A Network-Based Model of Passenger Transfer Flow between Bus and Metro
An Application to the Public Transport System of Beijing

Wang, Wenjing; Wang, Yihong; Correia, Gonçalo Homem de Almeida; Chen, Yusen

DOI
10.1155/2020/6659931

Publication date
2020

Document Version
Final published version

Published in
Journal of Advanced Transportation

Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 6659931, 12 pages
https://doi.org/10.1155/2020/6659931

Research Article

A Network-Based Model of Passenger Transfer Flow between Bus and Metro: An Application to the Public Transport System of Beijing

Wenjing Wang,[1,2] **Yihong Wang**,[3] **Gonçalo Homem de Almeida Correia**,[3] **and Yusen Chen**[3]

[1]*Beijing University of Technology, Beijing 100124, China*
[2]*Research Institute of Highway, Ministry of Transport, Beijing 100088, China*
[3]*Department of Transport and Planning, Delft University of Technology, P.O. Box 5048, GA Delft 2600, Netherlands*

Correspondence should be addressed to Yihong Wang; y.wang-14@tudelft.nl

Received 28 October 2020; Accepted 20 November 2020; Published 7 December 2020

Academic Editor: Nirajan Shiwakoti

In a multimodal public transport network, transfers are inevitable. Planning and managing an efficient transfer connection is thus important and requires an understanding of the factors that influence those transfers. Existing studies on predicting passenger transfer flows have mainly used transit assignment models based on route choice, which need extensive computation and underlying behavioral assumptions. Inspired by studies that use network properties to estimate public transport (PT) demand, this paper proposes to use the network properties of a multimodal PT system to explain transfer flows. A statistical model is estimated to identify the relationship between transfer flow and the network properties in a joint bus and metro network. Apart from transfer time, the number of stops, and bus lines, the most important network property we propose in this study is transfer accessibility. Transfer accessibility is a newly defined indicator for the geographic factors contributing to the possibility of transferring at a station, given its position in a multimodal PT network, based on an adapted gravity-based measure. It assumes that transfer accessibility at each station is proportional to the number of reachable points of interest within the network and dependent on a cost function describing the effect of distance. The R-squared of the regression model we propose is 0.69, based on the smart card data, PT network data, and Points of Interest (POIs) data from the city of Beijing, China. This suggests that the model could offer some decision support for PT planners especially when complex network assignment models are too computationally intensive to calibrate and use.

1. Introduction

In a public transport (PT) network, it is impossible to provide all passengers with a direct and unimodal PT service between all the stations and stops. Passengers sometimes have to transfer between different lines and often between different modes. A trip by PT could, therefore, involve one or even more transfers from one mode to another [1, 2]. In contrast to door-to-door service, inconvenient transfers can disrupt passenger travel and reduce the competitiveness of PT [3, 4]. A better transfer connection between modes has been shown to improve the level of service of PT in general and thus stimulate its overall usage [5–7]. To provide a better transfer connection, it is necessary to be able to quantify transfer flows, thus allowing smart transfer planning and management [8]. For example, if PT planning and management authorities want to understand pedestrian behavior at a transfer corridor and further improve connection efficiency, they need to estimate and predict the passengers' transfer flow [9]. Since the combination of bus and metro is a typical one in many cities, much research has focused on how to provide a better-integrated bus and metro system through such transfer connections [10, 11], which is also the focus of this paper.

Many rule-based algorithms have been developed to estimate transfer flow based on smart card data [12, 13], but they can only estimate the historical transfer flow of an existing station. To predict the transfer flow of a newly planned station, transit assignment models based on transit users' route choices have been used [1, 14, 15]. Discrete choice models have been used to explain the route choice of travelers based on utility maximization [16]. Such models search for the route choice set of travelers and calculate the probability of each choice, resulting in extensive calibration and computation time [17]. There are also studies using only network properties [18] to assign PT passenger flows, which provide a parsimonious alternative to existing passenger assignment models [19]. However, this type of approach has still not been used to model transfer flows and there is no research attempt to examine the relation between transfer flow and network properties. In this paper, we aim to fill this gap by establishing a model of transfer flow between metro and bus based on network properties.

Some network indicators can be obtained directly from the data [20, 21], such as transfer time and the number of bus lines around one metro station [22]. Apart from these relatively straightforward indicators, the most important network property introduced in this study is what we call transfer accessibility. This is a newly defined indicator for the radiation of a transfer station given its position in a bimodal PT network. Intuitively, this indicator represents the accessibility of a transfer station, which is proportional to the sum of potential interactions between all reachable metro stations and all reachable bus stops and inversely proportional to generalized travel cost of these interactions. The potential interaction is measured in terms of the potential production of a bus stop (or a metro station) plus the potential attraction of a metro station (or a bus stop). For both production and attraction, we use the number of points of interest (POIs) around each station (or stop) as a proxy, which is a dataset that is typically available nowadays. It should be noted that some research referred to the robustness of transfer connections within a station also as transfer accessibility [23], which should be distinguished from our concept.

Our approach to calculating transfer accessibility based on the sum of potential interactions is very similar to the measurement of gravity-based accessibility [24], which can be regarded as an analogy to Newton's gravitational law [25]. Namely, the exchange of people between two cities is directly proportional to the product of population and inversely proportional to the square of the distance between the two cities [24]. In this paper, we propose such a gravity-based model to estimate transfer accessibility and then use it as an explanatory variable to establish a regression model of station-level transfer flows.

The paper is organized as follows. First, the methodology is described, which includes the definition of transfer accessibility and the regression model for transfer flow prediction. Then, the PT data of Beijing used in our study is further explained. Following that, we present the application of our model to those data. In the final section, we draw conclusions and suggest directions for future research.

2. Methodology

We assume that the network properties of a station can be related to transfer flow between two modes of transportation. In this study, we aim to test this assumption. Since not all single features are normally distributed and a nonlinear relationship may exist between the independent and dependent variables [26], we take the logarithm of the variables to build the regression model if necessary. The model is presented as follows:

$$\log\left(y_j\right) = \beta_0 + \beta_1 \log\left(x_1\right) + \cdots + \beta_p \log\left(x_p\right) + \varepsilon, \quad (1)$$

where y_j is the transfer flow of station j, ε represents the error term, and x_p are the different explanatory variables that represent network properties.

Next, we select a group of network properties that are considered to be related to transfer flows. Based on a review of the existing literature, the following network properties are selected (more details in Section 2.2).

(i) Transfer accessibility (the new indicator)

(ii) Transfer time [27]

(iii) The number of bus stops around each metro station [28]

(iv) The number of bus lines per bus stop [22]

As summarized in Figure 1, a regression model is established to find the relationship between transfer flow and the four network attributes mentioned above, among which transfer accessibility needs to be calculated based on a gravity model. The gravity model assumes that transfer accessibility at each station is dependent on the number of reachable POIs, PT stops at this station, and a cost function describing the effect of distance. Its calculation process consists of five steps: for a station, (1) find all OD pairs that connect to this station, (2) calculate a proxy for potential trip interactions between every OD pair, specifically in terms of the number of POIs surrounding an origin station plus the one surrounding a destination station, (3) for each OD pair, multiply the interaction by a cost function that describes the effect of distance for each OD station pair, (4) filter out those OD pairs connected by direct transport, such as direct metro or bus lines, and (5) sum the calculation results over all the reachable OD station pairs to calculate gravity-based accessibility. The method can be applied in a PT network that includes bus stops and metro stations.

2.1. Dependent Variable. In this study, the dependent variable is the transfer flow. In order to compute transfer flow from smart card data, it is necessary to first identify what a transfer is. When commuters travel in PT networks using smart cards [29], the following data from each trip is available through smart card data: anonymous identities (IDs) of users, IDs of boarding and alighting stations, and timestamps.

During the past decade, different approaches have been proposed to identify transfers based on smart card data [30], many of which are rule-based approaches. For example,

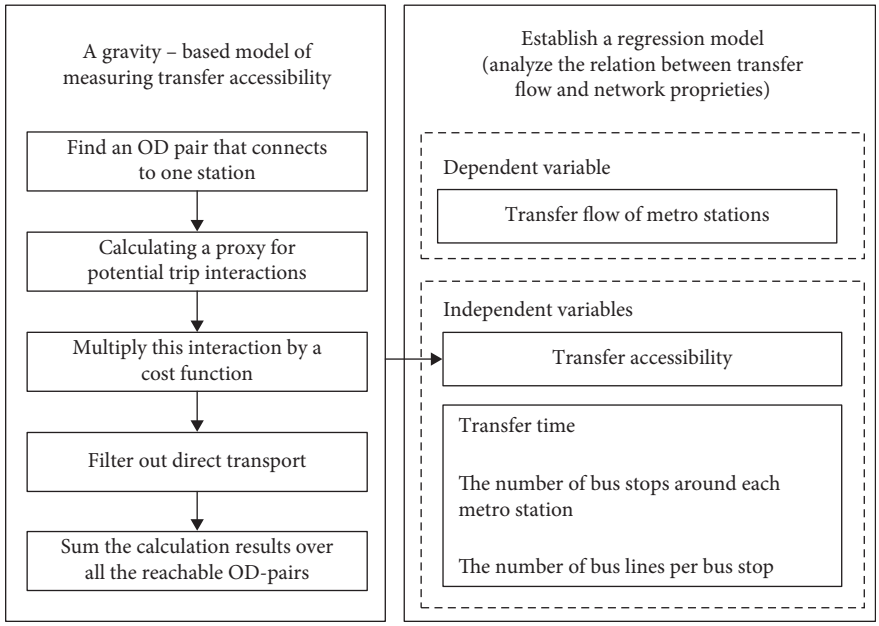

Figure 1: Main components of the developed methodology and overall research design.

different fixed time thresholds are set for the observed time gaps between consecutive trip legs/segments [31]. Transfer time thresholds ranging from 30 minutes to 90 minutes have been used for London to identify transfers with smart card data [12, 32]. Otherwise, transfer walking distance can also be applied. A maximum threshold of 750 meters on transfer distances was used to estimate transfers in London [33], and 400 meters in The Hague, Netherlands [13]. Some approaches further distinguish transfers from short activities, which incorporate the effects of denied boarding, transferring to a vehicle of the same line [13], and the circuitry of the path trajectories [34].

In this paper, we also identify transfers using a rule-based approach. The thresholds of transfer time and transfer distance are set to detect transfers based on smart card data. Our research area is the city of Beijing and we focus on the transfers between bus and metro. Firstly, the complexity of the Beijing PT network is similar to London and Shanghai. Based on the transfer data of London [12] and Shanghai [35], we can preliminarily determine that the transfer time is generally about 30 minutes for these large-scale cities. The maximum transfer distance is set at 2.5 km, based on the assumed maximum walking speed [33]. Secondly, in order to test whether 30 minutes are reasonable for Beijing, we analyzed the time interval of two adjacent trips of all passengers, where their trips interval is about 30 minutes and distance is within 2.5 km, based on Beijing smart card data. As shown in Figure 2, the time interval of 95% of trips is less than 25 minutes. Therefore, we set our threshold of transfer time as 25 minutes and the maximum transfer distance as 2.5 km. Following these rules, it is possible to estimate transfer flows through every metro station, based on smart card data.

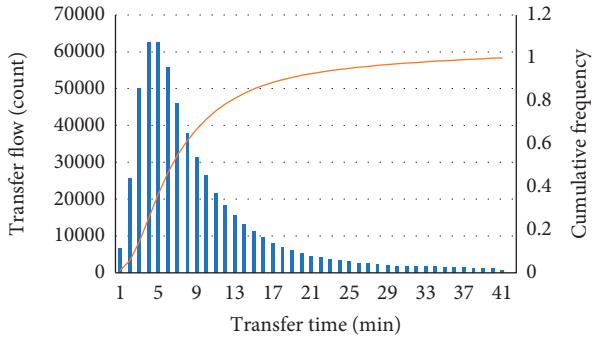

Figure 2: Cumulative frequency distribution of transfer flow on different transfer times.

There are many types of transfer, including internal transfers such as the ones within the metro system, and external transfers between bus and metro. We consider internal transfer between different metro lines as one trip segment since commuters only need to swipe their cards when they get in and out of a metro station and do not swipe their cards when they transfer between different metro lines. In our joint network of bus and metro, one-time transfers between metro and bus comprise the majority of the transfers, accounting for 91% of all transfers between metro and bus, based on Beijing smart card data (Figure 3). Thus, one-time transfers between metro and bus are our research focus in this paper.

2.2. Independent Variables. In our regression model that predicts transfer flow, there are four independent variables in total. The first independent variable is the transfer time of a trip between the bus and the metro, determined according

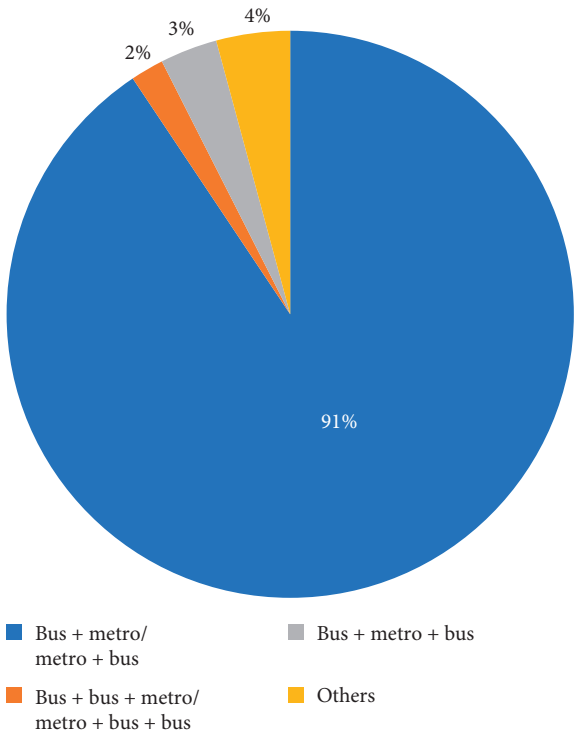

Figure 3: Different transfer patterns between metro and bus (note that the transfers within the metro system or bus system are not considered in this paper and are thus excluded).

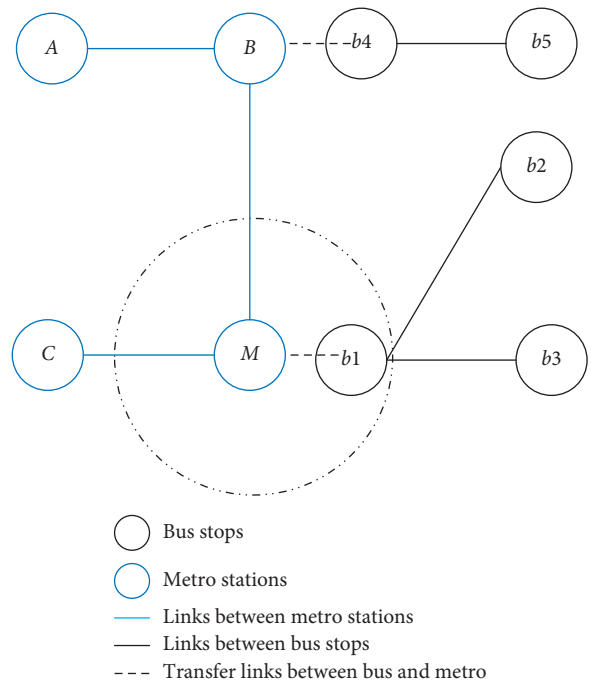

Figure 4: A PT toy network.

to the time interval of the traveler swiping their card. Based on the median of transfer times of all transfer trips through one metro station, the transfer time from a metro station to a bus stop (or vice versa) can be obtained. We use the median value of all empirical transfer times at one metro station to represent the general transfer time of this station. For a newly planned station, transfer time can be initially estimated based on the transfer distance and the estimated waiting time.

The second independent variable is the number of bus stops around one metro station, which reflects the potential opportunities for commuters to transfer. We set the radius as one kilometer and count the number of bus stops within this range from each metro station. The third independent variable is the number of bus lines per bus stop, which reflects the intensity of bus service at a bus stop next to the metro station. The assumption is that if there are more lines at one bus stop, there would be more transfer trips. We explain the first three as follows and will specify the last, the new one put forward in this paper. As it has been introduced before, a gravity-based model is proposed to measure transfer accessibility. This model assumes that transfer accessibility of each station is dependent on the number of reachable POIs in a city, data which is nowadays easy to obtain, and a cost function describing the effect of distance.

We use a toy PT network combining a bus network and a metro network to explain our definition. As illustrated in Figure 4, each node represents a metro station (a blue node) or a bus stop (a black node). There are four metro stations (A, B, C, and M) and five bus stops ($b1$, $b2$, $b3$, $b4$, and $b5$). A link between two bus stops or two metro stations exists if there are PT services connecting them. A dashed line represents the transfer connection between a bus stop and a metro station. For example, commuters can walk between bus stop b1 and metro station M to transfer and continue their trips.

In this gravity-based model, we focus on one transfer station and find all the OD pairs that can be connected through it. In our case, an OD pair should consist of one bus stop and one metro station. When we focus on one metro station, all possible transfer links from one metro station to different bus stops which are located around this metro station will be searched. In the PT toy network example (Figure 4), we focus on metro station M, which has a possible transfer link with bus stop $b1$. We assume that a trip is transferred from bus to metro; therefore, the origin node could be either bus stop $b2$ or $b3$, connected by a bus line to bus stop $b1$. The destination node could be either metro station A, B, or C, since all metro stations are interconnected, and commuters can travel from metro station M to any other metro station. There are 6 OD pairs connected through metro station M, including $b2$-A, $b2$-B, $b2$-C, $b3$-A, $b3$-B, and $b3$-C.

For one transfer metro station, we search for all potential OD pairs that are connected through this station. We use the number of POIs surrounding a metro station or a bus stop as a proxy for potential trip production or attraction. For metro station M in the above PT toy network, one needs to calculate the number of surrounding POIs of 6 OD pairs which are connected through this station. For example, the proxy potential trip interaction for metro station M between the OD pair "$b2$-A" is the sum of the number of POIs around bus stop $b2$ and metro station A. The total number of

company POIs and housing POIs is counted within a 500-meter radius [35] from each metro station and each bus stop.

An OD pair might be connected directly by a single PT mode. If that is the case, the amount of transfer flow between this OD pair would be reduced. Therefore, if one wants to estimate transfer demand [36] more accurately, the impact of direct transport should be removed. The number of metro stations, the number of bus lines, the travel time by bus [37], and the standard deviation of travel time will affect commuters' choices. We combine the four factors mentioned above to obtain the transfer demand impact factor $\zeta k(j)$:

$$\zeta k(j) = \begin{cases} 0, \mid m_k > 0, \\ \dfrac{1}{\left(n_k \times t_{\text{ktotal}}/t_{\text{kbus}} \times (1 - \text{std}_{\text{kbus}}/t_{\text{kbus}})\right)}, \mid m_k = 0,\ n_k > 0, \\ 1, \mid m_k = 0,\ n_k = 0, \end{cases} \tag{2}$$

where j is the current transfer station, and k is the k^{th} OD pair which is connected through station j. $\zeta k(j)$ denotes the transfer demand impact coefficient of the k^{th} OD pair transferring at station j. m_k and n_k are the number of metro lines and the number of bus lines, respectively, which can connect the k^{th} OD pair directly. t_{ktotal} is the total travel time of the k^{th} OD pair when commuters choose to transfer at station j. t_{kbus} is the average bus travel time of the k^{th} OD pair when commuters choose to travel by bus directly. $\text{std}_{\text{tkbus}}$ is the standard deviation of bus travel time on the k^{th} OD pair when commuters choose to travel by bus directly.

If some metro lines can directly connect the k^{th} OD pair, set $\zeta k(j) = 0$, and if there is neither a metro line nor a bus line between the k^{th} OD pair of station j, set $\zeta k(j) = 1$. Otherwise, $\zeta k(j)$ is determined by the effect of multiple parameters, including n_k, t_{ktotal}, t_{kbus}, and $\text{std}_{\text{tkbus}}$. Bus running times and running time variation will affect service reliability and will further affect the attractiveness of travel by bus [22]. Therefore, we can assume that the lower the standard deviation of bus travel time is, the more punctual and stable bus travel time will be, which should motivate commuters to use it [22]. The higher the number of bus lines between one OD pair, the higher the probability of having a good bus connection; this also motivates commuters to use the bus directly instead of transfer.

We use a combined cost function to model commuters' reluctance to travel a long distance. This function has the following form [25]:

$$\text{f}\left(c_{kj}\right) = c_{kj}^{n} \times \exp(-\beta \times c_{kj}), \tag{3}$$

where $\text{f}(c_{kj})$ is a generalized impedance function of travel distance with two parameters for calibration, and c_{kj} is the travel distance traveling through transfer metro station j between the k^{th} OD pair. The shape of this function for different values of its parameters is shown in Figure 5.

The values of n and β should be calibrated to calculate transfer accessibility based on the cost function. In Figure 2, if we focus on metro station M, b2-A is one of all the potential OD pairs which are connected through this station. In this case, the travel distance c_{kj} between the OD pair, "b2-A" is the sum of the distance b2-M and the distance M-A. Based on the estimated n, β, and this travel distance c_{kj}, it is possible to obtain the cost function between the OD pair "b2-A", which is not always decreasing. It first rises and then gradually decreases until it stabilizes near zero with the change in travel distance.

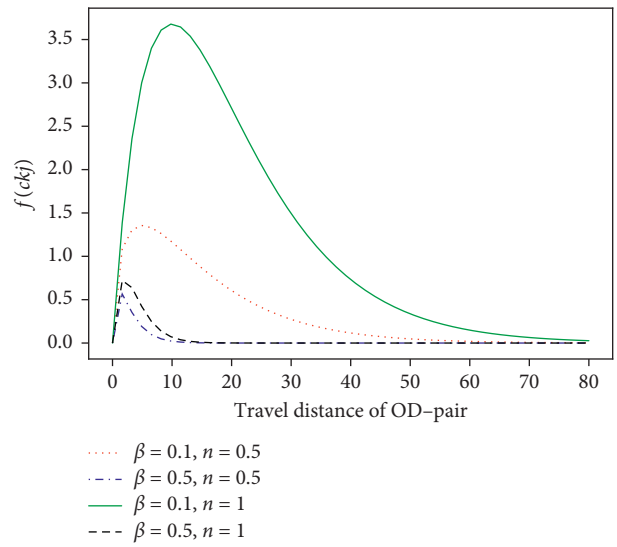

Figure 5: Cost functions with different parameters.

By summing the calculation results of accessibility of station j over all the potential OD pairs which are connected through this station, it is possible to obtain the transfer accessibility of station j. The definition of the transfer accessibility of metro station j is given as follows:

$$x_1(j) = \sum_{k=1}^{m} p_k(j) \times \zeta_k(j) \times f\left(c_{kj}\right), \tag{4}$$

where m is the number of OD pairs transferring at station j. k represents the k^{th} OD pair transferring at station j. $p_k(j)$ is the potential trip interactions of the k^{th} OD pair transferring at station j. $\zeta_k(j)$ denotes the transfer demand impact factor of the k^{th} OD pair transferring at station j. $f(c_{kj})$ is a cost function describing the effect of distance.

3. Application to the PT Network of Beijing

3.1. Data. The case study is conducted in the city of Beijing, the capital of China. Some basic information about Beijing and its network is shown in Table 1.

We use network data, smart card data, and POI data in our research. The number of bus stops around one metro station is counted within a one-kilometer radius from each metro station. In Figure 6, nodes represent metro stations, and the depth of color represents the number of bus stops nearby this metro station.

A smart card can be used by Beijing's travelers to board the metro, buses, and public bicycles. According to the National Report on Urban Passenger Transport Development [39], 67.4% of the travelers used a smart card when they travel by PT in Beijing in 2017. Therefore, smart card

Table 1: Basic information of PT network in Beijing.

Concept	Information
Area	16,410 square kilometers
Population	21.73 million [38]
The number of bus lines	886
The number of bus stops	About 32,000 (the same stop with different directions will be counted as different stops)
The number of metro lines	22 metro lines
The number of metro stations	370 metro stations [38]
Daily average PT passenger flow	19.55 million

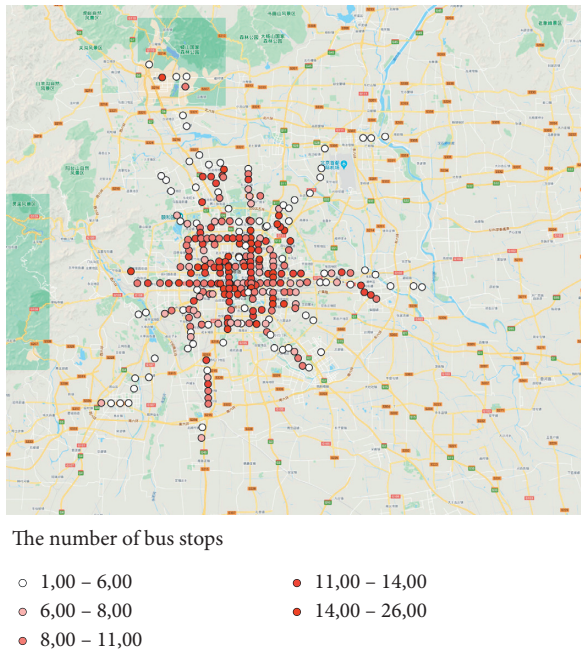

Figure 6: The metro network in Beijing, China, with the depth of color indicating the number of surrounding bus stops.

data can somehow be used as a representative sample of the PT passenger population at the time. Notably, our approach can also be applied to the latest PT data obtained from the new smartphone-based payment methods, such as NFC and QR codes, as long as they record the same type of information. Cardholders need to check in and check out when they travel in all PT systems [40]. As shown in Table 2, the data used in this paper is from September 4 to September 11 in 2017 (8 days). It contains the records of all the transactions completed by smart cardholders during this period. Travelers do not need to check out when they transfer within the metro system, but they do need to check out first and check in again if they transfer between metro and bus.

The POI data used in this paper were extracted from the Gaode Maps service, which is the Chinese equivalent of Google Maps [41]. About 1.2 million POIs of twenty categories can be obtained in Beijing. The available information of the POI data includes name, coordinates, and category. The twenty categories include residence and company. Three types of information are extracted from the original POI dataset for each metro station and bus stop, including the total number of surrounding POIs, the number of surrounding residence POIs, and the number of surrounding company POIs [35]. The number of POIs around the metro stations is indicated by the depth of color in Figure 7.

3.2. Data Preprocessing. We use the data from September 4, 2017, as an example to illustrate the preprocessing of the raw data. The number of bus card transactions on this day is 141,192,280 and the number of subway card transactions is 534,1597. Firstly, the anomalous data is removed, including the following cases: (1) when the line number is not available; (2) when there is a missing record of the boarding or alighting stop; (3) when the alighting time is earlier than the boarding time; (4) when the boarding and alighting are at the same stop on the same line; (5) when there is duplicate data; and (6) when the station ID is wrong. After data preprocessing, we obtain 5,070,457 valid bus records and 5,300,593 valid metro records. Consequently, the total number of bus and subway records is 10,371,050. Secondly, the data of users with two consecutive travel records are detected in the combined transit and metro records. We connect two adjacent trip records of the same user into one trip record, leading to three types of travel including a

Table 2: Information on smart card data used in this paper.

Data concept	Bus	Metro
Attributes	Card id; Card type; Line id; Boarding stop; Check-in time; Alighting stop; Check-out time;	Card id; Card type; Entry line id; Entry station; Check-in time; Exit line id; Exit station; Check-out time;
Size	More than 10 million transaction records per day	More than 5 million transaction records per day
Period	September 4 to September 11 in 2017	

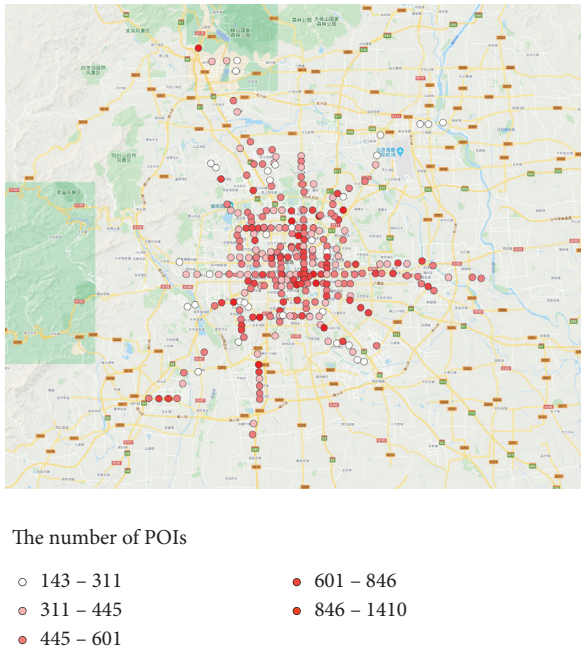

Figure 7: The number of POIs surrounding metro stations.

transfer: bus and bus trip, metro and metro trip, and bus and metro trip. We focus on bus and metro trips and obtain 1,082,269 records. Thirdly, the transfer time and transfer distance are calculated for these bus and metro trips. If the transfer time is less than 25 minutes and the transfer distance is less than 2.5 km for one trip record, we consider it to be a transfer trip. We obtain 566,978 transfer trip records. Similarly, we analyze the remaining 7 days of data to calculate the average transfer flow.

4. Results of the Case Study

4.1. Identifying Transfers and Calculating Variables. The transfer flow of all metro stations is shown in Figure 8, where it can be observed that stations with more transfer flow are not necessarily located in the city center.

As shown in Figure 9, transfer times range from 3 minutes to 25 minutes. Most of the transfers take around 8 minutes. The number of bus stops within a one-kilometer radius of each metro station ranges from 1 to 25. On average,

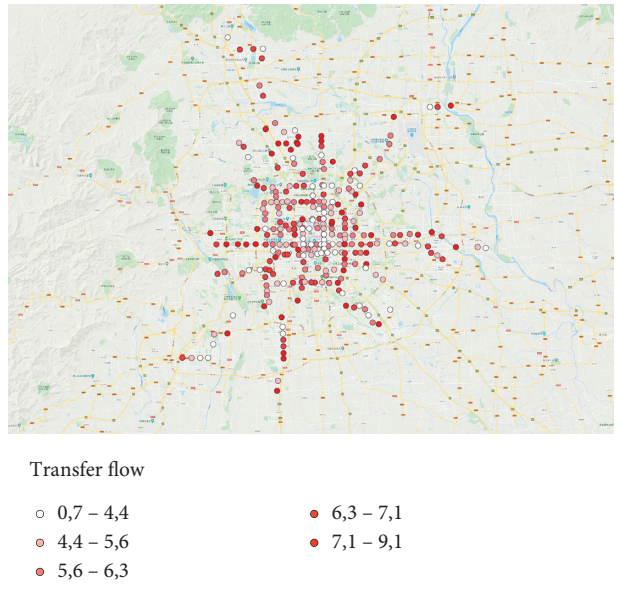

Figure 8: Transfer flow from bus to metro on one day.

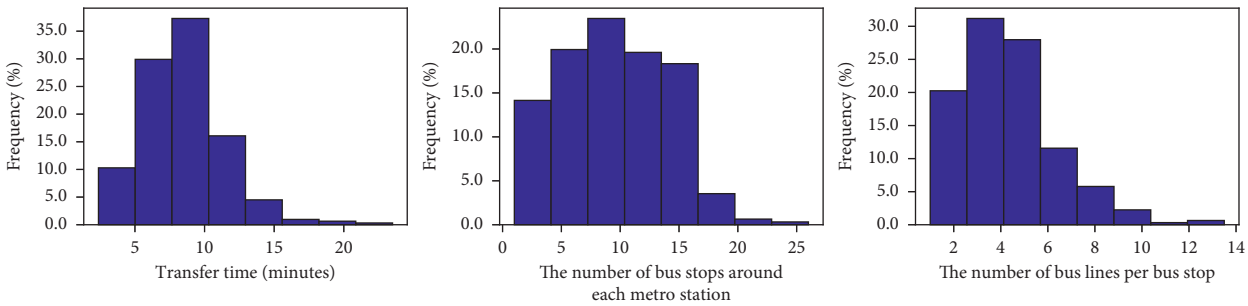

Figure 9: Frequency distribution histogram of three attributes.

there are around 8 bus stops near each metro station. The number of bus lines per bus stop varies from 1 to 13, whilst 3 to 5 seem to appear more often.

Before calculating the transfer accessibility, two parameters n and β in the cost function of the gravity-based model need to be determined in (3). Using (1), we estimate the model using the real PT data in Beijing. The R-squared accuracy that results from the different parameters is indicated by the depth of color in Figure 10. When $n = 5$ and $\beta = 0.1$, the evaluation results are the best; therefore, we use these values.

With 300 metro stations and more than 30,000 bus stops, there would be theoretically about 9 million OD pairs. Based on the formula, we can calculate the transfer accessibility of every metro station which is indicated by the color depth in Figure 11. It can be observed that some metro stations far from the center are highly accessible since some of them are the only connections to a lot of distant bus stops.

4.2. Correlation Analysis of Variables. The correlation between the independent variables was analyzed in Table 3. The correlations between transfer accessibility and other indicators are weak, except for the number of bus lines per bus stop, which is slightly higher. We still keep these two variables, since they both have a significant impact on model accuracy (more detail in Table 4).

4.3. Model Estimation. We established a regression model for each of the four independent variables and the transfer flow to explore the influence of every single predictive attribute. We show the relationship between every independent variable and the dependent variable in Figure 12. The four attributes all have a significant impact on the transfer flow.

In our final dataset, we have 306 metro stations. The data is split in 70%, as a training set, and 30%, as a test set. The model estimation results based on the training set are summarized in Table 5. All of the coefficients have their positive or negative signs as hypothesized and are all significant.

In general, the coefficients of three attributes including transfer accessibility, the number of bus stops, and the number of bus lines per bus stop are positive and significant in explaining the transfer flow. More bus lines and more bus stops would also lead to more transfer flow. Transfer flow decreases with the increase of transfer time.

We use cross-validation to evaluate our model in terms of R-square"(5)". K-fold method [42] was chosen to do cross-validation. In K-fold cross-validation, the original sample is randomly partitioned into K subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model and the remaining K-1 subsamples are used as training data. The cross-validation process is then repeated K times (the folds), with each of the K subsamples used exactly once as the validation data. The K results from the folds can then be combined to produce a single estimation. The advantage of this method over repeated random subsampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. We tested different k values and finally set $k = 6$.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (\overline{y} - y_i)^2}, \tag{5}$$

where $\hat{y}_i$ is predicted value of y using our model, y_i is the actual value of y, and $\overline{y}$ is the mean actual value of y. R-square reflects the extent to which the fluctuation of y can be described by the fluctuation of the independent variables of our model. The value range of R-square is from 0 to 1. The closer R-square is to 1, the more accurate the model is.

We test the prediction results with and without the proposed variable in Table 4. The accuracy of the model is 0.6032 without the variable "transfer accessibility" and 0.6935 with this proposed variable. The combination of the four variables we proposed can obtain higher accuracy. The model we proposed performs well, not only for explaining the data but also for predicting the transfer flows.

Furthermore, we use a residual plot to show the residuals on the vertical axis and the independent variable on the horizontal axis. As shown in Figure 13, the points in a residual plot are randomly dispersed around the horizontal axis, which proves that our linear regression model is appropriate for the data.

We also calculate the F-test [43] to evaluate the accuracy of the model. Our testing approach is illustrated as follows. We start with two hypotheses. H_0 is the null hypothesis that the lagged-variable model does not explain the variance in the transfer flow better than the intercept-only model. H_1 is the

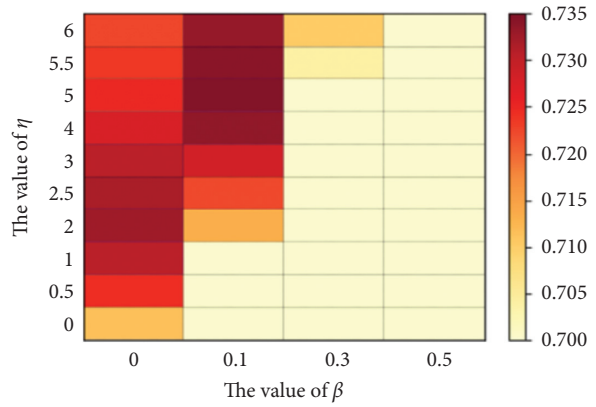

Figure 10: The accuracy of different parameter estimates on the regression model.

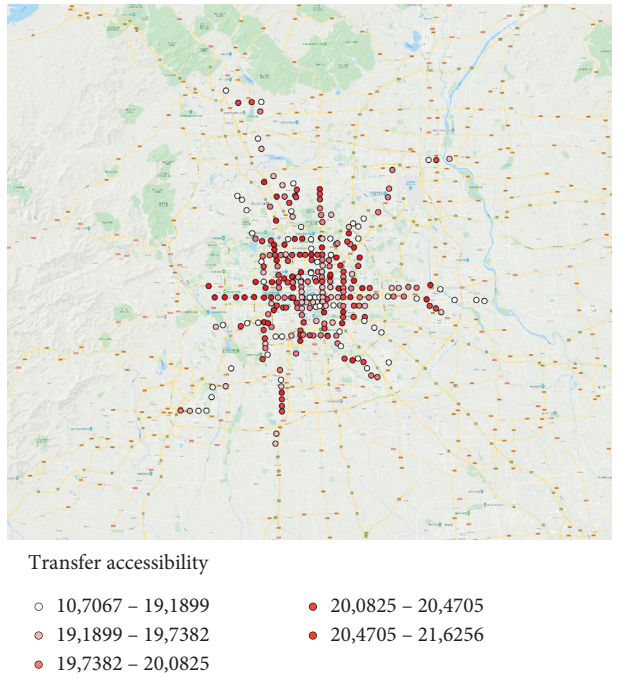

Figure 11: Transfer accessibility of different metro stations.

Table 3: The correlation between the independent variables.

Coef.		$x1$	$x2$	$x3$	$x4$
Transfer accessibility	$x1$		-0.2455	0.2616	0.5733
Transfer time	$x2$	—		0.0158	-0.3173
The number of bus stops around each metro station	$x3$		—		-0.0165
The number of bus lines per bus stop	$x4$			—	—

Table 4: The accuracy of the model with the different variable combination.

Variables	R^2
$x1x2x3,x4$	0.6935
$x2x3,x4$	0.6032
$x1x3,x4$	0.6333
$x1x2,x3$	0.6736
$x1x2,x4$	0.6875
$x1$	0.6170
$x2$	0.1739
$x3$	0.1020
$x4$	0.4506

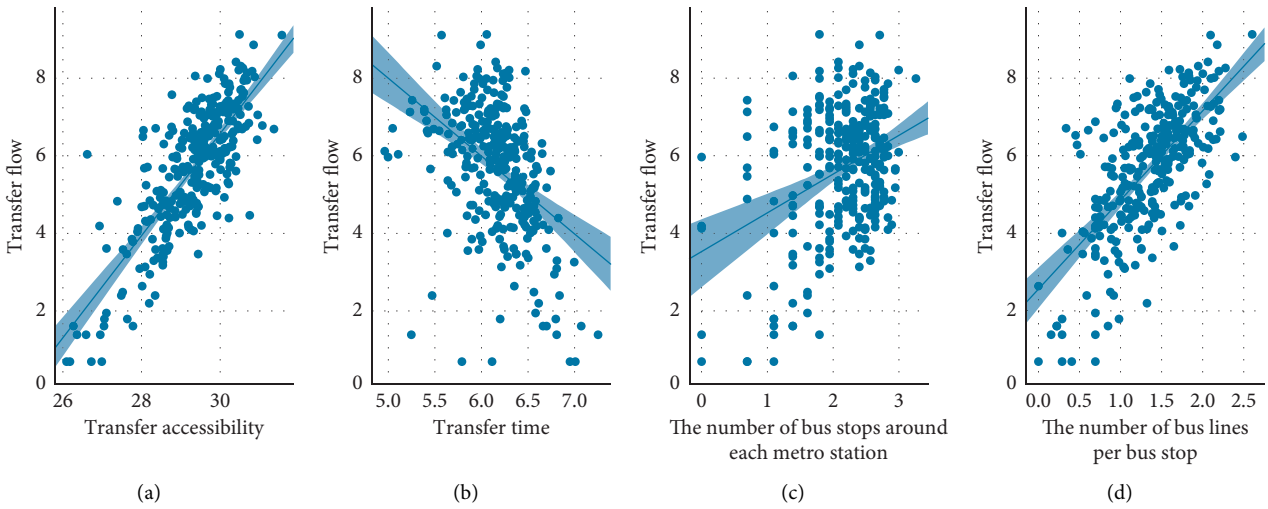

Figure 12: The relation between four attributes and transfer flow: (a) transfer accessibility, (b) transfer time, (c) the number of bus stops around each metro station, and (d) the number of bus lines per bus stop.

Table 5: Estimation results of the regression model based on the training set.

		Coef.	t	$P > \|t\|$
Transfer accessibility	$x1$	0.8701	10.194	0
Transfer time	$x2$	−1.1918	−7.981	0
The number of bus stops around each metro station	$x3$	0.2956	2.908	0.004
The number of bus lines per bus stop	$x4$	0.7236	4.691	0
No. of Observations			214	
R^2			0.6935	

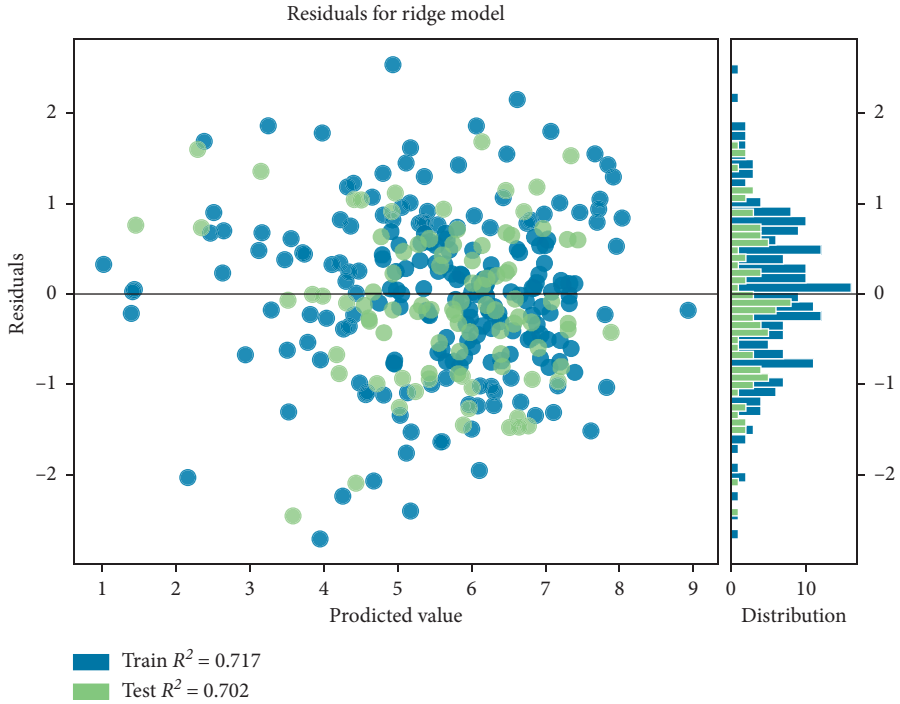

Figure 13: Residuals for ridge model.

alternate hypothesis that the lagged-variable model is better. We apply the F-test on the two models. In our example, the p value is 1.11e-80, which is an extremely small number. There is less than 1% chance that the F-statistic of 188.6 could have occurred by chance under H_0. Thus, we reject the Null hypothesis and accept the alternate hypothesis H_1 that the complex model can explain the variance in the dependent variable better than the intercept-only model.

5. Conclusion

In this paper, we have developed a regression model to explain how network-related attributes can be used to model transfer flow in a multimodal PT network. We conducted our case study in a joint bus and metro network in Beijing and several properties were shown to influence transfer flow between these two modes, namely, transfer accessibility, transfer time, and the number of bus lines per bus stop. Among them, the most important property we proposed was transfer accessibility, which was defined to represent the radiation of a station as a transferring hub, given its position in a multimodal PT network.

We believe that our method could be used not only for explaining transfer flow at existing stations but also for predicting transfer flow at newly planned stations. It provides a parsimonious alternative to existing passenger assignment models, which are mostly expensive, given the modeling required as well as data hungriness. Our model can be directly applied to the evaluation of the transfer flow at a new station in Beijing. The model can also be used for other cities as long as they have the same data available as we had, including smart card data, network data, and POI data. The innovation of our study lies in the new approach to modeling passenger transfer flow based on network properties. Also, transfer accessibility is a new concept, which might be useful for other PT research as well.

This work can still be improved in a few ways. Firstly, several features can be added to the existing methodology in the future. Cities with different sizes and thus with different PT network scales can be used to further validate the findings of this paper. Secondly, the number of passengers depends on the time and period. One can consider the temporal effects on transfer flow in future research. Finally, one-time transfers between metro and bus are our research focus in this paper, since it accounts for the majority of the transfers between metro and bus, but it would be interesting to explore the transferability of our model to other complex transfer types in the future.

Data Availability

The data can only be shared internally within the institute where the first author works.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] B. Si, M. Zhong, J. Liu, Z. Gao, and J. Wu, "Development of a transfer-cost-based logit assignment model for the Beijing rail transit network using automated fare collection data," *Journal of Advanced Transportation*, vol. 47, p. 512, 2010.

[2] R. Schakenbos, L. L. Paix, S. Nijenstein, and K. T. Geurs, "Valuation of a transfer in a multimodal public transport trip," *Transport Policy*, vol. 46, pp. 72–81, 2016.

[3] L. Groenendijk, J. Rezaei, and G. Correia, "Incorporating the travellers' experience value in assessing the quality of transit nodes: a Rotterdam case study," *Case Studies on Transport Policy*, vol. 6, no. 4, pp. 564–576, 2018.

[4] Z. Guo and N. H. M. Wilson, "Assessing the cost of transfer inconvenience in public transport systems: a case study of the London Underground," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 2, pp. 91–104, 2011.

[5] A. Fonzone, J.-D. Schmöcker, and F. Viti, "New services, new travelers, old models? Directions to pioneer public transport models in the era of big data," *Journal of Intelligent Transportation Systems*, vol. 20, no. 4, pp. 311–315, 2016.

[6] A. Ceder, S. Chowdhury, N. Taghipouran, and J. Olsen, "Modelling public-transport users' behaviour at connection point," *Transport Policy*, vol. 27, pp. 112–122, 2013.

[7] M. Bordagaray, L. dell'Olio, A. Fonzone, and Á. Ibeas, "Capturing the conditions that introduce systematic variation in bike-sharing travel behavior using data mining techniques," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 231–248, 2016.

[8] A. Gavriilidou and O. Cats, "Reconciling transfer synchronization and service regularity: real-time control strategies using passenger data," *Transportmetrica A: Transport Science*, vol. 15, no. 2, pp. 215–243, 2019.

[9] X. Ji, X. Zhou, and B. Ran, "A cell-based study on pedestrian acceleration and overtaking in a transfer station corridor," *Physica A: Statistical Mechanics and Its Applications*, vol. 392, no. 8, pp. 1828–1839, 2013.

[10] D. Akin, "Model for estimating increased ridership caused by integration of two urban transit modes: case study of metro and bus-minibus transit systems, istanbul, Turkey," *Journal of the Transportation Research Board*, vol. 1986, pp. 162–171, 2007.

[11] Y. Sun, X. Sun, B. Li, and D. Gao, "Joint optimization of a rail transit route and bus routes in a transit corridor," *Procedia - Social and Behavioral Sciences*, vol. 96, pp. 1218–1226, 2013.

[12] C. Seaborn, J. Attanucci, and N. H. Wilson, "Analyzing multimodal public transport journeys in London with smart card fare payment data," *Transportation Research Record*, vol. 2121, no. 1, pp. 55–62, 2009.

[13] M. D. Yap, O. Cats, N. Van Oort, and S. P. Hoogendoorn, "A robust transfer inference algorithm for public transport journeys during disruptions," *Transportation Research Procedia*, vol. 27, pp. 1042–1049, 2017.

[14] L. Du, G. Song, Y. Wang, J. Huang, M. Ruan, and Z. Yu, "Traffic events oriented dynamic traffic assignment model for expressway network: a network flow approach," *IEEE Intelligent Transportation Systems Magazine*, vol. 10, no. 1, pp. 107–120, 2018.

[15] C. Wang and H. Chen, "A trip chain based user equilibrium traffic assignment model with flexible activities scheduling order," *Journal of Traffic and Transportation Engineering*, vol. 4, pp. 1–10, 2016.

[16] T. Gosens and J. Rouwendal, "Nature-based outdoor recreation trips: duration, travel mode and location,"

Transportation Research Part A: Policy and Practice, vol. 116, pp. 513–530, 2018.

[17] P. V. Subba Rao, P. K. Sikdar, K. V. Krishna Rao, and S. L. Dhingra, "Another insight into artificial neural networks through behavioural analysis of access mode choice," *Computers, Environment and Urban Systems*, vol. 22, no. 5, pp. 485–496, 1998.

[18] Y. Hadas, "Assessing public transport systems connectivity based on Google Transit data," *Journal of Transport Geography*, vol. 33, pp. 105–116, 2013.

[19] D. Luo, O. Cats, and H. van Lint, "Can passenger flow distribution be estimated solely based on network properties in public transport systems?" *Transportation*, vol. 47, Article ID 0123456789, 2019.

[20] S. Xiao, X. C. Liu, and Y. Yinhai Wang, "Data-driven geospatial-enabled transportation platform for freeway performance analysis," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 2, pp. 10–21, 2015.

[21] Y. Wang, G. Correia, E. de Romph, and B. F. Santos, "Road network design in a developing country using mobile phone data: an application to Senegal," *IEEE Intelligent Transportation Systems Magazine*, vol. 31, no. 15, pp. 2–15, 2018.

[22] X. Chen, L. Yu, Y. Zhang, and J. Guo, "Analyzing urban bus service reliability at the stop, route, and network levels," *Transportation Research Part A: Policy and Practice*, vol. 43, no. 8, pp. 722–734, 2009.

[23] Y. Chen, B. Mao, Y. Bai, T. K. Ho, and Z. Li, "Timetable synchronization of last trains for urban rail networks with maximum accessibility," *Transportation Research Part C: Emerging Technologies*, vol. 99, 2019.

[24] S. Goh, K. Lee, J. S. Park, and M. Y. Choi, "Modification of the gravity model and application to the metropolitan Seoul subway system," *Physical Review E-Statistical, Nonlinear, and Soft Matter Physics*, vol. 86, no. 2, pp. 1–6, 2012.

[25] J. D. Ortúzar and L. G. Willumsen, *Modelling Transport*, Wiley, Hoboken, NJ, USA, 2011.

[26] K. Benoit, *Linear Regression Models with Logarithmic Transformations*, London School of Economics, London, UK, pp. 1–8, 2011.

[27] M. S. Chowdhury and S. I.-J. Chien, "Joint optimization of bus size, headway, and slack time for efficient timed transfer," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2218, no. 1, pp. 48–58, 2011.

[28] J. Gutiérrez, O. D. Cardozo, and J. C. García-Palomares, "Transit ridership forecasting at station level: an approach based on distance-decay weighted regression," *Journal of Transport Geography*, vol. 19, no. 6, pp. 1081–1092, 2011.

[29] X. Yang, H. Yin, J. Wu, Y. Qu, Z. Gao, and T. Tang, "Recognizing the critical stations in urban rail networks: an analysis method based on the smart-card data," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 1, pp. 29–35, 2019.

[30] A. Viallard, M. Trépanier, and C. Morency, "Assessing the evolution of transit user behavior from smart card data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 4, pp. 184–194, 2019.

[31] I. Semanjski, S. Gautama, R. Ahas, and F. Witlox, "Spatial context mining approach for transport mode recognition from mobile sensed big data," *Computers, Environment and Urban Systems*, vol. 66, pp. 38–52, 2017.

[32] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.

[33] J. B. Gordon, H. N. Koutsopoulos, N. H. Wilson, and J. P. Attanucci, "Automated inference of linked transit journeys in London using fare-transaction and vehicle location data," *Transportation Research Record*, vol. 2343, no. 1, pp. 17–24, 2013.

[34] N. Nassir, M. Hickman, and Z.-L. Ma, "Activity detection and transfer identification for public transit fare card data," *Transportation*, vol. 42, no. 4, pp. 683–705, 2015.

[35] Y. Wang, G. H. d. A. Correia, E. de Romph, and H. J. P. Timmermans, "Using metro smart card data to model location choice of after-work activities: an application to Shanghai," *Journal of Transport Geography*, vol. 63, pp. 40–47, 2017.

[36] L. Codeca, R. Frank, S. Faye, and T. Engel, "Luxembourg SUMO traffic (LuST) scenario: traffic demand evaluation," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 2, pp. 52–63, 2017.

[37] Y. Zheng, Y. Zhang, and L. Li, "Reliable path planning for bus networks considering travel time uncertainty," *IEEE Intelligent Transportation Systems Magazine*, vol. 8, no. 1, pp. 35–50, 2016.

[38] Ministry of Transport of the China, *National Report on Urban Passenger Transport Development*, Ministry of Transport of the China, Beijing, China, 2018.

[39] Ministry of Transport of the China, *National Report on Urban Passenger Transport Development*, Ministry of Transport of the China, Beijing, China, 2017.

[40] Y. Zhou, L. Yao, Y. Chen, Y. Gong, and J. Lai, "Bus arrival time calculation model based on smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 81–96, 2017.

[41] Y. Wang, G. H. d. A. Correia, B. van Arem, and H. J. P. Timmermans, "Understanding travellers' preferences for different types of trip destination based on mobile internet usage data," *Transportation Research Part C: Emerging Technologies*, March, vol. 90, pp. 247–259, 2018.

[42] J. D. Rodríguez, A. Pérez, and J. A. Lozano, "Sensitivity analysis of *K*-fold cross validation in prediction error estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 569–575, 2010.

[43] R. Christensen, "Significantly InsignificantFTests," *The American Statistician*, vol. 57, no. 1, pp. 27–32, 2003.