Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Indigenization of urban mobility

Zimo Yang [a,b], Defu Lian [b,c,d], Nicholas Jing Yuan [b], Xing Xie [b], Yong Rui [b], Tao Zhou [a,b,d,*]

[a] *CompleX Lab, Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China*

[b] *Microsoft Research, Beijing 100080, People's Republic of China*

[c] *Department of Computer Science and Technology, University of Science and Technology of China, Hefei, 230026, People's Republic of China*

[d] *Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China*

HIGHLIGHTS

- Identification of urban mobility pattern is very important for predicting and controlling spatial events.
- Native and non-native people show distinct mobility pattern in large cities.
- Behavioral analysis can largely improve the accuracy of location prediction.

ARTICLE INFO

Article history:
Received 3 June 2016
Received in revised form 12 October 2016
Available online 18 November 2016

Keywords:
Human dynamics
Urban mobility
Location prediction
Indigenization
Behavioral analysis
Demographical analysis

ABSTRACT

The identification of urban mobility patterns is very important for predicting and controlling spatial events. In this study, we analyzed millions of geographical check-ins crawled from a leading Chinese location-based social networking service (Jiepang.com), which contains demographic information that facilitates group-specific studies. We determined the distinct mobility patterns of natives and non-natives in all five large cities that we considered. We used a mixed method to assign different algorithms to natives and non-natives, which greatly improved the accuracy of location prediction compared with the basic algorithms. We also propose so-called indigenization coefficients to quantify the extent to which an individual behaves like a native, which depends only on their check-in behavior, rather than requiring demographic information. Surprisingly, the hybrid algorithm weighted using the indigenization coefficients outperformed a mixed algorithm that used additional demographic information, suggesting the advantage of behavioral data in characterizing individual mobility compared with the demographic information. The present location prediction algorithms can find applications in urban planning, traffic forecasting, mobile recommendation, and so on.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Understanding urban mobility of human is very important for disease control [1], city planning [2], and traffic forecasting [3], as well as for increasing business value in location-based services [4,5]. Individuals differ greatly in their mobility patterns [6], but aggregated analyses detect regular displacement distributions, which range from power laws

[*] Correspondence to: CompleX Lab, Web Sciences Center, University of Electronic Science and Technology of China, Xiyuan Street 2006, Chengdu 611731, People's Republic of China.
E-mail address: zhutou@ustc.edu (T. Zhou).

http://dx.doi.org/10.1016/j.physa.2016.11.101
0378-4371/© 2016 Elsevier B.V. All rights reserved.

[7,8] to exponential laws [9,10]. These statistical regularities may be the result of combining several parameters, including the preferential return mechanism embedded in individual behaviors [11], the structure of transportation systems [12,13], urban organization [14] and the constraints of travel costs [6,15].

In general, an in-depth understanding can be obtained by classifying people into groups according to their demographic features. For some specific measures, such as predictability, group-based differences are statistically insignificant [16], but recent empirical studies have shown that people with different jobs [6], different ages [17], different genders [17–19] and different purposes (tour or not) [20] have different mobility patterns. Thus, group-based analyses are helpful when addressing specific social problems, e.g., gender-specific mobility patterns can be used to quantify the equality between men and women [20,21], the leisure mobility of elderly people can be considered as an important indicator that characterizes their living conditions [22], and the urban mobility of children may affect their future integration into society [23].

These group-based analyses can be improved in two ways. First, the demographic information contains noise and bias, where the former is derived from false or out-of-date data whereas the latter is a result of inconsistencies between demographic features and behavioral patterns; thus, at the individual level, a girl could behave like a boy and an aged person could behave like a young man. Methodologically, we can uncover the quantitative differences in behavioral patterns among people in different demographic groups, and then consider the behavioral differences directly instead of demographic differences. This shift from demographic analyses to behavioral analyses can extend the potential applications to datasets without demographic information as well as facilitating an associated shift from group-based services to personalized services. Second, group-specific insights can be used to address relevant problems such as location prediction and travel recommendation, where the algorithmic performance verifies the significance of the analyses.

In this study, we investigated the different mobility patterns of natives and non-natives, which are increasingly important in terms of business value in location-based services [24] and social value when measuring the social integration of immigrants during globalization and urbanization processes [25,26]. Our study involved the intensive analysis of a large-scale dataset, which included millions of geographical check-ins in five large cities in China. The study makes three main contributions, as follows. (i) We identified the distinct mobility patterns of natives and non-natives, i.e., the distribution of the location visiting frequencies of non-natives was more heterogeneous than that of natives at the aggregated level, whereas the frequency distributions of natives were statistically more heterogeneous at the individual level. (ii) Compared with the basic algorithms, the accuracy of location prediction was improved greatly by assigning different algorithms to natives and non-natives. (iii) We developed behavior-based indices to characterize how an individual behaves like a native, i.e., indigenization coefficients, and we showed that a hybrid algorithm weighted using indigenization coefficients, which did not require any demographic information, improved the prediction accuracy greatly.

Our study is closely relevant to the real world, in particular, the prediction of locations can find wide applications. Years ago, it is already known as a critical part in traffic forecasting [3] and urban planning [27]. Very recently, thanks to the development of information and communication technologies such as mobile communication, the mobility prediction plays an increasingly important role in location-based recommendation [4,5,28–30]. In addition, the prediction methods can be improved by considering social ties between target users while such methods can also be applied in analyzing social networks [31–35].

2. Data

Our experiments were based on check-in records crawled from *Jiepang* (http://jiepang.com/), which is a leading Chinese location-based social networking service that is similar to *Foursquare* (https://foursquare.com/). It helps users to record and track all of their life activities, to connect with friends at specific moments, and to explore communities of people with similar interests. Tweets attaching the specific webpage URLs of *Jiepang* check-in locations are often shared on the *weibo* platform (http://weibo.com/). We crawled these tweets via weibo public APIs and extracted location check-in information from them, including locations, time-stamps and optional texts. In addition, we crawled users' hometown information from their Weibo profile pages. Notice that, all data were obtained via the open APIs. Therefore the use of these data has obeyed the corresponding terms and conditions. The dataset covers the activities of users from 5 August 2011 to 17 September 2012. The detailed techniques in crawling and cleaning the data can be found in Refs. [36,37]. The data were anonymized before this study, where both user identities and location entities are replaced by 128-bit MD5 numbers, and the full dataset is available from the link (we will make it available to the public after acceptance).

In the dataset, each item (i.e., a check-in) recorded the user ID (anonymized), check-in time, longitude, latitude, and name of a location. The latitude, longitude and name together comprise a specific and unique location, and thus determine its representation MD5. Those mentioned frequencies are computed based on the representation MD5s of locations. Given a city X, all of the users were divided into three classes: (A) users whose hometown and most frequently visited city was both X; (B) users whose hometowns were not X and their most frequently visited cities were not X; (C) users without hometown information or who did not belong to (A) or (B). The users in classes A and B were called natives and non-natives of city X, respectively, but the users in (C) were not considered in this study. We used these classifications to minimize the bias caused by users who might live and work for a long time in a city that differed from their hometown.

We targeted five big cities in China: Beijing, Shanghai, Nanjing, Chengdu, and Hong Kong. Users who checked in only once and locations that appeared only once were removed. After filtering, the data comprised 1,371,294 check-ins. The basic statistics for the dataset are shown in Table 1. It should be noted that user records could appear in several places

Table 1
Basic statistics for the data used in this study. For each of the five cities, N^u, N^u_y, N^u_n, N^l, N^l_y, N^l_n, N^c, N^c_y, and N^c_n represent the number of users, native users, non-native users, visited locations, locations visited by natives, locations visited by non-natives, check-ins, native check-ins, and non-native check-ins, respectively.

City	N^u	N^u_y	N^u_n	N^l	N^l_y	N^l_n	N^c	N^c_y	N^c_n
Beijing	11 077	6824	4253	28 100	26 882	8617	511 133	384 222	126 911
Shanghai	6 322	3847	2475	45 070	44 413	6594	782 677	734 719	48 958
Nanjing	2 132	177	1955	2 421	1 539	1747	18 757	7 934	10 823
Chengdu	2 320	173	2147	2 330	1 518	1542	21 952	7 976	13 976
Hong Kong	1 727	130	2597	2 460	1 036	2079	36 775	5 124	31 651

Table 2
Gini coefficients for the distributions of visiting frequencies at different locations. G^p_y and G^p_n denote the coefficients contributed by the native population and non-native population, respectively, while $\langle G^i_y \rangle$ and $\langle G^i_n \rangle$ denote the average coefficients for native individuals and non-native individuals.

City	G^p_y	G^p_n	$\langle G^i_y \rangle$	$\langle G^i_n \rangle$
Beijing	0.68	0.77	0.21	0.14
Shanghai	0.67	0.73	0.26	0.13
Nanjing	0.43	0.68	0.24	0.15
Chengdu	0.43	0.77	0.24	0.10
Hong Kong	0.46	0.82	0.21	0.16

because they could be a native of Beijing but also a non-native of Chengdu and Nanjing, whereas a check-in could only occur in one place.

3. Empirical analysis

As shown in Fig. 1, for each city, the distribution of the location visiting frequencies for non-natives is more heterogeneous than that for natives. This is because the non-natives tended to visit popular locations, such as the Imperial Palace in Beijing and the Bund in Shanghai. In contrast, natives usually check in repeatedly at locations of personal importance, and these locations are different for different natives. Therefore, the distribution of visiting frequencies contributed by all natives is relatively homogeneous.

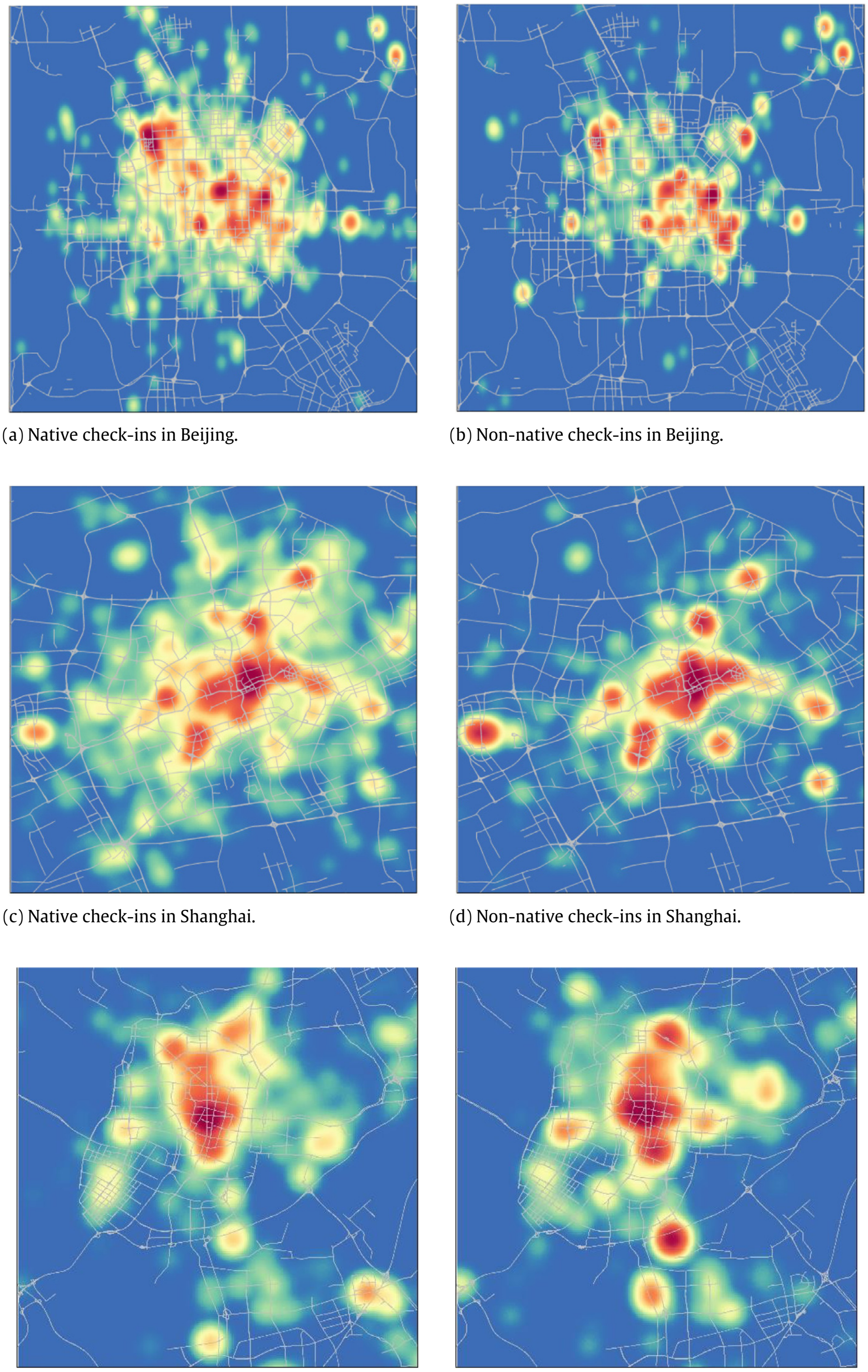
We used the well-known Gini coefficient [38] to quantify the heterogeneity of a visiting frequency distribution. The Gini coefficient measures the inequality among the values in a frequency distribution, which can theoretically range from 0 to 1, where 0 and 1 correspond to complete equality and inequality, respectively. A higher Gini coefficient indicates a more heterogeneous distribution, and vice versa.

The Gini coefficient is usually defined mathematically based on the Lorenz curve, which is used in characterizing the heterogeneity of a distribution. For example, in this paper, when drawing a Lorenz curve for the distribution of locations' visits (which can be contributed by all visitors, only native visitors, or only non-native visitors), we firstly rank all locations according to their visiting frequencies. As shown in Fig. 2, the x-axis stands for the fraction of locations where all locations are arranged from the less visited ones (left) to frequently visited ones (right), and the y-axis represents the fraction of cumulated number of visits on those locations. If every location has been visited exactly the same times, the Lorenz curve is the diagonal line as the red dash line in Fig. 2. The blue line is a usual example of Lorenz curve in real heterogeneous systems. For example, a point (0.4, 0.05) in the blue curve means the 40% of the least visited locations contribute only 5% of the total visits. Obviously, for a highly heterogeneous distribution, the area S under the Lorenz curve is very small, therefore we adopt the simple coefficient $G = 1 - 2S$ to quantify the extent of heterogeneity, which is the well-known Gini coefficient. If there are n locations with visiting frequencies of $f_1 \leq f_2 \leq \cdots \leq f_n$, the Gini coefficient is:

$$G = \frac{1}{n}\left\{ n + 1 - 2\left[\frac{\sum\limits_{i=1}^{n}(n+1-i)f_i}{\sum\limits_{i=1}^{n} f_i}\right]\right\} = \frac{2\sum\limits_{i=1}^{n} i f_i}{n\sum\limits_{i=1}^{n} f_i} - \frac{n+1}{n}. \tag{1}$$

As shown in Table 2, for every city that we considered, at the aggregated level, the Gini coefficient contributed by non-natives was higher than the Gini coefficient contributed by natives (i.e., $G^p_n > G^p_y$), which is in accordance with Fig. 1. The corresponding Lorenz curves are presented in Fig. 3.

At the individual level, a native individual tends to check in repeatedly at locations of personal importance, while a non-native individual does not stay long in an area and thus they did not check in multiple times at specific locations. Therefore, the visiting frequency distribution of a native person was significantly more heterogeneous than that of a non-native person. As shown in Table 2, this hypothesis was supported because the average Gini coefficient for all native individuals was higher than that for all non-natives, i.e., $\langle G^i_y \rangle > \langle G^i_n \rangle$, for every city under consideration. The significance of the two results,



(a) Native check-ins in Beijing.

(b) Non-native check-ins in Beijing.

(c) Native check-ins in Shanghai.

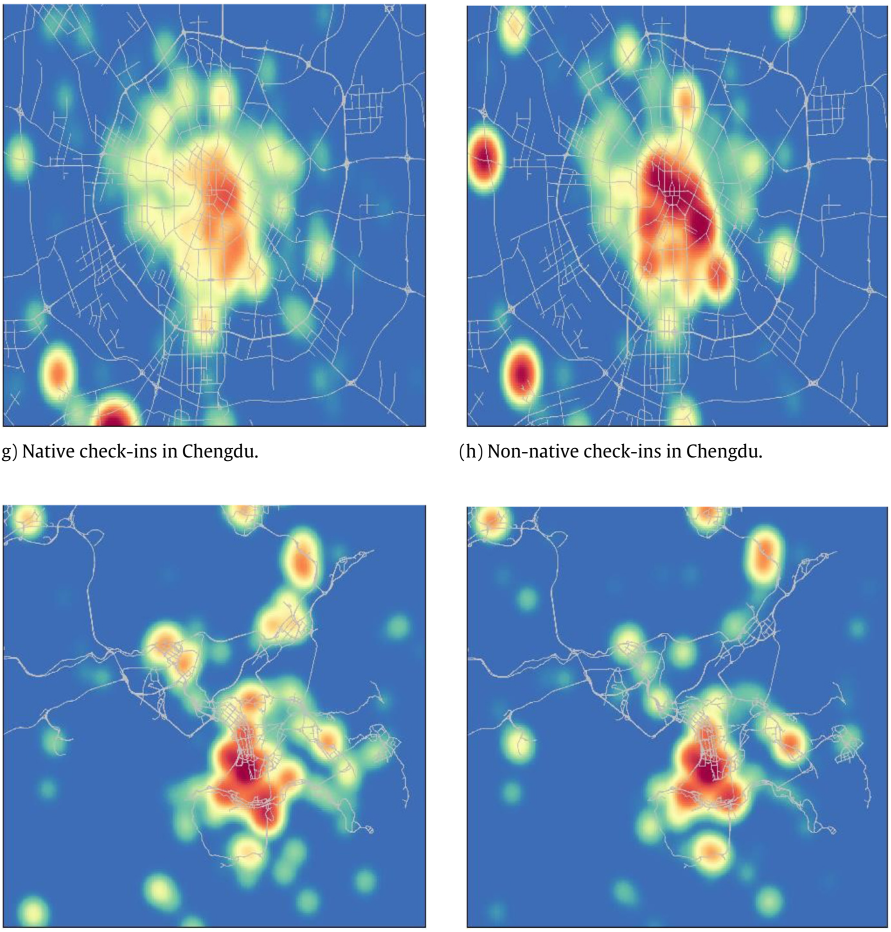
(d) Non-native check-ins in Shanghai.

(e) Native check-ins in Nanjing.

(f) Non-native check-ins in Nanjing.

Fig. 1. Density maps of location visiting frequencies for: (a) natives in Beijing, (b) non-natives in Beijing, (c) natives in Shanghai, (d) non-natives in Shanghai, (e) natives in Nanjing, (f) non-natives in Nanjing, (g) natives in Chengdu, (h) non-natives in Chengdu, (i) natives in Hong Kong, and (j) non-natives in Hong Kong. The frequencies were normalized for each plot and the areas colored in red have high densities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$G_n^p > G_y^p$ and $\langle G_y^i \rangle > \langle G_n^i \rangle$, has been validated by well-accepted statistical tests including t-statistics, degree of freedom and p-value [39].



(g) Native check-ins in Chengdu.

(h) Non-native check-ins in Chengdu.

(i) Native check-ins in Hong Kong.

(j) Non-native check-ins in Hong Kong.

Fig. 1. (*continued*)

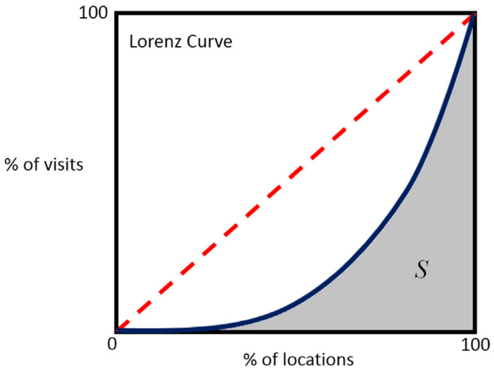


Fig. 2. An illustration of the Lorenz curve. The red dash line stands for the ideal case that every location has been visited exactly the same times, and the blue curve is an example Lorenz curve, with the shadowed area being S.

Natives and non-natives are also statistically distinguishable in other aspects. As shown in Fig. 4(a)–(d), the native individuals statistically have visited more distinct locations (i.e., larger N_D in average) and shared more check-ins (i.e., larger N_T in average). As shown in Fig. 4(e) and (f), the natives and non-natives have remarkable differences in their active periods, where an individual's active period is simply quantified by the number of days between his first and last check-ins. Firstly, a native user's active period is much longer than a non-native user, with a significant peak around 270 days for all five cities.

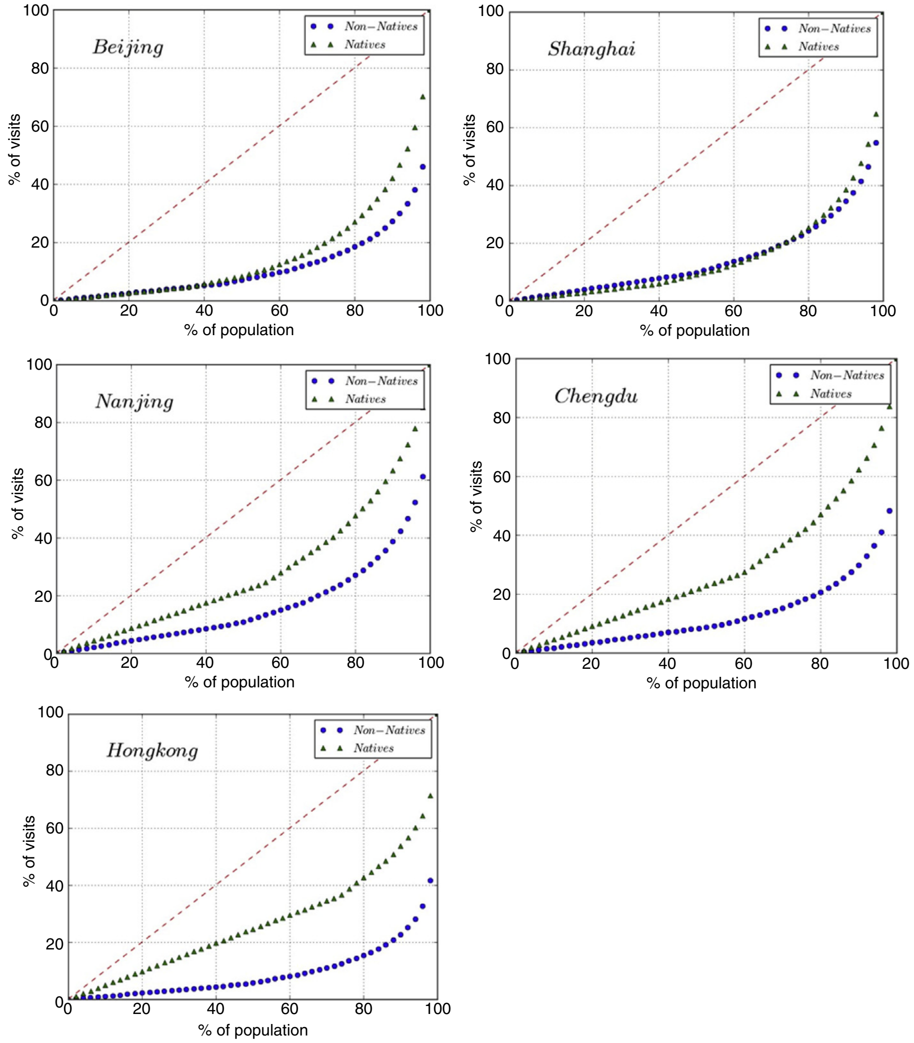


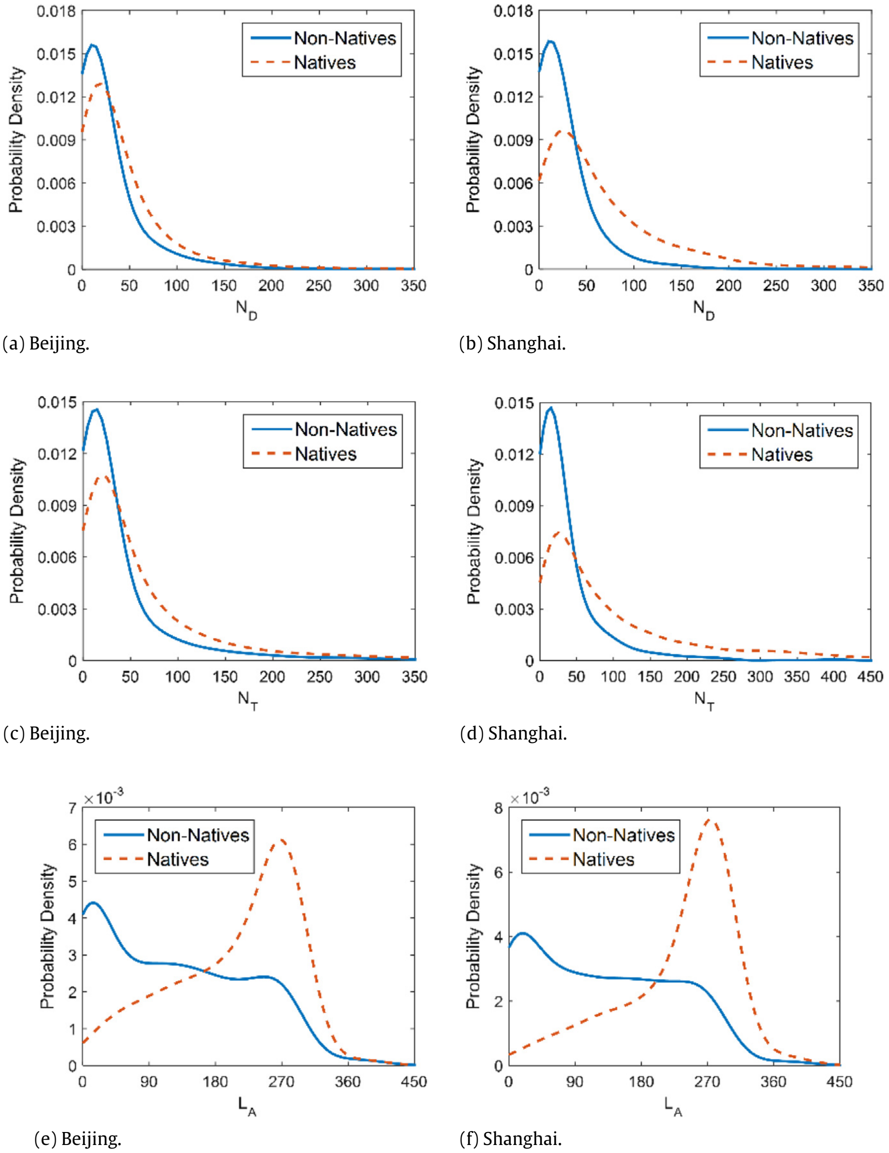
Fig. 3. Lorenz curves at the population level for Beijing, Shanghai, Nanjing, Chengdu and Hong Kong. The triangles and circles represent the results for natives and non-natives, respectively. In each city, the distribution of visited frequencies of locations that contributed by all non-natives is more heterogeneous than that contributed by all natives.

Secondly, a small but notable peak at about a week is observed in each distribution for non-native users, which is probably contributed by the tourists.

4. Location prediction algorithms

Location prediction is a core technique that underlies many significant location-based services [23] and other application [40]. Thus, many algorithms have been proposed to obtain highly accurate predictions [41], including collaborative analysis, Markov chain methods, linear regression, decision trees (e.g., M5 tree and T-pattern tree), neural networks, Bayesian networks, and other data mining approaches [4,5,42–49]. In our analysis, we used the two simplest methods for location prediction. The first is called the history-based method (HB), where given a target user, each location is scored directly as the number of check-ins it receives in the target user's records. The second is called the popularity-based method (PB), where the prediction score for a location is the total number of check-ins it receives in all of the users' records, regardless of how many times it appears in the target user's check-in list. A location with a higher score is assumed to be more likely to appear in the target user's future trajectory.

The check-ins were divided into two datasets: the training set contained 90% of the check-ins and the testing set comprised the remaining 10%. Such division can be in two ways: *random division* with the check-ins in testing set being selected randomly and *temporal division* where the testing set contains the 10% newest check-in records. Every record in the testing set satisfied two conditions: (i) the user had at least two check-ins in the target city; (ii) the location was associated with at least two check-ins. Given a target user, a location prediction algorithm produces a ranked list of all



(a) Beijing.

(b) Shanghai.

(c) Beijing.

(d) Shanghai.

(e) Beijing.

(f) Shanghai.

Fig. 4. Statistical differences between natives and non-natives. The red dash and blue solid curves stand for probability density functions for natives and non-natives, respectively. The six subplots respectively report the distributions of the number of distinct visited locations N_D in (a) Beijing and (b) Shanghai, the number of total check-ins N_T in (c) Beijing and (d) Shanghai, the length of active period L_A, namely the number of days between a user's first and last check-ins in (e) Beijing and (f) Shanghai. The average values $\langle N_D \rangle$ for natives in Beijing, non-natives in Beijing, natives in Shanghai, and non-natives in Shanghai are 40.94, 24.14, 69.19 and 21.82, respectively. The average values $\langle N_T \rangle$ for natives in Beijing, non-natives in Beijing, natives in Shanghai, and non-natives in Shanghai are 94.33, 42.35, 206.79 and 39.95, respectively. These distributions for the other three cities (i.e., Nanjing, Chengdu and Hong Kong) are similar and thus omitted.

locations, where those with high likelihoods of being visited by the target user occupy the top positions. It should be noted that this differs slightly from an e-commerce recommender system [50], where an item bought by the target user typically cannot be recommended again.

We evaluated the prediction accuracy using the AUC [51], namely the area under the receiver operating characteristics curve, which is similar to the ranking score [52]. The AUC measured how well a prediction algorithm could successfully distinguish relevant locations (those in the testing set) from the irrelevant locations. For a check-in (u, α) in the testing set, if the location α was ranked in the kth place among all N locations in the training set for the user u, its AUC value was $(N - k + 1)/N$. The AUC of an algorithm was defined as the average AUC value over all check-ins in the testing set. Clearly, for a random prediction, AUC $=$ 0.5, and for a prefect prediction, AUC $=$ 1. Therefore, the extent to which AUC exceeded 0.5 indicated the prediction accuracy. In addition to the AUC value, we also adopt the Recall value [53]. For every user, we provide a list of L predictions according to the algorithm. If a check-in (u, α) in the testing set was ranked within the top-L place, it is called correctly predicted. The Recall value is defined as the ratio of correctly predicted check-ins to the total

Table 3
Accuracy of location prediction quantified by AUC values. The six prediction algorithms presented here are history-based method (HB), popularity-based method (PB), demography-based method (DB), individual behavior-based method (IBB), collaborative-behavior-based (CBB) and logistic-regression-based method (LRB). All data points were obtained by averaging over 10 independent runs with random divisions into training and testing sets.

City	HB	PB	DB	IBB	CBB	LRB
Beijing	0.7830	0.8282	0.8307	0.9084	0.9077	0.9094
Shanghai	0.8462	0.8327	0.8580	0.9247	0.9234	0.9277
Nanjing	0.7855	0.7712	0.8193	0.8731	0.8638	0.8751
Chengdu	0.7914	0.8288	0.8601	0.8990	0.8921	0.9070
Hong Kong	0.7598	0.8875	0.8954	0.9262	0.9238	0.9292

Table 4
Accuracy of location prediction quantified by AUC values. The six prediction algorithms presented here are history-based method (HB), popularity-based method (PB), demography-based method (DB), individual behavior-based method (IBB), collaborative-behavior-based (CBB) and logistic-regression-based method (LRB). All data points were obtained by averaging over 10 independent runs with temporal divisions into training and testing sets.

City	HB	PB	DB	IBB	CBB	LRB
Beijing	0.7220	0.8286	0.8012	0.8928	0.8903	0.8953
Shanghai	0.8089	0.8321	0.8251	0.9094	0.9041	0.9102
Nanjing	0.7648	0.7692	0.8113	0.8583	0.8345	0.8645
Chengdu	0.7331	0.8229	0.8516	0.8911	0.8733	0.9054
Hong Kong	0.6745	0.9077	0.9022	0.9277	0.9267	0.9276

Table 5
AUC values for natives and non-natives under HB and PB methods. For each city, HB outperforms PB for natives while PB outperforms HB for non-natives.

	HB	PB
Beijing natives	0.8575	0.8272
Shanghai natives	0.8565	0.7936
Nanjing natives	0.8293	0.7406
Chengdu natives	0.8486	0.8014
Hong Kong natives	0.8665	0.8005
Beijing non-natives	0.7644	0.8923
Shanghai non-natives	0.8018	0.8569
Nanjing non-natives	0.6981	0.8316
Chengdu non-natives	0.7129	0.8790
Hong Kong non-natives	0.7417	0.9159

number of check-ins. In this paper, we only show the result at $L = 100$ (Recall@100). The recall values for other L and the Precision value [53] will give qualitatively the same results.

Tables 3 and 4 show the AUC values obtained using HB and PB. At the individual level, the visiting distribution of a native is statistically more heterogeneous than that of a non-native, so it is reasonable to expect that HB is more suitable for natives than non-natives. While at the population level, the visiting distribution contributed by all non-natives is more heterogeneous than that contributed by all natives, indicating that non-natives tend to visit some popular spots in a city and thus PB is more suitable for non-natives than natives. These inferences were supported by the empirical analysis. As shown in Table 5, the average prediction accuracy for natives was higher (or lower) than that for non-natives in every city under HB (or PB). Thus, we developed the so-called demography-based method (DB), which is a mixed algorithm that applies HB if the target user is a native person, or PB for a non-native person. As shown in Tables 3 and 4, the mixed algorithm outperformed both HB and PB.

Despite its considerable improvement, the demography-based method has two limitations. First, truthful demographic information is not easy to obtain in general online systems, and thus the applicability of the method is restricted. Second, some demographic evidence may be inconsistent with the behavioral patterns, where we aim to predict the latter, i.e., we want to quantify the extent to which an individual behaves like a native. Given that a native is more likely to visit some locations more times compared with a non-native (see the average Gini coefficients in Table 2), we propose an index I_i to count the ratio of repeated check-ins, i.e., for a user u, $N_T^{(u)}$ denotes the total number of u's check-ins and $N_D^{(u)}$ is the number of different locations visited by u; thus, the index is defined as:

$$I_i(u) = 1 - \frac{N_D^{(u)}}{N_T^{(u)}}. \tag{2}$$

For example, if the user u has seven check-ins at locations $\{A, B, C, A, B, A, D\}$, then $N_D^{(u)} = 4$ and thus $I_i(u) = 3/7$. This is an individual behavioral index because it only requires the behavioral information for an individual. Analogously, given that a native is less likely to visit popular locations than a non-native (see the Gini coefficients in Table 2), we propose an index I_c to count the average normalized popularity of a user's visits. First, we rank all of the locations in descending

Table 6
Accuracy of location prediction quantified by Recall with $L = 100$. The six prediction algorithms presented here are history-based method (HB), popularity-based method (PB), demography-based method (DB), individual behavior-based method (IBB), collaborative-behavior-based (CBB) and logistic-regression-based method (LRB). All data points were obtained by averaging over 10 independent runs with random divisions into training and testing sets.

Recall	HB	PB	DB	IBB	CBB	LRB
Beijing	0.5409	0.2247	0.5507	0.5974	0.5774	0.6054
Shanghai	0.6411	0.1938	0.6436	0.6639	0.6379	0.6781
Nanjing	0.6275	0.5712	0.6873	0.7849	0.7564	0.7957
Chengdu	0.6461	0.7151	0.7682	0.8383	0.8315	0.8396
Hong Kong	0.5669	0.8062	0.8248	0.8885	0.8851	0.8902

Table 7
Accuracy of location prediction quantified by Recall with $L = 100$. The six prediction algorithms presented here are history-based method (HB), popularity-based method (PB), demography-based method (DB), individual behavior-based method (IBB), collaborative-behavior-based (CBB) and logistic-regression-based method (LRB). All data points were obtained by averaging over 10 independent runs with temporal divisions into training and testing sets.

Recall	HB	PB	DB	IBB	CBB	LRB
Beijing	0.4253	0.2337	0.5061	0.5049	0.5207	0.5271
Shanghai	0.5746	0.2142	0.5899	0.6098	0.5998	0.6134
Nanjing	0.5912	0.5776	0.6781	0.7536	0.7386	0.7612
Chengdu	0.5387	0.684	0.7513	0.8248	0.8084	0.8319
Hong Kong	0.4788	0.8408	0.8425	0.8957	0.8917	0.8968

order according to their visiting frequencies (locations with the same frequency are ranked randomly). Next, a location's normalized popularity is defined by its ranking score, e.g., if the Great Wall is ranked fifth among all 28,100 locations in Beijing, its normalized popularity is 5/28 100. Thus, the I_c of a user is the average normalized popularity of the user's visits, where a location that appears several times should also be counted several times. If a user u's sequential check-in locations are $l_1, l_2, \ldots, l_{N_T^{(u)}}$, then the index is:

$$I_c(u) = \frac{1}{N_T^{(u)}} \sum_{k=1}^{N_T^{(u)}} R(l_k), \tag{3}$$

where $R(l_k)$ is the normalized popularity (i.e., ranking score) of location l_k. In contrast to I_i, I_c is a collaborative behavioral index because it requires the behavioral information for all users. I_i and I_c are called indigenization coefficients, and a larger value indicates a higher similarity to a native for both coefficients.

Thus, we propose an individual-behavior-based (IBB) method. Given a target user u, the predicted score for a location l is:

$$S_i(u, l) = I_i^{\alpha}(u)P(u, l) + [1 - I_i(u)]^{\alpha} Q(l), \tag{4}$$

where α is a free parameter, P is the normalized history score for user u and location l, which is defined as the ratio of the number of u's check-ins at location l relative to the total number of u's check-ins, and Q is the normalized popularity score for location l, which is defined as:

$$Q(l) = N(l)/\max_{l'} N(l'), \tag{5}$$

where $N(l)$ is the number of visits at location l by all users. Analogously, a collaborative-behavior-based (CBB) method scores a location l as follows:

$$S_c(u, l) = I_c^{\alpha}(u)P(u, l) + [1 - I_c(u)]^{\alpha} Q(l). \tag{6}$$

Tables 3 and 4 shows the AUC values for IBB and CBB with the optimal parameter α. Surprisingly, without using any demographic information, the accuracies of both IBB and CBB were improved greatly compared with the demography-based method.

Tables 6 and 7 show the recall values at $L = 100$ for all the above-mentioned algorithms, which are in accordance with the results of AUC values.

5. Discussion

Based on the distinct mobility patterns of natives and non-natives, we developed a mixed algorithm that uses demographic information, which greatly outperformed basic algorithms based on the visiting history of individuals and location popularity. We also developed two indigenization coefficients, I_i and I_c, for estimating the extent to which an individual behaves like a native. Without any demographic information, the simple hybrid algorithm weighted using each indigenization coefficient greatly improved the prediction accuracy compared with the mixed algorithm (DB). It

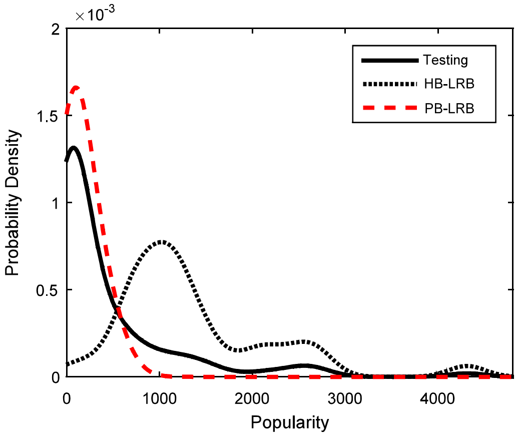


Fig. 5. The probability density functions of visiting frequencies of locations (i) in the testing set (back solid curve), (ii) having not been correctly predicted by HB while been correctly predicted by LRB (red dash curve), and (iii) having not been correctly predicted by PB while been correctly predicted by LRB (black dot curve). If a location has been appeared several times in different check-ins, it will be counted several times. This figure shows the result for Beijing, and the other four cities are similar.

is not surprising that grouping people before applying mixed or hybrid algorithms (e.g., see Ref. [54]) can yield more accurate predictions than those obtained with the basic algorithms. However, we must emphasize that it is not necessary to group people based on their attributes, but instead we can simply characterize them based on behavioral coefficients after determining the correlations between attributes and behavioral patterns. Shifting from demographic information to indigenization coefficients is only an example; indeed, we could build a user's profile based only on their behavioral records and select effective features based on an in-depth understanding of the relationships between attributes and behaviors. This method differs from attribute-based methods in two respects: (i) it can be applied without attribute information because knowledge of how the behavioral features are selected can be learned from other similar systems with attribute information [55]; (ii) if the selections are appropriate, a number of quantitative behavioral coefficients can provide a more accurate description of users than a static classification based on attributes. Furthermore, this method differs from mainstream machine learning methods (e.g., the ensemble learning method [56]) because it can provide insights in addition to predictions.

Although the current prediction accuracies are competitive (for example, even HB can provide slightly more accurate predictions than the one-order Markov chain method [43] that gives AUC values as 0.7819, 0.8460, 0.7723, 0.7858 and 0.7588 for Beijing, Shanghai, Nanjing, Chengdu and Hong Kong, respectively), the readers should be aware that this paper does not aim at the design of a better-performed yet complicated algorithm than other state-of-the-art ones, but to reveal the difference between mobility patterns of natives and non-natives, as well as to show the power of behavioral data analysis. It is very possible that a machine learning algorithm with carefully selected group of features via the feature engineering techniques can beat the current algorithm in prediction accuracy. Therefore, we emphasize phenomena and perspectives, rather than the details and refinements of algorithms. Indeed, our proposed prediction algorithms can be improved further in many ways. For example, the prediction accuracy of IBB and CBB can be improved further by introducing a stretched index to make the mean values of $P(u, l)$ and $Q(l)$ the same, although this would make the equation and learning process more complicated. We can also define an integrated indigenization coefficient:

$$I = \frac{1}{1 + \exp(-w_i I_i - w_c I_c)}, \tag{7}$$

where the parameters w_i and w_c can be learned from the logistic regression that best classifies natives and non-natives. As shown in Tables 3, 4, 6 and 7, this logistic-regression-based (LRB) method can improve the accuracy further, but it requires the demographic information and the learned parameters for different cities are very different.

We have recorded all the check-ins in the testing set which have not been correctly predicted by HB or PB, but been correctly predicted by LRB. These two sets are respectively denoted by HB → LRB and PB → LRB. Fig. 5 shows the distributions of visiting frequencies of corresponding locations in HB → LRB and PB → LRB for Beijing (other cities are similar). For PB → LRB, the result is trivial since only less attractive locations can be missed in PB, namely in the set PB → LRB, there are no popular locations. For HB → LRB, the distribution is remarkably boarder than that of all corresponding locations in the testing set, indicating that individuals also prefer to visit popular yet unvisited locations, which is a typical behavioral pattern of non-natives. Therefore, the LRB algorithm can be to some extent treated as a tradeoff of HB and PB. To compare HB or PB with IBB or CBB will achieve the similar results.

The current data is insufficient in two aspects: (i) the size of the dataset is not big enough, especially in Nanjing, Chengdu and Hong Kong, where the numbers of native users are very small; (ii) the covered time period is relatively short, only a litter bit more than one year. As possibly future works, if we have sufficiently long-term and large-scale datasets, we can measure the trend in the integrated indigenization coefficient and estimate whether a non-native person will behave like a native person after a specific time period and, if this is the case, how long the average person requires to act like a native. At the same time, taking into account the temporal information, one may design a time-aware indigenization coefficient that can more effectively distinguishing different behavioral patterns between natives and non-natives and lead to more accurate location prediction. Besides, the cross-region (i.e., data from different cities in different countries) and cross-data-type (i.e., GPS, RFID, Wi-Fi and other types of datasets) analyses are very helpful in validating the current findings and uncovering other mobility patterns.

Acknowledgments

The authors acknowledge their helpful discussions with Xiaoyong Yan. This research was supported partly by the National Natural Science Foundation of China under Grant Nos. 11222543 and 61433014, the Program for New Century Excellent Talents in University under Grant No. NCET-11-0070, and the CCF-Tencent open fund.

References

[1] P. Wang, M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabási, Understanding the spreading patterns of mobile phone viruses, Science 324 (2009) 1071–1076.
[2] J. Yuan, Y. Zheng, X. Xie, Discovering regions of different functions in a city using human mobility and POIs, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2012, pp. 186–194.
[3] R. Kitamura, C. Chen, R.M. Pendyala, R. Narayaran, Micro-simulation of daily activity-travel patterns for travel demand forecasting, Transportation 27 (2000) 25–51.
[4] V.W. Zheng, Y. Zheng, X. Xie, Q. Yang, Collaborative location and activity recommendations with GPS history data, in: Proceedings of the 19th International Conference on World Wide Web, ACM Press, New York, 2010, pp. 1029–1038.
[5] T.H. Dao, S.R. Jeong, H. Ahn, A novel recommendation model of location-based advertising: context-aware collaborative filtering using GA approach, Expert Syst. Appl. 39 (2012) 3731–3739.
[6] X.-Y. Yan, X.-P. Han, B.-H. Wang, T. Zhou, Diversity of individual mobility patterns and emergence of aggregated scaling laws, Sci. Rep. 3 (2013) 2678.
[7] M.C. González, C.A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns, Nature 453 (2008) 779–782.
[8] I. Rhee, M. Shin, S. Hong, K. Lee, S.J. Kim, S. Chong, On the levy-walk nature of human mobility, IEEE/ACM Trans. Netw. 19 (2011) 630–643.
[9] A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti, L. Giovannini, Statistical laws in urban mobility from microscopic GPS data in the area of Florence, J. Stat. Mech. Theory Exp. (2010) P05001.
[10] X. Liang, X.-D. Zheng, W.-F. Lü, T.-Y. Zhu, K. Xu, The scaling of human mobility by taxis is exponential, Physica A 391 (2012) 2135–2144.
[11] C.-M. Song, T. Koren, P. Wang, A.-L. Barabási, Modelling the scaling properties of human mobility, Nat. Phys. 6 (2010) 818–823.
[12] B. Jiang, J. Yin, S. Zhao, Characterizing the human mobility pattern in a large street network, Phys. Rev. E 80 (2009) 021136.
[13] X.-P. Han, Q. Hao, T. Zhou, B.-H. Wang, Origin of the scaling law in human mobility: hierarchy of traffic systems, Phys. Rev. E 83 (2011) 036117.
[14] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, C. Mascolo, A tale of many cities: universal patterns in human urban mobility, PLoS One 7 (2012) e37027.
[15] R. Gallotti, A. Bazzani, S. Rambaldi, Towards a statistical physics of human mobility, Internat. J. Modern Phys. C 23 (2012) 1250061.
[16] C.-M. Song, Z.-H. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility, Science 327 (2010) 1018–1021.
[17] H.A. Tran, A. Schlyter, Gender and class in urban transport: the cases of Xian and Hanoi, Environ. Urban 22 (2010) 139–155.
[18] D. Salon, S. Gulyani, Mobility, poverty, and gender: Travel 'Choices' of slum residents in Nairobi, Kenya, Transp. Rev. 30 (2010) 641–657.
[19] T.P. Uteng, T. Cresswell (Eds.), Gendered Mobility, Ashgate Publishing Limited, Surrey, 2008.
[20] F. Girardin, F. Calabrese, F. Dal Fiorre, A. Biderman, C. Ratti, J. Blat, Uncovering the presence and movements of tourists from user-generated content, in: International Forum on Tourism Statistics, 2008.
[21] S. Hanson, Gender and mobility: new approaches for informing sustainability, Gender Place Culture 17 (2010) 5–23.
[22] S. Fobker, R. Grotz, Everyday mobility of elderly people in different Urban settings: The example of the city of Bonn, Germany, Urban Stud. 43 (2006) 99–118.
[23] M. O'Brien, D. Jones, D. Sloan, M. Rustin, Children's independent spatial mobility in the Urban public realm, Childhood 7 (2000) 257–277.
[24] I.A. Junglas, R.T. Watson, Location-based services, Commun. ACM 51 (3) (2008) 65–69.
[25] B. Hernández, M.C. Hidalgo, M.E. Salazar-Laplace, S. Hess, Place attachment and place identity in natives and non-natives, J. Environ. Psychol. 27 (2007) 310–319.
[26] Z. Li, F. Wu, Residential satisfaction in China's informal settlements: A case study of Beijing, Shanghai, and Guangzhou, Urban Geogr. 34 (2013) 923–949.
[27] M. Horner, M. O'Kelly, Embedding economics of scale concepts for hub network design, J. Transp. Geogr. 9 (2001) 255–265.
[28] V.W. Zheng, B. Cao, Y. Zheng, X. Xie, Q. Yang, Collaborative filtering meets mobile recommendation: A user-centered approach, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence, AAAI Press, 2010, pp. 236–241.
[29] M. Ye, P. Yin, W.C. Lee, D.L. Lee, Exploiting geographical influence for collaborative point-of-interest recommendation, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, 2011, pp. 325–334.
[30] B. Liu, Y. Fu, Z. Yao, H. Xiong, Learning geographical preferences for point-of-interest recommendation, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2013, pp. 1043–1051.
[31] E. Cho, S.A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2011, pp. 1082–1090.
[32] J. Chang, E. Sun, Location3: How users share and respond to location-based data on social networking sites, in: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, AAAI Press, 2011, pp. 74–80.
[33] C. Cheng, H. Yang, I. King, M.R. Lyu, Fused matrix factorization with geographical and social influence in location-based social networks, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, AAAI Press, 2012.
[34] H. Gao, J. Tang, H. Liu, Exploring social-historical ties on location-based social networks, in: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, AAAI Press, 2012.
[35] A. Sadilek, H. Kautz, J.P. Bigham, Finding your friends and following them to where you are, in: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, ACM Press, New York, 2012, pp. 723–732.
[36] N.J. Yuan, F. Zhang, D. Lian, K. Zhang, S. Yu, X. Xie, We know how you live: exploring the spectrum of urban lifestyles, in: Proceedings of the First ACM Conference on Online Social Networks, ACM Press, New York, 2013, pp. 3–14.
[37] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, Y. Rui, GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2014, pp. 831–840.
[38] S. Yitzhaki, More than a dozen alternative ways of spelling Gini, Res. Econ. Inequal. 8 (1998) 13–30.
[39] J.A. Rice, Mathematical Statistics and Data Analysis, third ed., in: Duxbury Advanced, 2006.
[40] Z.-D. Zhao, Z. Yang, Z.-K. Zhang, T. Zhou, Z.-G. Huang, Y.-C. Lai, Emergence of scaling in human-interest dynamics, Sci. Rep. 3 (2013) 3472.

[41] D. Lian, X. Xie, F. Zhang, N.J. Yuan, T. Zhou, Y. Rui, Mining location-based social networks: A predictive perspective, IEEE Data Eng. Bull. 38 (2) (2015) 35–46.
[42] S.-C. Liou, H.-C. Lu, Applied neural network for location prediction and resources reservation scheme in wireless networks, in: Proceedings of International Conference on Communication Technology, IEEE Press, 2003, pp. 958–961.
[43] L. Song, D. Kotz, R. Jain, X. He, Evaluating location predictors with extensive Wi-Fi mobility data, in: Proceedings of the Twenty-Third Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM, IEEE Press, 2004, pp. 1414–1424.
[44] G. Yavas, D. Katsaros, Ö. Ulusoy, Y. Manolopoulos, A data mining approach for location prediction in mobile environments, Data Knowl. Eng. 54 (2005) 121–146.
[45] S. Akoush, A. Sameh, Mobile user movement prediction using Bayesian learning for neural networks, in: Proceedings of the International Conference on Wireless Communications and Mobile Computing, ACM Press, New York, 2007, pp. 191–196.
[46] A. Monreale, F. Pinelli, R. Trasarti, F. Giannotti, Wherenext: a location predictor on trajectory pattern mining, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2009, pp. 637–646.
[47] A. Noulas, S. Scellato, N. Lathia, C. Mascolo, Mining user mobility features for next place prediction in location-based services, in: Proceedings of the 12th IEEE International Conference on Data Mining, ICDM, IEEE Press, 2012, pp. 1038–1043.
[48] X. Lu, E. Wetter, N. Bharti, A.J. Tatem, L. Bengtsson, Approaching the limit of predictability in human mobility, Sci. Rep. 3 (2013) 2923.
[49] D. Lian, Y. Ge, F. Zhang, N.Y. Yuan, X. Xie, T. Zhou, Y. Rui, Content-aware collaborative filtering for location recommendation based on human mobility data, in: Proceedings of the 15th IEEE International Conference on Data Mining, ICDM, IEEE Press, 2015, pp. 261–270.
[50] L. Lü, M. Medo, C.H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems, Phys. Rep. 519 (2012) 1–49.
[51] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.
[52] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Bipartite network projection and personal recommendation, Phys. Rev. E 76 (2007) 046115.
[53] J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T. Riedl, Evaluating collaborative filtering recommender systems, ACM Trans. Inf. Syst. 22 (2004) 5–53.
[54] R. Burke, Hybrid recommender systems: Survey and experiments, User Modell. User-Adapt. Interact. 12 (2002) 331–370.
[55] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345.
[56] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.