Understanding travellers' preferences for different types of trip destination based on mobile internet usage data

Wang, Yihong; Correia, Gonçalo Homem de Almeida; van Arem, Bart; Timmermans, H. J.P.(Harry)

DOI
10.1016/j.trc.2018.03.009

Publication date
2018

Document Version
Final published version

Published in
Transportation Research Part C: Emerging Technologies

Transportation Research Part C 90 (2018) 247–259

Contents lists available at ScienceDirect

Transportation Research Part C

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

Understanding travellers' preferences for different types of trip destination based on mobile internet usage data

Yihong Wang[a,*], Gonçalo Homem de Almeida Correia[a], Bart van Arem[a], H.J.P. (Harry) Timmermans[b,c]

[a] Department of Transport and Planning, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands
[b] Department of the Built Environment, Section of Urban Systems and Real Estate, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[c] Department of Air Transportation Management, Nanjing University of Aeronautics and Astronautics, China

ARTICLE INFO

Keywords:
Mobile internet usage
Mobile phone data
Travel behaviour
Mobility analysis
Data fusion

ABSTRACT

New mobility data sources like mobile phone traces have been shown to reveal individuals' movements in space and time. However, socioeconomic attributes of travellers are missing in those data. Consequently, it is not possible to partition the population and have an in-depth understanding of the socio-demographic factors influencing travel behaviour. Aiming at filling this gap, we use mobile internet usage behaviour, including one's preferred type of website and application (app) visited through mobile internet as well as the level of usage frequency, as a distinguishing element between different population segments. We compare the travel behaviour of each segment in terms of the preference for types of trip destinations. The point of interest (POI) data are used to cluster grid cells of a city according to the main function of a grid cell, serving as a reference to determine the type of trip destination. The method is tested for the city of Shanghai, China, by using a special mobile phone dataset that includes not only the spatial-temporal traces but also the mobile internet usage behaviour of the same users. We identify statistically significant relationships between a traveller's favourite category of mobile internet content and more frequent types of trip destinations that he/she visits. For example, compared to others, people whose favourite type of app/website is in the "tourism" category significantly preferred to visit touristy areas. Moreover, users with different levels of internet usage intensity show different preferences for types of destinations as well. We found that people who used mobile internet more intensively were more likely to visit more commercial areas, and people who used it less preferred to have activities in predominantly residential areas.

1. Introduction

There is a recent trend in complementing or even replacing traditional travel survey data with new mobility-related data sources, such as GPS data, mobile phone traces and smart card transaction data (Chen et al., 2016; Demissie et al., 2013a; Iqbal et al., 2014; Ni et al., 2018; Toole et al., 2015; Wang et al., 2017; Wolf, 2006; Yue et al., 2014; Zhao et al., 2018). These trajectory-based data are getting popular for travel analysis because (1) they are inexpensive to collect; (2) they are usually up to date; and (3) most of them contain a large sample with observations that are longitudinal in time (Calabrese et al., 2013; Demissie et al., 2013b; Morency et al.,

* Corresponding author.
E-mail addresses: Y.Wang-14@tudelft.nl (Y. Wang), G.Correia@tudelft.nl (G.H.d.A. Correia), B.vanArem@tudelft.nl (B. van Arem), H.J.P.Timmermans@tue.nl (H.J.P.H. Timmermans).

https://doi.org/10.1016/j.trc.2018.03.009
Received 31 October 2017; Received in revised form 13 March 2018; Accepted 13 March 2018
Available online 23 March 2018
0968-090X/

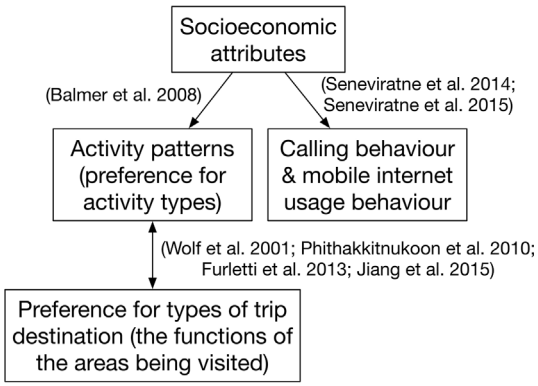


Fig. 1. The conceptual framework.

2007).

However, despite the potential advantages, these sources of information only include the spatial-temporal traces describing people's movements. If the aim is to understand travel behaviour from an activity-based perspective (Chen et al., 2016; Rasouli and Timmermans, 2014; Zhao and Zhang, 2017), the information of these data sets is usually very limited. For example, activity purpose of the trips is typically missing (Calabrese et al., 2013). Moreover, in traditional travel demand models, socioeconomic information is used to segment the population, and better explain the heterogeneity of activity-travel behaviour, including, but not limited to, activity patterns (Balmer et al., 2008) and location choice (Sivakumar and Bhat, 2007). However, in anonymous big data, socioeconomic information is unavailable mainly due to privacy reasons (Calabrese et al., 2014).

To deal with such problems, researchers have tried to combine different types of data in order to fill the gaps (Anda et al., 2017). In attempting to derive activity purpose information from trajectory data, there have been several applications fusing trajectory data with land use data, OpenStreetMap data or point of interest (POI) data (Dashdorj et al., 2013; Demissie et al., 2015; Wolf et al., 2004; Yuan et al., 2012). This geo-coded background knowledge can help estimating the function of an area, which can tentatively be connected to the type of activity that a visitor performed in that area (Furletti et al., 2013; Jiang et al., 2015; Phithakkitnukoon et al., 2010; Wolf et al., 2001). We referred to the main function of an area being visited as "type of trip destination" in this paper. The left chain in Fig. 1 shows how we derive the dependency of one's preference for destination types on socioeconomic attributes, based on literature review. Intuitively, such dependency exists in most cities. For example, it is common that some specific urban areas are more frequented by young people.

To partition the population using mobile phone data, Arai et al. (2014) and Bwambale et al. (2017) suggested utilizing calling behaviour such as calling frequency and duration to predict one's personal attributes. However, mobile phones are less used for calls today, making calling behaviour less useful, while simultaneously people are spending more time on services provided by mobile internet such as mobile apps (Richmond, 2012). Therefore, mobile internet usage behaviour, if available, could have a greater potential to reflect individuals' traits, such as gender and age (Seneviratne et al., 2015, 2014). The right chain in Fig. 1 shows the dependency of mobile internet usage behaviour on socioeconomic attributes.

As a whole, Fig. 1, which can be regarded as a conceptual framework, shows the relationship between mobile internet usage behaviour and preference for types of trip destination. Since they are both dependent on the socioeconomic attributes, even if the socioeconomic attributes are unobserved, they are still likely to be correlated with each other. Based on this hypothesis derived from the conceptual framework, our study aims to understand travellers' preferences for types of trip destination by means of segmenting them based on the preferred type of sites and applications visited through mobile internet as well as the level of visiting frequency, by fusing mobile phone traces and mobile internet usage data. We are allowed to do this study because of the data provided by the Shanghai Unicom WO+ Open Data Application Contest.[1]

Furthermore, mobile internet usage behaviour might sometimes be able to reflect even more information about a person, such as one's specific interests and lifestyles, than the traditional socioeconomic attributes do. At the same time, one's interests and lifestyles are regarded as the determinants of location choice through preference for different types of non-work activities (Wen and Koppelman, 2000). A more specific interest or lifestyle might be related to a more specific travel preference especially for non-work activities. For example, a foodie would visit more sites and applications about food, and meanwhile, he/she would also like to visit more restaurants in real life. We see the potential to explore such relationships by fusing mobile internet usage data and mobile phone traces, and we especially focus on the types of destinations for out-of-home non-work activities, designated herein as secondary activities for simplicity. Many studies have used mobile phone data to analyse users' home and workplace locations as well as commuting trips (Ahas et al., 2010; Alexander et al., 2015; Calabrese et al., 2011; Isaacman et al., 2011). However, trips for secondary activities have not often been analysed using this type of data, except in only a few studies (e.g., Järv et al., 2014; Huang and Levinson, 2015), which does not mean that they are not an important part of urban travel demand. In fact, they are taking a larger share than ever before, especially in large metropolitan areas (Wang et al., 2017).

[1] https://www.kesci.com/woplus/.

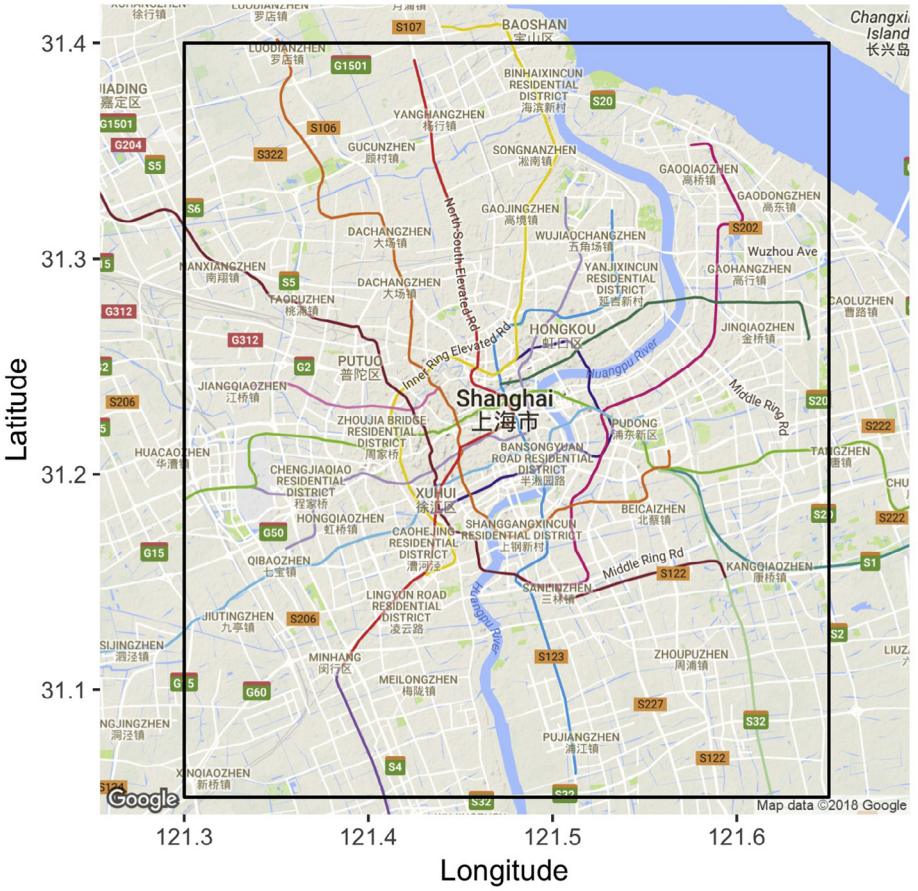


Fig. 2. The map of the target area.

The rest of this paper is organized as follows. First, we introduce the data used in our research. Next, we explain our research method. Then, the results are presented. Finally, we draw the conclusions, discuss the usefulness and limitation of our research, and point out the directions for future research.

2. Case study

In this paper, the case study is conducted in Shanghai, China. As one of the four directly-controlled municipalities of China, Shanghai is world famous for being a global financial centre and transport hub. The total area of Shanghai is 6340 square kilometres, and the population of Shanghai has exceeded 24 million. The city of Shanghai is divided into 16 districts. Except the Chongming district composed of three islands in the Yellow Sea, the other 15 districts lie on China's east coast. They are separated by Huangpu River into two banks, Pudong and Puxi, which literally mean the east bank and the west bank respectively in Chinese. Despite the crossing river, these two banks are well connected by several bridges, tunnels for cars and tunnels for metro. The boundary of our target area, covering the relatively more populous parts of Shanghai, is represented by the rectangle (about 1775 square kilometres) in Fig. 2, using the WGS 84 (EPSG:4326) reference coordinate system. Note that WGS 84/Pseudo-Mercator (EPSG:3857) is used as the projection system to calculate distance in this work.

2.1. Mobile phone data

The Shanghai Unicom WO+ Open Data Application Contest provides both the mobile phone traces and the mobile internet usage data of the same sample of the Shanghai Unicom users. Unicom is one of the three mobile carriers in China. It was reported that the total number of the Unicom mobile users had reached about 270 million in China by the beginning of 2017.

The mobile phone traces include the spatial-temporal records of 620 thousand sampled users moving within the city of Shanghai hour by hour from 12 a.m. 27th of December 2015, to 3 p.m. 6th of January 2016. Every time a user had a mobile phone activity (i.e., a call, a text message, a voice mail, or an internet connection), the location information and the timestamp of the activity would automatically be recorded with the anonymous user ID in the original database. However, the provided data were hourly aggregated for each user. To be specific, if a user was detected to have visited several locations within an hour, only the location where the user

stayed for the longest time would be known for that hour. It is also possible that a user did not have any mobile phone activity within an hour, thus leaving no location information. It is regarded as a missing trace for that user. The detected location information of an available trace is represented by a pair of coordinates using the WGS 84 (EPSG:4326) reference coordinate system. 4 digits of the longitude coordinate are stored after the decimal point, and 5 digits of the latitude coordinate are stored after the decimal point. According to the data provider, due to the inherent detection inaccuracy, the real location of a trace lies within the 200 m × 200 m square of which the centre is the detected point.

The mobile internet usage data include the page view counts of each user for different types of mobile apps and websites during the same study period. The page view counts of mobile apps and websites were merged for the same category, thus producing a total of 13 types of mobile internet contents: “finance”, “food”, “shopping”, “social news”, “housing”, “tourism”, “sports”, “car”, “entertainment”, “education”, “job seeking”, “game”, and “health”. The specific mobile apps and websites in each category were selected by the data provider. The users who never browsed any mobile internet contents are labelled with the tag “null”.

2.2. POI data

A POI is a specific point location associated with a pair of coordinates and some information about this location, such as name, category and description. The POI data used in our study were extracted from the Gaode Maps service,[2] which is the Chinese equivalent of Google Maps. The Gaode open platform allows the registered developers to obtain the POI data of a specific area through the application program interface (API). In our target area, about 260 thousand POIs of ten predefined categories can be obtained. The available information of the POI data includes name, coordinates and category. The ten categories are “hotel”, “sports and recreation”, “finance and insurance”, “residence”, “education”, “workplace”, “restaurant”, “car service”, “tourism”, and “health”.

3. Methodology

In Fig. 3, we present a flowchart of the proposed research method in this study. First, trip destinations chosen by the users for secondary activities can be extracted from mobile phone traces. Second, each trip destination can be labelled by the cluster of the grid cell calculated based on the POI data, and we can discover the users’ preferences regarding the types of trip destinations for secondary activities. Third, the favourite categories of mobile internet contents and the total usage intensity levels of the users can be derived from mobile internet usage data. Fourth, the relationships can be statistically tested between the users’ mobile internet behaviour and the users’ preferences for the types of trip destinations. We also perform a sensitivity analysis to examine to what extent results would be affected due to the inherent spatial inaccuracy of mobile phone traces.

3.1. Extracting trip information from mobile phone traces

3.1.1. Stay detection

To ensure that the problem of missing traces would not affect our analysis, we first filter the users and only focus on those who were traced at least 80% of the total hours. Also, we only focus on the users who were always traced within the prescribed boundary of the target area. To estimate the trips made by mobile users, it is important to distinguish stay locations (i.e., origins and destinations of trips) from pass-by locations, in the mobile phone traces (Ahas et al., 2010; Alexander et al., 2015; Wang and Chen, 2018; Zheng et al., 2009). Meanwhile, signal errors may lead to false movement of traces which do not represent actual movement of users (Çolak et al., 2015). The effects of such errors should be reduced as well. In this study, we adopt the main steps lately suggested by Alexander et al. (2015) to detect stay locations, whilst the parameter used in the third step is modified to suit our case:

1. For each user, we find the traces that are spatially close (within 300 m) to their subsequent observations and thus obtain the sets of geographically and temporally close traces. The medoid of the coordinates within each set is then calculated to update the locations of the traces.
2. The traces that are close in space but far apart in time need to be consolidated for each user as well. The complete-linkage hierarchical clustering algorithm is applied using 500 m as the threshold. In this algorithm, we first treat each point as a cluster and merge step by step the two clusters whose merger has the smallest diameter, until the smallest diameter reaches 500 m. The medoid of the coordinates within of each cluster is used to update the location of the traces.
3. To identify whether a user stays or passes through, a duration threshold should be chosen dependent on the assumed shortest activity duration as well as the sampling rate of the data. In our study, the sampling rate for each user is relatively small: at most one trace per hour. Hence, it is stipulated that at least two consecutive traces close in space can determine a stay point, which will necessarily lead to overlooking some short activities; however, this is the best that can be done to extract stay points with these data.

3.1.2. Detecting the trips for secondary activities

In this research, we will not study the trips toward home or work activities but focus on the trips for secondary activities. Thus, we need to detect them using the stay traces of the users. A possible way is to infer activity purposes based on the ground truth (i.e., the

[2] https://lbs.amap.com/.

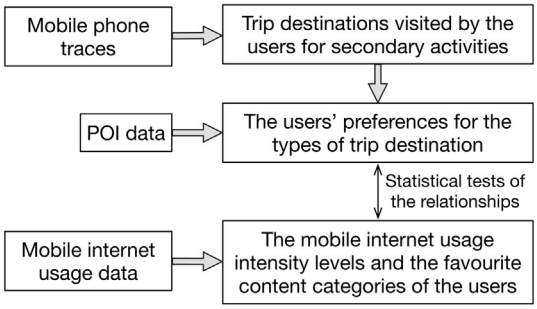


Fig. 3. The flowchart of the research method.

features related to a certain activity purpose). Those features can be sourced from either general knowledge, such as the fact that people mostly spend their night at home (Alexander et al., 2015; Nanni et al., 2013), or travel survey data, which provide more powerful evidence (Liu et al., 2013). In this study, since travel survey data are not available, we apply the rules suggested by Alexander et al. (2015) to detect the trips for secondary activities, and we choose the parameters that suit our case:

1. For each user, the home location is defined as the location with most stay traces from 7 p.m. to 8 a.m. on weekdays, on weekends, and on holidays.
2. The work location is defined as the place to which one travels the maximum accumulated distance from home, $\max(v*d)$, where v is the number of visits between 8 a.m. and 7 p.m. on weekdays during the study period, and d is the distance between a given place and home location. If the user visits the detected work location fewer than 2 days per week, it is not regarded as a work location.
3. It is assumed that the stay traces at the detected home location should be labelled as home activity. The same applies to labelling work activity, and the remaining stay traces are labelled as secondary activity. In this suggested approach, only stable home and workplace locations can be detected.

In the further analysis, we only focus on the users who had a home location and performed at least one secondary activity during the study period.

3.2. Clustering types of trip destinations for secondary activities

In this section, we further distinguish the trips for secondary activities in terms of the types of trip destinations. In traditional travel demand models, the purposes of secondary activities, such as eating out and shopping, can be used to distinguish the trips for secondary activities. However, it is very difficult to detect such purposes in mobile phone traces, especially without any travel survey data available as a reference. A compromising solution is to distinguish the trips for secondary activity based on geographical information of the visited area, such as land use (Wolf et al., 2001), and the POI data can be used to depict urban land use in a more detailed way (Jiang et al., 2015; Phithakkitnukoon et al., 2010; Yuan et al., 2012). Following this strategy, we define the types of trip destinations for secondary activities as follows.

A virtual grid reference can be constructed to divide the city (Demissie et al., 2015; Phithakkitnukoon et al., 2010). Each cell should be characterized and serves as a reference for determining the type of trip destination. For a cell $\in \{1,2,...,K\}$, the number of each type of POIs is calculated, named p_{kj}, where $j \in \{1,2,...,J\}$ indicates a POI type (e.g., restaurant or workplace). The number of POIs of each type is then ranked over all cells, and the percentile rank r_{kj} is calculated as the percentages of cells that have lower number of POIs of type j than cell k has. As a result, each cell k can be portrayed as a vector of the percentile ranks of all the POI types $\boldsymbol{r_k} = (r_{k1},r_{k2},...,r_{kJ})$.

In our study, a hierarchical clustering algorithm is applied to the vectors of all cells. We use the Pearson-correlation-based distance metric (Resnick et al., 1994; Xue et al., 2005) since we assume that the similarity between the functions of two areas can be reflected by the correlation between vector $\boldsymbol{r_k}$ and vector $\boldsymbol{r_{k'}}$, where $k' \in \{1,2,...,K\}\backslash\{k\}$. The distance $d_{kk'}$ between these two vectors, used in the clustering algorithm, is calculated in the following equation:

$$d_{kk'} = 1 - \frac{cov(\boldsymbol{r_k},\boldsymbol{r_{k'}})}{s(\boldsymbol{r_k})s(\boldsymbol{r_{k'}})} \tag{1}$$

where $cov(\boldsymbol{r_k},\boldsymbol{r_k'})$ is the covariance of $\boldsymbol{r_k}$ and $\boldsymbol{r_{k'}}$; $s(\boldsymbol{r_k})$ is the standard deviation of $\boldsymbol{r_k}$; $s(\boldsymbol{r_{k'}})$ is the standard deviation of $\boldsymbol{r_{k'}}$. Since correlation is scale-invariant, it is better to standardize $\boldsymbol{r_k}$ as $\hat{\boldsymbol{r}}_{\boldsymbol{k}}$ to represent the profile of a cell, whose element is calculated as follows:

$$\hat{r_{kj}} = (r_{kj} - \overline{r_k}) / \sqrt{\sum_{j\in J} (r_{kj} - \overline{r_k})^2} \tag{2}$$

where $\overline{\boldsymbol{r_k}}$ is the mean of $\boldsymbol{r_k}$. To find relatively more compact clusters of approximately equal diameters, we choose the complete-

linkage clustering method (Everitt et al., 2011; Sørensen, 1948). Consequently, each cell k can be related to a cluster $c \in \{1,2,...,C\}$. Cluster compactness can be assessed by the Dunn index (Dunn, 1973), which is the ratio of the smallest distance between observations in different clusters to the largest intra-cluster distance. Note that the distance used to calculate the Dunn index is still the Pearson-correlation-based distance defined previously. Intuitively, maximizing the Dunn index can help us select the optimal parameters and obtain the most distinctive urban area functions. In this case, the parameters to be selected include the number of clusters and the side length of the grid cells.

We pre-set the upper bound of the number of clusters as 10, equal to the number of dimensions of the POI data in this study, mainly for interpretation. In this study, we aim to interpret the statistical relationship between the preference for mobile internet contents and the preference for types of trip destination. Each type of trip destination is desired to have a distinctive characteristic. Thus, we expect our clusters to reflect the most distinctive urban functions. If the number of clusters is too large, the differences between some urban functions would possibly become very subtle, and the corresponding types of trip destination would be difficult to interpret.

Different from the other studies choosing an arbitrary value for the side length of the grid cells of the city, for example, 500 m (Phithakkitnukoon et al., 2010) and 800 m (Demissie et al., 2015), our study tests several values (i.e., 300 m, 400 m, 500 m, 600 m, 700 m and 800 m) as the side length of a grid cell. We only consider this range of values because the size of the grid cell should neither be too large nor too small. If it is too large, the defined function of a cell must become too rough; if it is too small, the detected destination of a trip would be very likely to lie in a wrong cell due to the inherent detection inaccuracy explained in Section 2.1. However, even if the size of a grid cell is very large, it is still possible that the detected destination and the real destination would lie in different grid cells. Thus, we will present the method to examine the impact of this issue on the final results in Section 3.5.

We generate the clustering results iteratively and choose the combination of the side length value and the number of clusters that can maximize the Dunn index and thus give us the most compact set of clusters. Consequently, each user has a set of trips for secondary activities during the study period. The coordinates of a trip destination can correspond to a grid cell and further correspond to a cluster c, which is defined to be the type of that trip destination.

3.3. Analysing mobile internet usage behaviour

Let f_{un} indicate the frequency of browsing a type of mobile internet content $n \in \{1,2,...,N\}$ (e.g., finance or shopping) through mobile apps and/or websites by an individual $u \in \{1,2,...,U\}$ across several days. Given this, two main indicators of one's mobile internet usage during a period can be derived: (1) the frequency of using all mobile internet contents $F_u = \sum_{n=1}^{N} f_{un}$, which reflects an individual's usage intensity, and (2) the relative preferences for using different types of mobile internet contents, expressed in terms of an N -dimensional vector $\boldsymbol{w_u} = (w_{u1},w_{u2},...,w_{uN})$, reflecting the different lifestyles and interests. We rank f_{un} for each n over all users and calculate w_{un} as the percentages of users who browse n less often than user u does.

Based on the total usage intensity, the population can be divided into three classes: (1) the "null" class, representing the people who never use any mobile internet service, (2) the "low intensity" class, representing the people whose usage intensity is lower than or equal to the median value of all non-zero total usage intensities, and (3) the "high intensity" class, representing the people whose usage intensity exceeds the median value of all non-zero total usage intensities. To segment the population using the preferences for specific contents, we find the content category n that maximizes w_{un} for a user u and use it to tag this user. Intuitively, such a tag is a user's favourite content category. For example, a user can predominantly be tagged as "shopping", "finance", etc.

3.4. Relating preferred types of trip destinations to mobile internet usage behaviour

In order to understand if there are statistically significant differences regarding the preferences for the different types of trip destinations among those who have different preferences for mobile internet content, we mainly use the statistical test of comparing two population proportions with independent samples (Ott et al., 2016), which is explained as follows.

The number of trips going to a destination of type c is aggregated over the users tagged as n regarding mobile internet content. The aggregate number of these trips is expressed as x_{cn}, and the total number of trips made by the users with the interest tag n is $x_n = \sum_c x_{cn}$. Then the proportion of trips to the destinations of type c made by the users with the interest tag n is $\rho_{cn} = x_{cn}/x_n$. On the other hand, the number of trips to the destinations of type c is aggregated over the remaining users who do not prefer n. The aggregate number of these trips is expressed as $x_{cn'}$, where $n' \in \{1,2,...,N\}\backslash\{n\}$, and the total number of the trips made by the remaining users is $x_{n'} = \sum_c x_{cn'}$. The proportion of trips to the destinations of type c made by the remaining users is $\rho_{cn'} = x_{cn'}/x_{n'}$. The two-tailed z-test, if following a normal distribution, is appropriate for our objective, which is to check whether the two proportions, ρ_{cn} and $\rho_{cn'}$, are different or the same, and the test statistic is given as follows:

$$Z_{cn} = (\rho_{cn} - \rho_{cn'}) / \sqrt{\rho^*_{cn}(1-\rho^*_{cn})(1/x_n + 1/x_{n'})} \tag{3}$$

where $\rho^*_{cn} = \frac{x_{cn} + x_{cn'}}{x_n + x_{n'}}$.

Based on the value of Z_{cn}, the significance of the difference can be derived, in terms of the corresponding p-value pv_{cn}. For every combination of c and n, we calculate the significance of the difference. Thus, there are $C \times N$ cases in total, causing the multiple comparisons problem: if a statistical analysis involves multiple simultaneous statistical tests, there will be more chances of rare events, increasing the likelihood of incorrectly rejecting a null hypothesis (Miller, 1981). Therefore, a stricter threshold of p-value should be used to reject a null hypothesis. In this study, we use the Bonferroni correction, which suggests dividing the original p-value

threshold by the number of hypotheses. In our case, we set the original p-value threshold as the typical one, 0.05, and the threshold after the Bonferroni correction is $0.05/(C \times N)$.

We construct an indicator of the significance of the preference $pref_{cn}$, explained in the following equation:

$$pref_{cn} = \begin{cases} \log_{10}(1/pv_{cn}) & if\ \rho_{cn} > \rho_{cn'}\ and\ pv_{cn} < 0.05/(C \times N) \\ -\log_{10}(1/pv_{cn}) & if\ \rho_{cn} < \rho_{cn'}\ and\ pv_{cn} < 0.05/(C \times N) \\ 0 & if\ pv_{cn} \geqslant 0.05/(C \times N) \end{cases} \tag{4}$$

The absolute value of this indicator is larger if the significance is higher. If the indicator is positive, it means that compared to the others, the users tagged by n significantly prefer to visit the destinations of type c. If the indicator is negative, it means that the users tagged by n significantly prefer not to visit the destinations of type c. If the indicator is zero, it means that there is no significantly different preference.

The same method can be applied to understand if there are statistically significant differences regarding people's preferences for different types of trip destinations among those who have a certain level of total mobile internet usage intensity.

3.5. Sensitivity analysis

Consider T_{ui} as the i th stay trace of an individual $u \in \{1,2,...,U\}$. The location of a stay trace can be represented by longitude lon_{ui} and latitude lat_{ui} in terms of metres using the WGS 84/Pseudo-Mercator (EPSG:3857) projection system. As mentioned in Section 2.1, the true location of a trace lies within the 200 m × 200 m square of which the centre is the detected point. Thus, it is possible that the true activity location does not lie in the correct grid cell. In this study, we assess the impact of such detection inaccuracy on the results regarding the statistical relationship between the preference for mobile internet contents and the preference for types of trip destination.

We assume that the longitude of the true location of a stay trace lon'_{ui} can be uniformly drawn inside the interval $[lon_{ui}-100, lon_{ui}+100]$, and the latitude of the true location lat'_{ui} can be uniformly drawn inside the interval $[lat_{ui}-100, lat_{ui}+100]$. We draw lon_{ui}' and lat_{ui}' of all the stay traces independently in 20 loops, except in the first loop where we set lon'_{ui} as lon_{ui} and set lat_{ui}' as lat_{ui}. In each loop, based on lon_{ui}' and lat_{ui}', the stay traces are assigned to their belonging grid cells. Consequently, the type of the trip destination corresponding to each detected trip can be determined in each loop. Finally, we mainly assess the two specific impacts on the results of the statistical relationships. First, we assess whether any conflicting significant results will be found in the 20 loops. Second, we examine whether the same significant relationships are robust enough to be found in more than 80% of the loops, namely 16 loops. We construct an indicator of the significance of the robust preference $pref'_{cn}$, explained in the following equation:

$$pref'_{cn} = \begin{cases} \log_{10}(1/\overline{pv_{cn}}) & if\ \rho_{cn} > \rho_{cn'}\ and\ pv_{cn} < 0.05/(C \times N)\ in\ more\ than\ 16\ loops \\ -\log_{10}(1/\overline{pv_{cn}}) & if\ \rho_{cn} < \rho_{cn'}\ and\ pv_{cn} < 0.05/(C \times N)\ in\ more\ than\ 16\ loops \\ 0 & Otherwise \end{cases} \tag{5}$$

where $\overline{pv_{cn}}$ is the average of the values of pv_{cn} in the loops, and $pv_{cn} < 0.05/(C \times N)$. The same method can also be applied to construct an indicator of the robust statistical relationships between mobile internet usage intensity and preferred types of trip destination.

4. Results and discussion

We processed the mobile phone traces using the method explained in Section 3.1. As a result, we obtained 26,535 target users meeting the specified criteria, and we detected their trips for secondary activities. Next, we clustered the grid cells using the method explained in Section 3.2. As shown in Fig. 4, based on the Dunn index, we can find that the clusters are best distinguished by setting the number of clusters as 6 or 7 and setting the side length as 500 m in our case. We chose the smaller number of clusters, 6, for interpretation.

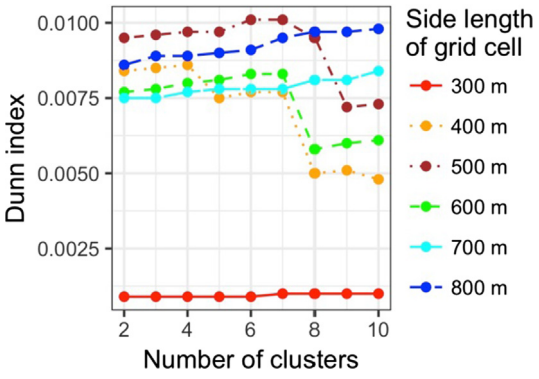


Fig. 4. The Dunn index used to determine the number of clusters and the side length of the grid cells.

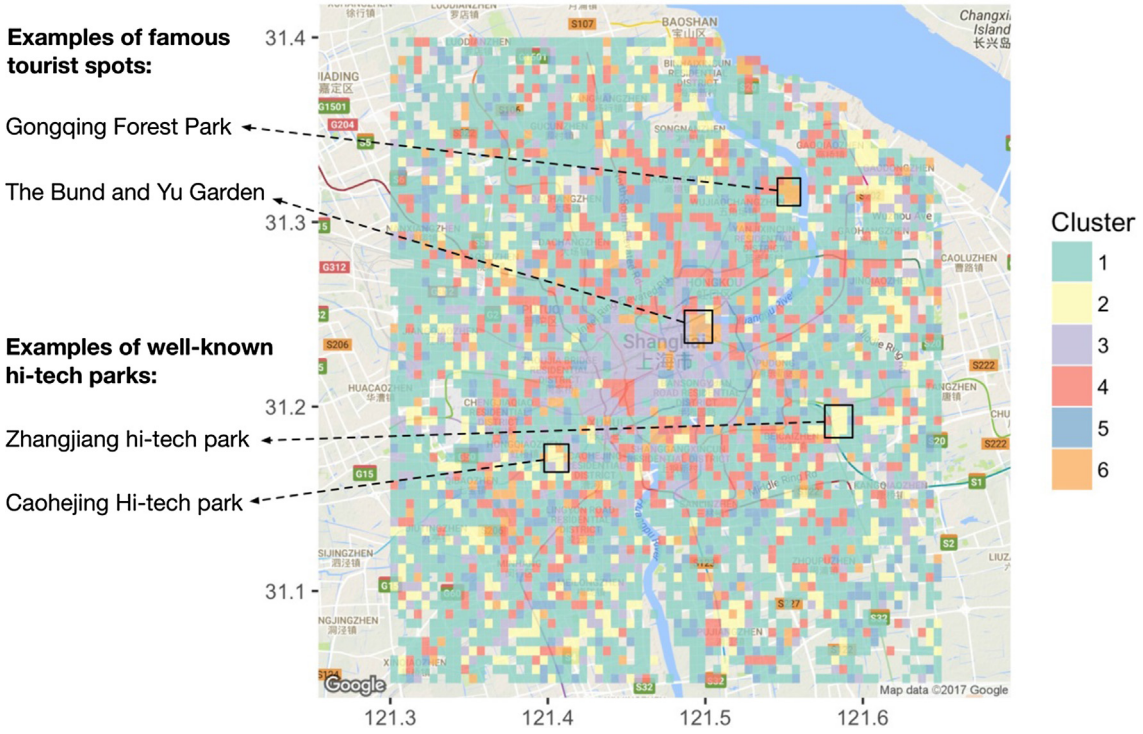


Fig. 5. The clustered grid cells of the city.

Fig. 5 geographically shows the computed clusters of the grid cells in the city. Table 1 shows the portrait of each cluster, where the profile chart depicts the first quartile, the median and the third quartile of $\widehat{n_{kj}}$ (see Section 3.2 for the definition) of the cells belonging to each cluster. $\widehat{n_{kj}}$ can indicate which POI types are dominant in a cell. For a cell k, the POI type j is relatively more influential if the value of $\widehat{n_{kj}}$ is higher.

In Table 1, it appears that the variances of $\widehat{n_{kj}}$ for restaurant, education and sports & recreation are relatively small among the clusters; hence, they are not the main factors making a cluster different from the others. On the other hand, the other POI types all seem to determine the characteristics of a cluster. Among them, the POI types of residence, workplace and tourism play the most important role in distinguishing the clusters. It can be observed that cluster 2, 3 and 6 represent the areas with relatively more workplaces, whilst cluster 1, 4 and 5 represent the areas with relatively fewer workplaces. Thus, it is not a surprise to see that the relative importance of residence is higher in cluster 1, 4 and 5. Among those more commercial clusters, cluster 3 is a special one since it seems to be all-round in terms of the relatively higher importance of most POI types including residence. Moreover, the relative importance of tourism is nearly zero in cluster 1, 2 and 5; in contrast, it is the highest in cluster 6, which thus seems to represent predominantly touristy areas.

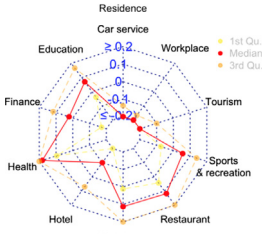
It can be observed in Fig. 5 that the city centre is mostly composed of the cells belonging to cluster 3, implying that the centre is more multifunctional compared to the remaining parts. In addition, we can find in Fig. 5 that the famous tourist spots such as Gongqing Forest Park, Yu Garden and the Bund all belong to cluster 6, and the well-known hi-tech parks such as Caohejing and Zhangjiang are assigned to cluster 2. These results make much sense based on our interpretation of the cluster profiles.

We also characterized each cluster with a few keywords in Table 1. Note that although cluster 1, 4 and 5 all somehow represent the predominant residential areas mainly outside the city centre, they are still different in terms of the relative importance of the other types of POIs within each cluster. In the areas belonging to cluster 1, there are relatively more POIs of car service in addition to residence. On the other hand, there are relatively more POIs of health in the areas belonging to cluster 5. The areas belonging to cluster 4 seem more multifunctional. They are almost similar with the areas belonging to cluster 3, except that they have a very low number of workplaces.

Fig. 6 presents the results of the statistical test between the mobile internet usage behaviour and the preferences for the types of trip destinations in the initial loop using the original spatial traces (see the explanation in Section 3.5). Intuitively, if a category of users like/dislike visiting the destinations of a specific type more significantly, the corresponding colour, representing the indicator $pref_{cn}$ (see the explanation in Section 3.4), will be deeper red/blue. Based on our definition of $pref_{cn}$ (equal to zero for the insignificant results), only the significant results are retained in the figure.

Next, we randomly draw the locations of the traces within the boundaries in 20 loops to examine the impact of mobile detection inaccuracy. We first found that there were no conflicting statistical relationships (e.g., ρ_{cn} is significantly larger than $\rho_{cn'}$ in one loop, but significantly smaller than $\rho_{cn'}$ in another loop) in these 20 loops. Fig. 7 further presents the results of the robust statistical relationships that held in more than 16 loops. Comparing Figs. 6 and 7, we can find that the preferences of people tagged by “health”,

Table 1
The portraits of the six clusters.

Cluster	Description		Profile chart (in terms of $\widehat{r_{kj}}$ of the cells belong to each cluster)
	Based on the profile	Based on the location (Fig. 5)	
1	Residential (with more car service POIs)	Mainly outside the centre	
2	Commercial (or industrial)	Mainly outside the centre	
3	All-round	Mainly in the centre	
4	Residential (more multifunctional)	Mainly outside the centre	
5	Residential (with more health POIs)	Mainly outside the centre	
6	Touristy and commercial	Some in the centre and some outside	

"shopping" and "social news" for certain types of trip destination did not hold in more than 16 loops, and the preferences of people who did not use any mobile internet were also sensitive to the possible spatial detection errors.

Results show that people who have different tastes in mobile internet content do have different preferences for different types of trip destinations. Some of the observed statistically significant results seem to be intuitive. More importantly, it seems that results can

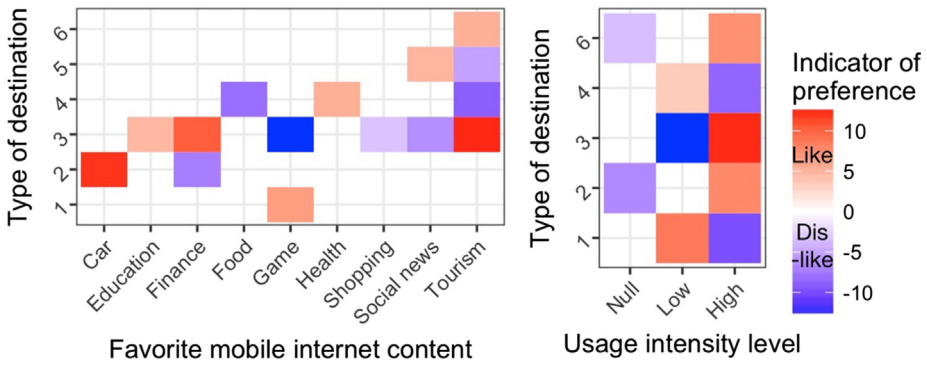


Fig. 6. The statistical relationships between the mobile internet usage behaviour and the preferences for the types of trip destinations in the initial loop using the original spatial traces.

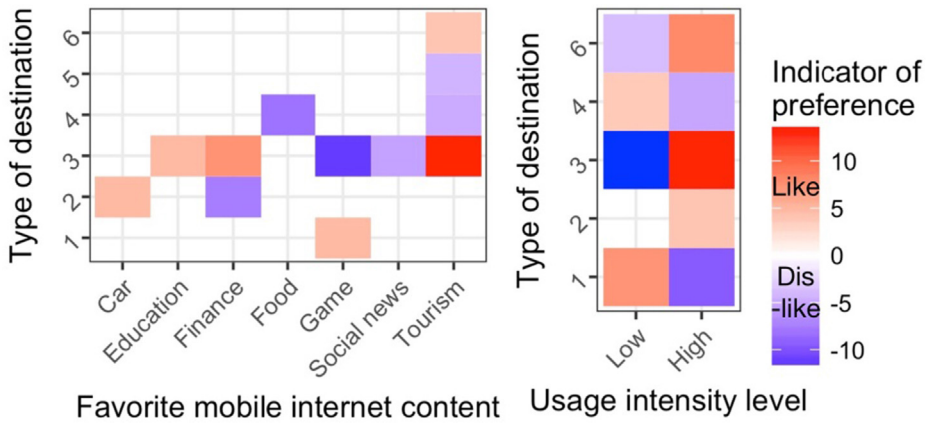


Fig. 7. The robust statistical relationships between the mobile internet usage behaviour and the preferences for the types of trip destinations in the 20 loops.

reflect some travel preferences that have never been captured in the existing literature using traditional travel survey data, mainly because travellers are now grouped based on their specific interests. For example, the destinations of type 6, which are mostly related to touristy areas, were only significantly visited by people tagged with the "tourism" label. Some existing studies have explored the travel preferences of car lovers, and they found that car lovers had their own preferences for residential or job location choice since it is flexible for them to travel farther, and they are not much willing to drive to downtown often (van Wee, 2009; van Wee et al., 2002). In our study, we found that car lovers also had their own preference for secondary activities, which shows that they would significantly like/need to visit the commercial or industrial areas far from the city centre. Despite being more multifunctional, the characteristics of the destination type 3, compared to the other types, are more similar with the concept of a CBD (central business district) since a CBD is usually located in the city centre, also being the most attractive part of the city. It was found using traditional travel survey data that younger people preferred to visit CBDs (Sivakumar and Bhat, 2007). In our study, we assume that users who preferred education and game contents are more likely to be younger, and they seem to have totally different preferences for the destination type 3. This implies that age might not be sufficient to segment the population for travel analysis, and mobile internet behaviour might be able to reflect a person's characteristics in more detail.

Users with different levels of total mobile internet usage intensity also had clearly different preferences for different types of destinations. Users who often used mobile internet preferred to visit the destinations of type 2, 3 and 6, and they did not visit often the destinations of type 1 and 4. In contrast, users who less often used mobile internet preferred to visit the destinations of type 1 and 4, and they did not visit often the destinations of type 3 and 6. It can be observed in Table 1 that cluster 2, 3 and 6 are distinct from cluster 1 and 4, as the former are more commercial, and the latter are only residential. Therefore, we can draw the conclusion that those who used mobile internet intensively were more likely to visit commercial areas, and those who used less mobile internet preferred to make trips to more residential areas. It is worth comparing our results with the results of an existing study using travel survey data (Giuliano et al., 2003). In that study, researchers found no significant difference between the land use of the places where the elderly and the non-elderly travelled. We have similar results in our case, if we assume people who never used mobile internet services were most likely to be the elderly. However, among those who used mobile internet services, the level of usage can be related to their preferred destination types.

It is also worth comparing our results with the relevant ones found in the recent studies using new big data sources. A research group from Estonia was able to access mobile phone traces associated to users' demographics, and by using such data, they conducted several studies to investigate the impact of ethnicity, age, and gender on activity locations and spaces (Järv et al., 2014; Silm et al.,

2017; Silm and Ahas, 2014). They found in their case studies that ethnicity had a significant influence on the spatial preferences of individuals for out-of-home non-work activities, and the ethnic segregation in activity spaces was higher in younger age groups. We believe that our results can complement the results of those studies since mobile internet usage data (the apps and webpages used) can characterize users in a different way. Also, it is arguably easier to re-identify users by using mobile phone traces associated to users' demographics, which is not desirable from a privacy perspective. Our approach seems to be able to distinguish different population segments at a relatively lower privacy risk. As another promising new mobility data source, social media data, including user-generated text, hash-tags, check-in information and even photos, can provide rich contexts of locations and users, allowing researchers to estimate more accurate activity purposes and find more specific interests of users (Hasan and Ukkusuri, 2014; Huang et al., 2017; Rashidi et al., 2017). Thus, it is also possible to characterize users and relate them to mobility behaviour using social media data. For example, Hasan and Ukkusuri (2015) used the Foursquare check-ins posted via Twitter to understand people's different attitudes and interests through activity locations. Also by using Twitter data, Abbasi et al., (2015) were able to identify the tourists in Sydney and at the same time find that they were more likely to visit touristy areas. However, an issue of using social media data for such analysis is that the users of a social media product may not be an unbiased sample of the general population of travellers, both demographically and geographically (Hasan and Ukkusuri, 2015), when compared to the general mobile phone users.

5. Conclusions and recommendations

This paper proposes a method to segment the population and understand travellers' preferences for types of trip destinations by fusing mobile internet usage data and mobile phone traces. The results of a case study, using a dataset from Shanghai, China, show that given one's favourite category of mobile internet content, the proportions of visiting some types of destinations were significantly higher, and the proportions of visiting some others were significantly lower. Many of these observed relationships were interpretable. For example, compared to the others, the users whose favourite content was "tourism" preferred to visit the touristy areas. Moreover, the users intensively using mobile internet were more likely to visit more commercial areas, and the users who used mobile internet less often would prefer to visit predominantly residential areas.

There are some limitations in this study which derive essentially from the data quality. The sampling rate of the mobile phone traces is relatively lower. As we have discussed in Section 3.1.1, we have to stipulate that at least two consecutive traces close in space can determine a stay point. This will necessarily lead to overlooking some short activities; however, it is the best approach for the available data. In addition, we only use the number of POIs to reflect the characteristics of an area; however, some additional information about the POIs can be added to improve our model. For example, in our case, we found that the number of restaurants is not very different in different areas of the region, but as we know, the quality of restaurants can be very diverse, and it is possible that the better restaurants are spatially distributed in a different way than the others. Thus, although our current model cannot distinguish the areas frequented by foodies, it may be possible to do so, by using more detailed data such as ratings or more specific categories of POIs.

People may also question about the potential privacy issues of such analysis since users generally do not want their mobility traces or mobile usage to be disclosed (Blondel et al., 2015). Despite such privacy risks, society can however benefit from using such big data for transport analysis and planning. Therefore, it is important to consider the extent to which such data should be pre-processed before being available for researchers or decision makers. For example, the data should be aggregated to prevent the privacy risks, whilst at the same time it should not be overly aggregated since it could cause the loss of information and make transport analysis not accurate. In our case, we think that the data provider found a good balance of data aggregation. They aggregated the mobile phone traces hour by hour. They also aggregated the specific websites and apps into several categories, and they only provided the frequency of each user visiting each category of websites and apps during a period. Even though the demographics of the users were removed, the aggregate mobile internet usage data can still help distinguish different population segments.

The significant and interpretable relationships found in this case suggest the potential of using mobile internet usage data to enhance the explanatory travel behaviour models in future research. Although we only explored the statistical significance in this case, several applications can be made based on the findings of our study. For example, mobile internet usage data may be used to predict mobile users' destination choice or for developing a travel behaviour model that would benefit from population partition, such as trip generation model and mode choice model.

Acknowledgment

We would like to express our gratitude to the Shanghai Unicom WO+ Open Data Application Contest for making the mobile phone data available for this research. We are grateful to the Yanxishe (a Chinese urban data research organization), who provided us the POI data extracted from the Gaode Maps service. Thanks go also to the TRAIL research school and the Dutch Organization for Scientific Research (NWO) for sponsoring the first author for his PhD study.

References

Abbasi, A., Rashidi, T.H., Maghrebi, M., Waller, S.T., 2015. Utilising location based social media in travel survey methods. In: Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks - LBSN'15. ACM Press, New York, New York, USA, pp. 1–9. http://dx.doi.org/10.1145/2830657.2830660.

Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. J. Urban Technol. 17,

3–27. http://dx.doi.org/10.1080/10630731003597306.
Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin – destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C 58, 240–250. http://dx.doi.org/10.1016/j.trc.2015.02.018.
Anda, C., Erath, A., Fourie, P.J., 2017. Transport modelling in the age of big data. Int. J. Urban Sci. 21, 19–42. http://dx.doi.org/10.1080/12265934.2017.1281150.
Arai, A., Witayangkurn, A., Kanasugi, H., Horanont, T., Shao, X., Shibasaki, R., 2014. Understanding user attributes from calling behavior. In: Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia - MoMM '14. ACM Press, New York, USA, pp. 95–104. http://dx.doi.org/10.1145/2684103.2684107.
Balmer, M., Meister, K., Nagel, K., 2008. Agent-based simulation of travel demand: Structure and computational performance of MATSim-T. In: The 2nd TRB Conference on Innovations in Travel Modeling.
Blondel, V.D., Decuyper, A., Krings, G., 2015. A survey of results on mobile phone datasets analysis. EPJ Data Sci. 4, 10. http://dx.doi.org/10.1140/epjds/s13688-015-0046-0.
Bwambale, A., Choudhury, C., Hess, S., 2017. Modelling trip generation using mobile phone data: a latent demographics approach. J. Transp. Geogr. In Press. http://dx.doi.org/10.1016/j.jtrangeo.2017.08.020.
Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput. 10, 36–44. http://dx.doi.org/10.1109/MPRV.2011.41.
Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transp. Res. Part C Emerg. Technol. 26, 301–313. http://dx.doi.org/10.1016/j.trc.2012.09.009.
Calabrese, F., Ferrari, L., Blondel, V.D., 2014. Urban sensing using mobile phone network data: a survey of research. ACM Comput. Surv. 47, 1–20. http://dx.doi.org/10.1145/2655691.
Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg. Technol. 68, 285–299. http://dx.doi.org/10.1016/j.trc.2016.04.005.
Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C., 2015. Analyzing cell phone location data for urban travel. Transp. Res. Rec. J. Transp. Res. Board 2526, 126–135. http://dx.doi.org/10.3141/2526-14.
Dashdorj, Z., Serafini, L., Antonelli, F., Larcher, R., 2013. Semantic enrichment of mobile phone data records. In: Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia - MUM '13. ACM Press, New York, USA, pp. 1–10. doi: 10.1145/2541831.2541857.
Demissie, M.G., Correia, G., Bento, C., 2015. Analysis of the pattern and intensity of urban activities through aggregate cellphone usage. Transp. A Transp. Sci. 11, 502–524. http://dx.doi.org/10.1080/23249935.2015.1019591.
Demissie, M.G., Correia, G., Bento, C., 2013a. Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. Transp. Res. Part C Emerg. Technol. 32, 76–88. http://dx.doi.org/10.1016/j.trc.2013.03.010.
Demissie, M.G., Correia, G., Bento, C., 2013b. Exploring cellular network handover information for urban mobility analysis. J. Transp. Geogr. 31, 164–170. http://dx.doi.org/10.1016/j.jtrangeo.2013.06.016.
Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. 3, 32–57. http://dx.doi.org/10.1080/01969727308546046.
Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Hierarchical Clustering. John Wiley & Sons, Ltd, pp. 71–110. http://dx.doi.org/10.1002/9780470977811.ch4.
Furletti, B., Cintia, P., Renso, C., Spinsanti, L., 2013. Inferring human activities from GPS tracks. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13. ACM Press, New York, USA, pp. 1. http://dx.doi.org/10.1145/2505821.2505830.
Giuliano, G., Hu, H.-H., Lee, K., 2003. Travel Patterns of the Elderly: the Rold of Land Use.
Hasan, S., Ukkusuri, S.V., 2015. Location contexts of user check-ins to model urban geo life-style patterns. PLoS One 10, e0124819. http://dx.doi.org/10.1371/journal.pone.0124819.
Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C Emerg. Technol. 44, 363–381. http://dx.doi.org/10.1016/J.TRC.2014.04.003.
Huang, A., Gallegos, L., Lerman, K., 2017. Travel analytics: Understanding how destination choice and business clusters are connected based on social media data. Transp. Res. Part C Emerg. Technol. 77, 245–256. http://dx.doi.org/10.1016/J.TRC.2016.12.019.
Huang, A., Levinson, D., 2015. Axis of travel: Modeling non-work destination choice with GPS data. Transp. Res. Part C Emerg. Technol. 58, 208–223. http://dx.doi.org/10.1016/J.TRC.2015.03.022.
Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transp. Res. Part C Emerg. Technol. 40, 63–74. http://dx.doi.org/10.1016/j.trc.2014.01.002.
Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A., 2011. Identifying important places in people's lives from cellular network data. In: International Conference on Pervasive Computing. Springer, Berlin, Heidelberg, pp. 133–151. http://dx.doi.org/10.1007/978-3-642-21726-5_9.
Järv, O., Ahas, R., Witlox, F., 2014. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. Transp. Res. Part C Emerg. Technol. 38, 122–135. http://dx.doi.org/10.1016/J.TRC.2013.11.003.
Jiang, S., Alves, A., Rodrigues, F., Ferreira Jr., J.C., Pereira, F., 2015. Mining point-of-interest data from social networks for urban land use classification and disaggregation. Comput. Environ. Urban Syst. 53, 36–46. http://dx.doi.org/10.1016/J.COMPENVURBSYS.2014.12.001.
Liu, F., Janssens, D., Wets, G., Cools, M., 2013. Annotating mobile phone location data with activity purposes using machine learning algorithms. Expert Syst. Appl. 40, 3299–3311. http://dx.doi.org/10.1016/j.eswa.2012.12.100.
Miller, R.G., 1981. Simultaneous Statistical Inference. Springer.
Morency, C., Trépanier, M., Agard, B., 2007. Measuring transit use variability with smart-card data. Transp. Policy 14, 193–203. http://dx.doi.org/10.1016/j.tranpol.2007.01.001.
Nanni, M., Trasarti, R., Furletti, B., Gabrielli, L., Van Der Mede, P., De Bruijn, J., De Romph, E., Bruil, G., 2013. Transportation planning based on GSM traces: A case study on Ivory Coast. In: Citizen in Sensor Networks. Springer, pp. 15–25. doi:10.1007/978-3-319-04178-0.
Ni, L., Wang, X., (Cara), (Michael) Chen, X., 2018. A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. Transp. Res. Part C Emerg. Technol. 86, 510–526. http://dx.doi.org/10.1016/J.TRC.2017.12.002.
Ott, L., Longnecker, M., Draper, J.D., 2016. An Introduction to Statistical Methods and Data Analysis. seventh ed. Cengage.
Phithakkitnukoon, S., Horanont, T., Lorenzo, G. Di, Shibasaki, R., Ratti, C., 2010. Activity-aware map : identifying human daily activity pattern using mobile phone data. In: Human Behavior Understanding. Springer, pp. 14–25. doi:10.1007/978-3-642-14715-9_3.
Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transp. Res. Part C Emerg. Technol. 75, 197–211. http://dx.doi.org/10.1016/J.TRC.2016.12.008.
Rasouli, S., Timmermans, H., 2014. Activity-based models of travel demand: promises, progress and prospects. Int. J. Urban Sci. 18, 31–60. http://dx.doi.org/10.1080/12265934.2013.835118.
Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. GroupLens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work - CSCW '94. ACM Press, New York, USA, pp. 175–186. http://dx.doi.org/10.1145/192844.192905.
Richmond, S., 2012. Smartphones hardly used for calls [WWW Document]. URL http://www.telegraph.co.uk/technology/mobile-phones/9365085/Smartphones-hardly-used-for-calls.html (Accessed 1.3.17).
Seneviratne, S., Seneviratne, A., Mohapatra, P., Mahanti, A., 2015. Your installed apps reveal your gender and more! ACM SIGMOBILE Mob. Comput. Commun. Rev. 18, 55–61. http://dx.doi.org/10.1145/2721896.2721908.
Seneviratne, S., Seneviratne, A., Mohapatra, P., Mahanti, A., 2014. Predicting user traits from a snapshot of apps installed on a smartphone. ACM SIGMOBILE Mob. Comput. Commun. Rev. 18, 1–8. http://dx.doi.org/10.1145/2636242.2636244.
Silm, S., Ahas, R., 2014. Ethnic differences in activity spaces: a study of out-of-home nonemployment activities with mobile phone data. Ann. Assoc. Am. Geogr. 104, 542–559. http://dx.doi.org/10.1080/00045608.2014.892362.

Silm, S., Ahas, R., Mooses, V., 2017. Are younger age groups less segregated? Measuring ethnic segregation in activity spaces using mobile phone data. J. Ethn. Migr. Stud. 1–21. http://dx.doi.org/10.1080/1369183X.2017.1400425.

Sivakumar, A., Bhat, C., 2007. Comprehensive, unified framework for analyzing spatial location choice. Transp. Res. Rec. J Transp. Res. Board 103–111. http://dx.doi.org/10.3141/2003-13.

Sørensen, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5, 1–34.

Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: Travel demand estimation using big data resources. Transp. Res. Part C Emerg. Technol. 58, 162–177. http://dx.doi.org/10.1016/J.TRC.2015.04.022.

van Wee, B., 2009. Self-selection: a key to a better understanding of location choices, travel behaviour and transport externalities? Transp. Rev. 29, 279–292. http://dx.doi.org/10.1080/01441640902752961.

van Wee, B., Holwerda, H., van Baren, R., 2002. Preferences for modes, residential location and travel behaviour: the relevance for land-use impacts on mobility. Eur. J. Transp. Infrastruct. Res. 2, 305–316.

Wang, F., Chen, C., 2018. On data processing required to derive mobility patterns from passively-generated mobile phone data. Transp. Res. Part C Emerg. Technol. 87, 58–74. http://dx.doi.org/10.1016/J.TRC.2017.12.003.

Wang, Y., Correia, G.H.D.A., de Romph, E., Timmermans, H.J.P., 2017. Using metro smart card data to model location choice of after-work activities: An application to Shanghai. J. Transp. Geogr. 63, 40–47. http://dx.doi.org/10.1016/j.jtrangeo.2017.06.010.

Wen, C.-H., Koppelman, F.S., 2000. A conceptual and methdological framework for the generation of activity-travel patterns. Transportation (Amst). 27, 5–23. http://dx.doi.org/10.1023/A:1005234603206.

Wolf, J., 2006. Applications of new technologies in travel surveys. In: Travel Survey Methods. Emerald Group Publishing Limited, pp. 531–544. http://dx.doi.org/10.1108/9780080464015-029.

Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary: experiment to derive trip purpose from global positioning system travel data. Transp. Res. Rec. J. Transp. Res. Board 1768, 125–134. http://dx.doi.org/10.3141/1768-15.

Wolf, J., SchöUnfelder, S., Samaga, U., Oliveira, M., Axhausen, K., 2004. Eighty weeks of global positioning system traces: approaches to enriching trip information. Transp. Res. Rec. J. Transp. Res. Board 1870, 46–54. http://dx.doi.org/10.3141/1870-06.

Xue, G.-R., Lin, C., Yang, Q., Xi, W., Zeng, H.-J., Yu, Y., Chen, Z., 2005. Scalable collaborative filtering using cluster-based smoothing. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '05. ACM Press, New York, USA, pp. 114. doi: 10.1145/1076034.1076056.

Yuan, J., Zheng, Y., Xie, X., 2012. 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 - KDD '12. ACM Press, New York, USA, pp. 186. doi:10.1145/2339530.2339561.

Yue, Y., Lan, T., Yeh, A.G.O., Li, Q.Q., 2014. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. Travel Behav. Soc. 1, 69–78. http://dx.doi.org/10.1016/j.tbs.2013.12.002.

Zhao, S., Zhang, K., 2017. Observing individual dynamic choices of activity chains from location-based crowdsourced data. Transp. Res. Part C Emerg. Technol. 85, 1–22. http://dx.doi.org/10.1016/J.TRC.2017.09.005.

Zhao, Z., Koutsopoulos, H.N., Zhao, J., 2018. Individual mobility prediction using transit smart card data. Transp. Res. Part C Emerg. Technol. 89, 19–34. http://dx.doi.org/10.1016/J.TRC.2018.01.022.

Zheng, Y., Zhang, L., Xie, X., Ma, W.-Y., 2009. Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th International Conference on World Wide Web - WWW '09. ACM, pp. 791. doi:10.1145/1526709.1526816.