_{i+1} - t_i < \Delta T$ (ΔT is a threshold value, in this paper pre-defined as 0.5 h), p_{i+1} is defined as a redundant point and removed. After eliminating the redundant records, we defined each of the remaining records as one 'occurrence' or 'visit' of a certain user. In this paper, all mobility measures are based on this definition.

Another point worth noting is the potential representativeness bias in CDR data. As demonstrated in previous studies (Fuchs and Busse 2009, Forbes 2014), big (geo)data such as location-based social media (LBSM) and georeferenced mobile phone data all have different representativeness issues and sampling biases across various population groups. First, people with limited phone activities are under-represented in CDR data. In this research, we eliminate users who had fewer than 10 phone calls during 9 days. Figure 3 demonstrates the distribution of phone call frequency in the remaining sample set. As can be seen, the majority of the population made 10–50 phone calls during the time span of the study, providing a calculable number of sample points for each user when computing the movement indicators in Section 3.2 (e.g., the activity space of users with only two sample points cannot be approximated as an ellipse). Another potential bias comes from the under/over-representation of demographic groups. For instance, the ratio between males and females in the sample set is 1:1.20, which is lower than the gender ratio (M:F = 1:0.96) published in the official statistical year book of province *H* (citation removed as requested by the data provider). This is potentially due to the higher call frequency of female users. Similar patterns exist for other age groups ('0–14': '15–64': '>64' = 1:1124.6:15.8). Compared to the official statistics ('0–14':'15–64': '>64' = 1:5.94:0.69), the age group '0–14' is heavily underrepresented due to the fact that children and teenagers younger than 14 years old are can rarely own a mobile phone in China. However, despite the potential sampling biases, this large set still

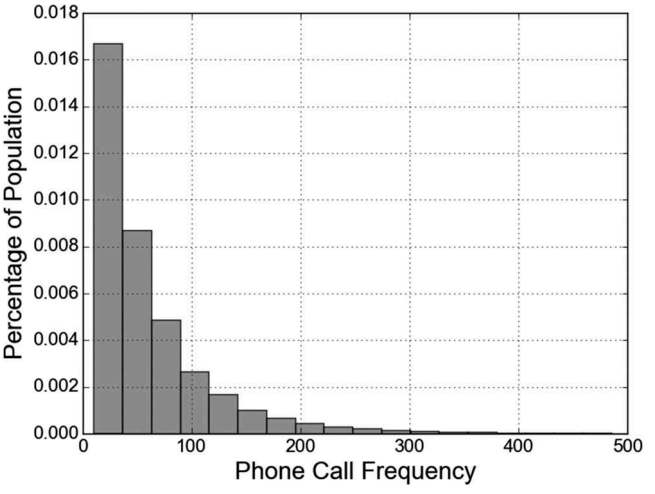

Figure 3. The distribution of phone call frequency.

ensures a statistically reasonable sample size (>5000) for each demographic group to conduct aggregated-level analyses in Section 4.2.2. Additionally, due to the wide usage of mobile phones, CDR data often provide a better demographic representation than LBSM data (Instagram, for instance, particularly attracts adults between the ages of 18 and 29, women and urban dwellers (Forbes 2014)). The detailed analyses on demographic groups are illustrated in Section 4.2.2.

3.2. *Methodology*

3.2.1. *Defining indicators*

Historically, activity space is measured based on various types of methods from an *ellipse-based* representation which focuses on both shape and directional distributions, e.g., standard deviational ellipse, radius of gyration (Song *et al*. 2010), a *convex hull-based* representation which outlines the shape of activity space (Harding *et al*. 2012), a *density-based* representation which provides more details for internal structure, to a *network-based* representation which utilizes road network data to construct paths between points (Sherman *et al*. 2005). As discussed in Section 2.2, activity space is characterized by both external descriptive statistics (e.g., shape, size) and internal structures (e.g., regularity). This research represents three aspects – – scale, shape, and randomness – – of activity behavior (Yuan *et al*. 2012). 'Scale' and 'shape' measure the basic descriptive characteristics of individual activity space, whereas 'randomness' depicts the internal structure of activity space by measuring the regularity of individual trajectories. As illustrated in previous research, the eigenvector-based approaches are capable of capturing both the magnitude and directional distributions of human mobility.

Radius. For each individual, we approximate the physical movement area based on the rotation of user trajectories (González *et al*. 2008). The eigenvectors of trajectories determine the principal axes $\hat{e}_1$ and $\hat{e}_2$, then the trajectories are approximated as ellipses, where

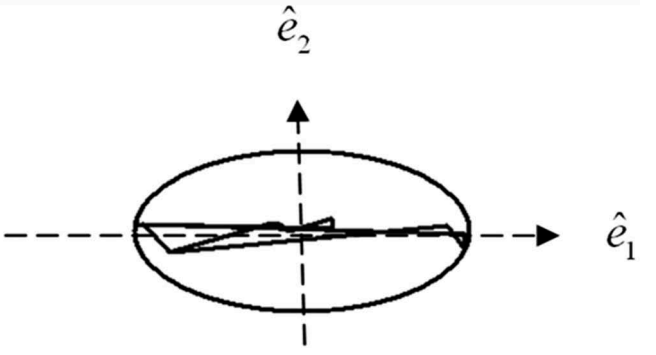

Figure 4. Transformation of trajectories (González *et al*. 2008).

$\hat{e}_1$ and $\hat{e}_2$ are the major and minor axes (Figure 4). For a given ellipse, the average value of the semi-major and semi-minor axes is considered as a measurement of moving radius:

$$R = \frac{|\hat{e}_1| + |\hat{e}_2|}{4}. \tag{2}$$

Shape index (1-eccentricity). Because user trajectories are approximated as ellipses, the movement eccentricity represents how much a particular trajectory deviates from being circular:

$$e = \sqrt{1 - (\frac{|\hat{e}_2|}{|\hat{e}_1|})^2}, \; e \in [0, 1]. \tag{3}$$

For instance, if $e \approx 1$, it is highly possible that the particular person mostly moves between work and home; therefore, the trajectory is close to a regular straight line. Here we are more interested in the bimodal nature of human trajectories, i.e., the fact that most people move between two major POIs in their daily life (Bagrow and Koren 2009). We use 1-eccentricity to represent how the movement deviates from a straight line, defined as the SI in our analysis.

Entropy. Entropy characterizes the heterogeneity of visitation patterns. Based on Song *et al*. (2010), movement entropy is calculated as

$$E = -\sum_{i=1}^{N} p_i \log_2 p_i, \tag{4}$$

where p_i is the probability that location i is visited by the user. N stands for the total number of distinct locations visited in a given trajectory (note that here the locations are only distinguished by coordinate pairs). For example, if a given person A has only visited the two locations P_1 and P_2 in the dataset, and each location has been visited five times, the entropy value is calculated as

$$E_A = -(0.5 * \log_2 0.5 + 0.5 * \log_2 0.5) = 1.0. \tag{5}$$

3.2.2. *Fitting weibull distributions*

This research utilizes the Weibull distribution to model the distribution of three indicators defined in Section 3.2.1. The probability density function is defined in Equation 1. The second step is to construct Weibull distribution models for the three indicators and interpret the variation among age and gender groups. As shown in Section 2.4, the scale parameter λ determines the magnitude of central tendency and the variance of the given variable, while *k* controls how the variance is distributed along with the value of a variable. Hence, the modeling parameters for each city provide a feasible method to compare the characteristics of cities by interpreting the distribution of their residents' activities. Section 4.1 depicts the model fitting process.

3.2.3. *Exploring demographic factors*

As stated by Beckmann (2000) and Nobis *et al*. (2005), human mobility patterns are restricted by individual level factors (e.g., age and gender) and 'supra-individual regime,' which include temporal order (e.g., day and night, the seasons), social conditions (e.g., economic, legal, cultural, and political regulations) and physical configurations (e.g., spatial distribution of urban infrastructure, and transportation networks). Here, we are also interested in how activity space correlates with explanatory factors (e.g., age, gender, and the built environment). Section 4.2 focuses on exploring the similarity and distinctions among age and gender groups in the usage of different urban regions. Kernel density plots are utilized to visualize the clustering patterns of different demographic groups. Since the distinction of urban and suburban area plays a crucial role in land use studies (Lewis 1959, Ahas *et al*. 2010), we also quantify the visiting patterns in urban and suburban areas for age and gender groups.

4. Analyses, results, and discussions

4.1. *Model fitting*

First the distributions are explored visually: the PDF of the three indicators are plotted in Figure 5. There are various types of distributions available for model fitting, such as skewed normal distribution or exponential distribution.

As discussed in Section 2.3, mathematically, it is easier to make a comparison if the three indicators can be fitted using the same type of distribution. In addition, the indication of the Weibull distribution (i.e., shape and scale parameters) extended from survival analysis opens a new perspective to quantitatively interpret human mobility patterns in cities. The model fitting results are listed in Table 3. Here we also employ the one-sample Kolmogorov–Smirnov Test to validate the goodness of fit. The constructed models passed the test at significance level $p = 0.01$.

The results in Table 3 can be interpreted from two perspectives: cross-indicator and cross-city (i.e., compare the fitted *k* and *λ* values horizontally and vertically in Table 3). Tables 4 and 5 list the theoretical and empirical indications of the model parameters. The former presents uniform patterns that exist among all 10 cities (or the majority of the 10 cities) regardless of the specific spatial setting of a certain city, whereas the latter focuses on the detailed and specific differences between cities. Note that it is less

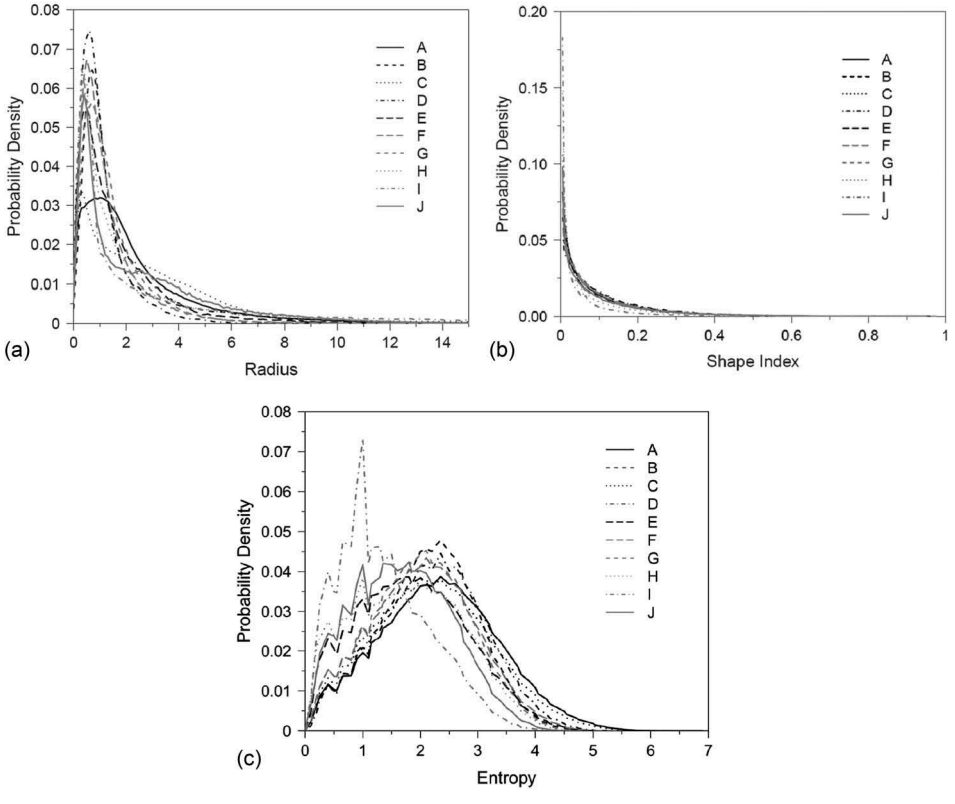

Figure 5. The PDF of the three measures: (a) radius; (b) SI; (c) entropy.

Table 3. The parameters of the Weibull distribution.

	k_1 (Radius)	λ_1 (Radius)	k_2 (SI)	λ_2 (SI)	k_3 (Entropy)	λ_3 (Entropy)
A	1.13	2.62	0.57	0.07	2.41	2.64
B	1.13	1.96	0.57	0.08	2.74	2.4
C	1.03	3.25	0.48	0.05	2.33	2.51
D	1.37	1.24	0.62	0.09	2.64	2.5
E	1.15	1.84	0.41	0.05	2.10	2.13
F	1.18	1.52	0.51	0.07	2.42	2.32
G	1.35	1.65	0.52	0.07	2.52	2.27
H	0.98	2.06	0.32	0.02	1.98	2.02
I	0.73	2.96	0.22	0.01	1.86	1.58
J	1.03	2.65	0.33	0.04	2.13	1.94
Average	1.11	2.18	0.46	0.055	2.31	2.23

meaningful to compare magnitude or dispersion between different variables (i.e., the λ values) when the shape parameter k varies, so Tables 4 and 5 focus on illustrating the indication of k values.

As can be seen, the model-fitting based on the Weibull distribution presents a method to mathematically model the distribution of activity space, which provides valuable input for urban-related studies. It is also feasible to identify outlier patterns based on Figure 5. For instance, city *I* behaves very differently compared to the other 9

Table 4. Interpretation of Weibull parameters (cross-indicator).

Model parameter	Theoretical indication	Empirical indication
$k_1 \approx 1$	The distribution of the radius is close to an exponential distribution, which further addresses the arguments from previous research regarding the power law distribution of human movements (Kang *et al*. 2012). The failure rate is close to constant when the radius increases.	The distinct k values of the three indicators (i.e., $k_1 \approx 1$, $k_2 < 1$, $k_3 > 1$) provide quantitative support to compare the empirical distributions of human activities from different perspectives. For instance, the failure rate is constant for radius $k_1 \approx 1$ and indicates that the spatial setting of the cities in the sample set potentially allows short-commuters to expand their movement radius as much as long-commuters. For SI ($k_2 < 1$), a decreasing failure rate shows that unlike radius, the majority of the population maintain (fail at) a very small SI value (i.e., activity close to a straight line). However, those who maintain a large SI (closer to a circle), are more likely to further increase the SI of their activity space compared to users with a smaller SI. This pattern exists for all 10 cities regardless of city settings and is potentially due to the bimodal nature of human mobility – human beings tend to maintain two major destinations (areas) in their daily life (Bagrow and Koren 2009). For movement entropy, $k_3 > 1$ indicates that low-entropy users are more likely to explore and increase the randomness of their activities; however, as entropy further increases, their potential of further expanding the randomness of activities decreases. Note that the increasing of SI and entropy are not necessarily correlated, since people who maintain a round-shaped activity space can also have fixed daily routes and low entropy values. This result can be utilized to better simulate human behavior in travel analysis and agent-based modeling.
$k_2 < 1$	The decay of SI is faster than exponential and the failure rate decreases when SI increases.	
$k_3 > 1$	The distribution of entropy is similar to a skewed normal distribution and the failure rate increases when entropy increases.	
λ values	λ values are mostly used to compare the magnitude and dispersion of each distribution when the k values remain the same; hence, theoretically it is not meaningful to cross-compare λ values for different variables.	N/A

cities for all three measurements, which indicates the potential for interesting constellations in urban patterns as illustrated.

4.2. *Gender and age analysis*

As mentioned in Section 1, human activities are restricted by a multitude of factors, including *super-individual factors* such as the built environment (e.g., cities), and *individual-level factors* such as age and gender. Based on these factors, it is feasible to identify particular patterns for population groups categorized by social attributes.

The remainder of this section focuses on analyzing how individual and super-individual factors affect the distribution of mobility patterns based on the Weibull distribution. Since it is not feasible to discuss the impact of all explanatory factors, we specifically consider two individual level factors (age and gender).

Table 5. Interpretation of Weibull parameters (cross-city).

Model parameter	Theoretical indication	Empirical indication
$k_{1X} < k_{1Y} < 1$ *(in this table, k_{iX}, k_{iY} indicate the k_i values of cities X and Y, i = 1,2,3)*	The failure rate of *X* decreases faster than *Y* when the radius increases.	City *X* indicates a faster decay in the beginning of the curve with persistent outliers who are more likely to further expand their movement radius. For instance, city *I* has the lowest shape parameter (k_1) among all 10 cities. A potential interpretation is that city *I* is a very special case where the majority of the area that falls within the administrative boundary is covered by forests, with a central urban area and multiple small towns/villages located in forests. These increased spatial constraints result in the decreased potential of random movement from much of the population of this 'Forest City', as it is known by its residents. However, a small portion of the residents have to commute between the city and its surrounding villages. This potentially resulted in the heavy-tailed distribution and a low shape parameter for movement radius.
$k_{1X} > k_{1Y} > 1$	The failure rate of *X* increases faster than *Y* when the radius indicator increases.	Compared to city *Y*, users with a smaller radius in city *X* are more likely to explore and increase the spatial coverage of their activities; however, as the radius further increases, their potential of further expanding the movement scale decreases. This may happen when the city has a fully functioning downtown area and a smaller portion of outlier population (e.g., long commuters), such as cities *D* and *G* – where users are less likely to travel to suburban areas (this finding is further confirmed by the urban/suburban ratio analysis in Section 4.2.2).
$k_{1X} < 1 < k_{1Y}$	The failure rate of *X* decreases whereas the failure rate of *Y* increases when the radius indicator increases.	Similar to the first two cases, city *X* indicates a faster decay in the beginning of the curve and a heavy-tailed distribution with persistent outliers who are more likely to further expand their movement radius, whereas in City *Y*, users with a smaller radius are more likely to explore and increase the spatial coverage of their activities.
$k_{2X} < k_{2Y} < 1$	The failure rate of *X* decreases faster than *Y* when the SI indicator increases.	City *X* indicates a faster decay in the beginning of the curve and a heavy-tailed distribution for SI. However, those who maintain a large SI are more likely to further increase the SI of their activity space (closer to a circle). For example, city *I* also has the lowest shape parameter for SI (k_2) among all 10 cities. This is potentially due to the special life style in a forest city that people have to visit multiple destinations for basic living needs (instead of following a bimodal movement pattern between home and work).
$k_{3X} > k_{3Y} > 1$	The failure rate of *X* increases faster than *Y* when the entropy increases.	Compared to city *Y*, users with smaller entropy values in city *X* are more likely to explore and increase the spatial coverage of their activities; however, as entropy further increases, their potential of further expanding the movement randomness decreases. For example, city *B* has the highest k_3 value among all 10 cities, indicating that the residents in this city prefer to explore new places when their entropy value is low (close to 0); however, there is a smaller portion of tail users with very large entropy values (i.e., people who travel more randomly, such as a salesman.)

(*Continued*)

Table 5. (Continued).

Model parameter	Theoretical indication	Empirical indication
λ values	A larger scale factor indicates a more dispersed distribution if the shape factor remains the same.	This can be best demonstrated by city *A* and city *B* in Figure 5(a), where city *A* shows a lower maximum point and a larger scale parameter compared to city *B*. Note that all distributions in Figure 5(a) have a maximum point, which occurs between 0.5 and 1.5 km for different cities, indicating that there are only a few people who maintain a very small radius (<0.5 km). The occurrence of maximum points can be influenced by various factors such as the size, shape, and economical status of the urban system. This is consistent with the findings in Kang *et al.* (2012), where the authors demonstrated that the PDF of movement radii first reaches a maximum point, and then the decay follows an exponential distribution.

4.2.1. *Fitting distributions by age and gender*

Figure 6 shows an example plot between age and average radius for each age group in city *A*. As can be seen, the heteroscedasticity appears to be much larger at both ends of the curve (where age <16 and age >70). One reason for this is that these two groups have lower sample sizes compared to other age groups; therefore, the data appear to be less stable. Another potential reason is that individuals below age 16 are considered as 'no capacity' or 'with limited capacity' for civil conduct in the Chinese legal system (http://www.shenzhenlawfirm.com/fg-e/general-01.htm); hence, they are not fully responsible for themselves financially and their activity patterns may partially reflect the wills of their legal guardians (i.e., parents) instead of their own (e.g., attend a book club after school instead of go partying). Although there is no upper age limit for civil conduct in China, the average life expectancy at Chinese retirement age (60) is 79 (http://apps.who.int/gho/data/?theme=main&vid=60340). To

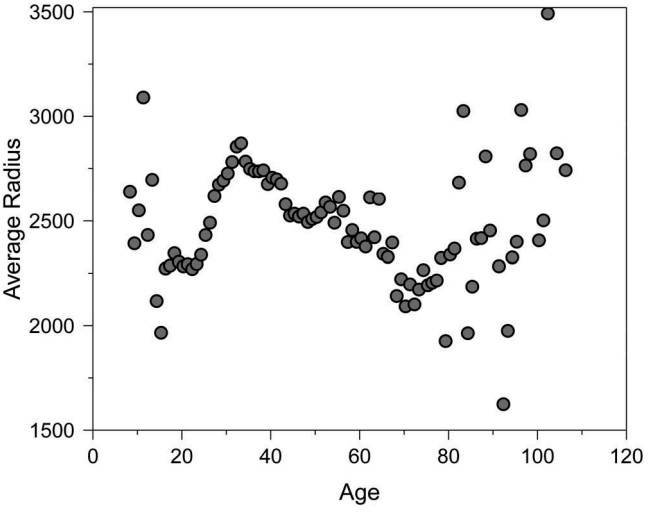

Figure 6. Correlation between age and radius in city *A*.

eliminate the heteroscedasticity and noise effect in the data, here we only consider individuals between ages 16 and 70 when constructing the regression models regarding age.

Figure 7 plots the correlation between age (16 ≤ age ≤70) and the average explanatory variables (radius/shape index/entropy) for both male and female users. The data are aggregated for 10 cities. Because the curves appear to be nonlinear, we also fit in a quadratic polynomial regression for each curve (solid line). Table 6 summarizes the descriptive statistics for Figure 7.

As can be seen, all three measurements indicate a similar pattern where the dependent variable first increases then decreases when the explanatory variable – – age – – increases. Note that the fitted lines only indicate general trends of the three variables, but the extreme values of the original lines appear to be in different age groups for three variables: mid-30 years for radius, mid-40 years for SI, and oscillate between mid-20 years to mid-40 years for entropy, indicating that middle-aged people show a more active mobility pattern (larger activity space, larger deviation from a straight line, and most random visiting patterns). However, the randomness of activities appears to be high for a wider age group (mid-20 years to mid-40 years). In addition, there appear to be noticeable differences between male and female users. As shown in Figure 7, male phone users have a larger activity space and a higher shape index and randomness than female

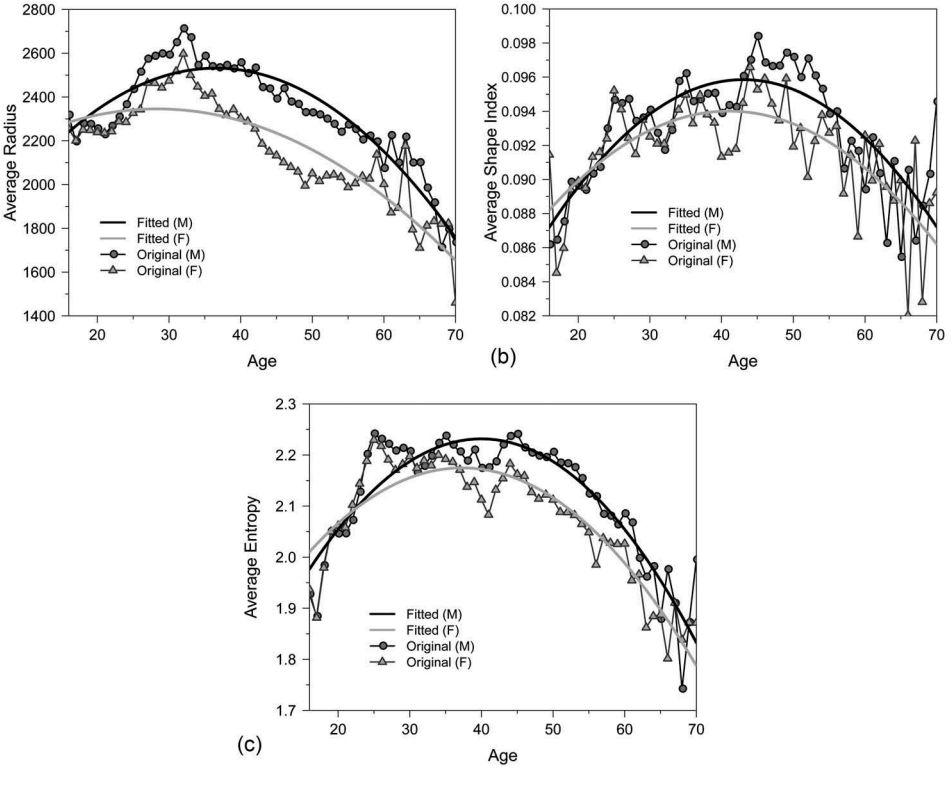

Figure 7. Correlation between age and the three indicators: (a) radius; (b) SI; (c) entropy.

Table 6. The descriptive statistics of three measurements for males and females.

	Male	Female
Quadratic parameter (radius)	−1.38	−0.798
Extreme point (radius)	36	28
Quadratic parameter (SI)	−2.37 e-05	−1.86 e-05
Extreme point (SI)	43	41
Quadratic parameter (entropy)	−0.000886	−0.000724
Extreme point (entropy)	39	37

Note: In a quadratic polynomial regression $y = ax^2+bx+c$, the quadratic parameter a controls the shape of the curve, the smaller the absolute value of a is, the closer the curve is to a straight line.

users, indicating that male users are more active regarding their physical movement. These results demonstrate a more comprehensive picture compared to previous studies for only one city (Yuan *et al*. 2012).

Similar to Section 4.1, in order to better interpret the probability distribution for males and females, as well as differences between age groups, we also fit the Weibull distribution to the following demographic groups:

- Male/female, 16 ≤ age ≤ 17: this group is considered as teenagers.
- Male/female, 18 ≤ age ≤22: college students.
- Male/female, 23 ≤ age ≤ 40: young professionals.
- Male/female, 41 ≤ age ≤ 59: middle-aged professionals.
- Male/female, 60 ≤ age ≤ 70: This group is considered as retired, due to the fact that the official retirement age in China is 60 for most professions.

Table 7 shows that both the shape and scale parameters regarding the radii of females are slightly smaller than those of males, indicating that male users in general have a larger activity space than females, which is consistent with the results shown in Figure 7. This is potentially due to the fact that the employment rate for males is substantially higher than that for females in China (79% versus 65% in 2007), which leads to the result that males serve as a more active labor force and maintain a higher mobility level (http://data.worldbank.org/indicator/SL.TLF.CACT.MA.ZS?page=1, http://data.worldbank.org/indicator/SL.TLF.CACT.FE.ZS?page=1). Note that the correlation between employment and activity patterns is only a hypothesis and aims to provide a preliminary indicator for social scientists. An extended analysis between occupation type and activity level is

Table 7. Comparison of parameters between age and gender.

	k_1 (Radius)	λ_1 (Radius)	k_2 (SI)	λ_2 (SI)	k_3 (entropy)	λ_3 (entropy)
Male (16–17)	0.94	2.15	0.43	0.05	2.22	2.15
Female (16–17)	0.95	2.17	0.42	0.05	2.23	2.14
Male (18–22)	1.01	2.26	0.48	0.06	2.38	2.31
Female (18–22)	1.00	2.25	0.49	0.06	2.41	2.32
Male (23–40)	1.06	2.61	0.48	0.06	2.28	2.48
Female (23–40)	1.05	2.45	0.49	0.06	2.34	2.45
Male (41–59)	1.06	2.45	0.49	0.06	2.33	2.47
Female (41–59)	1.04	2.16	0.49	0.06	2.37	2.39
Male (60–70)	1.02	2.02	0.41	0.05	2.23	2.19
Female (60–70)	0.98	1.85	0.40	0.05	2.16	2.14

needed to prove this assumption, which is beyond the scope of this study. Although a Kolmogorov–Smirnov test between gender groups shows statistical significance at the $p = 0.01$ level, there is no substantial difference regarding the model parameters of SI. The statistical significance of such small differences is majorly due to the large sample size in the case study, which is an inevitable issue in big data analytics.

Regarding the age groups, in general, teenagers and seniors have the lowest shape parameters for all three measures, indicating that these two groups have the heaviest tails (i.e., users clustered at a small value; however, for users with a larger mobility indicator, they are more persistent than other groups and have more outliers). Young and middle-aged professionals have the largest shape parameters, indicating that this group is the least heavy-tailed and has the lowest percentage of outliers. This can be due to multiple reasons: on the one hand, the differences in sample sizes are an inevitable factor; on the other hand, it is possible that the majority of young and middle-aged professionals are employed so the user profiles are less diverse (employment rate >90%, https://stats.oecd.org/Index.aspx?DataSetCode=LFS_SEXAGE_I_R). For scale parameters, it can be seen that middle-aged males have the highest randomness and largest activity space. This is another demonstration of the hypothesis regarding employment and activity patterns. Note that the level of randomness is calculated individually but the assumption of outliers is argued based on the distribution of a group, so they are not necessarily related to each other.

4.2.2. *Age, gender and the built environment*

The development of urban systems partitions the earth into different administrative districts with boundaries that restrict individual behaviors, such as configurations and social conditions (Ahas *et al*. 2015). As discussed in Section 2, due to privacy issues, the differences of urban space usage among demographic groups are rarely addressed in CDR data analysis. There have been several studies on modeling urban dynamic patterns from mobile connection datasets (e.g., the real time Rome project at the MIT SENSEable Lab[1]), but our research focuses on extracting the implications of various clustering patterns for demographic groups. This can provide fundamental support for urban planners to acquire first-hand movement information of population groups, as well as relating these patterns to the distribution of urban infrastructures and the Weibull fittings in Section 4.2.1.

First the division between urban and suburban areas in 10 cities is defined based on the LandScan 2008 population data (http://web.ornl.gov/sci/landscan/), which specifies worldwide population density at 1'' resolution. Here the largest region with top 3% population density is classified as urban area in each city. We also applied a low-pass filter to eliminate noise (Figure 8).

Table 8 demonstrates the ratio between urban and suburban visiting frequencies for different age and gender groups ('urban/suburban ratio'), which is utilized as an indicator to show users' visiting preferences in different urban areas, i.e., a higher urban/suburban ratio indicates that a certain group's activities are more 'urban-oriented' and less 'suburban-oriented'. Note that no time filtering is applied in this analysis (i.e., all data are included), as we are looking at general patterns across age/gender groups, not for patterns during a specific time span. As indicated in Table 8, there is no consistent pattern of urban/suburban ratio across age groups, i.e., in certain cities teenagers and

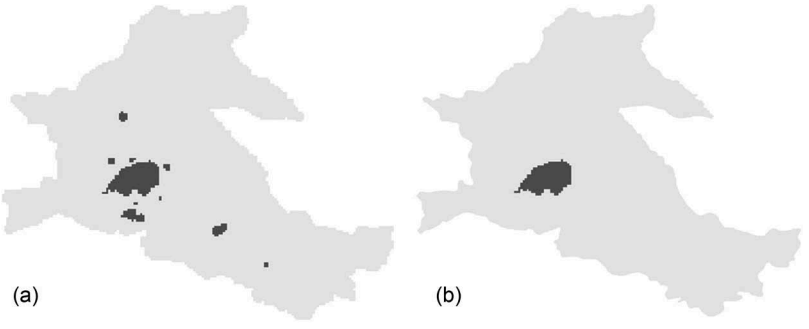

Figure 8. City *A* urban-suburban division (a) before filtering; (b) after filtering.

Table 8. Urban–suburban visiting frequency ratio by age and gender.

	16–17	18–22	23–40	41–59	60–70	Male	Female
A	2.26	3.25	3.02	3.16	2.50	3.18	2.99
B	2.11	2.66	2.91	2.99	3.71	3.16	2.73
C	1.67	1.55	1.55	1.46	1.68	1.57	1.49
D	7.26	6.29	5.88	6.10	6.01	6.40	5.71
E	2.58	2.63	2.07	2.03	1.81	2.21	2.00
F	1.91	2.39	2.36	2.49	2.51	2.47	2.36
G	6.79	6.89	6.49	7.00	5.94	7.12	6.34
H	2.47	2.25	2.31	2.44	2.05	2.51	2.22
I	1.87	2.08	1.94	2.10	2.62	2.17	1.92
J	1.23	1.51	1.97	2.44	1.78	2.19	1.91

seniors visit the urban area the most; however, in other cities, young and middle-aged professionals visit the urban area the most. The 10 cities in this study have very distinct urban functionality divisions, which potentially leads to the distinct visiting patterns of urban/suburban areas in age groups. For instance, city *A* is the home of more than three national/regional universities which are all located in the urban area; hence, the college students group appears to be the most 'urban-oriented' group. Another example is city *C*, where the local economy highly depends on petroleum and related industries. A large number of young and middle-aged professionals work in the oil fields in the suburban area, which explains their lower urban/suburban ratio.

However, the urban/suburban ratios of males and females are tested to be significantly different in all 10 cities (based on Wilcoxon signed rank test, significance level $p = 0.01$). Male users appear to visit urban areas more frequently than females. In other words, males are more active in general (possibly due to the social obligations of being employed); however, females have more access to the suburban areas possibly due to social duties such as taking care of the family. This case study aims to provide an initial insight on how the investigation of movement patterns of demographic groups can provide informative data for urban planners. Future research should look into more detailed patterns in each city and correlate these with different urban infrastructures.

On the other hand, the magnitude of urban/suburban ratios also varies substantially for different cities. A higher urban/suburban ratio indicates a more concentrated planning pattern (such as in cities *D* and *G*), and the residents have less access to the

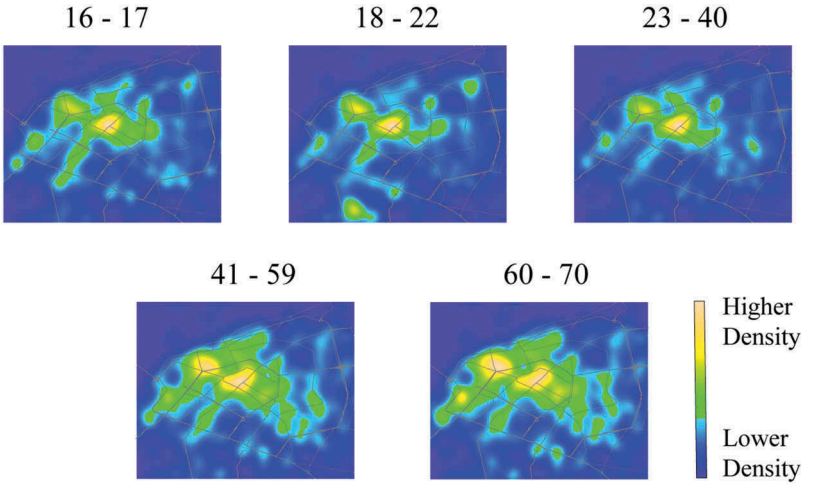

Figure 9. Kernel density distribution of age groups in city *A*.

suburban area. This may be the result of various factors, such as the spatial distribution of work opportunities, residential areas, education institutes, transportation infrastructure, recreational facilities etc., which is beyond the scope of this research.

To cross-validate these findings from a few case studies, we also plotted the kernel density estimation in city *A* (Figure 9) for five age groups (16–17, 18–22, 23–40, 41–59, and 60–70) during 2–5 pm (To be consistent with the CDR spatial accuracy in Section 3.1, here we utilize a search radius = 500 m). Although all age groups indicate clustering in the city center and the northwestern region, the oldest age groups (41–59 and 60–70) show more spread patterns, whereas the young professionals (23–40) show a focused cluster in the city center. This is consistent with the finding in Section 4.2.1 that the activity space measurements of seniors are more heavy-tailed distributed with more variations in daily activities (e.g., possibly due to less scheduling constraints after retirement). Another interesting pattern is the cluster of college students on the south side of the city, where several national and regional universities are located.[3] This further confirms the indication of Table 8 that the spatial distribution of different age groups tightly connects with the functionality of urban divisions and the availability of facilities (e.g., education facilities for students, employment opportunities for professionals, and leisure facilities for the retired).

To cross-validate the findings in Figure 9 and demonstrate that city *A* is not a special case in the sample set, we also plot the clustering patterns of two age groups (18–22 and 23–40) in a smaller city *E* with urban population less than 1 million (due to page limitation we are not able to plot the densities for all 10 cities). As shown in Figure 10, the college students group in city *E* also demonstrates a cluster which deviates from the city center and is consistent with the locations of higher education institutes in this city.

Similarly, Figure 11 plots the distribution for males and females in city *A*. As can be seen, during regular work hours (2–5 pm), females show a higher spread pattern than males. This provides an informative addition to the findings in Section 4.2.1 and Table 8, where the distribution of females appears to be more widespread. Although from

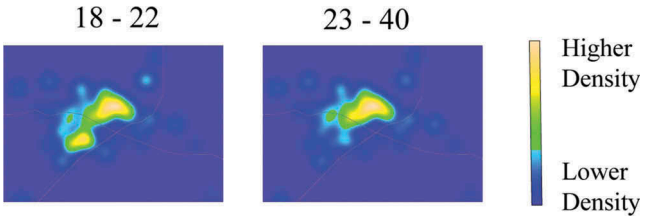

Figure 10. Kernel density distribution of age groups in city *E*.

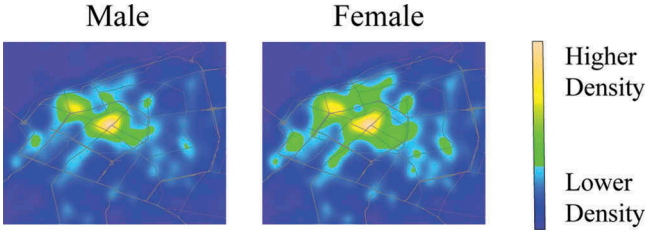

Figure 11. Kernel density distribution of gender groups in city *A*.

individual perspective, male phone users appear to have a larger activity space, and a higher SI and randomness than female users, from an urban perspective female users are more spread out and tend to make use of a wider range of urban areas. A potential explanation may involve a deeper investigation of the societal roles of gender groups in Chinese society. For instance, the higher occurrence of females in a suburban area (with smaller movement radii) may indicate that they are more family-oriented.

4.3. *Discussion of uncertainty*

Finally, it is important to highlight that there are different aspects of uncertainty related to human activity studies. These issues arise in our data mining process in different ways (Xia 2005, Yuan *et al*. 2012), including but not limited to:

- Natural variability of human activities: Although human mobility seems to be highly predictable (González *et al*. 2008, Song *et al*. 2010), randomness is an inevitable part of human motion.
- Inaccuracy/imprecision due to the limitation of available data. As discussed in Section 3.1, positional inaccuracy, sampling resolution and imprecision all contribute to the uncertainty of the data source.
- Imperfection of models and algorithms: As Box and Draper (1987, pp. 424) stated: 'Essentially, all models are wrong, but some are useful'. In this research the Weibull distribution is adopted to interpret the distribution of human activities, due to the fact that it is flexible to fit into curves with distinct shapes and conduct cross-variable comparisons. For a single variable, other models may be applicable such as

the lognormal distribution. The application of different models will inevitably have an impact on the uncertainty of the results. Another distinction worth noting is that between 'statistically significant difference' and 'substantial difference' (i.e., large difference in magnitudes). Due to the large sample size utilized in this research, small differences can be tested as 'statistically significant'. In future studies, it will be helpful to cross-validate the results with other data sources to test their robustness.

5. Conclusions and future work

ICTs have become increasingly influential in our society. The broader impacts of this research will yield an enhanced understanding of human activities for different urban systems and demographic groups, as well as provide novel methods for expanding the important and widely applicable area of geographic knowledge discovery. Mobile phone data used as an input in the analysis of human mobility has the potential to transform research in diverse fields, such as geography, transportation, planning, and economics.

The first contribution of this research comes from the methodology perspective when analyzing the mathematical distribution of mobility indicators. We explored the feasibility of applying existing methods (the Weibull distribution) to novel topics that have not been applied in the area of human mobility modeling. The indication of the Weibull distribution was extended from the temporal to the spatial dimension. We discovered the differences among age groups regarding the performance of tail (outlier) users from shape indices, which was further demonstrated by the density plots in Figures 9–11. The effectiveness of this model fitting was tested thoroughly based on a sample data set, which provides a valuable reference for future studies based on other georeferenced mobile phone datasets. The results demonstrate that this method is particularly suitable for comparing distributions with distinct shapes. The methodologies addressed in this research can also be extended to analyze similar datasets acquired from various social media, e.g., volunteered geographic information on Twitter or Facebook.

The second contribution consists of the various types of empirical results presented in Section 4 to analyze the spatial distribution of demographic groups. Three measurements were first defined to investigate the characteristics of user activity space. We also explored the correlation between activity spaces and various demographic factors, such as age, gender, and the built environment. This model fitting is informative for interpreting the distribution of activity space for different population groups. The case studies in Section 4.2.2 explored the usage of urban space in demographic groups and discovered that it is highly likely that the spatial distribution of different age groups tightly connects with the functionality of urban divisions and the availability of facilities. We also discovered significant differences between males and females regarding the visiting frequency of urban and suburban areas.

The results of this research will provide references to help policy makers understand the characteristics of individual mobility, as well as update environmental and transportation policies. For instance, we discovered the differences in urban and suburban visiting patterns between male and female users, and demonstrated the connections between specific urban facilities and the mobility patterns of various age groups. Due to page limitation, Figures 9–11 utilized cities *A* and *E* as an exemplary analysis instead of plotting all 10 cities in the

whole dataset. Future research directions include the validation of models for other cities and countries, and the comparison between different models. For follow-up studies it will be necessary to explore how explanatory variables, such as the cultural background or the shape of an urban system, impact the model fitting process. Another direction for future research is re-classifying urban divisions based on the usage of urban space by demographic groups. In addition, several potential methods can be adopted to quantitatively measure the described uncertainty issues, such as probability theories, Bayesian networks, and fuzzy sets. The designed methodologies and the extension of the Weibull models can also be applied to other types of big (geo)data, such as crowed-sourced LBSM, volunteered geographic information, and taxi trajectories from GPS devices.

Notes

1. When opening a new phone line, users can choose to provide or not provide their citizen ID which includes both the gender information and the date of birth. Detailed data are listed in Table 1.
2. China Statistical Yearbook, National Bureau of Statistics of China, 2008.
3. This conclusion was derived based on the landmarks on Google™ Map. Based on the request of data provider we are not able to provide the identifiable base maps from Google™ Map.

Acknowledgements

We thank Dr Yu Liu and Geosoft Lab at Peking University for providing the data. Gwen Raubal helped improving the grammar and style of this work. The reviewers provided excellent feedback, which helped us improving the content and clarity of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the Swiss National Science Foundation [project # 205121_141284].

References

Ahas, R., *et al*., 2010. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. *Transportation Research Part C-Emerging Technologies*, 18 (1), 45–54. doi:10.1016/j.trc.2009.04.011

Ahas, R., *et al*., 2015. Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *International Journal of Geographical Information Science*, 29 (11), 2017–2039.

Azevedo, T.S., *et al*., 2009. An analysis of human mobility using real traces. *Proceedings of the 2009 IEEE conference on Wireless Communications & Networking Conference*. Budapest, Hungary: IEEE Press, 2390–2395.

Bagrow, J.P. and Koren, T., 2009. Investigating bimodal clustering in human mobility. *International Conference on Computational Science and Engineering*. Vancouver, Canada, 944–947.

Beckmann, K. 2000. Umweltgerechtes Verkehrsverhalten beginnt in den Köpfen. ed. *Mobilitätsforschung für das 21*, Köln, Jahrhundert, 213–238.
Borrel, V., De Amorim, M.D., and Fdida, S., 2006, On natural mobility models. *Autonomic Communication*, 3854, 243–253.
Box, G.E.P. and Draper, N.R., 1987. *Empirical model-building and response surfaces*. New York: Wiley.
Brown, L.A. and Moore, E.G., 1970, The intra-urban migration process: a perspective. *Geografiska Annaler. Series B, Human Geography*, 52 (1), 1–13. doi:10.2307/490436
Candia, J., *et al*., 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41 (22), 1–11. doi:10.1088/1751-8113/41/22/224015
Caragliu, A., Bo, C.D., and Nijkamp, P., 2009. *Smart cities in Europe, Serie Research Memoranda 0048*.
Chainey, S., Tompson, L., and Uhlig, S., 2008, The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21 (1–2), 4–28. doi:10.1057/palgrave.sj.8350066
Cheyne, J.A. and Efran, M.G., 1972, The effect of spatial and interpersonal variables on the invasion of group controlled territories. *Sociometry*, 35 (3), 477–489. doi:10.2307/2786507
Fiore, F.D., *et al*. 2014. A set of perspectives on how mobile technology may affect travel. *Journal of Transport Geography*, 41, 97–106. doi:10.1016/j.jtrangeo.2014.08.014
Forbes, 2014. Scientists Warn About Bias In The Facebook And Twitter Data Used In Millions Of Studies.
Fuchs, M. and Busse, B., 2009, The coverage bias of mobile web surveys across European countries. *International Journal of Internet Science*, 4, 21–33.
Gao, S., *et al*., 2013. Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, 17 (3), 463–481. doi:10.1111/tgis.12042
Golledge, R.G. and Stimson, R.J., 1997. *Spatial behavior: a geographic perspective*. New York: Guilford Press.
González, M.C., Hidalgo, C.A., and Barabási, A.-L., 2008, Understanding individual human mobility patterns. *Nature*, 453 (7196), 779–782. doi:10.1038/nature06958
Gordon, P., Kumar, A., and Richardson, H.W., 1989, The influence of metropolitan spatial structure on commuting time. *Journal of Urban Economics*, 26 (2), 138–151. doi:10.1016/0094-1190(89)90013-2
Groves, P.D., 2013. *Principles of GNSS, inertial, and multisensor integrated navigation systems*. 2nd ed. Boston: Artech House.
Hägerstrand, T., 1970. What about people in regional science? *Papers of the Regional Science Association*, 24, 7–21. doi:10.1111/j.1435-5597.1970.tb01464.x
Harding, C., *et al*., 2012. Modeling the effect of land use on activity spaces. *Transportation Research Record*, 2323, 67–74. doi:10.3141/2323-08
Harvey, A.S. and Taylor, M.E., 2000, Activity settings and travel behaviour: A social contact perspective. *Transportation*, 27 (1), 53–73. doi:10.1023/A:1005207320044
Hasan, S., *et al*., 2013. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151 (1–2), 304–318. doi:10.1007/s10955-012-0645-0
Herrera, J.C., *et al*., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment. *Transportation Research Part C-Emerging Technologies*, 18 (4), 568–583. doi:10.1016/j.trc.2009.10.006
Horton, F. and Reynolds, D.R., 1971. Effects of urban spatial structure on individual behavior. *Economic Geography*, 47, 36–48. doi:10.2307/143224
Jiang, B. and Jia, T. 2011. Exploring human mobility patterns based on location information of US flights. arXiv:1104.4578.
Jiang, B. and Yin, J., 2013. Ht-Index for Quantifying the Fractal or Scaling Structure of Geographic Features, arXiv:1305.0883.
Kang, C., *et al*., 2012. Intra-urban human mobility patterns: an urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, 391 (4), 1702–1717. doi:10.1016/j.physa.2011.11.005
Knox, P.L. and McCarthy, L., 2012. *Urbanization: an introduction to urban geography*. 3rd ed. Pearson: Boston.

Leitch, R.D., 1995. *Reliability analysis for engineers: an introduction*. Oxford; New York: Oxford University Press.

Lewis, G.K., 1959, Changes in suburban land-use patterns. *Annals of the Association of American Geographers*, 49 (2), 194–195.

Liu, Y., *et al*., 2012. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, 14 (4), 463–483. doi:10.1007/s10109-012-0166-z

Lu, Y., 2000. *Spatial cluster analysis of point data: location quotients versus kernel density*. Portland, OR: *University Consortium of Geographic Information Science* (UCGIS) *Summer Assembly Graduate Papers*.

Mason, M.J. and Korpela, K., 2009, Activity spaces and urban adolescent substance use and emotional health. *Journal of Adolescence*, 32 (4), 925–939. doi:10.1016/j.adolescence.2008.08.004

Mazey, M.E., 1981, The effect of a physio-political barrier upon urban activity space. *Ohio Journal of Science*, 81 (5–6), 212–217.

Mennis, J. and Mason, M.J., 2011, People, places, and adolescent substance use: integrating activity space and social network data for analyzing health behavior. *Annals of the Association of American Geographers*, 101 (2), 272–291. doi:10.1080/00045608.2010.534712

Miller, H.J., 2009. Geographic data mining and knowledge discovery: an overview. *In*: H.J. Miller and J. Han, eds. *Geographic data mining and knowledge discovery*. 2nd ed. London: CRC Press, 3–32.

Morgan, E.C., *et al*., 2011. Probability distributions for offshore wind speeds. *Energy Conversion and Management*, 52 (1), 15–26. doi:10.1016/j.enconman.2010.06.015

Neubeck, K., 2004. *Practical reliability analysis*. Upper Saddle River, NJ: Prentice Hall.

Nobis, C., Lenz, B., and Vance, C., 2005. Communication and travel behaviour: two facets of human activity patterns. In: H. Timmermans, ed. *Progress in activity-based analysis*. Oxford, UK: Elsevier, 471–488.

Ott, T. and Swiaczny, F., 2001. *Time-integrative geographic information systems: management and analysis of spatio-temporal data* [online]. Berlin; New York: Springer. Available from: http://www.loc.gov/catdir/toc/fy033/2001266545.html; http://www.loc.gov/catdir/enhancements/fy0812/2001266545-d.html

Pendyala, R.M., Goulias, K.G., and Kitamura, R., 1991, Impact of telecommuting on spatial and temporal patterns of household travel. *Transportation*, 18 (4), 383–409. doi:10.1007/BF00186566

Phithakkitnukoon, S., *et al*. 2010. Activity-aware map: Identifying human daily activity pattern using mobile phone data. *In*: A.A. Salah, *et al*. eds. *HBU 2010*. Heidelberg: LNCS, Springer, 14–25.

Ratti, C., *et al*. 2007. Mobile landscapes: graz in real time. *In*: G. Gartner, W. Cartwright and M.P. Peterson, eds. *Location based services and teleCartography*. Berlin: Springer, 433–444.

Rhee, I., *et al*., 2011. On the levy-walk nature of human mobility. *Ieee-Acm Transactions on Networking*, 19 (3), 630–643. doi:10.1109/TNET.2011.2120618

Rinne, H., 2008. *The Weibull distribution: a handbook*. Boca Raton, FL: Chapman and Hall/CRC.

Salingaros, N.A., 1998, Theory of the urban web. *Journal of Urban Design*, 3 (1), 53–71. doi:10.1080/13574809808724416

Schönfelder, S., 2006. *Urban rhythms: Modelling the rhythms of individual travel behaviour*. (Doctoral). Swiss Federal Institute of Technology.

Schönfelder, S. and Axhausen, K.W., 2002. Measuring the size and structure of human activity spaces: the longitudinal perspective. *In: Arbeitsberichte Verkehrs- und Raumplanung*. Zürich: ETH Zürich, 135, IVT. DOI:10.3929/ethz-a-004444846.

Sherman, J.E., *et al*., 2005. A suite of methods for representing activity space in a healthcare accessibility study. *International Journal of Health Geographics*, 4 (1), 1–21. doi:10.1186/1476-072X-4-24

Silm, S. and Ahas, R., 2014, Ethnic differences in activity spaces: a study of out-of-home nonemployment activities with mobile phone data. *Annals of the Association of American Geographers*, 104 (3), 542–559. doi:10.1080/00045608.2014.892362

Singh, V., 1987, On application of the Weibull distribution in hydrology. *Water Resources Management*, 1 (1), 33–43. doi:10.1007/BF00421796

Song, C.M., *et al*., 2010. Limits of predictability in human mobility. *Science*, 327 (5968), 1018–1021. doi:10.1126/science.1177170

Talen, E., 1999, Sense of community and neighbourhood form: an assessment of the social doctrine of new urbanism. *Urban Studies*, 36 (8), 1361–1379. doi:10.1080/0042098993033

Traunmueller, M., Quattrone, G., and Capra, L., Mining mobile phone data to investigate urban crime theories at scale. ed. *International Conference on Social Informatics*, 2014, 396–411.

Warf, B., 2010. *Encyclopedia of geography*. Thousand Oaks, CA: Sage Publications.

Weibull, W., 1951, A statistical distribution function of wide applicability. *Journal of Applied Mechanics—Transactions of the ASME*, 18 (3), 293–297.

Xia, Y., 2005. *Integrating uncertainty in data mining*. Doctoral Dissertation. University of California.

Yuan, Y. and Raubal, M., 2012. Extracting dynamic urban mobility patterns from mobile phone data. *In*: N. Xiao, *et al*., ed. *Geographic information science – 7th international conference*. Columbus, OH: Springer, 354–367.

Yuan, Y. and Raubal, M., 2014, Measuring similarity of mobile phone user trajectories – a Spatio-temporal Edit Distance method. *International Journal of Geographical Information Science*, 28 (3), 496–520. doi:10.1080/13658816.2013.854369

Yuan, Y., Raubal, M., and Liu, Y., 2012, Correlating mobile phone usage and travel behavior - a case study of Harbin, China. *Computers, Environment and Urban Systems*, 36 (2), 118–130. doi:10.1016/j.compenvurbsys.2011.07.003

Yue, Y., *et al*., 2014. Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies. *Travel Behaviour and Society*, 1 (2), 69–78. doi:10.1016/j.tbs.2013.12.002

Appendix. Estimating parameters for Weibull distribution

Normally, the parameter estimation of a two-parameter Weibull distribution can be achieved by maximum likelihood estimation (MLE). The MLE functions for shape parameter k and scale parameter λ are defined as

$$\hat{\lambda}^k = \frac{1}{N}\sum_{i=1}^{N}(x_i{}^k - x_N^k) \tag{1}$$

$$\hat{k}^{-1} = \frac{\sum_{i=1}^{N}(x_i^k \ln x_i - x_N^k \ln x_N)}{\sum_{i=1}^{N}(x_i^k - x_N^k)} - \frac{1}{N}\sum_{i=1}^{N}\ln x_i \tag{2}$$

where $x_1, \ldots, x_k$ are sample values and N is the number of sample points.