Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 8246575, 30 pages
https://doi.org/10.1155/2021/8246575

Research Article

Local or Neighborhood? Examining the Relationship between Traffic Accidents and Land Use Using a Gradient Boosting Machine Learning Method: The Case of Suzhou Industrial Park, China

Yueming Yang, Hyungchul Chung, and Joon Sik Kim

Urban Planning and Design, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou Industrial Park, Suzhou, Jiangsu Province, China

Correspondence should be addressed to Hyungchul Chung; hyungchul.chung@xjtlu.edu.cn

Received 19 December 2019; Revised 24 December 2020; Accepted 13 January 2021; Published 27 January 2021

Academic Editor: Francesco Galante

In cities, road traffic accidents are critical endangerment to people's safety. A vast number of studies which are designed to understand these accidents' leading causes and mechanisms exist. The widely held view is that emerging analysis methods can be a critical tool for understanding the complex interactions between land use and urban transportation. Using a case study of Suzhou Industrial Park (SIP) in Suzhou, China, this paper examines the relationship between different land use types and traffic accidents using a gradient boosting model (GBM) machine learning method. The results show that the GBM can be used as an effective accident model for a variety of research and analysis methods by (1) ranking the influential factors, (2) testing the degree of interpretation of each variable as the complexity of iterations changes, and (3) obtaining partial dependence plots, among other methods. The findings of this study also suggest that land use types—including facility points—demonstrate differing degrees of influence at two geographical scales: local level and neighborhood level. In the ranking of relative importance at both scales, the variables of education institutions, traffic lights, and service institutions are all ranked high—with a more significant influence on the occurrence of accidents. However, residential land and land use mix variables differed significantly in both scales and showed a significant deviation compared to the other results. When adjusting the complexity of the decision tree, the local level is more suitable for measuring variables such as residential areas and green parks where pedestrians and vehicles have fixed mobility periods and moderate flows. On the contrary, the nearest neighborhood level is more suitable to a small number of variables related to public service facilities at fixed locations, such as traffic lights and bus stops. In the partial dependence plots, all variables, except educational institutions and residences, show a positive correlation for accidents in the fitting process. The results of this study can ideally help inform transportation planners to reconsider transport accident occurrence rates in the context of the proximity to various land use types and public service facilities.

1. Introduction

Traffic safety is a crucial issue affecting the quality of urban residential life. According to global statistics from the World Health Organization (WHO), around 3,700 people die per day due to road traffic collisions, and tens of millions suffer related injuries each year [1]. China has one of the highest rates of traffic accidents in the world, with more than 260 thousand fatalities annually. The WHO's 2015 global status report on road safety [2] indicates that 18.2 deaths per every 10,000 people occur in China, a statistic which also reflects the world average. However, China's rate is higher than the rest of the Western Pacific region's average of 16.9 deaths per every 10,000 people [2], and it only falls below Southeast Asia and Africa in the six major regions designated by the WHO (see Figure 1).

Traffic accidents threaten people's lives, in addition to generating substantial economic losses. In general, traffic accidents involve the subjective actor (the driver) and the objective environment (vehicles and roads). Exploring the

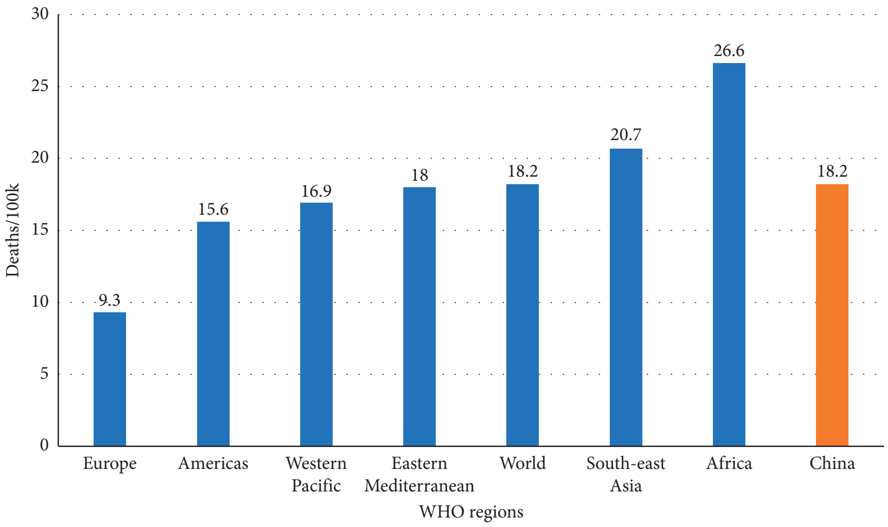


Figure 1: WHO statistics for death on the road in 2015.

causes and mechanisms of traffic accidents within this dynamic will help reduce their overall risk.

Earlier studies have already identified these specific influencing factors:

(1) The natural environment: cloudy or rainy weather, temperature, humidity, and visibility are proven to be related to traffic accidents [3, 4]; lousy weather (rain/fog/snow) also has a significant positive correlation with accidents [5].

(2) Road conditions: greater complexity of the road environment, including a number of intersections, road network density, and a number of vehicles, is likely to create more potential risks of an accident [6, 7].

(3) Human conditions: high employment and population densities in a given site lead to an increase in traffic flow and consequently increase accidents [8–10]; accordingly, sparsely populated areas have lower accident rates [11]. Age demographics and education levels will also affect the times and frequencies of people going out, thus influencing the conditions for accidents [11].

(4) The social environment: studies show the relationship between facility accessibility, land use, and traffic accidents. For example, industrial land, commercial land, and land mix are all positively correlated with traffic accidents [5]. Educational land use areas report varied results in the literature. For instance, evidence shows that educational land use has a significant impact on accidents [4], while a study found that education land use has the least magnitude among other factors on accidents [12]. Natural land use corresponds to the highest safety level, demonstrating the fewest risks [8, 13]. However, Zou et al. [14] uses truck crash severity data in New York City to examine whether traffic accidents are caused by land use patterns rather than land types. He points out that, in one case, both service employment and recreational employment occupy high-density land, but fewer traffic accidents occur in the service employment region. His study also demonstrates that this is subject to change in any given environment and becomes more evident in the recreational employment region. This illustrates the need to take more detailed land use into account when considering accident conditions.

Earlier research on traffic accidents can also be divided into micro- and macroscales [15, 16]. The former focuses on the road itself, such as crossroads or intersections [7, 13, 17] and highways [18, 19]; a closer look examines road length and width, vehicle speeds, and traffic flow, to list a few, as tools to optimize the road structure. On the macrolevel, most census tracts [4, 20, 21], traffic analysis zones (TAZs) [8, 22–24], and living communities [24] are used to identify social and economic factors (such as population and land use) that illustrate the spatial agglomeration of accidents.

Looking from different scales also tends to indicate diverse accident outcomes. Huang et al. [25] point out that detecting road facility as hotspots is more accurately an analysis tool than observing their entire encompassing region. Some crossroad traffic accident studies delineate the scales of 15 m, 60 m, and 75 m, respectively, and correspondingly reach different conclusions [26, 27]. Some results indicate that collisions are more likely to occur at distances of 100 to 200 ft. from intersections, while some of the experimental results are smaller, below 50 ft. Yu and Zhu [28] found that creating buffer zones around schools (with distances of 0.5, 1, 1.5, and 2 miles, respectively) will each impact security of the school zone differently. In this way, they demonstrate the tangible biases that examining land use at different scales present, given that every scale will incorporate a different range of influencing factors. Nevertheless, autocorrelation and heterogeneity of spatial effects must also be considered; regardless of scale, geographic units with higher internal similarities will achieve more stable statistical results.

Among these commonly used research scales, TAZ is the only regional system associated with transportation.

Compared with larger geographic scales, TAZ has better internal similarities in land use, road network, and traffic operation. In light of comparably smaller geographic scales, TAZ would link traffic data to produce more evident socioeconomic characteristics. The scale at which TAZ operates is also easy to integrate with the transportation planning process and is therefore used as a local level research scale in this paper.

In addition to discussing the scale of TAZ, this paper is going to address other research scales. In western literature, accident research in geographic scales also includes local areas [29], counties [16, 30, 31], and regions [32]. In the Chinese context, although the study of land form is often divided into administrative regions such as provinces and cities [33, 34] or terrain areas such as plateaus and hills [35–37], a consistently defined scope at which road network patterns are observed to impact traffic safety remains neglected in the literature. Therefore, it is necessary to explore different research scales that are more suitable for site characteristics and data. This research is a contribution to help fill this gap.

This paper attempts to utilize a scale that is relatively homologous to that of the TAZ model. The center of each TAZ is used as the centroid to generate Thiessen polygons using ArcGIS, thus avoiding the problem of missing or duplicated study areas that can easily be caused by buffers. The Thiessen polygons are irregularly shaped polygons with varying areas based on the centroid, and they are more spatially homogeneous than administrative boundaries. After studying collision models with different spatial units such as census tracts, state electoral divisions, developed grid cells, and natural area boundaries, some scholars have recommended Thiessen polygons because of their higher spatial performance [38–40]. The Thiessen polygon is, therefore, chosen as the research scale of the nearest neighborhood level apart from the TAZ area.

The different study scales may inadvertently create a modifiable areal unit problem (MAUP) or the issue of changing statistical properties due to differences in areal units. However, since both selected scales build on existing TAZ areas and do not involve adjustments in basic spatial units such as census tracts or urban structure based on major roadways, the changes in the statistical results should be modest at most [40]. The results from the two study scales (the local level of the TAZ area and the nearest neighborhood level bounded by the Tyson polygon) would then be comparatively analyzed.

Most of the methods employed in early quantitative studies of traffic accidents focus mainly on singular influencing factors. Gasparini [41] and Li et al. [42] first adopted Markov chain traffic accident statistical models to analyze the time factors of traffic accidents; Kim and Yamashita [43] compared the number of accidents per unit area in different land use and found that commercial geographic entities have the lowest level of traffic safety.

Later studies began to incorporate the generalized linear model (GLM) and used to study the relationship between various influencing factors and the frequency of accidents. The logistic model and the logit model were used multiple times in US studies to analyze accident severity [3, 18, 29, 44]; Kim et al. [45] and Dissanayake et al. [46] employed Poisson regression and negative binomial regression. They examined the respective relationships between geographic entities such as parks, businesses, schools, and high-density residential buildings with traffic accidents. It is not easy to measure the difference between various geographic units. Adjacent geographic units usually have spatial autocorrelation, yet spatial heterogeneity often occurs when they are far apart. Random parameters and spatial models such as Bayesian space models use spatial autoregressive models (SAMs) and geographically weighted regression (GWR) models to solve this issue via traffic safety spatial analyses [24, 47, 48].

It can be concluded that earlier research focuses on macroanalysis of accidents with the intent to produce statistics from their data by emphasizing single factors or multiple factors in the process. Since the 1990s, alongside the development of machine learning and the advancement of data mining technology, systematic causation is being studied increasingly often with the application of machine learning models to traffic accidents. Li and Shao [6] use backpropagation (BP) neural networks and the artificial neutral network (ANN) as methods to identify critical causal factors to the severity of injuries in traffic accidents. The neural network method incorporates the occurrence of traffic accidents as an input and output system. Influencing factors such as people, vehicles, roads, and the traffic environment are considered as input layer variables. The number of accidents or fatalities is operated as output layer variables. Through multiple corrections of parameters, a complex, nonlinear relationship network model between variables is established and more accurate traffic accident analysis results are obtained. Chong et al. [49] proposed testing artificial neural networks and decision trees to model the severity of traffic accident injuries. Experiments using datasets obtained from national automotive sampling systems showed that decision trees outperformed neural networks. Advanced algorithms that have already been applied to the monitoring of sudden traffic events include the probability neural network (PNN) and the support vector machines (SVMs) [50, 51]; Al-Ghamdi et al. [3] introduce a mixed model of wavelets transformation and logistic regression to their traffic events testing method. Further research suggests that the causes of traffic accidents are systematic and intrinsically linked. BP and ANN machine learning methods establish nonlinear networks with input variables. While the obtained results from these models are interpreted in terms of causal relationships, the outcome parameters can be compared either within the same dataset [49] or across different models.

Recent big data technology uses data mining and machine learning techniques to calculate traffic data, identify potential risk factors, and assist in offering targeted measures to avoid and prevent traffic accidents. This paper uses a new machine learning method known as the gradient boosting model (GBM) as a novel application to the traffic accident research field. In doing so, this research aims to explore the relationship between complex land use characteristics and traffic accidents. Using the same data source, the scale of the local level bounded by TAZ and the scale of the nearest neighborhood level bounded by the Thiessen polygons are separately counted to check the power of the model and the explanatory effect of the variables at the two locational levels.

The following section will introduce the mathematical model of the methodology used in this paper. The third section will then present the data and variables. The fourth section contains the experimental results and discussion and divided between the preparation of the GBM model, the interpretation degree of each variable, the explanatory power within the two scales in the case of a change in parameter "number of trees," and the partial dependence of each variable. Section 5 will presents the conclusions and limitations of this research.

2. Materials and Methods

2.1. Methodological Review. This section demonstrates a development from traditional statistical methods to data-driven methods. Mannering et al. [52] point that the choice of analysis method for crash data should take into account the trade-off between prediction capability and the causal nature of factors contributing to accidents. Traditional statistical methods have been relatively easy to use data as it presents accuracy in prediction and rationale in causality. With a big dataset, the data-driven approach should be primarily used. Other researchers suggest that cultivating new methodologies to address unobserved heterogeneity and endogeneity is beneficial for understanding accident determinants [53]. When selecting different methods, in addition to consideration for the dataset, implicit assumptions also need to be made based on the likelihood or severity of the accident [54]. This helps to embody different aspects of the accident mechanism and make more accurate safety decisions. The following is a detailed analysis of the choice of methods in this study.

Statistical models are designed to capture the relationship between independent and dependent variables as accurately as possible. The ordinary least squares (OLSs) method in its simplified form demonstrates a linear relationship, wherein the error term satisfies a normal/Gaussian distribution and satisfies homoscedasticity. Although the errors do not meet the condition of being normally distributed and homoscedastic, the generalized least squares (GLSs) method can use its link function to convert a number of target variables that satisfy a particular distribution condition into a linear model, thereby eliminating heteroscedasticity in linear relations. Weighted least squares (WLSs) can also be used to convert the model to a linear format by weighting the explanatory variables to eliminate their heteroscedasticity.

Despite their potential usefulness, Mannering and Bhat [53] note that simple linear regressions such as OLS, GLS, and WLS are seldom used as a method in accident research. Linear regression methods, in their varied forms, are only applicable to fit hyperplane datasets without using other factors as weights. Traffic accidents are the outcome of interweaving multiple influencing factors, which largely rely on the construction and solution of complex nonlinear problems [53, 55]. Linear regression models cannot adapt to capture complex patterns, and it is impractical to add interaction terms or use polynomials. Therefore, global regression based on linear models alone is not sufficient for this type of analysis. To address this issue, the GWR method could be used to build spatial models, which use locally weighted regression to enhance the accuracy of the results. It calculates weights by constructing spatial kernel functions and then uses local regression to intuitively reflect the nonstationary characteristics of geographic relationships [56]. However, in reality, the accurate modeling of complex geographic relationships requires increasingly nonstationary solution accuracy and computing power. If GWR is used, the model needs further improvements in proximity analysis, calculation of kernel weights, and optimization of bandwidth parameters, among other areas [38, 57].

In addition, various machine learning methods and spatial regression models have been increasingly used in traffic accident research due to their capacity for superfitting to nonlinear problems. Among them, support vector machine (SVM) methods use kernel functions for nonlinear classification [58–60]; hierarchical clustering algorithms divide traffic impacts into layers based on data distribution [61]; *K*-means clustering algorithms and GWR both perform cluster analysis based on the collection distance of sample points [62]; and deep learning is often applied to general graph models or hypergraph models without massive constraints [62], such as image recognition of traffic accidents in social media and black spot recognition in urban traffic safety [63–65]. However, the earlier studies present a lack of accuracy due to the errors and unobserved variances.

As this study looks at the impact of each land use type on traffic accidents and its pattern at different spatial levels, regression- and tree-based models are selected to address the complexity of issues and factors involved in accidents. The latter involves drawing multiple trees from top to bottom through multiple terminal nodes to visually represent the detailed effects of each factor in the model in a nested manner [66]. In one of these tree-based regression methods, boosting first builds multiple decision trees by an orderly sampling of the initial training set and then combines the multiple trees to slowly train the prediction model to improve the prediction performance [66]. The gradient boosting method (GBM) is used to implement this boosting technique.

This research identifies GBM as a better method over traditional methods such as generalized linear functions of all kinds since it can use different steps and a few critical parameters to help explain the loss function in the model. This loss function is the same as the rule of finding error patterns in the linear function to help describe the model more accurately. Therefore, when the interpretation of the model in some traditional methods is not accurate enough, GBM can learn nonlinear relationships to achieve better accuracy. GBM is also very receptive to outliers and is not sensitive to noisy data; it works to account for missing data while efficiently calculating. Additionally, the bagging algorithm, which also belongs to the tree-based algorithm, shares similarities with the characteristics of GBM. However, bagging uses a self-service sampling method (sampling method with replacement; duplicate samples may be taken) in building a decision-tree, which is less efficient than GBM. It is more suitable for data with fewer dimensions and higher accuracy requirements [62]. The land-type data obtained in this research is complex in its distribution, and the sizes of various types of land use vary greatly and are mixed with each other. Therefore, GBM would have a higher accuracy when sampling and is thus selected as the method for use in this research.

2.2. GBM Model. As previously mentioned, the gradient boosting model (GBM) is selected for the machine learning method used in this paper. Boosting algorithms are a commonly used machine learning method that can be applied to classify regression problems. GBM uses boosting to distinguish the strong from among weak classifiers and obtains the new model by training in the direction of gradient descent of the previously modeled loss function. Generally, an important criterion for evaluating the performance of a model is a loss function. The loss function essentially refers to the degree of the model's unreliability. As the loss function decreases, the model becomes more reliable and predictable. The best way to improve the model performance is to make the loss function decline in the direction of the gradient. Given the increasing difficulty in loss function optimization in previous machine learning models, Friedman [67] proposed the following gradient boosting algorithm. It acts as a greedy function approximation method designed to obtain the next model by training in the gradient descent direction of the current model loss function. The following is its mathematical derivation.

Firstly, the model is initially set up with ε as the coefficient and h as the assumed classification rule of the overall function $F(x)$:

$$F(x) = \sum_{n=1}^{N} \varepsilon_n h_n(x), \tag{1}$$

where $X = \{x_1, x_2, ..., x_n\}$ represents the independent variables in the input space and Y represents the response variable in the output space. Given a training dataset [50]$_1{}^N$, the purpose of which is to find a hypothesis function $F*\ (x)$ that maps the x function to y, and the difference between this hypothetical function and the real function can be represented by a loss function. The loss function $\Psi\ (y, F(x))$ is a nonnegative real-valued function of $F*\ (x)$ and Y, with the ultimate goal of minimizing the loss function:

$$F^*(x) = \arg_{F(x)} \min E_{y,x} \Psi\ (y, F(x)). \tag{2}$$

Then, in combination with equation (1), approach $F*\ (x)$ by a linear expansion in equation (2):

$$F(x) = \sum_{m=1}^{M} \beta_m h(x; a_m), \tag{3}$$

where the function $h(x; a_m)$ is a simple classifier with x and $a = \{a_1, a_2, ...\}$ is the parameter in the classifier function. However, the expansion coefficient $\{\rho_m\}M\ 0$ and the classifier parameter $\{a_m\}M\ 0$ are mainly obtained in the training data using the segment-by-segment training algorithm. The initial hypothetical function $F_0\ (x)$ is given first, and then, $m = 1, 2, ..., M$ iterates stepwise as in the following equations:

$$(\beta_m, a_m) = \arg\min_{\beta,a} \sum_{i=1}^{N} \Psi\ (y_i, F_{m-1}(x_i) + \beta h(x_i; a_m)), \tag{4}$$

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m). \tag{5}$$

Gradient boosting uses a two-step strategy to solve the loss function $\Psi(y, F(x))$ in equation (4). The first step is to put the function $h(x; a)$ into the least squares as (6) and get the current pseudoresidual:

$$a_m = \arg\min \sum_{i=1}^{N} [\tilde{y}_{im} - \rho h(x_i; a)]^2. \tag{6}$$

In the second step, given $h(x; a)$, the optimal value of the coefficient β_m is determined by the following formula:

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} \Psi\ (y_i, F_{m-1}(x_i) + \beta h(x_i; a_m)). \tag{7}$$

This strategy first replaces the difficult optimization problem by the least square method of equation (6), then optimizing loss function Ψ based on a simple parameter in equation (7). The gradient boosting model has achieved rapid development in recent years. Zhao et al. [26] reported a stochastic gradient decision tree based on GBM and constructed a decision tree model with two methods. Elsewhere, an extended end-to-end promotion tree system named XGBoost (extreme gradient boosting) model was proposed by Tianqi Chen in 2016 and has widely been used in image classification and loss estimation since then [68, 69].

2.3. Relative Importance of Factors. When predicting the coefficients of the independent variables in the model, it is difficult to rank the coefficients of the independent variables in the model. Moreover, multicollinearity frequently causes interactions between variables in the model, and autocorrelation tends to cause errors. This paper conducts relative weight analysis to solve these problems by sorting the importance of the fit of the model according to each independent variable. It also helps to clarify the multicollinearity between variables [70]. The symbol R_i (where $i = 1, 2, ..., n$) refers to the reliability set of the influencing factors of the entire traffic accident. $R\ (R_1, R_2, \ldots, R_n)$ is the first polynomial of the i-th influencing factor and its reliability. If taking the partial derivative with respect to R_i (where $i = 1, 2, \ldots, n$), the following equation is obtained:

$$I_i = \frac{\partial R(R_1, R_2, \ldots, R_n)}{\partial R_i}, \quad (\text{among them, } i = 1, 2, \ldots, n). \tag{8}$$

Then, I_i in equation (8) is the difference between the reliability of the entire set of influencing factors obtained by taking the maximum value 1 and the minimum value 0 in $R\ (R_1, R_2, \ldots, R_n)$ except for the influencing factor I, ceteris paribus. I_i is the maximum degree of influence of i (where $i = 1, 2, \ldots, n$) on the reliability of the set. With an attempt to compare the relative importance of each factor in the factor set, it is assumed that the reliability of each factor in I_i (where $i = 1, 2, \ldots, n$) is r, thus the weighting expression of the relative importance of each factor in the reliability of the factor set is as follows:

$$I_i^W(r) = \frac{I_i(r)}{\sum_{i=1}^{n} I_i(r)}, \quad (\text{among them, } i = 1, 2, \ldots, n). \tag{9}$$

Equation (9) gives the relative importance of each influencing factor in the reliability of the factor set under different reliability conditions. This method of measuring is applied and discussed by many scholars [71, 72]. They also point out its controversies like large instability, inability to respond to positive or negative correlations, and unclear quantifiers.

2.4. Partial Dependence. Partial dependence changes the value of the target feature while controlling other fixed variables and how the fitting result of the observation model changes. The idea is the marginal effect of variables on the predictions of machine learning models [67]. The estimation method of the partial function is

$$\widehat{f}_{xs}(xs) = \frac{1}{n} \sum_{i=1}^{n} \widehat{f}\left(xs, x_C^{(i)}\right), \tag{10}$$

where xs is the feature drawn in the partial dependence graph, while x_c is the actual eigenvalue of the feature other than the selected variable. These two types of features together constitute the feature space x. The assumption of partial dependence is that feature C is not related to other features in dataset S. n refers to the number of instances.

This function represents the effect of the selected explanatory variable and can be used to explain the "black box" model of GBM [73, 74]. Partial dependence resolves the issue that the importance of this indicator cannot reflect the positive and negative relationship to a certain extent. In general, the direction of the partial dependence plot reflects the directions of correlation between variables and outcome, whether it is positive or negative. Compared with the earlier GBM models, which could only plot the importance of variables in a ranked bar chart, the newer GBM model has added the function of partial dependence plot. The advantage of this method is that it is intuitive, easy to operate, and can explain causality; however, it can also be interpreted as impractical to show a complete distribution of features at times and assume that the calculated variables are independent of other variables [62].

3. Data and Model Building

3.1. Study Area and Data Source. This study selected Suzhou Industrial Park (SIP) as a study area. Established in 1992 and located in the eastern part of Suzhou City, SIP is adjacent to Kunshan City and contains both Jinji Lake and Dushu Lake. It is located in the east of the Taihu Plain in the Yangtze River Delta. The administrative region covers an area of 278 square kilometers, with a registered population of 0.576 million [75]. SIP has a multilayered transportation system with a dense network of highways, national and provincial highways, railways, waterways, and other transportation networks. In particular, the SIP transportation system is unique, as it traverses numerous waterways, including lakes and rivers. Suzhou Industrial Park is an important cooperation project between the governments of China and Singapore. It draws on the successful experience of advanced countries and regions in its development and management. In the functional land of the industrial park, the central business district is developed around Jinji Lake as the center of the SIP. Within the 80 square-kilometer boundaries of the "China-Singapore Cooperation Zone," four major functions such as business, science and technology innovation, tourism and vacation, high-end manufacturing, and international trade are included. The land use comprehensively covers various types of land use and is relatively conducive to the coordination with transportation compared to other cities in China.

This research uses the SIP's traffic accident data in 2016, with a record of 58,315 traffic accidents, which has been obtained from the SIP traffic police bureau. The spatial join function in ArcMap 10.5 was used to calculate the accident points in the TAZ unit to obtain the accident frequency distribution map (see Figure 2). The accidents are categorized into six levels based on the frequency of occurrence, and the six levels are separated by the color gradient. The map shows the following: (1) the degree of aggregation of the same level is low; (2) each level including the normal peak area of the accident presents a relatively discrete distribution; and (3) accidental peak areas are occurring in the dense TAZ areas, which are distributed in the north and southwest, respectively. In the two areas, the number of accidents in the adjacent TAZs differed significantly.

The following data of SIP are used in this study: traffic accident data, different types of land use data (e.g., residential land and educational land), points of interest (POI) (e.g., shopping and leisure places and financial outlets), and road facility data (e.g., traffic lights and intersections). Since the accident data obtained occurred in 2016, the data for land uses and road facilities are also selected for the same period in order to maintain consistency of the study and to explore the causes of accidents more accurately. Therefore, the study period is fixed at 2016.

3.2. Two Scales of the Analysis Unit. The analysis in this research involves two spatial units: the local level based on TAZ data and the nearest neighborhood level based on Thiessen polygons (see Figure 3).

Regarding the local level scales, TAZs within the SIP are selected as a spatial unit of analysis, and the number or density of various land use types within the TAZ regions is calculated. The TAZs with smaller areas are deleted because they create outliers and distort the data distribution in the independent variable. In terms of the nearest neighborhood

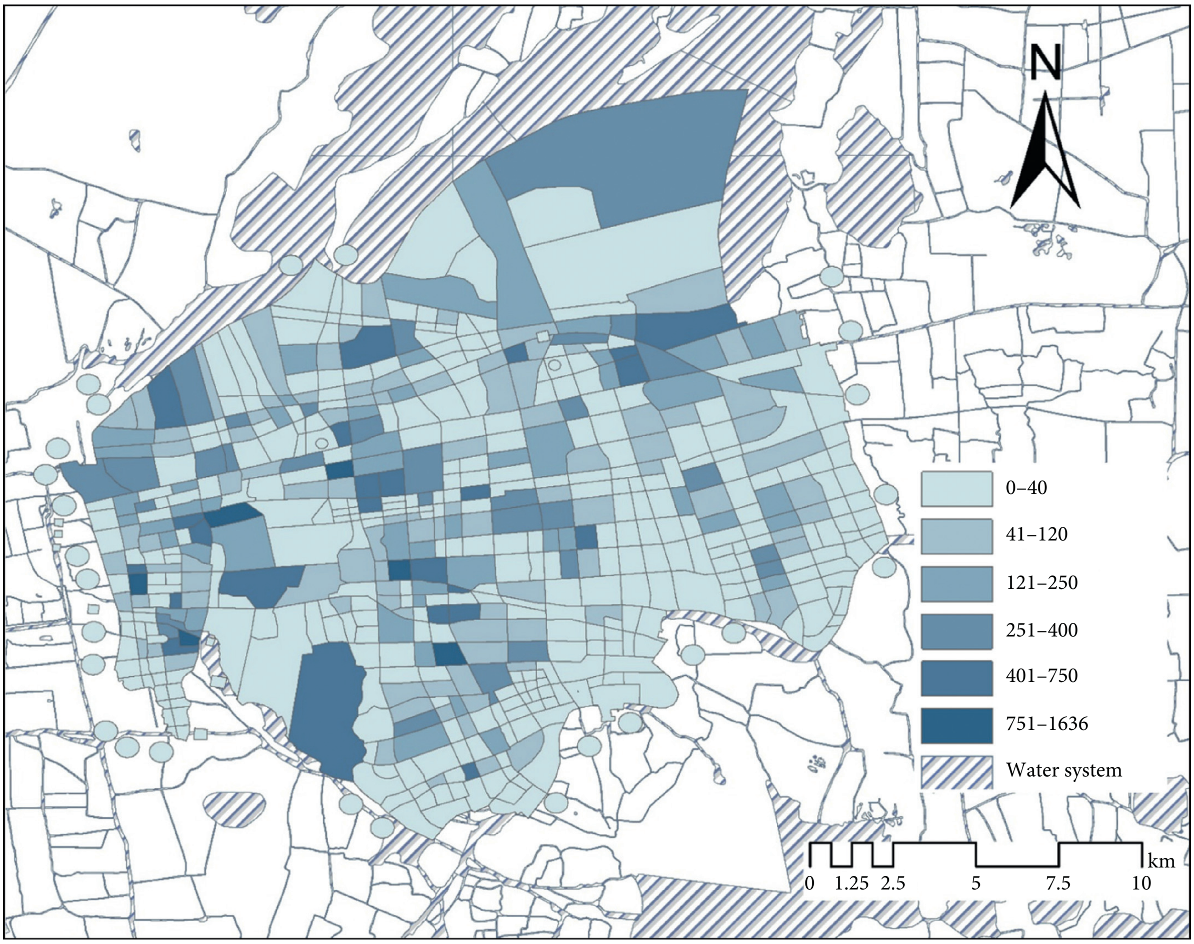


Figure 2: Accident frequency distribution map in Suzhou Industrial Park in 2016.

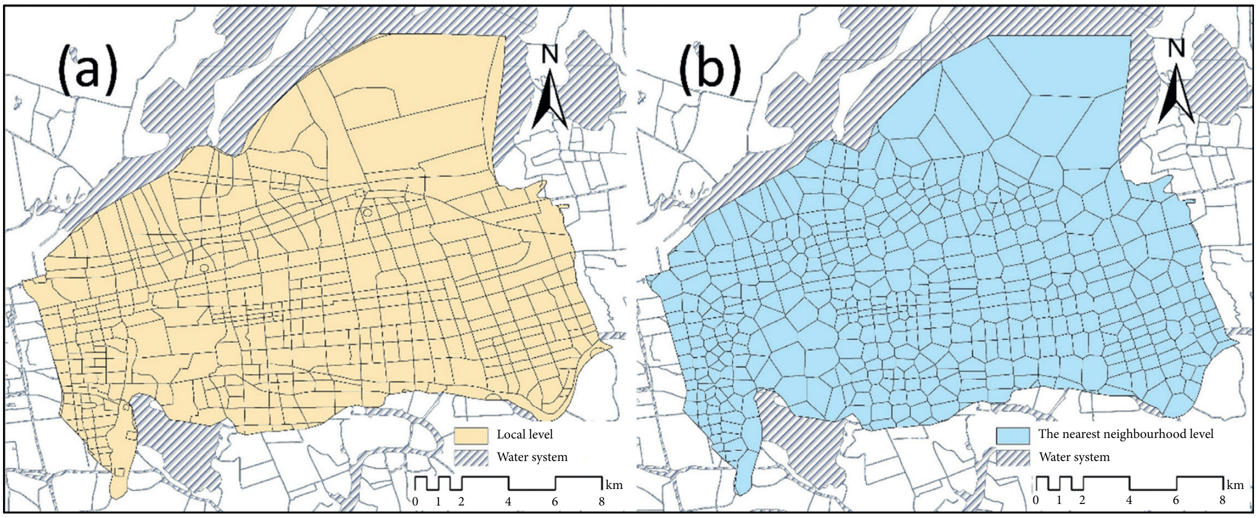


Figure 3: Two maps of spatial units at two-level scales: (a) local level TAZ area; (b) the nearest neighborhood level-Thiessen polygons.

scale, Thiessen polygons were created to avoid the potential analytical errors caused by overlapping areas that often occur in buffers. The TAZ's centroid is used as the central point to create Thiessen polygons within the SIP area. These two scales have similarities concerning general spatial location as they share the same center points and the geographic structure of the analysis unit. However, while TAZ has been used extensively in transportation-related analysis, the neighborhood level was rarely tested because it is not directly tied with regional spatial systems and transportation networks. Therefore, in this research, the nearest neighborhood scale with Thiessen polygons as an analysis boundary is worth exploring. Table 1 shows the basic descriptive statistics of the spatial units at both scales.

3.3. Variables. The variables used in this research are structured and explained in Table 2. The number of traffic accidents occurring at each of the two geographical levels comprises the dependent variables. These variables in 12 categories include transportation facilities such as "Intersection," "Trafficlight," and "Busstation"; residential living facilities such as "EduInstitu," "Financial," "Healthcare," and "Government"; and land use mix "*D*1" and "*D*2." With particular reference to the work of Yue et al. [15], this research uses neighborhood vibrancy to measure the degree of land mix, using Hill numbers to refer to the multidimensional POI mixed use. *D*1 and *D*2 calculate the exponential of the Shannon entropy and Simpson index to measure the diversity of residential, office, and commercial sites, respectively. Moreover, in order to control the density level of the POI points of the land type, the classification and regression tree (CART) method is used to divide individual land-type variable into segments representing different density levels and convert them into dummy variables. After integrating the data from both scales, all density dummy variables are binary for high density and low density. Low-density variables are used as references. The real variables representing the land type and the dummy variables controlling the density levels are included in the model of this study.

3.4. Model Building. The GBM model needs to set several parameters, including distribution, n.trees, interaction.depth, weights, n.minobsinnode, shrinkage, train.-fraction, cv.folds, keep.data, class.stratify.cv, and n.cores, to name a few. Some of these variables are set selectively, such as weights, n.minobsinnode, and some use the default value rather than setting purposely like n.trees (default is 100), interaction.depth (default is 1), and bag.fraction (default is 0.5). The most important and most frequently trained parameters are shrinkage, N.trees, and cv.folds. Reducing the rate requires more iterations, and it takes longer for larger data [67]. The empirical results have shown that shrinkage coefficients with smaller values ($v \leq 0.1$) exhibit better generalization errors. The n.trees is generally used with parameter of shrinkage. Lowering the shrinkage and adding more trees can improve the generalization ability of the model and avoid overfitting [76]. Cv.folds is the judgment method of the model. N-cv.folds go through a total of *n* experiments and obtain the accuracy index of the measurement algorithm after each test, which is used as an indicator to judge the merits of the algorithm.

Prior to adjusting parameters, important parameters are set to default values because initial default values help determine other parameters. The steps of the adjustment are as follows. Firstly, based on the accident distribution data, the model of Poisson distribution is confirmed. Subsequently, the learning rate is taken to be the original default number of 0.1. There are some evidences to suggest that a shrinkage rate of 0.001 will bring relatively low deviation [12, 77], and this study sets the parameter for shrinkage rate to 0.001. The CV method is used to detect different parameters with Poisson deviance as the representative of the loss. The CV number is set as 10, according to the characteristics of the data in this case. Interaction.depth indicates the integer of the maximum depth of each tree [67]. This parameter and n.minobsinnode are the trade-offs that together determine the performance of the model. If the two parameters are too large, it will easily lead to overfitting, but it leads to underfitting when they are too small. In the case where several other parameters are fixed, the lowest Poisson deviance is adopted to decide the value of interaction.depth is 15.

Table 1: Comparison of spatial units at two scales.

	Number of spatial unit	Max of area (km^2)	Min of area (km^2)	Average of area (km^2)	Std. dev.
Local level	471	15.74054	0.00037	0.50185	0.95272
Nearest neighborhood level	480	11.10298	0.00352	0.50780	0.75309

In summary, in order to verify the performance validity and stability of the predictive model, following a series of adjustments and experimental comparisons with other similarity indicators, the final model parameters are as follows in Table 3.

4. Results and Discussion

4.1. Relative Importance of Explanatory Variables. Relative importance is the role of the indicated feature in predicting the target response and can be used to visually quantify the contribution of each explanatory variable to the model. It is determined by the frequency of the features used in the segmentation points of the tree. The higher the frequency of use is, the higher the importance of the variable is [67]. The response of the eigenvalues or independent variables at the two scales is predicted according to the selected model parameters. Figure 4 illustrates that the relationship between Poisson deviance and iteration could be used to estimate the effect of the model parameters: both test error and train error decrease when the iteration increases. The model does not appear to display the problem of overfitting. If the data are underfit, signifying that the model learning ability is insufficient, it is also necessary to judge depending on whether or not the deviation of the abnormal value occurs. Figure 5 lists the contribution of all variables and their ranking. Under the existing parameters of the model, all variables have a nonzero contribution, and the tailing is longer. This means that, under the identified model and existing data, all land uses and POIs impact the distribution of the final accident frequency. This lateral also proves that the parameters of the model are valid.

A total value of 100 is allotted to each variable in both models. The relative influence plots (see Figure 4) show that

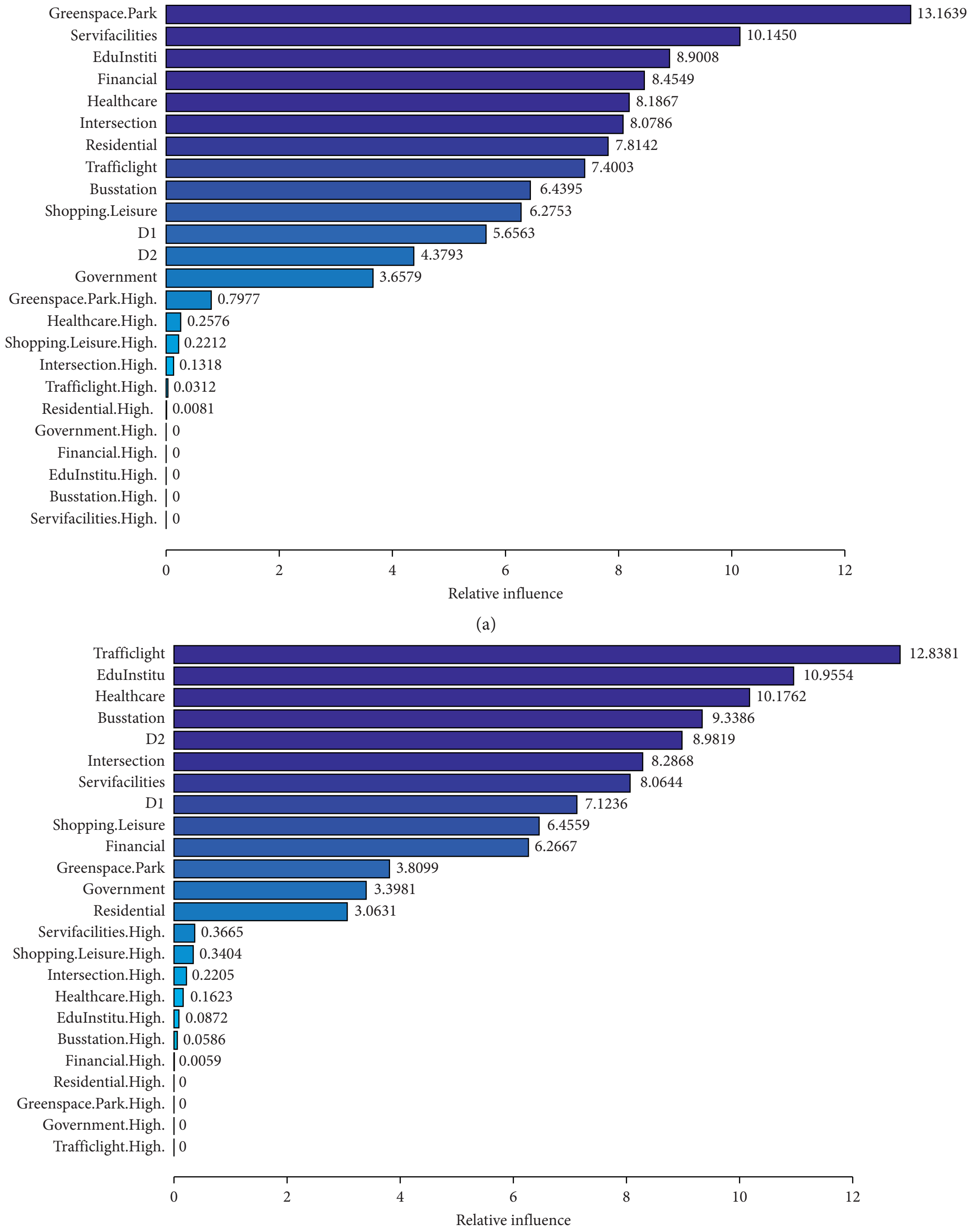
Table 2: List of variables and calculation.

Category	Variable	Description
Accident	Accident	Number of accidents within the study unit
Residential land density	Residential Residential (high) Residential (low) (ref)	Number of residences/unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of residential land use, where "Residential (low)" is set as a reference variable
Healthcare institutions density	Healthcare Healthcare (high) Healthcare (low) (ref)	Number of healthcare institutions/unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of land use by healthcare institutions, where "Healthcare (low)" is set as a reference variable
Greenspace and park density	Greenspace&Park Greenspace&Park (high) Greenspace&Park (low) (ref)	Total number of greenspace and parks /unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of land use by greenspaces and parks, where "Greenspace&Park (low)" is set as a reference variable
Governmental agencies density	Government Government (high) Government (low) (ref)	Number of governmental agencies/unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of land use by governmental agencies, where "Government (low)" is set as a reference variable
Financial services density	Financial Financial (high) Financial (low) (ref)	Number of financial services/unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of land use by financial services, where "Financial (low)" is set as a reference variable
Education institution density	EduInstitu EduInstitu (high) EduInstitu (low) (ref)	Number of education institutions/unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of land use by education institutions, where "EduInstitu (low)" is set as a reference variable
Intersection density	Intersection Intersection (high) Intersection (low) (ref)	Number of intersections/unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of the land use by intersections, where "Intersection (low)" is set as a reference variable
Traffic light density	Trafficlight Trafficlight (high) Trafficlight (low) (ref)	Number of traffic lights /unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of land use by traffic lights, where "Trafficlight (low)" is set as a reference variable
Bus station density	Busstation Busstation (high) Busstation (low) (ref)	Number of bus stations/unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented by low density and high density of land use by bus stations, where "Busstation (low)" is set as a reference variable
Service facilities density	Servifacilities Servifacilities (high) Servifacilities (low) (ref)	Number of service facilities/unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of land use by service facilities, where "Servifacilities (low)" is set as a reference variable
Shopping and leisure density	Shopping&Leisure Shopping&Leisure (high) Shopping&Leisure (low) (ref)	Total number of shopping stores and leisure outlets /unit area ($1/\mathrm{km}^2$) Two dummy variables used to measure density are, respectively, represented as low density and high density of land use by shopping and leisure places, where "Shopping&Leisure (low)" is set as a reference variable.
Land use mix	$D1$	Land use mix measurement [15]
	$D2$	Land use mix measurement [15]

Table 3: Parameter setting in model building.

Distribution	Poisson
n.trees	100 (default)
interaction.depth	15
Shrinkage	0.001
cv.folds	10

the real variables representing different land uses and land mixes are ranked higher, while the dummy variables controlling density levels, which do not exceed 1%, are ranked lower. This is the case for both models, which confirms that a high density of land use has little effect on the overall results of the model. The density dummy variables help to reconcile the completeness of the model and the persistence of the



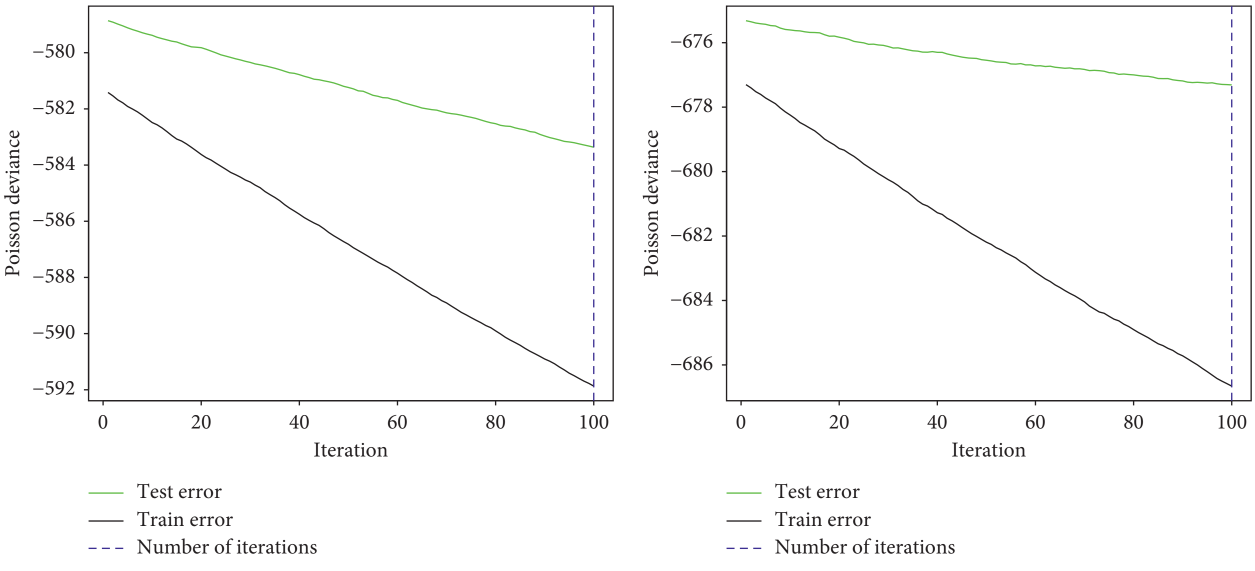
(a)

(b)

Figure 4: (a) Relative influence at local level. (b) Relative influence at the nearest neighborhood level.

variables, but since their importance values are too low relative to the real variables, they are not analyzed more extensively here.

The two-scale models (see Figure 4) have similarities and differences in the order of relative importance. With the relative contribution of 13.16% and 12.84%, "Greenspace.Park" and "Trafficlight" represent the most crucial variables in two types of geographical scales. As is demonstrated in other studies [6, 7], the factors most influential to the accident are road facilities such as traffic lights, road width, and distance to intersections. However, greenspace is less frequently identified as a decisive cause of traffic accidents. Subsequent "Servifacilities" and "EduInstitu" are also ranked as the second most important variables of the two models at 10.14% and 10.96%, respectively. The two models turn a very vital commonality up: the three



(a)

(b)

Figure 5: (a) Relative influence at local level. (b) Relative influence at the nearest neighborhood level.

variables of "Intersection," "Shopping.Leisure," and "Government" have a very steady order of importance in the two models. This is a feasible locus as the data sources of the two models are the same. This sort of alignment is consistent with some initial results [78, 79] that several types of land use are ranked higher in accident studies. The "Greenspace&Park" variable that appears afterward on the nearest neighborhood level model is considered an aberration since it contributes 13.16% in the local level model, which sorted the first but ranks last in the neighborhood level model with only 3.81%. The gap between the two is broad. Also, identified as aberrations are the "Residential" and "*D*2" variables. These two variables are ranked very differently at either scale. The contribution disparity is about 16%. "Greenspace&Park," "Residential," and "*D*2" are the three variables identified as aberrations because they are different in the rank order between the two scales, when compared to the results of other studies that also analyze their relevant importance (Ding et al. [12] and Saha et al. [79]), revealing that "*D*1" and "Greenspace&Park" are in the middle and rear positions, respectively, as the results in the nearest neighbor level in this study. However, the importance of "residential" variable is ranked lower than "*D*1" and "Greenspace&Park." This suggests that the model fits more accurately at the nearest neighbor level for the nonpoint land types. The variables "Servifacilities," "EduInstitu," "Financial," and "Healthcare" are very close in value although disparate in ranking at two scales. Their coefficients differ by 2% in both models.

In sum, the degree of explanation of the continuous variables obtained by calculating the density is higher than that of the dummy variables. Besides, under different models, the order of importance of factors is not exactly the same, and three aberrations appear. This shows that the distinction in the locational levels allows the model to establish utterly different utility functions during the construction and fitting phases. However, there are also some reasons that the autocorrelation leads to aberrations, and the parameters are underfitting. Some of the nonpoint land types in this study, such as "D1" and "Greenspace&Park" variables, ranked consistently with other studies. The degree of interpretation of the two geographical scale models is slightly inconsistent. "Servifacilities," "EduInstitu," "Financial," and "Healthcare" contribute a higher proportion of variables in the local-scale model than their nearest neighborhood scale model, indicating that the established parameters provide a more accurate description of the local-scale model. These results suggest that the performance of the variables at different scales and in specific land-use types may be explored further. It is also important to note that the dummy variables need to be introduced with care when applying such approaches because of their low degree of interpretation.

4.2. Change in Influencing Factors of Real Variables due to Increasing n.trees. As described in the previous section, the number of trees has the same effect as the number of iterations. As a rule of thumb, the number of iterative regression trees is generally set to a larger number because the "gbm.perf" parameter in "gbm" package can estimate the suitable number of iterations for prediction after the model is trained [80]. An increase in tree complexity would help improve the prediction bias and reduce the learning rate. In this study, the parameters of n.trees are not set especially when the model is determined, and the default number of this parameter of 100 is employed. Nevertheless, when the model is tested with a large tree value, the error of each variable is large. In this circumstance, the results of the setting of a number of trees by Ding et al. [12] are referenced. In his experiment, when the tree complexity is low, the explanatory degree of the variable is low, but the explanatory degree is stable as the tree complexity increases to eight. Thus, this study tests the changes in the interpretation of n.trees from 1 to 30 and compares the two levels (see Figures 6 and 7). The following shows comparisons of all variables representing land types and land mix between variables and at different scales.

4.3. Comparison between Variables. After adjusting the parameters, the line of the explanatory degree of the variable at each point is compared with the default parameter of the model. It is discernible that as the complexity of the tree increases, the degree of interpretation of all variables rises and falls around the default value of 100 in the range of trees from 1 to 30. There is no complete detachment although no variable goes steady after any number of trees 1 to 30. Variables have different amplitudes. Using the variables' sum of squares for error (SSE) to measure the magnitude of the shock (see Table 4), it is necessary to order the SSE numbers from small to large: Government < Greenspace.Park < Shopping.Leisure < Residential < *D*2<Busstation < Intersection < Financial < Traf ficlight < *D*1<Servifacilities < EduInstitu < Healthcare.

The statistical results of the SSE (Table 4) show that the SSE of all variables is at a small value, which indicates that the degree of interpretation of all variables does not fluctuate much as the number of trees increases. The parameters of the model are then fit to the data of this study. Except for "Servifacilities" and "EduInstitu," which reached 5%, the peaks and bottoms of most variables only differ by about 3%. Therefore, the number of SSE necessarily represents the smoothness of the curve from the observation (see Figure 6). However, the size ranking of SSE also shows some irrational results, with the transit facility variables "Busstation" and "Trafficlight" located in the moderate SSE values; "Greenspace&Park," "Shopping&Leisure," "Residential," "Servifacilities," "EduInstitu," and "Healthcare," which all belong to the dense POI points, are distributed at the larger and smaller ends of the SSE. In the urban area, "Busstation" and "Trafficlight" are factors with a fixed position and an accurate number and demonstrate high stability. Even if the model parameters are changed, parallel impacts on such variables are small. However, "Greenspace&Park," "Shopping&Leisure," "Residential," "Servifacilities," "EduInstitu," and "Healthcare"—factors which are numerous, high in level, and widely distributed—are similar facilities that reflect the daily life of residents. When such factors are used to adjust the parameters involving the complexity of the tree, the final fitting effects for the model are unstable to generate two types of small and large fluctuations.

In general, the complexity of the tree can identify and detect the utility of various variables in the model to a certain extent, which helps to improve the accuracy of the model fitting. However, in this study, the fitting of the model is not accurate enough as the number of iterations of the test is low. Although the interpretation of each variable still fluctuates within a reasonable range of the variable, the results of the model do not reach a consistent state under a certain number of iterations.

4.4. Comparison of Two Geographical Levels. If the statistical methods and models are identical, the partial dependence plots show that the variables' likelihood to occur is different at two levels. Suppose the degree of explanatory variables is adopted to represent the capacity of the locational scales. In that case, the following three types of situations can be divided into three categories shown in Figure 6:

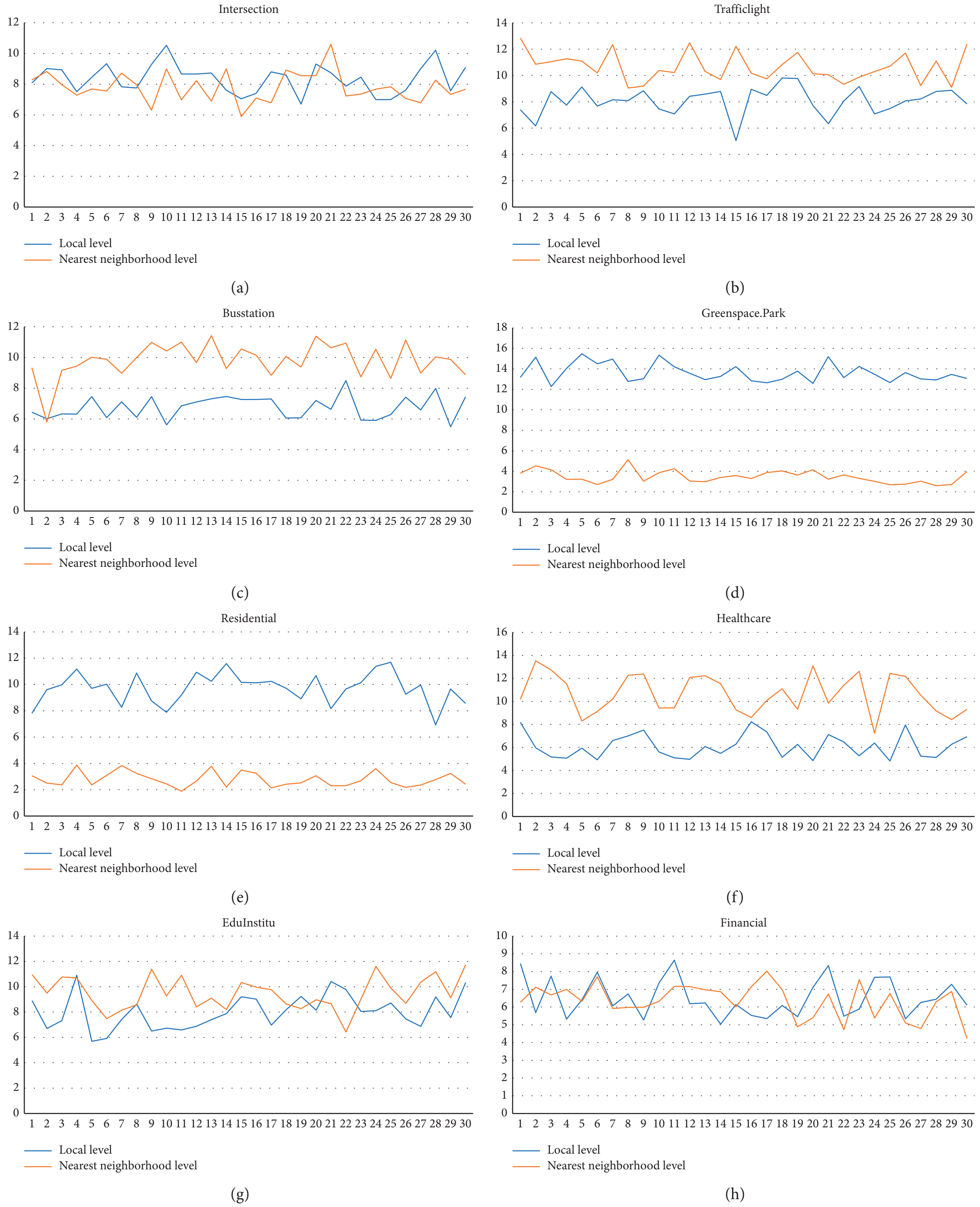
(1) For the variables of "Residential" and "Greenspace&Park," the blue curve representing the local level is above the orange curve of the neighborhood level, indicating that the capacity of the local level is high for these five variables rather than that of the nearest neighborhood level. It can be inferred that these five variables are more responsive to the TAZ region.

(2) Inversely, for "Trafficlight," "Busstation," "Healthcare," and "*D*2," the capacity of the nearest neighborhood level is higher than that of the local level. Notably, the neighborhood level is aggregated so that these four variables exceed the others.

(3) When it comes to the variables of "Shopping&Leisure," "Financial," "Government," "Intersection," "Servifacilities," "*D*1," and "EduInstitu," the two lines are repeated or interlaced, so it is hard to judge the suitability of either scale.

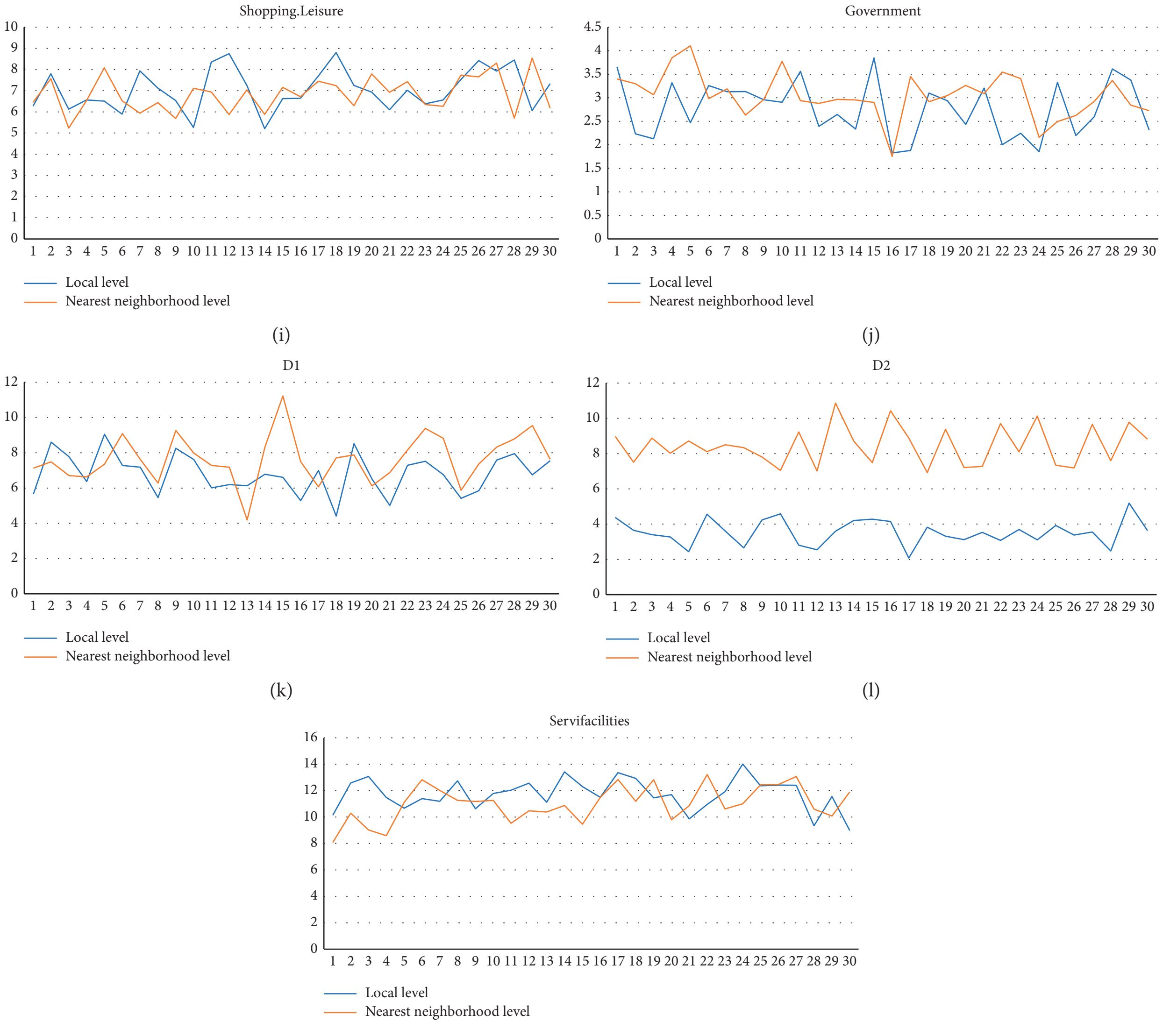
The "cale effect" in this research highlights several critical arguments. For instance, the research findings indicate a severe scale dependence on the study of model evaluation of traffic accidents. These analysis results of the local level bounded by TAZ must be interpreted with caution because there are only two variables "Residential" and "Greenspace&Park" that are available to demonstrate a higher importance at the local level scale. The other variables either show higher importance on the nearest neighborhood level or are mixed between the two scales. "Residential" and "Greenspace&Park" indicate the places where residents live and spend the most time daily. The mobility of pedestrians and vehicles is high at fixed times, such as during peak commuting hours and weekends. In contrast, among the variables with a high degree of explanation at the nearest neighborhood level, "Trafficlight" and "Busstation" are more about the infrastructure and public services invested and managed by the government. The spatially homogeneous character of the Thiessen polygon can be more evident in such sites. "Financial" and "Shopping&Leisure" have more attributes that are greatly affected by the surrounding environment. There may be dynamic changes in location. Hence, they showed a kind of crosscomplication in the two scales. The spatial heterogeneity and sensitivity in such sites will be reduced. Indeed, the accident effects of the two variables are not evidently different in both scales. Therefore, the nearest neighborhood level is more powerful than the local level due to its comprehensive description of variables as a byproduct of its larger sample sizes.

4.5. Change in Influencing Factors of Density Dummy Variables due to Increasing n.trees. The variation in the density



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 6: Continued.



(i)

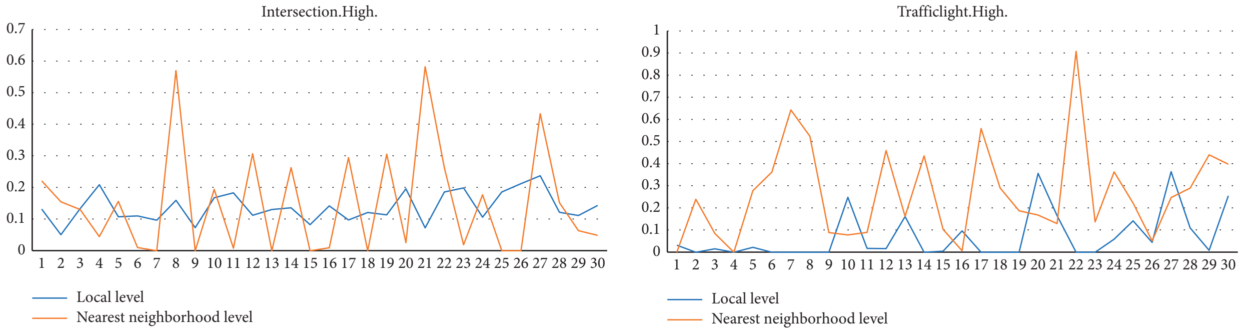
(j)

(k)

(l)

(m)

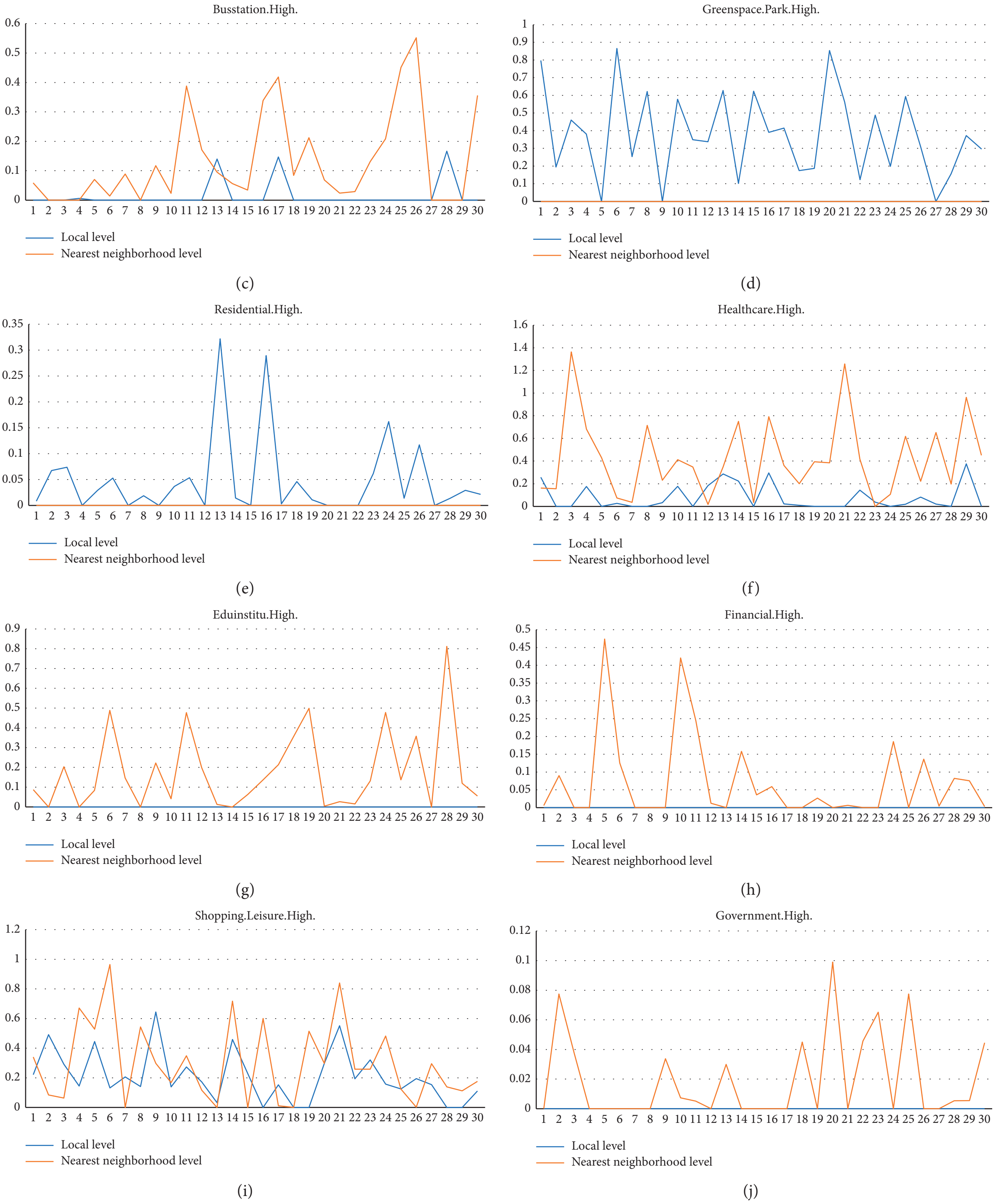
Figure 6: Variation in real variable interpretation and comparison of two scales as n.trees changing from 1 to 30.



(a)

(b)

Figure 7: Continued.



(c)

(d)

(e)

(f)

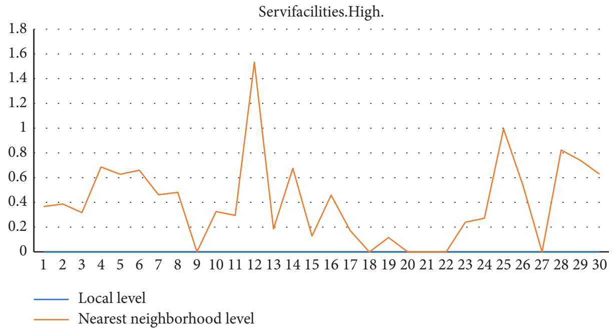
(g)

(h)

(i)

(j)

Figure 7: Continued.



(k)

Figure 7: Variation in density dummy variables and interpretation/comparison of two scales as n.trees changing from 1 to 30.

dummy variables as the number of trees increases is shown in Figure 7, wherein density dummy variables exhibit two distinct characteristics. The first characteristic involves their influencing factor lines tending to fluctuate more pointedly relative to the real variables. A possible explanation for this might be because their maximum degree of interpretation does not exceed 1%, yet their minimum value often reaches 0 in individual trees. Therefore, the visualization displays more dramatic fluctuations. The second characteristic is that the influencing factor line of some variables coincides with the *x*-axis and is constantly equal to 0. This means that the variables do not show any explanatory power in the model. "Greenspace&Park.High" and "Residential.High" show this at the nearest neighborhood scale; this also occurs in "EduInstitu.High," "Financial.High," "Greenspace&Park.High," and "Servifacilities.High" at the local scale. In this case, even though some explanatory power exists for these kinds of variables at the other scale, it is impossible to compare the results of the two scales to obtain real meaning. Relative to the real variables, the density dummy variables offer lower explanatory power consistently as the number of trees grows and varies irregularly, with limited practical utility for the model.

4.6. Partial Dependence Effect of Variables. The partial dependence plot is used below to describe the relationship between land use, road variables, and accident frequency. The partial dependence plot is a summary of the changes in all variables under the same conditions. The partial dependence plot cannot evaluate the statistical model directly, but it shows how the variation in the independent variables affects the process of model fitting [67]. Among all the variables, the partial dependence plots represent real variables of road structure facilities and land mixture (shown in Figure 8).

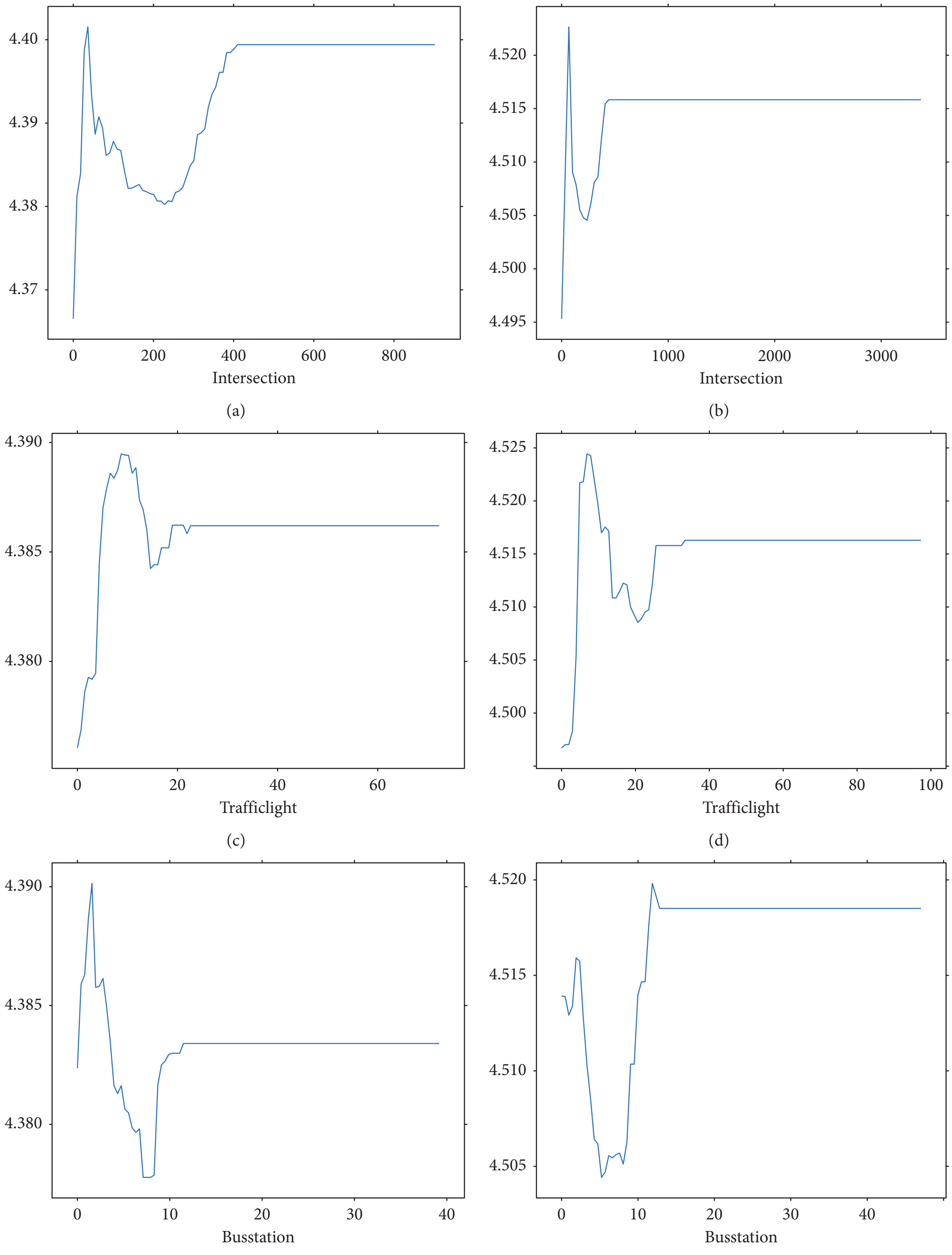
Road structure facilities in Figure 8 include three variables: "Trafficlight," "Intersection," and "Busstation." As discussed earlier, road structure facilities are usually defined as influential factors correlated with traffic accidents. Two variables fit the accident model to a stable stage when the coefficient value is small, even under different levels of scale: when there are more than 25 traffic lights per square kilometer, their influence on traffic accidents tends to be stable; similarly, the degree of interpretation does not change after over 15 bus stations per square kilometer. A particular case was observed when the intersection coefficient reaches a considerable value before the traffic accident stabilizes. The data reported here appear to support that this case occurs around the value of 400 at both levels.

However, as the density of transportation facilities increases, specific differences arise in the development trend per each level. Where traffic lights are sparse, the number of traffic accidents tends to increase sharply. The number of accidents declined briefly after 10 traffic lights per square kilometer and then settled at a value after 20 traffic lights per square kilometer, both at the local and the nearest neighborhood levels. The "Busstation" variable fluctuated several times at a density of less than 15 bus stops per square kilometer and stabilizes after that. The "Intersection" variable, on the other hand, fixed at the densities of around 400 after initially jolting upward and falling immediately after that. For both the "Busstation" and "Intersection" variables, the local and nearest neighborhood levels revealed very similar fluctuation ranges, but the fluctuation at the nearest neighborhood level was somewhat more drastic. This indicates that the nearest neighborhood scale is more suited for capturing the subtle effects of urban transport facility density on accident occurrences. A similar report on traffic accidents in Seattle shows a rising trend in both the 3-way and 4-way intersections [78], which is not evident in this study's findings. Taken together, these results suggest that the impact of intersections on traffic safety varies according to the complexity of their surrounding region. Nevertheless, for the "Trafficlight" variable, the two scales selected in this study produce similar effects of traffic lights essentially, while the result at the nearest neighborhood level better demonstrates the complex variation in bus stations and intersections.

For the two variables of mixed land use shown in Figure 8, both "*D*1" and "*D*2" reach a stability of 4.5 on the *Y*-axis after rising along the *X*-axis. However, from the curve, "*D*2" reaches stability "*D*1" earlier than relatively. This is because the *D*1 formula only measures the weight of each land use type, which is the exponential of the Shannon entropy. In the meanwhile, *D*2 considers the richness of the land use and the relative abundances of the POI [15]. The

Table 4: SSE of variables in two scales.

Government	Greenspace.Park	Shopping.Leisure	Residential	*D*2	Busstation	Intersection	Financial	Trafficlight	*D*1	Servifacilities	EduInstitu	Healthcare
16.91	34.99	47.76	48.83	50.53	52.40	53.03	59.26	66.22	90.79	93.73	101.03	112.89



(a)

(b)

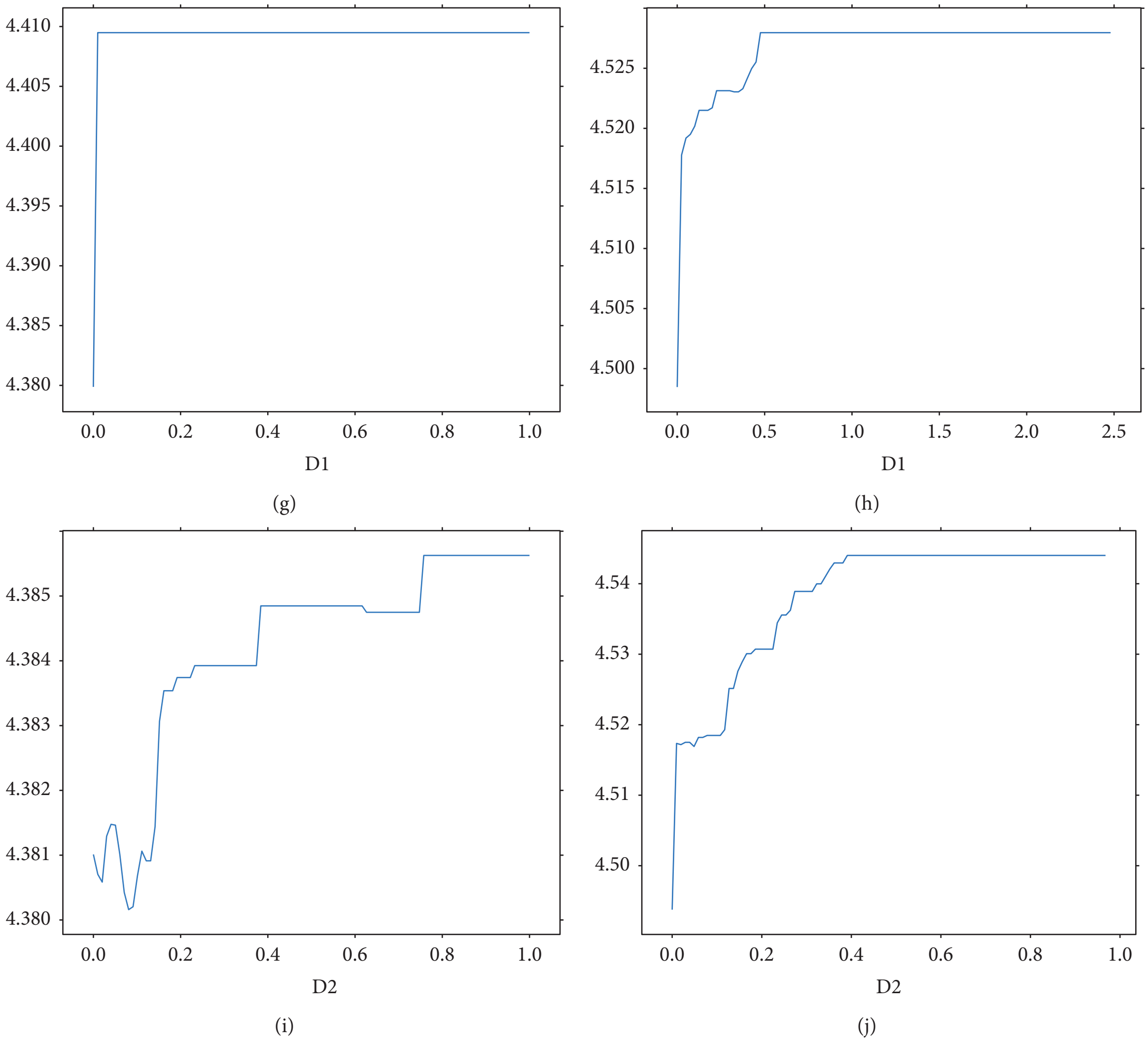
(c)

(d)

(e)

(f)

Figure 8: Continued.



(g)

(h)

(i)

(j)

Figure 8: Partial dependence plots of road structure facilities on accidents.

data used in the study confirms the diversity between three land use types containing residential, office, and commercial land; therefore, the tendency of $D2$ is more secure than that of $D1$. This result is consistent with the Chen and Zhou's work [78] on a crash frequency that shows a positive correlation between land use mix and accident. It also aligns with the increasingly stable trend of partial dependence in the Ding et al.'s study [12] on traffic accidents in Seattle.

The other variables represent 8 land uses shown in Figures 9 and 10. As a whole, all of these variables, except "Residential" and "EduInstitu," show a significant positive correlation with accidents. The partial dependence diagrams of these variables show an overall increasing and then stable trend after some fluctuations. The two variables of "Residential" and "EduInstitu" are negatively correlated at the local level although positively correlated at the nearest neighborhood scale. This suggests that several land types, "Healthcare," "Greenspace&Park," "Government," "Financial," "Servifacilities," and "Shopping&Leisure," may overall lead to an increase in the number of accidents at lower densities, while the number of accidents will not continue to increase after reaching a certain density. Thus, variables will likely function strategically at the turning point in the graphs.

In Figures 9 and 10, other than "Residential" and "EduInstitu," the six variables show consistent evolutionary trends at both the local and nearest neighborhood levels. It suggests that the partial dependency diagrams fully explain the role of each variable in the model with consistent influence across the two scales. The differences between the variables of "Healthcare," "Shopping&Leisure," and "Government" are highly uniform across the two scales. They have a slight decrease in the local level scale compared to their nearest neighborhood scale and reach stability after a range of "0–20," "0–60," and "0–30," respectively. For the "Financial" and "Greenspace&Park" variables, the nearest neighborhood scale is more revealing of their subtle changes before a turning point. It is particularly evident for the "Greenspace&Park" variable, as there is a rapid decrease in traffic accidents when the density of "Greenspace&Park" is

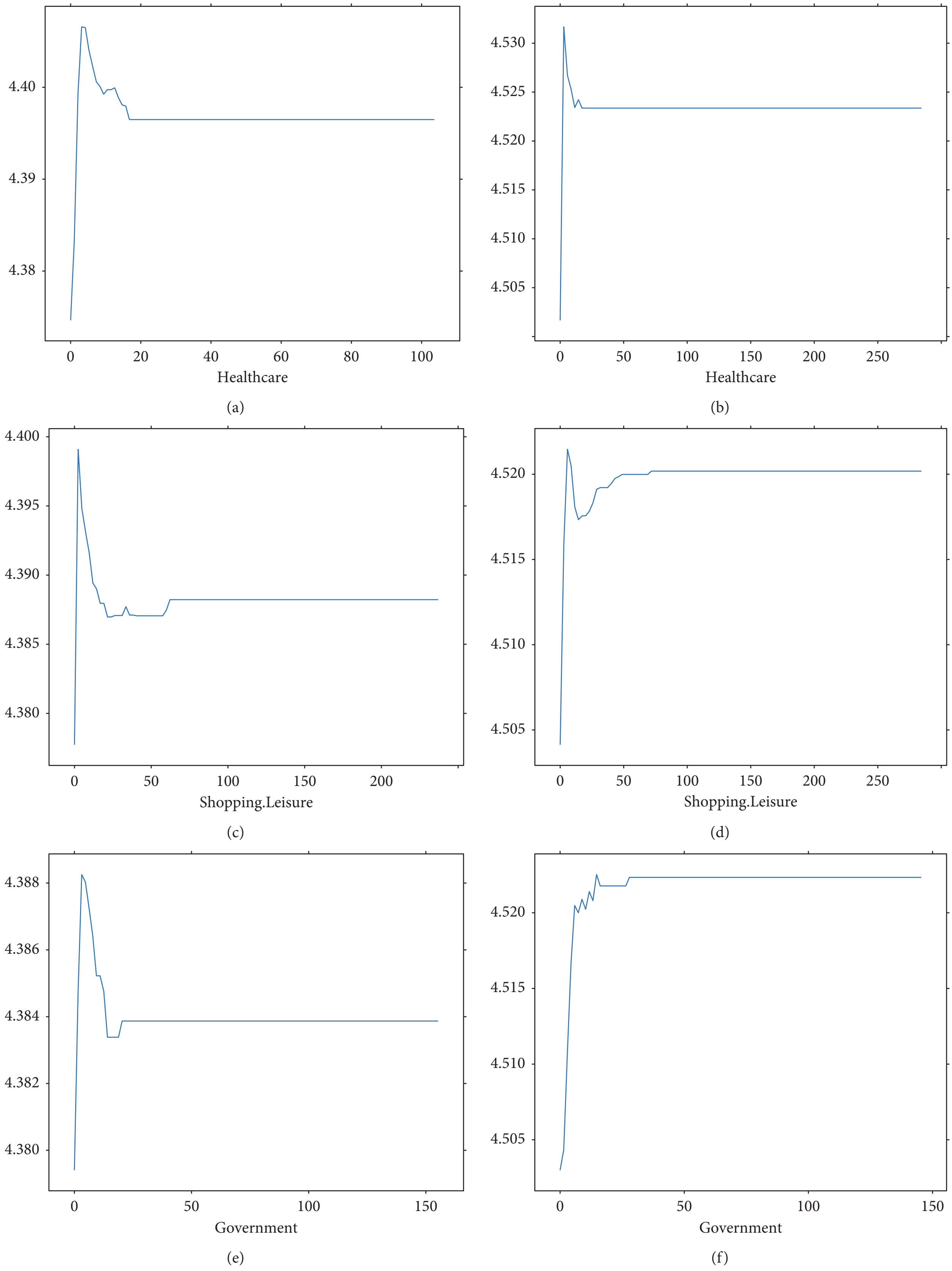
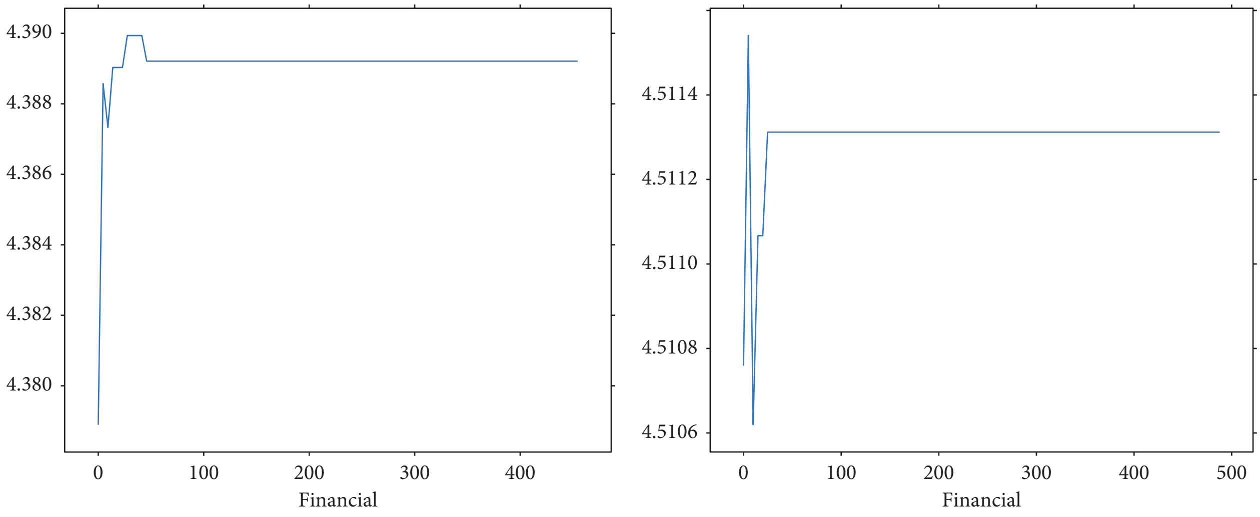


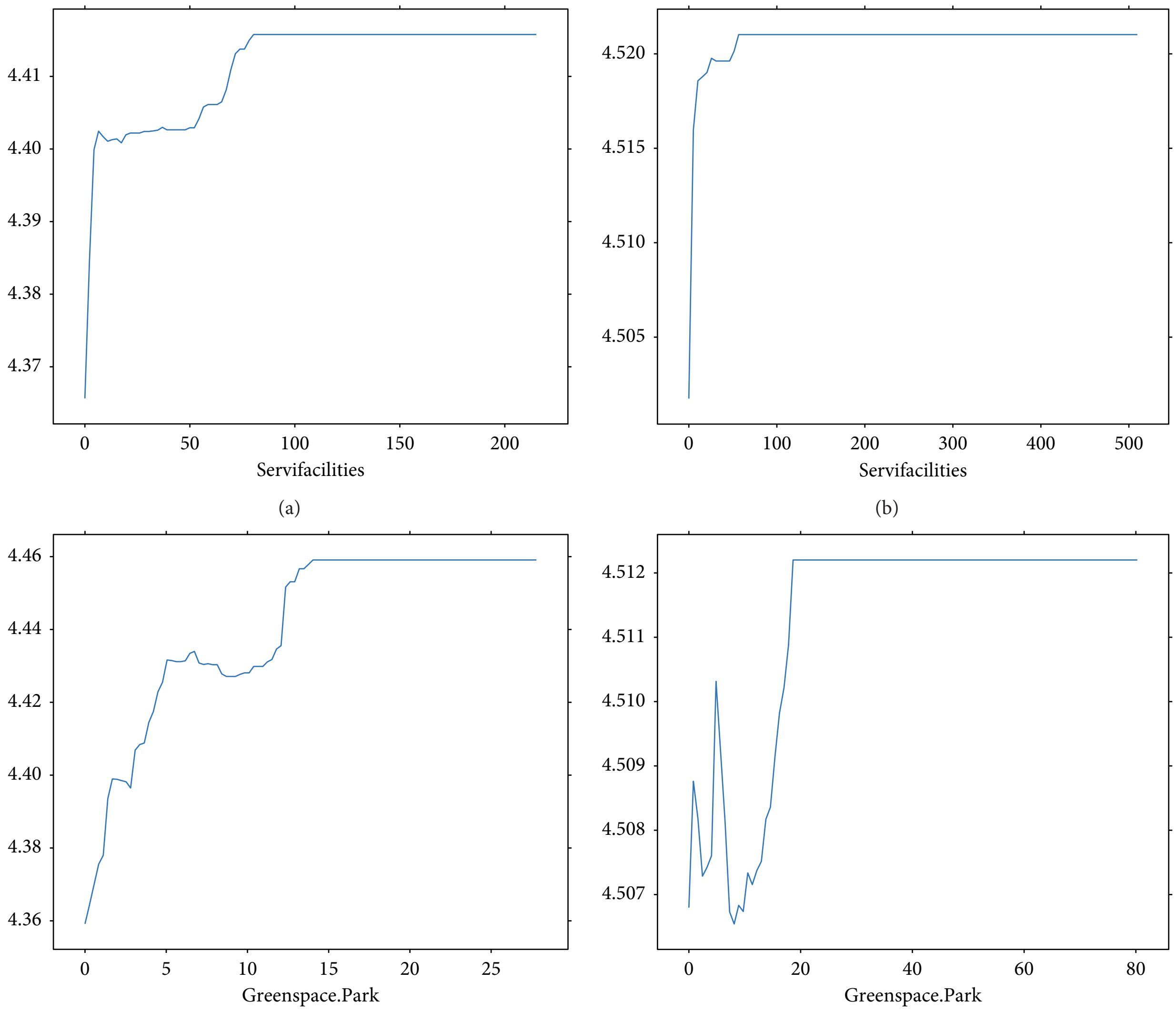
Figure 9: Continued.



(g)

(h)

Figure 9: Partial dependence plots of land use and POIs on accidents (a).



(a)

(b)

(c)

(d)

Figure 10: Continued.

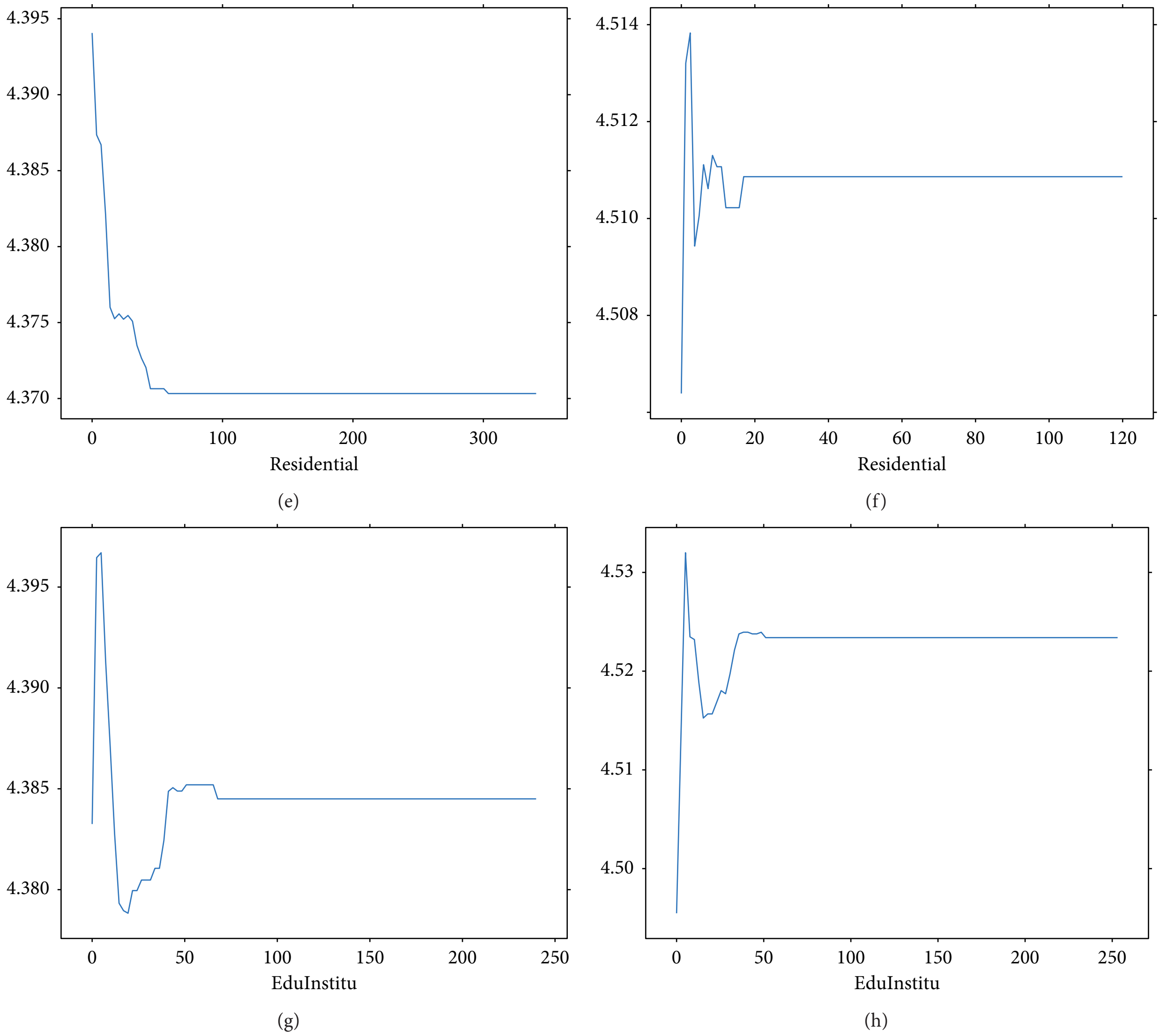
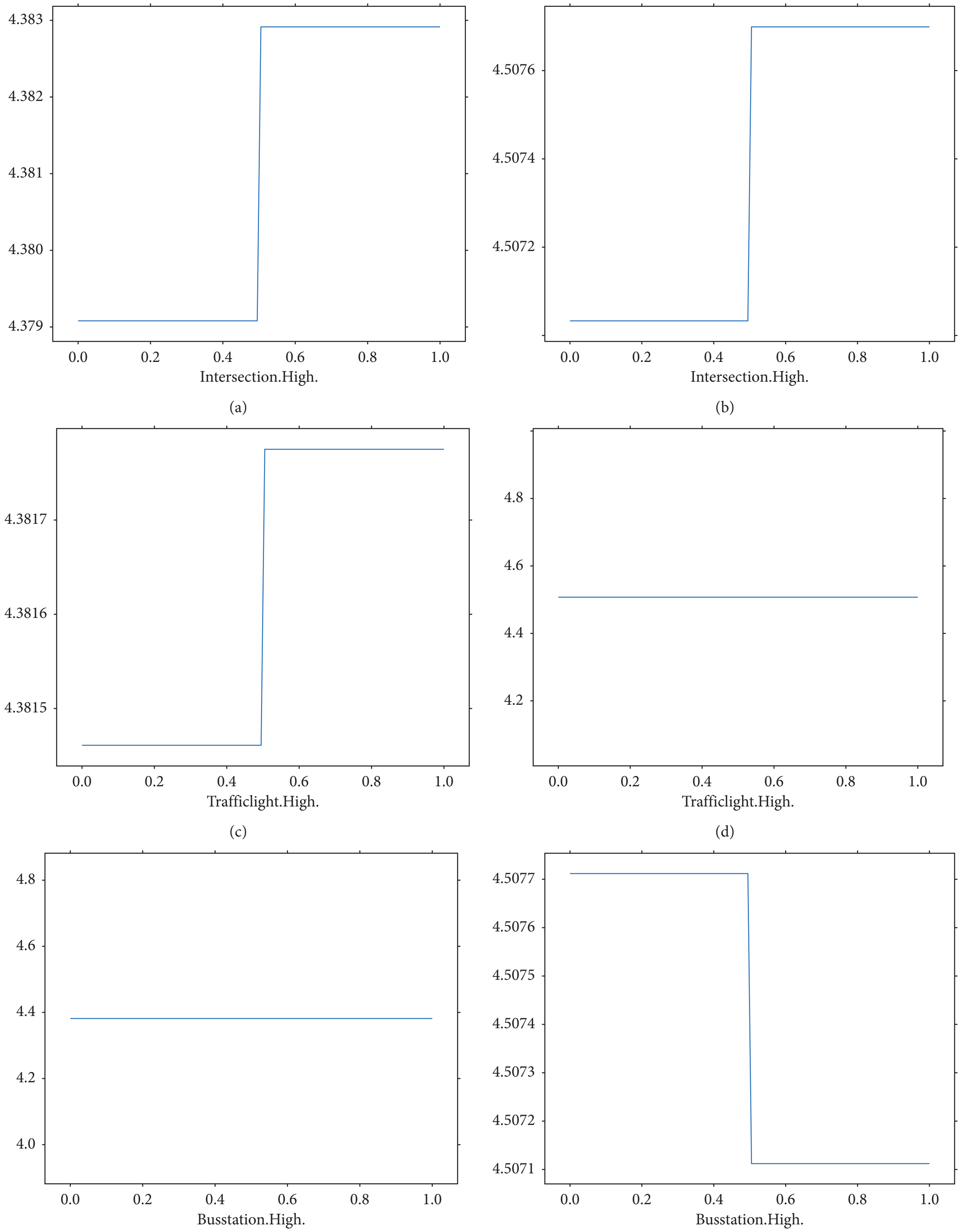


Figure 10: Partial dependence plots of land use and POIs on accidents (b).

around 10 per square kilometers. Therefore, to control for safer traffic conditions in the area, the land use of greenspace and parks can complement the role of transport facilities.

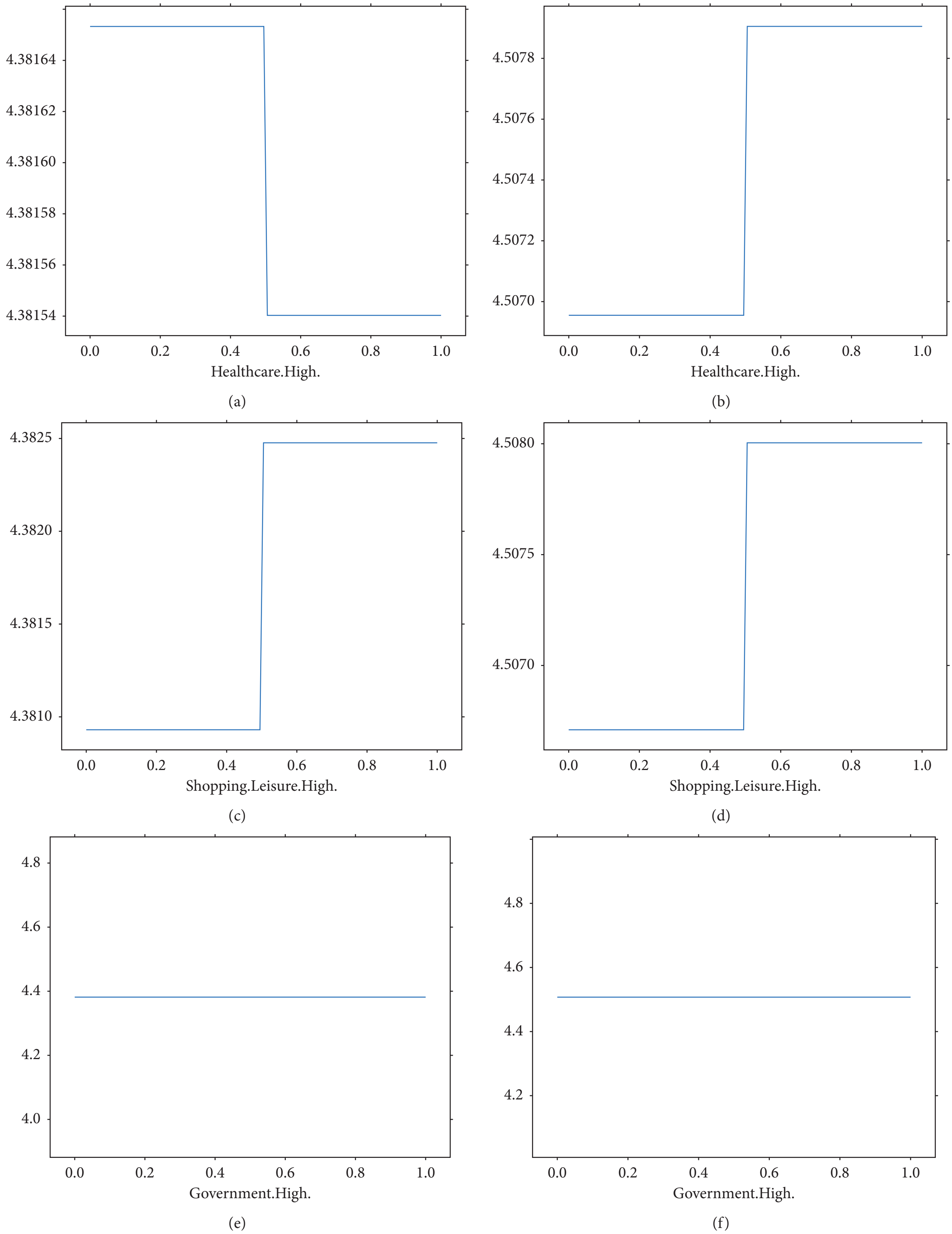
It is worth noting that the "Financial" and "Government" variables demonstrate opposite effects at the two different levels in Figure 9. If their stationary points are similar and their corresponding likelihoods also match, they show a steep decline in the local level and an abrupt rise at the neighborhood level. The partial dependence plot trend shows a negative correlation at the local level scale and a positive correlation at the nearest neighborhood level before reaching sustained stability. It is likely that the variation in the number of accidents is uncertain at both scales and may increase or decrease with a sudden boom in residential and educational institutions. However, the description of the accident relation with residential and educational units varies in the literature, and they sometimes conflict with each other. It has commonly been assumed that a positive correlation between the number of residential units (higher population density) and pedestrian crashes [10]. Other empirical cases, in contrast, demonstrate the direct opposite [6]. Similarly, some studies suggest that students are more likely to be at risk in areas with high school density due to irregular traffic crossing behaviors and low safety awareness. Despite this, Ukkusuri [4] presented contrasting experimental results. Nevertheless, the trend turning point in this study's partial dependency diagrams reveals that the overall number of accidents will no longer increase after the number of educational institutions reaches 70 per square kilometer and the number of residential areas reaches 50.

Concerning the partial dependence plot of the density dummy variables, all the variables show the linearities, as demonstrated in Figures 11–13. The three linear relationships are explained as follows:



(a) (b) (c) (d) (e) (f)

Figure 11: Partial dependence plots of density dummy variables of road structure facilities on accidents.



(a)

(b)

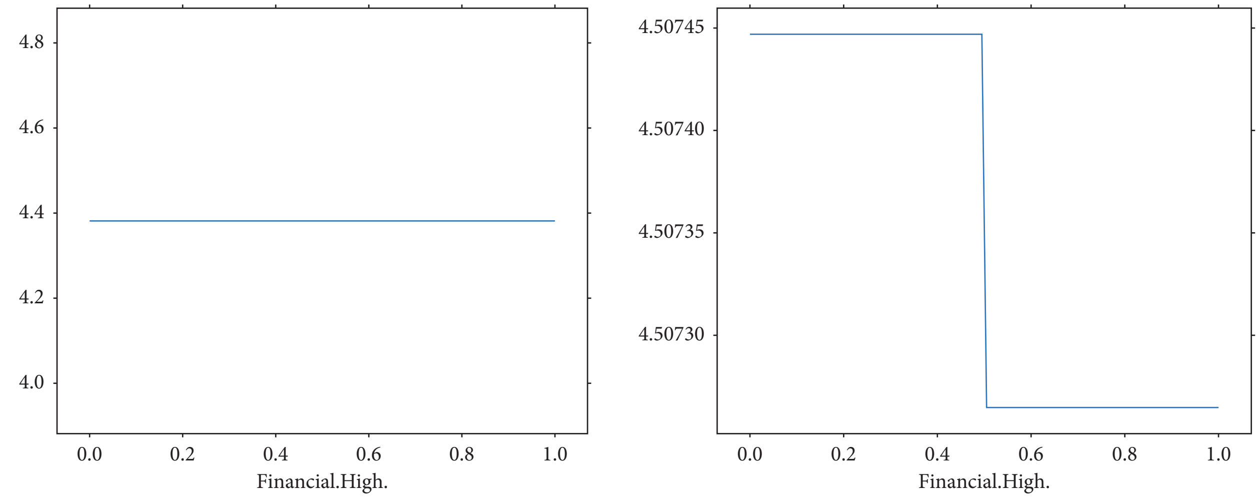
(c)

(d)

(e)

(f)

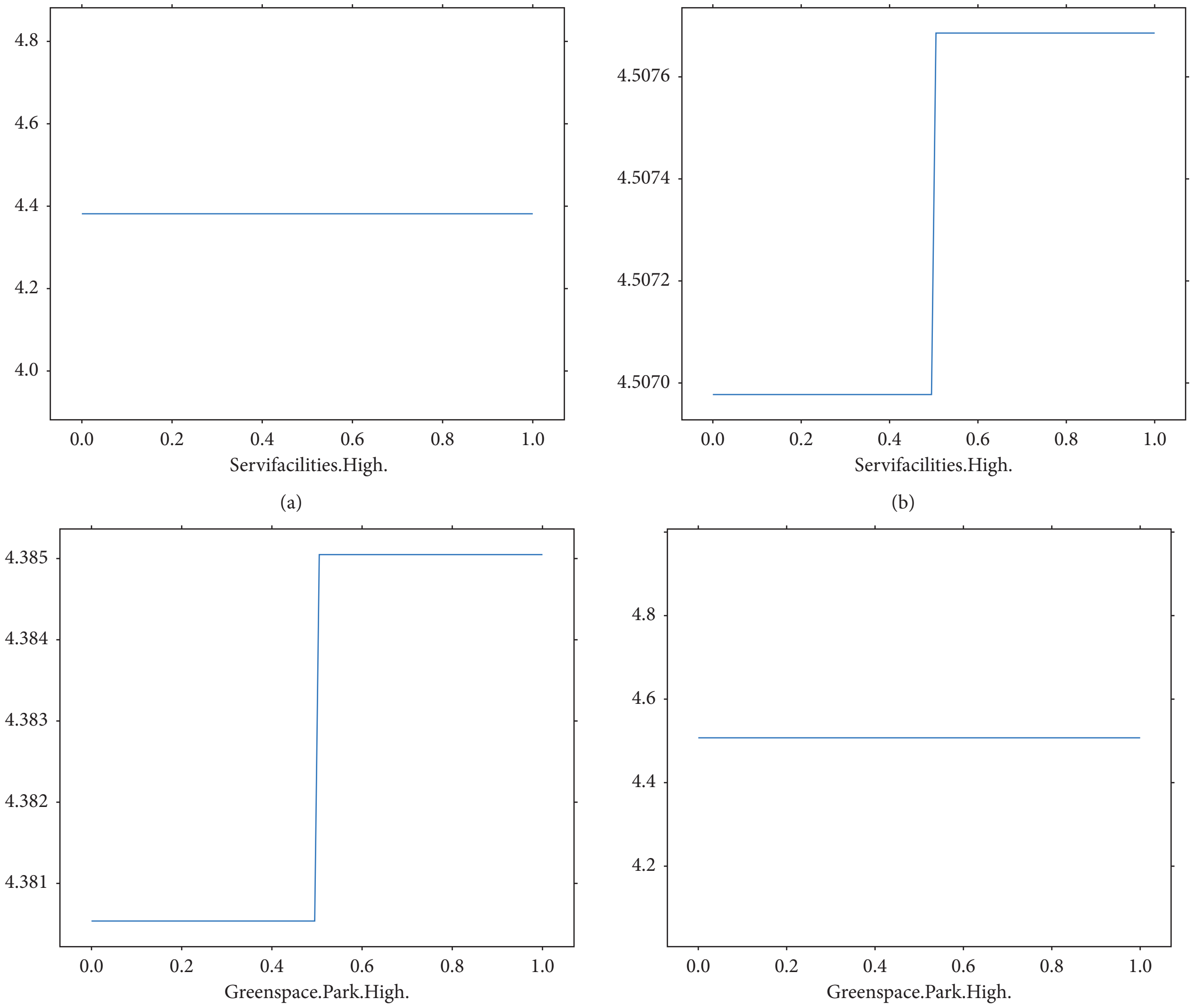
Figure 12: Continued.



(g)

(h)

Figure 12: Partial dependence plots of density dummy variables of land use and POIs on accidents (a).



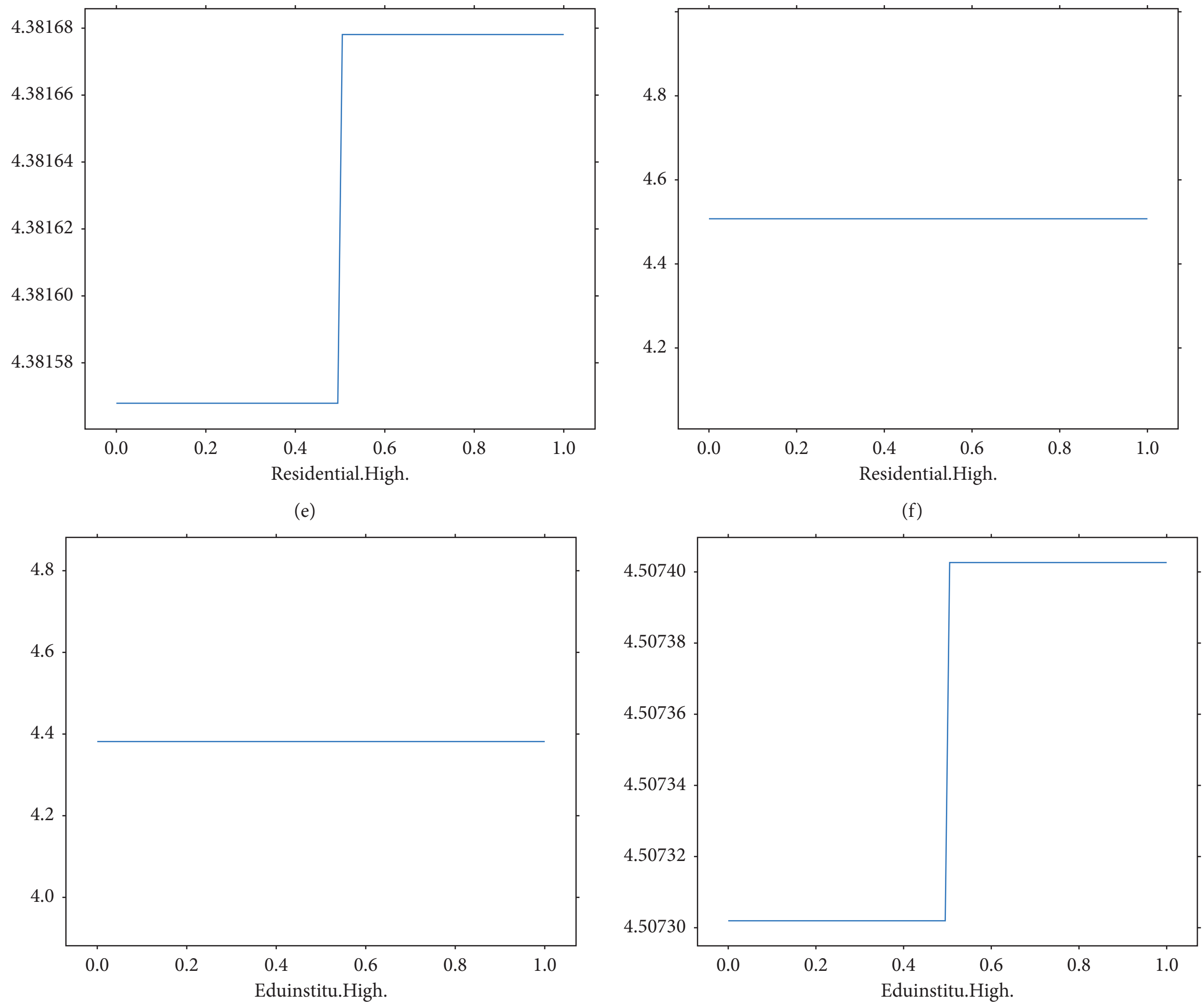
(a)

(b)

(c)

(d)

Figure 13: Continued.



(e)

(f)

(g)

(h)

Figure 13: Partial dependence plots of density dummy variables of land use and POIs on accidents (b).

(1) The plots of "Busstation.High," "Government.High," "Financial.High," "Servifacilities.High," and "EduInstitu.High" at the local level and "Greenspace&Park.High," "Trafficlight.High," "Government.High," and "Residential.High" at the nearest neighborhood level are linear parallel to the X-axis.

(2) However, "Healthcare.High" at the local level, as well as "Busstation.High" and "Financial.High," at the nearest neighborhood level present parallel lines with a high front and a low back, connected in the middle by a plumb line.

(3) All the other dummy variables have a plot similar to the previous one but are parallel lines with a low front and a high back and connected by a vertical line in the middle.

However, these three linearities neither describe a positive or negative correlation between high-density land uses and accidents nor present fluctuating intervals and meaningful turning points as the x-axis changes. The practical guide can hardly represent in the partial dependence lot of density dummy variables.

5. Conclusion

Road safety is critical to the health and wellbeing of people. To this end, a large and growing body of literature has investigated the leading causes and mechanisms of traffic accidents. Most research on traffic accidents has emphasized a complicated relationship between land use and urban transportation. In this study of the Suzhou Industrial Park (SIP), accident data provided by the SIP traffic police bureau were used to build a GBM machine learning model to identify the relationship between traffic accidents and land use. The research process includes the following:

(1) Processing the traffic accident data as well as land use and related facilities data.

(2) Establishing a GBM model on the frequency of traffic accidents following by determining the model parameters.

(3) Analyzing each land type variable's contribution to accident frequency and comparing these with the explanatory degree of the variables, as the number of iterations in the model grows.

(4) Discussing the estimated impacts of each variable on the accident intuitively according to the partial dependence plots at hand.

The study has highlighted factors affecting accidents, geographical scale exploration, and model operation. The GBM analysis was conducted at the local and neighborhood scales to explore the overall validity of the geographical levels and the model fitting. This also included the effect of variables of transportation facilities, land use, land mix, and density on the accident outcomes. The thirteen variables, including road facilities, land types, and some POI facilities, have been involved in two spatial scales that are bounded by TAZ units (local) and Thiessen polygons (nearest neighborhood). The results show that they all impacted accident occurrence at both scales, among which the more critical factors include categories of residential land, consumption and leisure land, and green parks. However, the experimental results at the two scales reflected vital differences and similarities at various experiment points. Among the rankings of relative importance, "Trafficlight," "EduInstitu," "Healthcare," "Intersection," and "Servifacilities" all have shown a degree of interpretation from 7% to 13% and existed in the crucial places of rankings on both scales. However, "Greenspace&Park," "Residential," and "D2" differed significantly and showed abnormality of the results. When adjusting the complexity of the tree, some variables such as "Residential" and "Greenspace&Park" appeared to be more influential at the local level, while the nearest neighborhood level showed more activity for the variables of "Trafficlight," "Busstation," "Healthcare," and "D2." In the partial dependence plots, the variables of "Residential" and "EduInstitu" showed accident frequencies at both scales. These results may be due to the fact that the spatial distribution of traffic accidents is uneven in SIP. Accident rates varied widely in each TAZ area. The large TAZ regions in the northern part of the study area and the dense TAZ regions in the southwest area showed the normal peak situation of the accidents, and the location distribution was scattered.

The local level has been seen as more suitable for measuring variables where pedestrians and vehicles have fixed mobility periods and moderate flows, such as residential areas and green parks. One the one hand, the nearest neighborhood level could be applied to a small number of variables related to public service facilities at fixed locations, such as traffic lights and bus stops. In other land uses such as financial networks, shopping, and leisure, where the sample size was extensive with a complexity of hierarchy, the scale could be modified according to the overall land use requirements. Therefore, this research suggests that it will be worth considering applying the nearest neighborhood scale with the boundary of Thiessen polygons in addition to the commonly used TAZ areas when examining traffic accidents or even traffic safety research of municipal engineering projects. When planning for a smaller geographical area, these different scale ranges might help confirm the settings and enrich the understanding of the study area's spatial structure to improve overall road safety.

Research on accident models has been developed using an advanced technique. GBM is a machine learning model that has been promoted to use rapidly in recent years. It particularly allows to validate existing models by ranking the importance of the coefficients and the variation in the model fit. Based on the introduction of multiple variables in the past studies, this research used the ordering of explanatory variables, tested the fitting degree of each variable by changing the parameter setting and partial dependence graphs, and comprehensively built an application model suitable for the current land use and road situation. Since a growing number of studies extensively analyze traffic accidents in different regions, the findings of this study could be compared with some of them to confirm its consistency and deviation. In this way, the analysis results of this study could be validated against others of its kind. The results of GBM included the coefficients of the variables under existing parameter settings. GBM was useful for this traffic accident study and positively contributed to understanding the relationship between urban form and traffic accidents. It is suggested that policymakers pay further attention to the benefits of using advanced methods in accident research than traditional means and understand the cause of this discrepancy to find the most efficient method in practice.

This study has several limitations that one should take into account for future studies. First, this study confirmed that the GBM model is only useful when it applies to regression and classification problems with the sufficient number of parameters from existing studies. Similar to other linear models, the coefficients of the variables were the only representative of the importance and influence of the dependent variable within one single model, and their values could not be used as an absolute reference for some practical applications. It is conceivable that if this model is applied to an emerging research subject, and the reliability of the GBM's result could be somewhat limited because it would not be able to produce absolute results. The application of this technique depends heavily on previous results and experience because determining variables (causality) and selecting the most suitable model parameters would be difficult. Therefore, this model might be beneficial for judging the relative fit of the identified variables, the relative importance between variables, and the internal interaction of the model parameters. When adjusting the complexity of the tree, the likelihood of variables fluctuated with the change in the number of trees but did not reach a stable value within the scope of the test. To ensure the integrity of the variables and the overall stability of the model, dummy variables that represent high density of land uses were introduced. However, the results of the density dummy variables were not satisfactory. The relative importance ranking was low. Also, the influence factors did not produce effective changes with the increasing number of trees, and they did not show meaningful fluctuation intervals and turning points in the partial dependence plot. Overall, this model could be best used for comparative purposes and might produce a more accurate model by adjusting parameters.

Second, the GBM occasionally presents accuracy and overfitting issues. The GBM predicts less accuracy than some regularization, polynomial regression, and partial regression methods [66]. It is also easy to overfit due to being relatively unconstrained in operation, causing a single decision tree to retain branches (without pruning) until it remembers the training data [62]. This needs to be treated carefully based on the different sizes and characteristics of the actual dataset when adjusting the parameters. As mentioned in the first point, it is worthwhile to explore varied applicable parameters to ensure the reliability of the model.

Third, the relatively small study area makes this finding less generalizable to other cities or regions of China, especially given the relatively unique layout of SIP although the exploration of geographical scale level is one of the important contributions of this study. In addition to the commonly used local level TAZ area, this research highlights the significance and usefulness of the nearest neighborhood level drawn from the Thiessen polygons zone, which can be used as a scientific and reasonable level scale. However, along with the results of this study, these levels have only been verified to apply to the Suzhou Industrial Park, and it might not be directly replicated to other regions in China.

Data Availability

The dataset utilized for this study is not publicly available due to the confidentiality agreement with the Suzhou Industrial Park government.

Disclosure

The contents of this study reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. Xi'an Jiaotong-Liverpool University assumes no liability for the contents or uses thereof.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was sponsored by a grant from Leap Grant Programme of the Research Institute of Future Cities at Xi'an Jiaotong-Liverpool University (ref. number: RIFC2019-01).

References

[1] WHO, *World Health Statistics 2015*, WHO, Geneva, Switzerland, 2019.

[2] WHO, *Data Visualization of the WHO Global Status Report on Road Safety 2015*, WHO, Geneva, Switzerland, 2019.

[3] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," *Accident Analysis & Prevention*, vol. 34, no. 6, pp. 729–741, 2002.

[4] S. Ukkusuri, L. F. Miranda-Moreno, G. Ramadurai, and J. Isa-Tavarez, "The role of built environment on pedestrian crash frequency," *Safety Science*, vol. 50, no. 4, pp. 1141–1151, 2012.

[5] J. Yang, W. Deng, J. Wang, Q. Li, and Z. Wang, "Modeling pedestrians' road crossing behavior in traffic system microsimulation in China," *Transportation Research Part A: Policy and Practice*, vol. 40, no. 3, pp. 280–290, 2006.

[6] J. Li and C. Shao, "Traffic accident forecast model based on BP neural network," (in CN), *in Computer and Communications*, vol. 24, no. 2, pp. 34–37, 2006, http://www.cnki.com.cn/Article/CJFDTOTAL-JTJS200602009.htm.

[7] E. Shesterov and A. Mikhailov, "Accident rates at signalized intersections," (in English), *Transportation Research Procedia*, vol. 20, pp. 613–617, 2017.

[8] S. S. Pulugurtha, V. R. Duddu, and Y. Kotagiri, "Traffic analysis zone level crash estimation models based on land use characteristics," *Accident Analysis & Prevention*, vol. 50, pp. 678–687, 2013.

[9] I. M. Abdalla, R. Raeside, D. Barker, and D. R. D. McGuigan, "An investigation into the relationships between area social characteristics and road accident casualties," (in en), *Accident Analysis & Prevention*, vol. 29, no. 5, pp. 583–593, 1997.

[10] C. Siddiqui, M. Abdel-Aty, and K. Choi, "Macroscopic spatial analysis of pedestrian and bicycle crashes," *Accident Analysis & Prevention*, vol. 45, pp. 382–391, 2012.

[11] Y. C. Macnab, "Bayesian spatial and ecological models for small-area accident and injury analysis," *Accident Analysis & Prevention*, vol. 36, no. 6, pp. 1019–1028, 2004.

[12] C. Ding, P. Chen, and J. Jiao, "Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: a machine learning approach," *Accident Analysis & Prevention*, vol. 112, pp. 116–126, 2018.

[13] A. Soltani and S. Askari, "Exploring spatial autocorrelation of traffic crashes based on severity," *Injury*, vol. 48, no. 3, pp. 637–647, 2017.

[14] W. Zou, X. Wang, and D. Zhang, "Truck crash severity in New York city: an investigation of the spatial and the time of day effects," *Accident Analysis & Prevention*, vol. 99, pp. 249–261, 2017.

[15] Y. Yue, Y. Zhuang, A. G. O. Yeh, J.-Y. Xie, C.-L. Ma, and Q.-Q. Li, "Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy," *International Journal of Geographical Information Science*, vol. 31, no. 4, pp. 658–675, 2017.

[16] Q. Zeng and H. Huang, "Bayesian spatial joint modeling of traffic crashes on an urban road network," *Accident Analysis & Prevention*, vol. 67, pp. 105–112, 2014.

[17] N. K. ChikkaKrishna, M. Parida, and S. S. Jain, "Identifying safety factors associated with crash frequency and severity on nonurban four-lane highway stretch in India," (in en), *Journal of Transportation Safety & Security*, vol. 9, no. sup1, pp. 6–32, 2017.

[18] J. C. Milton, V. N. Shankar, and F. L. Mannering, "Highway accident severities and the mixed logit model: an exploratory

empirical analysis," *Accident Analysis & Prevention*, vol. 40, no. 1, pp. 260–266, 2008.

[19] P. C. Anastasopoulos, F. L. Mannering, V. N. Shankar, and J. E. Haddock, "A study of factors affecting highway accident rates using the random-parameters tobit model," *Accident Analysis & Prevention*, vol. 45, pp. 628–633, 2012.

[20] K. J. Clifton and K. Kreamer-Fults, "An examination of the environmental attributes associated with pedestrian–vehicular crashes near public schools," vol. 39, no. 4, pp. 708–715, 2007.

[21] M. Wier, J. Weintraub, E. H. Humphreys, E. Seto, and R. Bhatia, "An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning," *Accident Analysis & Prevention*, vol. 41, no. 1, pp. 137–145, 2009.

[22] L. Ma and X. Yan, "Modeling zonal traffic accident counts with the regression under zero-adjusted inverse Gaussian distribution," *Procedia - Social and Behavioral Sciences*, vol. 138, pp. 452–459, 2014.

[23] K.-S. Ng, W.-T. Hung, and W.-G. Wong, "An algorithm for assessing the risk of traffic accident," *Journal of Safety Research*, vol. 33, no. 3, pp. 387–410, 2002.

[24] X. Wang, J. Yang, C. Lee, Z. Ji, and S. You, "Macro-level safety analysis of pedestrian crashes in Shanghai, China," *Accident Analysis & Prevention*, vol. 96, pp. 12–21, 2016.

[25] H. Huang, B. Song, P. Xu, Q. Zeng, J. Lee, and M. Abdel-Aty, "Macro and micro models for zonal crash prediction with application in hot zones identification," *Journal of Transport Geography*, vol. 54, pp. 248–256, 2016.

[26] Y. Zhao, H. Zhang, L. An, and Q. Liu, "Improving the approaches of traffic demand forecasting in the big data era," *Cities*, vol. 82, pp. 19–26, 2018.

[27] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in *Proceedings of the 2009 18th ACM conference on Information and knowledge management*, November 2009.

[28] C.-Y. Yu and X. Zhu, "Planning for safe schools," *Journal of Planning Education and Research*, vol. 36, no. 4, pp. 476–486, 2016.

[29] J.-K. Kim, S. Kim, G. F. Ulfarsson, and L. A. Porrello, "Bicyclist injury severities in bicycle–motor vehicle accidents," *Accident Analysis & Prevention*, vol. 39, no. 2, pp. 238–251, 2007.

[30] S.-P. Miaou, J. J. Song, and B. Mallick, "Roadway traffic crash mapping: a space-time modeling approach," *Journal of Transportation Statistics*, vol. 6, pp. 33–57, 2003.

[31] J. Aguero-Valverde and P. P. Jovanis, "Spatial analysis of fatal and injury crashes in Pennsylvania," *Accident Analysis & Prevention*, vol. 38, no. 3, pp. 618–625, 2006.

[32] S. Washington, J. Metarko, I. Fomunung, R. Ross, F. Julian, and E. Moran, "An inter-regional comparison: fatal crashes in the southeastern and non-southeastern United States: preliminary findings," *Accident Analysis & Prevention*, vol. 31, no. 1-2, pp. 135–146, 1999.

[33] S. Chen and S. Gao, "Study on the distance distribution of land use patterns and total travel volume in China's megacities," *Journal of Sichuan United University: Engineering Science*, vol. 3, no. 3, pp. 83–89, 1999.

[34] D. Wu, H. Mao, X. Zhang, and J. Huang, "Evaluation of urban land use efficiency in China," *Journal of Geography*, vol. 66, no. 8, pp. 1111–1121, 2011.

[35] F. Jiao, Z. Wen, and R. Li, "Evaluation of land type structure at county scale in loess hilly region (ansai)," *ISWC OpenIR*, vol. 12, no. 1, 2005.

[36] G. Song and L. Wang, "Key soil and water loss zones on the loess plateau are divided into land use zones," *ISWC OpenIR*, vol. 1, 1996.

[37] X. Liao, Z. Chen, H. Wang, and C. Luo, "Comprehensive land use zoning in Tibet," *Journal of Mountains*, vol. 27, no. 1, pp. 96–01, 2009.

[38] R. Amoh-Gyimah, M. Saberi, and M. Sarvi, "The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models," *Analytic Methods in Accident Research*, vol. 13, pp. 28–51, 2017.

[39] Y. Wang and K. M. Kockelman, "A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods," *Accident Analysis & Prevention*, vol. 60, pp. 71–84, 2013.

[40] Á. Briz-Redón, F. Martínez-Ruiz, and F. Montes, "Investigation of the consequences of the modifiable areal unit problem in macroscopic traffic safety analysis: a case study accounting for scale and zoning," *Accident Analysis & Prevention*, vol. 132, Article ID 105276, 2019.

[41] M. Gasparini, "Markov chain Monte Carlo in practice," *Technometrics*, vol. 39, no. 3, p. 338, 1997.

[42] X. Li, N. Zhang, and J. Gefu, "Grey-markov model for forecasting road accidents," *Journal of Highway and Transportation Research and Development*, vol. 3, no. 4, 2003.

[43] K. Kim and E. Yamashita, "Motor vehicle crashes and land use: empirical analysis from Hawaii," *Transportation Research Record*, vol. 1784, pp. 73–79, 2002.

[44] D. B. Fambro, K. Fitzpatrick, and R. J. Koppa, "Determination of stopping sight distances," in *NCHRP Report, Transportation Research Board,* Washington, DC, USA, 2021, https://trid.trb.org/view.aspx?id=476601.

[45] K. Kim, I. M. Brunner, and E. Y. Yamashita, "Influence of land use, population, employment, and economic activity on accidents," *Journal of the Transportation Research Board*, vol. 1953, no. 1, pp. 56–64, 2006.

[46] D. Dissanayake, J. Aryaija, and D. M. P. Wedagama, "Modelling the effects of land use and temporal factors on child pedestrian casualties," vol. 41, no. 5, pp. 1016–1024, 2009.

[47] A. Hadayeghi, A. Shalaby, and B. Persaud, "Development of planning-level transportation safety models using full bayesian semiparametric additive techniques," *Journal of Transportation Safety & Security*, vol. 2, pp. 45–68, 2010, https://escholarship.org/content/qt7721r7vh/qt7721r7vh.pdf?t=oqepmx.

[48] A. Matkan, A. Shariat, B. Mirbagheri, and M. Shahri, "Explorative spatial analysis of traffic accidents using geographically weighted Poisson regression model for urban safety planning," in *Proceedings of the the 3rd International Conference on Road Safety and Simulation*, Indianapolis, IN, USA, September 2011, https://www.researchgate.net/publication/237078451_Explorative_Spatial_Analysis_of_Traffic_Accidents_Using_Geographically_Weighted_Poisson_Regression_Model_for_Urban_Safety_Planning.

[49] M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using decision trees and neural networks," *IADIS International Conference on Applied Computing*, vol. 2, 2004.

[50] B. Abdulhai and S. G. Ritchie, "Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network," *Transportation Research Part C: Emerging Technologies*, vol. 7, no. 5, pp. 261–280, 1999.

[51] F. Yuan and R. L. Cheu, "Incident detection using support vector machines," *Transportation Research Part C: Emerging Technologies*, vol. 11, no. 3-4, pp. 309–328, 2003.

[52] F. Mannering, C. R. Bhat, V. Shankar, and M. Abdel-Aty, "Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis," *Analytic Methods in Accident Research*, vol. 25, p. 100113, 2020.

[53] F. L. Mannering and C. R. Bhat, "Analytic methods in accident research: methodological frontier and future directions," *Analytic Methods in Accident Research*, vol. 1, pp. 1–22, 2014.

[54] F. Mannering, "Temporal instability and the analysis of highway accident data," *Analytic Methods in Accident Research*, vol. 17, pp. 1–13, 2018.

[55] Q. Zeng, H. Huang, X. Pei, and S. C. Wong, "Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks," *Analytic Methods in Accident Research*, vol. 10, pp. 12–25, 2016.

[56] A. Stewart Fotheringham, M. Charlton, and C. Brunsdon, "The geography of parameter space: an investigation of spatial non-stationarity," *International Journal of Geographical Information Systems*, vol. 10, no. 5, pp. 605–627, 1996.

[57] W. Qin and J. Wang, "Exploring spatial relationship non-stationary based on GWR and GIS," in *Geoinformatics 2006: Geospatial Information Science*Vol. 6420, International Society for Optics and Photonics, Bellingham, WA, USA, 2006.

[58] J. Zhu and M. Xiao-Ping, "Safety evaluation of human accidents in coal mine based on ant colony optimization and SVM," *Procedia Earth and Planetary Science*, vol. 1, no. 1, pp. 1418–1424, 2009.

[59] K. Zheng, Y. Chen, Y. Jiang, and S. Qiao, "A SVM based ship collision risk assessment algorithm," *Ocean Engineering*, vol. 202, Article ID 107062, 2020.

[60] A. B. Parsa, H. Taghipour, S. Derrible, and A. Mohammadian, "Real-time accident detection: coping with imbalanced data," *Accident Analysis & Prevention*, vol. 129, pp. 202–210, 2019.

[61] L. Zheng and T. Sayed, "Bayesian hierarchical modeling of traffic conflict extremes for crash estimation: a non-stationary peak over threshold approach," *Analytic Methods in Accident Research*, vol. 24, Article ID 100106, 2019.

[62] C. Molnar, Interpretable Machine Learning.

[63] Z. Zhang, Q. He, J. Gao, and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," *Transportation Research Part C: Emerging Technologies*, vol. 86, pp. 580–596, 2018.

[64] Z. Fan, C. Liu, D. Cai, and S. Yue, "Research on black spot identification of safety in urban traffic accidents based on machine learning method," *Safety Science*, vol. 118, pp. 607–616, 2019.

[65] L. Li, B. Du, Y. Wang, L. Qin, and H. Tan, "Estimation of missing values in heterogeneous traffic data: application of multimodal deep learning model," *Knowledge-Based Systems*, vol. 194, Article ID 105592, 2020.

[66] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, Berlin, Germany, 2013.

[67] J. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, pp. 11–28, 2000.

[68] S. Wang, P. Dong, and Y. Tian, "A novel method of statistical line loss estimation for distribution feeders based on feeder cluster and modified XGBoost," *Energies*, vol. 10, no. 12, p. 2067, 2017.

[69] X. Ren, H. Guo, S. Li, S. Wang, and J. Li, *A Novel Image Classification Method with CNN-XGBoost Model*, pp. 378–390, Springer, Berlin, Germany, 2017.

[70] S. Tonidandel and J. M. Lebreton, "Determining the relative importance of predictors in logistic regression: an extension of relative weight Analysis," *Organizational Research Methods*, vol. 13, no. 4, pp. 767–781, 2010.

[71] R. F. Martell, D. M. Lane, and C. Emrich, "Male-female differences: a computer simulation," *American Psychologist*, vol. 51, no. 2, pp. 157-158, 1996.

[72] J. W. Johnson and J. M. Lebreton, "History and use of relative importance indices in organizational research," *Organizational Research Methods*, vol. 7, no. 3, pp. 238–257, 2004.

[73] D. R. Cutler et al., "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007, https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-0539.1.

[74] J. Elith, S. Ferrier, F. Huettmann, and J. Leathwick, "The evaluation strip: a new and robust method for plotting predicted responses from species distribution models," *Ecological Modelling*, vol. 186, no. 3, pp. 280–289, 2005, http://www.sciencedirect.com/science/article/pii/S0304380004006180.

[75] SIPM Committee. "Suzhou Industrial Park Profile," http://www.sipac.gov.cn/szgyyq/yqjj/common_tt.shtml.

[76] J. Brownlee, "Overfitting and underfitting with machine learning algorithms," in *Machine Learning Mastery*, 2016, https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms.

[77] Y.-S. Chung, "Factor complexity of crash occurrence: an empirical demonstration using boosted regression trees," *Accident Analysis & Prevention*, vol. 61, pp. 107–118, 2013.

[78] P. Chen and J. Zhou, "Effects of the built environment on automobile-involved pedestrian crash frequency and risk," *Journal of Transport & Health*, vol. 3, no. 4, pp. 448–456, 2016.

[79] D. Saha, P. Alluri, and A. Gan, "Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees," *Accident Analysis & Prevention*, vol. 79, pp. 133–144, 2015.

[80] J. Brownlee, How to Configure the Gradient Boosting Algorithm, ed, 2016, https://machinelearningmastery.com/configure-gradient-boosting-algorithm/.