Arabian Journal of Geosciences (2018) 11: 173
https://doi.org/10.1007/s12517-018-3496-4

ORIGINAL PAPER

Exploring the potential of location-based social networks data as proxy variables in collective human mobility prediction models

Omid Reza Abbasi[1] · Ali Asghar Alesheikh[1]

Received: 23 June 2017 / Accepted: 15 March 2018 / Published online: 18 April 2018

Abstract

The study of human mobility has gained much attention in recent years. To date, various models have been developed to predict human mobility patterns for intra- and/or inter-city cases. These models incorporate the populations as proxy variables in the place of real variables which cannot be observed easily. However, inaccuracies in predicting human mobility within cities are usually encountered. One source of inaccuracies in intra-city scenarios arises from the fact that cities' populations are influenced by people from other areas. Therefore, population cannot be regarded as a good proxy variable for movement modeling. The objectives of this article are to introduce new proxy variables for use in current models for predicting human mobility patterns within cities, and to evaluate the accuracy of the predictions. In this study, we have introduced new proxy variables, namely, venues and check-ins, extracted from location-based social networks (LBSNs). In order to evaluate the models, we have compared our results with empirical data obtained from taxi vehicles, based on trip distances and destination population distributions. The Sørensen similarity index (SSI) and *R*-squared measures were also used to compare the performances of models using each variable. The results show that all models with LBSN variables can capture real human movements better within Manhattan, New York City. Our analytical results indicated that the predicted trips using LBSN data are more similar to the real trips, on average, by about 20% based on the SSI. Moreover, the *R*-squared measures obtained from regression analyses were enhanced significantly.

Keywords Human mobility · Mobility pattern prediction · Location-based social networks · Manhattan · OD matrix

Introduction

The prediction of human mobility patterns has various applications in urban planning (Camagni et al. 2002), land use management (Agarwal et al. 2002), traffic engineering (Jiang et al. 2009), emergency management (Bagrow et al. 2011), the spread of biological diseases (Brockmann et al. 2009; Prothero 1977; Wesolowski et al. 2012), the spread of mobile phone viruses (Wang et al. 2009), and location-based services (Buhalis and Amaranggana 2013). Many researchers have focused on the study of human mobility in different contexts, including intra-urban (Kang et al. 2012), inter-urban (Liu et al. 2014), individual (Gonzalez et al. 2008), or collective (Peng et al. 2012; Zheng et al. 2015) scenarios. Modeling trip distributions has a long history, beginning with the introduction of the well-known intervening opportunities (IO) model presented in the 1940s (Stouffer 1940). The existing human mobility prediction models can be categorized into parametric and parameter-free models. The former category includes gravity (Zipf 1946), rank-based (Noulas et al. 2012), and IO models (Stouffer 1940). Recently, a tendency toward modeling human mobility without the need for using adjustable parameters has been observed. Population-weighted opportunity (PWO) models (Yan et al. 2014) and radiation models (Simini et al. 2012) are two examples of this category. These models do not contain any adjustable parameters, and generally only need the spatial distribution of the population as their input (Yan et al. 2014).

Parameter-free models for predicting human mobility, such as PWO and radiation models, assume that people tend to select a destination that has relatively more benefits or opportunities. However, because of the difficulty in measuring each destination's opportunities, the models assume that the number of opportunities in a destination is proportional to its

✉ Ali Asghar Alesheikh
alesheikh@kntu.ac.ir

[1] Department of GIS, Faculty of Geodesy and Geomatics Engineering, K. N. Toosi University of Technology, Tehran, Iran

population. The models also assume that the number of trips departing from an origin is proportional to its population (Yan et al. 2014). In fact, the population plays the role of a proxy variable in the models. However, these assumptions might not be valid, especially in the case of boroughs. Difficulties can be mainly attributed to the interactions between neighboring boroughs (Masucci et al. 2013). For example, some trips occurring in Manhattan might be made by the residents of the neighboring boroughs (e.g., Bronx). On the other hand, place-based variables such as census tract population do not have sufficient temporal resolution to capture the true opportunities associated with the temporal resolution of the real data.

In the context of human mobility pattern prediction, models try to capture the underlying patterns behind people's movements in a system of zones. The design of the zones in a city should be similar to that of authority areas (e.g., census tracts), so that the results of applying models are directly applicable to the city. In a model, when the real data for a variable are not available, a proxy variable is used. Hence, the proxy variable should be a good representative of the real variable. Accurate numbers of produced and attracted trips in zones are not usually available, necessitating the use of a proxy variable, which is commonly the population of the zone. The aim of this paper is to evaluate the utilization of location-based social network (LBSN) data and places of interest (POIs) as proxy variables in the models. LBSNs are special types of social networks, where the users are able to share their locations and activities with each other as check-ins. POIs are the places in a city where people routinely perform their activities. These include stores, restaurants, airports, museums, clubs, hotels, offices, banks, and so on. The accurate positions of these so-called POIs or venues can also be extracted from LBSNs. The current assumptions made about population as a proxy variable in human mobility prediction models do not take the real conditions of intra-city areas into account. In contrast with the existing models, our study assumes that the number of opportunities in a zone is proportional to the number of places that an individual may find useful or interesting. In addition, we assume that the number of trips departing from an origin is proportional to the number of check-ins located in that zone. Considering the positive relationship between trips toward a destination and the check-ins located in it, we believe that LBSN data reflect statistics that are closer to reality than that those resulting from population data. From a decision-making process perspective, it is clear that when a person makes a decision about going to a destination, he or she does not evaluate the populations of the origin and the destination. The intervening places of interest, however, play a vital role in his or her decision-making process. Some researchers (Hasan et al. 2013; Li et al. 2016; Noulas et al. 2012) have leveraged geosocial network data to understand collective or individual human mobility patterns. Agryzkov et al. (2017) tried to answer the question of whether the data generated by Foursquare users are in agreement with activities within the city. In another study by Hristova et al. (2016), Foursquare data were used to analyze the social media footprints of attendees of sports games, in order to identify temporal, spatial, and microeconomic patterns. Noë et al. (2016) utilized Foursquare data to study the relationships between the personalities of users and the way they choose a place to visit. They concluded that people with a similar personality are more likely to visit a specific category of places. Despite vast and emerging research on LBSN data and mobility, the direct use of LBSN data as an alternative to population in human mobility pattern prediction models has not been evaluated, especially in intra-city scenarios where the interactions among many parts of the city are remarkable. Abbasi et al. (2017) used geosocial data as proxy variables within a rank-based model and concluded that they have good potential for utilization in this field. However, their results should be validated against other, more established models of human mobility prediction, such as gravity and IO models.

One of the most challenging issues in modeling trip distribution in the case of parameterized models is the availability of accurate data. Numerous studies on the subject have used various kinds of data sources, such as cellular networks (Caceres et al. 2007), GPS-enabled taxis (Peng et al. 2012), vehicle identification data (Zhou and Mahmassani 2006), and Bluetooth technology (Barceló et al. 2010). Wireless location technologies (WLTs) have also been used in several studies (Caceres et al. 2007). These data sources generally involve some issues, such as privacy concerns, low accuracy of positioning techniques, sample size, matching the region of analysis with regions used in the positioning method, and so on. For instance, positioning using a set of connected Bluetooth-enabled devices should be done only with the prior consent of users. The sample size is also a major issue in these data collection techniques. As the proxy variable is an alternative for use when real data are unavailable, the proxy data should be easy to collect. Since the LBSN data can be categorized as volunteered geospatial information (VGI), collecting them is a relatively easy task. In this study, check-in data have been extracted from the Foursquare social network through its application programming interface (API). According to its website,[1] more than 50 million people use Foursquare every month, so the penetration rate of its check-ins is higher than that of other LBSN services such as GeoLife and Loopt. Moreover, the positioning technique used in the LBSN is based on the built-in GPS sensors of smartphones. Therefore, the platial accuracy of such data is higher than that of the other sources.

In this study, we have computed four models for predicting human mobility, using both population (the standard proxy

[1] https://www.foursquare.com

variable) and LBSN data (the proposed proxy variable), in Manhattan. We have evaluated the models via real data obtained from taxi vehicles, using destination constraints (e.g., trip distances and destination population distributions) and some numerical measures (e.g., Sørensen similarity index (SSI), *R*-squared, and cosine similarity index).

The remainder of this article is organized as follows. In the following section, the materials and methods used in the study are introduced and the details of methods used in our evaluation section are presented. Then, results are provided, together with some discussion. The final section concludes the article and outlines future work.

Materials and methods

Study area

Manhattan is the most densely populated borough of New York City (NYC) and is one of the world's major commercial and financial centers. More than 1.5 million people live in Manhattan, which has a land area of about 60 km^2. The high density, the presence of various land uses, and huge interactions with neighboring boroughs (The Bronx, Brooklyn, Queens, and Staten Island) result in high mobility in Manhattan. In addition, there is a huge influx of daily commuters from New Jersey, Connecticut, and NYC suburbs such as White Plains and Long Island, who are surely making various trips within Manhattan throughout the day. It is worth noting that these people are not considered to be residents of Manhattan, and therefore, they are not reflected in population statistics reports. Therefore, we considered Manhattan as our study area for predicting human mobility patterns. Manhattan and its neighboring boroughs are shown in Fig. 1. In this article, we considered 288 census tracts within Manhattan as origin and destination zones for trips.

Data sets

In this study, we have used the US census counts made in April 2010 by the US Census Bureau, to extract the population distribution of Manhattan. For the sake of compatibility, the census tracts were also selected as trip zones.

In order to predict human mobility using LBSN data, check-in data for 18 months (from April 2012 to September 2013) from the Foursquare social network were used (Yang et al. 2015). This data set includes two large text files in which the data on check-ins and venue locations are stored. There are more than 33 million check-ins for 3.7 million venues within the files. The venue data set contains the venue ID used by the Foursquare system, the venue location, and the venue category name. The check-in data set contains the ID for the venue where the check-in occurred, an anonymous user ID and time information. To extract Manhattan data from the data set, a point-in-polygon analysis was performed. The data set contains 333,819 check-ins for Manhattan. Moreover, the locations of the POIs for which check-ins occurred, were extracted.

We used travel records for taxi passengers to evaluate the prediction accuracy of the models. The data set was collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab and Livery Passenger Enhancement

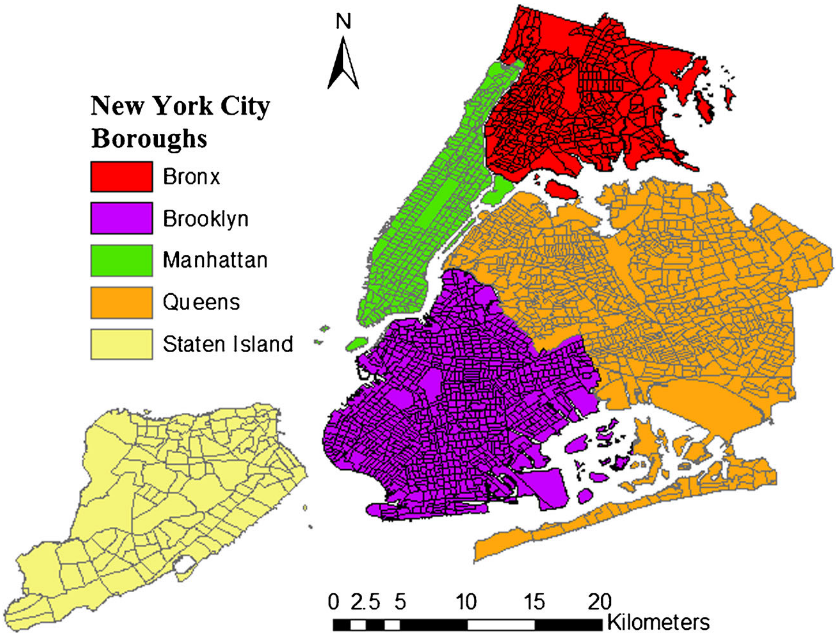


Fig. 1 Manhattan and its neighboring boroughs

Programs (TPEP/LPEP). The data set contains pick-up and drop-off times and locations, passenger counts, trip distances, and some other fields relating to payment. These data were collected throughout September 2013 and included about 800,000 trips. This data set includes both yellow and green taxis in Manhattan. Green taxis are allowed to pick up passengers only in Upper Manhattan and other boroughs. Therefore, in order to capture a more complete pattern, we merged data from green and yellow taxis. However, because only trips starting and ending in Manhattan should be accounted for, all the trips to (from) Manhattan from (to) other regions were filtered out. Usually, finer-resolution data sets suffer from having many zero counts. About 40% of our taxi data set consisted of zero counts.

Models

All models predicting human mobility patterns try to capture the decision-making process of travelers. This process is simulated in terms of the probability of going from one zone to another. This section introduces the models used in this study and outlines their relationships and the differences between them.

Gravity model

Analogous to Newton's law of gravity, the gravity model is a well-known framework with applications in various fields, particularly in spatial economics (Matyas 1998). As it is a parameterized model, it relies on the ground truth data for calibrating its parameters. The gravity model assumes that the flow between an origin and a destination is proportional to their attractions (in the literature, population is assumed to be a good representative of attraction), and decreases as the distance between them increases. The following equation is a common version of gravity model, called the doubly constrained gravity model:

$$T_{ij} = A_i T_i B_j T_j f\left(r_{ij}\right), \tag{1}$$

where T_i is the total number of trips departing from location i, T_j is the total number of trips arriving at location j, $f(r_{ij})$ is a function of the distance r_{ij}, and $A_i = 1/\sum_j B_j T_j f(r_{ij})$ and $B_j = 1/\sum_i A_i T_i f(r_{ij})$ are balancing factors that are dependent on each other. The balancing factors are calculated via an iterative procedure, which demands high computational effort. To simplify the calculations, one of the balancing factors can be set equal to one. This leads to a simpler form of gravity model, known as a singly constrained gravity model. In this study, we used a power distance decay function and a singly constrained (origin-constrained) gravity model, in which the trip distribution is described as:

$$T_{ij} = T_i \frac{m_j r_{ij}^{-\beta}}{\sum\limits_{k\neq i}^{N} m_k r_{ik}^{-\beta}}, \tag{2}$$

where β is an adjustable parameter, m_j is the population of the destination zone, N is the total number of zones in the city and the other variables are the same as in Eq. (1). In order to determine the parameter β of the model, we used taxi passenger trips as the ground truth data. The parameter β should be assigned a value that yields the best fitted distribution to the ground truth data. For this purpose, several numerical algorithms (Easa 1993; Evans 1971; Hyman 1969; Openshaw 1976; Williams 1976) have been developed. Due to its higher efficiency (Celik 2010), Hyman's calibration algorithm (Hyman 1969) was employed as the method for calibrating the model. Hyman's method aims to minimize the difference between the average cost of travel predicted by the model and the observed average cost of travel. The cost of travel in the gravity model is the distance between the origin and the destination. Therefore, the following equation should be minimized (Yan et al. 2014):

$$E(\beta) = \left|\bar{r}(\beta) - \bar{r}\right| = \left|\frac{\sum_i\sum_j T_{ij}(\beta) r_{ij}}{\sum_i\sum_j T_{ij}(\beta)} - \frac{\sum_i\sum_j T_{ij} r_{ij}}{\sum_i\sum_j T_{ij}}\right| \tag{3}$$

where $\bar{r}(\beta)$ is the average distance of predicted trips using parameter β and $\bar{r}$ is the average distance of observed trips. Since providing a direct solution for this equation is not straightforward, the algorithm uses an initial approximation for the parameter and utilizes an iterative procedure to solve the equation.

IO model

In the IO model (Stouffer 1940), unlike the gravity model, there is no direct use of distances between origins and destinations; only opportunities are considered. The IO model is defined as:

$$T_{ij} = T_i \frac{e^{-\alpha\left(S_{ij} - m_j\right)} - e^{-\alpha S_{ij}}}{1 - e^{-\alpha M}}, \tag{4}$$

where α is, again, the adjustable parameter of the model which should be determined using ground truth data, M is the total population in the city, and S_{ij} is the population within a circle centered at the destination, with a radius equal to the distance between the origin and the destination zone. In fact, in this model, the effect of distance has been latently modeled by using this variable.

PWO model

The PWO model (Yan et al. 2014) is a parameter-free model that requires the population distribution for predicting human mobility in cities. It is derived from a stochastic decision-making process and tries to predict an individual's destination based on opportunities. If the attractions are assumed to be inversely proportional to the populations of destinations and origins, the gravity model becomes a PWO model. The number of trips from location i to location j is computed as:

$$T_{ij} = T_i \frac{m_j \left(\frac{1}{S_{ji}} - \frac{1}{M} \right)}{\sum_{k \neq i}^{N} m_k \left(\frac{1}{S_{kj}} - \frac{1}{M} \right)}, \tag{5}$$

where T_i is the number of trips departing from origin i and m_i and m_j are the populations of the origin and the destination, respectively. The other variables are the same as in the previous equations.

Radiation model

The IO model is based on the assumption that the probability of traveling from one location to another is proportional to the population of the destination. Changing this to the ratio of the population of the destination j and the total population of the origin i and the destination j, yields the radiation model. The radiation model is also a parameter-free model for predicting human mobility and is computed as Simini et al. (2012):

$$T_{ij} = T_i \frac{m_i m_j}{\left(m_i + s_{ij}\right)\left(m_i + m_j + s_{ij}\right)}, \tag{6}$$

Note that in Eq. (6), s_{ij} is the population within a circle whose center is the trip origin and whose radius is r_{ij}. This model originates from diffusion dynamics (Kang et al. 2015).

The flow diagram below (Fig. 2) shows the procedure required for applying human mobility models to a city.

Results and discussion

Initially, we performed some preliminary analyses on the characteristics of POIs and the population distribution in Manhattan. The distribution of POIs in each tract, together with their populations, is presented in Fig. 3. It can be seen from the figure that the POIs are denser in the southern parts of the borough, possibly due to the high density of the built environment. In addition, the center of business and government of New York City is located in this region. However, since Upper Manhattan is mainly a residential area, southern parts are not as densely populated. Furthermore, Upper Manhattan is not a major center of tourism in NYC, resulting in lower numbers of associated trips. Therefore, the characteristics of the POI distribution in Manhattan are dissimilar to those of the population distribution.

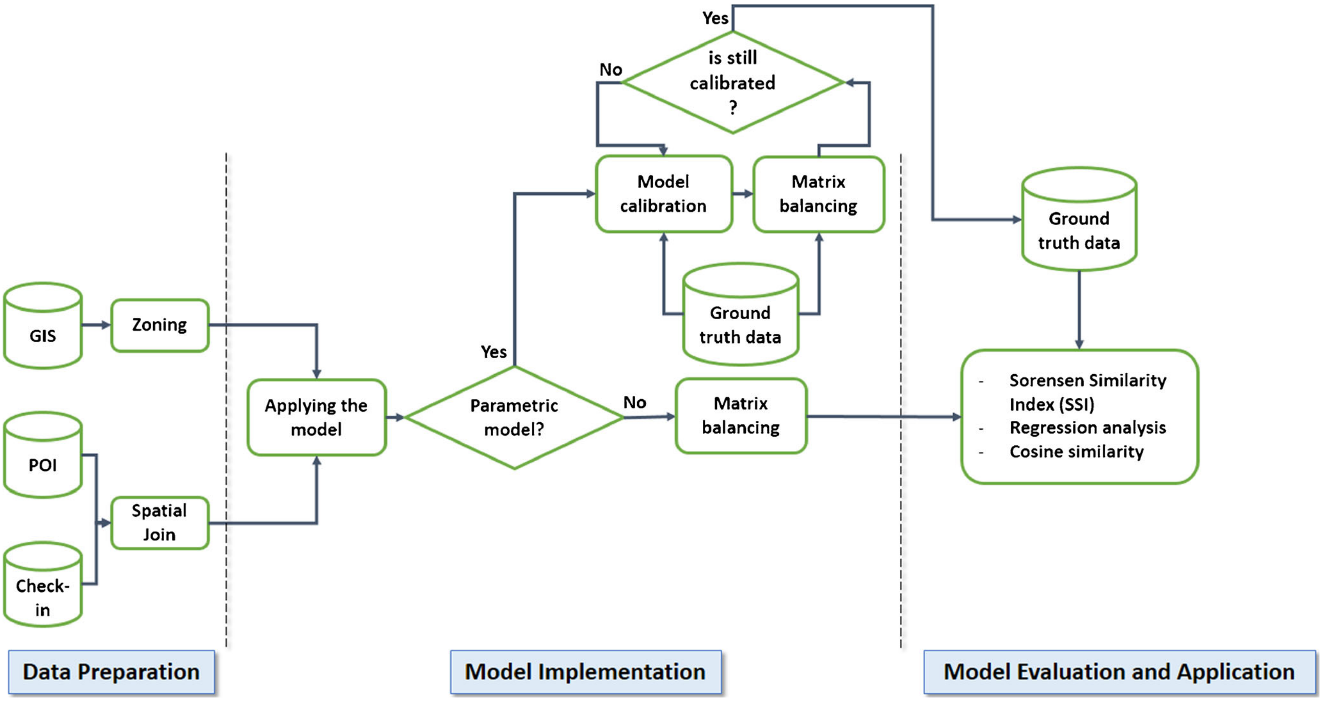


Fig. 2 The flow diagram for applying a human mobility model in a city

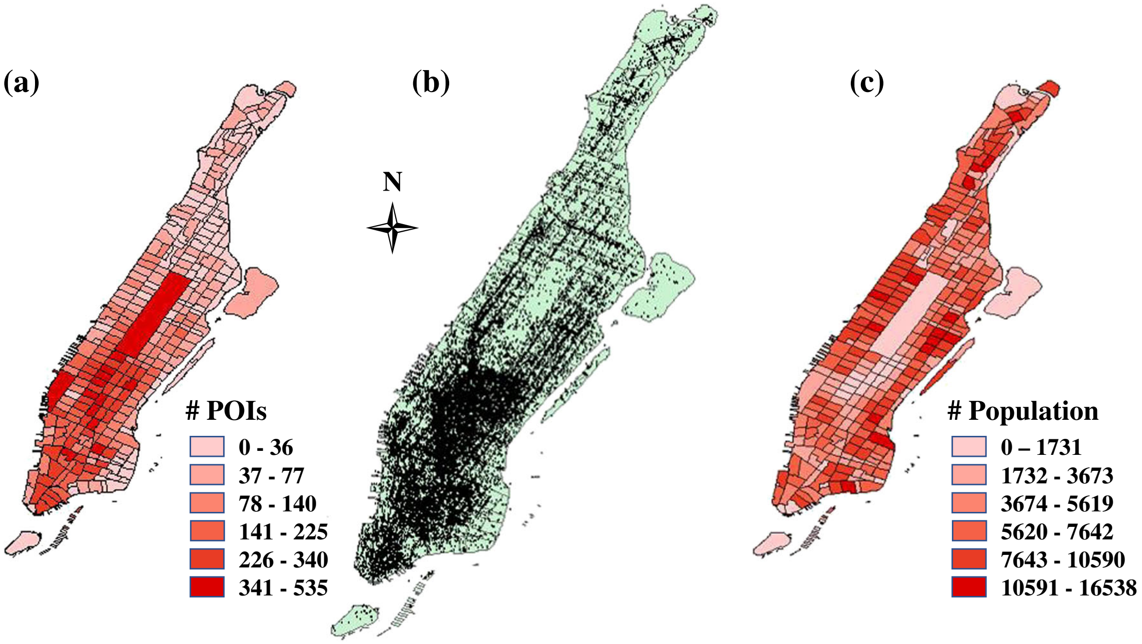
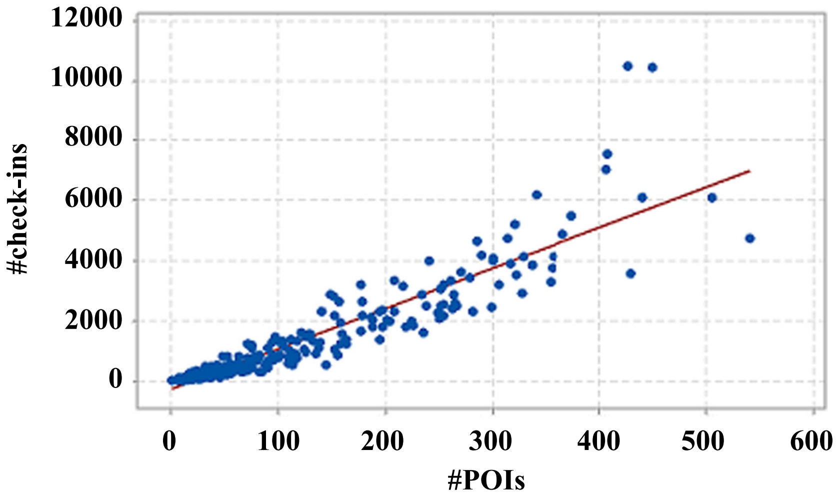


Fig. 3 **a** Color-coded map of the number of POIs in each tract, **b** distribution of POIs in Manhattan, and **c** color-coded map of population in each tract

As shown in Fig. 4, there is a relatively high positive relationship (R-squared = 0.844) between the number of POIs and the number of check-ins in a zone. Therefore, the more the POIs in an area, the higher the number of check-ins, which is analogous to the assumption that the higher the population in a zone, the higher the number of trips departing from it (Simini et al. 2012). Hence, it seems that our assumptions about trips are valid. In all the models introduced in the previous section, the number of check-ins occurring in a zone has the potential to be a proxy for the total number of trips produced in that zone (T_i), based on our assumptions. In a similar manner, the numbers of POIs located in the zones can act as proxies for the attractions of the zones. In addition, S_{ij} is computed using the POIs located in the aforementioned circle. To ensure that the total predicted fluxes and the total observed fluxes are matched, a normalization factor κ is also introduced into the models.

In order to compare the assumptions, we performed a distance distribution analysis for the trips. The distance between the origin and the destination is an important factor in traveling. In addition, the trip distance distribution can provide important evidence to urban and regional planners and other decision-making authorities within a city. With the help of this analysis, the effect of trip distance on the probability of traveling can be statistically studied. Figure 5 shows the probability of traveling between two locations at a distance r, produced by different models, using population variables and LBSN data. The plot has a logarithmic scale.

Fig. 4 Statistical relationship between number of POIs and number of check-ins in a zone



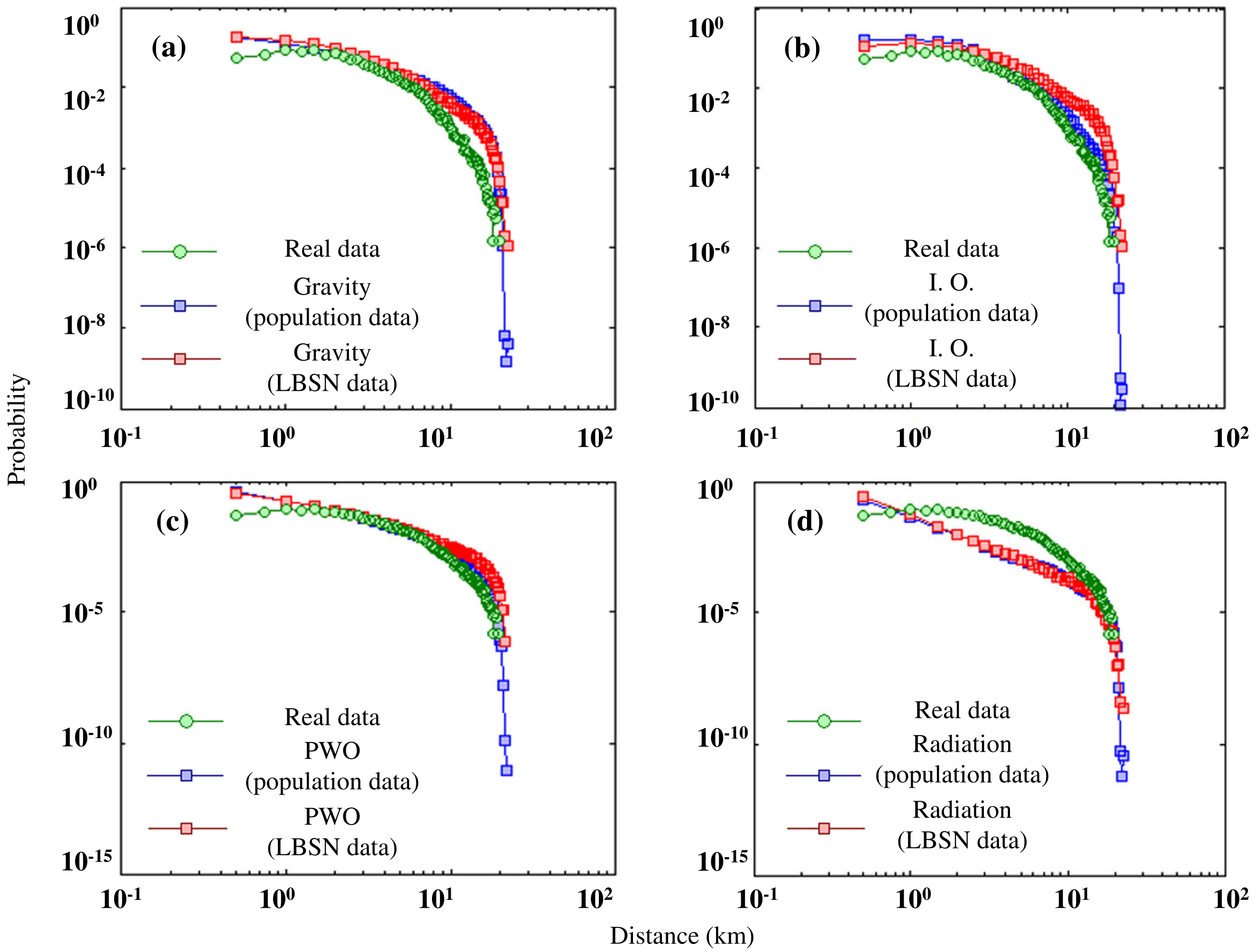


Fig. 5 Probability of traveling from location *i* to location *j* with respect to the distance between them

As shown in Fig. 5, when the models are based on the population, they show a more abrupt decay than when the inputs are based on LBSN data, suggesting that for long distances, the original assumptions underestimate the probability of making trips. Using LBSN data, the models predict the probability of making long trips more accurately. Apart from this, the other parts of the plots show no significant differences.

As far as the managerial decision-making process in a city is concerned, the population characteristics of the city represent aspects which have remarkable impacts on human mobility (Yan et al. 2014). We compared the probability of traveling from an origin to a destination with population *m*, produced by the models, with the empirical data (Fig. 6). This gave us a valuable measure of how much the population of a destination is representative of its attractions.

Figure 6 reveals that our assumptions regarding the use of LBSN data within cities are much closer to reality than the assumptions made in existing models (i.e., the use of population). It can be seen from Fig. 6 that the population-based models underestimate the probability of traveling to zones with low population. This happens very frequently at an intra-city level. There are some zones in Manhattan (e.g., Central Park) which have few residents (according to the Census Bureau's survey, Central Park has only one resident), but due to the land use, many trips are directed toward them. The existing models fail to predict the trips in such regions. As noted earlier, the northern parts of Manhattan are more populated, but trade centers are mainly located in Lower Manhattan. Thus, when predicting mobility via population, the probability of traveling to highly populated zones is overestimated. However, the predictions using LBSN data accurately match the real data for all models.

Furthermore, we conducted a test using the two-sampled Kolmogorov-Smirnov hypothesis (KS test) to determine whether the two samples of data could have come from the same underlying distribution at the 5% significance level. In essence, the KS test tries to determine if two samples differ

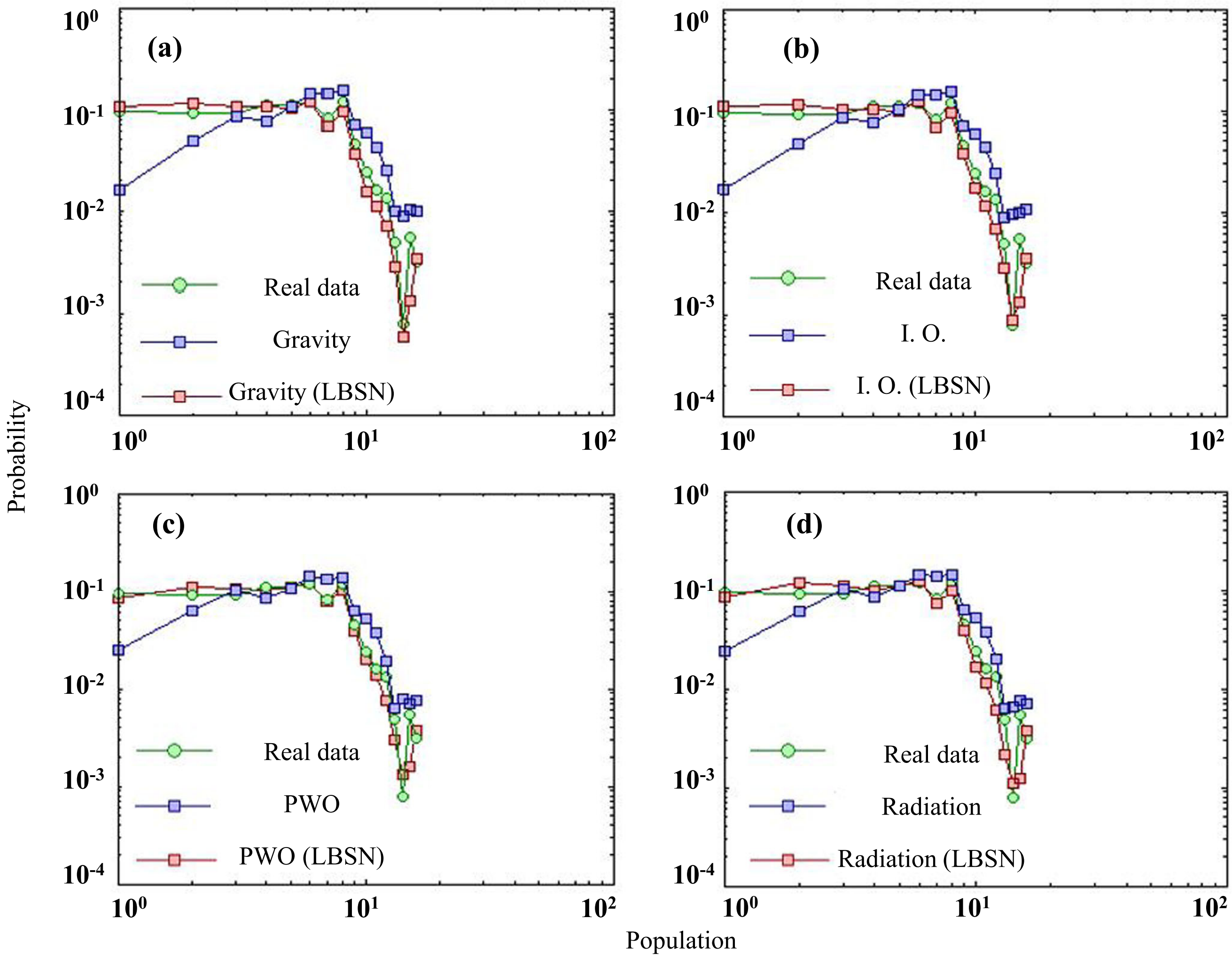


Fig. 6 Probability of traveling from location *i* to location *j* with respect to the population of the destination

significantly. It is a non-parametric hypothesis test and the underlying distribution of the samples need not be known. Tables 1 and 2 summarize the *P* values resulting from the KS test. The participating samples in the test are observed taxi trips and the estimated trips from each model.

As can be inferred from the tables, where the null hypothesis for the sample resulting from population data is accepted, the same is true for the estimated trips from the LBSN data. In most cases, the *P* values of the test for LBSN data are higher than those for population data.

In addition to the plots given above, we completed our evaluations using some numerical measures. The SSI is a similarity measure which evaluates the amount of closeness between two sample data sets. It has been used in this study to quantify the similarity of predicted and actual trips. The index is defined as (Lenormand et al. 2012):

$$SSI = \frac{2\sum_{i}^{N}\sum_{j}^{N}\min\left(T_{ij},T_{ij}^{'}\right)}{\sum_{i}^{N}\sum_{j}^{N}T_{ij}+\sum_{i}^{N}\sum_{j}^{N}T_{ij}^{'}}, \tag{7}$$

where T_{ij} and $T_{ij}^{'}$ are the actual and predicted trip flows, respectively, from location *i* to location *j*. The value of SSI is

Table 1 Two-sampled KS test results for trip distribution based on distance probability distribution

KS test	Gravity (LBSN)	Gravity	PWO (LBSN)	PWO	Radiation (LBSN)	Radiation	IO (LBSN)	IO
P value	0.210	0.078	0.362	0.947	0.0442	0.0240	0.210	0.078
H_0	Accept	Accept	Accept	Accept	Reject	Reject	Accept	Accept

Table 2 Two-sampled KS test results for trip distribution based on destination population probability distribution

KS test	Gravity (LBSN)	Gravity	PWO (LBSN)	PWO	Radiation (LBSN)	Radiation	IO (LBSN)	IO
P value	0.709	0.945	0.945	0.709	0.945	0.945	0.709	0.945
H_0	accept	accept	accept	accept	accept	accept	accept	accept

between zero and one, with zero indicating complete disagreement and one indicating equality. Figure 7 shows the SSI values of the models for Manhattan.

As Fig. 7 shows, regardless of the data used, the radiation model has the lowest index value, indicating poorer agreement with the real data. This is consistent with the results of previous studies on the subject, suggesting that the radiation model has limited capabilities for predicting human mobility in intra-urban scenarios (Liang et al. 2013; Masucci et al. 2013), as is the case in our study. The SSI value for the PWO is slightly worse than that for gravity and IO models. However, unlike the gravity and IO models, the PWO model requires no parameters to be determined. Nevertheless, it should also be noted that, for planners, geographers, economists, and many others, the parameters provide context and have explanatory power. Results from the LBSN data are more similar to the real data (except for the radiation model), on average, by about 20%.

The scatterplot of each model is shown in Fig. 8. These plots have a log-log scale, so that more details can be seen when the values are within a broad range. The blue dots in each diagram indicate the number of modeled trips against the number of observed trips for all origin-destination pairs. The red line passing through the clouds of blue points is the identity line ($y = x$) and indicates the equality of predicted and observed trips. As can be seen, the point clouds obtained from the LBSN data tend toward the identity line, showing good agreement of the results obtained from LBSN data with real observations, whereas the upper point clouds are more diffused over the plot area.

Further, we studied the performances of models based on the *R-squared* measure obtained from the regression analysis of each plot. Figure 9 demonstrates significant differences in the two data sets used. Again, LBSN data performed much better than the population data. Note that the relative differences between the bars in Figs. 7 and 9 in the case of LBSN data, are preserved, showing the stability of the models when using LBSN data.

In order to analyze the results in a more detailed manner, we computed the cosine similarities between origin-destination matrices at zone level, rather than at the level of the whole city. Firstly, the rows and columns of each matrix were partitioned. Then, the cosine similarities between corresponding rows and columns in each matrix were computed. To compute cosine similarities, each row (column) is considered asa vector in a 288-dimensional space (i.e., the dimension of the space is equal to the number of zones). If the angle between this vector and the corresponding vector extracted from the ground truth matrix in the space is equal to zero, there is complete similarity (identity). Conversely, if the two vectors are in opposite directions, the value of the index will be −1. Since the trip distribution matrix is a non-negative matrix, the index ranges in practice from zero to one, indicating parallel and perpendicular vectors,

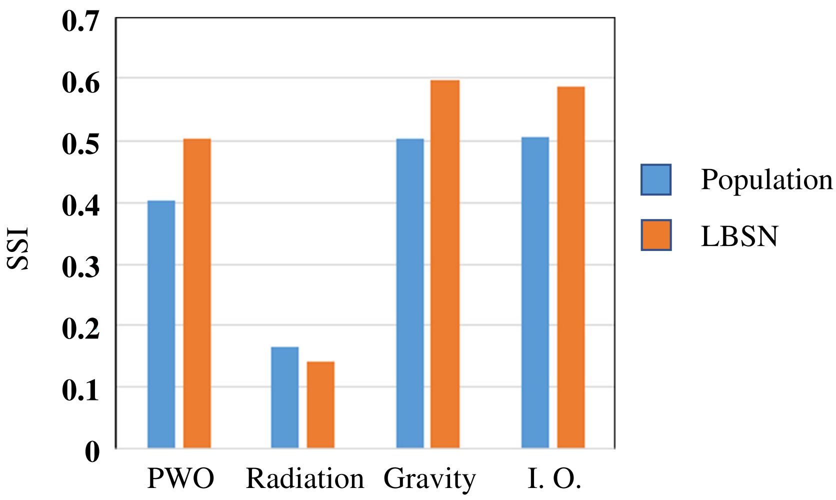


Fig. 7 Comparison of performances of models based on SSI. SSI is an index to quantify the similarity between two data sets

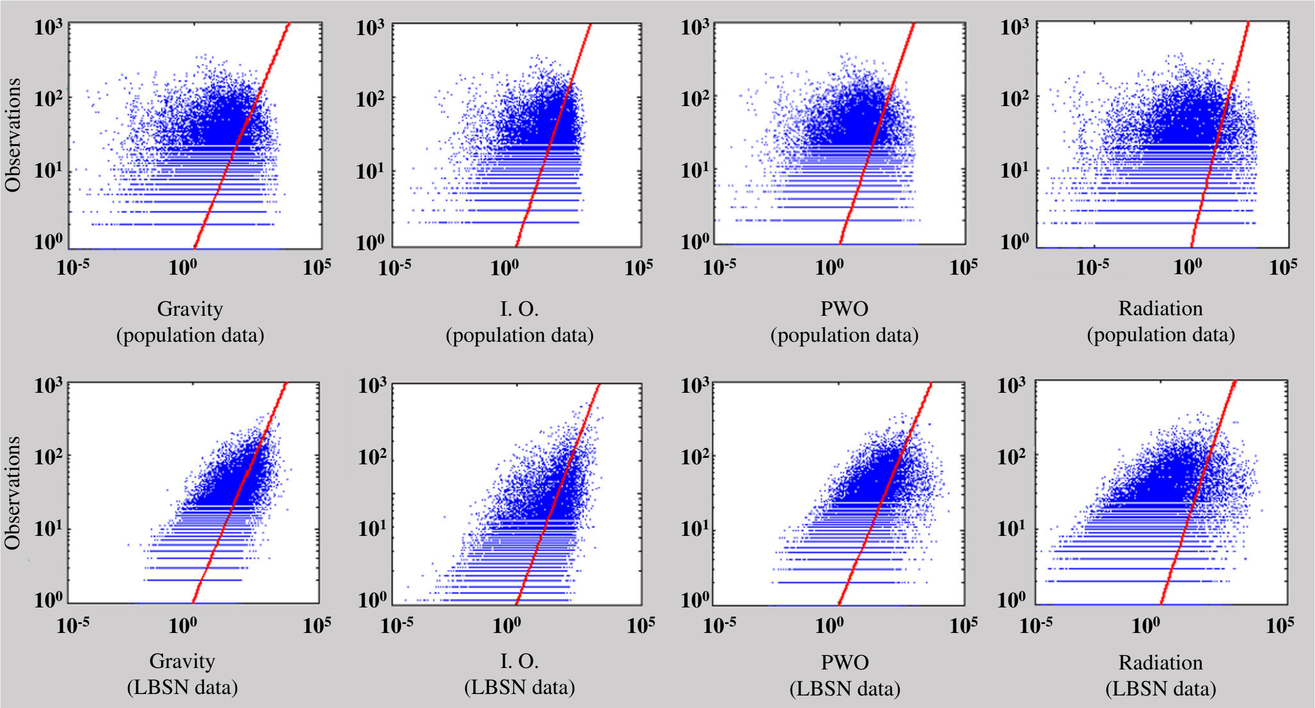


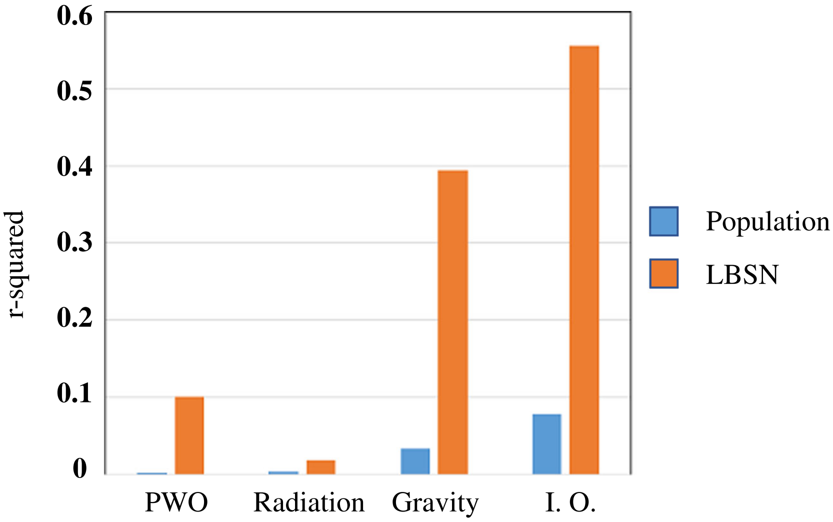
Fig. 8 Comparing the observed fluxes with predicted fluxes. The red line is the identity line

respectively. Figures 10, 11, 12, and 13 show the frequency histograms of the cosine similarities for rows and columns.

The red line indicates the mean value of the histogram (μ) and the blue bounds show the interval $\mu-\sqrt{2}\sigma$ to $\mu+\sqrt{2}\sigma$, where σ is the standard deviation. According to Chebyshev's inequality, at least 50% of values lie within the blue area. The histograms show an overall improvement in the predictions, except for the case of the radiation model.

In order to see to what extent particular types of check-ins are incorporated in mobility modeling in Manhattan, we aggregated check-ins occurring at similar locations into seven categories, i.e., eating out, shopping, religious affairs, recreational activities, educational and academic activities, job-related activities, and other activities. The plot below (Fig. 14) shows the contribution of each category to the mobility modeling in Manhattan. Because check-in numbers play the role of coefficient in the models, they directly affect the results. As shown in the figure, a significant proportion of the check-ins in our data set relates to shopping and eating out.

Fig. 9 Comparison of performances of models based on the *R*-squared measure resulting from regression analysis



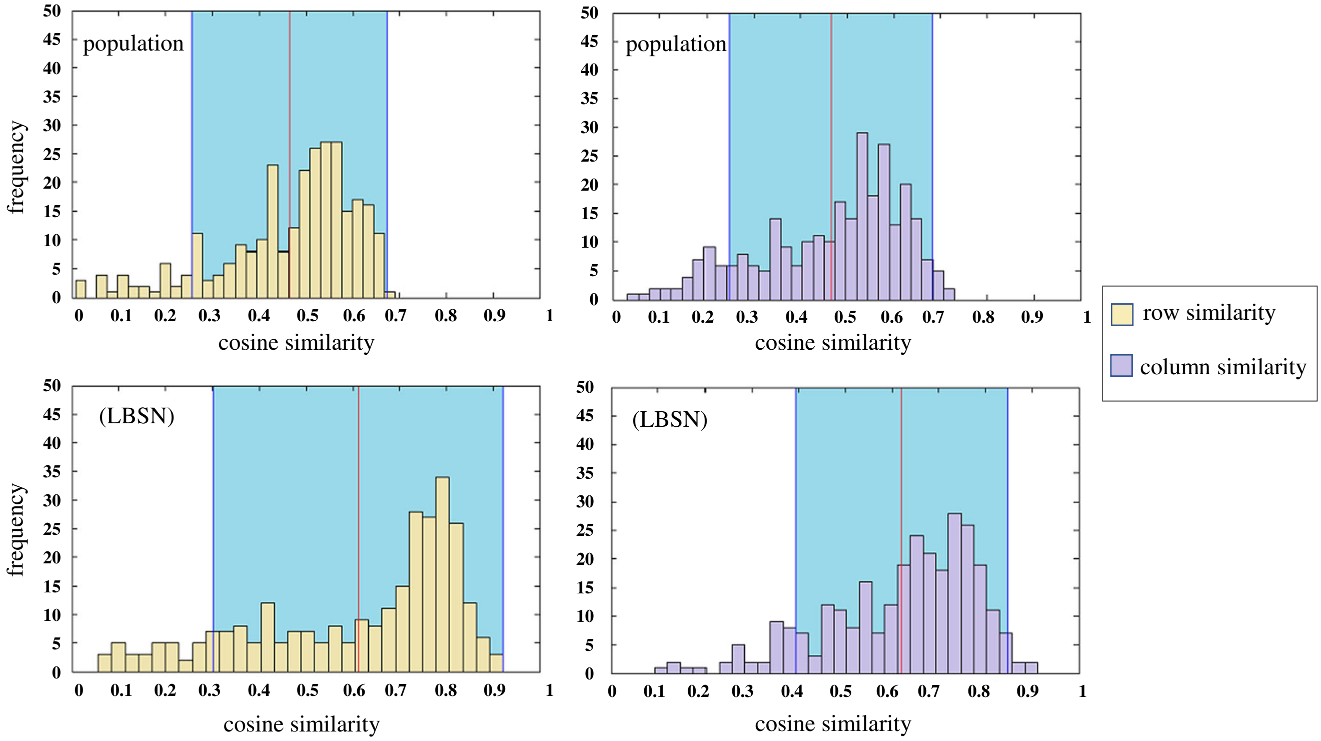


Fig. 10 Frequency histograms of cosine similarities for **a** rows and **b** columns of OD matrices from the gravity model

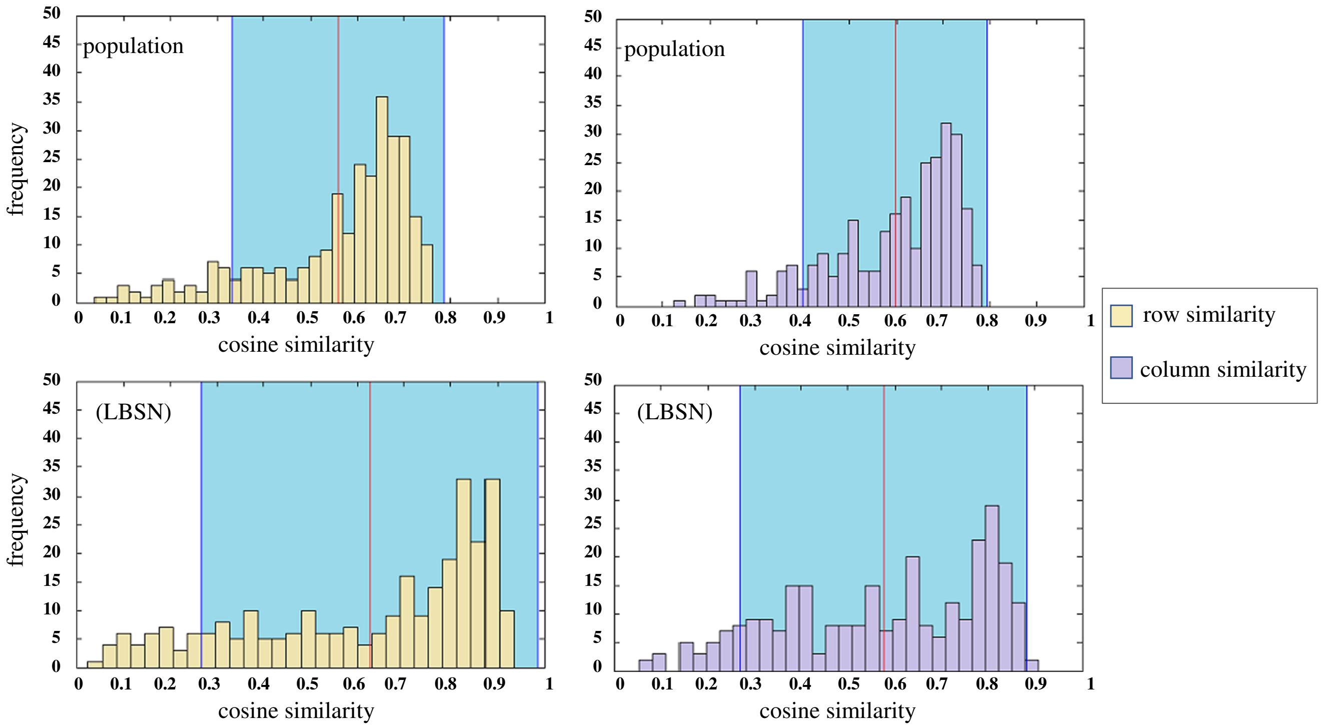


Fig. 11 Frequency histograms of cosine similarities for **a** rows and **b** columns of OD matrices from the IO model

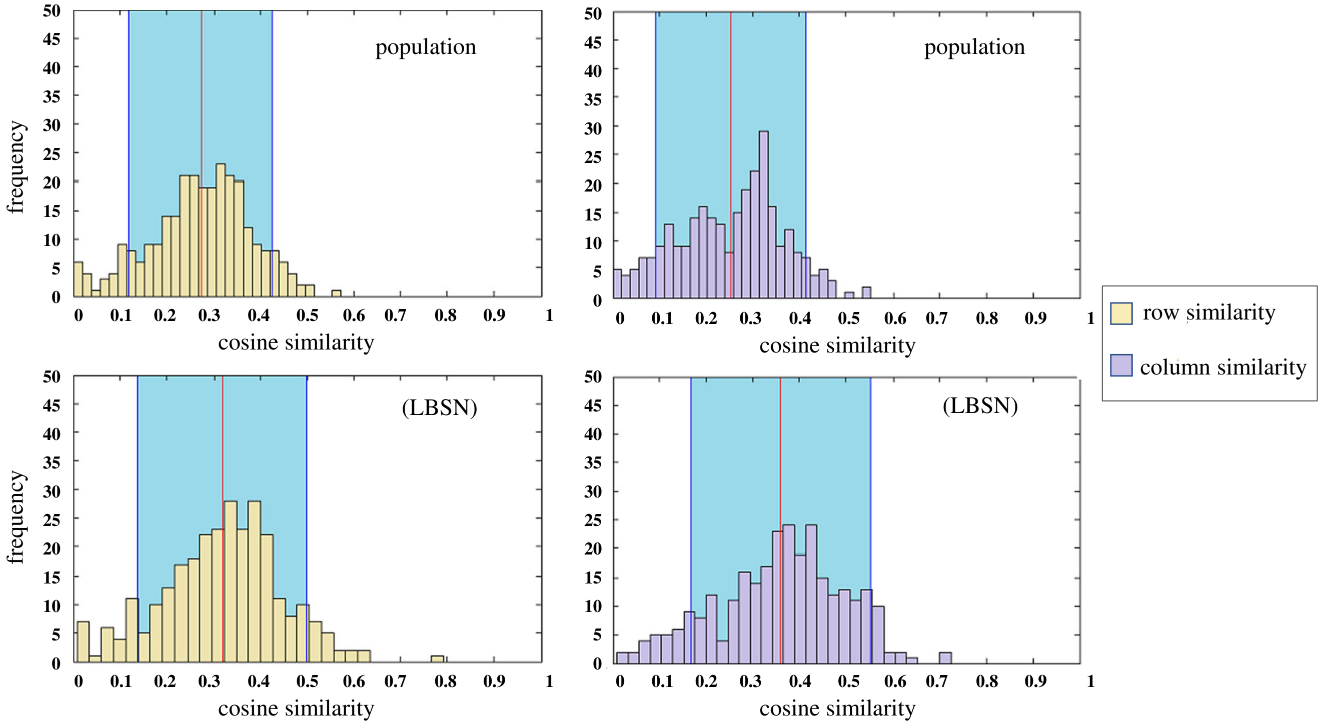


Fig. 12 Frequency histograms of cosine similarities for **a** rows and **b** columns of OD matrices from the PWO model

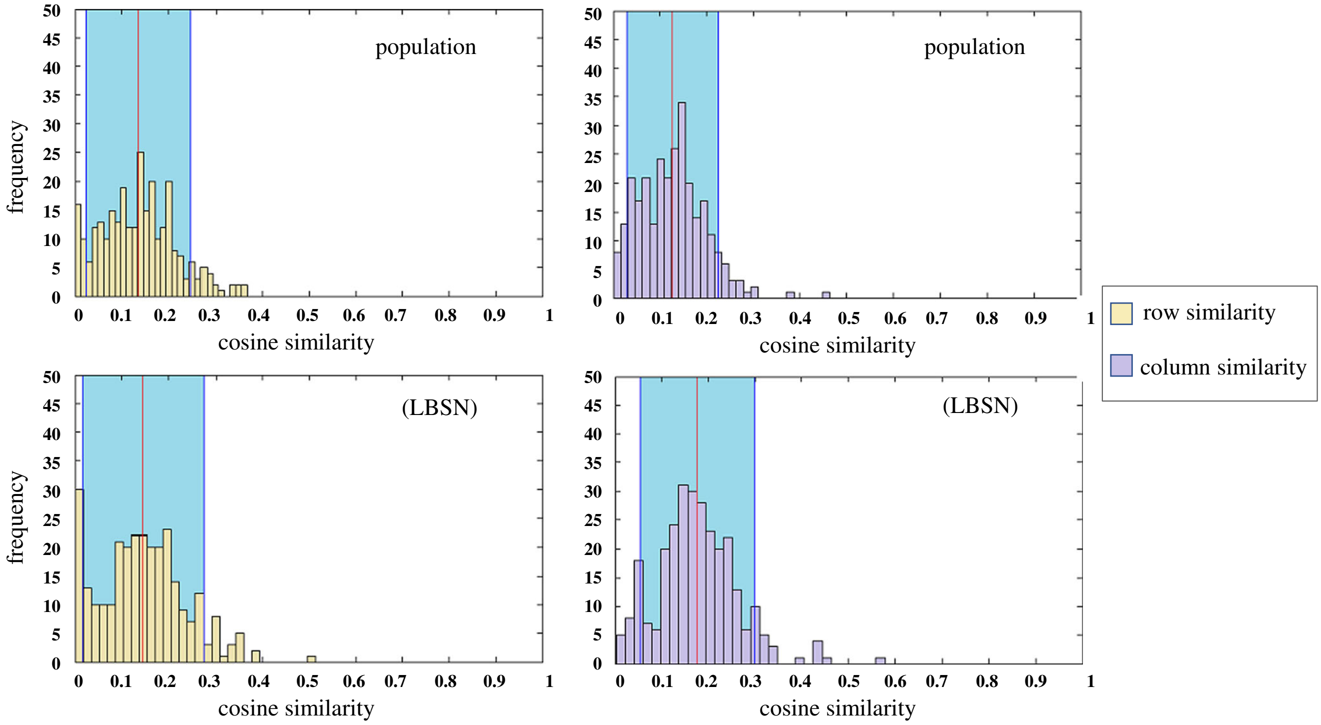


Fig. 13 Frequency histograms of cosine similarities for **a** rows and **b** columns of OD matrices from the radiation model

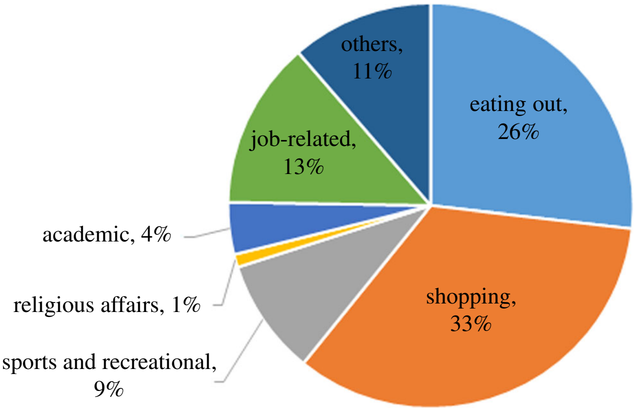


Fig. 14 The contribution of each check-in category to mobility modeling in Manhattan

Conclusions

In this article, we used LBSN data to predict human mobility patterns in Manhattan, NYC. Different boroughs have many interactions with each other, and people do not generally live and work in the same area. Thus, there are offsets in terms of population and activities. This can result in reduced mobility prediction accuracies. Since the LBSN data are inherently more directly related to trips than population data, we used data from check-ins as proxy variables to predict the human mobility within Manhattan. In this paper, we explored the predictive potential of the existing human mobility models by replacing the population variable by POIs and check-ins. In this way, we changed the possibly unrealistic assumptions about population within cities, while preserving the simplicity of the models. Our proposed assumptions resulted in improved performance. Results from evaluation measures revealed that all models using the proposed assumptions achieved overall accuracies much better than when the original assumptions were used. LBSN data led to patterns that were, on average, 20% more similar to the real observations based on SSI. Moreover, the accuracy of predictions was enhanced significantly according to the *R*-squared measure obtained from regression analysis. Future work could evaluate the applicability of such an approach on different spatial scales, such as for inter-city mobility. Utilizing the LBSN data can lead to more accurate predictions of human mobility within cities. This study adopts some basic assumptions. For instance, the comparison of the results of human mobility prediction models against taxicab journeys is common in the literature. The assumption is that taxi trips are representative of people's movements within the city. However, this may not be true. It is also assumed that the reported check-ins in the data set are genuine. The activities in LBSNs might not be representative of all types of activities in the real world. These considerations may have an influence on the results of our analyses. However, LBSN data have the potential to be used successfully as proxy variables in the models, instead of more static variables such as population.

References

Abbasi OR, Alesheikh A, Sharif M (2017) Ranking the City: the role of location-based social media check-ins in collective human mobility prediction. ISPRS International Journal of Geo-Information 6:136

Agarwal C, Green GM, Grove JM, Evans TP, Schweik CM (2002) A review and assessment of land-use change models: dynamics of space, time, and human choice

Agryzkov T, Martí P, Tortosa L, Vicent JF (2017) Measuring urban activities using Foursquare data and network analysis: a case study of Murcia (Spain). Int J Geogr Inf Sci 31:100–121

Bagrow JP, Wang D, Barabasi A-L (2011) Collective response of human populations to large-scale emergencies. PLoS One 6:e17680

Barceló J, Montero L, Marqués L, Carmona C (2010) Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. Transp Res Rec: J Transp Res Board 2175:19–27

Brockmann D, David V, Gallardo AM (2009) Human mobility and spatial disease dynamics. In: Reviews of nonlinear dynamics and complexity. Wiley-VCH Verlag GmbH & Co. KGaA 2:1–24

Buhalis D, Amaranggana A (2013) Smart tourism destinations. In: Information and communication Technologies in Tourism 2014. Springer, pp 553–564

Caceres N, Wideberg J, Benitez F (2007) Deriving origin destination data from a mobile phone network. Intel Transp Syst, IET 1:15–26

Camagni R, Gibelli MC, Rigamonti P (2002) Urban mobility and urban form: the social and environmental costs of different patterns of urban expansion. Ecol Econ 40:199–216

Celik HM (2010) Sample size needed for calibrating trip distribution and behavior of the gravity model. J Transp Geogr 18:183–190

Easa SM (1993) Urban trip distribution in practice. I: conventional analysis. J Transp Eng 119:793–815

Evans AW (1971) The calibration of trip distribution models with exponential or similar cost functions. Transp Res 5:15–38

Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453:779–782

Hasan S, Zhan X, Ukkusuri SV Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, 2013. ACM, p 6

Hristova D, Liben-Nowell D, Noulas A, Mascolo C If You've got the money, I've got the time: spatio-temporal footprints of spending at sports events on Foursquare. In: Tenth international AAAI conference on web and social media, 2016

Hyman G (1969) The calibration of trip distribution models. Environ Plan 1:105–112

Jiang B, Yin J, Zhao S (2009) Characterizing the human mobility pattern in a large street network. Phys Rev E 80:021136

Kang C, Liu Y, Guo D, Qin K (2015) A generalized radiation model for human mobility: spatial scale, searching direction and trip constraint. PLoS One 10:e0143500

Kang C, Ma X, Tong D, Liu Y (2012) Intra-urban human mobility patterns: an urban morphology perspective. Physica A: Stat Mech Appl 391:1702–1717

Lenormand M, Huet S, Gargiulo F, Deffuant G (2012) A universal model of commuting networks. PLoS One 7:e45985

Li M, Westerholt R, Fan H, Zipf A (2016) Assessing spatiotemporal predictability of LBSN: a case study of three Foursquare datasets. GeoInformatica:1–21. https://doi.org/10.1007/s10707-016-0279-5

Liang X, Zhao J, Dong L, Xu K (2013) Unraveling the origin of exponential law in intra-urban human mobility. Sci Rep 3

Liu Y, Sui Z, Kang C, Gao Y (2014) Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PLoS One 9:e86026

Masucci AP, Serras J, Johansson A, Batty M (2013) Gravity versus radiation models: on the importance of scale and heterogeneity in commuting flows. Phys Rev E 88:022812

Matyas L (1998) The gravity model: some econometric considerations. World Econ 21:397–401

Noë N, Whitaker RM, Chorley MJ, Pollet TV (2016) Birds of a feather locate together? Foursquare checkins and personality homophily. Comput Hum Behav 58:343–353

Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PLoS One 7:e37027

Openshaw S (1976) An empirical study of some spatial interaction models. Environ Plan A 8:23–41

Peng C, Jin X, Wong K-C, Shi M, Liò P (2012) Collective human mobility pattern from taxi trips in urban area. PLoS One 7:e34487

Prothero RM (1977) Disease and mobility: a neglected factor in epidemiology. Int J Epidemiol 6:259–267

Simini F, González MC, Maritan A, Barabási A-L (2012) A universal model for mobility and migration patterns. Nature 484:96–100

Stouffer SA (1940) Intervening opportunities: a theory relating mobility and distance. Am Sociol Rev 5:845–867

Wang P, González MC, Hidalgo CA, Barabási A-L (2009) Understanding the spreading patterns of mobile phone viruses. Science 324:1071–1076

Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO (2012) Quantifying the impact of human mobility on malaria. Science 338:267–270

Williams I (1976) A comparison of some calibration techniques for doubly constrained models with an exponential cost function. Transp Res 10:91–104

Yan X-Y, Zhao C, Fan Y, Di Z, Wang W-X (2014) Universal predictability of mobility patterns in cities. J R Soc Interface 11:20140834

Yang D, Zhang D, Chen L, Qu B (2015) NationTelescope: monitoring and visualizing large-scale collective behavior in LBSNs. J Netw Comput Appl 55:170–180

Zheng Z, Rasouli S, Timmermans H (2015) Two-regime pattern in human mobility: evidence from GPS taxi trajectory data. Geographical Analysis 48:157–175

Zhou X, Mahmassani HS (2006) Dynamic origin-destination demand estimation using automatic vehicle identification data. IEEE Trans Intell Transp Syst 7:105–114

Zipf GK (1946) The P 1 P 2/D hypothesis: on the intercity movement of persons. Am Sociol Rev 11:677–686