Examining the relationship between built environment and metro ridership at station-to-station level

Citation for published version (APA):

Gan, Z., Yang, M., Feng, T., & Timmermans, H. J. P. (2020). Examining the relationship between built environment and metro ridership at station-to-station level. *Transportation Research Part D: Transport and Environment*, 82, [102332]. <https://doi.org/10.1016/j.trd.2020.102332>

Document license:

TAVERNE

DOI:

[10.1016/j.trd.2020.102332](https://doi.org/10.1016/j.trd.2020.102332)

Document status and date:

Published: 01/05/2020

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Examining the relationship between built environment and metro ridership at station-to-station level

Zuoxian Gan^a, Min Yang^{b,*}, Tao Feng^c, Harry J.P. Timmermans^{c,d}

^a College of Transportation Engineering, Dalian Maritime University, Dalian 116026, China

^b Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, Nanjing 211189, China

^c Urban Planning and Transportation Group, Department of the Built Environment, Eindhoven University of Technology, 5600MB Eindhoven, the Netherlands

^d Department of Air Transportation Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

ARTICLE INFO

Keywords:

Built environment
Station-to-station ridership
Non-linear effect
Gradient boosting regression trees
Metro

ABSTRACT

Very few studies have examined the impact of built environment on urban rail transit ridership at the station-to-station (origin-destination) level. Moreover, most direct ridership models (DRMs) tend to involve simple a prior assumed linear or log-linear relationship in which the estimated parameters are assumed to hold across the entire data space of the explanatory variables. These models cannot detect any changes in the linear (or non-linear) effects across different values of the features of built environment on urban rail transit ridership, which possibly induces biased results and hides some non-negligible and detailed information. Based on these research gaps, this study develops a time-of-day origin-destination DRM that uses smart card data pertaining to the Nanjing metro system, China. It applies a gradient boosting regression trees model to provide a more refined data mining approach to investigate the non-linear associations between features of the built environment and station-to-station ridership. Data related to the built environment, station type, demographics, and travel impedance including a less used variable – detour, were collected and used in the analysis. The empirical results show that most independent variables are associated with station-to-station ridership in a discontinuous non-linear way, regardless of the time period. The built environment on the origin side has a larger effect on station-to-station ridership than the built environment on the destination side for the morning peak hours, while the opposite holds for the afternoon peak hours and night. The results also indicate that transfer times is more important variables than detour and route distance.

1. Introduction

Experiences with the historical high automobile dependence in developed countries, such as the United States and some European countries, indicate that car-based transportation systems have resulted in a series of social and environmental challenges, including urban sprawl, traffic congestion, land waste, energy consumption, and air pollution. For example, accounting for one-fourth to one-third of total CO₂ emissions from fossil fuels makes the transport sector the second largest source of CO₂ emissions (IEA, 2016). Furthermore, road traffic accounts for about 75% of transportation-related CO₂ emissions (Ao et al., 2019). As an effective and

* Corresponding author.

E-mail addresses: zxgan@dlmu.edu.cn (Z. Gan), yangmin@seu.edu.cn (M. Yang), T.Feng@tue.nl (T. Feng), H.J.P.Timmermans@tue.nl (H.J.P. Timmermans).

<https://doi.org/10.1016/j.trd.2020.102332>

sustainable option to counter sprawl, reduce traffic congestion and cut down air pollution, transit-oriented development (TOD), catering to the development ideology of Smart Growth and New Urbanism, has become popular in urban planning practice (Cervero et al., 2002, 2004). Consequently, the importance of public transit usage has been spotlighted over the past decades. As one of the most popular alternatives to automobile dependence, urban rail transit has been rapidly developed in many cities around the world (Baum-Snow and Kahn, 2000; Loo and Li, 2006). Due to substantial economic growth, population explosion, availability of cheaper vehicles, and a faster increase in motorized transportation, developing countries are more likely to be lured and trapped in the malignant development of automobile dependence if the local governments do not prioritize public transit over the automobile. Public transit priority policy has been advocated in many countries and the construction upsurge of urban rail transit is expanding from the western countries to emerging markets (Zhao et al., 2014). For example, by the end of 2018, the total length of urban rail transit lines in China reached 5766.6 km, involving in 35 cities.

For the analysis of project viability and sustainability of urban rail transit, it is of paramount importance for policymakers to predict urban rail transit ridership (Cardozo et al., 2012). Therefore, in addition to the general four-step models (McNally, 2000) and activity-based models (Rasouli and Timmermans, 2014a,b), several dedicated ridership models have been developed. Although these models lack a behavioral foundation, are aggregate in nature and do not consider competition among modes, their data requirements are minimal and they are easy to apply. Recently, with the available travel data derived from smart cards and widely available geographic information systems (GIS), these so-called DRMs have become quite popular in transportation planning practice. These models are typically based on a priori assumed statistical relationships between ridership and aggregate areal data. Numerous studies have investigated the effects of the built environment on urban rail transit ridership based on such statistical associations (e.g., Kuby et al., 2004; Sohn and Shim, 2010; Gutiérrez et al., 2011; Choi et al., 2012; Cardozo et al., 2012; Zhao et al., 2014; Durning and Townsend, 2015; Jun et al., 2015; Liu et al., 2016; Kepaptsoglou et al., 2017; Gan et al., 2019b; Liu et al., 2019). These empirical studies typically assume that population/employment density, presence of residential/employment areas, land use mix, road density, intersection number, intermodal connection, and the distance to city center are key factors influencing transit ridership. Originally, many of these models rely on ordinary least squares regression analysis, assuming the relationship with ridership is linear or loglinear. However, realizing ridership data violate the assumptions underlying OLS regression analysis, more advanced regression models have been applied. For example, Chu (2004) adopted Poisson regression model to examine the effects of features of the built environment on urban transit station boarding; Kuby et al. (2004) and Estupiñán and Rodríguez (2008) used two-stage least squares (2LSL) to predict transit station ridership; Sohn and Shim (2010) applied structural equation model to identify direct and indirect relationships among variables in generating metro station ridership.

Many studies (hundreds if not thousands) have touched on the relationships between urban rail transit ridership at the station level and the built environment. However, in sharp contrast, very few studies have investigated the impact of the built environment on urban rail transit ridership at the station-to-station level. To the best of the authors' awareness, only four recent studies published in mainstream academic journals in transportation explored how features of the built environment affect urban rail transit ridership at the station-to-station level (Duncan, 2010; Choi et al., 2012; Zhao et al., 2014; Iseki et al., 2018). This huge difference in the number of urban rail transit ridership studies at the station level and at the station-to-station level suggests that the influence of the built environment on station-to-station-level ridership warrants further exploration. Moreover, station-to-station analysis enables urban and transportation planners to distinguish the impact of origin factors and destination factors on urban rail transit ridership from destination factors. All four abovementioned studies employed a priori assumed multiplicative model (or the hierarchical model based on multiplicative form) to investigate the effects of various features of the built environment on urban rail transit ridership at the station-to-station level.

Since early this century, the commonly used methods in travel behavior analysis, based on the general linear model have slowly but gradually been supplemented by data mining methods. The potential relevance of these methods has increased with the emergence of big data, such as smart cards. Rather than a priori assuming a particular function that is fitted to the data, data mining methods explore associations in the data, which tends to result in more refined, non-linear relationships between the dependent and independent variables. Perhaps the most ambitious model along these lines, is Albatross (Arentze and Timmermans, 2004), which is a complex computational process model, fully based on probabilistic decision trees (Arentze and Timmermans, 2007). Other examples include Bayesian networks (Janssens et al., 2004), automatic interaction detection (Strambi and Van De Bilt, 1998), multivariate adaptive regression splines (Rasouli and Timmermans, 2012), random forests (e.g., Rasouli and Timmermans, 2014b; Zong and Zhang, 2019), dynamic decision trees (Kim, et al., 2018) and gradient boosting regression trees (e.g., Ding et al., 2016, 2018), to name a few data mining methods. Despite the increasing popularity of these methods, accumulated experiences in travel behavior in general and in predicting ridership in particular with data mining methods are still relatively limited.

Against this background, this study aims to enrich the existing literature on transit ridership by investigating in a more refined manner than the aggregate statistical models based on a single functional form the effects of the built environment on urban rail transit ridership at the station-to-station level by mining smart card data using a gradient boosting regression tree algorithm (GBRT). Although this algorithm has been gradually introduced into travel demand research, studies on disentangling the non-linear relationships between the features of built environment and travel demand by using data mining algorithms are still in their infancy. It is relevant therefore to explore whether this data mining algorithm outperforms traditional regression methods in travel demand research. Therefore, this study also contributes to the body of knowledge about the use of data mining algorithms in transportation research.

To achieve this objective, one-month smart card data of Nanjing Metro, China, land use data from the local urban planning bureau and road network data derived from the OpenStreetMap were collected. The ridership at the station-to-station level is analyzed for four time periods, namely morning peak hours, midday, afternoon peak hours and night. Then, both a conventional DRM –

multiplicative model, and the gradient boosting regression trees algorithm are applied to assess the relative importance of different features of the built environment on transit ridership and examine in a more refined manner how the built environment influences ridership at the station-to-station level, while controlling for station type at both the origin and destination, demographic characteristics, and travel impedance. The relative importance and partial dependence plots of each independent variable are presented.

Thus, the main contributions of the present study concern the following aspects. First, this study enriches the sparse existing literature on urban rail transit ridership at the station-to-station level through a comprehensive understanding of the effects of various influential factors, including built environment, station type, demographics characteristics, and travel impedance for different time periods (AM and PM rush hours, midday, and night). Second, this study adds to the limited experience in travel behavior research with the application of GBRT in forecasting station-to-station ridership. The remainder of this paper is organized as follows. [Section 2](#) presents a literature review of the relationship between urban rail transit ridership and the built environment. [Section 3](#) gives an overview of the study area and data collection. The modeling approach is introduced in [Section 4](#), while [Section 5](#) provides the results. The final section draws conclusions and discusses implications of this study's findings.

2. Literature review

2.1. Station level studies

Urban rail transit ridership analysis is generally performed at the station level, and hence the impact of features of the station-area built environment on station-level urban rail transit ridership received much attention. Higher population and employment density increase the likelihood that individuals use urban rail transit. Using cross-sectional data on average weekday boarding of rail transit from nine US cities, [Kuby et al. \(2004\)](#) applied the OLS model to investigate the relationships between features of the built environment and ridership at the station level. Both population and employment density were found to be significantly and positively associated with the number of urban rail boardings. In a similar fashion, [Durning and Townsend \(2015\)](#), and [Zhao et al. \(2014\)](#) found that station-area population and employment density have significant and positive effects on urban rail transit ridership at the station level in developed countries such as Canada, and in developing countries such as China. However, a recent study based on the GBRT model revealed that population density and employment density play key roles in station ridership only when they fall into specific ranges, while they have trivial effects on station ridership when they exceed these ranges ([Ding et al., 2019](#)).

As for land use mix, the findings are mixed. By applying an integration of distance-decay functions and a multiple regression model, respectively a mixed geographically weighted regression model, respectively, [Gutiérrez et al. \(2011\)](#) and [Jun et al. \(2015\)](#) found this variable to be positively associated with urban rail transit ridership. By contrast, [Ryan and Frank \(2009\)](#), [Cardozo et al. \(2012\)](#), [Durning and Townsend \(2015\)](#) and [Liu et al. \(2016\)](#) found that the effect of land use mix on urban rail transit ridership is insignificant. [Ding et al. \(2019\)](#) further pointed out that the effect of land use mix on station ridership becomes effective only when this index is larger than 0.5. Other studies explored the association between specific land use density and station-level urban rail transit ridership. For example, using data collected in Canada's five largest cities, [Durning and Townsend \(2015\)](#) applied a bootstrapped OLS regression model to investigate the impact of specific land use densities on the number of station boardings, controlling for station attributes, service attributes and socioeconomic characteristics. Residential ratio, commercial ratio, and government-institutional ratio were found to have significantly positive effects on ridership, while the impact of the open area ratio, park area ratio and resource-industrial ratio were insignificant. Some other studies examined the relationship between the proportion of different types of land use and urban rail transit ridership, yielding inconsistent and fragmented findings (e.g., [Sohn and Shim, 2010](#); [Sung and Oh, 2011](#); [Zhao et al., 2014](#); [Tu et al., 2018](#); [Gan et al., 2019b](#)).

Yet other studies focused on the impact of street network design on urban rail transit ridership since street network design has been shown to influence individuals' willingness to walk to urban rail transit stations and therefore ridership ([Ewing and Cervero, 2010](#); [Liu et al., 2020](#)). Street network characteristics such as intersection density, road length, and road density are normally selected as independent variables and have been proved to be significantly associated with urban rail transit ridership. For instance, using one-week data for all stations in the Shenzhen metro and bus systems, [Tu et al. \(2018\)](#) found that increased road density encourages a larger number of people to travel by bus and metro. Similar studies conducted by [Zhao et al. \(2014\)](#) and [Gan et al. \(2019b\)](#) also concluded that road density (or road length) is positively and significantly associated with urban rail transit station ridership. [Durning and Townsend \(2015\)](#) observed a positive relationship between rapid transit station ridership and intersection density based on data collected in Canada's five largest cities. By contrast, [Jun et al. \(2015\)](#) showed that intersection density has a significantly negative impact on metro station ridership in Seoul, South Korea, while [Ding et al. \(2019\)](#) found that intersection density is positively associated with station ridership only when it is within the range of 10–18.

Distance to city center, a measure of regional accessibility, has also been examined in the literature. In two empirical studies pertaining to Nanjing, China, the effects of distance to city center on metro station ridership have been found insignificant ([Zhao et al., 2014](#); [Gan et al., 2019b](#)). However, based on the daily boardings data of light rail transit for the Baltimore Metro collected, Maryland, distance to city center was found to have a negative effect on ridership. Using Metrorail data for the Washington metropolitan area, [Ding et al. \(2019\)](#) found that distance to city center has a non-linear effect on ridership. Once the distance to the city center exceeds 5 km, the influence of this independent variable become trivial. Another aspect of the built environment, namely intermodal connection factors such as car park and rides, and number of bus lines or bus stops has also been investigated in many studies ([Sohn and Shim, 2010](#); [Cardozo et al., 2012](#); [Zhao et al., 2014](#); [Durning and Townsend, 2015](#); [Liu et al., 2016](#); [Ding et al., 2019](#)). In general, these variables tend to be positively correlated with urban rail transit ridership.

Urban rail transit ridership is affected not only by the built environment, but also by station type. Therefore, station type

variables, mainly terminal and transfer stations have been included in several studies. Compared to other types of stations, terminal and transfer stations were found to have a higher ridership due to their larger catchment area (Kuby et al., 2004; Zhao et al., 2014; Ding et al., 2019). Prior studies also documented how socio-economic characteristics such as median household income within station catchment area, car-free households, ethnicity (e.g., share of White population), age and gender balance influence urban rail transit station ridership (Ryan and Frank, 2009; Durning and Townsend, 2015; Jun et al., 2015; Liu et al., 2016; Ding et al., 2019).

2.2. Station-to-station level studies

Studies on transit ridership at the station-to-station level can be regarded as an extension of studies on ridership at the station level. The selected explanatory variables are the same: built environment characteristics, station type, and socio-economic characteristics. However, these variables are calculated for both origin and destination. Therefore, the DRM applied in station-to-station ridership analyses is usually called origin-destination DRM. Moreover, station-to-station ridership is also influenced by travel impedance (trip-specific) variables such as transfer times, and route distance. For example, based on the number of weekday riders for 2002 extracted from Bay Area Rapid Transit System, Duncan (2010) employed a multiplicative model and a Poisson model to find that the numbers of housing units on the origin side and the number of jobs on the destination side, the number of connecting buses on both sides, and a terminal dummy are positively associated with station-to-station ridership. Moreover, he found that transfer times, fare, and route distance have significant and negative effects on station-to-station ridership. Choi et al. (2012), Zhao et al. (2014), and Iseki et al. (2018) developed the origin-destination DRM by time of day and investigate the effects of independent variables on station-to-station ridership for different time periods. They concluded that features of the built environment correlated with station-to-station ridership vary by between origin and destination and across time of day. For instance, Choi et al. (2012) found that during AM peak hours the population on the origin side plays a key role in station-to-station ridership, while during PM peak hours the population on the destination side is highly associated with station-to-station ridership. The results indicate that the effects of residential area, business/office area, employment, and a CBD dummy on station-to-station ridership also vary by trip sides and time of day. Similar conclusions can be found in Zhao et al. (2014) and Iseki et al. (2018).

Choi et al. (2012) examined the impacts of land use mix at the origin and destination on station-to-station ridership, but found that there is no significant relationship between land use mix and station-to-station ridership, regardless of trip side and time of day. However, they pointed out that providing good bus connection services on both the origin and destination sides at all time periods can increase metro ridership. The results reported in Zhao et al. (2014) support this finding. In addition, all the above-mentioned studies showed that travel impedance variables such as transfer times, travel time, route distance, and fare have significantly negative effects on station-to-station ridership. Although these studies made a substantial stride in forecasting urban transit ridership at the station-to-station level, all assume that independent variables have a pre-defined association (i.e. log-linear) with station-to-station ridership. These models assume that the estimated parameter of each independent variable applies to the full range of that variable. However, these models do not allow to explore any more refined, discontinuous non-linear effects on ridership.

In summary, this literature review on urban rail transit ridership analysis suggest that: (1) Most prior studies on urban rail transit ridership have been focused on the station level, while studies on ridership at the station-to-station level are rare; (2) The effects of features of the built environment on urban rail transit ridership received much attention at both station level and station-to-station level. While, station-to-station ridership is also influenced by the travel impedance and few studies have examined the impact of detour on transit ridership; (3) Linear and log-linear regression methods have been the primary and most widely used models for ridership analysis at the station level and station-to-station level. These models assume that independent variables follow a consistent pre-defined association (e.g., linear or log-linear) with transit ridership; (4) Data mining allows exploring the existence of discontinuous non-linear relationships. Only one recent study (Ding et al., 2019) has applied this approach. At the same time, a growing number of studies have shown that the effect of a variable on travel behavior may differ for different ranges of this variable (e.g., Rasouli and Timmermans, 2014a,b; Ding et al., 2018, 2019; Wu et al., 2019). Therefore, more refined, data mining of smart card data may be a more promising way of examining the relationship between features of the built environment and station-to-station ridership.

To this end, this study uses data derived from various sources in Nanjing City, China, and applies a data mining algorithm to investigate the refined, discontinuous non-linear relationships between built environment and station-to-station ridership for different time periods, while controlling for station type, demographics characteristics, and travel impedance including transfer times, detour, and route distance.

3. Study area and data collection

In recent years, transportation infrastructure construction in big and mega Chinese cities has rapidly expanded in response to urban sprawl and increasing travel demand. Nanjing is one of such cities where urban rail transit is widely used after the completion of its first metro line (Line 1) in September 2005. By 2020, there will be 15 urban rail transit lines in operation or under construction and the total mileage will reach 520.2 km. With all 258 stations, including 71 transfer stations, the share of passenger volume of public transportation accounting for the total motorized travel will grow up to around two-thirds, and urban rail transit will account for 45% of total passenger volume of public transportation in the year 2020.¹ In 2015, six metro lines were in operation with 112

¹ http://www.sdpc.gov.cn/zcfb/zcfbtz/201505/t20150520_692548.html.

Fig. 1. Study area and Nanjing metro system in 2015.

stations and 225 km tracks, mainly serving the inner-city and connecting the core districts and the periphery (Fig. 1). In 2015, the entire annual ridership of the metro system exceeded 717 million, accounting for more than 34% of the public transportation passenger volume.

However, despite of the impressive number of users, the development of the metro system is still confronted by various problems. For example, according to one-month SCD data derived from Nanjing Metro Corporation (NMC) in April 2015, Xinjiekou Station (the transfer station of metro lines 1 and 2) has the largest average boarding (119,912 passengers) and alighting (124,380 passengers), while approximately 26 stations have less than 3,000 boarding and alighting per day. Moreover, less than 6% of the OD-station pairs account for more than 48% of the station-to-station passenger volume in the morning and afternoon rush hours. Considering the goal of optimizing metro usage, it is necessary to predict the travel demand between origin and destination stations, and the effects of influential factors.

The SCD records were obtained from the AFC system of Nanjing metro system and include all trips paid for one-way tickets,

Table 1
Descriptive statistics for the independent and dependent variables.

Variables	Mean	S.D.	Min	Max	Source
OD flows					
7:00–9:00	31.102	99.775	1	4,281	NMC
11:00–13:00	13.855	36.978	1	958	NMC
17:00–19:00	27.247	75.908	1	2,612	NMC
21:00–23:00	10.200	39.325	1	1,986	NMC
Built environment					
Population density (1000 person per km ²)	10.21	9.51	0.57	46.67	Original data from NUPB & measured in GIS
Residential ratio (%)	28.08	18.71	0	65.91	Original data from NUPB & measured in GIS
Business/commercial ratio (%)	9.18	9.40	0	61.59	Original data from NUPB & measured in GIS
Industrial/manufacturing ratio (%)	9.93	12.07	0	54.41	Original data from NUPB & measured in GIS
Land use mix	0.54	0.22	0.01	0.86	Original data from NUPB & measured in GIS
Road density (km/km ²)	9.12	2.81	2.31	16.87	Original data from OpenStreetMap & measured in GIS
Number of intersections	12.56	7.48	2	39	Original data from OpenStreetMap & measured in GIS
Number of bus lines	9.74	7.16	0	44	Original data from Baidu Map & measured in GIS
Distance to city center (km)	16.17	12.47	0	60.38	Measured in GIS
Station type					
Terminal dummy	N/A	N/A	0	1	Observed
Transfer dummy	N/A	N/A	0	1	Observed
Demographics					
Housing price (*10 ³ yuan/m ²)	27.96	10.54	6.83	56.48	Original data from Lianjia.com & measured in GIS
Travel impedance variables					
Transfer times	1.08	0.75	0	3	Observed
Detour	1.42	0.46	1	8.26	Measured in GIS
Route distance (km)	25.46	16.69	0.77	100.53	Measured in GIS

registered and anonymous cards. The total number of trip records in April 2015 was more than 43 million. Each record includes ticket id, tap-in time and station id, tap-out time and station id, and travel duration, while personal information (e.g. age, gender and education) is not available. Therefore, station-to-station ridership can be counted by the time of day. We focus on ridership on weekday; the records of non-weekdays are not considered in the present paper. The average daily weekday station-to-station ridership for four time periods, morning rush hours (7:00–9:00), midday (11:00–13:00), afternoon rush hours (17:00–19:00), and night (21:00–23:00) is used, eliminating the possible impact of extraordinary events in a specific day. Only station-to-station ridership (OD flows) which is equal to or larger than 1 is analyzed in the study. Table 1 displays the data summary for the four selected time spans. It can be observed that station-to-station ridership during rush hours is higher than for midday and night. It also shows significant differences in the OD flows distribution between different OD station pairs. For example, during the morning rush hours (7:00–9:00), one OD station pair has the largest average flow with 4,281 passengers, while 41% OD station pairs have < 5 passengers.

To investigate the correlations between metro station-to-station ridership and the built environment, while controlling for station type, demographics, and travel impedance attributes, data were collected from various sources such as NUPB (Nanjing Urban Planning Bureau), OpenStreetMap, Baidu Map, and Lianjia.com in the year of 2015. The main variables related to the built environment are commonly measured by the four Ds (Cervero et al., 2004; Ewing and Cervero, 2010; Ao et al., 2019; Ding et al., 2019). In this study, four Ds – density, land use diversity, design, and distance to city center are utilized. Population density is used to measure density, as people's activity demand generates daily travel. Since different categories of land use result in different mobility patterns, the proportions of three main land use types are considered, namely residential, business/commercial, and industrial/manufacturing ratios. In addition, land use mix, which is calculated in terms of the entropy of land use, is included as an index of land use diversity (Cervero et al., 2004; Gan et al., 2019a). Road density, namely the ratio of the total length of roads (km) to the total catchment area of a metro station (km²), and the number of intersections in the catchment area of a metro station are used as indicators of design. The number of bus lines is selected as a measure to judge intermodal connection (cf. Cardozo et al., 2012; Choi et al., 2012; Zhao et al., 2014).

Station type is another important factor influencing ridership (Choi et al., 2012; Durning and Townsend, 2015). In this study, two dummy variables (terminal dummy and transfer dummy) are set. Their references corresponded to the non-terminal stations and non-transfer stations. Demographics characteristics, especially median household income, is generally recognized to impact citywide travel. However, it is difficult to obtain to data on median household income for a small-scale area (i.e., station catchment area, community). Instead, average housing price crawled from Lianjia, a popular real estate web site in China (<http://www.lianjia.com>), is used as a proxy variable for representing household income within stations' catchment areas. Different from ridership analysis at the station level, all above-mentioned independent variables are measured for both origin and destination metro stations.

Transfer times and route distance are utilized to explore the effects of travel impedance factors on station-to-station ridership (cf. Choi et al., 2012; Zhao et al., 2014). In addition, detour is used because it captures different network information. For example, the route distances from Baijiahu Station to Tianyuanxilu Station and from Baijiahu Station to CPU (China Pharmaceutical University) Station are similar (11.2 km vs 11.8 km) (Fig. 2). However, the straight line distance from Baijiahu Station to Tianyuanxilu Station is much smaller than that the distance from Baijiahu Station to CPU Station (1.3 km vs 9.5 km). For this reason, it is likely that many people will choose taking the metro from Baijiahu Station to CPU Station, while not from Baijiahu Station to Tianyuanxilu Station.

Fig. 2. Illustration of the detour between metro stations.

The detour is defined as:

$$\text{detour} = \frac{d_{cd}^{\text{route}}}{d_{cd}^{\text{linear}}} \quad (1)$$

where d_{cd}^{route} denotes the route distance between metro stations c and d and d_{cd}^{linear} is the straight line distance between metro stations c and d . The value of detour is always equal to or larger than 1.

Data on built environment and demographics (housing price) were collected for station catchment areas which are defined as the area within a 800 m walking distance². Apart from terminal dummy, transfer dummy, and transfer times, other independent variables are calculated in a geographic information system (ArcGIS version 10.3). Table 1 summarizes the details of the sample and their source.

4. Methodology

4.1. Multiplicative model

As discussed in the literature review section, the DRM model with a multiplicative form has been commonly applied to investigate the influence of selected explanatory variables by origin and destination separately (Choi et al., 2012; Iseki et al., 2018). The multiplicative model is defined as follows.

$$R_{cd} = \emptyset \prod_{p=1}^P X_{cp}^{\alpha_p} \prod_{p=1}^P X_{dp}^{\beta_p} \prod_{n=1}^N Y_{cdn}^{\theta_n} \quad (2)$$

where R_{cd} denotes the ridership from station c to station d , X_{cp} is the p th explanatory variable of origin station c and X_{dp} is the p th explanatory variable of destination station d , Y_{cdn} denotes the n th travel impedance variable from station c to station d , $\emptyset$ is the scale parameter, α_p , β_p and θ_n are coefficients to be estimated. Eq. (2) can be easily transformed into a linear form by taking the logarithm on both sides of equation. Thus, it has been estimated using conventional methods such as ordinary least squares regression analysis.

4.2. Gradient boosting regression trees model

As discussed, rather than a priori hypothesizing a particular function form such as the loglinear relationship implied by the multiplicative model, the aim of this study is to predict station-to-station ridership in a more refined manner, captured specific non-linear effects for different ranges of the selected explanatory variables. Different data mining methods can be applied to effect. Because our goal is not to derive a behaviorally founded model, but rather to use the smart card to find the best prediction, ensemble methods tend to show best results. Particularly, Friedman's gradient boosting regression trees algorithm is used (Friedman, 2001). Gradient boosting is a machine learning technique, which produces a prediction model in the form of an ensemble of models, in this case regression trees. The objective of the algorithm is to minimize a loss function. The regression tree can be defined as follows:

² Although there is no exact criterion for the distance threshold of stations' catchment areas, 800m (about 0.5mile) is widely accepted in the literature (e.g., Kuby et al., 2004; Zhao et al., 2014; Liu et al., 2016). Moreover, 800m is defined as the threshold distance in an official design guidance of urban rail—"Guidelines for Planning and Design of Urban Rail" released by MOHURD (MOHURD, 2015).

$$f_m(x) = \sum_{j=1}^J b_{jm} I(x \in R_{jm}) \quad (3)$$

where $I(x \in R_{jm}) = \begin{cases} 1, & \text{if } x \in R_{jm} \\ 0, & \text{otherwise} \end{cases}$, J denotes the number of leaves for each tree, b_{jm} is the constant value for the corresponding region R_{jm} , the input space is split into disjoint regions $R_{1m}, R_{2m}, \dots, R_{Jm}$ by the tree.

Thus, a GBRT model uses regression trees as a base and updates the approximation function $f(x)$ by minimizing the expected value of loss function $L(y, f(x))$. The commonly used loss function, namely squared-error loss function $L(y, f(x)) = (y - f(x))^2$ is adopted. Based on the gradient descent direction, the GBRT model with m regression trees is utilized to update $f(x)$ as follows.

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^J \rho_m b_{jm} I(x \in R_{jm}) \quad (4)$$

$$\rho_m = \underset{\rho}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, f_{m-1}(x_i) + \sum_{j=1}^J \rho b_{jm} I(x \in R_{jm})) \quad (5)$$

where ρ_m is estimated by minimizing the expected value of the loss function.

Overfitting can be prevented by controlling the number of iterations, but a more effective strategy is incorporating a shrinkage parameter (also called learning rate) into function $f(x)$. As a simple regularization strategy, shrinkage parameter ξ ($0 < \xi < 1$) can avoid overfitting by scaling the contribution of each tree. Then, Eq. (4) is updated as follows:

$$f_m(x) = f_{m-1}(x) + \xi \sum_{j=1}^J \rho_m b_{jm} I(x \in R_{jm}) \quad (6)$$

Although a smaller shrinkage parameter can reduce the effect of overfitting, more iterations are required to make the training error converge and obtain the optimal model, which will increase the running time of model. Therefore, there is a tradeoff between the value of the shrinkage parameter and the number of iterations M . Empirical evidence suggests that a small ξ ($\xi < 0.1$) with a large M is preferable (Hastie et al., 2005; Zhang and Haghani, 2015). The number of nodes in a tree, namely tree complexity J , is another important parameter for influencing the performance of GBRT models. We usually expect very large trees, but increasing J will raise the computation complexity of model. Hastie et al. (2005) suggests that $4 \leq J \leq 8$ is sufficient and generally works well. Furthermore, they point out that the GBRT model is not sensitive to the exact choice of J within the range of 4–8.

GBRT models cannot produce exact estimated coefficients and confidence intervals, but it can quantify the relative importance of each independent variable based on the optimal model. The relative importance of a variable x_i can be described as follows:

$$\begin{cases} I_{x_i}^2 = \frac{1}{M} \sum_{m=1}^M I_{x_i}^2(T_m) \\ I_{x_i}^2(T_m) = \sum_{j=1}^{J-1} d_j \end{cases} \quad (7)$$

where j is the number of internal nodes, m represents the m th iteration, and d_j denotes the improvement in squared error from making the j th split. The sum of the relative importance of all explanatory variables equals 100%.

Partial dependence plots produced by the GBRT model can visualize the associations between the dependent and independent variables. It illustrates the marginal effect of an explanatory variable on the predicted response while controlling for other independent variables in the given model. Based on the partial dependence plot, we can figure out whether or not independent variables within certain ranges, have stabilizing effects on the OD flows between metro stations. In other words, these plots provide an opportunity to investigate the refined non-linear effects of features of the built environment on station-to-station ridership.

5. Results

Following common practice in data mining, a cross-validation procedure was adopted to develop the GBRT model. The sample was randomly partitioned into five subsets. Four different subsets were used for extracting the model, while the remaining subset was used to test how well the model could predict. Based on the experimental findings of previous studies, we tested models with different values of shrinkage (0.1, 0.05, 0.01, 0.005) and tree complexity (3, 4, 5, 6) in the 5-fold cross-validation procedure. The evaluation results based on pseudo- R^2 showed that the optimal model in different step of 5-fold cross-validation (e.g., former 80% of the entire dataset as training samples and the reminder 20% as testing samples versus former 20% of the entire dataset as testing samples and the reminder 80% as training samples) were corresponding to different shrinkage and tree complexity, while the differences in pseudo- R^2 values were not big. Generally, we found the pseudo- R^2 of the models were relatively high when tree complexity and shrinkage were respectively set as 5 and 0.01 for all the four time periods. Considering the evaluation results and computational time, the three parameters related to shrinkage, tree complexity, and number of trees were set respectively as 0.01, 5 and 10000. After 9,886, 9,861, 9,996 and 8,955 boosting iterations, the models for the four time periods (7:00–9:00, 11:00–13:00, 17:00–19:00 and 21:00–23:00) achieved their best results. The values of testing pseudo- R^2 for the GBRT models were 0.874, 0.913, 0.884 and 0.680,³

³ The R^2 for the testing data tends to be smaller than for the training data due to the generalization ability of model. Therefore, we only report R^2

while the corresponding values for the traditional multiplicative models were 0.623, 0.639, 0.658 and 0.544⁴ (Table 2). It indicates that the superior prediction performance of the GBRT models compared to the multiplicative model.

5.1. Relative importance of independent variables

Fig. 3 shows the relative importance of the independent variables in predicting station-to-station ridership. The results illustrate that the joint contribution of the features of the built environment on the origin side to predicting station-to-station ridership for the four different time periods are about 42%, 33%, 32% and 24%, respectively, whereas on the destination side these variables contribute to 27%, 32%, 37% and 55%, respectively. The collective contribution of the remaining explanatory variables (housing price and station type on both sides, and the travel impedance variables) are 31%, 35%, 31% and 21%, respectively. It is evident that the built environment on the origin side has a larger effect on station-to-station ridership than the destination side has for morning rush hours, while this is reversed for afternoon rush hours and night. Fig. 3 also shows that, except for the night time, the three travel impedance variables collectively contribute to more than 21% of the total effect on station-to-station ridership. It indicates that travel impedance factors play an important role in the prediction of OD flows.

With respect to morning rush hours (7:00–9:00), land use mix and number of bus lines on the origin side, population density on the destination side, and transfer times are the four most important explanatory variables, with contributions more than 10%, respectively. However, for afternoon rush hours (17:00–19:00), population density on the origin side, transfer times, route distance, and number of bus lines on the destination side are the top four factors with the highest relative importance. Furthermore, it is found that population density on both sides, and travel impedance attributes (transfer times, detour and route distance) have relatively high and robust effects on station-to-station ridership, regardless of the time period. For the four time spans, the relative importance of population density all rank in the top ten, with contributions more than 4.4%. The relative importance of the three travel impedance factors are between 3.2% and 11.4%, which places them in the top twelve. It indicates that population density (no matter on the origin or destination side) and travel impedance factors play important roles in predicting station-to-station ridership.

As for the three main land use types, Fig. 3 illustrates that the relative importance of the business/commercial ratio is much higher than that of residential and industrial/manufacturing ratios. The relative importance of land use mix is highly volatile for the four time periods. Land use mix has a great relative importance on the origin side (13.54%, rank 1) but a very small relative importance on the destination side (1.22%, rank 17) for the morning rush hours. In contrast, the relative contribution of land use mix on the destination side is much higher than that on the origin side for afternoon rush hours (0.86% vs 7.31%) and night (0.58% vs 14.71%). It indicates that as time goes by in a day, the dominant influence of land use mix gradually changes from the origin side to the destination side.

Fig. 3 also reports that the relative importance of road density on both sides in the context of station-to-station ridership is highest for midday. In general, the relative importance of other variables related to design, namely number of intersections, is small and constant for the four time periods as all are less than 4.1%. With respect to the number of bus lines, a bigger relative contribution is found for the origin side for the morning peak hours (11.11% vs 0.60%) and midday (11.50% vs 5.84%), while for the destination side for the afternoon peak hours (2.32% vs 7.82%) and night (0.45% vs 15.29%). Compared to other features of the built environment, the differences between the relative importance of distance to city center for different time periods are relatively small. It is found that distance to city center on the origin side has a bigger relative importance for afternoon rush hours and night (6.33% and 6.81%, respectively), while the relative importance of distance to city center on the origin side is higher for morning peak hours (5.96%). In terms of station type and demographics, it is shown that the relative importance of these factors is relatively small, with relative contributions less than 3%. Terminal stations and housing price on the origin side have a higher relative importance for morning rush hours, while their relative importance on the destination side is bigger for afternoon rush hours. Among the three travel impedance variables, transfer times has the highest relative importance. It indicates that, compared to detour and route distance, people are more sensitive to transfer times between different metro lines. Fig. 3 also shows that detour plays a non-trivial role in predicting station-to-station ridership.

5.2. Non-linear impact of independent variables on station-to-station ridership

5.2.1. Features of the built environment

The estimated coefficients of the conventional multiplicative model can be interpreted as the average marginal effects of the independent variables on the logarithm of station-to-station ridership. The partial dependence plots produced by the GBRT model provide a more refined analysis of the non-linear and possible threshold effects of the independent variables on station-to-station ridership. In other words, partial dependence plots enable us to figure out that within which particular range the effects of the independent variables on the prediction are high and relatively low within the remaining ranges.

Fig. 4 presents the effects of population density on predicting station-to-station ridership for four different time periods, while

(footnote continued)

for the testing data here.

⁴ In order to verify the possibly presence of multi-collinearity, variance inflation factors (VIF) were calculated for the independent variables. The results showed that there is evident multi-collinearity since all VIFs are all smaller than 6. Following the recommendation of an anonymous reviewer, we also ran OLS and 2SLS models but found that multiplicative model outperform these two models.

Table 2
Estimation results of the multiplicative model.

Independent variables	7:00–9:00		11:00–13:00		17:00–19:00		21:00–23:00	
	Coefficient	p-Value	Coefficient	p-Value	Coefficient	p-Value	Coefficient	p-Value
<i>Origin</i>								
Population density	0.467	0.000	0.290	0.000	0.331	0.000	0.564	0.000
Residential ratio	0.045	0.006	−0.054	0.000	−0.183	0.000	−0.127	0.000
Business/commercial ratio	−0.198	0.000	−0.062	0.000	0.109	0.000	0.154	0.000
Industrial/manufacturing ratio	−0.046	0.000	−0.101	0.000	−0.086	0.000	−0.112	0.000
Land use mix	0.128	0.000	−0.030	0.190	0.048	0.051	−0.243	0.000
Road density	0.134	0.021	0.225	0.000	0.020	0.717	0.716	0.000
Number of intersections	0.037	0.400	0.991	0.014	0.136	0.001	−0.238	0.000
Number of bus lines	0.255	0.000	0.235	0.000	0.137	0.000	0.115	0.000
Distance to city center	0.214	0.000	−0.088	0.000	−0.155	0.000	−0.001	0.959
Terminal dummy	0.483	0.000	0.581	0.000	0.483	0.000	0.226	0.000
Transfer dummy	0.054	0.206	0.511	0.000	0.393	0.000	0.805	0.000
Housing price	−0.035	0.390	−0.131	0.000	0.375	0.000	0.023	0.669
<i>Destination</i>								
Population density	0.236	0.000	0.338	0.000	0.580	0.000	0.349	0.000
Residential ratio	−0.182	0.000	−0.033	0.024	0.047	0.002	−0.001	0.593
Business/commercial ratio	0.130	0.000	−0.024	0.039	−0.130	0.000	−0.137	0.000
Industrial/manufacturing ratio	−0.082	0.000	−0.110	0.000	−0.091	0.000	−0.018	0.100
Land use mix	−0.028	0.272	−0.084	0.000	0.016	0.512	−0.044	0.179
Road density	−0.008	0.893	0.316	0.000	0.198	0.000	0.178	0.010
Number of intersections	0.177	0.000	−0.034	0.402	0.153	0.008	0.028	0.582
Number of bus lines	0.128	0.000	0.229	0.000	0.212	0.000	0.314	0.000
Distance to city center	−0.291	0.000	−0.120	0.000	0.213	0.000	0.241	0.000
Terminal dummy	0.561	0.000	0.474	0.000	0.379	0.000	0.253	0.000
Transfer dummy	0.328	0.000	0.636	0.000	0.447	0.000	0.158	0.000
Housing price	0.316	0.000	−0.195	0.000	−0.247	0.000	0.184	0.000
<i>Travel impedance</i>								
Transfer times	−1.064	0.000	−0.876	0.000	−0.985	0.000	−0.944	0.000
Detour	−1.360	0.000	−0.856	0.000	−1.173	0.000	−0.519	0.000
Route distance	−0.116	0.000	−0.226	0.000	−0.267	0.000	−0.239	0.000
Constant	0.518	0.105	1.368	0.000	0.511	0.093	−2.860	0.000
Number of samples	10,041		9248		9870		6485	
R^2	0.624		0.640		0.659		0.546	
Adjusted R^2	0.623		0.639		0.658		0.544

Fig. 3. Comparison of relative importance between different time periods.

Fig. 4. The effects of population density on station-station ridership.

controlling for all other independent variables in the models. All the plots show a non-linear association between population density and station-to-station ridership, regardless of on which side and which time periods. Furthermore, it is evident that population density on both sides have positive effects on station-to-station ridership, which is consistent with the results produced by the multiplicative model (Table 2). According to the range of values for the vertical axes, population density on the origin side produces higher OD flows during the PM rush hours, while population density on the destination side generates higher OD flows during the AM rush hours. Taking population density on the origin side as an example for the peak period (Fig. 4c), population density on the origin side has a trivial effect when it is less than about 26 thousand persons per km^2 , while station-to-station ridership increases substantially when population density on the origin side exceeds 26 thousand persons per km^2 . For the off-peak period, taking Fig. 4f as an example, when population density on the destination side grows from 6 to 9 thousand persons per km^2 , station-to-station ridership increases at about one person; then population density on the destination side has weak impact when it grows from 9 to 19 thousand persons per km^2 ; station-to-station ridership increases linearly when population density on the destination side increases from 19 to 38 thousand persons per km^2 . A negative association between population density on the destination side and station-to-station ridership exists when population density exceeds 29 thousand persons per km^2 , for unknown reasons (Fig. 4g).

The partial dependence plots for residential ratio are presented in Fig. 5. It shows that residential ratio on the origin side is positively associated with station-to-station ridership during morning rush hours but negatively correlated with station-to-station ridership during afternoon rush hours and off-peak periods, which is consistent with the results of the multiplicative models. Taking Fig. 5a as an example, residential ratio on the origin side has a linear effect (albeit fluctuating sometimes) on station-to-station ridership when it increases from 0 to around 0.33; then the effect becomes negative when the residential ratio is between 0.33 and 0.48; station-to-station ridership is positively associated with residential ratio when this index exceeds 0.48. As for residential ratio on the destination side, residential ratio is negatively associated with station-to-station ridership during AM rush hours and midday. However, the relationship between residential ratio and station-to-station ridership during PM rush hours and night does not show a clear trend (Fig. 5g and h). The impact of residential ratio on the destination side reaches the highest points at about 0.35 and then

Fig. 5. The effects of residential ratio on station-station ridership.

Fig. 6. The effects of business/commercial ratio on station-station ridership.

decreases for both PM rush hours and night. The range of values for the vertical axis in Fig. 5h is relatively small.

The partial dependence plots for business/commercial ratio are presented in Fig. 6. It is noted that business/commercial ratio shows a negative correlation with station-to-station ridership when it exceeds 0.3. This negative association appears counterintuitive. A possible reason could be that the majority of business/commercial ratios within station catchment areas are smaller than 0.3 and only four samples are larger than 0.3 in the study area (see Table 2, the mean value is 9.18% and the standard deviation is 9.40%), and therefore the outliers may produce this result. Nevertheless, Fig. 6a and c show that the relationship between business/commercial ratio on the origin side and station-to-station ridership is negative during AM peak hours while positive during PM peak hours when the business/commercial increases from 0 to 0.3. Fig. 6e and g report a reverse relationship between business/commercial on the destination side and station-to-station ridership during peak hours. Moreover, Fig. 6 demonstrates the non-linear effects of business/commercial ratio (on both origin and destination sides) on station-to-station ridership. For instance, Fig. 6e shows that the business/commercial ratio on the destination side has weak effect on station-to-station ridership when it is smaller than 0.16, and station-to-station ridership increases rapidly when it grows from 0.16 to 0.3. Fig. 6c presents a similar regularity. It is worth noting that both relative importance and partial dependence plots (compare the ranges of values for the vertical axes in Figs. 5–7) suggest that the business/commercial ratio has a larger effect on station-to-station ridership than the residential and industrial/manufacturing ratios.

Fig. 7 indicates that the industrial/manufacturing ratio on both the origin and destination sides have negative effects on station-to-station ridership, regardless of the time periods. In general, all plots in Fig. 7 show that station-to-station ridership reach its low points when the industrial/manufacturing ratio is between 0.30 and 0.37. Furthermore, most plots in Fig. 7 show that: first, industrial/manufacturing ratio has trivial effect on station-to-station ridership when it increases from 0 to 0.25; then, station-to-station ridership decreases linearly and sharply as the industrial/manufacturing ratio moves to the range 0.25–0.30; and then the industrial/manufacturing ratio has the smallest effect on station-to-station ridership; finally, the curves become horizontal. Therefore, the industrial/manufacturing ratio has threshold effects on station-to-station ridership. Only when the ratio exceeds the value of 0.25, the

Fig. 7. The effects of industrial/manufacturing ratio on station-station ridership.

Fig. 8. The effects of land use mix on station-station ridership.

negative relationship with station-to-station ridership becomes clear, which is not reflected in the results of the conventional multiplicative DRM model (Table 2).

The eight plots in Fig. 8 illustrate the relationship between land use mix on the different sides and station-to-station ridership for different time periods. The curves are nearly horizontal in most plots, indicating that station-to-station transit ridership is not related to land use mix on either side. This is in line with Choi et al. (2012), who also found that land use mix has no significant positive effect on metro ridership in Seoul City.

The effects of road density on both the origin and destination sides on station-to-station ridership for different time periods are shown in Fig. 9. In general, these plots indicate that road density on both sides is positively correlated with station-to-station ridership, regardless of time of day. This pattern is in line with Tu et al. (2018). Furthermore, compared to the estimation results in Table 2, the partial dependence plots illustrate more details about this positive relationship. Taking Fig. 9g as an example, road density on the destination side has a trivial impact on station-to-station ridership when it is less than 9 km/km². Then, ridership increases when road density grows from 9 to 10 km/km²; and the effect becomes trivial when road density moves from 10 to 15.5 km/km². Finally, station-to-station ridership increases substantially when road density on the destination side exceeds 15.5 km/km².

Fig. 10 summarizes the associations between station-to-station ridership and number of intersections on the origin and destination side for the four chosen time periods. Fig. 10b, f and g have a similar non-linear pattern: first steady then decreasing and then increasing (a U-shaped manner), while Fig. 10c and e have another similar non-linear pattern: first steady then increasing and then decreasing (an inverted U-shaped manner). The non-linear relationship between station-to-station ridership and the number of intersections for the night time slot differs from the other three time periods. According to the ranges of vertical axes, it is found that number of intersections on the origin side has a larger effects on station-to-station ridership during PM peak hours than during other time periods, while number of intersections on the destination side has a larger impact on station-to-station ridership during peak hours than during off-peak hours. Taking Fig. 10e as an example, the average station-to-station ridership is about 6 when number of

Fig. 9. The effects of road density on station-station ridership.

Fig. 10. The effects of number of intersections on station-station ridership.

intersections on the origin side is less than 17, while it increases to 15 when the number of intersections on the origin side is within 20 to 38. The plots in Fig. 10 show that number of intersections has a threshold effect on station-to-station ridership.

Fig. 11 displays the relationship between station-to-station ridership and number of bus lines on the origin and destination side for different time periods. Consistent with the literature (e.g., Choi et al., 2012; Zhao et al., 2014; Durning and Townsend, 2015), the results of the multiplicative model suggests that number of bus lines has a significant and positive effect on metro ridership at station-to-station level, regardless of the time periods (Table 2). However, the partial dependence plots produce by the GBRT models show that the significant and positive effects may be amplified and highlighted by a few outliers (e.g., number of bus lines > 40). Thus, the real relationship between number of bus lines and metro ridership would be hidden or overstated. For example, the horizontal curves with slight fluctuation, as number of bus lines less than 40, indicate that number of bus lines have a trivial effect on station-to-station ridership during midday period (Fig. 11c and g). Nevertheless, the remaining plots show that, excluding the outliers (number of bus lines > 40), the number of bus lines has a positive influence on station-to-station ridership for rush hours and night.

The partial dependence plots for distance to city center are shown in Fig. 12. Fig. 12a, g and h indicate that distance to city center on the origin side is positively associated with station-to-station ridership during AM rush hours, while distance to city center on the destination side has a positive effect on metro ridership at the station-to-station level during PM rush hours and night. On the other hand, Fig. 12b, c, e and f demonstrate that distance to city center on the origin side is negatively related to station-to-station ridership during midday and PM peak hours, while distance to city center on the destination side has negative effects on metro ridership at station-to-station level during AM peak hours and midday. This finding reflects the main direction of urban mobility in each time period such as inbound commuting trips moving into the city center during AM rush hours and outbound commuting trips moving away from the city center during PM rush hours. Distance to city center on the origin side has the largest effect on station-to-station ridership in the afternoon, while distance to city center on the destination side has the largest effect on station-to-station ridership in the morning. Taking Fig. 12e as an example, within the range of 0–5.5 km, station-to-station ridership decreases sharply (at about 40); and the effect of further increase in distance to city center on the destination side becomes inappreciable when it exceeds 5.5 km.

Fig. 11. The effects of number of bus lines on station-station ridership.

Fig. 12. The effects of distance to city center on station-station ridership.

5.2.2. Other independent variables

The partial dependence plots for station type are two straight lines since station type variables are dummy variables. The partial dependence plots for demographics and the three travel impedance variables are shown in Fig. 13. The plots indicate that the effects of housing price on both the origin and destination side on station-to-station fluctuates. Taking Fig. 13e as an example, it is shown that housing price on the destination side has a trivial influence on station-to-station ridership when it is less than 14×10^3 yuan/m²; station-to-station ridership then decreases slightly when housing price is between 14 and 18×10^3 yuan/m²; and then remains constant when housing price is within the range of 18– 29×10^3 yuan/m². However, increasing housing price on the destination side between 29 and 44×10^3 yuan/m² enhances metro ridership at the station-to-station level; and further when housing price on the destination side exceeds 44×10^3 yuan/m², it negatively impacts station-to-station ridership.

The effects of the three travel impedance variables are summarized in Fig. 14. With respect to transfer times, the multiplicative model indicates a significant and negative relationship with station-to-station ridership for different times of day. However, the partial dependence plots show the effects are all in a non-linear rather than linear way, regardless of the time period. The plots (Fig. 14a-d) show that station-to-station ridership decreases substantially when transfer times increases from 0 to 1, while the effect of further increasing in transfer times becomes small when it exceeds 1. This implies that the marginal effect progressively decreases as transfer times grows and individuals are more sensitive to transfer or not rather than transfer once or twice. Fig. 14e-h demonstrate that detour is negatively related to station-to-station ridership. In general, detour has a trivial effect on metro ridership at station-to-station level when it exceeds 2 (the exact cut-down point for night is around 2.5), regardless of time of day. It indicates that detour has a threshold effect on station-to-station ridership. Actually, the negative effects of detour on metro station-to-station ridership within the range of 1 to 2 is not always the same. The plots show that, station-to-station ridership decreases sharply when detour grows from 1 to about 1.2 (the average value of detour in Nanjing metro system is 1.42, see Table 1), while the negative effects become much mild as detour is between 1.2 and 2. The partial dependence plots for route distance illustrate that route distance is not always negatively associated with station-to-station ridership: there is a positive relationship when route distance is less than

Fig. 13. The effects of housing price on station-station ridership.

Fig. 14. The effects of transfer times, detour, and route distance on station-station ridership.

approximately 6 km. This finding is somewhat counterintuitive and inconsistent with the results of the multiplicative model in the present paper and previous studies (Choi et al., 2012; Zhao et al., 2014). We have to proclaim that the estimation results produced by traditional DRM model can only reflect the major negative influence trend of route distance (or travel time) on station-to-station ridership, and therefore may hidden the details. It should be noted that metro is appropriate for middle- and long-distance trips, and passengers whose trip distances are too long or too short will seldom choose metro to travel. All the partial dependence plots for route distance depict that this factor affects station-to-station ridership in a non-linear manner and the influence trends are quite similar for different time periods. When route distance is less than 6 km, an increasing in route distance will raise station-to-station ridership; then there is a reverse relationship between them when route distance is within the range of 6–30 km (especially within the range of 6–20 km). However, beyond the range, route distance has a trivial effect on station-to-station ridership.

6. Discussion and conclusions

By applying the GBRT model, the present paper examined the effects of features of the built environment on metro station-to-station ridership by different times of day (AM rush hours 7:00–9:00, midday 11:00–13:00, PM rush hours 17:00–19:00, and night 21:00–23:00), using smart card data from the Nanjing metro system, China. Four categories of independent variables, including built environment characteristics on the origin and destination side, station type on the origin and destination side, housings price on the origin and destination side, and travel impedance factors were used for modelling station-to-station ridership. The choice of this data mining method was motivated by the desire to analyse the relationships between features of the built environment and ridership in a more refined manner, based on the contention that this relationship discontinuous, non-monotonic, unlike continuous as common direct ridership models commonly assume.

This study found several interesting and meaningful results. First, the GBRT model is superior to the multiplicative model in predicting metro station-to-station ridership for all time periods considered. Second, the results of the GBRT model indicate that features of the built environment on the origin side have larger effects on station-to-station ridership than on the destination side during the morning rush hours, while the opposite is found during the afternoon rush hours and night. Number of bus lines on the origin side, population density on the destination side, and transfer times are the three most important explanatory variables for station-to-station ridership in the morning rush hours. Population density on the origin side, transfer times, route distance, and number of bus lines on the destination side are the top four factors with highest relative importance for the afternoon rush hours. Third, most independent variables indeed are associated with station-to-station ridership in a discontinuous, non-linear way, regardless of the time periods. Fourth, among the three travel impedance variables, transfer times shows the strongest relationship with station-to-station ridership.

Thus, the experience with the application of the GBRT data mining method in this study confirms the potential relevance of this algorithm to predict station-to-station ridership and describe in more detail the form of the relationship. It has the advantage that no a priori assumptions have to be made about the functional form of the relationship, but that the best fitting discontinuous form is

derived from the data. The result is a more refined expression and improved prediction of ridership. However, the results of this study also demonstrate a potential disadvantage of these data-driven data mining methods. The data-driven nature of the algorithm may lead to counterintuitive results for some ranges of explanatory variables. This caused by sparse data, outliers and/or the specific, unbalanced spatial distribution of the independent variables. It implies that results of the GBRT cannot be accepted uncritically. However, the results can be used to explore the relationship in a more refined way. Counterintuitive findings should be detected and the underlying data in the relevant range should be inspected. The model can be improved either by trying different settings or by estimating the effect for splines defined on the basis of the outcomes of the trees.

The findings of this study help policymakers and urban planners to better understand how the factors such as features of the built environment and travel impedance attributes influence station-to-station ridership. For travel demand modelers, the estimation results, including relative importance and partial dependence plots, provide a better sketch planning tool for predicting metro station-to-station ridership than that of conventional DRM models, thereby assisting metro station planning. The results demonstrate that population density plays a key role in metro ridership at station-to-station level. Therefore, high-density housing, business and commercial land development should be predominant in station catchment areas when developing new areas and satellite towns. The relative importance of population density and business/commercial ratio on both origin and destination sides are large for nearly all the times of day, indicating that they are strongly associated with station-to-station ridership. This suggests urban planners should pay more attention to the planning of business/commercial land use which is the important trip productions and attractors. Although the extent of land use mix does not account for an increase in station-to-station ridership in a specific time period, the partial dependence plots of residential ratio, business/commercial ratio, and industrial/manufacturing ratio indicate that excessive development of specific land type around metro station on origin or destination side will decrease metro ridership. Increasing road density will improve station-to-station ridership. However, it should be realized that road supply (mainly high-level municipal roads) also stimulate car ownership and usage (Li et al., 2010). Considering the sparse road densities around most suburban metro stations and the suburbanization of population in Chinese cities, reasonable ratios of different road hierarchies should be carefully kept in mind. More secondary, residential road and relatively small communities (blocks) are needed in the suburban metro station areas. The non-linear relationship between number of intersections and station-to-station ridership demonstrates that increasing intersection density to improve connectivity could encourage rail transit travel and promote ridership. While this positive effect only works within a certain range of the number of intersections. Furthermore, the effect becomes negative when intersection density around metro station exceeds the threshold value. This finding is also helpful for urban planning.

Since the number of bus lines has a high relative importance and is positively related to station-to-station ridership for most time periods, the connectivity between bus and metro should be paid enough attention by transportation planners and operators. The finding that both transfer times and detours have significant and negative effects on station-to-station ridership indicates that low connectivity efficiency, and improper metro network layout would reduce the utility of metro travel. This suggests that local government can decrease inconvenient metro transfers by directly connecting two important areas (e.g., city center, subsidiary center, and railway station) via one metro line. Furthermore, the relationship between route distance and station-to-station ridership implies that in the middle trip length range people are more likely to use metro, and this may shed a light on the layout of metro network and urban functional areas. Finally, different features of the built environment affecting station-to-station ridership at different time of day (AM rush hours, midday, PM rush hours, and night) and by different magnitudes indicates the different influencing mechanisms of station-to-station ridership between different time periods. This can provide a reference for travel demand optimization in urban rail transit management and built environment balance in urban planning, based on the refined ridership analyses rather than simply regarding the determinant of the total daily ridership.

Although this study contributed to some under-researched topics related to predicting metro ridership and led to some interesting results, some aspects need further investigation. First, the number of studies on direct station-to-station ridership model is still very limited and our study does not change that observation. Hence, replication studies should be welcome to accumulate knowledge about the performance of GBRT. To assess the generalizability of the approach and the specific results, it would be of interest to apply the model to smart card data of other cities that vary in size and complexity and connectivity of the metro network, and also in terms of the spatial distribution of the built environment variables. Second, in this study we compared the GBRT model only with the multiplicative model. It would be relevant for practitioners, however, to extend this comparison of modeling approaches to other data mining methods, other more advanced conventional models and to the more comprehensive activity-based models. DRMs do not consider competition between transportation modes and mostly reflect the correlations between explanatory variables and dependent variable rather than the direct causality. Thus, it is relevant to judge whether this limitation critically affects the forecasts and policy recommendations of the DRMs. Third, we only explored the effects of built environment, station type, and travel impedance attributes on station-to-station ridership in this paper, while travel impedance of competing modes (i.e., bus and car), travel behavior changes during the day, and the interactions between independent variables especially between the features of the built environment may be also associated with station-to-station ridership. This issue should be further investigated. Fourth, it is interesting to determine which “D” is influencing more metro station-to-station ridership. Therefore, more relevant information on the features of built environment need to be collected in comprehensive manner. Fifth, we examined the direct effects of the features of built environment and demographics, while all these variables might be influenced by the metro system investment. Although the VIF test indicates the endogeneity issues are not strong enough in our study, we could not deny the potential endogeneity issue and call for similar studies based on other models such as path analysis model in the future. Moreover, this study focused on the station-to-station ridership of metro systems. It will be meaningful to extend the study to other transportation modes such as bus, tram. The station-to-station ridership of integrated multimodal transportation systems should be considered in future work.

CRediT authorship contribution statement

Zuoxian Gan: Software, Data curation, Methodology, Writing - original draft. **Min Yang:** Project administration, Funding acquisition, Supervision, Writing - review & editing. **Tao Feng:** Methodology, Writing - review & editing. **Harry J.P. Timmermans:** Supervision, Writing - review & editing.

Acknowledgments

This study is supported by the National Key Research and Development Program of China (No. 2018YFB1601300) and National Natural Science Foundation of China (No. 71771049). The authors would like to thank the editor and the anonymous reviewers whose helpful comments and valuable suggestions on the earlier version led to significant improvement of this manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.trd.2020.102332>.

References

- Ao, Y., Yang, D., Chen, C., Wang, Y., 2019. Effects of rural built environment on travel-related CO₂ emissions considering travel attitudes. *Transportation Res. Part D: Transport Environ.* 73, 187–204.
- Arentze, T.A., Timmermans, H.J.P., 2004. A learning-based transportation oriented simulation system. *Transportation Res. Part B: Methodological* 38, 613–633.
- Arentze, T.A., Timmermans, H.J.P., 2007. Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. *Transportation Res. Part B: Methodological* 41, 772–783.
- Baum-Snow, N., Kahn, M.E., 2000. The effects of new public projects to expand urban rail transit. *J. Public Econ.* 77 (2), 241–263.
- Cardozo, O.D., García-Palomares, J.C., Gutiérrez, J., 2012. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Appl. Geogr.* 34, 548–558.
- Cervero, R., Ferrell, C., Murphy, S., 2002. Transit oriented development and joint development in the United States: a literature review. *Transit Cooperative Research Program: Research Results Digest*. Transportation Research Board, Washington, DC.
- Cervero, R., Murphy, S., Ferrell, C., Goguts, N., Tsai, Y.H., Arrington, G.B., Boroski, J., Smith-Heimer, J., Golem, R., Peninger, P., Nakajima, E., Chui, E., Dunphy, R., Myers, M., McKay, S., 2004. *Transit-Oriented Development in the United States: Experiences, Challenges, and Prospects*. Transportation Research Board, Washington, DC.
- Choi, J., Lee, Y.J., Kim, T., Sohn, K., 2012. An analysis of metro ridership at the station-to-station level in Seoul. *Transportation* 39 (3), 705–722.
- Chu, X., 2004. *Ridership Models at the Stop Level*. University of South Florida, National Center of Transit Research.
- Ding, C., Cao, X.J., Naess, P., 2018. Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. *Transportation Res. Part A: Policy Pract.* 110, 107–117.
- Ding, C., Cao, X., Liu, C., 2019. How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify non-linear thresholds. *J. Transp. Geogr.* 77, 70–78.
- Ding, C., Wu, X., Yu, G., Wang, Y., 2016. A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data. *Transportation Res. Part C: Emerging Technol.* 72, 225–238.
- Duncan, M., 2010. To park or to develop: trade-off in rail transit passenger demand. *J. Planning Educ. Res.* 30 (2), 162–181.
- Durning, M., Townsend, C., 2015. Direct ridership model of rail rapid transit systems in Canada. *Transportation Res. Rec.: J. Transportation Res. Board* 2537 (1), 96–102.
- Estupiñán, N., Rodríguez, D.A., 2008. The relationship between urban form and station boardings for Bogotá's BRT. *Transportation Res. Part A: Policy Pract.* 42 (2), 296–306.
- Ewing, R., Cervero, R., 2010. Travel and the built environment: A meta-analysis. *J. Am. Planning Assoc.* 76 (3), 265–294.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 1189–1232.
- Gan, Z., Feng, T., Wu, Y., Yang, M., Timmermans, H.J.P., 2019a. Station-based average travel distance and its relationship with urban form and land use: An analysis of smart card data in Nanjing City, China. *Transport Policy* 79, 137–154.
- Gan, Z., Feng, T., Yang, M., Timmermans, H.J.P., Luo, J., 2019b. Analysis of metro station ridership considering spatial heterogeneity. *Chin. Geographical Sci.* 20 (6), 1065–1077.
- Gutiérrez, J., Cardozo, O.D., García-Palomares, J.C., 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *J. Transp. Geogr.* 19 (6), 1081–1092.
- Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. *Math. Intelligencer* 27 (2), 83–85.
- IEA (International Energy Agency), 2016. *CO2 Emissions from Fuel Combustion Highlights 2016*. Paris.
- Iseki, H., Liu, C., Knaap, G., 2018. The determinants of travel demand between rail stations: A direct transit demand model using multilevel analysis for the Washington DC Metrorail system. *Transportation Res. Part A: Policy Pract.* 116, 635–649.
- Janssens, D., Wets, G., Brijs, T., Vanhoof, K., Arentze, T., Timmermans, H., 2004. Improving performance of multiagent rule-based model for activity pattern decisions with Bayesian networks. *Transp. Res. Rec.* 1894 (1), 75–83.
- Jun, M.J., Choi, K., Jeong, J.E., Kwon, K.H., Kim, H.J., 2015. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul. *J. Transp. Geogr.* 48, 30–40.
- Kepaptsoglou, K., Stathopoulos, A., Karlaftis, M.G., 2017. Ridership estimation of a new LRT system: direct demand model approach. *J. Transp. Geogr.* 58, 146–156.
- Kim, S., Rasouli, S., Timmermans, H.J.P., Yang, D., 2018. Estimating panel effects in probabilistic representations of dynamic decision trees using Bayesian generalized linear mixture models. *Transportation Res. Part B: Methodological* 111, 168–184.
- Kuby, M., Barranda, A., Upchurch, C., 2004. Factors influencing light-rail station boardings in the United States. *Transportation Res. Part A: Policy Pract.* 38, 223–247.
- Li, J., Walker, J.L., Srinivasan, S., Anderson, W.P., 2010. Modeling private car ownership in China: investigation of urban form impact across megacities. *Transp. Res. Rec.* 2193 (1), 76–84.
- Liu, C., Erdogan, S., Ma, T., Ducca, F.W., 2016. How to increase rail ridership in Maryland: Direct ridership models for policy guidance. *J. Urban Plann. Dev.* 142 (4), 04016017.
- Liu, S., Yao, E., Li, B., 2019. Exploring urban rail transit station-level ridership growth with network expansion. *Transportation Res. Part D: Transport Environ.* 73, 391–402.
- Liu, Y., Yang, D., Timmermans, H.J.P., de Vries, B., 2020. Analysis of the impact of street-scale built environment design near metro stations on pedestrian and cyclist road segment choice: A stated choice experiment. *J. Transp. Geogr.* 82 102570.
- Loo, B.P., Li, D.Y., 2006. Developing metro systems in the People's Republic of China: policy and gaps. *Transportation* 33 (2), 115–132.
- McNally, M., 2000. The four-step model. In: Hensher, D.A., Button, K.J. (Eds.), *Handbook of Transport Modelling*. Pergamon, Oxford, pp. 33–52.

- MOHURD, 2015. Guidelines for Planning and Design of Urban Rail. Ministry of Housing and Urban-Rural Development of People's Republic of China, Beijing.
- Rasouli, S., Timmermans, H.J.P., 2012. A multivariate adaptive regression spline emulator of travel time predicted by a complex activity-based model of travel demand. In: Mak, H.-Y., Lo, H.K. (Eds.), *Proceedings 17th International Conference Hong Kong Society of Transportation*, Hong Kong, China, pp. 404–410.
- Rasouli, S., Timmermans, H., 2014a. Activity-based models of travel demand: promises, progress and prospects. *Int. J. Urban Sci.* 18 (1), 31–60.
- Rasouli, S., Timmermans, H.J.P., 2014b. Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *Eur. J. Transport Infrastructure Res.* 14, 412–424.
- Ryan, S., Frank, L.F., 2009. Pedestrian environments and transit ridership. *J. Public Transportation* 12 (1), 39–57.
- Sohn, K., Shim, H., 2010. Factors generating boardings at metro stations in the Seoul metropolitan area. *Cities* 27, 358–368.
- Strambi, O., Van De Bilt, K.A., 1998. Trip generation modeling using CHAID, a criterion-based segmentation modeling tool. *Transp. Res. Rec.* 1645 (1), 24–31.
- Sung, H., Oh, J.T., 2011. Transit-oriented development in a high-density city: Identifying its association with transit ridership in Seoul, Korea. *Cities* 28 (1), 70–82.
- Tu, W., Cao, R., Yue, Y., Zhou, B., Li, Q., Li, Q., 2018. Spatial variations in urban public ridership derived from GPS trajectories and smart card data. *J. Transp. Geogr.* 69, 45–57.
- Wu, X., Tao, T., Cao, J., Fan, Y., Ramaswami, A., 2019. Examining threshold effects of built environment elements on travel-related carbon-dioxide emissions. *Transportation Res. Part D: Transport Environ.* 75, 1–12.
- Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. *Transportation Res. Part C: Emerging Technol.* 58, 308–324.
- Zhao, J., Deng, W., Song, Y., Zhu, Y., 2014. Analysis of metro ridership at station level and station-to-station level in Nanjing: an approach based on direct demand models. *Transportation* 41 (1), 133–155.
- Zong, W., Zhang, J., 2019. Use of smartphone applications and its impacts on urban life: A survey and random forest analysis in Japan. *Sustain. Cities Soc.* 49 101589.