DOI: 10.1111/tgis.12450

WILEY Transactions in GIS

RESEARCH ARTICLE

Exploring the effectiveness of location-based social media in modeling user activity space: A case study of Weibo

Yihong Yuan | **Xujiao Wang**

Department of Geography, Texas State University, San Marcos, Texas

Correspondence
Yihong Yuan, Department of Geography, Texas State University, San Marcos, TX 78666.
Email: yuan@txstate.edu

Abstract

Location-based social media (LBSM) has been widely utilized to supplement traditional survey methods in modeling human activity patterns. However, there has not been sufficient study to assess the reliability of these data in deriving human movement. This research aims to evaluate how data collection duration and sample sizes affect the reliability of LBSM data in activity modeling based on two indicators: radius of gyration (ROG) and entropy. We use a linear regression model with logarithmic transformation to approximate how the magnitude of each indicator changes with different data collection durations—from 1 to 12 months. The results indicate that both ROG and entropy increase when the amount of data increases. However, the rate of increase slows down and approaches zero eventually. We also approximated the limit values and verified that with 12-month data, we are at approximately >95% magnitude of the limit values for both indicators in all three cities. The clustering analysis also demonstrated that there are outlier users who exhibit distinct patterns. This case study focuses on three Chinese cities (Beijing, Shanghai, and Guangzhou) and provides a useful reference to explore the balance point between data effectiveness and an appropriate sample size from LBSM data.

1 | INTRODUCTION

In the past few decades, various research fields, including geography, transportation, computational physics, and computer science, have made much progress in the theories, methods, and applications of human activity analysis (Aggarwal & Ryoo, 2011; Gonzalez, Hidalgo, & Barabasi, 2008; Song, Qu, Blumm, & Barabasi, 2010). Originally, human activity studies were rooted in the study of modeling the general patterns of human mobility in space and time, such as *Random Walk* and its many derivatives (Borrel, De Amorim, & Fdida, 2006). These models have an impact on all phenomena and activities driven by human mobility, including urban planning and agent-based modeling approaches in transportation (Gonzalez et al., 2008). However, many mobility models are constructed at an abstract and general level, and therefore do not focus on representing concrete human activities (Yuan, Raubal, & Liu, 2012). In addition, previous studies also focus on the theories and methods in mining individual trajectories (Zheng & Zhou, 2011), including Intra-trajectoryl studies (i.e., defining measurements and indicators to model the inherent characteristics of human trajectories) (Ahmed, Karagiorgou, Pfoser, & Wenk, 2015; Zhao & Xu, 2009) and *inter-trajectory* studies (i.e., measuring trajectory similarity among individuals) (Abraham & Lal, 2010; Xia, Wang, Zhang, Kim, & Bae, 2011; Zhang, Huang, & Tan, 2006).

Unlike general mobility models in computational physics or trajectory analysis in computer science, activity studies in geography are usually georeferenced to a specific geographic location within geographic/projected coordinate systems (e.g., a specific city), and they focus on analyzing the activity participation of populations under different geographic and temporal contexts. Jones, Koppelman, and Orfueil (1990) defined activity analysis as a framework of modeling the similarities and distinctions of travel as daily or multi-day patterns for population groups with different lifestyles. Among all human activity studies, modeling activity space is an important topic to determine the spatial distribution of human behavior (Golledge & Stimson, 1997), which is defined as the local areas within which people travel during their daily activities (Mazey, 1981). Researchers have focused on both the morphology and the internal structure of human activity space—the former measures its basic characteristics (e.g., size, shape, etc.) and the latter emphasizes the reasons for which an activity space forms (e.g., regularly visited locations) (Golledge & Stimson, 1997; Schönfelder & Axhausen, 2002).

Traditional human activity analysis often utilizes travel surveys and questionnaires for data collection (Levinson & Kumar, 1995; Schönfelder & Axhausen, 2002). However, collecting such data can be costly, time consuming, and may only cover a relatively small sample set in a limited spatial environment. Meanwhile, the development of location-based social media (LBSM), defined as "Social Network Sites that include location information" (Roick & Heuser, 2013), has provided more flexibility for researchers regarding where, when, and how to collect human activity behavior. Studies utilizing LBSM for analyzing both individual and aggregated patterns have grown rapidly (Cho, Myers, & Leskovec, 2011; Gao & Liu, 2015; Hasan, Zhan, & Ukkusuri, 2013; Zagheni, Garimella, Weber, & State, 2014). Although many studies have attempted to classify neighborhoods (such as the Livehoods project; Cranshaw, Schwartz, Hong, & Sadeh, 2012) and/or extract activity anchor points (e.g., "home" and "work") (Qu & Zhang, 2013) from LBSM, there is a lack of understanding about the morphology (e.g., shape, size) and the internal structure (e.g., movement randomness) of activity space from such user-contributed datasets (Malleson & Birkin, 2014).

Similar to other types of big (geo) data, LBSM data also have different data quality issues, such as accuracy, precision, and sampling biases across various data sources. Hence, determining the appropriate data size, duration, and sampling resolution is crucial for designing a statistically sound study. However, in practice, these factors are often determined arbitrarily when using LBSM to analyze activity patterns, and there has yet to be a systematic study of how users' activity spaces change with different sample sizes from LBSM. In general, while larger sample sizes can provide more location information for a certain population group, researchers often seek an appropriate data collection duration, which achieves a balance between the details of information and computation efficiency/data collection cost. This article aims to address that issue.

Therefore, the objective of this research is twofold. First, methodologically, we provide an extendable data mining strategy to optimize sample size in future studies. We start by defining two indicators, radius of gyration (ROG) and entropy, to reflect both the morphology (size) and the internal structure (randomness) of activity space. Second, empirically, we quantify the "effectiveness" of LBSM at modeling human activity space. Here, we define "effectiveness" as the stability of activity space indicators with different amounts of data used. We will also investigate how this effectiveness manifests itself in different user categories in terms of the measurement of their activity space. Note that in this study, "data quantity" and "data collection duration" are used interchangeably. We chose data collection duration (e.g., 1 month, 2 months) instead of the exact number of check-ins (e.g., 1,000 records, 2,000 records) for two reasons: (1) to be consistent with other social media analysis, as most data collection campaigns are conducted based on a chronological circle (e.g., weeks or months); and (2) to collect our user data under the same study period so that they are comparable.

The remainder of this article is organized as follows: Section 2 describes related work in the areas of activity space modeling, LBSM data analysis, and its data quality issues; Section 3 introduces the fundamental research design, including the dataset and our 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 | Modeling human activities from LBSM

The development of social networking sites such as Twitter, Flickr, Instagram, and Facebook provides more opportunities to analyze human activity patterns based on crowd-sourced big geodata (Cho et al., 2011; Haffner, Mathews, Fekete, & Finchum, 2018; Hawelka et al., 2014; Malleson & Birkin, 2014; Yuan & Medel, 2016). Unlike traditional survey or individually collected GPS data, LBSM datasets cover a larger sample size and can easily be accessed by application programming interfaces (APIs), therefore they provide a rich resource for researchers to analyze human activity patterns from both aggregated (urban) and individual perspectives (Cao et al., 2015; Liben-Nowell, Novak, Kumar, Raghavan, & Tomkins, 2005; Yuan & Medel, 2016). From the urban perspective, researchers have investigated how user activities in different cities exhibit universal properties, as cities have been shown to be scaled versions of each other, despite their cultural and historical differences (Cho et al., 2011). Many studies attempted to link human mobility to urban structure and activities (Bawa-Cavia, 2011; Cranshaw et al., 2012; Liu et al., 2015; Mohammady & Culotta, 2014). For example, Phithakkitnukoon, Horanont, Di Lorenzo, Shibasaki, and Ratti (2010) developed an activity-aware map to investigate the most probable activity associated with each urban district. The results provide transportation and urban planners more accurate data to plan for a more sustainable city.

In addition to the urban-oriented research, previous studies have demonstrated the power of LBSM data in analyzing individual-oriented activity behavior and constructing mobility models (Cho et al., 2011). As stated in Jones et al. (1990), activity analysis is a framework for analyzing travel behavior as daily or multi-day patterns derived from lifestyles and activity participation among the population. Activity-based studies incorporate more spatial, temporal, and social constraints, in contrast to traditional human mobility studies, which are closely tied to the construction of generalized mobility and aim to describe general patterns and basic laws (Elliott & Urry, 2010). For example, environmental constraints are usually built into microeconomic models as parameters when conducting travel forecasting simulations. LBSM has attracted global users to share their whereabouts on daily, weekly, and long-term temporal scales, making LBSM particularly suitable for modeling individual patterns such as activity scheduling, social network structure, and location prediction. For instance, Hasan et al. (2013) analyzed the temporal distribution of different categories of user activities. Cho et al. (2011) studied how social constraints

like friendship influence individuals' movements. The results indicate that the farther a user travels, the more likely his/her movement is influenced by a friend who lives close to the travel destination.

Furthermore, recent studies also combined individual and urban-oriented research by investigating spatial semantics from user-generated content on LBSM, as well as its impact on modeling social relations or delineating thematically distinct hotspots in urban systems (Bennett & Agarwal, 2007; Goodchild & Li, 2012; Liu et al., 2015). For instance, Jenkins, Croitoru, Crooks, and Stefanidis (2016) investigated the emergence of unique topics at different locations and the identification of urban hotspots based on semantic signatures. Another study by Steiger, Resch, and Zipf (2016) analyzed how relationships among people can be discovered by modeling their activities using a trans-disciplinary approach combing spatial, temporal, and semantic dimensions. These studies went beyond the spatial dimension to incorporate the semantics of LBSM in interpreting user activity behaviors.

Among all activity-based studies, 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 that people travel within while performing their daily activities (Becker et al., 2013; Mazey, 1981; Yuan & Raubal, 2016). Traditionally, researchers have investigated both the morphology and the internal structure of activity space—the former measures its basic characteristics (e.g., size, shape, etc.) and the latter emphasizes the reasons for which an activity space forms (e.g., regularly visited locations) (Golledge & Stimson, 1997; Schönfelder & Axhausen, 2002). Previous studies used various measurements to approximate the external morphology of human activity space, such as ellipse-based measures (e.g., standard deviation ellipse and confidence ellipse) (Schönfelder & Axhausen, 2002), convex hull (Lee, Davis, Yoon, & Goulias, 2016), ROG (Song et al., 2010), and so on. Researchers also developed various methodologies to explore the internal structure of activity space. As Golledge and Stimson (1997) pointed out, there are three determinants of activity space for a given individual: (1) the position of the individual's home location; (2) regularly visited activity locations (i.e., points of interest, POIs) such as work location, grocery stores, park, cinemas, and so on; and (3) travel between and around the pegs, such as the accessibility of public transport to regularly visited locations. Therefore, many previous studies concentrated on extracting activity anchor points and individual differences of visiting these points, as well as understanding the formation of activity spaces (Ahas et al., 2015; Long & Nelson, 2013; Malleson & Birkin, 2014; Phithakkitnukoon et al., 2010; Silm & Ahas, 2014; Xu et al., 2015, 2016, 2015, 2016). In addition to external characteristics, researchers also applied various measures to quantify the internal structure of individual activity space, such as network-based measures (e.g., standard travel time polygon and shortest-path spanning tree) (Schönfelder & Axhausen, 2010; Sherman, Spencer, Preisser, Gesler, & Arcury, 2005) and density-based/probability-based measures [e.g., an activity surface created from kernel density estimation (Kwan, 2000), or an entropy value showing the probabilistic distribution of visiting different POIs (Song et al., 2010; Yuan et al., 2012)].

In the big data era, the widespread use of location-based technologies has provided rich geographic information to measure individual activity space, including but not limited to georeferenced mobile phone data, LBSM check-in data, Bluetooth data, and so on. However, many of the aforementioned studies focused on georeferenced mobile phone data, such as call detailed records (CDRs), due to the wide usage of cell phones. As mentioned in Section 1, although previous studies have utilized LBSM for certain activity analyses, such as neighborhood classification (Cranshaw et al., 2012), there has not been extensive analysis of the effectiveness of such self-reported datasets on deriving user activity space (Malleson & Birkin, 2014). Because there are a large number of activity space indicators, we chose one external indicator (radius of gyration, ROG) and one internal indicator (entropy) as a case study to examine how the magnitude of these measurements changes with the amount of LBSM data used, as both indicators are commonly used in activity space studies and have proven to be more robust to outlier points (Song et al., 2010). For example, ROG has been widely used to represent the spatial dispersion and activity range of individual daily activities (Xu et al., 2015). Entropy is often used to indicate the randomness of activity patterns, which is invaluable in determining the likelihood of users returning to previously visited locations and predicting future trips (Song et al., 2010). The goal of this study is not to review the robustness of every activity

space indicator. Instead, we aim to propose a data processing strategy that can be extended to other datasets and activity space measures in order to help researchers optimize their research design.

2.2 | Data quality issues of LBSM

Compared to georeferenced mobile phone data such as CDRs, LBSM data often have more reliable spatial accuracy (e.g., 5–10 m from a built-in smart phone GPS module versus 300–500 m from a cell tower position) (Calabrese, Ferrari, & Blondel, 2015). However, a series of potential data issues from LBSM data can also affect the reliability of analysis results when deriving human activity patterns from these datasets (Kaisler, Armour, Espinosa, & Money, 2013). Spatial data quality, such as accuracy/precision, resolution, completeness, and consistency (Veregin, 1999), plays a fundamental role in geographic analysis, therefore it is crucial to assess the reliability of LBSM data for human activity studies (Spielman, 2014). Previous research mostly analyzed LBSM quality issues from the following two perspectives.

- *Data accuracy and validity.* Noise and abnormality exist in most LBSM datasets (Kaisler et al., 2013). For example, fake accounts and check-ins can be automatically generated by a computer program instead of an actual user, which inevitably jeopardizes the reliability of such datasets in human activity studies. It is possible for a machine-generated account to post a large number of tweets from the same location. Although researchers have investigated algorithms to detect spams on social network sites (SNSs) (Saini, 2014), identifying fake check-ins still remains one of the biggest challenges in LBSM data quality assessment.
- *Data completeness and availability.* The completeness issue of LBSM data can be addressed from multiple perspectives. First, most studies utilize data collected during a given time span. In reality, data size is often arbitrarily defined in human mobility studies, which inevitably affects the quality of results (Cuzzocrea, Song, & Davis, 2011). Second, LBSM data follows a power law distribution, where the majority of users post sporadically and a very small portion of users check in frequently (Zafarani, Abbasi, & Liu, 2014). This inevitably affects the data completeness for certain users (Wu, Zhi, Sui, & Liu, 2014). The uneven distribution of LBSM data also exists across geographies, where users tend to create a disproportionate number of posts from certain locations, resulting in a sampling bias in space (Preoţiuc-Pietro & Cohn, 2013). For example, an exploratory analysis in Austin, TX found that Twitter users tend to post in recreational areas and the airport (Yuan, Wei, Chow, & Hagelman, 2017). Third, the characteristics of social media sites attract different user groups, which may result in demographic biases in mining, analyzing, and modeling LBSM data (Tufekci, 2014). Twitter, for instance, has proven to be more appealing to young people (Longley, Adnan, & Lansley, 2015). The selection bias from LBSM datasets remains a challenge in acquiring an unbiased and complete sample of the entire population (Hawelka et al., 2014; Longley et al., 2015; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2012; Sloan, Morgan, Burnap, & Williams, 2015).

As discussed in Section 1, this research aims to address LBSM data quality from a data quantity perspective. We aim to analyze how LBSM data collection duration affects the modeling of human activity space, and how this effect can be clustered into different patterns. Section 3 describes the dataset and our methodology.

3 | RESEARCH DESIGN

3.1 | Dataset

To ensure the generalizability of the results, we selected three Chinese cities (Beijing, Shanghai, and Guangzhou) in highly developed metropolitan areas as our study area. Beijing is the capital and the cultural, political, and economic center of China, with a population of 21.7 million based on the 2016 census data. Shanghai is a global

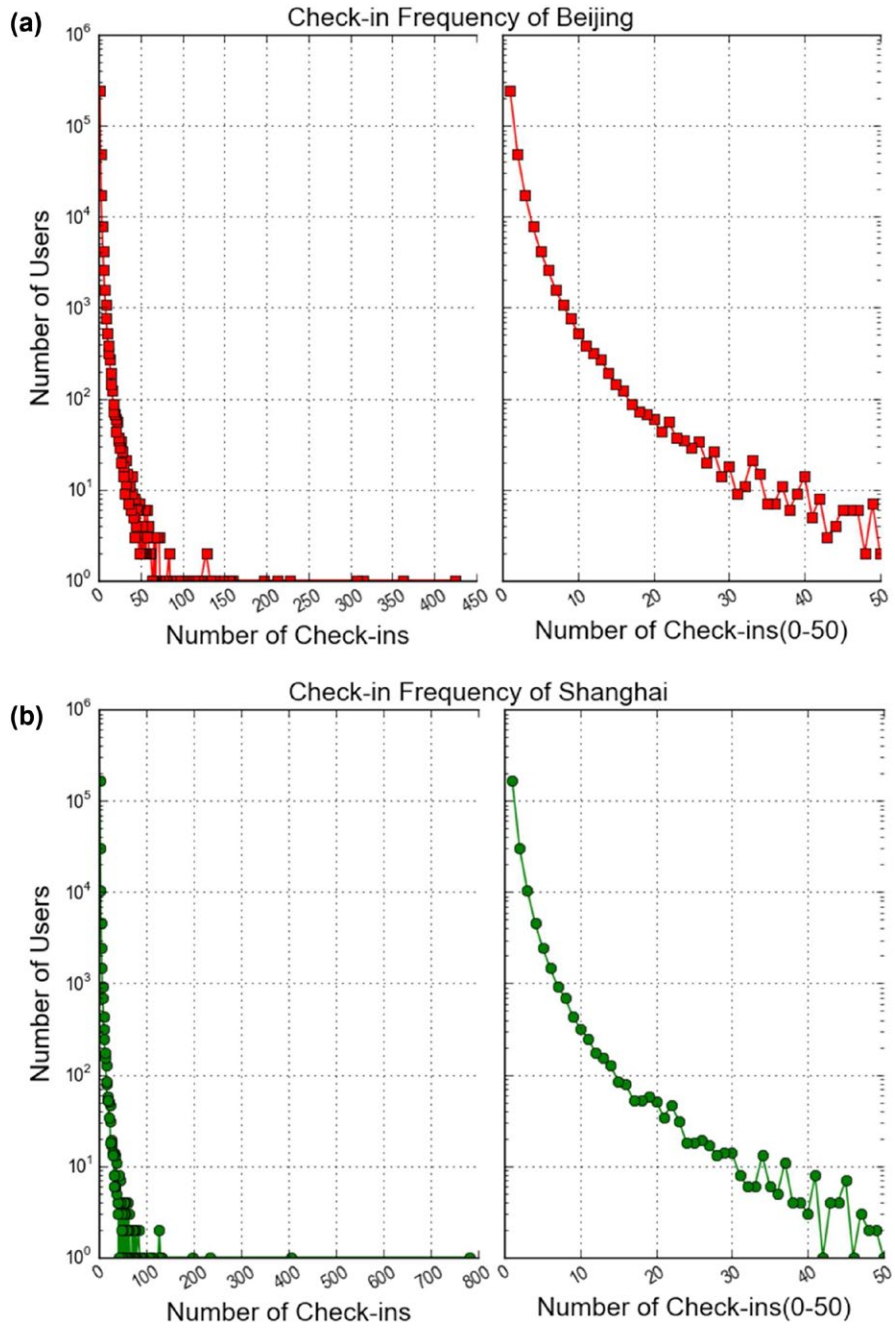

FIGURE 1 The frequency distribution of Weibo check-in data: (a) Beijing; (b) Shanghai; and (c) Guangzhou. For each city, the left sub-figure shows the distribution of all the data, and the right sub-figure shows a "zoom-in" view of users with fewer than 50 check-ins

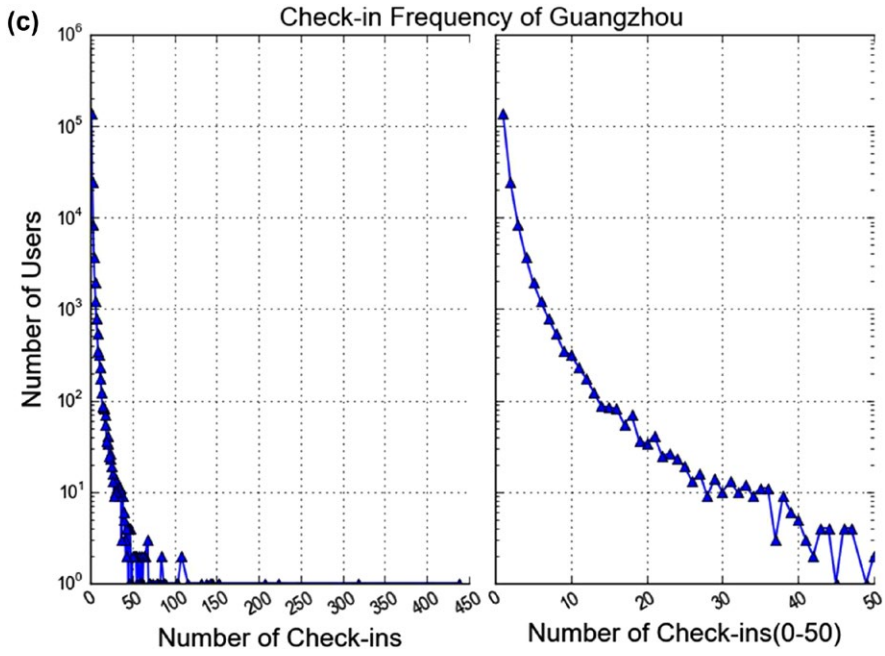

FIGURE 1 Continued

financial hub known as the "Pearl of the Orient," with a population of 24.15 million. Guangzhou is a crucial port city northwest of Hong Kong, with a population of 12.70 million (National Bureau of Statistics of China, 2016). Note that we only included the population within the city limit, instead of the entire metropolitan area. We extracted 1.18 million georeferenced Weibo posts (i.e., check-ins) for all three cities from April 2015 to March 2016 through the official Weibo API. Due to the power law nature of social media usage, many of these posts are from users who rarely use social media (Figure 1).

To ensure that we have adequate information to extract individual users' activity patterns, we only consider users with at least 10 check-ins during the study time span. Table 1 and Figure 2 show the number of users and check-ins after data cleaning. As shown in Figure 2, Weibo check-ins indicate a strong seasonal pattern, where October is the most active month. This is potentially due to the 9-day-long national holiday (the National Day) in China toward the beginning of October, during which Chinese residents often spend leisure time with their family. Figure 3 visualizes the geographic distribution of these check-ins based on a point density plot. As can be seen, this study uses the administrative boundary of the three cities instead of focusing on the central municipality areas for two reasons:

1. In Chinese megacities like Beijing or Shanghai, many local residents live in suburban towns or districts and work in the city center, so their activity space goes beyond the central municipality area (Na, Kwan, & Chai, 2015; Xu et al., 2015). This is particularly important for cities like Shanghai, where only 39% of check-ins are from the central urban area (Figure 4).
2. By using administrative boundaries, our study area is consistent with other studies that focus on analyzing urban mobility in these three cities using various types of big (geo) data (Ge, Shao, Xue, Zhu, & Cheng, 2017; Liu, Wang, Xiao, & Gao, 2012; Xu et al., 2015).

TABLE 1 Weibo metadata after data cleaning

City	Number of check-ins	Number of users
Beijing	48,997	2,775
Shanghai	33,194	1,781
Guangzhou	30,671	1,646

The extracted JavaScript Object Notation data from Weibo APIs contain various fields. Because this study focuses on analyzing human activity space, we only utilize the unique identifier (uid), the coordinates of check-in locations, and the timestamp of a check-in. Table 2 shows a few sample records.

3.2 | Methodology

As mentioned in Section 1, this article aims to explore the effectiveness of LBSM data for modeling human activity space. We are particularly interested in answering two questions. First, based on aggregated data from the entire sample set, how does data quantity affect different activity space indicators? We calculate the magnitude of two activity space indicators (ROG and entropy) based on different amounts of data. We then conduct a model fitting to approximate the limit of activity space indicators in Beijing, Shanghai, and Guangzhou. Second, how does the impact of data quantity vary for different users? For example, are there any outlier users whose activity space indicators behave differently compared to the majority? This second part of the methodology focuses on exploring the impact of data quantity on modeling activity space from an individual perspective, so we use Beijing as a case study. The procedure of the analysis is listed in Figure 5, and the details of each step are illustrated as follows.

3.2.1 | Define indicators

User activity space is defined by a great quantity of characteristics, such as scale, shape, structure, direction, and so on (Gonzalez et al., 2008). Due to the large number of indicators, here we choose the following two indicators to measure both the external and internal characteristics of activity spaces.

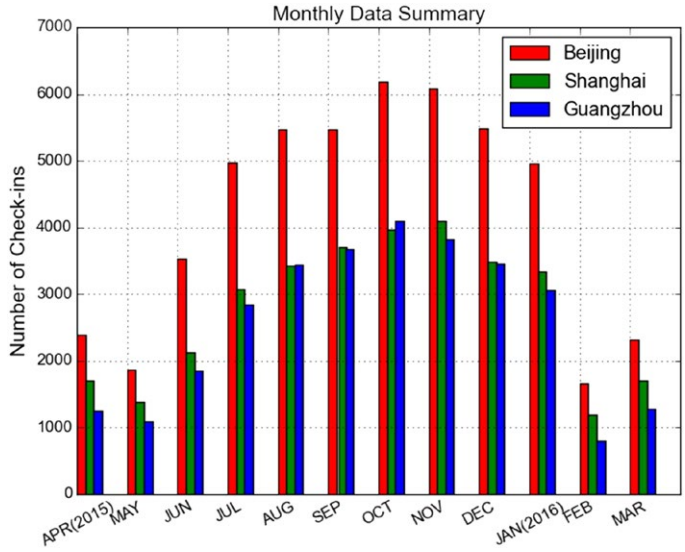

FIGURE 2 Monthly check-in data after data pre-processing

Radius of gyration

As mentioned by Gonzalez et al. (2008), ROG is considered an indicator of activity scale and a measurement of the external morphology, defined as:

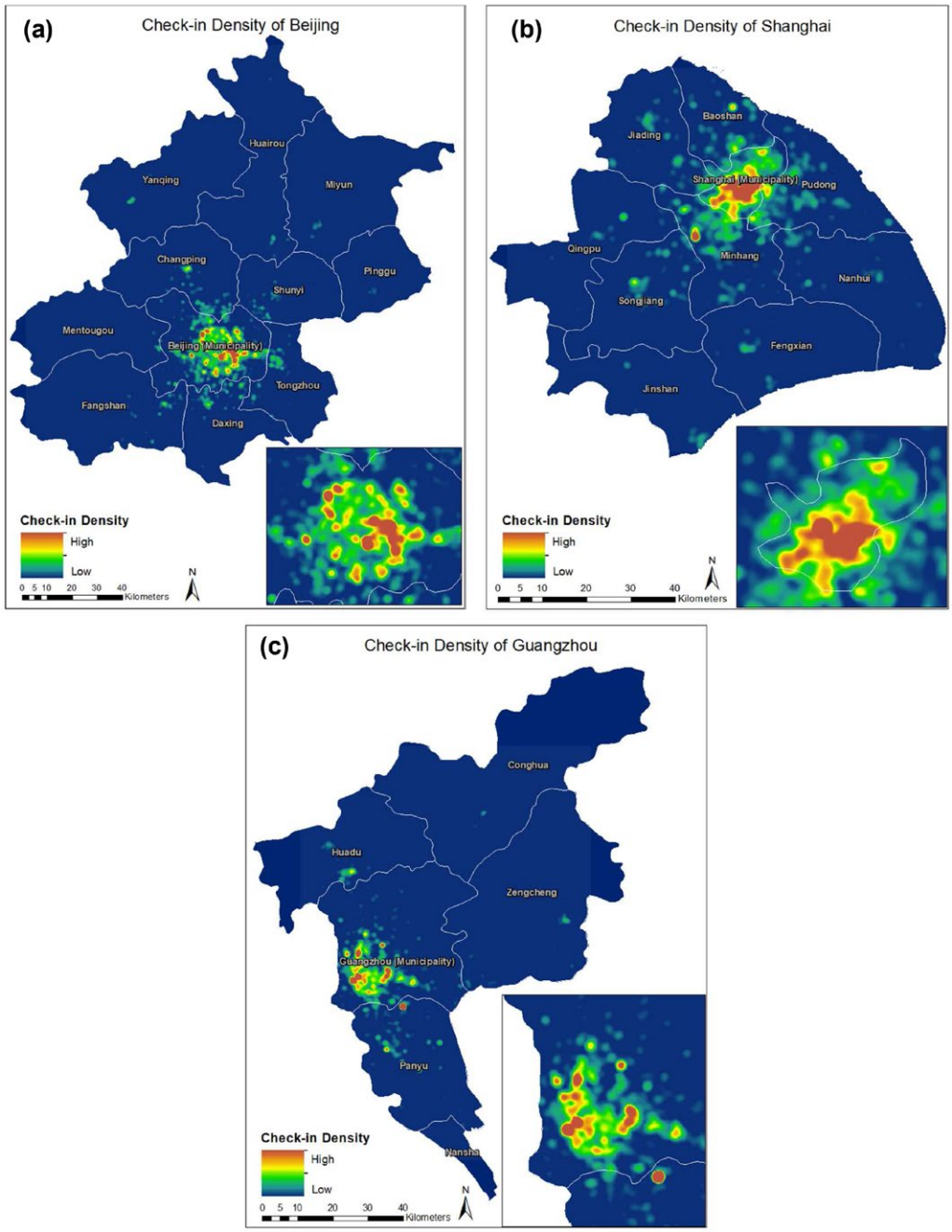

FIGURE 3 Weibo check-in data distribution: (a) Beijing; (b) Shanghai; and (c) Guangzhou

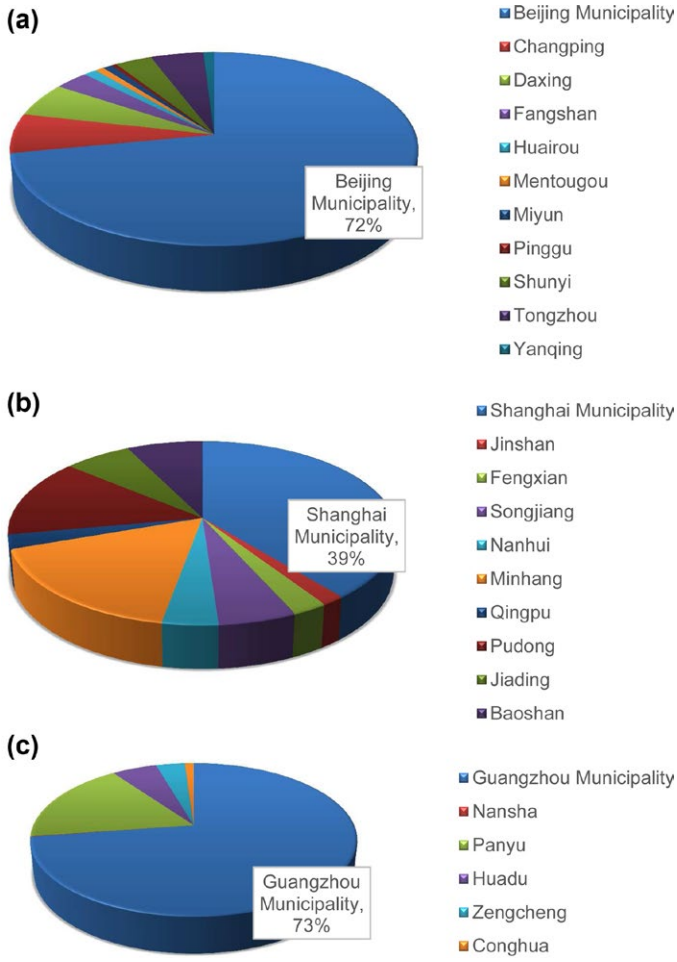

FIGURE 4 Percentage of check-ins in urban districts: (a) Beijing; (b) Shanghai; and (c) Guangzhou

TABLE 2 Example check-in records

uid[a]	Timestamp	Longitude	Latitude
187811******	2015-06-25 05:51:53	116.599239	39.908899
520391******	2015-11-11 11:27:09	116.419662	40.090118

[a]We removed the last few digits of user IDs due to privacy concerns.

$$\mathrm{ROG} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\vec{r_i} - \vec{r_m}\right)^2} \tag{1}$$

where n refers to the number of check-in locations of a given user; r_i is the geographical coordinate of each check-in location; and r_m refers to the centroid of all check-in points of a given user.

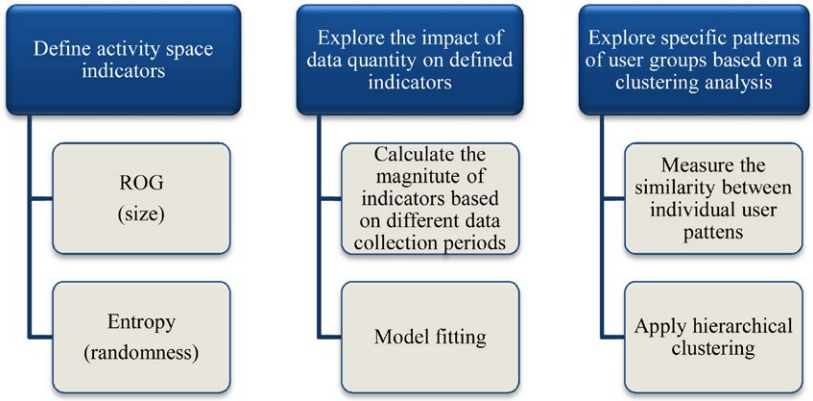

FIGURE 5 Flow chart of methodology

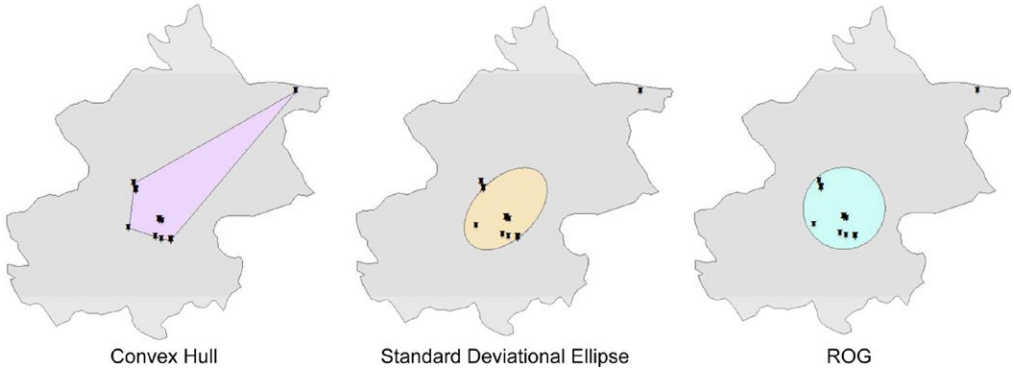

FIGURE 6 Approximating user activity space from different methods

As discussed in Section 2.1, researchers have proposed several other indicators to measure the morphology of activity space. However, many of these indicators are very sensitive to outlier points (Xu et al., 2015). Figure 6 shows the approximated activity space of an example user based on three methods (convex hull, standard deviational ellipse, and ROG). As can be seen, ROG is the least sensitive measurement to the outlier point located in northeast Beijing. Moreover, ROG also provides one single measurement (i.e., the radius) to represent the scale of the activity space.

Entropy

Entropy characterizes the heterogeneity of user activity patterns (Song et al., 2010). The formula is derived as follows:

$$E = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{2}$$

where p_i refers to the probability of a given user checking in at the same place i, and N stands for the total number of places where this user checked in. It is considered an indicator for the internal structure and randomness of activity space.

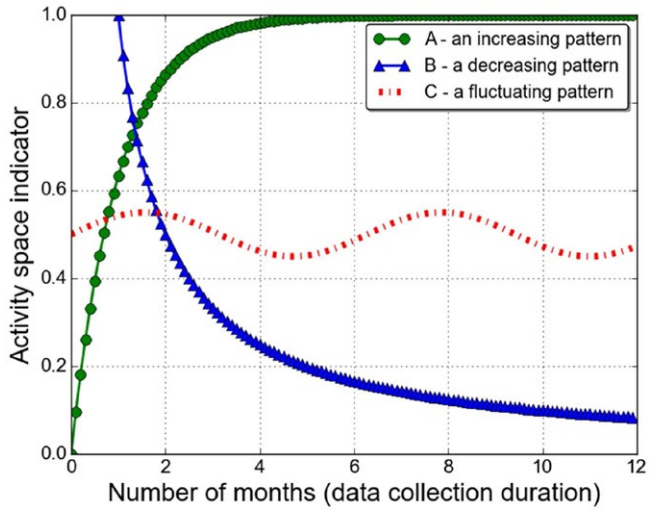

FIGURE 7 Simulated curves

3.2.2 | Model fitting

To explore the impact of data quantity in modeling LBSM user activity space, we examine how the magnitude of both indicators (ROG and entropy) changes with the amount of data used (from 1 month's to 12 months' data). We are interested in whether an indicator approaches a limit as the number of months increases, and if so, how to approximate this limit using a mathematical model. For example, the indicator may increase, decrease, or fluctuate as the amount of data increases (Figure 7). Understanding how the indicators change provides useful insight for choosing an appropriate data size in future studies.

3.2.3 | Clustering analysis

The previous step addresses the impact of data quantity on activity space modeling from an aggregated perspective. In addition, the curves in Figure 7 ("activity curves") can be plotted for each user to further explore how this influence varies at an individual level. We conduct a clustering analysis to extract both representative curves and outliers for the two indicators (ROG and entropy). Because this second part of the analysis aims to capture the similarities and outliers in individual activity patterns, we use Beijing as an exemplary study.

Here, we use the dynamic time warping (DTW) method to measure the similarity between activity curves in the clustering analysis. DTW is a robust distance measure widely applied in the fields of speech recognition and computer engineering, which is able to match stretched or distorted time series (Keogh & Ratanamahatana, 2005; Senin, 2008). Figure 8 shows an example of Euclidean distance and DTW distance when measuring the distance between two curves *A* and *B*. As can be seen, a non-linear (elastic) DTW alignment generates a more intuitive similarity measure by shifting curve *B* to the right to optimize the alignment of two curves. Methodologically, DTW aims to construct a distance matrix between each node pair of curves *A* and *B* and find the best alignment between time series *A* and series *B*, which is the shortest path through the distance matrix (Figure 9). Therefore, DTW is more robust in dealing with distortion, lags, and displacement in time series, which can be a common issue in activity patterns analysis (Yuan & Raubal, 2012).

For example, Figure 10 shows three example ROG activity curves from our sample set. As can be seen, activity curves *A* and *B* indicate a similar increasing pattern with a time lag, whereas *C* demonstrates a fluctuating

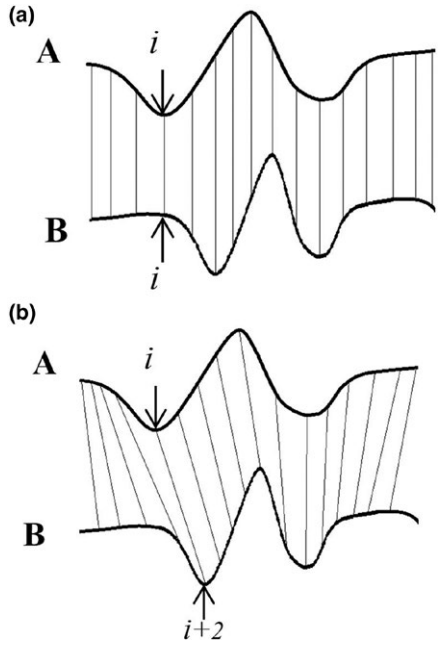

FIGURE 8 Measuring the distance between time series: (a) Euclidean distance; and (b) DTW distance

trend after the sixth month. Table 3 presents the distance measures between each pair of the three curves based on three commonly used similarity measures (DTW, Euclidean, and Fréchet; Eiter & Mannila, 1994). As shown in the results, the distance between *A* and *B* is larger than the distance between *B* and *C* based on a DTW algorithm, indicating that curves *A* and *B* are more similar. This is consistent with our common sense. However, both Euclidean and Fréchet distances indicate the opposite result, and neither of the two is able to effectively capture the similarity between curves *A* and *B*. This example further demonstrates that DTW is

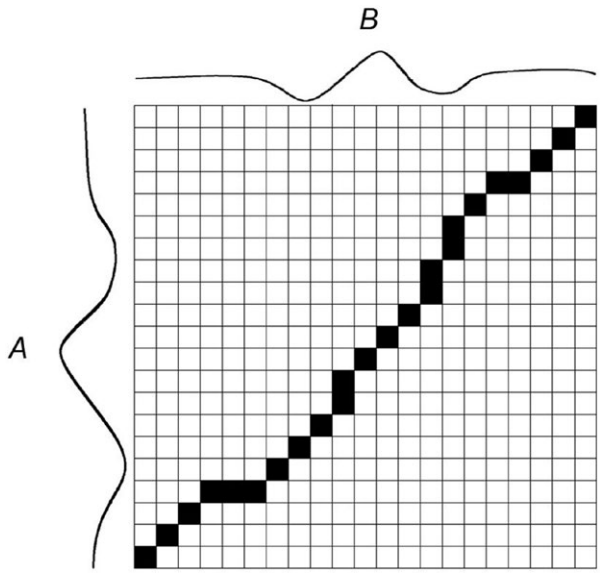

FIGURE 9 Constructing a distance matrix in the DTW algorithm

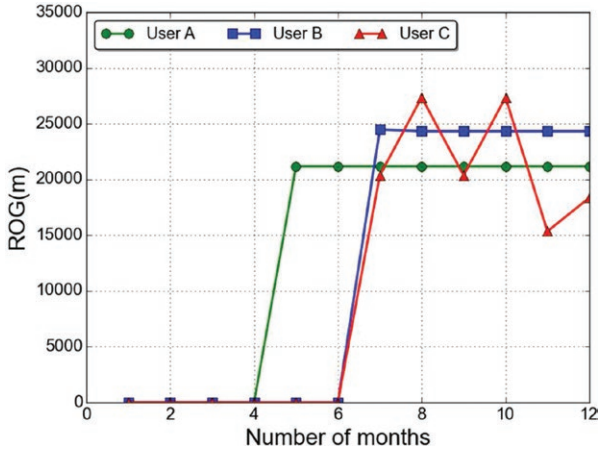

FIGURE 10 Three example user activity curves

TABLE 3 Comparing DTW, Euclidean, and Fréchet distances

	DTW	Euclidean	Fréchet
Distance (*A*,*B*)	0.4133	0.3092	0.3092
Distance (*B*,*C*)	0.4408	0.2565	0.1293

better at differentiating various curve shapes with distortion and time lag, which is a common data issue in our analysis.

4 | RESULTS AND DISCUSSION

4.1 | Generic analysis

As a generic analysis, to explore the correlation between activity space indicators and the amount of data utilized, we calculate the average ROG and entropy values with different data collection durations (from 1 month's data up to 12 months' data).

Figure 11 shows how ROG and entropy values change with different amounts of data used. As can be seen, in all three cities, both indicators show an increasing trend with a longer data collection period; however, the increasing trend slows down and the indicator approaches a limit value as the amount of data continues to grow. This is consistent with the assumption of time geography (Hägerstrand, 1970), where an individual's daily activity space is restricted to a certain spatial range due to physical constraints (e.g., moving speed), administrative boundaries, lifestyles, and so on.

To further quantify the change of activity space indicators with different data collection durations, we plot (Figure 12) the percentage of increase (p_i) of the indicator:

$$p_i = \frac{X_{i+1} - X_i}{X_i} \tag{3}$$

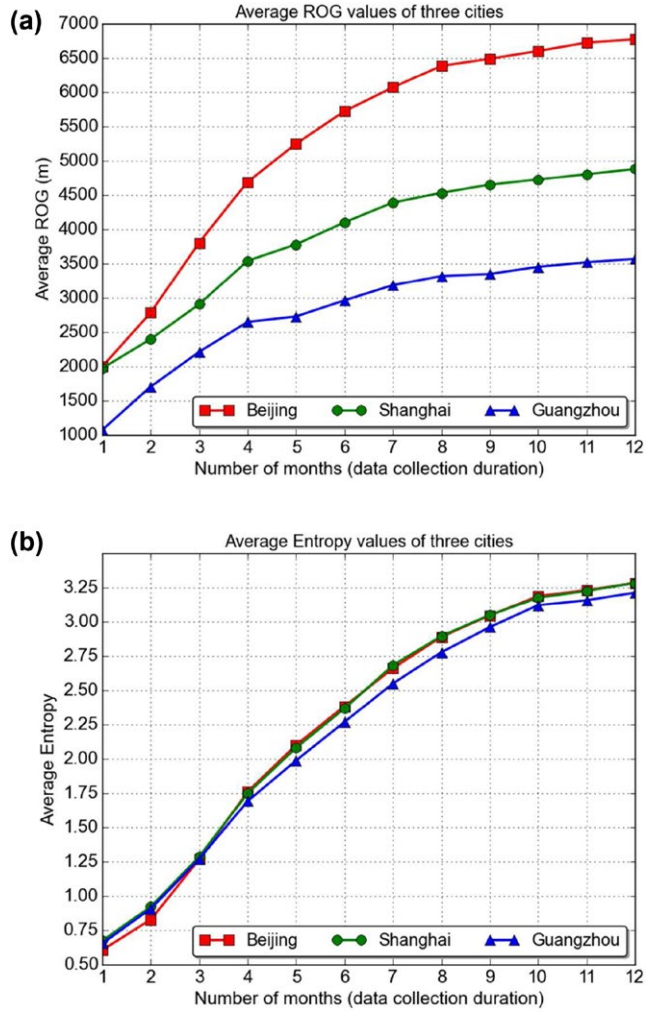

FIGURE 11 Average activity space indicators with different data collection durations: (a) ROG; and (b) entropy

where X_i stands for the value of indicator X calculated using i months' worth of data. As can be seen from Figure 11, the increase of both ROG and entropy values decreases when the amount of data increases. Since the correlation does not appear to be linear, we apply a logarithmic transformation on p_i and construct the following regression model:

$$\log(p_i) = am + b \tag{4}$$

where m is the number of months of data used in the analysis, and a and b are the coefficient and intercept of the fitted regression model (Table 4).

As can be seen, when the number of months increases, p_i is close to zero, indicating that both indicators are approaching a limit value as the data size increases. Based on the fitted regression models, we approximated this limit value based on a 120 month simulation. The approximated limit values are considered "the true value" of an indicator when the data collection duration is unlimited. Tables 5–7 show a comparison between observed ROG and entropy and the simulated limit values.

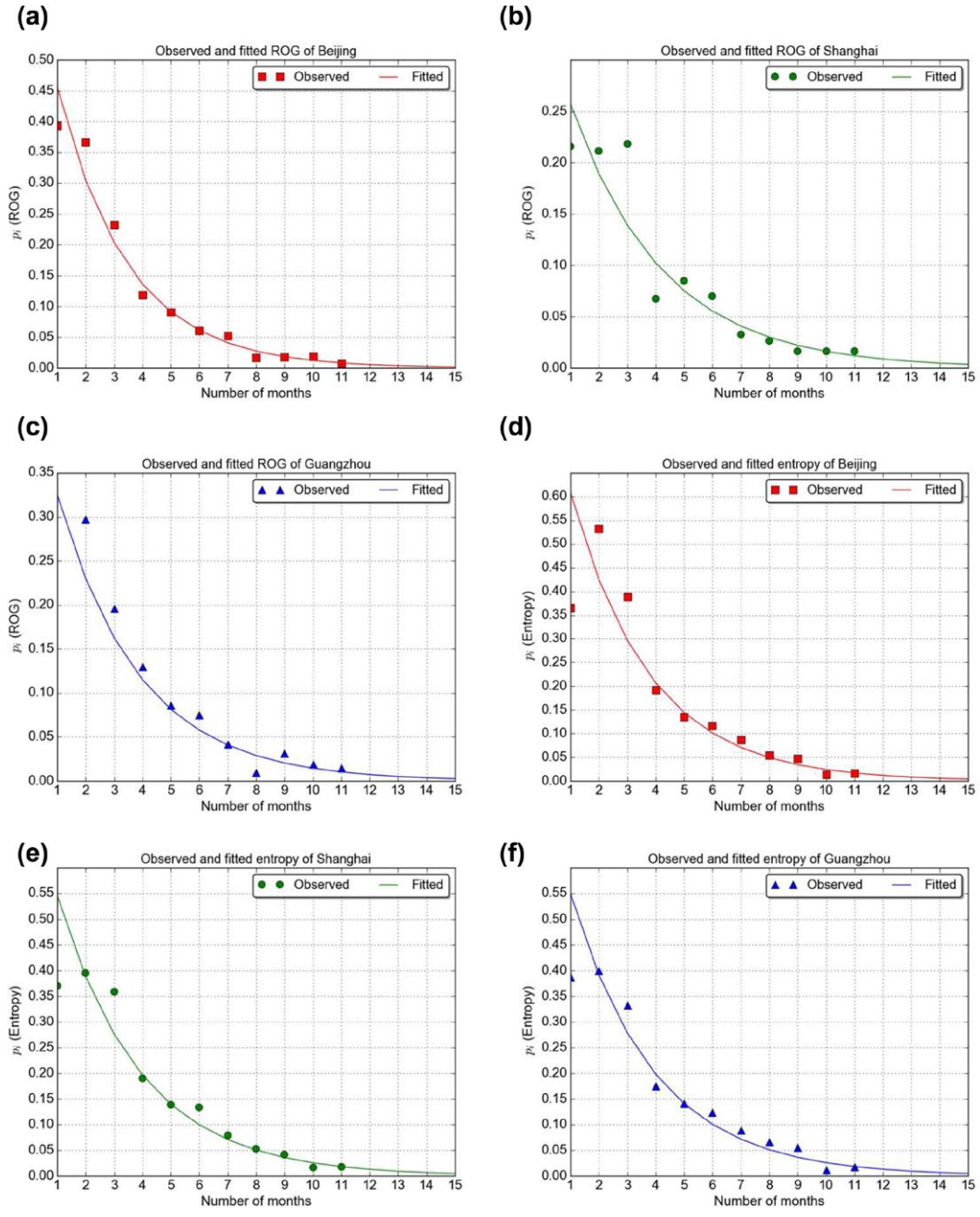

FIGURE 12 Observed and fitted p_i: (a) ROG (Beijing); (b) ROG (Shanghai); (c) ROG (Guangzhou); (d) entropy (Beijing); (e) entropy (Shanghai); and (f) entropy (Guangzhou)

As can be seen, when using 12 months' data, the calculated data is very close to the approximated limit value (over 96% for both ROG and entropy in all three cities). This result can be interpreted from multiple perspectives. On the one hand, the simulated ROG and entropy limits provide quantitative evidence to interpret user activity scale and randomness in a certain city. For example, in Beijing, the average ROG is approximately 6.88 km, which is larger than the limits in Shanghai (5.00 km) or Guangzhou (3.64 km). This is potentially determined by the city's

TABLE 4 Parameters of fitted regression models

		Coefficient (a)	Intercept (b)	Coefficient of determination (R^2)
Beijing	ROG	−0.4027	−0.3823	0.968
	Entropy	−0.3596	−0.1358	0.941
Shanghai	ROG	−0.3077	−1.0494	0.935
	Entropy	−0.3412	−0.2632	0.957
Guangzhou	ROG	−0.3463	−0.7771	0.936
	Entropy	−0.3407	−0.2558	0.918

TABLE 5 Comparison of observed indicators and the approximated limit (Beijing)

Number of months (1–12)	Average ROG (m)	ROG value/simulated ROG limit (%)	Average entropy	Entropy value/simulated entropy limit (%)
1	1999.33	29.04	0.6055	17.74
2	2784.05	40.44	0.8271	24.23
3	3805.92	55.29	1.2675	37.14
4	4691.59	68.15	1.7604	51.58
5	5248.10	76.24	2.0997	61.52
6	5721.44	83.12	2.3833	69.83
7	6069.68	88.17	2.6599	77.94
8	6383.70	92.74	2.8892	84.66
9	6489.79	94.28	3.0453	89.23
10	6602.11	95.91	3.1876	93.40
11	6724.35	97.68	3.2307	94.66
12	6772.00	98.38	3.2839	96.22
Limit value	6883.71	100	3.4129	100

size, planning, and structure. However, the entropy values in the three cities appear to be similar, indicating that the size and structure of the cities has little impact on the randomness of individual activity space. The same analysis can be extended to different cities to study the impact of urban setting on activity space. On the other hand, we further confirmed that 1 year's data is capable of capturing over 95% of variability in both ROG and entropy, whereas 6 months' worth of data can only approach 70%–80% of the limit value. Future studies can adopt a similar methodology to determine a balance point between data quantity and analytical precision. In addition, the difference between ROG and entropy is worth noting. For example, with 6 months' worth of data in Beijing, the calculated ROG is able to reflect over 83.12% of the limit value; however, for entropy, this proportion drops to 69.83%, suggesting that various indicators may have a different level of sensitivity toward data quantity. Similar patterns exist for Shanghai and Guangzhou, indicating that entropy values require a longer data collection period to stabilize.

4.2 | Clustering analysis

In addition to exploring the impact of data quantity on activity space from an aggregated perspective, we are also interested in how LBSM users may respond differently to various data collection durations. As mentioned in

TABLE 6 Comparison of observed indicators and the approximated limit (Shanghai)

Number of months (1–12)	Average ROG (m)	ROG value/simulated ROG limit (%)	Average entropy	Entropy value/simulated entropy limit (%)
1	1974.22	39.51	0.6732	19.84
2	2400.45	48.04	0.9222	27.18
3	2906.91	58.18	1.2864	37.91
4	3541.28	70.88	1.7479	51.51
5	3780.70	75.67	2.0794	61.28
6	4101.40	82.09	2.3667	69.75
7	4388.91	87.84	2.6829	79.07
8	4530.99	90.69	2.8949	85.32
9	4650.74	93.08	3.0480	89.83
10	4727.55	94.62	3.1751	93.58
11	4804.86	96.17	3.2254	95.06
12	4881.48	97.70	3.2835	96.77
Limit value	4996.32	100	3.393	100

Section 3.2, we use Beijing users as a case study to explore individual variability. Figure 13 depicts different activity curves from two categories of users in Beijing: (1) users with more than 100 check-ins during the study period; and (2) users with only 10 check-ins (i.e., users who barely meet our threshold in the data cleaning process). As can be seen, the first group of users shows a stronger variation in ROG values. This is potentially due to the small sample size in the first category (i.e., there are only 20 users with more than 100 check-ins in our sample set). On the other hand, these users are potentially more active, so it is possible that their activity range fluctuated more during the study period and needs more data points to stabilize.

TABLE 7 Comparison of observed indicators and the approximated limit (Guangzhou)

Number of months (1–12)	Average ROG (m)	ROG value/simulated ROG limit (%)	Average entropy	Entropy value/simulated entropy limit (%)
1	1073.16	29.51	0.6534	19.67
2	1705.67	46.90	0.9065	27.29
3	2213.66	60.87	1.2687	38.20
4	2648.4	72.82	1.6914	50.92
5	2728.24	75.02	1.9875	59.84
6	2963.07	81.47	2.2682	68.29
7	3184.79	87.57	2.5501	76.77
8	3317.81	91.23	2.7773	83.61
9	3349.84	92.11	2.9599	89.11
10	3453.73	94.97	3.1210	93.96
11	3520.08	96.79	3.1570	95.05
12	3570.47	98.18	3.2127	96.72
Limit value	3636.84	100	3.3215	100

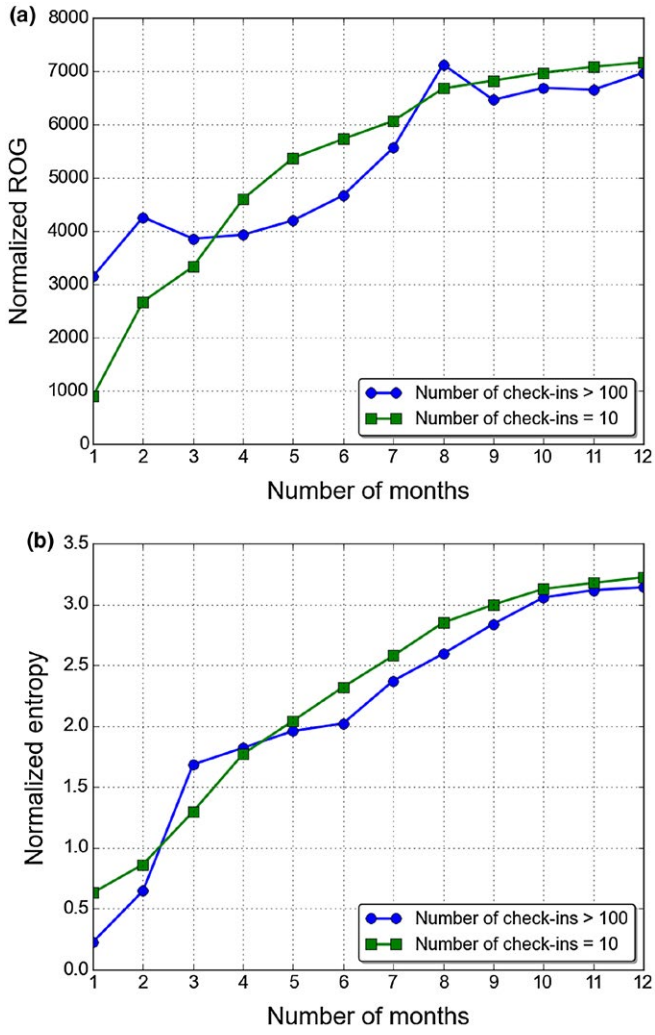

FIGURE 13 Comparing activity curves with different check-in frequencies: (a) ROG; and (b) entropy

To quantitatively cluster these individual patterns, we first normalized all activity curves to the range of [0,1] for individual users in Beijing and calculated the DTW distance among these curves. The output is a distance matrix storing the dissimilarity between each pair of users. We use the same distance matrix as the input of a hierarchical clustering analysis to identify similar groups and outliers in terms of how individual activity space indicators change with different data collection durations. When applying a hierarchical clustering method, the number of clusters is often affected by specific applications (Yuan & Raubal, 2012). As an example case study, we adopt the criteria discussed in Mardia, Kent, and Bibby (1979), where numCluster (the number of clusters) = max(2; sqrt(n/2)), and n is the number of users. Here we only include clusters with at least three users in the interpretation.

As can be seen from Table 8, when setting the number of clusters at 38, there are only two clusters with more than two users, and the majority fall into one cluster. Figure 14 shows the average activity curve of cluster *A*, which is very similar to the aggregated curve of the entire dataset in Figure 11a. This demonstrates a strong uniformity in terms of how user activity space indicators respond to different data collection durations. It further

TABLE 8 ROG clustering analysis summary

Cluster (>2 users)	Number of users in the cluster
A	2,730
B	3

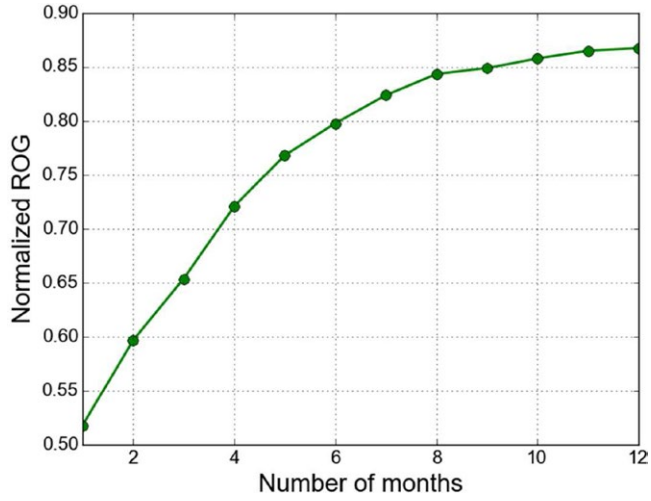

FIGURE 14 Average activity curve of cluster *A*

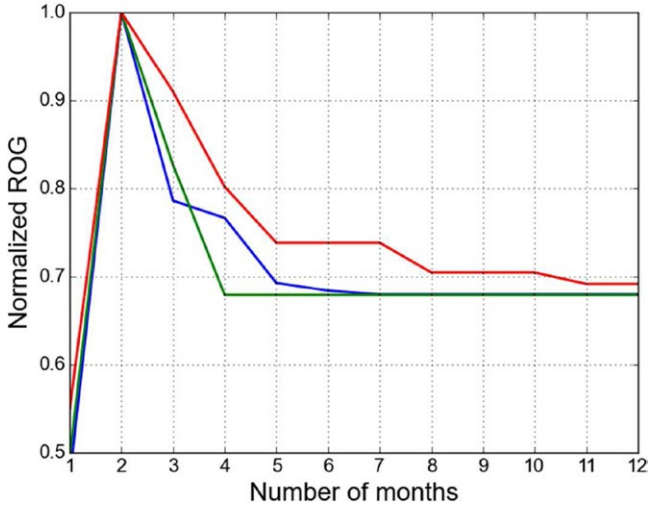

FIGURE 15 An outlier cluster *B*

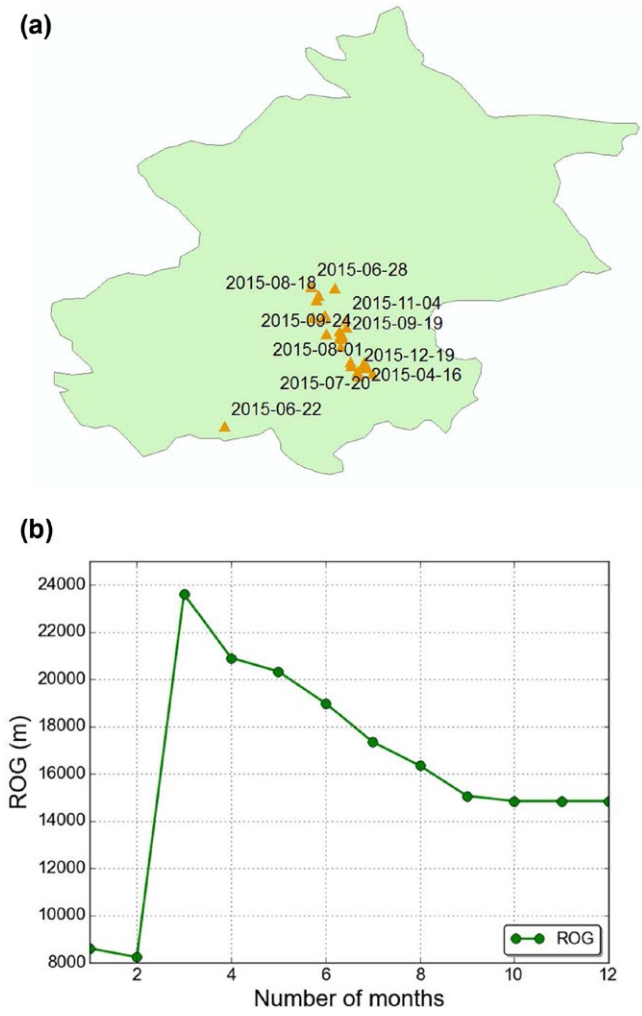

FIGURE 16 An example outlier user: (a) trajectory; and (b) activity curve

confirms that the methodology of simulating limit values in Section 4.1 is appropriate for the majority of users in our dataset.

Inevitably, there are outliers in the clustering results. As shown in Figure 15, the ROG of users in cluster *B* peaks at the second month of the data collection period (May 2015), which corresponds to a 3 day national holiday (Labor Day) at the beginning of May, when people in China often take extra vacation days to travel with their family. After that, the ROG in this group shows a declining trend, indicating that the users' activity space goes back to a relatively stable spatial range after the Labor Day holiday.

This outlier cluster also demonstrates that more collected data does not necessarily correspond to a larger ROG. Figure 16 shows the check-in locations of one example Weibo user, who mostly appeared in central Beijing, whereas there was one outlier point near the southern boundary of the Beijing study area on June 22, 2015. As can be seen in Figure 16, the calculated ROG values peak at 2 months and decrease when more check-in points are included (i.e., the user is back to his/her "normal" activity pattern), as more check-ins around central Beijing decrease the weight of the outlier point when calculating ROG.

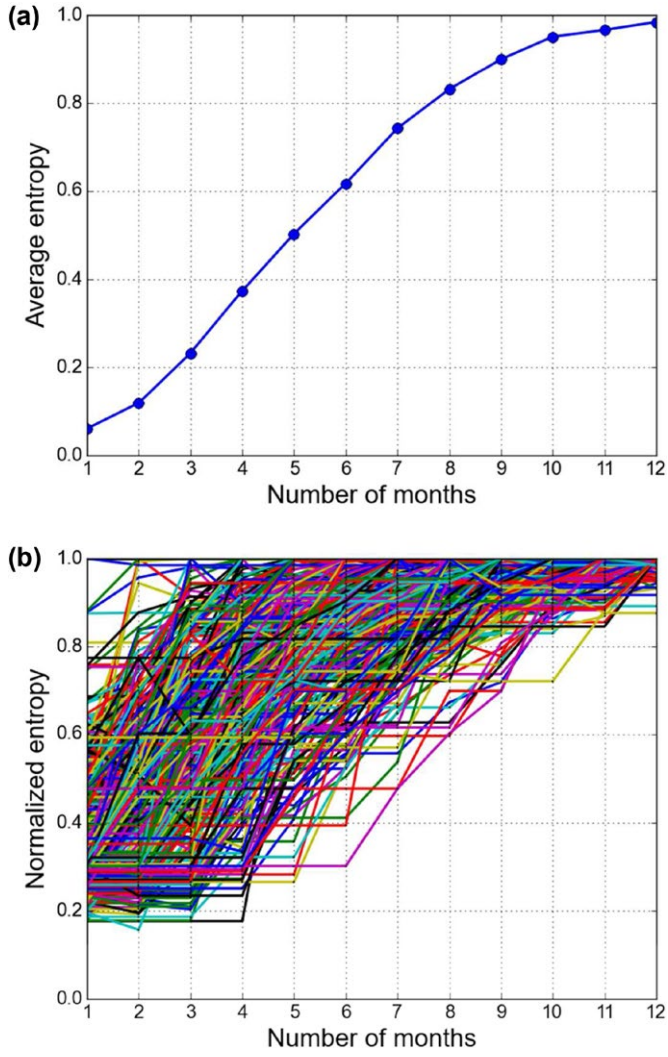

FIGURE 17 Increasing trend of entropy: (a) average activity curve; and (b) example users

Like ROG, we also clustered the entropy values and the results indicate a similar pattern to the ROG, where the majority of users (2,732 out of 2,775) show similarly increasing patterns when more data are collected (Figure 17). Based on the definition of entropy, it represents the level of randomness of user activity patterns. Hence, Figure 17 shows that, for the majority of users, more data reveals more randomness of their acitvity patterns.

However, Figure 18 shows two outlier clusters with a decreasing trend, where more check-in data at regular places actually improve the regularity of user patterns. Note that in Figure 18a, the entropy values reached a steady point because the users did not contribute sufficient data in the second half of the study period, whereas in Figure 18b, the entropy values of this cluster decrease gradually when the amount of data increases, which indicates that the longer the study period is, the higher level of regularity the data can reveal about these outlier users.

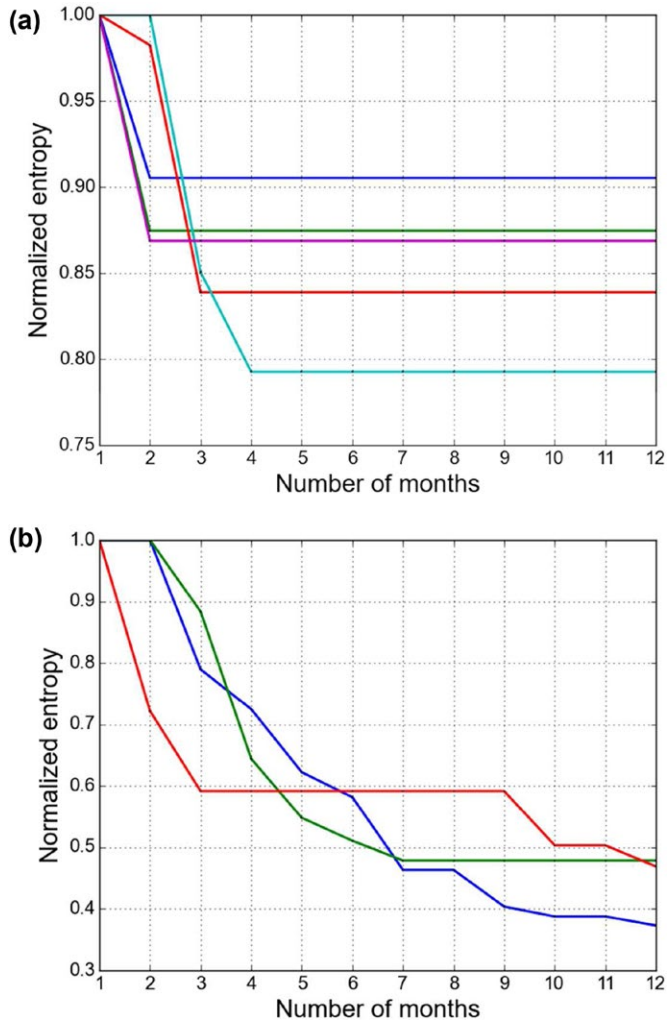

FIGURE 18 Entropy outlier clusters: (a) outlier cluster *B*; and (b) outlier cluster *C*

5 | CONCLUSIONS

This article explores the effectiveness of LBSM in modeling human activity space. More specifically, we tested how different amounts of check-in data affect the calculation of two activity space indicators, ROG and entropy, in three Chinese cities. We conducted two analyses to explore the change of activity space indicators from both an aggregated and an individual perspective.

- First, we fitted linear regression models with logarithmic transformation to reflect the correlation between the change of indicator values and the amount of data used. The results indicate that both ROG and entropy increase when the amount of data increase. However, the proportion of increase slows down and approaches zero eventually. We also approximated the limit values and verified that with 12-months' data, we are at approximately >95% magnitude of the limit values for both indicators in all three cities.
- Moreover, we also went one step further and investigated different user patterns based on a clustering analysis of Beijing users. The results indicate that even though for the majority of users, their ROG and entropy values

grow when the data collection duration becomes longer, we can still observe certain patterns where outlier points are penalized with a longer data collection window, resulting in a decreasing trend for activity space indicators. This result is helpful for investigating the heterogeneity of user behavior patterns and exploring how data collection strategies should be personalized for different target users in future studies. For example, for users who travel more during holidays, researchers should conduct more extensive data collection to stabilize any activity measurements that may be affected by outlier points (e.g., overseas travel). The analysis on the regularity behind individual behavior also provides new insight for modeling human activity space in the age of instant access, which is a crucial issue in human geography.

As mentioned in Section 1, this research contributes from both the methodological and the empirical perspectives. We examined a methodology to analyze the efficiency of LBSM in modeling human activity space, which can be used to optimize data collection in future studies. In addition, we also explored the activity space patterns for three of the largest cities in a rapidly developing country. The aggregated activity patterns and outliers provide valuable input for urban planners and policymakers to understand the dynamics of urban residents in three densely populated Chinese cities. We foresee that the broader impact of this research will yield an enhanced understanding of applying LBSM data in human activity studies and other widely applicable areas of geography, such as transportation and urban planning.

There are several limitations to this study that are worth further investigation. For example, due to the demographic biases of social media users, most active LBSM users are young people who are enthusiastic about the newest technologies, so the data used in this study is not a randomly selected sample of the entire urban population. In addition, this study extracts geotagged posts directly based on location, so we did not differentiate between residents and travelers. Appendix A describes a pilot study based on how likely a user is to be a resident of Beijing. However, since the scope of this article focuses on introducing a data pre-processing strategy instead of explaining the pattern of residents in a particular city, we did not eliminate users who are potentially travelers. It is also possible that computer algorithms (instead of real users) automatically generate certain Weibo accounts. Additionally, even though human activity patterns can be predictable, randomness is still an inevitable component of human mobility (Song et al., 2010), which leads to the difficulties and challenges in ground-truthing human activity studies. Future studies should focus on generating a more systematic framework to deal with the uncertainty issues of modeling user activity patterns from LBSM. We also plan to explore more activity indicators, such as the shape and movement direction of activity spaces. The methods and analysis proposed in this study can also be applied to other social media platforms to test their robustness and extensibility. Due to the limitation of data sizes in less developed areas in China, we did not investigate the patterns in smaller cities or rural areas. Future studies can also explore the similarity/dissimilarity between cities in various stages of development when the data becomes available.

REFERENCES

Abraham, S., & Lal, P. S. (2010). Trajectory similarity of network constrained moving objects and applications to traffic security. In H. Chen, M. Chau, S. Li, S. Urs, S. Srinivasa, & G. A. Wang (Eds.), *Intelligence and security informatics, PAISI 2010* (Lecture Notes in Computer Science, Vol. 6122, pp. 31–43). Berlin, Germany: Springer.

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys*, *43*(3), 1–43.

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., … Zook, M. (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.

Ahmed, M., Karagiorgou, S., Pfoser, D., & Wenk, C. (2015). *Map construction algorithms* (1st ed.). New York, NY: Springer Science+Business Media.

Bawa-Cavia, A. (2011). Sensing the urban: Using location-based social network data in urban analysis. In *Proceedings of the First Workshop on Pervasive Urban Applications*. San Francisco, CA.

Becker, R., Volinsky, C., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., … Varshavsky, A. (2013). Human mobility characterization from cellular network data. *Communications of the ACM*, *56*(1), 74–82.

Bennett, B., & Agarwal, P. (2007). Semantic categories underlying the meaning of 'Place'. In S. Winter, M. Duckham, L. Kulik, & B. Kuipers (Eds.), *Spatial information theory: COSIT 2007 (Lecture Notes in Computer Science*, Vol. 4736, pp. 78–95). Berlin, Germany: Springer.

Borrel, V., De Amorim, M. D., & Fdida, S. (2006). On natural mobility models. In I. Stavrakakis & M. Smirnov (Eds.), *Autonomic communication: Second International IFIP Workshop, WAC 2005, Athens, Greece, October 2-5, 2005, Revised Selected Papers* (Lecture Notes in Computer Science, Vol. 3854, pp., 243–253). Berlin, Germany: Springer.

Calabrese, F., Ferrari, L., & Blondel, V. D. (2015). Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys*, *47*(2), 25.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment & Urban Systems*, *51*, 70–82.

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, CA: ACM.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). Livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social* Media, Dublin, Ireland: AAAI.

Cuzzocrea, A., Song, I.-Y., & Davis, K. C. (2011). Analytics over large-scale multidimensional data: the big data revolution! In *Proceedings of the 14th ACM International Workshop on Data Warehousing and OLAP*. Glasgow, Scotland, UK: ACM.

Eiter, T., & Mannila, H. (1994). *Computing discrete Fréchet distance*. Retrieved from https://www.kr.tuwien.ac.at/staff/eiter/et-archive/cdtr9464.pdf

Elliott, A., & Urry, J. (2010). *Mobile lives*. London, UK: Routledge.

Gao, H., & Liu, H. (2015). *Mining human mobility in location-based social networks*. London, UK: Morgan & Claypool.

Ge, W., Shao, D., Xue, M., Zhu, H., & Cheng, J. (2017). Urban taxi ridership analysis in the emerging metropolis: Case study in Shanghai. *Transportation Research Procedia*, *25*, 4916–4927.

Golledge, R. G., & Stimson, R. J. (1997). *Spatial behavior: A geographic perspective*. New York, NY: Guilford Press.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782.

Goodchild, M. F., & Li, L. N. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120.

Haffner, M., Mathews, A. J., Fekete, E., & Finchum, G. A. (2018). Location-based social media behavior and perception: Views of university students. *Geographical Review*, *108*(2), 203–224.

Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, *24*(1), 7–21.

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the second ACM SIGKDD International Workshop on Urban Computing*. Chicago, IL: ACM.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography & Geographic Information Science*, *41*(3), 260–271.

Jenkins, A., Croitoru, A., Crooks, A. T., & Stefanidis, A. (2016). Crowdsourcing a collective sense of place. *Plos One*, *11*(4), e0152932.

Jones, P., Koppelman, F., & Orfueil, J. P. (1990). Activity analysis: State-of-the-art and future. In P. Jones (Ed.), *Developments in dynamic and activity-based approaches to travel analysis* (pp. 34–55). Aldershot, UK: Avebury.

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In *Proceedings of the 46th Hawaii International Conference on System Sciences*. Wallea, Maui, HI.

Keogh, E., & Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowledge & Information Systems*, *7*(3), 358–386.

Kwan, M. P. (2000). Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: A methodological exploration with a large data set. *Transportation Research Part C: Emerging Technologies*, *8*(1–6), 185–203.

Lee, J. H., Davis, A. W., Yoon, S. Y., & Goulias, K. G. (2016). Activity space estimation with longitudinal observations of social media data. *Transportation*, *43*(6), 955–977.

Levinson, D., & Kumar, A. (1995). Activity, travel, and the allocation of time. *Journal of the American Planning Association*, *61*(4), 458–470.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(33), 11623–11628.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, *105*(3), 512–530.

Liu, Y., Wang, F., Xiao, Y., & Gao, S. (2012). Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape & Urban Planning*, *106*(1), 73–87.

Long, J. A., & Nelson, T. A. (2013). A review of quantitative methods for movement data. *International Journal of Geographical Information Science*, *27*(2), 292–318.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment & Planning A*, *47*(2), 465–484.

Malleson, N., & Birkin, M. (2014). New insights into individual activity spaces using crowd-sourced big data. In *Proceedings of the ASE Big Data, Social Communication, and Cyber Security Conference*, Stanford, CA: ASE.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. New York, NY: Academic Press.

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

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2012). Understanding the demographics of Twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Barcelona, Spain: AAAI.

Mohammady, E., & Culotta, A. (2014). Using county demographics to infer attributes of Twitter users. In *Proceedings of the ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*. Baltimore, MD: ACL.

Na, T., Kwan, M.-P., & Chai, Y. (2015). Urban form, car ownership and activity space in inner suburbs: A comparison between Beijing (China) and Chicago (United States). *Urban Studies*, *53*(9), 1784–1802.

National Bureau of Statistics of China. (2016). *China Statistical Year Book*. Retrieved from https://www.stats.gov.cn/tjsj/ndsj/2016/indexeh.htm

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., & Ratti, C. (2010). Activity-aware map: Identifying human daily activity pattern using mobile phone data. In A. A. Salah, T. Gevers, N. Sebe, & A. Vinciarelli (Eds.), *Human behavior understanding* (Lecture Notes in Computer Science, Vol. 6219, pp. 14–25). Berlin, Germany: Springer.

Preoţiuc-Pietro, D., & Cohn, T. (2013). Mining user behaviours: A study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference*. Paris, France: ACM.

Qu, Y., & Zhang, J. (2013). Regularly visited patches in human mobility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris, France: ACM.

Roick, O., & Heuser, S. (2013). Location based social networks – Definition, current state of the art and research agenda. *Transactions in GIS*, *17*(5), 763–784.

Saini, J. S. (2014). A study of spam detection algorithm on social media networks. In G. Sai Sundara Krishman, R. Anitha, R. S. Lekshmi, M. Senthil Kumar, A. Bonato, & M. Grana (Eds.), *Computational intelligence, cyber security and computational models* (Advances in Intelligent Systems & Computing, Vol. 246, pp. 195–202). Berlin, Germany: Springer.

Schönfelder, S., & Axhausen, K. W. (2002). Measuring the size and structure of human activity spaces the longitudinal perspective. *Arbeitsbericht Verkehrs- und Raumplanung*, *135*.

Schönfelder, S., & Axhausen, K. W. (2010). *Urban rhythms and travel behaviour: Spatial and temporal phenomena of daily travel*. Farnham, UK: Ashgate.

Senin, P. (2008). *Dynamic time warping algorithm review*. Retrieved from https://seninp.github.io/assets/pubs/senin_dtw_litreview_2008.pdf

Sherman, J. E., Spencer, J., Preisser, J. S., Gesler, W. M., & Arcury, T. A. (2005). A suite of methods for representing activity space in a healthcare accessibility study. *International Journal of Health Geographics*, *4*(1), 24.

Silm, S., & 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.

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who Tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *Plos One*, *10*(3), e0115545.

Song, C. M., Qu, Z. H., Blumm, N., & Barabasi, A. L. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), 1018–1021.

Spielman, S. E. (2014). Spatial collective intelligence? Credibility, accuracy, and volunteered geographic information. *Cartography & Geographic Information Science*, *41*(2), 115–124.

Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, *30*(9), 1694–1716.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*. Ann Arbor, MI: AAAI.

Veregin, H. (1999). Data quality parameters. In P. A. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind. (Eds.), *Geographical information systems* (Vol. 1, pp. 188–189). New York: NY: Wiley.

Wu, L., Zhi, Y., Sui, Z. W., & Liu, Y. (2014). Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data. *Plos One*, *9*, e97010.

Xia, Y., Wang, G. Y., Zhang, X., Kim, G. B., & Bae, H. Y. (2011). Spatio-temporal similarity measure for network constrained trajectory data. *International Journal of Computational Intelligence Systems*, *4*(5), 1070–1079.

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., & Li, Q. (2015). Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. *Transportation*, *42*(4), 625–646.

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Lu, F., Chen, J., … Li, Q. (2016). Another tale of two cities: Understanding human activity space using actively tracked cellphone location data. *Annals of the American Association of Geographers*, *106*(2), 489–502.

Yuan, Y., & Medel, M. (2016). Characterizing international travel behavior from geotagged photos: A case study of Flickr. *Plos One*, *11*(5), e0154885.

Yuan, Y., & Raubal, M. (2012). Extracting dynamic urban mobility patterns from mobile phone data. In N. Xiao, M.-P. Kwan, M. Goodchild, & S. Shekhar (Eds.), *Geographic Information Science: 7th International Conference, GIScience 2012, Columbus, Ohio, Proceedings* (Lecture Notes in Computer Science, Vol. 7478, pp. 354–367). Berlin, Germany: Springer.

Yuan, Y., & Raubal, M. (2016). Analyzing the distribution of human activity space from mobile phone usage: An individual and urban-oriented study. *International Journal of Geographical Information Science*, *30*(8), 1594–1621.

Yuan, Y., Raubal, M., & Liu, Y. (2012). Correlating mobile phone usage and travel behavior – A case study of Harbin, China. *Computers, Environment & Urban Systems*, *36*(2), 118–130.

Yuan, Y., Wei, G., Chow, T. E., & Hagelman, R. R. (2017). Investigating social media landscapes from Twitter: A case study of Austin, Texas. In *Proceedings of the 25th International Conference on Geoinformatics*. Buffalo, NY: CPGIS.

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: An introduction*. New York, NY: Cambridge University Press.

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from Twitter data. In C.-W. Chung, A. Border, K. Shim, & T. Suel (Eds.), *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web* (pp. 439–444). Seoul, Korea: ACM.

Zhang, Z., Huang, K. Q., & Tan, T. N. (2006). Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In *Proceedings of the 18th International Conference on Pattern Recognition* (Vol. 3, pp. 1135–1138). Hong Kong, China: IEEE.

Zhao, X. L., & Xu, W. X. (2009). A clustering-based approach for discovering interesting places in a single trajectory. In *Proceedings of the Second International Conference on Intelligent Computation Technology and Automation* (Vol 3, pp. 429–432). Changsha, China: IEEE.

Zheng, Y., & Zhou, X. (2011). *Computing with spatial trajectories*. New York, NY: Springer.

How to cite this article: Yuan Y, Wang X. Exploring the effectiveness of location-based social media in modeling user activity space: A case study of Weibo. *Transactions in GIS*. 2018;22:930–957. https://doi.org/10.1111/tgis.12450

APPENDIX A: COMPARING URBAN RESIDENTS WITH TRAVELERS

In a pilot study to explore user backgrounds in Beijing, we divide all users in our sample set into the following three categories.

- "Residents": Users with more than 80% of their check-ins located in Beijing.
- "Uncertain": Users with more than 20% and fewer than 80% of check-ins located in Beijing.
- "Travelers": Users with fewer than 20% of their check-ins located in Beijing.

The ratio of the three groups in our sample set is "Residents": "Uncertain": "Travelers" = 0.845: 0.151: 0.004. As shown in Figure A1, the "Residents" and "Uncertain" groups indicate very similar patterns, whereas the "Travelers" group shows a higher variation for both ROG and entropy values. This is potentially due to the limited sample size (for both the number of users and the number of records from each traveler). The results further confirm the existence of

outlier users as discussed in Section 4.2, and provide a possible explanation for the outliers, that travelers inevitably exhibit different patterns from local residents.

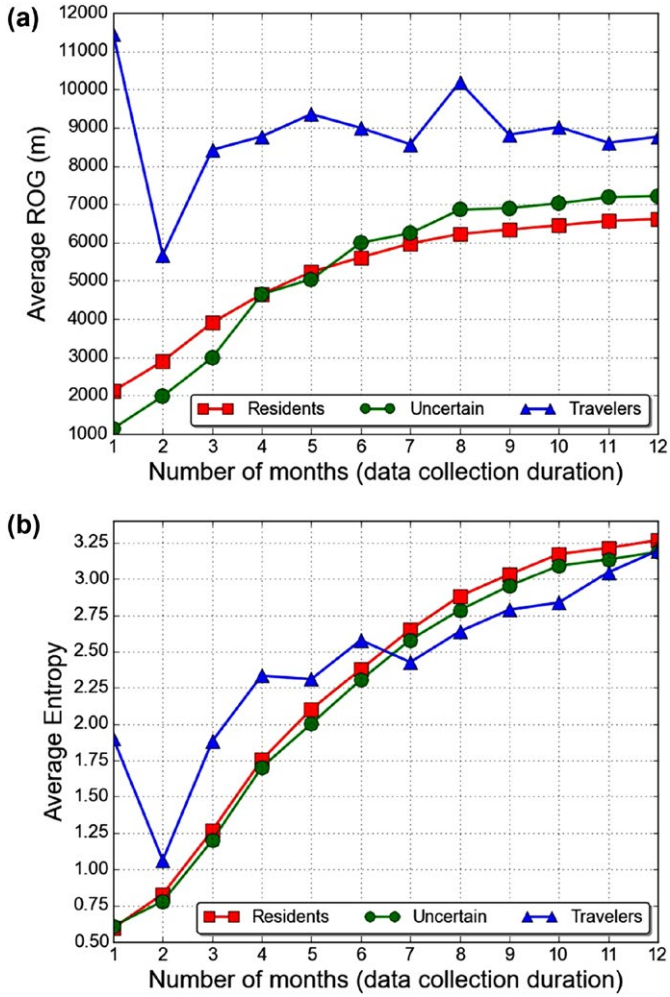

FIGURE A1 Comparison between residents and non-residents: (a) ROG; and (b) entropy