Journal of Transport Geography 82 (2020) 102560

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Exploring the non-linear associations between spatial attributes and walking distance to transit

Tao Tao*, Jueyu Wang, Xinyu Cao

Humphrey School of Public Affairs, University of Minnesota, Twin Cities., 301 19th Ave S, Minneapolis, MN 55455, United States

ARTICLE INFO

Keywords:
Machine learning
Walking behavior
Station area planning
Built environment
Land use

ABSTRACT

When examining environmental correlates of walking distance to transit stops, few studies report the importance of spatial attributes relative to other factors. Furthermore, previous studies often assume that they have linear relationships with walking distance. Using the 2016 Transit On Board Survey in the Minneapolis and St. Paul Metropolitan Area, this study adopted the gradient boosting decision trees method to examine the relationships between walking distance and spatial attributes. Results showed that spatial attributes collectively have larger predictive power than other factors. Moreover, they tend to have non-linear associations with walking distance. We further identified the most effective ranges of spatial attributes to guide stop area planning and stop location choice in the region.

1. Introduction

Walking access to transit plays a critical role in determining transit use and guiding transit planning. Transit users often walk to and from transit stops (Hsiao et al., 1997). They are less likely to use transit if stops are not within a walkable distance (Loutzenheiser, 1997; Zhao et al., 2003). Moreover, scholars and practitioners often use walking distance to transit to determine transit catchment/service areas, which are fundamental to ridership prediction and the assessment of economic development associated with transit (El-Geneidy et al., 2014). This study provides a nuanced understanding of transit users' walking distance and helps guide spatial planning around stops, as well as stop location choice.

Many studies examine environmental correlates of walking distance to transit and conclude that many spatial characteristics (such as density and diversity) around transit stops are associated with walking distance (Loutzenheiser, 1997; Maghelal, 2011). These findings offer critical implications for stop area planning. However, as discussed in the next section, the literature leaves a couple of critical questions unanswered. To begin with, a limited number of studies assess how large a role spatial attributes play in affecting travel behavior (Van Wee and Handy, 2016), and few emphasize walking distance to transit. The assessment is central to the efficacy of using land use policies to shape travel behavior through spatial planning (Stevens, 2017a). Moreover, previous studies often assume that spatial attributes have linear relationships with walking distance. However, changing the built environment may be infertile in altering travel behavior once certain thresholds of spatial attributes are reached (Ding et al., 2018). Accordingly, planners are eager to know the effective ranges of spatial attributes. For example, how dense is enough to influence transit riders' walking distance?

To fill these two gaps, this study employs gradient boosting decision trees (GBDT) on the 2016 Transit On Board Survey data in the Minneapolis-St. Paul (Twin Cities) Metropolitan Area. It answers two sets of questions: 1) How important is the collective contribution of spatial attributes to predicting walking distance, relative to the social environment, individual characteristics, and trip features? 2) Are the associations between spatial attributes and walking distance linear? Are there any thresholds that these attributes affect walking distance most effectively?

Answers to these research questions could offer important planning implications from the following two aspects. First, assessing the collective contribution of spatial attributes can offer a better understanding of the extent to which spatial attributes affect walking distance to transit, and quantifying the relative importance of individual spatial attributes can guide planners how to prioritize them when planning resources are inadequate. Second, identifying threshold effects of spatial attributes can inform planners their most effective ranges for stop area planning and stop location choice. Another important feature of this study is that we built our research on the Smart Location Database of the Environmental Protection Agency, a nationally-available dataset in the US context, and shared the programming codes (Tao, 2018).

* Corresponding author.
E-mail addresses: taotao@umn.edu (T. Tao), wang5931@umn.edu (J. Wang), cao@umn.edu (X. Cao).

https://doi.org/10.1016/j.jtrangeo.2019.102560
Received 16 October 2018; Received in revised form 3 October 2019; Accepted 3 October 2019
Available online 21 October 2019
0966-6923/

Planners in other regions could readily apply our models to their local data, and inform their spatial and transit planning.

The paper is organized as follows. We review the literature of walking distance to transit and identify the gaps in Section 2. We introduce the data and the GBDT approach in Section 3. Section 4 presents the results. In the final section, we summarize key findings and discuss associated implications.

2. Literature review

Many studies explore the correlates of walking distance to transit stops and shed light on how to influence transit users' willingness to walk. This is important to transit ridership as walking is a primary means to reach transit. Conventionally, planners assume that riders would like to walk 400 m to reach a bus stop and 800 m to reach a rail station (Gutiérrez and García-Palomares, 2008; Hsiao et al., 1997; Zhao et al., 2003). These distances define the size of the catchment area of transit stops. However, a growing number of studies question these rules of thumb and scrutinize walking distance (or time) to transit stops and associated factors.

The literature suggests that walking distance is associated with the spatial attributes surrounding transit stops, demographic characteristic of transit users, and the attributes of transit service. First, spatial attributes are correlates of walking distance. Previous studies show that walking distance is affected by population density (El-Geneidy et al., 2014; Jiang et al., 2012), job density (Wang and Cao, 2017), intersection density (El-Geneidy et al., 2014; Wang and Cao, 2017), and sidewalk density or availability (Maghelal, 2011; Tilahun and Li, 2015). Moreover, transit users are willing to walk longer in a pedestrian-friendly area (Jiang et al., 2012; O'Sullivan and Morrall, 1996). Second, individual characteristics affect walking distance. For example, young people and men tend to walk a longer distance to reach transit stops than seniors and women, respectively (Alshalalfah and Shalaby, 2007; El-Geneidy et al., 2014). Auto ownership, household income, and household size also have positive effects on walking distance because households with more vehicles, higher incomes, and more members are more likely to be choice riders and less likely to live close to transit (Alshalalfah and Shalaby, 2007; El-Geneidy et al., 2014). Third, transit service features are significant predictors of walking distance. Transit users tend to walk a longer distance to transit stops that have more frequent services and shorter waiting time (Alshalalfah and Shalaby, 2007; O'Sullivan and Morrall, 1996). They are also willing to walk longer if their total trip length is longer, but they will walk less if they need to make transfers (El-Geneidy et al., 2014).

Although many studies substantiate that spatial attributes have statistically significant relationships with walking distance, limited attention is paid to the practical importance of these findings. In fact, effect size matters (Ziliak and McCloskey, 2004). In the realm of planning, several scholars engage in a heated debate on the efficacy of using land use policies to influence travel behavior in the *Journal of the American Planning Association* (Ewing and Cervero, 2017; Nelson, 2017; Stevens, 2017b). Among others, Wang and Cao, (2017) study walking distance of transit egress trips in the Twin Cities and estimate the elasticity of each spatial variable, the effect size. However, they did not discuss the collective contribution of these variables. Land use changes are often multi-dimensional; the collective effects of these changes could be substantial even if the impact of a single spatial variable is moderate (Ewing and Cervero, 2017). Planners are interested in knowing how much effect on travel behavior they could bring collectively if they implement a set of land use instruments (Van Wee and Handy, 2016). Furthermore, because individual characteristics and transit service features confound the relationship between spatial attributes and walking distance to transit stops, how important is the collective effect of spatial attributes relative to these other factors? Assessing the relative contribution of different factors is also an important topic in the land use-travel behavior literature (Cao, 2019; Mokhtarian and Van Herick, 2016; Singh et al., 2018).

Most studies assume that the associations between spatial attributes and walking distance are linear (Jiang et al., 2012; Townsend and Zacharias, 2010; Zhao and Deng, 2013). Although some apply generalized linear regression (El-Geneidy et al., 2014; Wang and Cao, 2017), the pre-defined relationship may mask the nuanced connections between different spatial attributes and walking distance to transit. A growing number of scholars have question about the assumption that spatial attributes have linear or pre-defined relationships with travel behavior (Van Wee and Handy, 2016). Some studies reject the assumption. For instance, Ding et al. (2018) find that the influence of population density on driving distance is saturated at 30 persons per hectare in Oslo and the effect of distance to the city center varies at different intervals of the distance and is also saturated after a threshold is reached. Wang and Cao, (2017) conclude that the effects of spatial attributes on walking distance vary by stop location. For instance, job accessibility tends to have a larger effect for stops located in suburban employment centers than for those in downtown areas and non-downtown areas. This finding implies a non-linear relationship between job accessibility and walking distance. However, the relevant exploration of walking distance to transit is limited. In particular, each of spatial attributes may have differing non-linear relationships with walking distance. If true, the "one size fits all" assumption will produce erroneous results on the associations between spatial attributes and walking distance, and hence will mislead planning practice.

3. Data and method

3.1. Data and variables

This study investigates urban local bus users' walking distance from home to the boarding stop in the Twin Cities (Fig. 1). The data source is the 2016 Transit On Board Survey, conducted by the Metropolitan Council, the Metropolitan Planning Organization in the region. The data were collected on weekdays from April 2016 to February 2017. In the survey transit users were asked to provide the following information: trip purposes, starting and ending locations, access and egress modes, transit routes, and demographic characteristics. Besides paper-based questionnaires, respondents could use a laptop to complete the survey, thus avoiding manual data entry. Furthermore, respondents could choose their starting and ending locations through an interactive map, enhancing location accuracy.

In this research, we studied home-based trips: either trip destinations or originations are home. Walking distance between home and the transit stop, the dependent variable in this study, was measured as the shortest path in the street network through the Network Analysis function in ArcGIS. Due to some probable errors, the trips with a walking distance longer than one mile (1609.3 m) were removed from the analysis. After data pre-processing, 7887 trips were included in the study. About 53% of these trips were made in the early morning (6 am–9 am) or the late afternoon (3 pm–6:30 pm), 35% were during the midday (9 am–3 pm), and 12% were in the evening (6:30 pm–9 pm). As to trip direction, 58% of these trips started from home, and 42% were heading to home. Moreover, 38% of these trips were from home to work, and 28% traveled in the opposite direction. The mean walking distance is about 317 m.

Table 1 defines four categories of independent variables, and Table 2 presents their descriptive statistics. Trip attributes and demographics were from the Transit On Board Survey. We obtained spatial attributes and socioeconomic characteristics directly from the Smart Location Database of the United States Environmental Protection Agency (EPA, http://www.epa.gov/smartgrowth). In this study, spatial attributes refer to the spatial characteristics related to the built environment, which could be manipulated by urban planners. These two sets of variables are measured at the census block group (CBG) level, where transit stops are located. We intentionally chose to use this

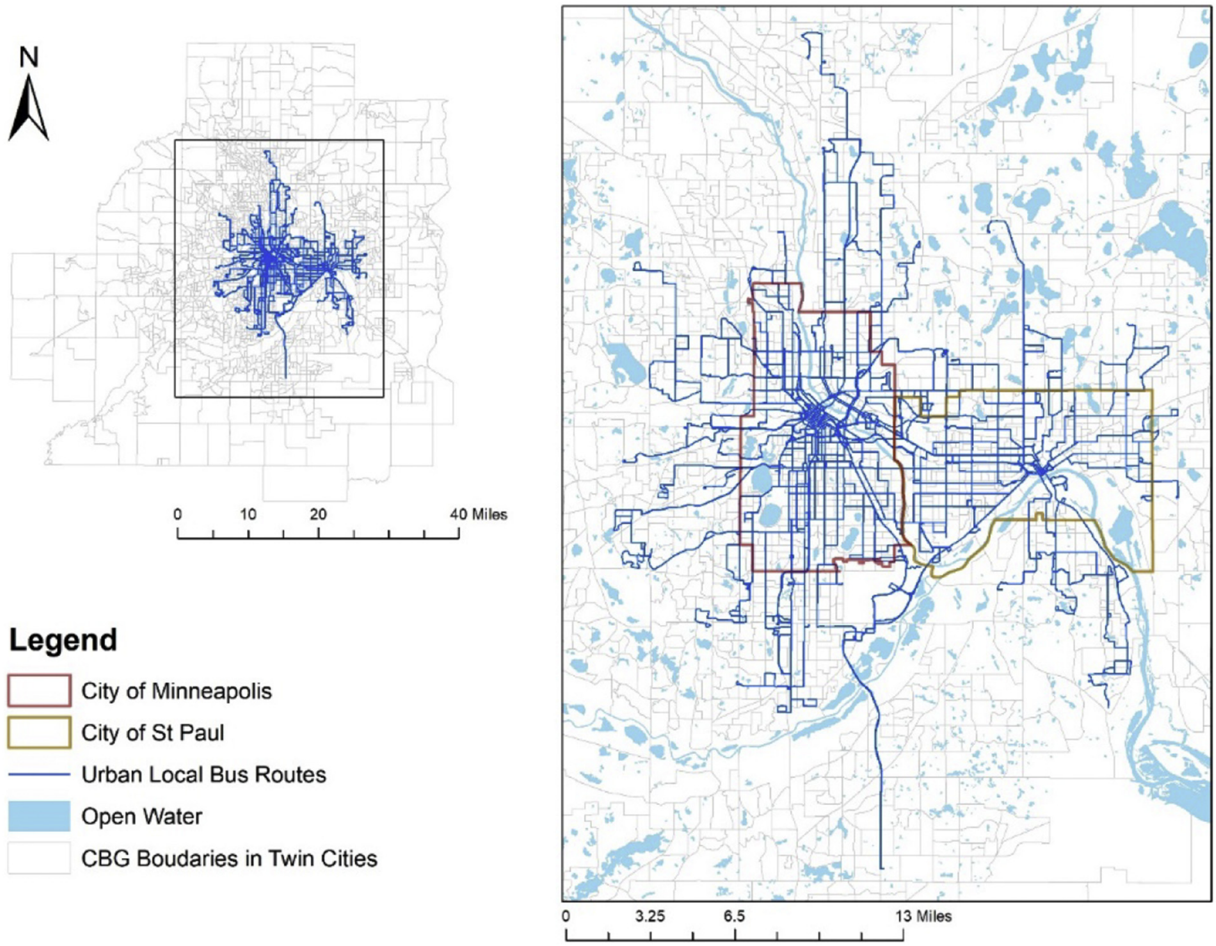

Fig. 1. Study Area.

database so that transit planners in other regions can duplicate our study with minimal effort.

3.2. Method

We used GBDT to examine the relationships between spatial attributes and walking distance, controlling for other factors. The approach was originally developed to predict and interpret data in computer science, and has recently been applied in the field of transportation (Ding et al., 2019; Ma et al., 2017; Wu et al., 2019). It combines both decision tree and gradient boosting methods (Elith et al., 2008; Friedman, 2001).

The decision tree method is to partition an observation space into several regions based on some specific rules. It identifies the regions having the most "homogenous" features and fit a constant to every region for prediction (Elith et al., 2008, p. 803). As the example shown in Fig. 2, we use two criteria to categorize the observations into three regions. The two criteria are whether variable X is smaller than constant a, and whether variable Y is smaller than constant b. Variable X and Y are independent variables, which could be spatial characteristics, socioeconomic characteristics, or other variables in this study. Then we use the average walking distance of the observations in each region to make predictions. The partitioned regions are named terminal nodes or leaves of a decision tree.

The gradient boosting method is to aggregate many weak (or simple) models into one strong (or complex) model in a sequential procedure (Elith et al., 2008). It is usually applied to search the optimal solution (e.g., the minimum value) to a problem with a wide range. As the example shown in Fig. 3, after a starting point is randomly selected, the next step is to find the 'steepest' route to reach the minimum value as quickly as possible. The slope (or direction) of this 'steepest' route is called gradient. After several steps, we could reach the minimum value in the curve.

Integrating these two methods, the main target of the GBDT approach is to combine many trees linearly and sequentially to make the prediction have the best performance. Usually a loss function is defined to evaluate the performance of a model.

We applied the "gbm" package in the R programming language to estimate our model. Ridgeway (2019) designed the package based on Friedman's (2002, 2001) work. The algorithm can be described as follows. At the setup stage, we need to specify several parameters, including training dataset with N observations, the number of iterations T, the depth of each tree K, and the shrinkage parameter λ. We then initialize a function $\hat{f}(\boldsymbol{x})$ as a constant (usually, the mean $\bar{y}$), and set

$$\hat{f}(\boldsymbol{x}) = argmin_{\rho} \sum_{i=1}^{N} \psi(y_i, \rho) \tag{1}$$

where $\psi(y_i, \rho)$ is the loss function for the estimation function of the ith observation, and ρ is the step length. We used the Gaussian distribution in our research, and the corresponding loss is RMSE (root mean squared error).

For iteration t $(1 \leq t \leq T)$, we first calculate the negative gradient to ensure that the value of the loss function ψ will decrease in the next iteration (Friedman, 2001).

$$z_i = -\frac{\partial}{\partial f(\boldsymbol{x}_i)} \psi(y_i, f(\boldsymbol{x}_i)) \Big|_{f(\boldsymbol{x}_i) = \hat{f}(\boldsymbol{x}_i)} \quad i = 1, \ldots, N \tag{2}$$

Table 1
Variable definition.

Variable	Description
Walking distance	Street network distance (meter) from home to the transit stop
Trip attributes	
Peak hour	A dummy variable equaling to 1 if the starting time of the trip is during peak hours (6:00–9:00 am or 4:00–6:30 pm)
Trip distance	Street network distance (mile) from home to the trip destination
Transfer	Number of transfers during the trip
Work destination	A dummy variable equaling to 1 if the destination of the trip is a workplace or a school
Socioeconomic attributes (Census Block Group)	
Working aged population	Percentage of population that is working aged
Households with zero cars	Percentage of zero-car households
Low-wage worker	Percentage of low-wage workers among all workers
Spatial attributes (Census Block Group)	
Population density	Gross population density (people/acre) on unprotected land
Job density	Gross employment density (jobs/acre) on unprotected land
Job and household entropy	Job and household entropy (based on number of activities generated by occupied housing and all five employment categories: retail, office, industrial, service, and entertainment)
Pedestrian network density	Network density in terms of facility miles of pedestrian-oriented links per square mile
Intersection density	Intersection density in terms of multi-modal intersections having four or more legs per square mile
Respondents' demographic attributes	
Male	A dummy variable equaling to 1 if the respondent is male
White	A dummy variable equaling to 1 if the respondent is Caucasian
African American	A dummy variable equaling to 1 if the respondent is African American
Youth	A dummy variable equaling to 1 if the respondent is "under 18 years old"
Senior	A dummy variable equaling to 1 if the respondent is "over 65 years old"
Household income	1 if less than $15,000
	2 if $15,000–$24,999
	3 if $25,000–$34,999
	4 if $35,000–$59,999
	5 if $60,000–$99,999
	6 if $100,000–$149,999
	7 if $150,000–$199,999
	8 if $200,000 or more
Job status	A dummy variable equaling to 1 if the respondent has a full-time or part-time job
Vehicle	A dummy variable equaling to 1 if the respondent has access to a vehicle
Driving license	A dummy variable equaling to 1 if the respondent has a driver's license
Transit pass	A dummy variable equaling to 1 if the respondent has a transit pass

Friedman (2002) suggested using subsamples from the training dataset, instead of the whole training dataset. This subsampling improves the performance of the algorithm significantly. Ridgeway incorporated this into the gbm package and selected $p \times N$ observations randomly from the training dataset, where p is the subsampling rate. This parameter is 0.5 by default in the package and usually generates a smaller deviance (Ridgeway, 2007). For those selected observations, fit a regression tree with K terminal leaves, $g(x) = E(\tilde{z} \mid \boldsymbol{x})$ and compute the optimal step length of each leaf, $\rho_1, \ldots, \rho_K$, as

$$\rho_k = argmin_\rho \sum_{\boldsymbol{X}_i \in S_k} \psi(y_i, \hat{f}(\boldsymbol{x}_i) + \rho) \tag{3}$$

where S_k is the subset of the training dataset, which lies in leaf k.

In the last step of the current iteration, update $\hat{f}(\boldsymbol{x})$ using Eq. (4) and start the next iteration

$$\hat{f}(\boldsymbol{x}) \leftarrow \hat{f}(\boldsymbol{x}) + \lambda \rho_{k(\boldsymbol{x})} \tag{4}$$

where $k(\boldsymbol{x})$ is the index of the regression tree leaf in which an observation with feature $\boldsymbol{x}$ would fall. We iterate the above process for T times to obtain our final estimation function.

GBDT is flexible in that users could change several parameters as needed. Generally, three parameters are important to the model and require regularization. The shrinkage λ, also known as learning rate, is the portion of the contribution of each decision tree that will be added to the final model at each iteration. A smaller shrinkage is preferable since it could improve the prediction of a model, but at the same time it increases the number of trees needed to fit the model and thus the estimation is more time-consuming. We used 0.001 in our research to balance computing time and prediction performance, as suggested by Ridgeway (2019), p. 7. The depth of tree K indicates the complexity of a tree structure. It is the number of terminal nodes in the tree. It takes more time to fit a complex tree. Another important parameter is the number of iterations T. Although a higher number of iterations improves the prediction performance of a model, the model is easily to be overfitting. Overfitting occurs when the model fits too closely to the training dataset but not good to other datasets. We introduced a five-fold cross validation method to alleviate the problem of overfitting. That is, the original dataset is divided into five parts, among which four parts are used to train models, and the remaining one is for testing. The model's fitness to the test dataset will be applied to evaluate its performance.

Compared to traditional methods (such as linear regression and generalized linear regression), the GBDT approach has several advantages. It offers a more accurate prediction, is less vulnerable to outliers, can accommodate missing data of independent variables, and can help address the multicollinearity issue (Ding et al., 2018; Elith et al., 2008). Besides its strong predictive power, the GBDT approach can be used to explain the relationships between variables because it "can provide information concerning the underlying relationship between the inputs x and the output y variable" (Friedman, 2001, p. 1216).

The GBDT approach is particularly useful to address our research questions for two reasons. First, it can generate the relative importance of independent variables (Friedman, 2001). The relative importance represents the extent to which an independent variable contributes to predicting the dependent variable. Because the relative importance of all variables adds up to 100%, we can assess the contribution of one variable or a group of variables, relative to other variables. More

Table 2
Descriptive statistics of variables.

Variable	Mean	Standard deviation	Min	Max
Walking distance (m)	317.24	300.66	0.09	1607.75
Trip attributes				
Peak hour	0.45	0.50	0	1
Trip distance (mile)	4.84	3.21	0.13	32.68
Transfer	0.31	0.53	0	3
Work destination	0.67	0.47	0	1
Socioeconomic attributes				
Working aged population	79.1%	9.8%	46.9%	99.6%
Households with zero cars	16.9%	14.3%	0	89.5%
Low-wage worker	26.8%	5.8%	12.5%	51.9%
Spatial attributes				
Population density (people/acre)	13.96	7.91	0.09	43.60
Job density (job/acre)	5.34	6.32	0	35.83
Job and household entropy	0.52	0.21	0	0.97
Pedestrian network density (mile/Sq.mile)	18.51	5.69	3.30	40.30
Intersection density (intersection/Sq.mile)	14.51	17.12	0	119.44
Demographic attributes				
Male	0.51	0.50	0	1
White	0.58	0.49	0	1
African American	0.29	0.46	0	1
Youth	0.06	0.24	0	1
Senior	0.05	0.22	0	1
Household income	3.43	1.68	1	8
Job status	0.78	0.41	0	1
Vehicle	0.30	0.46	0	1
Driving license	0.58	0.49	0	1
Transit pass	0.11	0.31	0	1

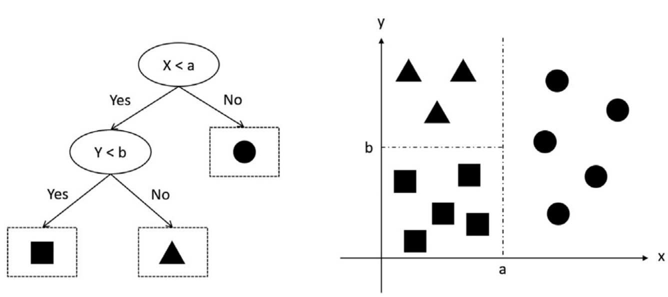

Fig. 2. An example of regression trees.

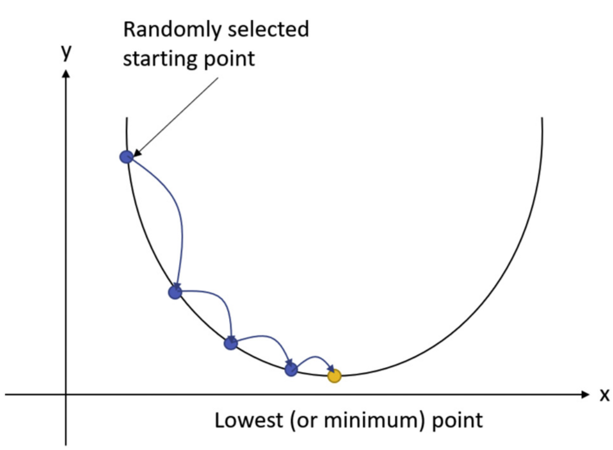

Fig. 3. An example of gradient boosting.

importantly, it can produce a partial dependence plot to illustrate the relationship between the dependent variable and an independent variable, controlling for other independent variables (Friedman, 2001). This plot can show whether the relationship is non-linear.

This approach also has a few limitations. It cannot offer statistical inference for independent variables; GBDT is unable to provide *p*-values to show the significance level of independent variables. By contrast, we emphasized practical significance of independent variables in this study. Another one is overfitting. As discussed before, we applied the cross-validation method to address the influence of overfitting. Lastly, this method can model only the relationships between variables within the observation space. Readers should be cautious when they try to extrapolate the relationships. However, this drawback also exists in other statistical models.

4. Results

4.1. Model regularization

We set the shrinkage λ as 0.001, the maximum number of iterations as 50,000, and the depth of tree K from 1 to 49. A five-fold cross-validation was applied to determine the best number of iterations. We compared model performance using RMSE. As the depth of tree increases, both RMSE and the number of iterations decrease (Fig. 4). Because RMSE decreases substantially when the depth of tree is smaller than 35, we chose 35, with a relatively small RMSE of 295.2, as the depth of tree. The model was converged after 2559 iterations. The Pseudo R^2 is 0.192.

4.2. The relative importance of independent variables

Table 3 presents the relative importance of all the independent variables in predicting walking distance to transit stops in the form of percentage. The sum of all the relative importance of these variables is 100%. Among all categories of the independent variables, spatial attributes collectively contribute to 41.6% of the prediction, which is larger than the collective contribution of socioeconomic attributes (23.5%); trip attributes (20.2%); and demographics (14.7%). Therefore, spatial attributes have the largest power in predicting walking distance among the variables tested. This shows the efficacy of affecting transit users' walking behavior through planning.

In terms of individual spatial attributes, job and household entropy (an indicator of land use mix) has the largest contribution (with a relative importance of 10.5%) in predicting walking distance to transit stops, followed by population density, pedestrian network density, job density, and intersection density. Population density, pedestrian network density, and job density have similar predictive power, around 8%.

Among other variables, trip distance has the largest contribution (16.0%). All three socioeconomic attributes also play an important role; their contributions range from 7.2% to 8.5%. Rider demographics have a relatively limited influence. Specifically, household income has the largest predictive power among all demographics, with a relative importance of 6.2%. The individual contribution of all other variables does not exceed 2%.

4.3. The associations between spatial attributes and walking distance

We use a partial dependence plot to demonstrate how a spatial attribute is associated with the predicted walking distance, controlling for all other independent variables. Besides the fitted curves, we smoothened them to highlight the general trend of the relationship. Fig. 5 presents the partial dependence plot of population density. The general trend is that population density is negatively associated with walking distance, consistent with El-Geneidy et al. (2014). As population density increases, transit users tend to walk a shorter distance. In densely populated areas, more transit stops are needed to meet the travel demand of more users. Consequently, transit users could walk a shorter distance to reach one of those stops. In particular, when population

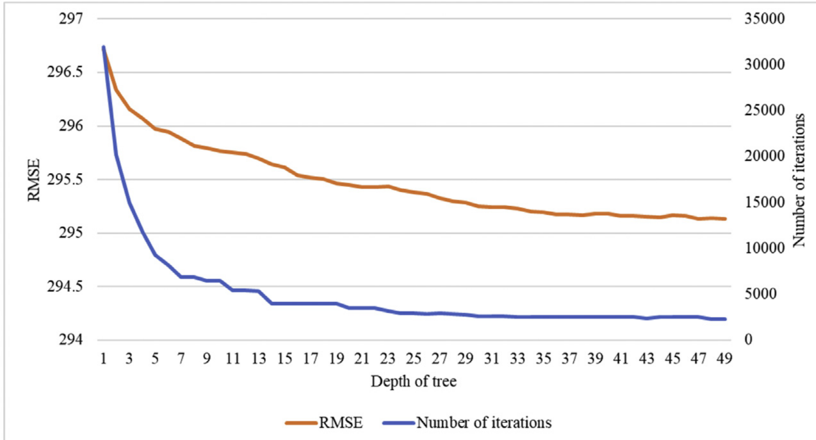

Fig. 4. RMSE and number of iterations versus tree depth.

Table 3
Relative importance of all the variables.

Variable	Relative importance (%)	Total (%)
Trip attributes		
Peak hour	1.2	20.2
Trip distance	16.0	
Transfer	1.9	
Work destination	1.1	
Socioeconomic attributes		
Working aged population	8.5	23.5
Households with zero cars	7.2	
Low-wage worker	7.7	
Spatial attributes		
Population density	8.9	41.6
Job density	8.0	
Job and household entropy	10.5	
Pedestrian network density	8.5	
Intersection density	5.8	
Demographic attributes		
Male	1.5	14.7
White	1.0	
African American	0.8	
Youth	0.7	
Senior	0.3	
Household income	6.2	
Job status	1.0	
Vehicle	1.1	
Driving license	1.4	
Transit pass	0.8	
Total		100

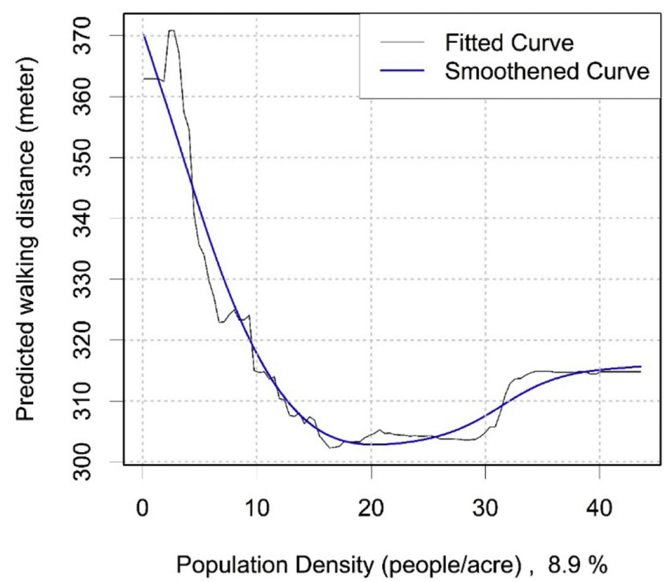

Fig. 5. Partial dependence plot of population density.
The relative importance of the variable is presented in the label for the horizontal axis, same for the following partial dependence plots.

density increases from 0 to 18 people/acre, walking distance drops substantially from 371 m to 303 m. Walking distance becomes stable after 18 people/acre. After about 30 people/acre, walking distance has a slight increase from 303 m to 315 m, with unknown reasons. From the perspective of stop area planning, densifying transit serving areas to 18 people/acre or larger helps shorten the walking distance to transit stops. Fig. 6 identifies the CBGs where population density is not less than 18 people/acre in the region. We found that most of these CBGs are located in the downtowns and adjacent areas. Because the vast majority of CBGs do not meet the threshold, there is a great potential to shorten users' walking distance by densifying the areas along existing transit routes.

Job density is positively associated with walking distance (Fig. 7). As job density increases, people tend to walk a longer distance to reach transit stops. On one hand, more jobs near residential areas may make these areas more pedestrian-friendly (such as main streets) and hence encourage walking. On the other hand, high job density around transit stops may push residential developments to the periphery of business establishments. Residents may have to walk a longer distance to reach transit stops. When job density increases from 0 to 10 jobs/acre, walking distance increases from 301 to 320 (note that there is some noise at the starting point of the fitted curve). Walking distance is relatively stable from 10 to 20 jobs/acre but increases rapidly again after 20 jobs/acre. The curve is saturated at about 30 jobs/acre. For stop area planning, increasing job density to 10 jobs/acre could help increase rider willingness to walk longer while for stop location choice, the locations with job density larger than 10 jobs/acre should be prioritized to enlarge the service area. Fig. 8 illustrates the CBGs where job density is larger or equal to 10 jobs/acre. To maximize transit service areas, the job-rich areas along existing transit routes, particularly those in urban areas, could be potential locations for transit stops.

Job and household entropy around transit stops is positively correlated with walking distance (Fig. 9). However, it is the larger mix that is associated with walking distance. In particular, when the index increases from 0 to 0.7, the walking distance is relatively constant. In the interval from 0.7 to 1.0, it rises dramatically from 309 m to 450 m. Fig. 10 presents the areas where job and house entropy is larger or equal to 0.7 in the region. The highly mixed areas along transit routes

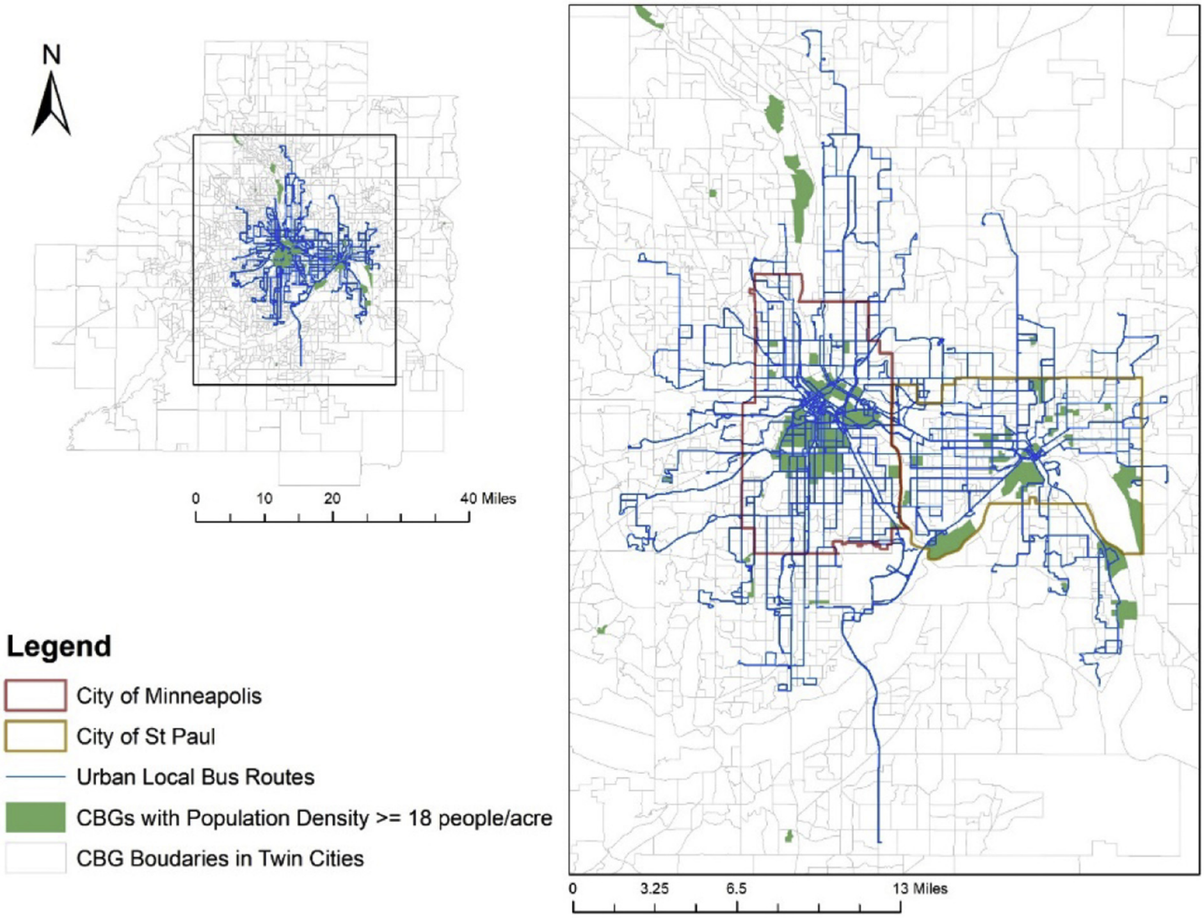

Fig. 6. Population density distribution.

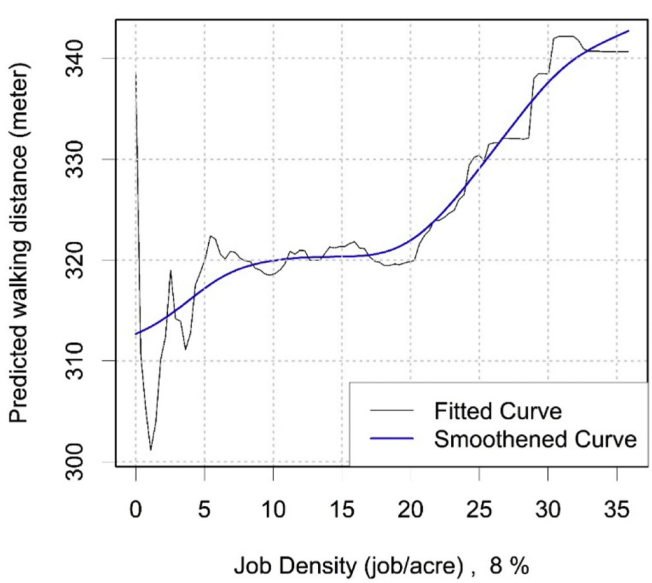

Fig. 7. Partial dependence plot of job density.

could be prioritized for transit stops to enlarge transit service areas. It is worth noting that although many CBGs outside of transit service areas have higher entropy index, they are not suitable for frequent transit services.

Fig. 11 shows that pedestrian network density has a positive relationship with walking distance. This result is consistent with Maghelal (2011). Higher pedestrian network density can encourage transit users to take a longer walk to reach transit stops. Pedestrian network density has almost a linear relationship with walking distance in the interval from 0 to 33 miles per square mile. After that, walking distance increases to 355 m, the saturation level. Fig. 12 shows that there is a positive association between intersection density and walking distance, consistent with El-Geneidy et al. (2014). Higher intersection density around transit stops indicates better network connectivity, and hence increases riders' willingness to walk a longer distance to stops. After intersection density reaches 90 per square mile, walking distance becomes stable at 346 m. In the Twin Cities, only several CBGs (not shown) have reached the thresholds of pedestrian network density (33 miles per square mile) and intersection density (90 per square mile). For the sake of enlarging transit service areas, there is a large room to improve these two spatial attributes.

4.4. Comparison with linear regression

We estimated a linear regression model and compared it with the GBDT model (Table 4). The partial dependence plots and the signs of the estimated coefficients in the linear regression demonstrate the relationship between independent variables and walking distance. As shown in Table 4, linear regression presents consistent results with the GBDT model in terms of spatial attributes. However, the R^2 of the linear regression model is 0.027, much lower than the pseudo R^2 of the GBDT model (0.192). This is not surprising because the GBDT model considers the non-linear relationships between independent variables and the dependent variable, and improves the overall goodness of fit.

We also compared the relative contribution of independent variables to walking distance. In linear regression, adding one independent

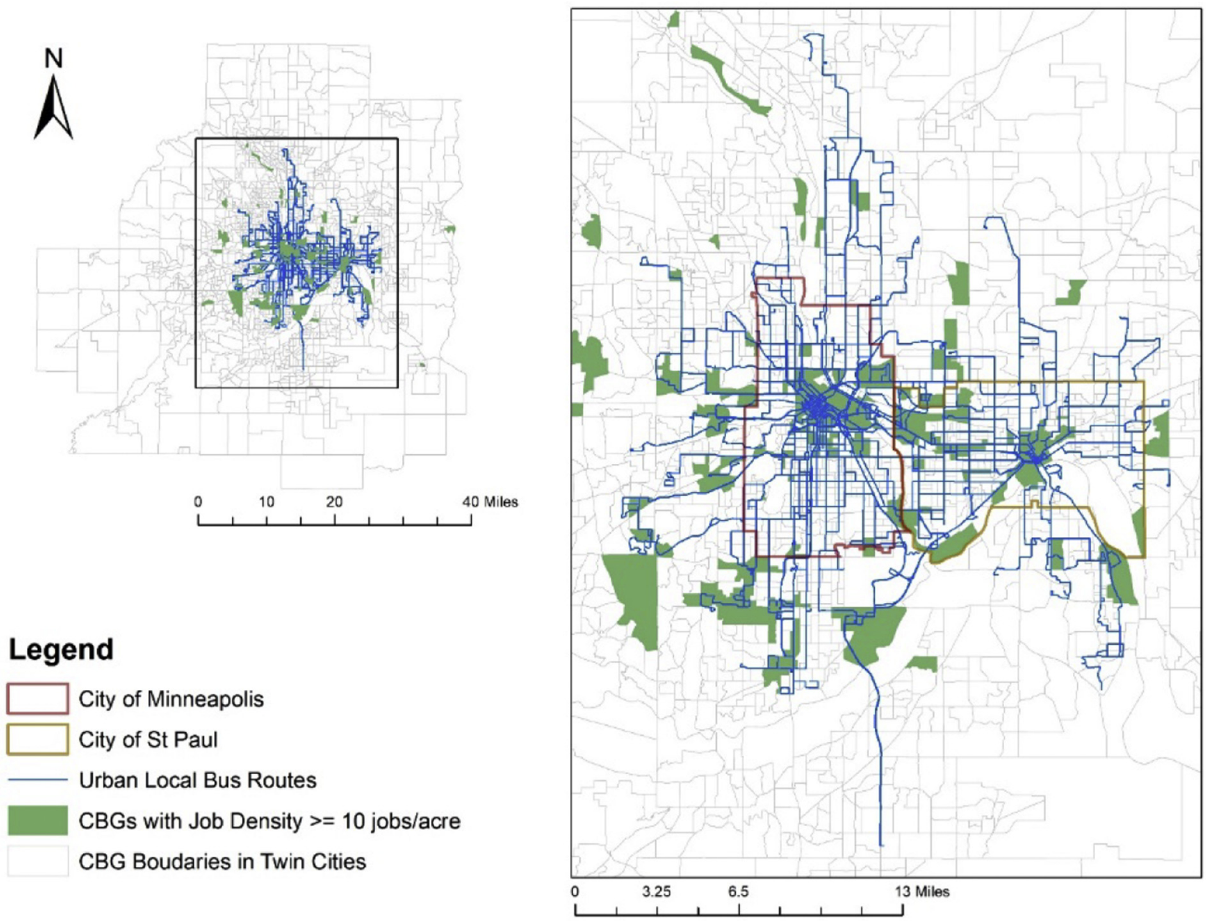

Fig. 8. Job density distribution.

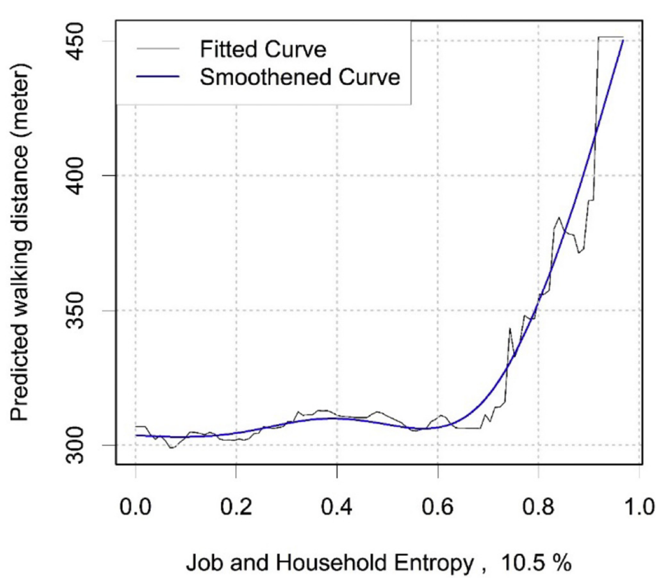

Fig. 9. Partial dependence plot of job and household entropy.

variable will improve R^2. The share of this improvement in the total R^2 is counted as the relative importance of the independent variable. Because the order of adding one variable affects the improvement, we computed the average improvement of the variable by considering all possible orders. This computation was carried out using the "relaimpo" package in R. Overall, the relative importance of independent variables differs between the GBDT model and the linear regression. Because the GBDT model showed the non-linear associations between independent variables and walking distance, the linearity assumption of the regression model is flawed. Therefore, the relative importance produced by the linear model is erroneous.

5. Conclusions

Using the 2016 Transit On Board Survey data in the Twin Cities, we examined the importance of spatial attributes to transit users' walking distance to stops and their non-linear associations. This study contributes twofold to the literature and planning practice. First, it assesses the collective importance of spatial attributes relative to other influential factors and evaluates the efficacy of using land use and transit planning to shape transit users' walking distance to access transit stops. Second, it investigates the non-linear relationships between spatial attributes and walking distance and identifies the most effective ranges of these attributes on walking distance.

The results showed that spatial attributes play a more important role in predicting walking distance than trip attributes, users' characteristics, and the socioeconomic environment. This underscores the critical role of spatial planning in influencing walking distance and provides supportive evidence for local planners to change the built environment around stops. Furthermore, all individual spatial dimensions tested here are important to riders' walking choice. Relatively, intersection density has a smaller contribution than other four spatial attributes, namely, population density, job density, mixed-use index, and pedestrian network density.

The partial dependence plots demonstrated that some spatial attributes have clear threshold effects on walking distance to stops, and that the non-linear patterns vary by variable (for example, population

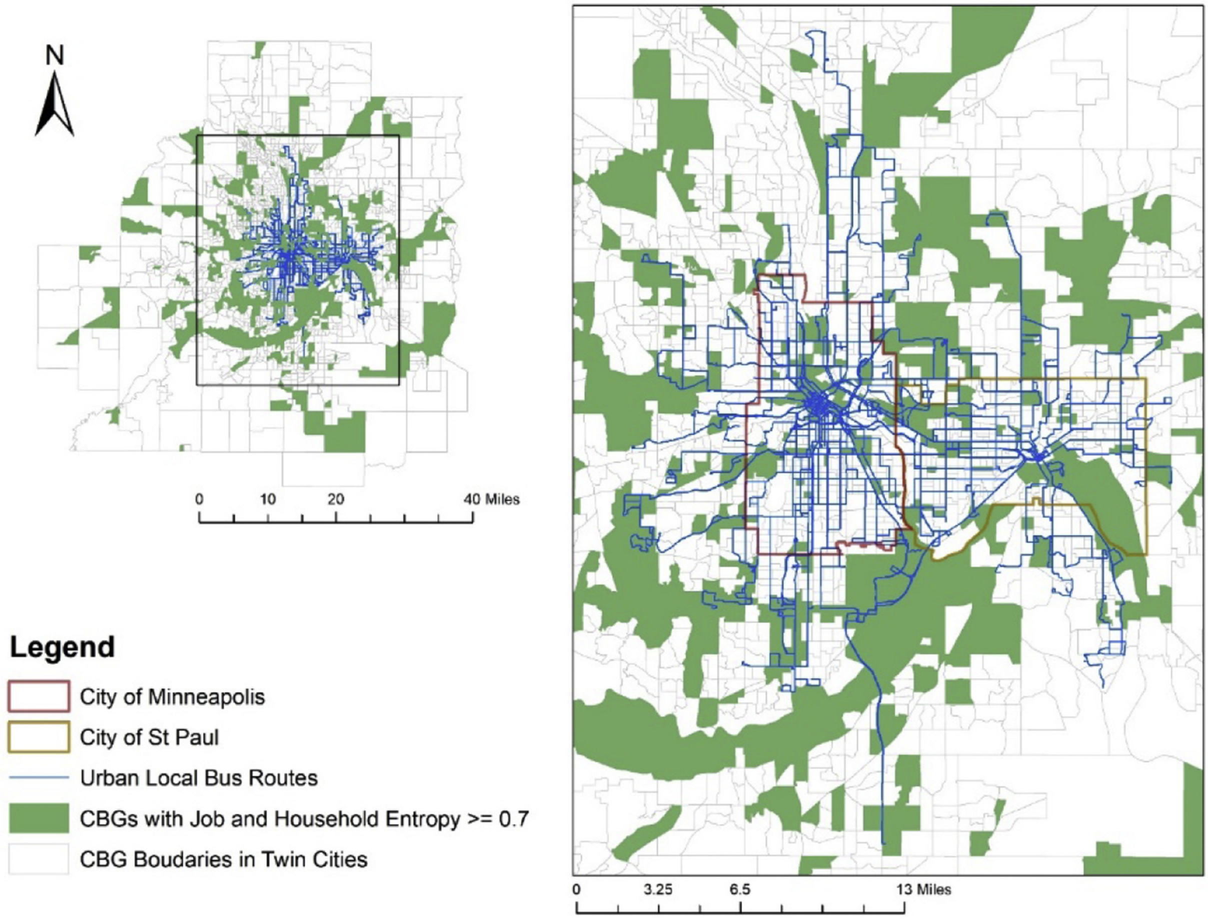

Fig. 10. Job and household entropy.

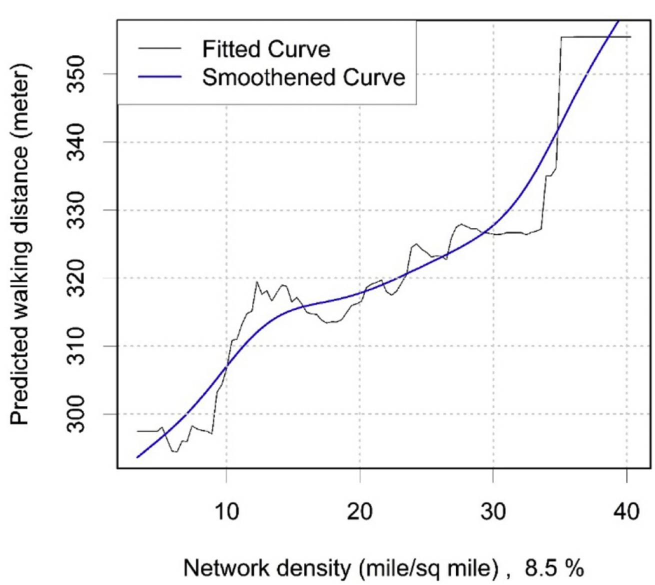

Fig. 11. Partial dependence plot of pedestrian network density.

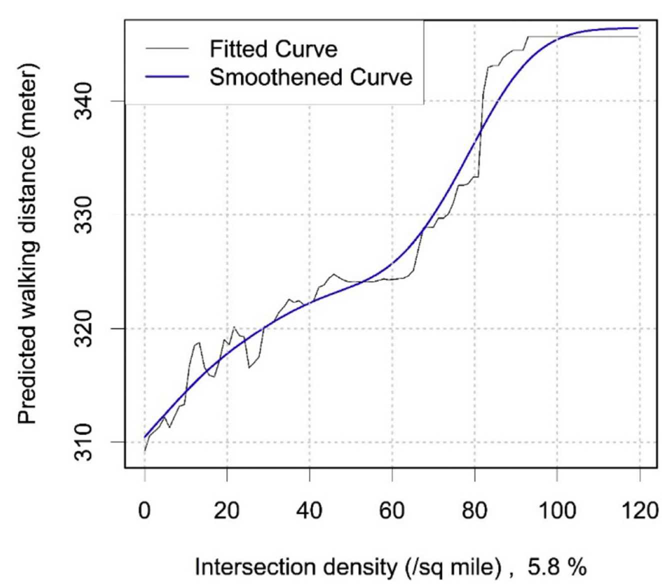

Fig. 12. Partial dependence plot of intersection density.

density and job-household entropy). First, these findings challenge the linearity assumption commonly adopted in the studies about environmental correlates of transit riders' walking distance. Future studies should consider the non-linear associations. Second, the GBDT approach is more effective in revealing the complex non-linear relationships between spatial attributes and walking distance than traditional models (such as linear regression and generalized linear regression), because the latter have a limited capacity to uncover the varying patterns.

Among the five Ds, population density is the only variable that is negatively associated with walking distance. The negative relationship is likely to be an outcome of transit planners' intentional effort to

Table 4
Comparison between GBDT and linear regression.

Variable	GBDT			Linear regression		
	Relationship	Relative Importance	Sum	Estimates	% R^2	Sum
Trip attributes						
Peak hour		1.2	20.2	−15.9	1.8	46.4
Trip distance		16.0		11.2	33.3	
Transfer		1.9		−36.7	5.9	
Work destination		1.1		19.6	5.4	
Socioeconomic attributes						
Working aged population		8.5	23.5	100.0	3.4	4.9
Households with zero car		7.2		4.9	0.4	
Low-wage worker		7.7		150.8	1.1	
Spatial attributes						
Population density	Negative	8.9	41.6	−1.6	4.6	31.7
Job density	Positive	8.0		0.9	6.2	
Job and household entropy	Positive	10.5		91.3	17.2	
Network density	Positive	8.5		1.6	1.1	
Intersection density	Positive	5.8		0.6	2.6	
Demographic attributes						
Male		1.5	14.7	13.6	2.4	17.0
White		1.0		18.3	0.8	
African American		0.8		20.3	1.2	
Youth		0.7		29.0	0.8	
Senior		0.3		−30.2	2.9	
Household income		6.2		1.0	0.1	
Job status		1.0		20.6	2.8	
Vehicle		1.1		−22.0	3.0	
Driving license		1.4		−12.1	1.3	
Transit pass		0.8		22.1	1.7	
R^2	0.192			0.027		

deploy more stops to meet transit demand in densely populated areas. From the perspective of stop location choice, placing stops closer to transit users helps reduce their walking distance, and in turn promote transit ridership (Gutiérrez et al., 2011; Hess, 2009). From the perspective of stop area planning, increasing population density around stops helps lower their walking distance. However, because densification increases development costs and may face oppositions from surrounding residents (not in my back yard), excessive densification may not be desirable. This study shows that 18 persons/acre is sufficient to optimize walking distance.

By contrast, job density, job and household entropy, pedestrian network density, and intersection density are positively associated with walking distance. These positive associations suggest that appropriate land use planning around stops has the potential to enlarge transit catchment areas. In particular, promoting employment densification, mixed-use development, and multi-modal street connectivity helps. To increase transit service areas, 10 jobs/acre should be the minimum planning goal for local centers; and for large employment centers, 30 jobs per acre should be the planning goal. Furthermore, land use should be sufficiently mixed to be effective in encourage transit riders to walk a longer distance. Because greater multi-modal network connectivity always helps enlarge transit service areas, grid street patterns with sidewalks are desirable for stop area planning.

The results presented in this study should be interpreted with caution. First, the data are cross-sectional, so the relationships found here are more of correlations than causality. This does not differ from most studies on the topic in the literature, however. Second, because this study is the first one that identifies the effective ranges of spatial attributes, the thresholds found in the Twin Cities may not be transferable to other regions with different sizes, different built environments, and different transit supplies. The generalizability merits further investigation. We encourage planners in other regions to carry out similar studies using our study protocol with the EPA data and the R programming code (Tao, 2018). To facilitate this comparison, we intentionally used the EPA data, which are readily available for the whole nation. However, spatial attributes at the CBG level may not accurately reflect the built environment around a transit stop, particularly when it is close to the CBG boundary. Alternatively, some studies used a buffer of the stop to define its surrounding area (El-Geneidy et al., 2014; Maghelal, 2011). Future research should examine whether CBG-based spatial attributes and buffer-based ones produce consistent results.

References

Alshalalfah, B.W., Shalaby, A.S., 2007. Case study: relationship of walk access distance to transit with service, travel, and personal characteristics. J. Urban Plan. Dev. 133, 114–118. https://doi.org/10.1061/(ASCE)0733-9488(2007)133:2(114).

Cao, X.J., 2019. Examining the effect of the Hiawatha LRT on auto use in the twin cities. Transp. Policy 81, 284–292. https://doi.org/10.1016/j.tranpol.2018.04.011.

Ding, C., Cao, X. (Jason), Næss, P., 2018. Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. Transp. Res. Part A: Policy. Pract. 110, 107–117. https://doi.org/10.1016/j.tra.2018.02.009.

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. https://doi.org/10.1016/j.jtrangeo.2019.04.011.

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., Surprenant-Legault, J., 2014. New evidence on walking distances to transit stops: identifying redundancies and gaps using variable service areas. Transportation 41, 193–210. https://doi.org/10.1007/s11116-013-9508-z.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Ewing, R., Cervero, R., 2017. "Does compact development make people drive less?" The nswer is yes. J. Am. Plan. Assoc. 83, 19–25. https://doi.org/10.1080/01944363.2016.1240044.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Gutiérrez, J., García-Palomares, J.C., 2008. Distance-measure impacts on the calculation of transport service areas using GIS. Environ. Plan. B: Plan. Design 35, 480–503. https://doi.org/10.1068/b33043.

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, 1081–1092. https://doi.org/10.1016/j.jtrangeo.2011.05.004.

Hess, D.B., 2009. Access to public transit and its influence on ridership for older adults in two U.S. cities. J. Transp. Land Use 2, 3–27. https://doi.org/10.5198/jtlu.v2i1.11.

Hsiao, S., Lu, J., Sterling, J., Weatherford, M., 1997. Use of geographic information system for analysis of transit pedestrian access. Transp. Res. Rec. 1604, 50–59. https://doi.org/10.3141/1604-07.

Jiang, Y., Christopher Zegras, P., Mehndiratta, S., 2012. Walk the line: station context, corridor type and bus rapid transit walk access in Jinan, China. J. Transp. Geogr. 20, 1–14. https://doi.org/10.1016/j.jtrangeo.2011.09.007.

Loutzenheiser, D.R., 1997. Pedestrian access to transit: model of walk trips and their design and urban form determinants around Bay Area rapid transit stations. Transp. Res. Rec. 1604, 40–49. https://doi.org/10.3141/1604-06.

Ma, X., Ding, C., Luan, S., Wang, Y., Wang, Y., 2017. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. IEEE Trans. Intell. Transp. Syst. 18, 2303–2310. https://doi.org/10.1109/TITS.2016.2635719.

Maghelal, P.K., 2011. Walking to transit: Influence of built environment at varying distances. ITE Journal 81 (2), 38–43. https://search.proquest.com/docview/851791472 accessed 8.11.19.

Mokhtarian, P.L., Van Herick, D., 2016. Quantifying residential self-selection effects: a review of methods and findings from applications of propensity score and sample selection approaches. J. Transp. Land Use 9, 9–28. https://doi.org/10.5198/jtlu.2016.788.

Nelson, A.C., 2017. Compact development reduces VMT: evidence and application for planners—comment on "does compact development make people drive less?". J. Am. Plan. Assoc. 83, 36–41. https://doi.org/10.1080/01944363.2016.1246378.

O'Sullivan, S., Morrall, J., 1996. Walking distances to and from light-rail transit stations. Transp. Res. Rec. 1538, 19–26. https://doi.org/10.3141/1538-03.

Ridgeway, G., 2007. Generalized boosted models: a guide to the gbm package. Compute 1, 1–12. Available online at. http://finzi.psych.upenn.edu/R/library/gbm/doc/gbm.pdf.

Ridgeway, G., 2019. Package ' gbm ' [WWW Document]. The R Project for Statistical Computing. https://cran.r-project.org/web/packages/gbm/gbm.pdf accessed 8.11.19.

Singh, A.C., Astroza, S., Garikapati, V.M., Pendyala, R.M., Bhat, C.R., Mokhtarian, P.L., 2018. Quantifying the relative contribution of factors to household vehicle miles of travel. Transp. Res. Part D: Transp. Environ. 63, 23–36. https://doi.org/10.1016/j.trd.2018.04.004.

Stevens, M.R., 2017a. Response to commentaries on “Deos compact development make people drive less?”. J. Am. Plan. Assoc. 83, 151–158. https://doi.org/10.1080/01944363.2016.1240044.

Stevens, M.R., 2017b. Does compact development make people drive less? J. Am. Plan. Assoc. 83, 7–18. https://doi.org/10.1080/01944363.2016.1240044.

Tao, T., 2018. Analyzing people's walking distance to access transit stops with the method of Gradient Boosting Decision Tree [WWW Document]. GitHub repository. https://vtao1989.github.io/DisToTransit_statistics/ (accessed 8.11.19).

Tilahun, N., Li, M., 2015. Walking access to transit stations. Transp. Res. Rec. 2534, 16–23. https://doi.org/10.3141/2534-03.

Townsend, C., Zacharias, J., 2010. Built environment and pedestrian behavior at rail rapid transit stations in Bangkok. Transportation 37, 317–330. https://doi.org/10.1007/s11116-009-9226-8.

Van Wee, B., Handy, S., 2016. Key research themes on urban space, scale, and sustainable urban mobility. Int. J. Sustain. Transp. 10, 18–24. https://doi.org/10.1080/15568318.2013.820998.

Wang, J., Cao, X., 2017. Exploring built environment correlates of walking distance of transit egress in the Twin Cities. J. Transp. Geogr. 64, 132–138. https://doi.org/10.1016/j.jtrangeo.2017.08.013.

Wu, X., Tao, T., Cao, J., Fan, Y., Ramaswami, A., 2019. Examining threshold effects of built environment elements on travel-related carbon-dioxide emissions. Transp. Res. Part D: Transp. Environ. 75, 1–12. https://doi.org/10.1016/j.trd.2019.08.018.

Zhao, J., Deng, W., 2013. Relationship of walk access distance to rapid rail transit stations with personal characteristics and station context. J. Urban Plan. Dev. 139, 311–321. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000155.

Zhao, F., Chow, L.-F., Li, M.-T., Gan, A., Ubaka, I., 2003. Forecasting transit walk accessibility : a regression model alternative to the buffer method. Transp. Res. Rec. 1835, 16. https://doi.org/10.3141/1835-05.

Ziliak, S.T., McCloskey, D.N., 2004. Size matters: the standard error of regressions in the American Economic Review. J. Socio-Econ. 33, 527–546. https://doi.org/10.1016/J.SOCEC.2004.09.024.