Detecting urban road network accessibility problems using taxi GPS data

JianXun Cui a, Feng Liu b, Davy Janssens b, Shi An a, Geert Wets b, Mario Cools c

a School of Transportation Science and Engineering, Harbin Institute of Technology, 73, Huanghe Road, Nangang District, Harbin 150090, China
b Transportation Research Institute (IMOB), Hasselt University, Wetenschapspark 5, bus 6, B-3590 Diepenbeek, Belgium
c LEM, Université de Liège, Chemin des Chevreuils 1, Bât B52/3, 4000 Liège, Belgium

ABSTRACT

Urban population growth and economic development have led to the creation of new communities, jobs and services at places where the existing road network might not cover or efficiently handle traffic. This generates isolated pockets of areas which are difficult to reach through the transport system. To address this accessibility problem, we have developed a novel approach to systematically examine the current urban land use and road network conditions as well as to identify poorly connected regions, using GPS data collected from taxis. This method is composed of four major steps. First, city-wide passenger travel demand patterns and travel times are modeled based on GPS trajectories. Upon this model, high density residential regions are then identified, and measures to assess accessibility of each of these places are developed. Next, the regions with the lowest level of accessibility among all the residential areas are detected, and finally the detected regions are further examined and specific transport situations are analyzed.

By applying the proposed method to the Chinese city of Harbin, we have identified 20 regions that have the lowest level of accessibility by car among all the identified residential areas. A serious reachability problem to petrol stations has also been discovered, in which drivers from 92.6% of the residential areas have to travel longer than 30 min to refill their cars. Furthermore, the comparison against a baseline model reveals the capacity of the derived measures in accounting for the actual travel routes under divergent traffic conditions. The experimental results demonstrate the potential and effectiveness of the proposed method in detecting car-based accessibility problems, contributing towards the development of urban road networks into a system that has better reachability and more reduced inequity.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

1.1. Accessibility measures and the limitations in terms of travel data

With the continuing urbanization of the world’s population and the economic growth of cities, spatial urban areas expand and new communities and activity locations are decentralized. However, existing transport networks have not developed at the same pace as urban growth, generating isolated pockets of areas which are difficult to reach by the transport systems. There is an urgent need to further understand the changing land use structure and transport conditions, in order to accurately identify poorly accessible areas and improve the accessibility of these places as well as the accessibility of the cities as a whole (e.g. Gwilliam, 2013).

Accessibility is defined as the ease and extent to which land-use and transport systems enable individuals to reach activities and destinations by means of certain transport modes, e.g. the number of jobs accessible within 30 min by car (e.g. Geurs and Wee, 2004). It takes into account not only travel efficiencies (e.g. driving speeds and congestion) but also the distribution of land-use and activity locations across the transport network. It is considered as a key dimension of quality of life and a priority for sustainable urban transport management and planning.

Various methods have been developed for measuring accessibility, and they are classified into four major categories, including infrastructure-based (e.g. DETR, 2000), location-based (e.g. Anderson et al., 2013), person-based (e.g. Neutens et al., 2012) and utility-based (e.g. Gulhan et al., 2013) measures. As this current study focuses on the identification of regions with low levels of accessibility, the location-based method is employed. The measures examine accessibility from a location point of view, and incorporate land-use components and travel time constraints. They have been widely used in urban planning and geographical studies (e.g. Novak and Sullivan, 2014). To compute location-based measures, the geographical area under investigation is first divided into a set of small analysis zones or study regions. Alongside, activity regions are represented with the size or number of their associated activity opportunities. Next, the physical separation between each pair of the study and activity regions is measured in terms of...
Travel times. Based on the derived times, accessibility measures for each of the study regions are ultimately obtained.

In the location-based measure building process, the essential element lies in the derivation of travel times. So far, three types of data have been utilized to derive this parameter for car-based trips, including travel surveys, static sensors, and GIS-based (Geographic Information System based) estimation. The existing accessibility research (Anderson et al., 2013; Owen and Levinson, 2012) can represent the state-of-art of such utilization. In these studies, the entire urban area is divided into 1236 zones, and a matrix of travel times between each pair of the origin and destination zones (OD) is derived. The methods as well as the travel data employed in each step of the entire OD-travel-time-matrix generation process can be described in Fig. 1.

Despite this comprehensive process as demonstrated in Fig. 1, the data used for the computation has raised concerns about the accuracy of the resultant OD-travel-time-matrix. First, travel surveys, which document daily activities and travel of a small sample of individuals, have a number of limitations (e.g. Liu et al., 2014). Secondly, due to the high cost, static sensors are usually installed on freeways, leading to the data only covering the high-capacity roads but shedding little light on the traffic in the rest of the city (e.g. Jenelius and Koutsopoulos, 2013). Thirdly, GIS-based methods are used to search for shortest paths, based on the assumption that travelers try to minimize travel times between their origins and destinations. However, as acknowledged by the authors of the above-described studies, this is not necessarily the case. Evidence from other research (e.g. Wolf et al., 2004) has also shown that individuals do not actually choose the shortest routes but more often the ones longer than assumed, due to a variety of other reasons, such as congestion and travel time reliability.

Due to these data constraints, the entire process of the OD-travel-time-matrix construction is not only labor and computation intensive, but it also requires the adoption of different models and hypotheses, increasing the level of uncertainties in the final estimated travel times. In addition, the analysis is restricted to a statistical average day; as a result, the measures are incapable of distinguishing accessibility among various types of days. Furthermore, the adoption of a constant speed for all the local roads ignores traffic variations at different times and across each individual link. Consequently, they cannot accurately reveal accessibility in various time periods of a day and therefore not well account for congestion.

1.2. Taxi GPS data

The advancement of Global Positioning Systems (GPS) has created the opportunity to use this technology as a new travel data collection method. For travelers who carry GPS devices with their vehicles, the accurate travel routes and travel times can be monitored automatically, providing detailed spatial and temporal travel information and near real time traffic conditions. Particularly, in many major cities around the world, GPS devices are already installed in taxis originally for taxi dispatch systems, thus no additional cost is incurred for the data collection.

Taxi GPS data has been employed for a range of studies, particularly for travel demand modeling (e.g. Lu and Li, 2014), travel time estimation and policy analysis (e.g. Castro et al., 2012; Li et al., 2012; Qian et al., 2014; Zheng et al., 2011). Regarding travel time estimation, the existing studies can be classified into two major classes. (i) Direct data-based travel time estimation, in which travel times of each major road are regarded as normally distributed and the mean of the observed values is used as the estimation of the resultant travel times for the link (e.g. Günthermann et al., 2004; Mustary et al., 2012). (ii) Indirect network-based travel time modeling, in which travel times are modeled based on taxi GPS data, while taking into account road network information and specific traffic situations (e.g. Feng et al., 2014; Rahmani and Koutsopoulos, 2013). With respect to policy analysis based on taxis GPS data, Zheng et al. (2011) proposed a method to detect flawed planning regarding newly built roads and subways in the road network in the Chinese capital of Beijing. Castro et al. (2012) developed an approach to predict traffic density and congestion levels to estimate the effect of emissions on air quality in the context of transport planning and environmental monitoring.

However, despite the variety of researches and applications, taxi GPS data has so far not been explored for accessibility analysis. Owen and Levinson (2012) pointed out that the direct traffic measurement using GPS as opposed to sophisticated modeling is the most promising way for ongoing accessibility analysis. Despite the fact that the potential value of large-scale GPS data in accessibility research has been underlined, a method, which is based on this type of data and which systematically assesses the accessibility of the entire urban road network as well as detects poorly accessible regions, has not yet been developed.

1.3. Research contributions

This study addresses the above-mentioned limitations with respect to the development of reliable methods for the systematic examination of accessibility in the urban road network. Specifically, a set of measures is developed that evaluate accessibility for each of the regions of the city, based on taxi GPS data. The obtained measures can then be used to identify regions with low levels of accessibility, and the specific land use and

---

**Fig. 1. The process of computing the OD-travel-time-matrix.**

- **Travel time estimates of each road link in the transport network**
  - Freeway travel times
  - Freeway ramp delay times
  - Arterial road travel times
  - Local street road travel times

- **OD travel time matrix construction**
  - Shortest travel routes

- **Data sources and methods**
  - Static sensors: Methods: travel speed direct computation, and indirect imputation such as linear extrapolation.
  - Static sensors: Methods: Poisson modeling.
  - Travel surveys and/or static sensors: Methods: travel demand modelling and prediction between each OD pair, traffic assignment to arterial roads, and travel time estimation.
  - Methods: assumption of an overall speed of 37km/h.

---
The development of urban land use and population growth. In many cities worldwide, urban transport problems in the detected regions are further investigated. Compared to traditional techniques, the proposed method offers the following advantages. (i) It builds more temporal and spatial sensitive measures, capturing varied traffic conditions across different time periods of the day as well as between weekdays and weekends. (ii) The results provide more objective and up-to-date measures, catching up with the fast pace of urban land use development and population growth. (iii) The massive GPS data enables better modeling of travel time distributions, leading to the derived results more accurately reflecting the true accessibility situations. (iv) In many cities worldwide, GPS devices are already installed in taxis, generating no extra financial costs in terms of data collection, making it a cost-effective approach and easily transferable to the cities. (v) In particular, in this study, the GPS data recorded from all taxis licensed in Harbin is explored. The data provides a unique opportunity for the examination into the accessibility across an extensive part of the road network.

The remainder of this paper is organized as follows. Section 2 describes the GPS data and Section 3 details the methodology. A case study is carried out in Section 4 and the experimental results are compared against a baseline model in Section 5. Section 6 examines the GPS data coverage issue and Section 7 gives discussions for future research. Finally, Section 8 ends this paper with major conclusions.

2. Data description

The GPS data was collected in the Chinese city of Harbin, in which all licensed taxis (approximately 16,000) are equipped with GPS devices as a part of security measures to protect drivers from being assaulted. The devices record the positions of the taxis every 30s (second) during the day and every 2 min at night, generating data of 1.6 GB in size and 24 million GPS points each day. The data includes a list of variables, ranging from vehicle positions to hardware status indicators (e.g. cameras), and to security-related information (e.g. alarm signals). According to the GPS data, the average number of passenger trips for a taxi is 30 trips per day. This generates a total of 0.48 million passenger trips. Although this figure could under-estimate the total taxi demand, as some passengers ultimately might switch to other transport modes or re-plan their trips (e.g. due to long waiting times), the comparison between the number of the actually realized passenger trips and the number of trips made by private cars still reflects the importance of taxis in undertaking the urban travel demand. In Harbin, there are roughly 1 million private cars. If we assume an average of 2.41 trips per day for each private car in line with the regional statistics (Mei and Gui, 2009), all the private vehicles produce 2.41 million trips. As a result, the taxi passenger trips account for 16.6% of the total personal travel made by both taxis and private cars within the urban area each day.

In our study, the GPS data collected between July and September in 2013 is used, and the variables include taxi vehicle ids, recording times and coordinates of GPS points, and status messages indicating whether or not passengers are on board. Alongside the GPS trajectories, a digital map of the entire road network of the city is also utilized. This dataset includes the coordinates and types of all activity locations (19 types in total) in the urban area.

3. Methodology

3.1. Overview of the approach

The method is composed of 4 major steps, including (i) modeling city-wide taxi passenger travel patterns, (ii) identifying high density residential regions and building accessibility measures, (iii) detecting regions with the lowest level of accessibility among the obtained residential areas, and (iv) further examining the land use and transport situations of the detected regions. Prior to the analysis, a preliminary step is conducted for passenger trip identification.

3.2. Passenger trip identification

A GPS trajectory from a taxi during a day can be described as a sequence of time-ordered GPS points, referred as \( p_1(t_1, s_1), p_2(t_2, s_2), \ldots, p_n(t_n, s_n) \), where \( n \) is the total number of the points. Each \( p_k(k=1,\ldots,n) \) contains a geospatial coordinate set \( l_k = (x_k, y_k) \), time stamp \( t_k \), and status message \( s_k \) equal to 1 when the taxi is ‘occupied’ by clients and 0 when the taxi is ‘idle’ and the driver is looking for clients. The driving speed at each point, i.e. \( Speed_k \), is calculated as the ratio between the Euclidian distance from the point to its next point and the time interval of these two points as follows:

\[
Speed_k = \frac{ED(l_{k+1}, l_k)}{t_{k+1} - t_k}
\]

Due to random errors, raw GPS data usually contains wrong records that are incompatible with the physical phenomena of the traffic. In the data cleaning process, the GPS points, which have the coordinates as zero or \( Speed_k \) higher than a threshold, i.e. \( TH_{Speed} \), are deleted. From the remaining traces, passenger trips are identified based on \( s_k \). In our study, in order to analyze taxi passenger travel patterns, only the passenger trips, in which taxis are occupied by customers, are used. In addition, the taxis with passengers on board are also expected to better reflect real traffic situations, e.g. driving speeds.

For each of the obtained passenger trips (i.e. \( p_1(l_1, s_1), p_2(l_2, s_2), \ldots, p_n(l_n, s_n) \)), identified as Trip, two variables are computed according to formula 2, including the travel time \( d_{Trip} \) and the route directness or circuitry \( r_{Trip} \), i.e. the ratio between the actual travel distance and the Euclidian distance.

\[
d_{Trip} = t_e - t_b
\]

\[
r_{Trip} = \frac{\sum_{k=1}^{n-1} ED(l_k, l_{k+1})}{ED(l_1, l_n)}
\]

3.3. Taxi passenger travel pattern modeling

The entire urban area is divided into \( Grid_X \times Grid_Y \) disjoint regions using a grid-based method, with each region as \( r_i(i=1,\ldots, Grid_X \times Grid_Y) \). The temporal dimension of trips is classified into different time periods of a day, i.e. \( TimeP \), and different types of the day, i.e. \( DayT \). Based on the spatial and temporal division, the start and end positions along with the start times of each of the trips are projected into the corresponding regions and periods. A 5-dimensional passenger OD-travel-pattern-matrix, i.e. \( OD(r_i, r_j, TimeP, DayT) \), is constructed, with each element accommodating all trips that leave from \( r_i \) end in \( r_j \) and start within \( TimeP \) on the Day with the type of \( DayT \).

From the OD-travel-pattern-matrix, two variables, including \( m_0 \) and \( m_d \), which describe the average number of trips that either start in \( r_i \) or end in \( r_j \) in TimeP over all survey days with DayT, are obtained. In order to derive the average travel times, the trips with \( r_{Trip} \) higher than a threshold, i.e. \( TH_{Trip} \), are first removed from the OD-travel-pattern-matrix. The total number of the remaining trips from \( r_i \) to \( r_j \) in TimeP over all survey days with DayT, referred as \( m_{ij} \), is then counted. If \( m_{ij} \) is higher than another parameter, i.e. \( TH_{ij} \), the distribution of \( d_{Trip} \) of all these trips is approximated by a Normal distribution \( N(d_{Trip}|\mu, \sigma) \), with \( \mu \) and \( \sigma \).
representing the mean and standard variance, respectively, computed as follows.

\[
N \left( d_{\text{trip}} \mid u_{ij}, \sigma_j \right) = \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\left( \frac{\left(d_{\text{trip}} - u_{ij}\right)^2}{2\sigma_j^2} \right)}
\]

\[
u_j = u_j(r_i, r_j, \text{TimeP, DayT}) = \frac{\sum_{i=1}^{m_j} d_{\text{trip}}}{m_j}
\]

\[
\sigma_j = \sqrt{\frac{\sum_{i=1}^{m_j} \left(d_{\text{trip}} - u_j\right)^2}{m_j - 1}}
\]

In this process of inferring travel time distributions, we assume that, most of taxi drivers are honest and typically find out the fastest routes to send passengers to a destination, and the travel times derived from the passenger trips represent the real travel duration needed between the origin and destination places. However, for the few remaining drivers who might deliberately give passengers a roundabout trip, or who have to bring passengers to multiple destinations in the case of taxi sharing, the observed travel times are longer than the actually required times. The variable \( r_{\text{trip}} \) is thus used to differentiate between a trip that reflects the actual travel duration and the other that could take a much longer time. The trips that are characterized by an unexpected high value of \( r_{\text{trip}} \) are excluded from the travel time calculation. The other parameter \( T_{\text{Htrip}} \) is employed to ensure that, a sufficient number of trips between the two considered regions has been observed in order to infer the travel time distribution that is more accurate and better representative of the general travel conditions.

### 3.4. Accessibility measure building

Using the above estimated travel demand and travel time distributions, location-based accessibility measures, i.e. the contour measures in particular, are computed. The advantages of contour measures are related to the operationalization, interpretability and communicability criteria. The measures are relatively easy to interpret for researchers and policy makers, as no assumptions are made on a person’s perception of travel, land-use and their interaction (e.g. Geurs and Wee, 2004). In this process, the study and activity regions are first generated, based on which accessibility measures are derived.

To select study regions, the average number of passenger trips that start in the morning or end at night on a day in a region, encoded in \( m_0 \) or \( m_d \) in these corresponding time periods, is used. The regions, which have both \( m_0 \) and \( m_d \) greater than a threshold value, i.e. \( TH_M \), are regarded as high density residential areas and chosen as the study regions. Besides, all the activity locations in the city are projected into the regions according to the positions of the places. The regions, which accommodate at least one activity location, form the activity regions.

A contour measure for a region is defined as the total number of activity locations that could be reached from the region by a certain transport mode within a specific travel time. According to traditional methods (e.g. Anderson et al., 2013), the measure for \( r_i \) to all activity locations of a certain type (i.e. \( c \)) within a time constraint (i.e. \( TH_T \)), referred as \( AC.m_i \), as well as the measure to the activities of all types, referred as \( AC.m_i \), can be obtained in formula 4.

\[
AC.m_i = \sum_{c} AC.m_{ici}
\]

Where, \( u_{ij} \) represents the mean travel time from \( r_i \) to the activity region \( r_j \), \( u_{ij} \) is the total number of activity locations of type \( c \) in \( r_j \), and \( TotalOfReg(c) \) denotes the total number of regions that contain activity locations of type \( c \). \( AC.m_{ici} \) is the sum of the number of activity locations of each region \( r_j \) across the urban area that has the mean travel time to \( r_i \) (i.e. \( u_{ij} \)) shorter than or equal to \( TH_T \).

Based on the probability distribution of \( d_{\text{trip}} \) of all trips from \( r_i \) to \( r_j \), we modify formula 4 in the following manner.

\[
AC.p_c = \frac{TH_T}{\sum_{c} TotalOfReg(c)} \left( \int_{-\infty}^{TH_T} N \left( d_{\text{trip}} \mid u_{ij}, \sigma_j \right) d(d_{\text{trip}}) \right)
\]

\[
AC.p_i = \sum_{c} AC.p_{ci}
\]

In formula 5, \( \int_{-\infty}^{TH_T} N \left( d_{\text{trip}} \mid u_{ij}, \sigma_j \right) d(d_{\text{trip}}) \) is the cumulative distribution function for the normal probability density distribution \( N \left( d_{\text{trip}} \mid u_{ij}, \sigma_j \right) \) described in formula 3. This function calculates the probability that \( d_{\text{trip}} \) is found to have a value less than or equal to \( TH_T \). \( AC.p_i \) and \( AC.p_c \) represent the probability-based contour measures to the activity type \( c \) and to the combination of all types, respectively.

Both \( AC.m_i \) and \( AC.p_i \) characterize the accessibility of a region by taking into account land use and transport characteristics of the area. They are expressed with the total number of activity locations that are reachable within a certain time limit, which is further determined by the travel distances between the study region and the activity locations as well as the quality of the road infrastructure linking these places. A low value of the measures signals a problem of long time travel in order to reach activity destinations, due to long travel distances and/or bad traffic conditions, e.g. congestion.

Despite the commonality of these two types of measures, differences exist in terms of the utilization of the travel times and the subsequently derived results. \( AC.m_i \) uses \( u_{ij} \) as the only time information to compute the accessibility between \( r_i \) and each of the activity regions \( r_j \). Consequently, based on the comparison of \( u_{ij} \) and \( TH_T \), the trips between the two regions are either entirely included in the measures if \( u_{ij} \) is less than or equal to \( TH_T \) or excluded if \( u_{ij} \) is larger than \( TH_T \). This makes the measures very sensitive to the selection of \( TH_T \). Meanwhile, \( AC.m_i \) disregards the probability distribution of \( d_{\text{trip}} \) in the sense that even if \( u_{ij} \) is less than or equal to \( TH_T \), still a certain percentage of the trips between the two corresponding regions could have longer travel times than \( TH_T \), and that when \( u_{ij} \) is larger than \( TH_T \), a part of the trips could be realized with travel times shorter than or equal to \( TH_T \). The ignoring of the different travel times (i.e. the travel time distribution) thus leads to the measures not fully reflecting the actual accessibility situations. In contrast, \( AC.p_i \) takes into account the probability distribution of \( d_{\text{trip}} \) (i.e. not only \( u_{ij} \) but also \( r_{\text{trip}} \)) and models the exact part of the trips that are performed within \( TH_T \). Thus, \( AC.p_i \) can more accurately mirror the true accessibility situations of the corresponding region. The differences between these two types of measures are further demonstrated in the case study section (i.e. Section 4.3).

### 3.5. Regions with low accessibility detection

Based on \( AC.p_i \), all the study regions are sorted; the regions with the measures lower than a certain percentage of the mean measure across all the study areas, i.e. \( TH_A \), are considered as the areas with the lowest level of accessibility. From these regions, people have to travel for a longer time to reach activity destinations, due to longer distances and/or
worse traffic conditions along the way, than they do from the remaining regions with higher accessibility values.

3.6. Specific land use and transport problem examination

An in-depth examination into the land use and transport situations surrounding each of the detected regions \( r_i \) is conducted in the following two aspects. (i) The geographic distribution of \( r_i \) and all the activity regions of the city. This is to examine the spatial distance of \( r_i \) from the activity places, especially from the city center where most activity entities are established. (ii) Accessibility measures of each individual type. This analysis investigates whether the low accessibility of the region is caused by certain particular activity types or by a general low level of reachability over all types.

4. Case study

In this section, adopting the proposed approach and using the data described in Section 2, we carry out a case study and demonstrate the effectiveness of the method in detecting accessibility problems.

4.1. Passenger trip identification

In the raw GPS dataset, 0.08% of the total points have the coordinates as zero, and 1.14% have Speed\( k \) higher than \( TH_{\text{speed}} \) set as 120 km/h, i.e. the maximum speed limit in China. All these error records are first deleted. The passenger trips are then identified; on average, 475,026 trips are obtained each day.

Fig. 2a describes the distribution of the average of Speed\( k \), over each half an hour throughout all the weekdays. It shows clear variations in driving speeds over different time periods of the day, signaling the importance of the temporal partition, as the different driving conditions lead to varied travel times and consequently different accessibility measures for a same pair of study and activity regions across different time periods. Based on this observed speed distribution, the day can be divided into 4 periods, including 7:30 am–8:30 am (morning), 8:30 am–16 pm (day), 16–18 pm (evening), and 18 pm–7 am (night). The average speed in each of these slots is 18.42, 21.40, 18.82 and 27.28 km/h, respectively.

Fig. 2b further illustrates the distribution of Speed\( k \) in each of the obtained periods. The differences among the mean speeds of the periods are examined using an ANOVA (Analysis of Variance) procedure. The p-value of the overall F-test is smaller than 0.0001, indicating that significant differences in the mean speeds exist. In addition, pairwise comparisons have been carried out to examine the deviations in mean speeds for each pair of the periods. The resulting p-values of these pairwise comparisons are all smaller than 0.0001, except the one from the pair of the morning and evening periods, which equals 0.107. Thus, based on the pairwise comparisons one can conclude that substantial differences in driving speeds exist between the different time periods, except for the morning and evening rush hour, during which similar low driving speeds are observed.

When the derived time periods are examined with the actual traffic situations in Harbin, it can be noted that this day segmentation is well consistent with the actual reporting, in which the morning and evening rush hour is delimited as the interval of 7–8:30 am and 16–18 pm respectively (e.g. Jiao and Li, 2012). This further demonstrates the ability of the GPS data in revealing the real traffic situations.

4.2. Travel pattern modeling

During the model construction, the entire city is divided into regions; \( Grid_X \) and \( Grid_Y \) decide the total number of the study units. The larger these values are, the higher the spatial resolution reaches, but the less the number of observed trips between the regions is. In order to derive results that are statistically sound and representative, these two variables are set as 40 respectively, resulting in a total of 1600 regions with each being 1.87 km² in size. When compared with other existing studies, it is noted that the average size of the study units varies depending on the study areas and transport modes. It ranges from 2.14 km² in the Twin Cities for car-based accessibility analysis (Anderson et al., 2013), to 0.146 km² in Denizli of Turkey for public transit-based accessibility studies (Gulhan et al., 2013), and to 0.026 km² and 0.050 km² in two small towns of Sweden for accessibility assessment on a combination of walks/bikes and autos (Makref and Folkesson, 1999).

Based on this spatial partition, along with the previously defined 4 temporal periods, the OD-travel-pattern-matrix \( OD(r_i,r_j,\text{TimeP},\text{DayT}) \) is constructed using the passenger trips, where \( ij \) = 1600, \( TimeP = 4 \), \( DayT = \text{weekdays/weekends} \), and \( Day = 65 \text{ weekdays/26 weekend days} \). Between different time periods of the day or different types of days, the activity location distribution may not change, but the travel times between a same pair of study and activity regions could differ to a large extent. In the case study, the accessibility in the morning of weekdays is analyzed, but the same process can be repeated to the remaining periods and weekends. The procedure is identical, but the results are likely to be different.

From the OD-travel-pattern-matrix, the average daily number of trips that either start in \( r_i \) in the morning or end in \( r_j \) at night, i.e. \( mo_i \) and \( md_j \), are obtained. Regarding the travel time distributions, the first parameter \( TH_i \) is set as 5% percentile of \( T_{\text{trip}} \) of all the trips, i.e. 2.78. The overall distribution of \( T_{\text{trip}} \) is presented in Fig. 3, with the mean as 1.46. The first parameter \( TH_{\text{trip}} \) is designated as 65, such that the considered region pairs have undergone at least one trip in the morning per day. For each of the region combinations filtered out by these two thresholds, \( u_i \) and \( c_{ij} \) are computed. A test on the statistical significance of the assumption that \( d_{\text{trip}} \) follows the Normal distribution is conducted. The
resultant p-values range from 0.48 to 0.99 across all these pairs, with 96.8% higher than 0.8. This implies that the hypothesis is valid and that the observed travel times indeed follow the Normal distribution.

A further investigation reveals that the relative low p-values (e.g. p-values < 0.8) are mainly caused by the trips with $d_{\text{trip}}$ longer than $u_{ij} + 3\sigma_{gi}$. Such rare instances only happen with the probability of 0.003 and are thus deleted. These outliers can be attributed to the likelihood that the corresponding taxis stop for a long time in the middle of the trips. In this scenario, the travel times are much longer than they actually are, but the route directness of these trips still reflects the actual geographic situations. Thus, these trips are not detected by the previous filtering process via $THT_{\text{trip}}$, which only eliminates the travel that takes a spatial detour. After all outliers from each of these region combinations are identified and removed, $u_{ij}$ and $\sigma_{gi}$ of all the remaining trips are re-computed and used for the subsequent analysis.

4.3. Accessibility measure building

Among all the 1600 regions, 108 (6.8%) are screened out as the study regions with $TH_T$ specified as 20, and 241 (15.1%) form the activity regions. Across all 2628 pairwise combinations between these regions, $m_0$ is higher than 65 previously set up for $TH_T_{\text{trip}}$. For each of these study regions, $AC_{mi}$ and $AC_{pi}$ are computed. Different cut-off values for $TH_T$ have been used in literature, depending on the type of considered activity types (e.g. Anderson et al., 2013). In this experiment, 30 min is adopted for the integrated analysis on all types.

Fig. 4a–b illustrate the travel time distribution of two typical study regions (i.e. $r_{12}$ and $r_{32}$) to the activity region (i.e. $r_a$) accommodating 524 facilities. As shown in Fig. 4a, the travel time $d_{\text{trip}}$ of all the trips between $r_{12}$ and $r_a$ follows the distribution of $N(d_{\text{trip}}, 26.5, 3.29)$. The mean $u_{ij}$ of 26.5 is shorter than the 30 min threshold, leading to $AC_{mi}$ to this particular activity region, referred as $AC_{mi-r_a}$ as 524. However, according to the distribution of $N(d_{\text{trip}}, 26.5, 3.29)$, the cumulative probability of $d_{\text{trip}}$ being longer than 30 min is 0.14. This suggests that, despite the shorter $u_{ij}$ than 30 min, still 14% of the total trips between $r_{12}$ and $r_a$ are conducted in travel times longer than this time limit. Thus, $AC_{mi}$ dismisses the probability of 0.14 at which trips longer than 30 min could occur between these two considered regions. Consequently, the obtained measure over-estimates the actual accessibility by 73.36. On the contrary, in Fig. 4b, $u_{ij}$ and $\sigma_{gi}$ between $r_{32}$ and $r_a$ are 36.56 and 10.67, respectively. The longer $u_{ij}$ than 30 min simply renders $AC_{mi-r_a}$ as 0, despite the cumulative probability of 0.27 at which travel could be realized within this time limit. The obtained measure thus results in an under-estimation of the actual accessibility by 141.48. The more opportunities an activity region provides, the larger extent the over- or under-estimation could be.

Given the potential over- or under-estimation by $AC_{mi-r_a}$ concerning each individual activity region $r_a$, if a study region has $u_{ij}$ shorter than or equal to (or longer than) $TH_T$ to more activity regions which contain more activities, the overall value of $AC_{mi}$ of this region tends to be over- (or under-) estimated accordingly.

Apart from the potentially biased accessibility estimation by $AC_{mi}$, this measure is also very sensitive to the selection of $TH_T$. This can be further demonstrated by the previous example. The accessibility for each of these two study regions $r_{12}$ and $r_{32}$ is measured only in two values, depending on the specification of $TH_T$. For $r_{12}$, the measure $AC_{mi-r_a}$ is either 524 if $TH_T$ is less than or equal to 26.5 min, or 0 if $TH_T$ is larger than this time duration. In the case of $r_{32}$, this measure is 524 if $TH_T$ is less than or equal to 36.56 min, or 0 if otherwise.

In contrast, $AC_{pi}$ takes into account the probability distribution of $d_{\text{trip}}$, leading to the measures accounting for the exact percentage of trips that could be realized within $TH_T$. Thus, the measure can accurately reveal the real accessibility situations of the corresponding regions. For instance, according to $AC_{pi}$, the precise measures of the regions $r_{12}$ and $r_{32}$ to $r_a$ are 450.64 and 141.48, respectively.

Fig. 5a describes the correlation between $AC_{mi}$ and $AC_{pi}$ for a same region, with $R$ as 0.96. It demonstrates that, while these two measures have an overall close link, deviation on the individual region level exists. While most of the study regions have lower $AC_{pi}$ than $AC_{mi}$, a small percentage (10.1%) exhibits the opposite trend. Fig. 5b further illustrates the distribution of these differences, with the average deviation $(AC_{mi-AC_{pi}})$ as 86.7.

4.4. Regions with the lowest level of accessibility detection

According to $AC_{pi}$, considerable variations in accessibility exist across the study regions. For example, the mean and standard variance of this variable are 6289 and 3005, with the maximum and minimum as 13,096 and 1002 respectively and the ratio between them as 13.1. In the context of the city with all 16,042 activity locations, the regions on average are able to reach 39.2% of all the services in 30 min by car, with the best and worst regions reaching 81.6% and 6.2% respectively.

As comparison, the probability-based contour measure with $TH_T$ as 40 min is computed. The mean of this measure is 9841, accounting for
61.3% of all the activity services. An increase by 22.1% in the number of accessible activities is gained at the cost of 10 min driving longer. The correlation \( R \) between these two measures is 0.98, suggesting that, while longer thresholds generally lead to higher absolute measures for a region, the measures with different time constraints have a close association and produce similar results in terms of relative measures among the regions.

Based on \( AC_{pi} \), all the study regions are sorted; \( THAC \) defines the percentage of the mean of the measure, i.e. \( u_{AC} \), such that a study region is considered as the area with low levels of accessibility if the corresponding \( AC_{pi} \) of the region is less than \( THAC \) multiplied by \( u_{AC} \). The specification of this parameter depends on the particular situations of the study city (e.g. the activity distributions relative to residential locations) as well as the inequality levels of accessibility across different regions. For a city with more evenly distributed accessibility, the variance of \( AC_{pi} \) is smaller, a higher value of \( THAC \) can be considered; whereas in an urban area featured with higher degrees of accessibility deviation, the variance of \( AC_{pi} \) is larger, a lower value of this threshold could thus be adopted. Furthermore, the setting of \( THAC \) is also related to the level of accessibility problems that a city aims to investigate, as different values of \( THAC \) for a same urban area lead to a different number of regions that are filtered out. The lower this threshold is, the less the number of the selected regions tends to be, and the worse the accessibility problems of these areas could be. Fig. 6 describes the distribution of \( AC_{pi} \) across all the 108 study regions, showing that 2, 7, 11, 22, 18 regions have less than 20%, 20–40%, 40–60%, 60–80% and 80–100% of \( u_{AC} \) (i.e. 6289). In this case study, \( THAC \) is set as 60%, as a result, 20 regions are pinpointed. These regions have the measures lower than 3612, accounting for only 57.4% of the average measure across all the study areas, capable of reaching maximum 22.5% of the total activities in the city. Fig. 7 depicts the geographic distribution of the detected regions, identified with the rank of the places assigned according to \( AC_{pi} \) in a descending order. It is noted that most of the detected regions are located outside the city center enclosed with the red oval, where most activities are established.

4.5. Specific land use and transport problem examination

The region with the lowest measure (i.e. \( rank = 1 \)) is examined, in order to demonstrate the typical problems of these detected areas. This region, referred as \( rp \), is located in the grid positions of 8×24 (see Fig.7), accommodates 137 activity locations and generates 30 trips in the morning per day. It reaches only 6.2% of all the activities within 30 min by car. The integrated-spatial-area, which is the spatial area of all the activity regions that have the average travel times from \( rp \) less than or equal to 30 min, is outlined in the large black polygon. A distance gap between the integrated-spatial-area and the city center is noted. Further examination reveals that the driving speed from \( rp \) towards its eastern direction is very low, e.g. only 16.9 km/h to the activity region \( rc \) located in this center. The slow movement hampers people quickly reaching the activity concentrated places. In order to overcome this problem, the traffic conditions between \( rp \) and \( rc \) is

---

**Fig. 5.** The correlation between \( AC_{mi} \) and \( AC_{pi} \) (a) and the distribution of the differences between them (b) Note: in Fig. a, the black spots and red stars represent \( AC_{mi} \) and \( AC_{pi} \) respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Fig. 6.** The distribution of \( AC_{pi} \) across all the 108 study regions. Note: The black curve represents the fitted normal density curve on the histogram, with the mean \( u_{AC} \) and the variance as 6289 and 3005 respectively. \( AC_{pi} \) on the x-axis is expressed with the percentage of \( u_{AC} \), ranging from 20% to 220%.

---
12.32 km; a driving speed higher than 24.64 km/h is thus required, if people are able to reach \( r_c \) within 30 min by car.

The overall accessibility problems are also manifested by the difficulties to reach many individual types. For instance, among all the 19 activity categories, this region has the lowest measure to 17 types, except ‘financial center’ and ‘public place’. In particular, regarding the ‘petrol station’, the measures are 0 under both the 30 min and 40 min thresholds while 2 with this parameter being extended to 60 min, suggesting that no petrol stations are reachable within 40 min. Among all the 147 stations provided by the city, 139 (94.6%) are outside the city and only 8 are scattered on the edge of the urban area as indicated with small black rectangles in Fig. 7. This problem also occurs to many other study regions, evidenced by the statistics that only 8 study regions (7.4%) are reachable to a certain station in 30 min by auto. Drivers in most parts of the city have to travel a long time in order to refill their vehicles.

5. Comparison of the results against a baseline model

To examine the performance of the proposed approach, we compare the results with those obtained under a baseline model. The baseline model considers the Euclidian distance between the centroids of the study and activity regions as well as the average travel speed between these places (\( v_{ij} \)) derived from the GPS data. Specifically, the expected travel times, i.e. \( e_{uij} \), are computed as follows.

\[
e_{uij} = e_{uij}(r_i, r_j, TimeP, DayT) = \frac{ED(r_i, r_j)}{v_{ij}}
\]

\[
v_{ij} = v_{ij}(r_i, r_j, TimeP, DayT) = \frac{\sum_{m=1}^{m_j} (v_{trip})}{m_j}
\]

Using \( e_{uij} \), the contour measure for \( r_n \), namely \( AC_{bi} \), is calculated according to formula 4.

Fig. 8a describes the correlation between \( e_{uij} \) and \( u_{ij} \) (a) as well as between \( AC_{bi} \) and \( AC_{mi}/AC_{pi} \) (b). Note: in Fig. a, the black stars and red points denote \( e_{uij} \) and \( u_{ij} \) respectively; in Fig. b, the black points, blue crosses and red stars represent \( AC_{bi} \), \( AC_{mi} \) and \( AC_{pi} \) respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
two regions, $e_{ij}$ is thus anticipated to be shorter. This characteristic is well reflected in this figure, where all $e_{ij}$ are shorter than their $w_{ij}$ counterparts. Furthermore, the extent of the differences (i.e. $e_{ij} - w_{ij}$) differs across the region pairs, with the average as 5.9 min while the minimum and maximum as 0.75 min and 19 min, respectively.

Fig. 8b demonstrates the deviations in the accessibility measures, showing that $AC_{b_i}$ is larger than $AC_{m_i}/AC_{p_i}$ for all study regions. Similarly, the degree of the differences varies across the regions, with the mean, minimum and maximum as 638, 0 and 2340 between $AC_{b_i}$ and $AC_{m_i}$ as well as 724, 0 and 2222 between $AC_{b_i}$ and $AC_{p_i}$ respectively. Thus, compared to the baseline model, the proposed method generates long travel times and low measures, as it is able to account for the actual routes that are decided based on various factors, e.g. route directness and traffic conditions.

6. The coverage issue

To examine the extent of the taxi GPS data in covering the road network, we conduct final analysis using the entire month of GPS data in September. In this process, all the distinct combinations between the study and activity regions which witness at least one trip in the morning a day, are first identified for each weekday. The derived results are then aggregated over multiple days, leading to two variables being generated, including the number of distinct region pairs $NumberOfODs$ and the number of considered days $NumberOfDays(NumberOfODs = 1...21)$. Fig. 9 describes the correlation between these two variables. The function $NumberOfODs = 6247 \times \log(1.57 + 1.09 \times NumberOfDays)$ is found to best fit the points, with the Pearson’s chi-squared test ($\chi^2$) as 0.0224 and p-value greater than 0.999. Based on this model, out of all the 26,028 study–activity region pairs, a minimum of 57.74 days is required to cover all these combinations. Thus, the 3 month of data used in our study is sufficient.

7. Discussions

7.1. Implications of the detected regions in the accessibility problems of the general road network as well as of the public transit system

This study explores taxi passenger trips, and it thus best characterizes taxi travel demand patterns and taxi availability in the city. However, as described in Section 2, in a typical Chinese city like Harbin, taxis play an important role in taking on the urban travel demand in the road network, e.g. 16.6% of the total personal travel by both taxis and private cars. That is, among the total vehicle flow for private trips on the roads, nearly 1 in 5 is taxis. Furthermore, the correlation coefficients between the number of passenger trips generated in the morning and attracted at night (i.e. $mo$ and $md$) from a region and the total number of activity locations of all types in the region, are 0.85 and 0.83, respectively. The high coefficients indicate a close link between the number of activity locations in a region and that of passenger trips it witnesses. Meanwhile, a region with more activity establishment is likely to attract not only taxi riders, but also people coming by other modes, e.g. private cars. All the above analysis suggests that the taxi travel demand patterns can represent the real travel demand for activities in the road network to a high degree, and that the detected regions by the taxi GPS data reflect road accessibility problems not only about taxi passenger travel but also concerning potentially large demand from private car users.

Moreover, while the proposed method analyzes accessibility by the auto mode and is applicable to the road network, it should be noted that, at least to a certain extent, the detected regions could also reflect problems with public transit systems (e.g. subways and buses) inside these areas. The regions identified with low accessibility measures are characterized with two major factors: high taxi demand and long travel times to reach activity locations by car. Regarding the taxi demand, the underlying reasons for people to take taxis, apart from the need for trips under strict time constraints (e.g. heading to an airport), include the availability to public transit systems as well as the reliability and comforts of taking the transit modes. Thus, the high taxi demand in these areas could signify a problem in terms of the public transit systems as well.

7.2. Paths for future research

The paper focuses on the development of a novel method to measure accessibility across different regions using taxi GPS data and on the examination of the effectiveness of the proposed method in identifying regions with low levels of accessibility in a city. It aims at shedding light on the potential and effectiveness of using GPS data for accessibility analysis, and this work can be situated as the first step towards this exploration. Further research is required to enhance the proposed method. First, the current study develops a model to explore the relationship between $NumberOfODs$ and $NumberOfDays$. However, the particular formula of this model depends on the specification of several other thresholds, including $Grid\_X$, $Grid\_Y$, $THu$ and $THup$. The higher values of these parameters would lead to more detailed and accurate measures, but calling for a larger GPS dataset and more observation days. This is especially true for the parameters $Grid\_X$ and $Grid\_Y$. While most of the existing accessibility studies employ a fixed zoning system that is identical to the one used for conventional travel data collection and travel demand modeling, the high spatial sensitivity of the GPS data allows more flexibility for the division of the urban area controlled by these two parameters. In this study, 1600 units are used, however, a finer partition and smaller units could be applied, depending on the level of details at which the accessibility is investigated. In the meantime, the resolution of the study units also depends on the size of the GPS dataset, as the amount of data in each unit affects the statistical significance and accuracy of the derived results in representing the true accessibility conditions of the area. A further investigation into how and to what extent the derive results are influenced by these two parameters along with the other thresholds would be important in providing a guideline for the parameter selection issue.

Secondly, this study use contour measures for the identification of regions with the lowest level of accessibility, the other distinguishable sub-category of location-based measures is potential measures, which use an impedance function, typically the negative exponential function, to model the declining attractiveness of activities to a destination region as travel times between the two regions increase (e.g. Anderson et al., 2013). However, similar to traditional contour measures, the existing methods only consider the single time point $u_{ij}$, disregarding $\sigma_{ij}$. This leads to a same measurement for two different region pairs that share an identical $u_{ij}$ but differ in $\sigma_{ij}$. However, according to the negative exponential function, the quantity of changes in the attractiveness resulted from the changes in travel times is not evenly distributed. The longer the travel times are, the more diminishing the attractiveness is. The two previously mentioned region pairs should thus be assigned with different measures; the pair, which has larger $\sigma_{ij}$ and therefore contains more trips in shorter times, should get a higher level of attractiveness and
consequently higher accessibility. The incorporation of $\sigma_{ij}$ into the calculation of potential measures is able to consider the dispersion of the travel time distributions and thus more accurately model the effects of travel times on the attractiveness of activity locations.

Thirdly, in the case study, only the accessibility in the morning of weekdays is analyzed. As described in Section 4.2, the travel times between a same pair of study and activity regions could differ between different time periods of the day or different types of days, leading to varied accessibility outcomes. Thus, it is important for further research to apply the developed method to the remaining time periods and types of day to examine the variation of accessibility situations. The variation in accessibility measures for a region could lead to the design of policies that are better tailored to the temporal factors. For instance, the opening and closing times of certain activity locations could be adjusted, taking into account the specific accessibility measures of the corresponding activity types in the corresponding time periods and corresponding types of days.

Lastly, while the detected regions are likely the areas with general low levels of road network accessibility as previously discussed, a high density residential region suffering from road accessibility problems may not be discovered by the proposed method, if the region is reached with fewer trips by taxis. This problem can be addressed by the integration of the current method with other types of modes, e.g. private cars, buses and/or subways, if GPS data from these sources is available. This combination would provide another powerful way of improvement in the current method, in terms of detecting all poorly accessible regions of the entire urban area over all these types of modes as well as across each individual type.

8. Conclusions

The achievement of good transport accessibility and equity in the distribution of urban services is one of the supreme goals for transport managers and urban planners. Alongside urban expansion, accessibility issues are becoming more imminent, particularly in many developing countries. To meet this increasing challenge, we have developed a method which measures accessibility to different urban activity services across the city by car. Compared with existing approaches, the proposed method is capable of providing more representative, objective, cost-effective, and temporal and spatial sensitive measures. The constant generation of the taxi GPS data also allows the derived results to be timely updated, catching up with the fast pace of urban land use development and population growth.

The developed method can be used to systematically analyze accessibility across an urban road network, as well as to identify regions with serious car-based accessibility problems, thus assisting policy makers in improving the present road network conditions. It can also be employed to assess the effect of realized land use or transport policies on accessibility, by comparing the measures derived using the taxi GPS data collected both before as well as after the implementation of these schemes. Furthermore, it can be applied to predict accessibility conditions for future policy scenarios, providing insights into how different designs would lead to varied impacts on accessibility and transport inequity. For instance, if the policy scenario involves the construction of new activity locations or the relocation of existing ones, the layout of the new planned locations could be integrated into the distribution of the existing activity services. Based on the combined activity distribution, accessibility measures can be computed using the proposed method.

On the contrary, if the policy considers the construction of new road infrastructure in a region or across several regions, e.g. a highway, the travel speed particularly in the building area would be mostly affected, leading to the decrease in travel times between this area and other parts of the city. The driving speed and travel time of the new road can be estimated based on its designed length and capacity. While the travel times of the existing roads, which would be replaced by the planned new infrastructure, could be computed based on the length of these roads as well as the present driving speeds revealed by the GPS data. The difference between the above obtained two travel times leads to the approximation of the amount of travel time reduction, which can then be utilized to update the existing travel time between the corresponding regions. Based on the obtained new travel times, accessibility for the considered policy scenario can be predicted using the proposed method.

Apart from the improvement on road network conditions as well as the evaluation on road transport plans, the proposed method can also be utilized to help enhance public transit systems. As discussed in Section 7.1, the high taxi demand for the detected regions could signal a potential problem in the public transit systems. A further investigation into the exact causes of the high taxi demand would be carried out, in the case of problems being identified with the transit systems, appropriate policy measures should be taken in order to enhance the conditions of the transit modes in these areas.

Besides the potential applications of the approach, this study also makes important contributions from a theoretical point of view. Particularly, a new accessibility measure, i.e. $AC_{pi}$, has been developed based on the detailed travel time distributions. The traditional measure, i.e. $AC_{mi}$, uses $u_i$ as the only time signature to compute accessibility, regardless of the variance of the travel times. This makes $AC_{mi}$ very sensitive to the selection of $TH_i$. In contrast, $AC_{pi}$ accounts for $\sigma_{ij}$ of the travel times and models the exact part of the trips that occur within $TH_i$, no matter to which value $TH_i$ is assigned. Thus, $AC_{pi}$ can more accurately mirror the true accessibility situations of the regions.

With more and more various urban vehicles being installed with GPS devices worldwide, this study along with its possible extension can be easily and widely transferable to the cities, thus paving a way for the development of a new, effective and cheaply realized accessibility analysis method that supports the urban land use and transport network development into a system that is fully integrated and easily reachable to activity services across the city.

Acknowledgements

The authors would like to acknowledge the support of McKinsey & Consulting Company Inc., Shanghai and Urban China Initiative (UCI) through the UCI Grant 2013.

References
