The Journal of Transport and Land Use http://jtlu.org
Vol. 12 No. 1 [2019] pp. 201–220

Analysis of trip generation rates in residential commuting based on mobile phone signaling data

Fei Shi
Nanjing University
shifei@nju.edu.cn

Le Zhu
Nanjing University
mszhule@outlook.com

Abstract: In this paper, mobile phone signaling data are first processed to extract information such as the trip volume and spatial distribution from the starting point to the termination point. This information is then used to identify the residential and employment locations of users. Next, multiple Thiessen polygons based on cell towers are aggregated into Traffic Analysis Zones (TAZs) to minimize differences between the actual cell tower coverage and the theoretical coverage. Then, based on TAZ cluster analysis involving transport accessibility and commuting population density, multiple stepwise regression is applied to obtain the commuting trip production rates and attraction rates for overall residential land and each subdivided housing type during the peak morning hours. The obtained commuting trip generation rates can be directly applied to local transport analysis models. This paper suggests that as information and data sharing continue, mobile phone signaling data will become increasingly important for use in future trip rate research.

Keywords: Trip rate, signaling data, commuting trip, trip generation, trip production, trip attraction

Article history:
Received: July 17, 2018
Received in revised form: August 31, 2018
Accepted: November 27, 2018
Available online: March 28, 2019

1 Introduction

1.1 About trip generation

Trip generation is the first step in the conventional four-step transport forecasting process (followed by trip distribution, mode choice, and route assignment) and is widely used for forecasting travel demands. In this process, the number of trips originating at or destined for a particular Traffic Analysis Zone (TAZ) is predicted. Typically, trip generation analysis focuses on residences, and residential trip generation is a function of land use or the social and economic attributes of households.

The first zonal trip generation (and its inverse, attraction) analysis in the Chicago Area Transportation Study (CATS) followed the "decay of activity intensity with distance from the central business district (CBD)" thinking, current at the time. Data from extensive surveys were combined and interpreted on a-distance-from-CBD scale. For example, commercial land use in ring 0 (the CBD and vicinity) was found to generate 728 vehicle trips per day in 1956. That same land use in ring 5 (approximately 17 km or 11 mi from the CBD) generated approximately 150 trips per day (Meyer and Miller, 2001).

http://dx.doi.org/10.5198/jtlu.2019.1431
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.

Residential trip generation rate analysis is often performed using statistical regression. Person, transit, walking, and auto trips per unit of time are regressed based on variables thought to be explanatory, such as household size, number of workers in the household, persons in an age group, type of residence (single family, apartment, etc.), and other factors. Typically, regressions are performed at the aggregate/zone level. The data were collected by conducting a single day of peak-period person counts and vehicle counts at all doorway and garage entrances to each property, as well as surveying as many people as possible about their mode of access to or from the site.

1.2 Why to use mobile phone data?

Nowadays, the statement of Big Data Era's coming is quite common. Various data from information-based means such as Internet have become important resources for new economic interests and technological innovation. The mobile phone data is one of the most important types in all sorts of big data. Over the years, with the rapid development of communication technology and the increasing openness of mobile operators to data, mobile phone data has been drawing more and more attention. And it can be applied in many fields such as urban planning, transportation planning, smart city and traffic management. Undoubtedly, mobile phone data will become an important tool for urban planning, construction and management. But overall, the current research on mobile phone data is relatively insufficient. It's necessary to deeply find the essence behind phone data which can be used for urban and transportation research. This article will use mobile phone data to explore trip generation rates and be expected to obtain more satisfactory results.

1.3 Why to focus on this topic?

Ellys and Reid (1982) cautioned the application of United States' official trip generation rates to cities in developing countries. The cities in developing countries have unique growth and travel patterns that are different from those in developed countries (Wilfred, et al., 2015; Filitowish, 2011; Tilsdor, 2012). Developing countries require trip rate indices tailored to their own cities instead of simply adopting existing indices from developed countries. Additionally, because it is difficult for trip rate indices for a country, province or even metropolis to accurately represent the trip rate characteristics of a specific area, targeted surveys and analyses should be conducted to obtain actual trip rates (Wilfred, et al., 2015).

The urbanization rate in China will soon reach 60%. Rapid urbanization spatially expands cities, and residential areas require re-planning to accommodate large numbers of migrant workers. The rebuilding of shanty areas is another important Chinese government action required to improve the quality of human housing and life. In 2017, the state council executive meeting hosted by Premier Li Keqiang made a decision to implement a 3-year (2018~2020) shanty area rebuilding program; this strategic breakthrough will rebuild 15 million units in various shanty areas and positively affect up to 100 million residents. The urbanization process and this extensive shanty area rebuilding scheme will significantly alter the distribution of urban residential land and urban traffic patterns. Therefore, investigating the trip rates of residential areas is extremely important for in-depth traffic analysis, decongestion strategies, and other relevant research (e.g., Traffic Impact Assessment).

2 Literature review

In this section, the existing research findings are reviewed from various perspectives, such as the factors that influence the trip rate, trip rate manuals and relevant analysis methods.

2.1 Influencing factors

Trip rate studies normally define a mathematical relationship between the trip volume and the influential characteristic parameters. A physical factor such as land use represents one type of influential factor. The land use and intensity affect the trip volume and selection of destinations. When land use category variables are selected as independent variables, the equation is highly explanatory and easy to understand. The most classic example is the trip rate manual published by the Institute of Transportation Engineers (ITE). The trip generation informational report published by ITE provides trip generation rates for numerous land use and building types. Wu et al. (2013) proposed adding accessibility, as an independent variable, into the trip demand model because undoubtedly, location and accessibility are important influential factors (Soltani et al., 2012; Agyemang-Duah, 1997).

Non-physical factors include socioeconomic and demographic characteristics. The main related factors include income, age, car ownership, possession of a driver's license, the number of family members, household structure, the number of employees, and disadvantaged groups, such as the disabled or elderly (Berki and Monigl, 2017; Roorda et al., 2010; Jan-Dirk Schmöcker et al., 2005; Pitsiava-Latinopoulou et al., 2001; Agyemang-Duah, 1997). Population and employment factors are alternative representations of land use.

Some studies have considered both types of influential factors mentioned above (land use and socioeconomic characteristics), e.g., trip rate studies in Log Angeles and Portland (Lee, 2016; Tian and Ewing, 2017). Ewing et al. (1996) reviewed previous findings and suggested that the trip rate is affected by accessibility and the population density (Safwat and Magnanti, 1988; Hanson, 1982; Landrock, 1981); however, they also suggested that, at least theoretically, accessibility and the population density are correlated, i.e., areas of high density can be less accessible than areas of low density, and the total trip rate could decrease as the population density increases.

2.2 Related manual

Some works related to this study are Trip Generation Manuals. The most well-known related work in the western world is the 10th edition of Trip Generation. ITE procedures estimate the number of trips entering or exiting a site at a given time (sometimes the combined number entering and exiting is estimated). ITE rates are functions of type of development, gross floor area, number of gas pumps, number of dwelling units, or other standard measurable aspects, usually included in site plans. These functions are typically in the form of Trips=a+b*Area or Trips=a+bln(Area), and they do not consider location, competitors, complementary factors, the cost of transport, or many other potentially important factors. They are often estimated based on very few observations (a non-statistically significant sample). Many localities require their use to ensure adequate public facilities for growth management and subdivision approval. Many US agencies rely on ITE's approach as a defensible method for assessing the impacts of new development (Clifton et al., 2015; Bochner et al., 2011). Ohlms (2016) suggested that ITE rates do not include agri-tourism land trip rates; therefore, a survey and analysis were conducted in Virginia as a supplement. Dey et al. (1994) proposed that ITE rates be treated as prior information and that the trip rate be updated based on effective local trip generation data via the Bayes statistical method. The most significant advantages of this method are that limited local data can be fully utilized, the sample volume is moderate and updated data effectively reflect the local reality. Influenced by some North American experts (Schneider et al., 2013; Clifton et al., 2013; Dock et al., 2015), recent versions of ITE rates have continuously improved, e.g., individual-based trip rate analysis was added (Currans, 2017). Notably, ITE rates originate from trip survey to single building or site, which is significantly different from 4-step trip generation forecasting, such as discrepancies between techniques.

In China, well-known works include the *Trip Rate Manual of the Beijing Transport Institute* (BTI) published at the beginning of this century. This manual was based on trip survey data for approximately one thousand buildings with a variety of land uses in the central portion of Beijing. Similar to ITE rates, BTI rates also apply the unitary regression equation for a single land use to obtain the corresponding trip rate, and the correlation coefficient R_2 ranges from 0.5-0.8. BTI rates provide trip rate ranges and the average trip rates based on limited surveys; however, determining how to select a proper trip rate from the recommended range for a given site plan is not specified.

2.3 Regression analysis

Linear regression analysis, such as multiple linear regression for various socioeconomic and demographic characteristics or various land uses, is often used in trip rate analysis (Wilfred et al., 2015). However, Mayer and Miller (2001) suggested that in socioeconomic and demographic characteristic-based multiple regression, multicollinearity between explanatory variables, such as income and automobile ownership, could lead to incorrect analysis conclusions. To obtain a simple regression equation and effective regression parameters, some studies have employed multiple stepwise regression (Quintero et al., 2016; Ko, 2013; Pan, 2008; Quintero et al., 2016; Arabani and Amani, 2007).

Nonlinear regression has also been conducted in previous studies. For instance, an ordered probit model (Lim and Inivasa, 2011; Roorda et al., 2010; Jan-Dirk Schmöcker et al., 2005; Schmöcker et al., 2005) was employed to analyze the effects of factors such as age, family structure, transport modes and the disabled population on the trip rate. An ordered response model was employed to forecast trip rates in Toronto and achieved satisfactory results (Bwambale, 2015; Agyemang-Duah, 1997). Other studies employed neural network models, compared accuracies of common models in forecasting trip rates during hurricane evacuations and found that the forecast results based on logistic and neural network models were most accurate (Wilmot and Mei, 2004).

Relevant reviews have provided details of regression analyses in the western world (e.g., Currans, 2017). In China, research on obtaining trip rates via regression analysis is still in an early stage, and prominent issues include the lack of considering the significant of results or explanatory variable multicollinearity.

2.4 Classification analysis

In addition to analyzing trip rate without consideration of different purposes, some studies have analyzed various trip purposes, including commuting, school attendance, tourism, shopping and entertainment (Berki and Monigl, 2017; Wilfred et al., 2015; Tian and Ewing, 2017; Lim and Inivasa, 2011; Quintero et al., 2016; Agyemang-Duah, 1995, 1997), because the trip purpose has a strong influence on the travel destination and mode choice, which are analyzed in subsequent steps of the modeling process (Everett, 2009). Other studies have analyzed trip modes, e.g., analyses of the trip rate for walking (Lee, 2016; Tian and Ewing, 2017). The classification analyses used were primarily based on data source characteristics and the research objective (Ewing et al., 1996).

A cross-classification method has also been used. For example, when transportation planning was conducted in London, UK, in 1963, there were 6 categories of family income, 3 categories of automobile ownership, and 6 categories of family structure, resulting in 108 family classifications. In the study, the average trip rate was calculated for each category. This is a common analysis approach in textbooks, and the advantage is that the effects of socioeconomic characteristics on trip rates are minimized. At the beginning of this century, Everett (2009) attempted to use this method, which requires highly detailed data obviously.

2.5 Commuting trip rate analysis based on mobile phone data

Analyses of commuting trip rates have been conducted in many studies. Moreover, analyses of home-based-other (HBO) trip rates and non-home-based (NHB) trip rates have been included in various studies (Berki and Monigl, 2017; Wilfred, 2015). When the trip purpose is known based on questionnaire survey data, it is easy to obtain trip rates for various purposes. When using mobile phone signaling data, it is difficult to obtain a detailed trip purpose; however, commuting trip rates can be analyzed after determining home and workplace locations via mobile phone signaling data. Mobile phone signaling data, which includes the passively recorded spatiotemporal trajectories of millions of users, have recently emerged as promising inputs for transport analyses. However, mobile phone signaling data are mostly used in transportation detection, origin-destination (OD) matrix analysis, population statistical analysis, urban expansion boundary identification and urban passenger flow corridor identification (Çolak et al., 2015; Demissie et al., 2016). Currently, literature on the application of mobile phone data in trip rate analysis is relatively scarce. Bwambale et al. (2017) analyzed the trip rates of various groups (gender, work state and age) using an ordered response logit model and a hybrid model with mobile phone data.

2.6 Summary and discussion

Undoubtedly, trip generation manuals in the United States and China have guided more in-depth research and innovation in this research field in the two countries. However, because calculating the trip rate involves the collection, processing and analysis of data from a limited number of selected cities and single buildings, trip rates normally vary over a wide range. In actual applications, it is difficult to select a proper value in the range. Additionally, some regression equations have low fitting degrees but are still recommended (Shoup, 2003). If trip rates are used directly without considering the local trip characteristics, the result could deviate, possibly quite significantly, from the actual situation. Everett (2009) investigated the transferability of trip rates and suggested that applying the same trip rates for cities with different scales could lead to significant error. Mayer and Miller (2001) suggested that ITE rates should be applied to local transport planning projects after modification.

Socioeconomic and demographic characteristic-based analysis requires relevant data. However, in China, this type of data is difficult to obtain and significantly changes annually, particularly in the case of income, number of employees, automobile ownership and driver's license ownership. Therefore, analysis and research methods from the western world cannot always be appropriately applied to Chinese cities. Although some studies have suggested that the number of employees is a better independent variable than the floor area (Macababbad and Regidor, 2009), in China, it is easier to obtain floor area data via building surveys, and the availability of floor area data allows researchers in China to conduct trip rate research.

As suggested by Ewing et al. (1996), trip rates are affected by accessibility and the population density. Such a theory concludes that high accessibility results in a high level of desire to travel in North American cities. However, despite what was suggested by Ewing et al. in the literature, in China, accessibility and the population density may not be correlated. Many high-density areas in Chinese cities have high accessibility via public transport but low accessibility via automobiles. Therefore, it is very difficult to conclude that overall, the accessibility (a weighted score based on the trip mode shares for public transport and automobile accessibility) and population density have a significant correlation.

Mobile phone signaling data represent a great opportunity for trip rate research because researchers will not be constrained by the limited amount of trip data from conventional manual surveys. Additionally, the survey workload and cost are significantly reduced. For example, in the United States, data collection costs can easily exceed the annual budget of a metropolitan planning organization (MPO) in a small or medium-sized area (Bwambale,2017). However, current mobile phone location data cannot provide point-level precision; locations are identified by cell towers, which is a challenge for subsequent research. Each cell tower radiation range involves multiple buildings and blocks with different land use categories. Therefore, unitary regression is not applicable, and analyses of trip rates for different land use types can only be based on multiple regression analysis. Compared with nonlinear regression, linear regression has more intuitive explanatory capability and is more convenient. Therefore, multiple linear regression is conducted in the following sections.

3 Data and methodology

3.1 Study area

In this paper, the central city area of Kunshan in Jiangsu Province is selected as an example. Kunshan, a county-level city with an urban area of 928 km2, a population of 1.66 million and a central city area of approximately 150 km2, consistently ranks No.1 among the top one hundred most developed counties in China. Notably, the city is adjacent to Metropolitan Shanghai. In 2017, Kunshan's GDP per capita was 32 thousand US dollars, or 1.7 times the level in Shanghai.

3.2 Mobile phone data processing

3.2.1 Mobile phone signal data cleaning and preprocessing

4G mobile phone signaling data have numerous advantages, such as supplying sufficient information with high reliability and real-time availability. Mobile terminals connect to cell towers periodically or randomly and actively or passively. These connections are recorded by the cell tower and become mobile phone signaling data. Mobile phone signaling data primarily include the unique identifier of a mobile phone, time, cell tower number and event type (outbound call or inbound call, power on or power off, texting, getting access to Internet, etc.). However, due to the defects of the mobile network and external interference from the natural and built environment, source location data contain large amounts of "noise", which significantly affects location data extraction and subsequent analysis. Therefore, errors and abnormal data should be identified and removed to obtain high-quality mobile phone signaling data that can be effectively utilized.

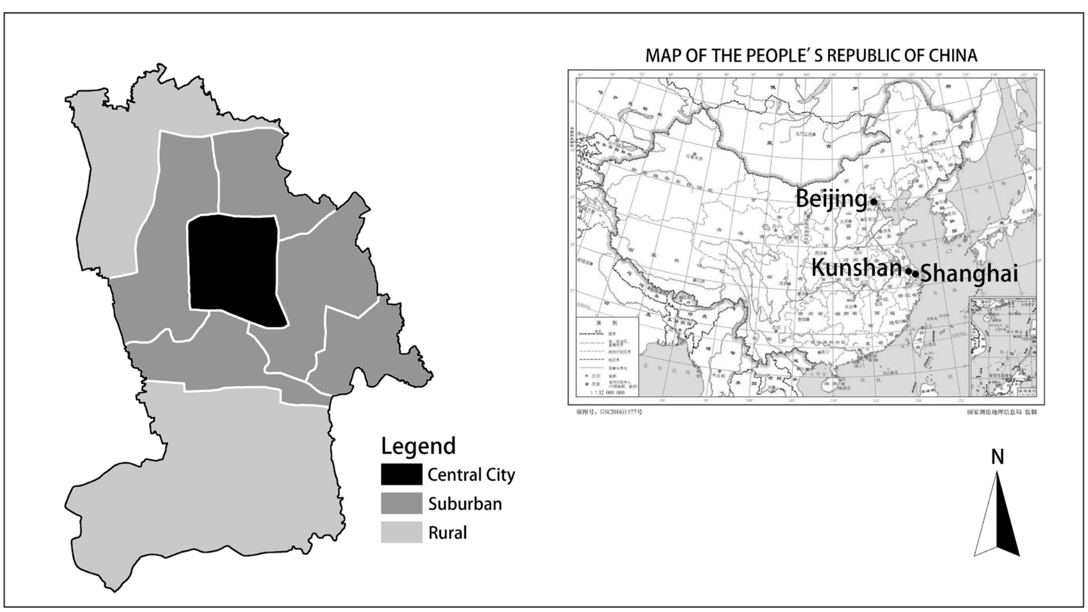

Figure 1. Map of Kunshan

This work, in practice, will need programming and cost plenty of time. Unsteadiness of cell tower signals, probably due to built environment and cell tower setting, easily results in Ping-pang Effect (or Table Tennis Effect), i.e., mobile signal switches back and forth in two adjacent cell towers. Although the service coverage of a cell tower is approximately 600–1000 meters, the phenomenon of signaling drifting sometimes happens, i.e. a mobile may connect to a further cell tower rather than the nearest cell tower. However, we can identify these two scenes mentioned-above through programming, and then clean up raw data.

In this paper, approximately 30 days of 4G mobile data from China Mobile (the largest mobile communication operator in China) are used. The data were collected in May and June 2017 in the central city of Kunshan and were distributed among 1296 cell towers. The original data set included an average of 2500 records of individual signaling data daily. After data preprocessing and cleaning, there were approximately 120 effective data records on average to provide supportive data for the subsequent precise identification of individual trip characteristics. Preprocessing raw data by deleting redundant data and records is a significant step. For instance, many mobile signaling data and records (r_1,r_2,…r_n) will be created when the mobile is located in the same coverage of a cell tower for several hours. However, we just need r_1 and rn to estimate time consumption of a mobile or a person in the service coverage of that cell tower. So the records from r_2 to r_{n-1} can be deleted due to their uselessness for trip rates analysis. It is easily understandable that the number of records and signaling data needed in this study decreases significantly after data preprocessing. 15min threshold value, which is acceptable in mainland China, is used to isolate stay and pass-by areas.

Notably, (1) a record is a series of information or data recorded by cell tower when a mobile is being used. A record can be understood as a parameter set including parameters such as International Mobile Subscriber Identification Number (IMSI), Location Area Code (LAC), cell tower ID, longitude and latitude of cell tower, event type, beginning time, ending time and so on. (2) Types of events include outbound call or inbound call, power on or power off, texting, get access to Internet, etc.

- Besides texting, outbound call or inbound call, getting access to Internet via mobile can produce records including location and time too. In fact, records can be created when beginning and ending, sending and receiving data during the whole process of getting access to Internet.
- Records can be created when location updated or the cell tower affiliated by a mobile changed.
- Mobile terminals can automatically connect to cell tower when no records have been created for over 40 minutes. Therefore, new location and time information are recorded.
- Records will be created when Wechat, Gaode Map or other APP in mobiles are used.

Records can be created due to different events when a mobile is in the same service coverage of cell tower. So there are much lengthy data. Frequency of recording data reaches an average of one item per 20 seconds if users get access to Internet frequently.

3.2.2 Residential location and employment location identification

In a residential area, most people are sleeping from 24:00 ~ 7:00. During this period, the cell towers mobile phones attach to are relatively constant. This characteristic can be used for residential location identification; however, there are two potential scenarios.

In the first scenario, a residential area may be adjacent to two or more cell towers. During the nighttime, the mobile phone may select different cell towers, which is known as cell tower hopping. Therefore, during the overnight period, if the time attached to a single cell tower exceeds 2 hours, the residential location can be identified.

In the second scenario, according to statistics (Zhou, et.al., 2017), approximately 20% of users will power off their mobile phones when they sleep. Therefore, the mobile phone data are missing for this group during the overnight period. To accurately identify the residential locations of members of this group, the criterion is set based on whether the power-off location is the same as the power-on location. Specifically, if the distance between the cell tower that received the last signal on the previous day and the cell tower that received the first signal on the next day is less than 800 m, then the resident is at the same location or lives at this location.

To identify employment locations, in this paper, the typical workday hours of 9:00–12:00 and 14:00–17:00 are selected. If the user attaches to a particular cell tower for the longest portion of this period and the daily average attachment time exceeds 3 hours, then this cell tower is identified as the work place. For those whose employment and residence locations are the same or whose work location is not fixed due to their job type, the associated travel behaviors do not follow conventional commuting patterns, and the proportion of such individuals is small; therefore, this scenario is not considered.

3.2.3 Extraction of commuting trip data during morning peak hours

Based on the previous criteria for residence and employment location identification, the commuting trip data from the residence location to the employment location during morning peak hours are extracted to create the commuting OD Matrix during morning peak hours and form a basic data set for trip rate research in this paper.

3.3 Preparation before regression analysis

3.3.1 Aggregation of cell tower coverage to TAZ

Only when cell towers are mapped to urban geographical spaces can the spatial movements of mobile phone users be obtained and the associated trip parameters be estimated. Cell towers are deployed based on the density of users. The coverage radius is approximately 100–500 m in urban areas and approximately 400–1000 m in suburban areas (Li, et. al., 2016). However, because cell tower coverage is affected by topography and buildings, it is difficult to accurately determine the coverage space. Therefore, in this paper, a TAZ is defined as the study unit, and each TAZ contains coverage areas for multiple cell towers.

A detailed description of the method is as follows. (1) In ArcGIS, Thiessen polygons are easily generated based on cell towers. (2) Each TAZ boundary should coincide with the boundary of cell tower Thiessen polygons to facilitate trip volume counting. Additionally, TAZs with irregular shapes, such as long strips, bends or horns, should be minimized as much as possible. (3) Each TAZ should contain 5 or more Thiessen polygons based on cell towers to reduce error in identifying accurate cell tower coverage. In regions with low cell tower densities or in suburban areas, the number should not be less than 3. The number of Thiessen polygons in the same TAZ is a little hard to decide because there are two major considerations. The first is that we hope there are more Thiessen polygons in the same TAZ to eliminate the error of service coverage of cell towers as much as possible, which means the number of TAZs decreases. The second is that we need more TAZs to conduct cluster analysis and regression analysis in the following analysis process. Therefore, we have to select a moderate number of Thiessen polygons in the same TAZ and get a reasonable number of TAZs; more than 80 TAZs may be suitable in this study for subsequent analysis. Finally, in this paper, the coverage areas of 1296 cell towers are aggregated into 98 TAZs (see Figure 2). The TAZ division method used in this paper is different from the traditional method, and because it is based on cell towers, it is called TAZ-CB (cell based).

3.3.2 Cluster analysis based on transport accessibility and the commuting population density

Based on previous experience, accessibility and population density may affect trip rates, two indices, transport accessibility and the commuting population density, are analyzed and calculated in this paper. Based on the subsequent results, the TAZ-CBs in the research scope are subjected to cluster analysis to identify location characteristic differences and obtain more accurate trip rates.

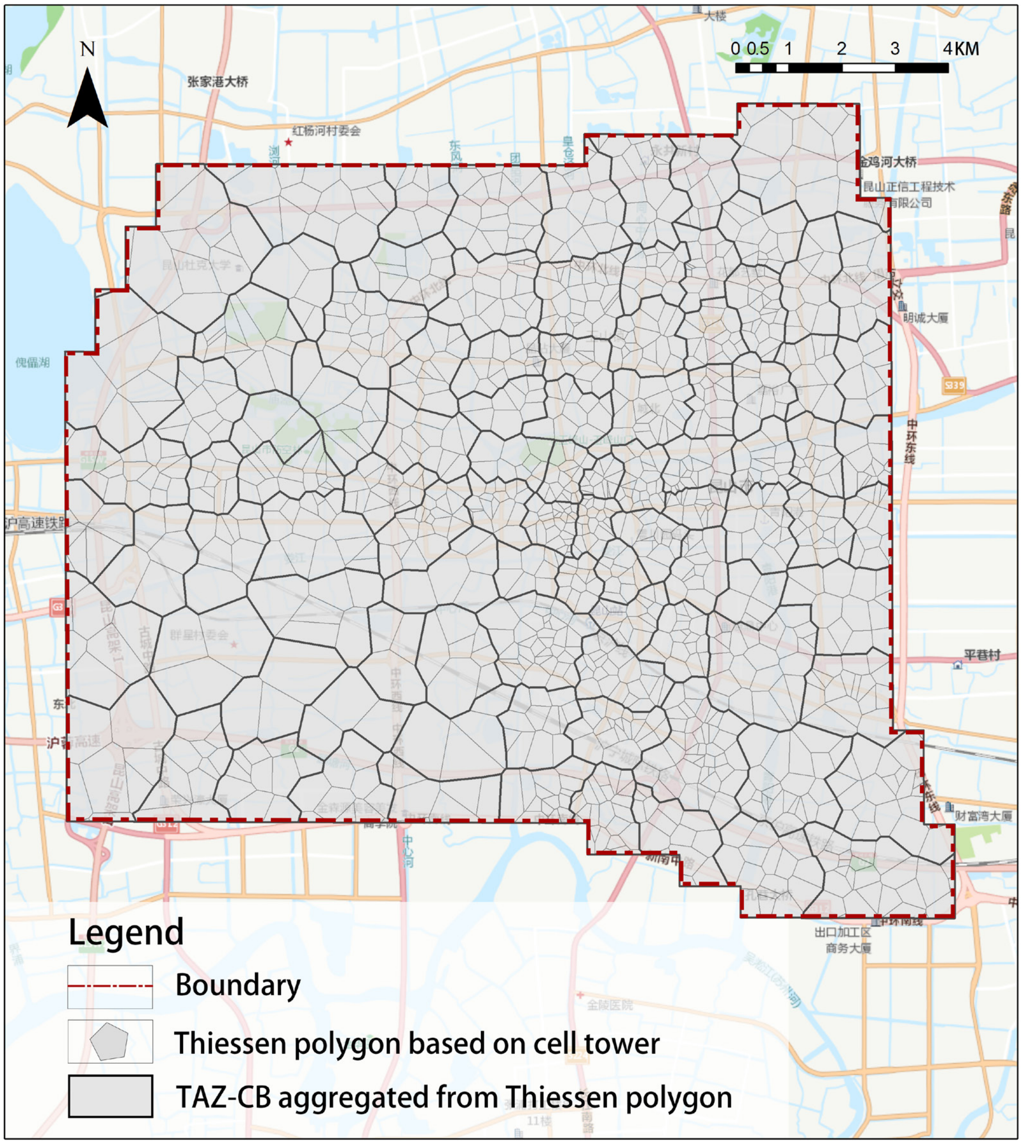

Figure 2. Cell tower coverage and TAZ-CB diagram

Under normal conditions, a transport location advantage results in a high volume of resident activity. In this paper, a Python program calls the path planning API of the map provider, Gaode Map or Amap, to obtain public transport and automobile accessibility information(Moreno-Monroy et al., 2017; Singleton, 2014), i.e., the time consumption for each trip between TAZ-CB centroids. Then, based on the trip mode share (public transport is 14.8% and automobile is 29.7%) in the central city areas of Kunshan in 2017, a weighted comprehensive transport accessibility score is obtained for subsequent analysis (see Figure 3).

In general, a high commuting population density results in high trip rates per unit area of residential buildings. Because this study focuses on commuting trips, the commuting population density is selected. The commuting population density is defined as the number of commuters divided by the TAZ-CB land area in which commuters are located (see Figure 4).

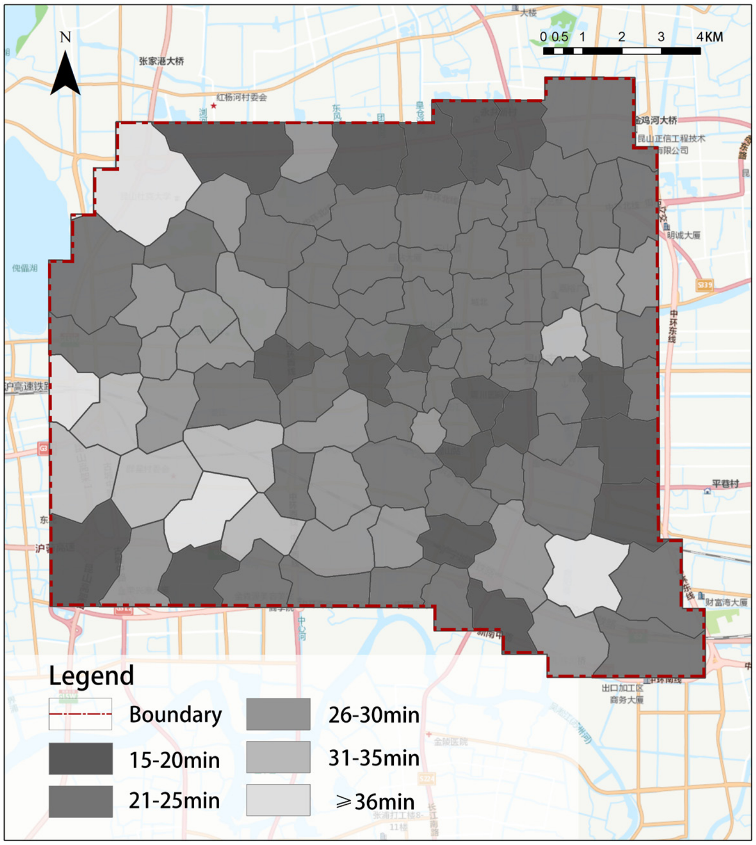

Figure 3. Transport accessibility diagram

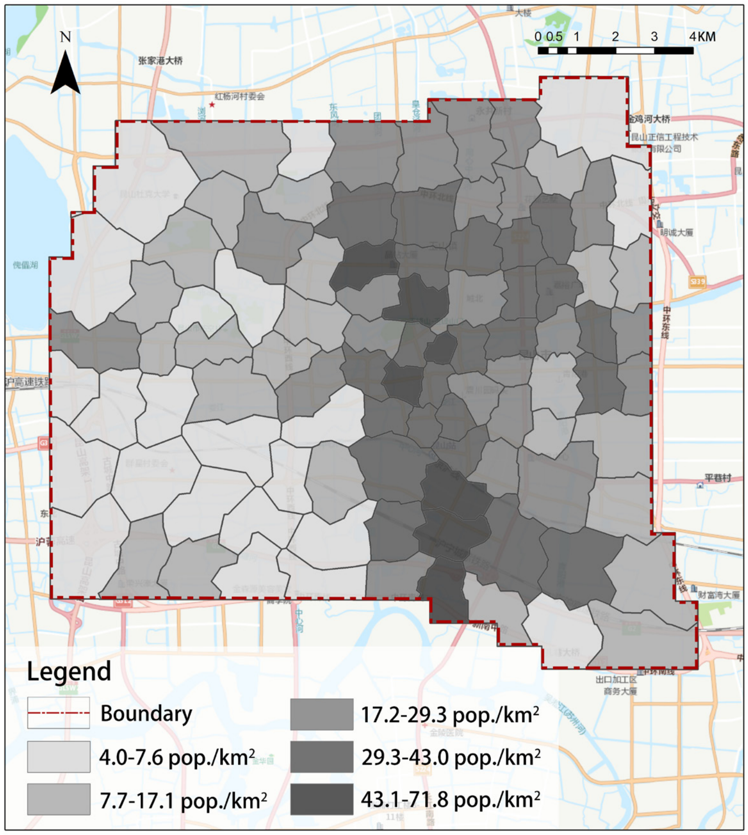

Figure 4. Commuting population density diagram

In this study, floor area data are based on digitized Kunshan land and architecture census data from 2014. However, because the mobile phone signaling data were collected in 2017 and the residential areas in some TAZ-CBs were still under construction or did not exist in 2014, some TAZ-CBs were removed in the subsequent analysis.

Constrained by the number of TAZ-CBs and the requirements of the number of the regression equations, cluster analysis in SPSS software defined only 2 classifications. The clustering results are as follows. The TAZ-CBs characteristics of Classification One include superior accessibility and a higher commuting population density, and 54 TAZ-CBs are included in this group. The TAZ-CBs characteristics of Classification Two include inferior accessibility and a lower commuting population density, and 31 TAZ-CBs are included in this group (see Table 1). Because land census data and mobile phone data were collected in different years, 13 TAZ-CBs with substantial residential building development or no residences since 2014 were removed from this study (see Figure 5).

Table 1. Cluster information

		Two TAZ-CB Classifications	
		Classification One	Classification Two
Characteristics		Superior accessibility, high commuting population density	Inferior accessibility, low commuting population density
Number of TAZ-CBs		54	31
Cluster center	Commuting trip accessibility (minutes)	20.88	26.45
	Commuting trip population density (thousand persons/km^2)	2.76	1.79

Note: 13 TAZ-CBs with substantial residential building development or no residences since 2014 were removed from the analysis.

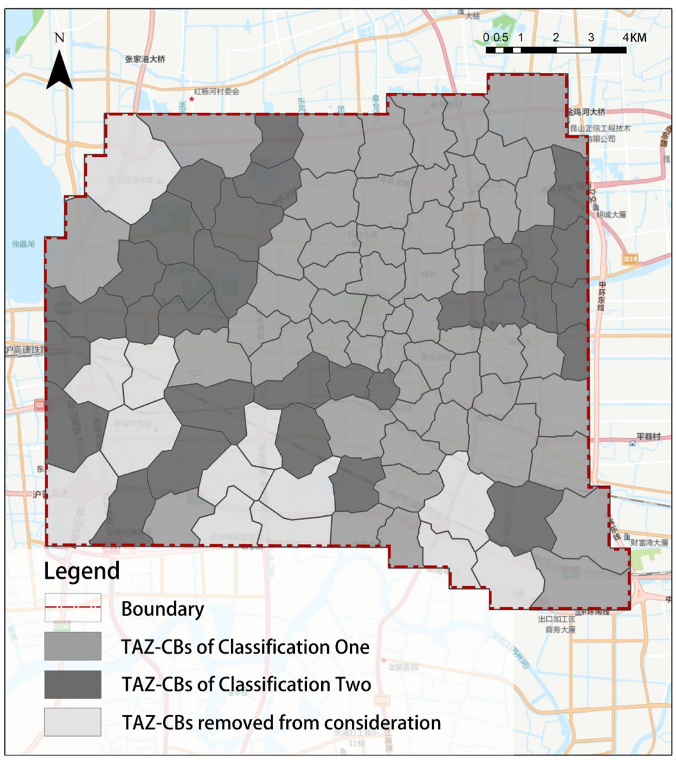

Figure 5. Diagram of cluster analysis results

3.4 Model specification

We employed the multinomial stepwise regression model, which has been widely used to study trip generation, in this research. The general form of a trip generation model is as follows:

$$T_i = f(x_1; x_2; x_3; \ldots; x_i; \ldots; x_k)$$

where *xi* values are prediction factors or explanatory variables. The most common form of the trip generation model is the following linear functional form:

$$T_i = a_0 + a_1x_1 + a_2x_2 + \ldots + a_ix_i + \ldots + a_kx_k$$

where *ai* values are the coefficients of the regression equation and can be obtained by regression analysis.

The core idea of stepwise regression is to introduce variables into the model one by one. After a new variable is introduced, all existing variables in the regression model are verified. After verification, insignificant variables are deleted. This process continues until no new variables can be introduced and all explanatory variables in the regression model are significant. Multiple stepwise regression can ensure that the final explanatory variables are significant and optimal and can be used to screen and remove variables that may cause multicollinearity.

4 Results and discussion

In this paper, stepwise multiple linear regression is employed to analyze the trip rate using commuting trip OD data during the peak morning hours in the central city areas of Kunshan. Subdivided housing types, including apartments, houses, common multistory and high-rise buildings, mixed commercial buildings and residential real estate are considered to obtain the trip rates for different housing types. Specifically, the floor area of each subdivided housing type is introduced into the regression model as an independent variable, and corresponding effective trip rates are obtained for each subdivided housing type. Because there are offices in some residential areas in Kunshan, regression analysis yields both the commuting trip production rate and the trip attraction rate. The results are listed in Table 2 and Table 3.

Table 2. Regression results for the commuting trip production rates (unit: passengers/thousand m²•peak hour)

Land use			Regression analysis taking land-use categories as explanatory variables			Regression analysis taking sub-divided housing types and other land-use categories		
Code		Category	Overall TAZ-CBs	TAZ-CBs in Classification One	TAZ-CBs in Classification Two	Overall TAZ-CBs	TAZ-CBs in Classification One	TAZ-CBs in Classification Two
R		Residential land	1.97**	2.02**	1.30**	×	×	×
	R_1	Type 1 residential land with houses, villas	×	×	×	1.50*	1.63*	—
	R_2	Type 2 residential land with common multi-storied or high-rise buildings	×	×	×	1.77**	1.71**	1.16**
	R_3	Type 3 residential land with relatively rough buildings	×	×	×	4.29*	—	7.37*
	R_a	A type of residential land with apartments	×	×	×	17.02**	26.66**	16.73**
	R_b	mixed commercial and residential land	×	×	×	—	—	—
	R_x	Affiliated land for residential area	×	×	×	—	—	—
A		Public management and public services land	—	—	—	—	—	—
B		Business Service land	1.48**	1.35*	—	1.21**	0.99*	—
G		Green land	—	—	—	—	—	—
M		Industrial land	1.63**	1.56**	2.91**	1.39**	1.43**	1.39*
S		Roads and transport facilities land	—	—	—	—	—	—
U		Public utility land	—	—	—	11.81*	13.70*	—
W		Logistics land	—	—	38.75**	—	—	—
Model fit statistics		R Square	0.911	0.917	0.953	0.953	0.956	0.959
		Adjusted R Square	0.906	0.912	0.945	0.946	0.952	0.954
		Std. Error of the Estimate	906.345	974.263	514.947	668.199	719.244	474.628

Note: "×" denotes a variable that is not involved in the regression; "—" denotes that the regression coefficient is insignificant and not listed; "**" represents p<0.01; "*" represents p<0.05.

Table 3. Regression results for the commuting trip attraction rates (unit: passengers/thousand m2•peak hour)

Land use			Regression analysis taking land-use categories as explanatory variables			Regression analysis taking subdivided housing types and other land-use categories as explanatory variables		
Code		Category	Overall TAZ-CBs	TAZ-CBs in Classification One	TAZ-CBs in Classification Two	Overall TAZ-CBs	TAZ-CBs in Classification One	TAZ-CBs in Classification Two
R		Residential land	1.01**	1.10**	0.54**	×	×	×
	R_1	Type 1 residential land with houses, villas	×	×	×	—	—	—
	R_2	Type 2 residential land with common multi-storied or high-rise buildings	×	×	×	0.85**	0.70**	0.49**
	R_3	Type 3 residential land with relatively rough buildings	×	×	×	—	—	—
	R_a	A type of residential land with apartments	×	×	×	9.48**	22.43**	8.79**
	R_b	mixed commercial and residential land	×	×	×	—	—	—
	R_x	Affiliated land for residential area	×	×	×	—	—	—
A		Public management and public services land	2.40**	2.77**	—	2.33**	1.86*	1.55*
B		Business Service land	2.27**	2.31**	1.89**	2.27**	2.05**	1.60**
G		Green land	—	—	—	—	—	—
M		Industrial land	2.33**	2.28**	3.21**	2.21**	2.18**	2.23**
S		Roads and transport facilities land	—	—	—	—	—	—
U		Public utility land	14.74*	—	—	15.97*	14.94*	—
W		Logistics land	—	—	—	—	—	—
Model fit statistics		R Square	0.890	0.875	0.940	0.900	0.898	0.968
		Adjusted R Square	0.882	0.865	0.932	0.893	0.888	0.963
		Std. Error of the Estimate	812.461	966.936	412.141	772.723	879.501	304.437

Note: "×" denotes a variable that is not involved in the regression; "—" denotes that the regression coefficient is insignificant and not listed; "**" represents p<0.01; "*" represents p<0.05.

The tables show that regression coefficients of residential land (R), common multistory or high-rise buildings (R_2) and apartments (R_a) have high significance, but the regression coefficients of some housing types cannot be adopted. The main reason for this limitation is as follows: R1 and Rb lands in the central city of Kunshan are relatively scarce (shown in Figure 6); thus, the low sample size affects the regression results. In addition, it is easy to understand that, for a specific type of residential land, such as houses, clear regularity may not exist for the trip rate. Therefore, it is difficult to obtain reliable trip rates with high significance for those housing types.

The total trip rates of R, R_2 and R_a and partial trip rates of R_1 and R_3 are obtained by regression analysis. Among them, the trip rates of R_2 are similar to those for overall residential land, largely because 88% of the residential floor area is associated with the R_2 land type. The trip rate of apartment land significantly exceeds that of other housing types because many apartments are located in Kunshan to meet the housing needs of migrant workers with a number of over 300 thousand from the north of Jiangsu Province and other provinces. This type of residential land has a much higher commuting population density than other housing types, which results in a significantly higher trip rate.

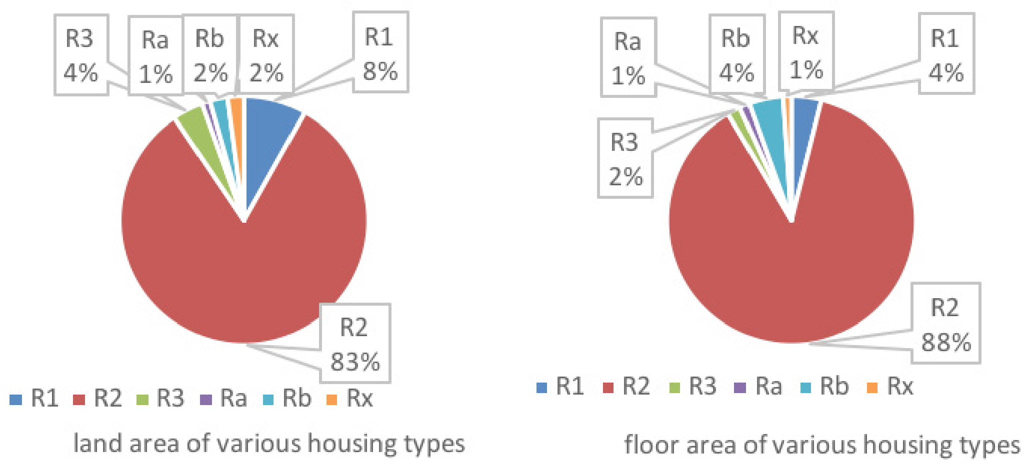

Figure 6. Proportions of the land and floor areas of various housing types in the research scope

Table 2 shows that for multiple types of residential land, the trip rates of different housing types vary substantially. For instance, the commuting trip rates for R_1, R_2 and R_3 are 0.8–2.2 times the value for overall residential land. Therefore, investigating the trip rate of overall residential land alone cannot precisely reflect the trip characteristics of each subdivided housing type. From the perspective of precision and practicality, the trip characteristics of each subdivided housing type should be analyzed in detail.

Based on different location conditions, the commuting trip production rates for overall residential land are 2.02 (Classification One) and 1.30 (Classification Two), which match the location characteristics, i.e., the high accessibility and commuting population density in Classification One lead to a high trip rate. The regression analysis of each housing type yields a similar conclusion, and the trip rate of Classification One always exceeds that of Classification Two.

The common inference is that the regression coefficient of all research units without classification should represent the overall average. Therefore, when a classification scheme is applied, theoretically, the regression coefficients of different classifications should fall on either side of the overall average. In this paper, when overall residential land is defined as the independent variable, the regression result matches this inference; however, the regression result of R_2 does not agree with this trend. One possible reason for this difference is that although stepwise multiple regression is applied in each regression analysis process, the regression results have different precisions. These differences reduce the comparability among regression coefficients and make it difficult to match the inference. We have no plan to adjust the trip rate of each subdivided housing type to satisfy the condition of the abovementioned inference because the trip rates of overall residential land and various housing types are regression values with highest level of precision for the corresponding scenarios.

Moreover, the commuting trip attraction rate for residential land is obtained in this paper. Specifically, the trip attraction rate for overall residential land is 1.01, and the trip attraction rate for R_2 is 0.85. The precondition of these trip rates is that various enterprises are located in residential areas in many cities, including Kunshan, and they include technology companies, education agencies and entertainment companies. Therefore, the use function of residential land may change. In addition, there are a large number of nannies and housekeepers that cannot be ignored in the commuting trip rate analysis.

5 Conclusion

In this paper, the information in mobile phone signaling data, such as the trip volume and spatial distribution from starting point to terminating point, is fully utilized. TAZ-CBs, which differ from traditional TAZs and comprise the coverage of multiple cell towers, are created to minimize the difference between the actual coverage and the theoretical coverage. A cluster analysis of transport accessibility, based on the path planning data of map providers, and the commuting population density, based on residential location identification, were conducted to obtain high-precision commuting trip production rates and attraction rates for overall residential land and each subdivided housing type during the peak morning hours. The trip rates with high significance obtained in this paper can be applied in local transport analysis models.

Notably, (1) the regression analysis in this paper only focuses on mobile phone data during peak morning hours, whereas the peak trip rates of some land types, such as business service land and public management land, do not occur during peak morning hours. (2) The trip rate in this paper is not based on an entire sample because the market share of China Mobile in Kunshan is 69.4%. We can easily calculate actual commuting trip rates by trip rates in Table 2 and Table 3 divided by 0.694. One limitation is that trip data are obtained from the mobile operator; due to privacy reasons, researchers cannot obtain the socioeconomic characteristics of mobile phone owners, and the obtained trip rates do not reflect trip characteristics of socioeconomic groups. Most likely, transport accessibility and the commuting population density may reflect some information related to housing prices and income. Mobile phone signaling data can be used to identify trip volumes but not trip modes. Therefore, currently, the trip rate measured by the number of vehicles is unavailable. In fact, trip rates, as measured by the number of vehicles, are more suitable for cities in North America because trip modes in China are more diverse.

Almost everyone carries a mobile phone. Signaling data provide unprecedented detail regarding population movement and have the potential for use in precise, quantitative analyses. As information and data sharing continues, mobile phone signaling data are expected to become an important source of basic data for future trip rate and transportation research.

Acknowledgements

This research work was supported by National Natural Science Foundation of China (No.51778277).

References

Agyemang-Duah, K., & Hall, F. L. (1997). Spatial transferability of an ordered response model of trip generation. *Transportation Research Part A Policy & Practice, 31*(5), 389–402.

Agyemang-Duah, K., Anderson, W. P., & Hall, F. L. (1995). Trip generation for shopping travel. *Transportation Research Record, 1453*, 12–20.

Arabani, M., & Amani, A. B. (2007). Evaluating the parameters affecting urban trip-generation. *Iranian Journal of Science & Technology Transaction B Engineering, 31*(5), 547–560.

Berki, Z., & Monigl, J. (2017). Trip generation and distribution modelling in Budapest. *Transportation Research Procedia, 27*, 172–179.

Bochner, B. S., Hooper, K., Sperry, B., & Dunphy, R. (2011). *Enhancing internal capture estimation for mixed-use developments.* NCHRP Report 684. Washington, DC: Transportation Research Board for the National Academies.

Bwamble, A., Choudhury, C. F. & Sanko, N. (2015). Modelling car trip generation in the developing world: The tale of two cities. Transportation Research Board 94th Annual Meeting, Washington, DC.

Bwambale, A., Choudhury, C., & Hess, S. (2017). Modelling trip generation using mobile phone data: A latent demographics approach. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2017.08.020

Clifton, K. J., Currans, K. M., & Muhs, C. D. (2013). Evolving the Institute of Transportation Engineers' trip generation handbook: A proposal for collecting multi-modal, multi-context, establishment-level data. *Journal of the Transportation Research Board, 2344*, 107–117.

Clifton, K. J., Currans, K. M., & Muhs, C. (2015). Adjusting ITE's trip generation handbook for urban Context. *Journal of Transport and Land Use, 8*(1) 5–29.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiratta, S. R., & González, M. C. (2015). Analyzing cell phone location data for Urban travel. *Transportation Research Record, 2526*(3), 126–135.

Currans, K. M. (2017). Issues in trip generation methods for transportation impact estimation of land-use development. *Journal of Planning Literature, 32*(4), 335–345.

Demissie, M. G., Antunes, F., Bento, C., Phithakkitnukoon, S., & Sukhvibul, T. (2016). Inferring origin-destination flows using mobile phone data: A case study of Senegal. 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology. IEEE, June 28–July 1, Chiang Mai, Thailand.

Dey, S. S., & Fricker, J. D. (1994). Bayesian updating of trip generation data: Combining national trip generation rates with local data. *Transportation, 21*(4), 393–403.

Dock, S., Cohen, L. Rogers, J. D., Henson, J. Weinberger, R., Schrieber, J., & Ricks, K. (2015). Methodology to gather multimodal urban trip generation data. Paper presented at 94th Annual Meeting of the Transportation Research Board, Washington, DC.

Ellys, B. & Reid, F. (1982). Critique of ITE trip generation rates and alternative basis for estimating new area traffic. *Transportation Research Record 1874*, 1–2.

Everett, J. D. (2009). An investigation of the transferability of trip generation models and the utilization of a spatial context variable. Doctoral dissertation. Knoxville, TN: University of Tennessee.

Ewing, R., Deanna, M., & Li, S. C. (1996). Land-use impacts on trip generation rates. *Transportation Research Record, 1518*(1), 1–6.

Filitowish, N. (2011). *Effectiveness of applying in-site trip rates for trip generation modeling in developing cities*. Urban transportation research series. Cape Town: Cape Town University.

Hanson, S. (1982). The determinants of daily travel-activity patterns: Relative location and sociodemographic factors. *Urban Geography, 3*, 179–202.

Ko, J. (2013). Vehicle trip generation rates for office buildings under urban settings. *ITE Journal, 8*(2), 41–45.

Landrock, J. N. (1981). Spatial stability of average daily travel times and trip rates within Great Britain. *Transportation Research A, 15A*, 55–62.

Lee, J. (2016). Impact of neighborhood walkability on trip generation and trip chaining: Case of Los Angeles. *Journal of Urban Planning & Development, 142*(3), 05015013.

Li, Z.-F., Yu, L., Gao, Y., et. al., (2016). Extraction method of temporal and spatial characteristics of residents' trips based on cellular signaling data, *Transportation Research, 2*(1), 51–57. (In Chinese)

Lim, K. K., & Inivasan, S. (2011). Comparative analysis of alternate econometric structures for trip generation models. *Transportation Research Record, 2254*, 68–78.

Macababbad, R. J. R., & Regidor, J. R. (2009). A study on the local trip generation characteristics of government office land use. *Proceedings of the Eastern Asia Society for Transportation Studies, 7*. https://www.jstage.jst.go.jp/browse/eastpro/2009/0/_contents/-char/ja/

Meyer, M. D., & Miller, E. J. (2001). *Urban transportation planning: A decision-oriented approach* (2nd edition). New York: Mc-Graw-Hill Companies.

Moreno-Monroy, A. I., Lovelace, R., & Ramos, F. R. (2017). Public transport and school location impacts on educational inequalities: Insights from São Paulo. *Journal of Transport Geography, 67*, 110–118.

Ohlms, P. B. (2016). Trip generation at Virginia agritourism land uses. Charlottesville, VA: Virginia Transportation Research Council.

Pan, T. (2008). Assignment of estimate average annual daily traffic on all roads in Florida. *Graduate Theses and Dissertations*. Retrieved from https://scholarcommons.usf.edu/etd/442/

Pitsiava-Latinopoulou, M., Tsohos, G., Basbas, S. (2001). Trip generation rates and land use—Transport planning in urban environment. Paper presented at the 7th International Conference on Urban Transport and the Environment in the 21st Century, Lemnos, Greece.

Quintero, P. A., Diaz, G. M., & Moreno, E. G. (2016). Trip generation by transportation mode of private school, semi-private and public. Case study in Merida, Venezuela. *Transportation Research Procedia, 18*, 73–79.

Roorda, M. J., Páez, A., Morency, C., Mercado, R., & Farber, S. (2010). Trip generation of vulnerable populations in three Canadian cities: A spatial ordered probit approach. *Transportation, 37*(3), 525–548.

Safwat, K. N. A., & Magnanti, T. L. (1988). A combined trip generation, trip distribution, modal split, and trip assignment model. *Transportation Science, 22*(1), 14–30.

Schmocker, J. D., Quddus, M. A., Noland, R. B., & Bell, M. G. H. (2005). Estimating trip generation of elderly and disabled people: Analysis of London data. *Transportation Research Record, 1924*, 31–60.

Schneider, R. J., Shafizadeh, K., Sperry, B. R, & Handy, S. L. (2013). Methodology to gather multimodal trip generation data in smart-growth areas. *Transportation Research Record, 2354*, 68–85.

Shoup, D. C. (2003). Truth in transportation planning. *University of California Transportation Center Working Papers, 2003,* 6(1).

Singleton, A. A. (2014). GIS approach to modelling CO2 emissions associated with the pupil-school commute. *International Journal of Geographical Information Science, 28*(2), 256–273.

Soltani, A., Saghapoor, T., Izadi, H., & Pakshir, A. (2012). Trip generation and its relationship with land-use diversity: Case studies of four urban districts in Shiraz metropolitan area. *Journal of Urban–Regional Studies and Research, 3*(12).

Tian, G., & Ewing, R. A. (2017). A walk trip generation model for Portland, OR. *Transportation Research Part D Transport & Environment, 52*, 340–353. (In Chinese)

Tilsdor, M. (2012). *Provision and effective use of alternative modes of transport in the developing cities*. Kampala, Uganda: Makerere University, Department of Architectural and Physical Planning.

Wilfred, G., Bwire, H., Mattsson, L., & Jonsson, D. (2015). Effects of land use on trip generation in urban areas: Comparison between estimated trip generation rates and planning practices. Paper presented at the 34th Annual Southern African Transport Conference, Dar es Salaam, Tanzania.

Wilmot, C. G., & Mei B. (2004). Comparison of alternative trip generation models for hurricane evacuation. *Natural Hazards Review, 5*(4), 170–178.

Wu, D., Sall, E., & Newhouse, S. (2012). Trip rates and accessibility: Gleaning basic planning information from activity-based travel demand model. *Transportation Research Record, 2303*(1), 81–88.

Zhou, T., Zhao, B., & Yu, B. (2017). Transportation big data analysis methodology based on CRISP-DM: An example of cellular signaling and RFID data in Chongqing. *Urban Transport of China, 2017*(5), 15–20. (In Chinese)