The Journal of Transport and Land Use http://jtlu.org
Vol. 12 No. 1 [2019] pp. 375–394

Identifying residential and workplace locations from transit smart card data

Yuan Tian
School of Transportation Science and Engineering
Harbin Institute of Technology
ytian.phd@hotmail.com

Stephan Winter
Department of Infrastructure Engineering, University of Melbourne
winter@unimelb.edu.au

Jian Wang
School of Management, Harbin Institute of Technology
wang_jian@hit.edu.cn

Abstract: Public transit is highly promoted worldwide to reduce traffic congestion. An evidence-based planning of stop locations and routes with regard to residential and workplace locations can reduce walking distances to transit and the number of transfers, which can improve service quality of public transport and thus increase ridership. This paper proposes a novel method of identifying residential and workplace locations from smart card data. The proposed method identifies relevant stops first and then refines their catchments to narrow down residential and workplace locations in three steps: defining constraints from the design of the public transport network, movement logic, and land use. In 2017, we tested the method using Beijing smart card data. The results show close to 69% residential locations inference rates and more than 72% workplace locations inference rates. The mean value of inferred areas is approximately 20% of the areas derived by traditional methods. Available data on alighting stops verify the inferred results at least for flat fare systems.

Keywords: Smart card data, public transport, location identification, transit OD

Article history:
Received: June 1, 2017
Received in revised form: March 6, 2018
Accepted: May 15, 2018
Available online: May 17, 2019

1 Introduction

The development and improvement of public transport is seen as an important approach to reduce traffic congestion and car dominated environmental pollution. Evidence-based planning of stop locations and routes with regard to residential and workplace locations can improve service quality of transit and thus lead to an increase in its ridership (Ceder, Butcher, & Wang, 2015). However, to know the travel demand of transit riders is a premise for this planning. Currently transit demand is obtained through transit surveys, automatic passenger counter (APC) system or through smart card data. Both methods get origin-destination (OD) data of stop-level granularity. The locations where transit riders start their travels, i.e., the places they are living or working, remain unknown. State-of-the-art methods assume

http://dx.doi.org/10.5198/jtlu.2019.1247
ISSN: 1938-7849

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

that these places are within walking range of the OD stops, which is a granularity similar to transit stop distances, and thus too coarse for transit stop planning. This paper, however, will suggest a novel method of reasoning that will refine the granularity substantially.

By overlapping the identified individual passengers' residential and workplace areas, the density of travel demand of the whole city can be obtained. The distances between identified travel demand areas and adjacent transit stops enable to identify redundancy and gaps at the transit route and transit network levels, thus providing the evidence to adjust the location of certain transit stops. Such inference could give transit planners a more accurate perspective on transport demand, thus improving transit planning and transit service levels.

The method is based on considering various constraints implied by the transit network. In detail, the approach suggests four steps: (a) identifying residential and workplace stops of an individual, (b) intersection of all residential and workplace stops of an individual, (c) considering not only the boarding stops, but also their contrast sets – the other stops nearby that have not been used for boarding, and (d) adding fine tuning from parcel-level land-use maps. This paper is going to prove the hypothesis that the transit stops from passengers' travel records enable to reconstruct possible residential and workplace locations with significantly finer granularity than traditional methods of walking ranges around residential and workplace stops.

In order to investigate the hypothesis, two weeks of 250,000 transit riders' smart card data from Beijing will be used for experimental verification. This dataset contains both boarding and alighting stop records. The experiment is designed in two parts. In the first part only, the boarding stops will be used to infer residential and workplace locations, and the alighting stops will be used for the verification of inferred locations. This part provides realistic evidence for those smart card systems where only boarding stops are recorded, i.e., flat fare systems (such as London's or Melbourne's). In the second part, both boarding and alighting stops are used in the reasoning process. The second part provides realistic evidence for those smart card systems where both boarding and alighting stops are recorded, i.e., distance-based fare systems (such as Beijing's or Singapore's). The experimental outcomes prove the validity of the reasoning method.

The remainder of this article is organized as follows. In the next section research background and related methods are reviewed. Then Section 3 introduces the methodology of residential and workplace locations identification. Section 4 uses the dataset from Beijing as case study for methodology verification. Section 5 is the results and discussion. Section 6 provides the conclusions drawn in this study and suggests future research directions.

2 Background

Traditional transit demand modelling starts with trip-generation and mode split steps. It obtains zone level transit demand from transit travel surveys (or, more recently, smart card systems), and then distributes demand at zone centroids or uniformly across a zone. However, both distributions accept an aggregation error. Furth, Mekuria, and SanClemente (2007) proposed a parcel-level modelling method that could improve the distribution. Parcel-level demand estimation begins with surveyed on-off counts at transit stops, and then allocates the counts to each individual parcel by measuring trip generation rates (considering land-use type, size and location factors). However, the parcel-level demand modelling involves some practical issues. First, some jurisdictions do not maintain or provide access to parcel-level population data, and where it is available it may be out-of-date or inaccurate. Secondly, travel survey data, which can obtain riders' residential or workplace locations, typically represents only a small sample of the population, introducing another inaccuracy of population demand modelling. And thirdly, on-off travel surveys cannot represent a continuous and dynamic transit demand.

Smart card automatic fare collection (AFC) is a popular fare-management tool that has been implemented in many cities around the world in recent years. Utilizing the detailed spatial-temporal travel information of smart card data is a new way to obtain transit demand. It is population-wide data rather than a sampled subset, collected continuously over time rather than at sampled points in time, and substantially more accurate about travel times and stops. It represents a large volume of revealed preference data that allows travelers' behavior to be modelled with higher accuracy than by using traditional survey data (Jánošíková, Slavík, & Koháni, 2014). However, even smart cards are either anonymous or not linking an address (parcel) to a particular usage, for the same reasons. Some cities use a distance-based fare implemented in their AFC systems such as Singapore (Zhong, Arisona, Huang, Batty, & Schmitt, 2014) and Brisbane (Tao, Rohde, & Corcoran, 2014). For the distance-based fare systems, passengers have to touch their smart cards on card readers both when checking in and checking out. Boarding information and alighting information are both recorded. This way, accurate stop-level OD matrices can be obtained. An alternative AFC system applies flat fares. In this case, transit riders just need to touch on their cards when boarding vehicles, and information about their alighting is not collected. Since the location of the boarding stop, and the travelled distance do not matter, only some flat fare AFC systems are integrated with AVL (Automatic Vehicle Location) systems, and only then can record transaction location when boarding. In order to obtain complete OD data from flat fare systems, some destination estimation algorithm are proposed (Chu & Chapleau, 2008; Nunes, Galvao Dias, & Falcao e Cunha, 2015; Trépanier, Tranchant, & Chapleau, 2007). In some cities AFC systems and AVL systems were designed separately, for example in San Diego (Munizaga & Palma, 2012) and London (Wang, Attanucci, & Wilson, 2011). Here, only the boarding time is recorded directly but without the boarding location. Only after integration with the existing AVL such flat fare AFC systems provide stop-level OD matrices with accurate origin location but with inferred alighting destinations.

However, for planning the location of transit stops, or for adjusting transit routes, the transit planners need OD data in form of residential and workplace locations of transit riders rather than their OD stops. Axhausen, Schonfelder, Wolf, Oliveria, and Samaga (2004) proposed for car drivers a method that could identify their location and purpose of trips' origins and destinations, by using GPS traces of vehicles. However, for transit riders, their walking traces from origin to boarding stop and from alighting stop to destination remain unknown as long as they do not voluntarily track themselves while walking, e.g., by mobile travel surveys (Cottrill et al., 2013). So generally the location of travel origin or destination cannot be obtained directly. Long, Zhang, and Cui (2012) proposed a residential and workplace stop inference method that uses smart card data, travel survey data and a land-use map. By using their method, individual transit riders' residential and workplace stops could be inferred, but where the passengers live or work remained unknown.

3 Methodology

This section develops the methodology of deriving residential and workplace locations from transit smart card data as a four-step-process: the identification of residential or workplace stops and their catchment areas (3.1), and the refinement of these areas in three steps (3.2-3.4).

3.1 Identifying transit riders' travel stops of residential and workplace

3.1.1 Trip chain construction

A trip chain is defined as a series of trips made by a traveler on a daily basis and is considered a useful way to demonstrate travelers' behaviors (McGuckin & Nakamoto, 2004). Ma, Wu, Wang, Chen, and Liu (2013) generated passengers' trip chains using Beijing smart card data. Based on the constructed trip chains, they applied a series of data mining approaches to extract the passengers' travel patterns and travel regularities.

Before identifying trip chains, original smart card records need to be converted into a format of individual daily trip chains. For flat fare smart card systems, only boarding information is recorded, so the format of the daily trip chain is:

Card ID,
Date,
Boarding time sequence (1st, 2nd… nth),
Boarding stop sequence (1st, 2nd… nth),
Route sequence (1st, 2nd… nth).

For distance-based smart card systems and subway smart card systems, both boarding and alighting information are recorded, so the format of daily trip chain is:

Card ID,
Date,
Boarding time sequence (1st, 2nd… nth),
Alighting time sequence (1st, 2nd… nth),
Boarding stop sequence (1st, 2nd… nth),
Alighting stop sequence (1st, 2nd… nth),
Route sequence (1st, 2nd… nth)

3.1.2 Residential and workplace transit stops identification of flat fare smart card systems

To infer residential or workplace locations by using smart card data, an individual's travel stops close to these locations need to be identified first. They will be called residential stops and workplace stops, respectively. For the average individual, the first boarding stop of each day has a high likelihood to be a transit stop near the rider's residential place. From 2005 Beijing household travel survey results, 99.5% of the boarding stops of the first trips are close to passengers' residential place (Long et al., 2012). As with other transport demand research, the presented model starts a day at 4am, given that many activities of a day can reach beyond midnight (Munizaga, Devillaine, Navarrete, & Silva, 2014). Some people have work patterns, such as night shifts, that lead to different travel patterns. This may result in reciprocal errors of identifying residence and workplace. As this section of the population is a small proportion, it will not have much impact on the overall travel demand analysis, and some of these misclassifications will be possible to be picked up in land-use analysis (3.4).

The presented method, but also other current methods accept that there can be more than one used transit stop around residence or workplace in general. These stops typically belong to different transit routes leading to different places. From the point of view of access to land and activities, other stops in the immediate vicinity, serving other directions and other routes, may also provide access to the same land uses and activities (Lee, Hickman, & Tong, 2013). For example, commuters often board at a fixed transit stop on weekday mornings to commute to work but may board other transit routes at weekends to go shopping or a restaurant. The collected different transit stops can be utilized for residential or workplace location inference.

Figure 1 illustrates the method of identifying residential and workplace stops from smart card data. First, the first boarding stops from each daily trip chain of the individual smart card are extracted. If more than one first boarding stop is identified, also their frequency is captured. The stop with highest ratio will be chosen as frequent residential boarding stop. Similarly, (Ma, Liu, Wen, Wang, & Wu, 2017) assumed that only the first trip and last trip of an individual transit rider for each day contribute to their commuting behavior. They counted the number of occurrences of each first trip's origin stop, the stop with the largest number of occurrences is considered the most frequent stop of residence. Similarly, the largest number of occurrences of each last trip's destination stop is considered the most frequent stop of workplace. The next step will utilize the identified frequent residential stop – which is the one most trusted – to filter out other residential stops from the first boarding stops. This step eliminates those (infrequent) occurrences of starting the day from somewhere else than usual. As residential stops should be within a walkable distance from riders' residence locations, the residence should be within the overlap area of the walkable distance ranges of these stops. All first boarding stops that overlap in their range with the frequent residential stop are classified as residential stops.

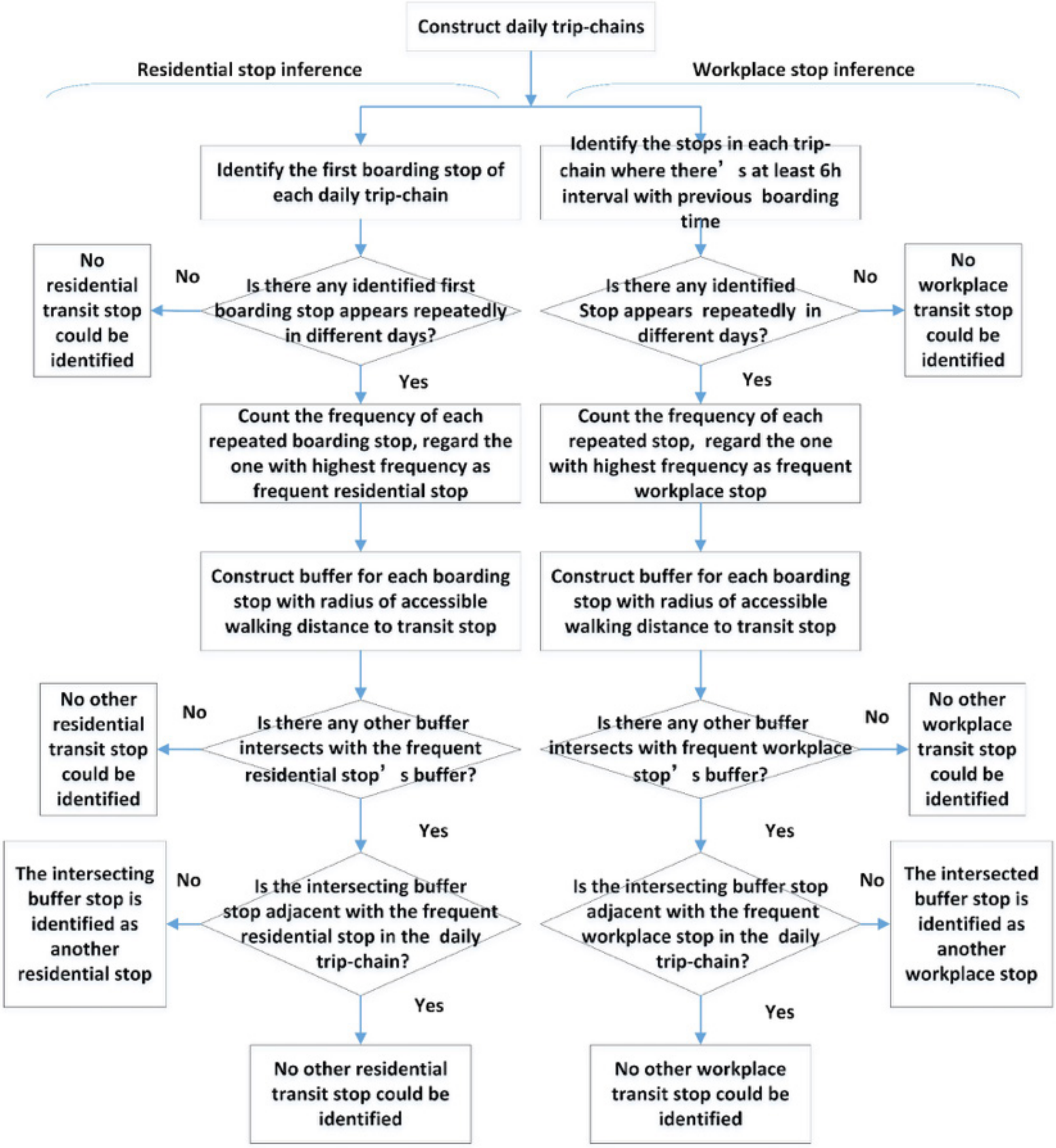
Construct daily trip-chains
Residential stop inference
Workplace stop inference
Identify the first boarding stop of each daily trip-chain
Identify the stops in each trip-chain where there' s at least 6h interval with previous boarding time
No
Is there any identified first boarding stop appears repeatedly in different days?
No residential transit stop could be identified
Is there any identified Stop appears repeatedly in different days?
No
No workplace transit stop could be identified
Yes
Yes
Count the frequency of each repeated boarding stop, regard the one with highest frequency as frequent residential stop
Count the frequency of each repeated stop, regard the one with highest frequency as frequent workplace stop
Construct buffer for each boarding stop with radius of accessible walking distance to transit stop
Construct buffer for each boarding stop with radius of accessible walking distance to transit stop
No other residential transit stop could be identified
No
Is there any other buffer intersects with the frequent residential stop' s buffer?
Is there any other buffer intersects with frequent workplace stop' s buffer?
No
No other workplace transit stop could be identified
Yes
Yes
The intersecting buffer stop is identified as another residential stop
No
Is the intersecting buffer stop adjacent with the frequent residential stop in the daily trip-chain?
Is the intersecting buffer stop adjacent with the frequent workplace stop in the daily trip-chain?
No
The intersected buffer stop is identified as another workplace stop
Yes
Yes
No other residential transit stop could be identified
No other workplace transit stop could be identified

Figure 1. Residential and workplace transit stops identification method of flat fare smart card systems

Walk access distance to transit stops is generally used for determining transit service areas. Service areas are used to help understand transit demand and identify gaps and redundancies in the existing transit network (El-Geneidy, Grimsrud, Wasfi, Tétreault, & Surprenant-Legault, 2014). Buffers at 400m around bus stops and 800m around rail stations are commonly used in the public transit research and industry to determine service areas that transit riders could access by foot (Biba, Curtin, & Manca, 2010; El-Geneidy, Tetreault, & Surprenant-Legault, 2010; Hess, 2009; Zhao, Chow, Li, Ubaka, & Gan, 2003). These ranges are supported by Daniels and Mulley (2013), who found the mean walking distance to bus stops is 461m, with a 75th percentile at 566m, and the mean walking distance to rail stations is 805m, with a 75th percentile at 1018m. Similarly, El-Geneidy et al. (2014) found the 85th percentile walking distance to bus services at 524m, and to rail stations 1259m by using detailed OD survey information of Montreal, Canada. Hoback, Anderson and Dutta (2008) performed Monte Carlo simulation in geographic information system with random addresses, the calculated walking distance from home to bus stop follows the street pattern is 0.36 miles (0.58 km) and average total walking distance is 0.8 miles (1.27 km) per round trip. However, the distances to transit stops are longer in Tehran, because in addition of walking, people often use local taxis with low fares to access transit services. Research found access distances to bus stops and subway stations of 1.5 km and 1.9 km respectively (Shirzadi Babakan & Alimohammadi, 2016). These results indicate that the walk access distance to transit service varies according to the different regions, the service being offered and the stop locations in the region. This paper will take 600m as bus stop walkable distance and 1000m as subway station walkable distance based on the specific circumstances of Beijing and in order to ensure that the buffers include all possible transit riders' origin and destination locations. Though this bigger buffer may overestimate the transit service areas at the first step of the method, the latter two steps will narrow down the initial areas.

Identifying workplace stops follows a similar process. This paper will identify workplace stops based on time difference between two consecutive travel records. Full time jobs typically cover a stationary period of about eight hours, although other work patterns and part time jobs may come with shorter periods of stationarity. For the particular area in the experiment, however, the 2005 Beijing household travel survey results show that 96% of the 27,550 respondents work over six hours. A threshold of six hours is also supported by literature (Long et al., 2012): If two consecutive boarding stops have a six-hour time interval, the latter one is identified as potential workplace stop. Identifying a frequent workplace stop first, then this one is utilized to identify other workplace stops with the process of residential stop identification.

3.1.3 Residential and workplace transit stops identification of distance-based smart card systems

For distance-based systems and subway smart card systems, both boarding and alighting stops can be recorded, so alighting stops can be used for further identification of residential and workplace locations. For residential stops identification, both the first boarding stop of a day and the previous day's last alighting stop can be identified from daily trip-chains. As Figure 2 illustrates, the first step is to judge whether these two stops are close to each other. By constructing walking access buffers for first boarding stop and last alighting stop, the buffers with overlap will be identified as residential transit stop pair. Then, calculating the frequency of identified stop pairs, the highest pair will be identified as frequent residential stops. Based on identified frequent residential stop pair, other residential stops could be identified. Constructing buffers for all of the stop pairs (the first boarding stop of a day and the previous day's last alighting stop) in the dataset, those intersecting with frequent residential stops are identified as potential residential stops. The last step will validate the other identified residential stops by using trip-chains. Only if an intersecting alighting stop is connected to the frequent residential stop by the same travel record, the alighting stop is discarded from the inference of residential locations.

For workplace stops identification, the process is similar with residential stops identification method, but now using the time difference between alighting stop and next boarding stop in trip-chains. Identifying the frequent transit stop pair supports then to find other workplace stops through walking access range overlaps.

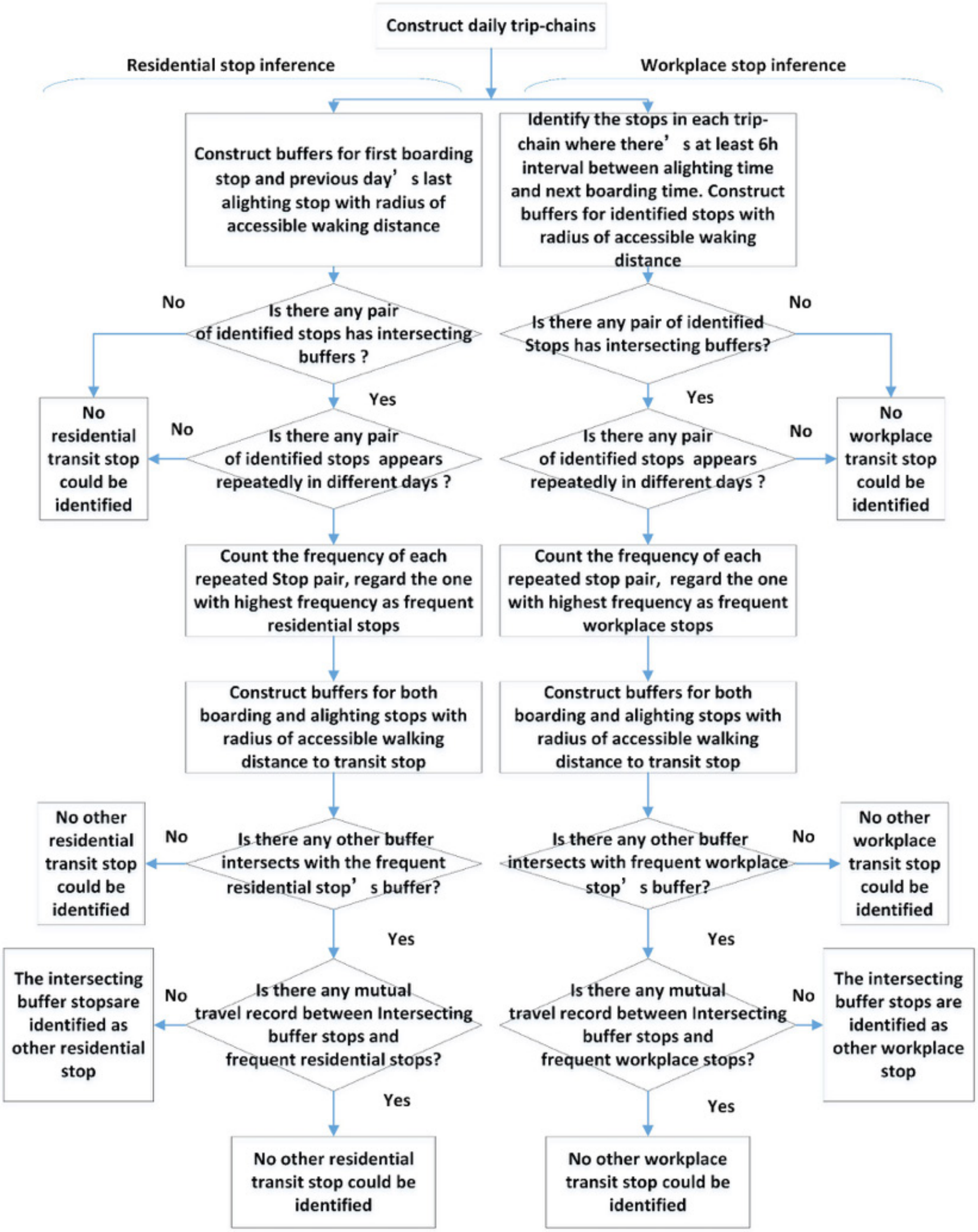
Construct daily trip-chains
Residential stop inference
Workplace stop inference
Construct buffers for first boarding stop and previous day' s last alighting stop with radius of accessible waking distance
Identify the stops in each trip-chain where there' s at least 6h interval between alighting time and next boarding time. Construct buffers for identified stops with radius of accessible waking distance
Is there any pair of identified stops has intersecting buffers ?
Is there any pair of identified Stops has intersecting buffers?
No
Yes
No residential transit stop could be identified
No workplace transit stop could be identified
Is there any pair of identified stops appears repeatedly in different days ?
Count the frequency of each repeated Stop pair, regard the one with highest frequency as frequent residential stops
Count the frequency of each repeated stop pair, regard the one with highest frequency as frequent workplace stops
Construct buffers for both boarding and alighting stops with radius of accessible walking distance to transit stop
No other residential transit stop could be identified
No other workplace transit stop could be identified
Is there any other buffer intersects with the frequent residential stop' s buffer?
Is there any other buffer intersects with frequent workplace stop' s buffer?
The intersecting buffer stopsare identified as other residential stop
The intersecting buffer stops are identified as other workplace stop
Is there any mutual travel record between Intersecting buffer stops and frequent residential stops?
Is there any mutual travel record between Intersecting buffer stops and frequent workplace stops?

Figure 2. Residential and workplace transit stops identification method of distance-based smart card systems

3.2 Initial residential and workplace location determination method by using identified multi-transit stops

The following process will consider a flat fare smart card system; a distance-based fare system has similar process. After the individual rider's residential (workplace) stops are identified from historical smart card data, the overlapping area intersected by walking access buffers of identified stops will be identified as residential (workplace) area.

In Figure 3, the red and blue points represent identified residential stops that are not adjacent stops in the daily trip-chains (red: a frequent residential boarding stop; blue: identified infrequent residential boarding stop from other first boarding stops). The overlapping buffer area (in yellow) is the initially inferred residential area.

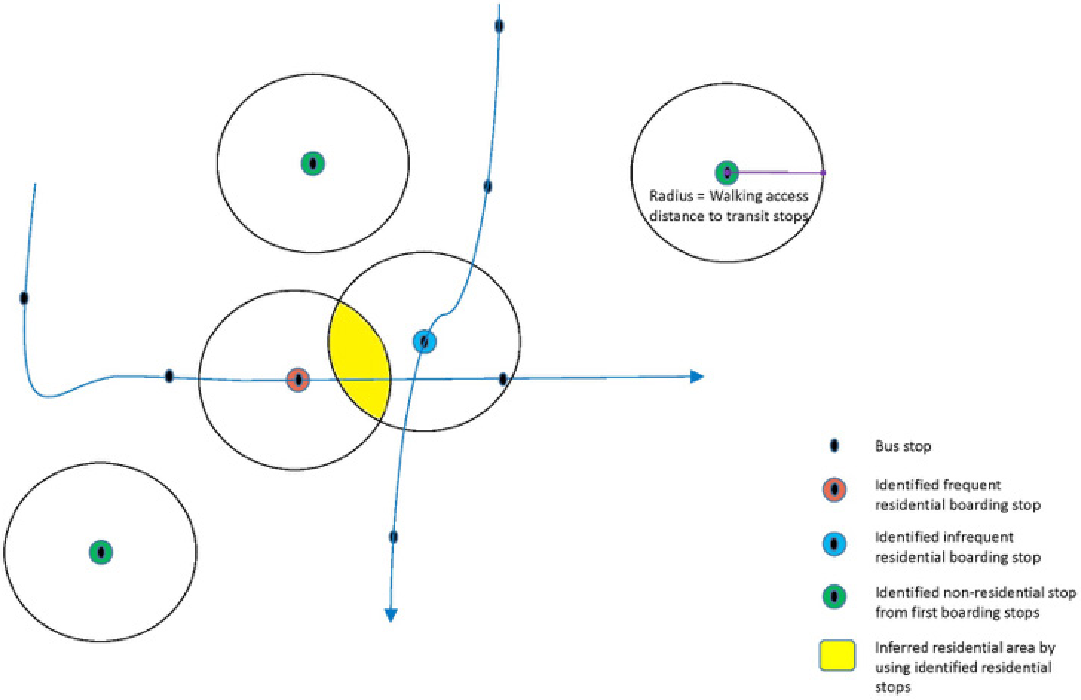
Radius = Walking access distance to transit stops
Bus stop
Identified frequent residential boarding stop
Identified infrequent residential boarding stop
Identified non-residential stop from first boarding stops
Inferred residential area by using identified residential stops

Figure 3. Residential (workplace) locations determination by using identified residential (workplace) stops

3.3 Further residential and workplace location determination by using adjacent stops of the same transit route

The previous step utilizes only the used transit stops to infer residential or workplace locations. In the next step, also the contrast set, i.e., the transit stops that riders do not use, are utilized for further inference. Transit riders board on one particular transit stop but not the previous or next stop of the same bus route because their residence or workplace is closer to this boarding stop. Thus, the previous stop and next stop of the residential or workplace stop can be a further constraint.

The constraint is realized by establishing the perpendicular bisectors between the identified residential (workplace) transit stops and their adjacent stops of the same route. From the perpendicular bisector to residential (workplace) stop and its adjacent stop of the same route have same distance. As the distance of segmentation, the areas from perpendicular bisector towards the residential (workplace) stop could contain the residence (workplace). The residential (workplace) location should be between the two perpendicular bisectors, thus the initially inferred residential (workplace) location will be narrowed down. Figure 4 illustrates how the initially inferred location (shown in Figure 3) is narrowed down; the yellow part is the updated residential location.

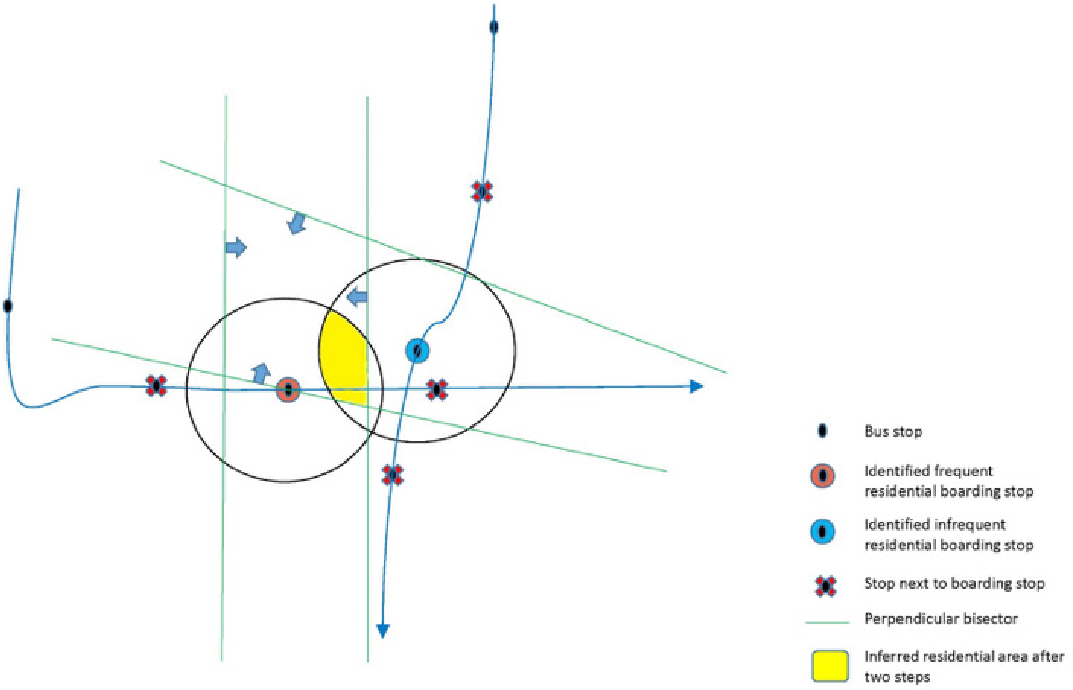
Bus stop
Identified frequent residential boarding stop
Identified infrequent residential boarding stop
Stop next to boarding stop
Perpendicular bisector
Inferred residential area after two steps

Figure 4. Narrowing down residential (workplace) locations by using adjacent stops as constraints

However, if the identified frequent residential (workplace) boarding stop and infrequent residential (workplace) boarding stop are belong to same route, and the stops are not adjacent in daily trip-chains, this situation will be seen that the transit riders live (work) close to both stops. For this situation, the algorithm will skip this step and go straight to section 3.4.

3.4 Final residential and workplace location determination by combining parcel-level land-use map

Parcel-level databases are built for taxation and land-use planning (Furth et al., 2007). Parcel-level land-use maps contain detailed information of land-use types, which can be utilized for further residential or workplace locations inference. Thus, the last step of the presented method combines the previous results with a parcel-level land-use map. The assumption is that residential locations should be within the parcels that show residential land use, such as residential or educational (to cater for students living in dormitories), and workplace locations should be within parcels that show work-related land use, such as industrial, commercial or government use.

4 Case study

4.1 Data description

The data available for case study comes from Beijing. The dataset consists of both bus data and subway data. Beijing, the capital city of China, had 22 subway lines and over 1000 bus routes by the end of

2017. Beijing Transit introduced a smart card system in May 2006. Since December 2014, all buses in Beijing accept the unified distance-based AFC system: transit riders need to touch on their smart card to contactless card reader devices when entering the bus and touch off when exiting. The bus drivers operate the electronic bus stop reporter system manually, such that when passengers touch on or touch off their smart cards, the boarding and alighting locations could be identified by combining the smart card system readings and the bus stop reporter system readings. The information recorded includes card ID, card type, transaction date, bus line, boarding stop, boarding time, alighting stop and alighting time. Similarly, subway AFC system store transit riders' full trip information.

The sample dataset consists of 250,000 different passengers' 7,283,866 records of two typical travel weeks from Monday October 9, 2017 to Sunday October 22, 2017. All the transit riders in this dataset have at least 16 total records and at least one bus travel record.

In addition, a parcel-level land-use map has been used, which was from Nov 2013. The map data consists of seven types of land use, as Figure 5 shows: residential land, green space, commercial establishments, transport facilities, education, government, and firms.

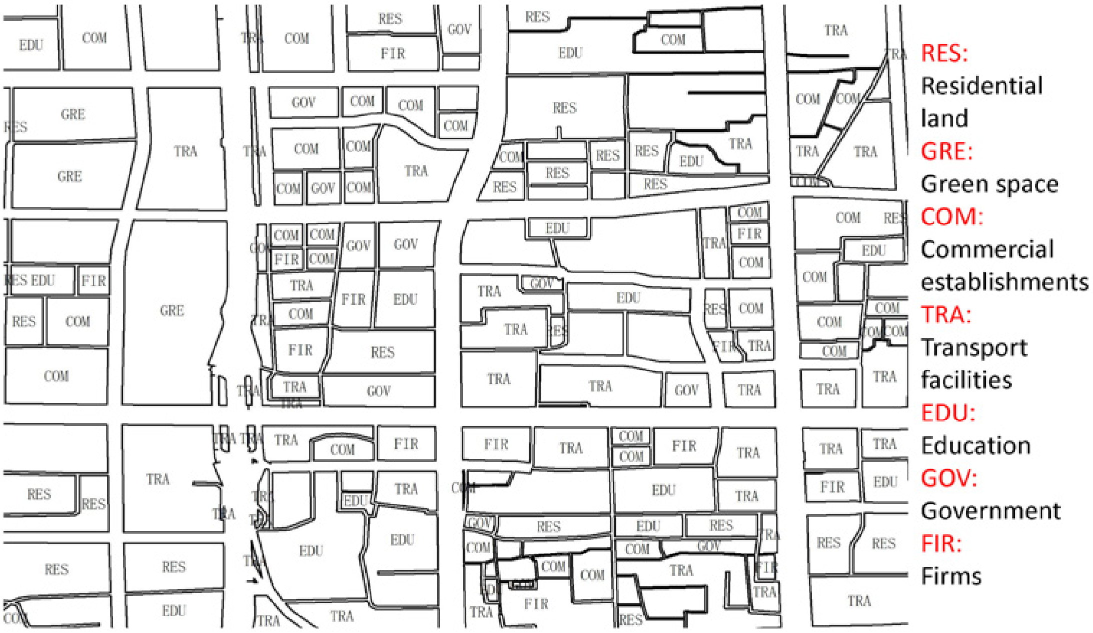

Figure 5. An illustration of Beijing parcel-level land-use map

4.2 Residential and workplace transit stop identification

To identify frequent transit travel stops of individual riders, multi-day data is required. Therefore, only the transit riders who have seven days or more recorded travels are used for analysis. Of these transit riders, only those with five boarding records of same bus stop are chosen (minimum value for frequent travel stop judgement). This leaves 159,356 transit riders for residential and workplace location inference. The method is applicable for the other riders as well if longer time spans are observed.

For flat fare smart card systems, only boarding records are used for inference, thus alighting stops are ignored in the first experiment. Identifying the first boarding stops in daily trip-chains and calculating the frequency of first boarding stops, the stop with highest frequency will be identified as frequent residential stop. 146,501 transit riders were identified that they have frequent residential stop. For workplace stop identification, the boarding stops that have 6h intervals with their previous boarding records in the daily trip-chains will be identified as potential workplace stops. The potential workplace stop with highest frequency of one transit rider in this dataset will be identified as frequent workplace stop. 89,529 transit riders were identified that have frequent workplace stop.

For the experiment on distance-based smart card systems, both boarding and alighting records will be used for inference. As Figure 2 shows, if buffers of a stop pair (the first boarding stop and the previous day's last alighting stop) have overlapping areas, the stop pair will be identified as residential stops. The most frequent residential stop pair will be selected as frequent residential stop pair. For workplace stops identification, the stop pairs (alighting stop, next boarding stop) that have overlapping buffer areas and a time interval between them of at least six hours will be identified as potential workplace stop pair, and those with the highest frequency will be selected as frequent workplace stop pair.

The identification of infrequent residential (workplace) stops is completed according to the processes of Figure 1 (flat fare systems) and Figure 2 (distance-based systems).

The identification results of residential stops and workplace stops of flat fare systems are shown in Table 1.

Table 1. The residential and workplace stops identification results of flat fare systems

	Residential stop identified results	Workplace stop identified results
No. of total transit riders who have frequent stop	146,501	89,529
No. of transit riders who have one identified frequent stop and no infrequent stop	52,597	40,375
No. of transit riders who have one frequent stop and one infrequent stop	66,007	38,650
No. of transit riders who have one frequent stop and two infrequent stops	20,863	9,230
No. of transit riders who have one frequent stop and at least three infrequent stops	7,034	1,274

4.3 Residential and workplace location determination

The experiment is executed using ArcGIS 10.1, geographic coordinate system is WGS 1984 and a projection to UTM Zone 52N according to the geographic location of Beijing. Figure 6 shows some of the inferred residential locations after the first step: The points are identified transit riders' frequent residential stops and corresponded other residential stops. The bigger circles are buffers with a radius of 1000m walk access distance to subway station, and the smaller circles are buffers of 600m bus stop walk access distance. Each overlapping area of intersecting buffers contains one transit rider's possible residential location.

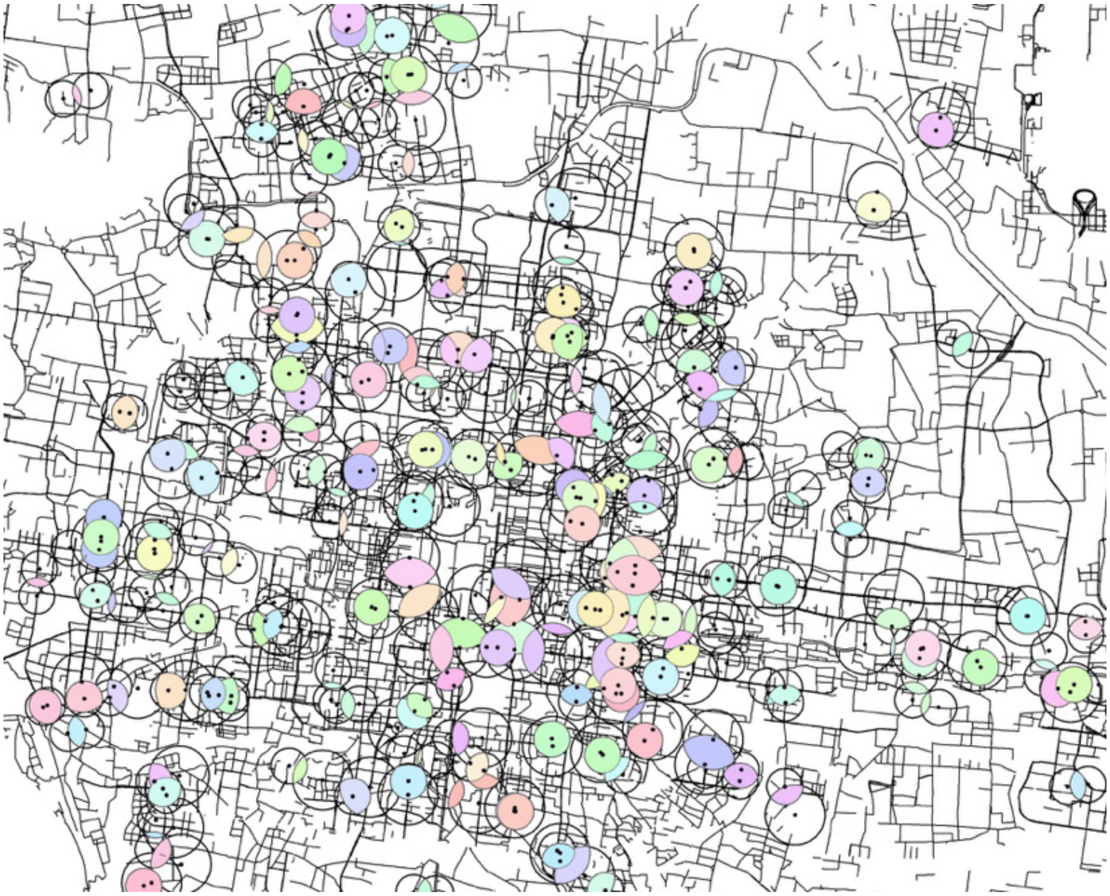

Figure 6. An illustration of the inferred residential locations by using identified residential stops

The previous stop and next stop of the same boarding route are utilized to narrow down the initial inferred areas. To establish perpendicular bisectors between stops, functionality in ArcGIS is used (Thiessen polygons): The identified residential (workplace) stop with its previous stop and its next stop are the three to be divided points, such that the Thiessen polygons consist primarily of two perpendicular bisectors derived by the three points. The area between two perpendicular bisectors contains the inferred residential (workplace) location.

Some transit riders are boarding or alighting at bus terminus, which means there is no previous stop or next stop of the identified bus stop. In these cases, only one perpendicular bisector can be established to narrow down the identified residential (workplace) stops. Figure 7 shows some results of inferred residential locations after the first two steps.

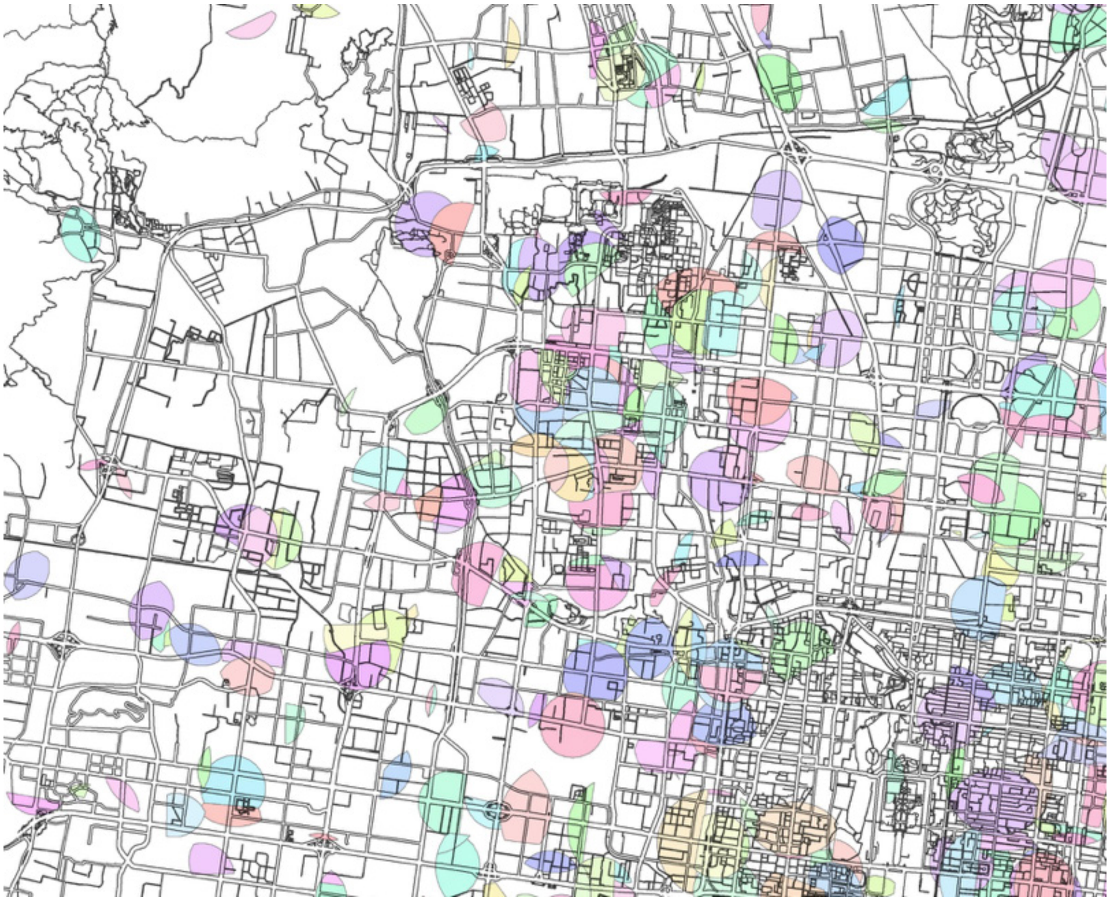

Figure 7. Parts of the inferred residential locations after two steps

At last, the parcel-level land-use map is used to intersect with previously inferred results. As some parcels' land-use property are missing values. When intersecting with residential (workplace) areas, the missing values were filled with residential(workplace) property values in order not to lose the final results. However, this missing value imputation method would also lead to a higher inferred results than the real situation. Figure 8 shows some of the finally inferred residential locations by only using the boarding records.

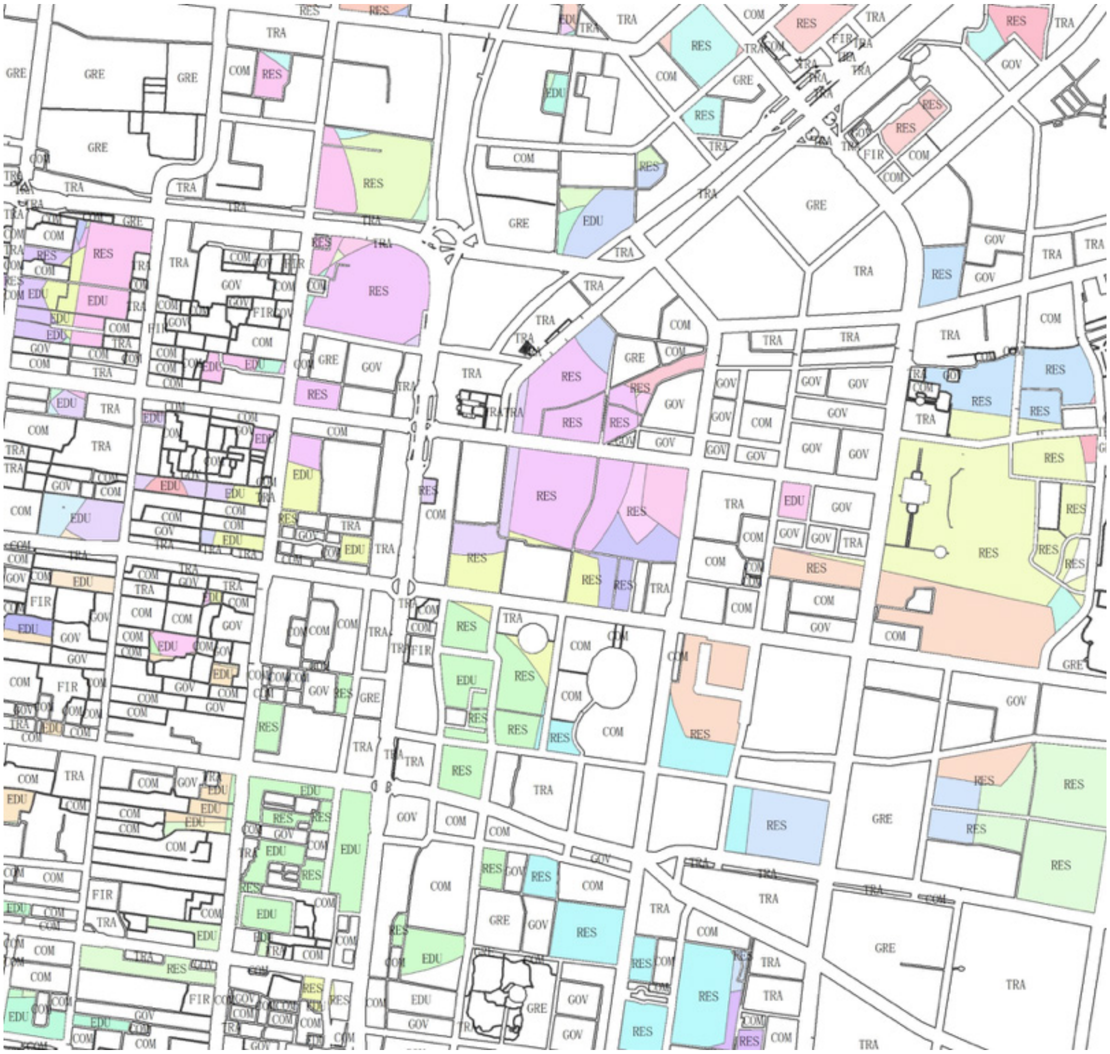

Figure 8. An illustration of the residential locations inference results

5 Results and discussion

5.1 Residential and workplace locations inference resul

Table 2 shows the identified results for flat fare systems. 159,356 transit riders had been filtered out for analysis. Of these, 146,501 transit riders have frequent residential stops, and 89,529 transit riders have identified frequent workplace stop.

For transit riders, whose frequent residential (workplace) stop can be identified from smart card records, 70.67% of 146,501 transit riders have inferred residential locations and 78.17% of 89529 transit riders have inferred workplace locations after the whole inference process.

Table 2. Residential and workplace location identification results of flat fare systems

Identification steps description	No. of transit riders whose residential location is identified	No. of transit riders whose workplace location is identified
Transit riders used for analysis	159,356	159,356
Residential (workplace) stops can be identified	146,501	89,529
After first step	146,501	89,529
After second step	130,047	77,926
After third step	103,535 (RES, EDU)	69,989 (COM, EDU, GOV, FIR)

Table 3 shows the inferred results of distance-based systems, for which both boarding and alighting records are used for inference. The same 159,356 transit riders are used for analysis. 104,165 transit riders can be identified meeting the conditions of having at least four counts of same residential stop pair (first boarding stop of the day and previous day's last alighting stop). And 62,704 transit riders have at least four counts of same workplace stop pair (alighting stop and the next boarding stop which have at least 6h interval in daily trip-chain).

Table 3. Residential and workplace location identification results of distance-based systems

Steps description	No. of transit riders whose residential location is identified	No. of transit riders whose workplace location is identified
Transit riders used for analysis	159,356	159,356
Residential (workplace) stops can be identified	104,165	62,704
After first step	104,165	62,704
After second step	92,830	59,372
After third step	71,727 (RES, EDU)	45,326(COM, EDU, GOV, FIR)

For transit riders, whose frequent residential (workplace) stops can be identified from smart card records, 68.86% of the 104,165 transit riders have inferred residential locations after the inference process, and 72.29% of the 62,704 transit riders have inferred workplace locations.

If just inferring from smart card data records by current best practice, the residential or workplace location of a transit rider is estimated to be within a 600m bus stop buffer area (1.13 km^2) or 1km subway station buffer area (3.14 km^2) around their boarding (alighting) stop. By using the inference method proposed in this paper, the inferred area is greatly narrowed down.

Figure 9 shows the inferred results of residential and workplace locations. The blue columns represent the inferred results of the residential locations, and the red ones are the workplace locations. For flat fare system inference results, the mean value of inferred residential area is 0.3356 km^2, only accounts for a proportion of 29.70% of initial 1.13 km^2(600m bus stop buffer area), and a proportion of 10.68% of initial 3.14km^2(1km subway station buffer area). The mean value of inferred workplace location is 0.3959 km^2, a little bigger than the inferred residential location. It accounts for 35.04% of initial bus stop buffer area and 12.60% of initial subway station buffer area.

For distance-based system, the inference results are smaller than flat fare system overall. The mean value of inferred residential area is 0.2674 km^2, only accounts for a proportion of 23.66% of initial 1.13 km^2 bus stop buffer area, and a proportion of 8.52% of initial 3.14 km^2 subway station buffer area. The mean value of inferred workplace location is 0.3565 km^2, accounts for 31.55% of initial bus stop buffer area and 11.35% of initial subway station buffer area.

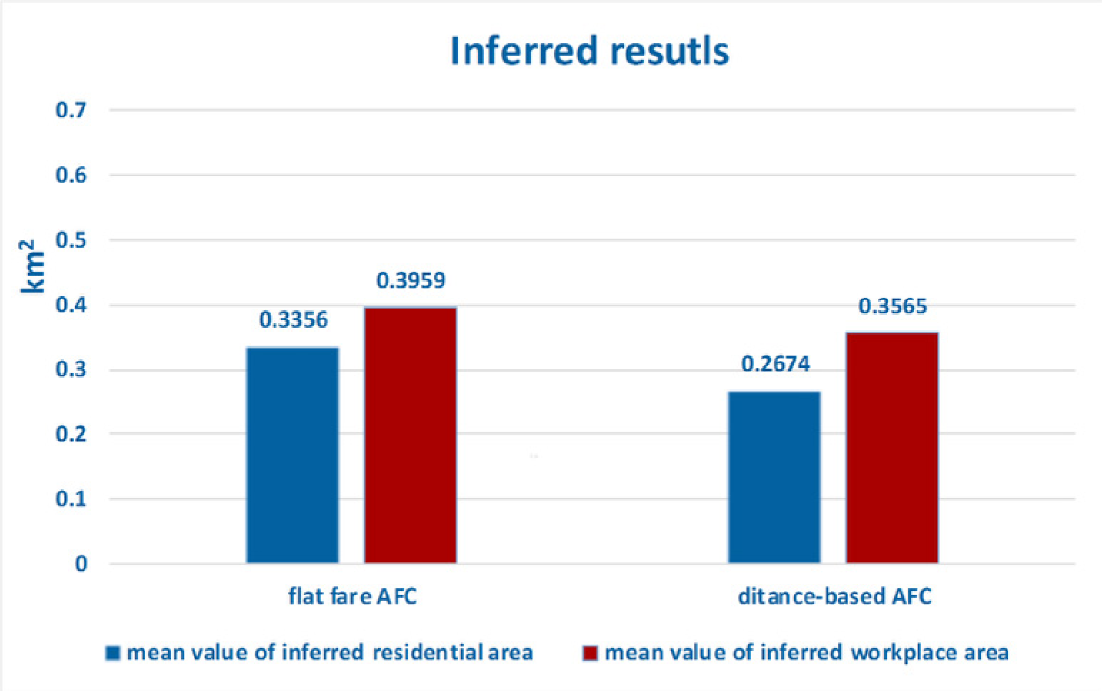

Figure 9. Inferred results of residential and workplace areas

5.2 Validation of inference results

To make the reasoning method persuasive, the inferred location results need to be validated. There is no detailed residential or workplace address for validation provided in the data set (for privacy reasons), so a validation method that is using the data itself is applied. For flat fare systems only, boarding records were used for location inference. Thus, alighting records can be used for validation. For this purpose, the residential and workplace stops were identified from alighting records for the 103,535 transit riders whose residential locations were identified and the 69,989 riders whose workplace locations were identified. The previous alighting stop of identified residential (workplace) boarding stop in trip-chains could be potential corresponding stop. Verifying the distance between identified residential (workplace) boarding stop and corresponding previous alighting stop with the distance value of 1200m if both are bus stops (twice of walk access distance). If both are subway stations the value is 2000m, and if one is bus stop and one is subway station the value will be 1600m. If distance between two corresponding stops more than the threshold values, a missing trip by other transport modes (such as taxi) will be identified.

After distance verification, some of the verified corresponding alighting stops are the same as the residential (workplace) boarding stops and some are different ones. Using the same stops to verify results will be meaningless, so only new appeared stops will be selected for verification. Constructing buffers with 600-meter radius for new bus stops and 1000-meter radius for subway stations. If the constructed buffer has an overlap with the corresponding inferred residential (workplace) locations, the inferred results will be seen verified.

Figure 10 shows parts of the verification results. The yellow circles are buffers constructed by new bus stops, the blue circles are buffers constructed by new subway stations and the red areas represent the overlapping areas that overlapped by constructed buffers and the inferred locations, which indicate that the results have been verified.

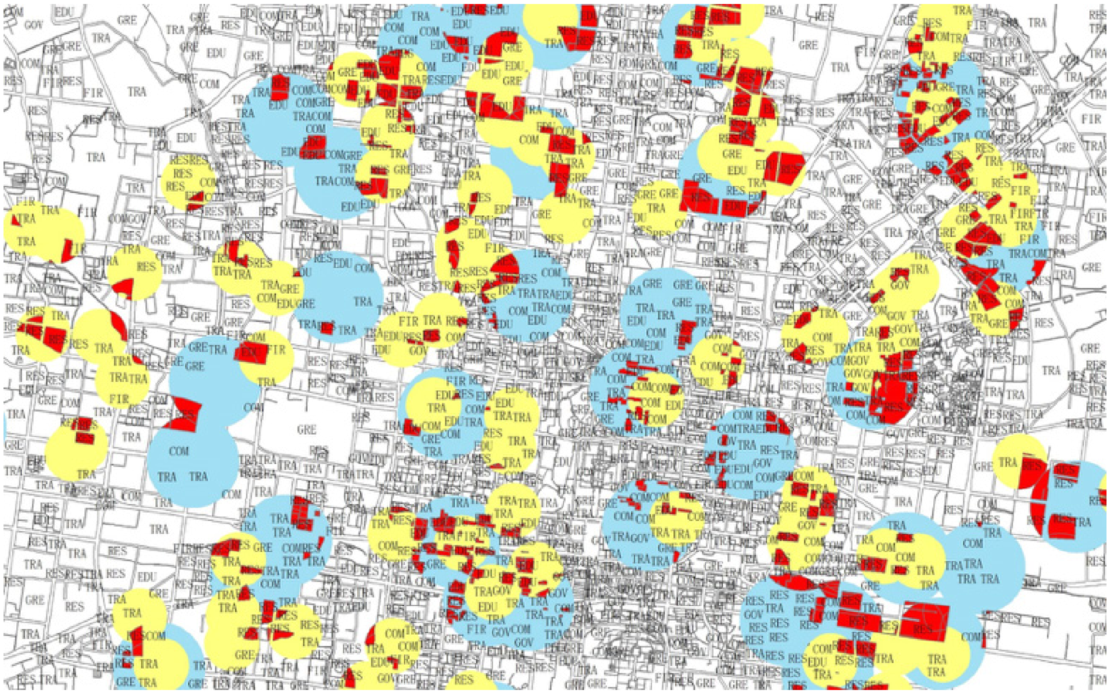

Figure 10. An illustration of the validation results obtained by using new appeared alighting stops

Table 4 shows the statistical results of validation. For transit riders, whose residential location is identified, 72,033 of them have different alighting stops corresponding to their identified residential boarding stops. By using these alighting stops for validation, 59,786 transit riders' inferred residential locations have been verified. For the 26,786 transit riders who have different alighting stops corresponding to their identified workplace boarding stops, 23,955 workplace locations have been verified.

Table 4. Verification of results

Steps description	No. of transit riders whose residential location is identified	No. of transit riders whose workplace location is identified
Transit riders who have new alighting stops	72,033	26,786
The inferred results be verified	59,786	23,955

5.3 Discussion

This paper proposed a novel reasoning method that can infer the residential and workplace locations of individual transit riders with significantly finer granularity than current best practice. Compared with transit demand surveys, which can obtain individual residential and workplace addresses but are generally limited by high costs and small sample sizes, the method proposed in this paper enables to obtain transit demand data from the total number of smart card users, including its variability over time and more comprehensive information with regard to travel times and locations. Compared to current best practice using smart card data, the method proposed in this paper infers residential and workplace locations instead of just stop-level OD or their areas of influence.

However, it should be noted that the initial experimental dataset is composed by 250,000 transit riders, but for only less than 42% of their final residential locations could be inferred, and for less than 28% the workplace locations could be inferred. These low proportions indicate that the experimental dataset of two weeks of data may not sufficient with regard to detecting regularity in travel patterns. Transit agencies, however, have years of smart card datasets, such that in practice the method should lead to significantly increased success rates. The experiment also used a parcel-level land-use map from November of 2013, but smart card data from October 2017. During the four years, some land use can be expected to have changed, impacting on the results.

The reasoning method applied Euclidean distances (circular buffers), however, Euclidean buffers may overestimate the service area of a stop compared to network buffers (El-Geneidy et al., 2014). Though network buffers are better approximations of the actual service areas, their accuracy is also limited by the level of map detail, especially for pedestrians. Off-street shortcuts and path inside parcels are often missing and may lead to underestimate service areas. While overestimated areas can be narrowed down by subsequent inference steps, underestimated survey areas may miss overlaps, and thus classifications as potential residential or workplace stops from the beginning. Since the used road map of Beijing was not detailed enough for pedestrian movement, we did not compare the impact of applying network buffers or Euclidean buffers.

6 Conclusions

This paper contributes a novel method of reasoning to identify where a transit rider lives or works by using public transit smart card data, aiming to refine the results of current methods significantly. The mean value of inferred areas is approximately 20% of the areas defined by current best practice methods. The new method, as much as the current best practice, requires data sets over longer periods of time in order to detect regular travel behavior. The inferred results, however, at least for flat fare systems could be validated by available data of alighting stops.

The reasoning method proposed in this paper can provide transit managers and planners a way to infer residential and workplace locations of transit riders. By further overlapping the population-wide coverage of these residential and workplace locations, some analysis such as travel demand hotspots analysis can be achieved, which will be a strong guidance to adjust transit network and relocate stops.

Future work will apply the reasoning method to datasets covering a longer period. With such data hotspots of residential or workplace travel demand can be identified and combined with actual transit network to do research related to transit adjustment, to identify redundancy and gaps at the route and transit network levels, to determine optimal stop spacing, and to improve transit service quality after analyzing passengers' travel behavior.

Acknowledgements

This paper was completed during the first author's visit at the University of Melbourne. This research was partly supported by the National Natural Science Foundation of China (General Program 51578199), and partly supported by China Scholarship Council (CSC) Program sponsored by the Ministry of Education in China, which are gratefully acknowledged.

Data availability

See supplemental files.

References

Axhausen, K., Schonfelder, S., Wolf, J., Oliveria, M., & Samaga, U. (2004). *Eighty weeks of gps traces, approaches to enriching trip information*. Paper presented at the Transportation Research Board Annual Meeting, Washington, DC.

Biba, S., Curtin, K. M., & Manca, G. (2010). A new method for determining the population with walking access to transit. *International Journal of Geographical Information Science, 24*(3), 347–364.

Chu, K. K. A., & Chapleau, R. (2008). Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record: Journal of the Transportation Research Board, 2063*, 63–72. doi: 10.3141/2063-08

Ceder, A. A., Butcher, M., & Wang, L. (2015). Optimization of bus stop placement for routes on uneven topography. *Transportation Research Part B: Methodological, 74*, 40–61.

Cottrill, C., Pereira, F., Zhao, F., Dias, I., Lim, H., Ben-Akiva, M., & Zegras, P. (2013). Future mobility survey: Experience in developing a smartphone-based travel survey in singapore. *Transportation Research Record: Journal of the Transportation Research Board, 2354*, 59–67.

Daniels, R., & Mulley, C. (2013). Explaining walking distance to public transport: The dominance of public transport supply. *Journal of Transport and Land Use, 6*(2), 5–20.

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., & Surprenant-Legault, J. (2014). New evidence on walking distances to transit stops: Identifying redundancies and gaps using variable service areas. *Transportation, 41*(1), 193–210.

El-Geneidy, A. M., Tetreault, P., & Surprenant-Legault, J. (2010). *Pedestrian access to transit: Identifying redundancies and gaps using a variable service area analysis.* Paper presented at the Transportation Research Board 89th Annual Meeting, Washington, DC.

Furth, P. G., Mekuria, M. C., & SanClemente, J. L. (2007). Parcel-level modeling to analyze transit stop location changes. *Journal of Public Transportation, 10*(2), 5.

Hess, D. B. (2009). Access to public transit and its influence on ridership for older adults in two US cities. *Journal of Transport and Land Use, 2*(1), 3–27.

Hoback, A., Anderson, S., & Dutta, U. (2008). True walking distance to transit. *Transportation planning and technology, 31*(6), 681–692.

Jánošíková, Ľ., Slavík, J., & Koháni, M. (2014). Estimation of a route choice model for urban public transport using smart card data. *Transportation planning and technology, 37*(7), 638–648.

Lee, S., Hickman, M., & Tong, D. (2013). Development of a temporal and spatial linkage between transit demand and land-use patterns. *Journal of Transport and Land Use, 6*(2), 33–46.

Long, Y., Zhang, Y., & Cui, C. (2012). Identifying commuting pattern of Beijing using bus smart card data. *Acta Geographica Sinica, 67*(10), 1339–1352.

Ma, X., Liu, C., Wen, H., Wang, Y., & Wu, Y.-J. (2017). Understanding commuting patterns using transit smart card data. *Journal of Transport Geography, 58*, 135–145.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies, 36*, 1–12. doi: 10.1016/j.trc.2013.07.010

McGuckin, N., & Nakamoto, Y. (2004). *Trips, chains and tours—using an operational definition.* Paper presented at the National Household Travel Survey Conference, Washington, DC.

Munizaga, M., Devillaine, F., Navarrete, C., & Silva, D. (2014). Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies, 44*, 70–79. doi: 10.1016/j.trc.2014.03.008

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research*

Part C: Emerging Technologies, 24, 9–18. doi: 10.1016/j.trc.2012.01.007

Nunes, A. A., Galvao Dias, T., & Falcao e Cunha, J. (2015). Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation. *IEEE Transactions on Intellegent Transportation Systems, 17*(1), 133–142.

Shirzadi Babakan, A., & Alimohammadi, A. (2016). An agent-based simulation of residential location choice of tenants in Tehran, Iran. Transactions in GIS, 20(1), 101–125.

Tao, S., Rohde, D., & Corcoran, J. (2014). Examining the spatial–temporal dynamics of bus passenger travel behavior using smart card data and the flow-comap. *Journal of Transport Geography, 41,* 21–36.

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems, 11*(1), 1–14. doi: 10.1080/15472450601122256

Wang, W., Attanucci, J. P., & Wilson, N. H. (2011). Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation, 14*(4), 7.

Zhao, F., Chow, L.-F., Li, M.-T., Ubaka, I., & Gan, A. (2003). Forecasting transit walk accessibility: Regression model alternative to buffer method. *Transportation Research Record: Journal of the Transportation Research Board,1835*, 34–41.

Zhong, C., Arisona, S. M., Huang, X., Batty, M., & Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science, 28*(11), 2178–2199.