Published in final edited form as:
Accid Anal Prev. 2018 March ; 112: 116–126. doi:10.1016/j.aap.2017.12.026.

Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: A machine learning approach

Chuan Ding[a], **Peng Chen**[b,*], **Junfeng Jiao**[c]

[a]School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure System and Safety Control, Beihang University, Beijing, China

[b]School of Architecture and Urban Planning, Harbin Institute of Technology Shenzhen Campus, Shenzhen, China

[c]School of Architecture, University of Texas at Austin, Austin, USA

Abstract

Although a growing body of literature focuses on the relationship between the built environment and pedestrian crashes, limited evidence is provided about the relative importance of many built environment attributes by accounting for their mutual interaction effects and their non-linear effects on automobile-involved pedestrian crashes. This study adopts the approach of Multiple Additive Poisson Regression Trees (MAPRT) to fill such gaps using pedestrian collision data collected from Seattle, Washington. Traffic analysis zones are chosen as the analytical unit. The effects of various factors on pedestrian crash frequency investigated include characteristics the of road network, street elements, land use patterns, and traffic demand. Density and the degree of mixed land use have major effects on pedestrian crash frequency, accounting for approximately 66% of the effects in total. More importantly, some factors show clear non-linear relationships with pedestrian crash frequency, challenging the linearity assumption commonly used in existing studies which employ statistical models. With various accurately identified non-linear relationships between the built environment and pedestrian crashes, this study suggests local agencies to adopt geo-spatial differentiated policies to establish a safe walking environment. These findings, especially the effective ranges of the built environment, provide evidence to support for transport and land use planning, policy recommendations, and road safety programs.

Keywords

Pedestrian crash frequency; Built environment; Machine learning; Variable importance; Non-linear effects

[*]Corresponding author. chenp5@uw.edu (P. Chen).

1. Introduction

Walking is a type of popular aerobic physical activity, which has numerous health benefits such as weight control, lower risk of heart disease, stroke, depression, and some cancers (Centers for Disease Control and Prevention, 2012). However, exposure to automobile collisions and the risk of being injured in such accidents can discourage walking. Among various factors, built environment plays a key role in attracting pedestrians. Not all walking environments are identically safe, and various built environment features are related to different levels of collision risks. For example, residents in rural areas generally do less walking and have a higher risk of being involved in pedestrian collisions due to more constrained walking environment access and higher traffic operating speeds. Recent research using the pedestrian danger index to evaluate the quality of walking environments found that 8/10 of the most dangerous metro areas for walking are in Florida and 19/20 of the most dangerous metro areas for walking are clustered in the southern United States (US) (Smart Growth America & National Complete Streets Coalition, 2017). Cities in the southern US are generally more sprawled as compared to the northeastern and the west coastal US cities, which may be not safe for walking. To deepen our understanding, investigating and promoting a safe walking environment is critical to the success of pedestrian injury and death reduction.

According to data reported in 2009, walking accounts for roughly 10.4% of all trips in the US, and mostly for errands and social and recreational purposes (Alliance for Biking and Walking, 2016; Bureau of Transportation Statistics, 2016). Although the number of traffic-related injuries and deaths has steadily declined in recent years, the number of injured and killed pedestrians does not follow the same trend. Data shows that 70,000 people were injured, and 5376 people were killed in automobile-involved pedestrian crashes in 2015. Pedestrian deaths accounted for roughly 15% of all traffic-related fatalities in the US. In addition, the number of pedestrian deaths in 2015 was noted as the biggest single-year increase. There were only 61,000 pedestrian injuries and 4795 deaths in 2006. A 11.48% has increased from 2006 to 2015 (The National Highway Traffic Safety Administration, 2015). Although authorities have the goodwill of promoting walking for health and environment benefits, pedestrian safety is worthy of greater inputs.

How can we promote a safer walking environment? A large body of literature has discussed the relationship between pedestrian crash outcomes and built environment factors. These outcome measurements include but are not limited to pedestrian crash frequency (Chen and Zhou, 2016; Miranda-Moreno et al., 2011; Narayanamoorthy et al., 2013; Siddiqui et al., 2012; Ukkusuri et al., 2011; Ukkusuri et al., 2012; Wang and Kockelman, 2013), injury severity, (Abay, 2013; Aziz et al., 2013; Clifton et al., 2009; Islam and Hossain, 2015; Kim et al., 2008; Mohamed et al., 2013; Tarko and Azam, 2011; Zahabi et al., 2011), and exposure and risk (Chen and Zhou, 2016; Moudon et al., 2008; Moudon et al., 2011; Schneider et al., 2010; Schneider et al., 2013; Wang et al., 2017). In addition, a handful of studies have identified spatial clusters of pedestrian crashes (Dai, 2012; Dai et al., 2010; Pulugurtha et al., 2007).

Several studies have comprehensively reviewed work concerning pedestrian crash frequency (Chen and Zhou, 2016; Pulugurtha and Sambhara, 2011; Ukkusuri et al., 2011; Wier et al., 2009), and therefore this study only focuses on recent advancements. The conventional research is built upon an agreed framework to identify various effects, including (1) roadway design, such as the densities of different types of streets (sidewalks, local streets, and arterial routes), intersections, and bus stops/stations; (2) land use, such as land use mixture, different types of land use including residential, industrial, schools, parks or activity centers, and parking; other land use measurements examined are zone size, and urban versus rural; (3) population characteristics, such as densities of population and employment, the densities of different sub-groups including senior citizens, children and teenagers, and other sociodemographic measurements such as median income, poverty rate, and race; (4) travel demand measurements, such as walking miles traveled, walking mode share, traffic volume, and trip forecast; (5) traffic controls, such as speed limit (Cai et al., 2016; Chen and Zhou, 2016; Miranda-Moreno et al., 2011; Moudon et al., 2008; Moudon et al., 2011; Narayanamoorthy et al., 2013; Pulugurtha and Sambhara, 2011; Siddiqui et al., 2012; Ukkusuri et al., 2011; Ukkusuri et al., 2012; Wang and Kockelman, 2013; Wier et al., 2009), as shown in Table 1.

In terms of methodology, early research has mostly employed negative binomial (NB) models to identify various correlations between pedestrian crash frequency and explanatory variables, which can handle the issue of data over-dispersion (Miranda-Moreno et al., 2011; Pulugurtha and Sambhara, 2011). For example, Miranda-Moreno et al. (2011) used a standard NB model, a generalized GNB model, and a latent-class NB model to investigate how the built environment affects both pedestrian activity and collision frequency. Pedestrian crashes are random events, and many factors remain unobserved. In this context, random effects models are therefore applied to account for the unobserved heterogeneity (Ukkusuri et al., 2011). More recent research includes multiple random effects to account for spatial autocorrelations in addition to the unobserved heterogeneity (Chen and Zhou, 2016; Narayanamoorthy et al., 2013; Siddiqui et al., 2012; Wang and Kockelman, 2013). Recently, the Bayesian hierarchical intrinsic conditional autoregressive model has been used to analyze crash counts (Chen and Zhou, 2016). A spatial modeling approach provides a chance to capture spatial autocorrelations.

Cross-sectional models are usually used to analyze pedestrian crash counts aggregated over multiple years and time series analysis is rarely used to account for temporal autocorrelations inside pedestrian crash counts. The reasons may include: first, the number of reported pedestrian crashes in each analytical unit is small if split by multiple years, and the resulting data is likely to involve the issue of the excess of zeros. An alternative way to solve this problem is to aggregate pedestrian crashes to a larger unit, such as from block groups to census tracts. However, the results may not offer insightful strategies for local safety improvement due to the large size of the analytical unit. Additionally, implementing the statistical analysis at a larger scale may lead to regression towards the mean. Finally, changes to the built environment between two consecutive years are usually trivial, and it is not necessary to employ time series models as an analytical method. If underreported minor collisions are better documented, using time series models to examine factors impacting pedestrian crash counts may be rewarding.

A methodological advancement of road safety research is the introduction of machine learning techniques, such as the Chi-squared Automatic Interaction Detection (CHAID) decision tree approach to rank the relative importance of contributing factors (Prati et al., 2017). To date, limited research has applied similar methods to study pedestrian crash frequency. The advantage of employing machine learning approaches includes, at least, the following four aspects. First, machine learning approaches can handle big data. In the future, authorities may develop smartphone apps to collect collision data, and minor collisions could be better reported by individuals. When processing large amounts of data, computational efficiency is a major challenge. Second, machine learning techniques are more sensitive to outliers in the sample and capture the interactions among variables. Analytical efficiency could be greatly improved. Third, discrete variables with many categories are more properly handled by machine learning techniques in contrast to conventional regression models (Prati et al., 2017). Fourth, although elasticities can be computed for explanatory variables and used for treatment evaluation, the decision tree approach provides an alternative way to rank factors. Despite these merits, the major weakness of machine learning techniques is its incapacity in causal inferences.

With the goal of providing an alternative way to examine various effects associated with pedestrian crashes, a recently developed methodology named Multiple Additive Poisson Regression Trees (MAPRT) model is introduced in this study. MAPRT model can rank the relative importance of many contributing factors and it also can identify their non-linear effects on pedestrian crash frequency. The following sections include methodology, results, conclusions, and a discussion of policy implications. Research limitations and potential advancement in future research is also elaborated upon.

2. Methodology

2.1. Multiple additive Poisson regression trees model

This study adopts a machine learning approach, the Multiple Additive Poisson Regression Trees (MAPRT) model, to investigate built environment effects on automobile-involved pedestrian crash frequency. Assuming x is a set of explanatory variables (i.e., built environment attributes) and $f(x)$ is an approximation function of the response variable y (i.e., pedestrian crash frequency), this method estimates a function as an additive expansion of a basis function $h\,(x; a_m)$, as noted in Eq. (1) (Chung, 2013; Ding et al., 2016; Friedman, 2001; Friedman et al., 2001; Saha et al., 2015).

$$f(x) = \sum_{m=1}^{M} f_m(x) = \sum_{m=1}^{M} \beta_m h(x; a_m) \tag{1}$$

where a_m is the mean of split locations and the terminal node for each splitting variable in an individual decision tree $h\,(x; a_m)$, β_m represents weights given to the nodes of each tree. For the Poisson outcome, parameters a_m and β_m are determined by minimizing a specified loss function, as shown in Eq. (2).

$$L(y, f(x)) = -2\sum_{i=1}^{M}\left[y_i f(x_i) - \exp f(x_i)\right] \tag{2}$$

To estimate the parameters, a_m and β_m, Friedman et al., 2001 proposed the gradient boosting approach, and the algorithm is well documented in several studies (Ding et al., 2016; Friedman et al., 2001). The model is built in a stage-wise fashion and is updated by minimizing the expected value of the loss function in Eq. (2). As the number of iterations increases, the minor fluctuations in training data are exaggerated, thus resulting in the poor prediction performance of testing data. To overcome such over-fitting problems (Friedman et al., 2001), learning rate ξ $(0 < \xi \leq 1)$, also called shrinkage, is used to scale the contribution of each base tree model, as shown in Eq. (3).

$$f_m(x) = f_{m-1}(x) + \xi \cdot \beta_m h(x; a_m), \text{ where } 0 < \xi \leq 1 \tag{3}$$

Smaller shrinkage values better minimize the loss function. However, it requires a larger number of trees to be added to the model. Another important parameter for the MAPRT approach is the tree complexity, referring to the number of splits (i.e., the number of nodes) that is used to fit each decision tree. To capture interactions among explanatory variables, it is necessary to increase the tree complexity. In this study, the tree complexity is the focus.

2.2. Relative importance of factors

The MAPRT model can handle different types of explanatory variables, capture interactions among them, and fit complex non-linear relationships (Chung, 2013; Guelman, 2012; Saha et al., 2015). Since each explanatory variable has a different effect on the response variable, it is often helpful to quantify their relative importance to make a comparison. However, two fundamental objectives of predictive learning, accuracy and interpretability, do not always coincide. Generally, most machine learning algorithms are seen as 'black-box' procedures. Drawing on insights and techniques from both statistical and machine learning methods, the tree-based ensemble methods not only achieve strong predictive performance, but also identify and interpret relevant variables and interactions (Friedman et al., 2001). This MAPRT model can provide an alternative way of sorting influential factors and identify the non-linear effect of each factor, which is helpful for drawing policy recommendations and initiating safety programs.

For a single decision tree T, Breiman et al. propose the following measure as an approximation of the relative importance of x_κ in predicting the response variable (Breiman et al., 1984), see Eq. (4).

$$I_\kappa^2(T) = \sum_{t=1}^{J-1}\hat{\tau}_t^2 I(v(t) = \kappa) \tag{4}$$

where the summation is over the non-terminal node t of the J-terminal node tree T, x_κ is the splitting variable associated with the node t, an d$\hat{\tau}_t^2$ is the corresponding empirical improvement in squared error as a result of splitting. Instead of fitting a single "best" model, the tree-based ensemble method strategically combines multiple simple tree models to optimize the predictive performance. For a collection of decision trees$\{T_m\}_1^M$, obtained through a gradient boosting approach, Eq. (3) can be generalized by its average over all additive trees, as shown in Eq. (5):

$$I_\kappa^2 = \frac{1}{M}\sum_{m=1}^{M} I_\kappa^2(T_m) \tag{5}$$

2.3. Partial dependence plots

After most relevant variables have been identified, the next step is to understand the nature of the dependence of the approximation $f(x)$ on their joint values. Graphical rendering of $f(x)$ as a function of its arguments provides a comprehensive summary of its dependence on the joint value of input variables (Hastie et al., 2009). Following the studies conducted by Friedman et al., 2001 and Hastie et al. (2009), partial dependence plots can be mathematically defined as follows:

Suppose S is a subset of p explanatory variables, such that $S \subset \{x_1, x_2, \ldots, x_p\}$. Let C be a complement to S, such that $S \cup C = \{x_1, x_2, \ldots, x_p\}$. The model approximation function, $f(x)$, depends upon p explanatory variables: $f(x) = f(x_S,\ x_C)$. The partial dependence of S explanatory variables on the approximation function $f(x)$ can be defined as follows:

$$f_S(x_s) = E_{x_c} f(x_S, x_C) \tag{6}$$

and can be estimated by

$$\bar{f}_S(x_S) = \frac{1}{N}\sum_{i=1}^{N}[f(x_S, x_C)] \tag{7}$$

where $\{x_{C1}, x_{C2}, \ldots, x_{CN}\}$ are the values of X_C occurring over all observations in the training data. In other words, to calculate the partial dependence of a given variable, the entire training set must be utilized for every set of joint values in X_S. It should be noted the partial dependence functions defined in Eq. (6) represent the effect of X_S on $f(x)$ after accounting for the average effects of the other variables X_C on $f(x)$ (Hastie et al., 2009). Partial dependence functions can be used to interpret the results of the MAPRT model, and can serve as a useful description regarding the chosen subset's effect on $f(x)$.

3. Data sources

This study examines how built environment factors are correlated with pedestrian crashes at an area level. According to the previous literature, built environment features are quantified by factors of the road network and land use. In this study, the rich built environment data provides an opportunity to examine the effects of some under-investigated factors, such as the density of stop signs, the proportion of steep areas, and the zonal average posted speed limit.

This study uses data collected from the city of Seattle, Washington. This data consists of two major components, pedestrian crash records and built environment features. The pedestrian crash records are obtained from the Seattle Department of Transportation (SDOT) during the period from January 2008 to December 2012. In total 2186 pedestrian crashes are reported and geocoded. The spatial distribution of these crashes is illustrated in Fig. 1. Two variables, walking mode share and the number of trips for all modes, are the major output of a regional activity-based travel demand model (Puget Sound Regional Council, 2014). The other data sets were obtained from three agencies, including SDOT, King County, and Puget Sound Regional Council. Traffic analysis zone (TAZ) is selected as the unit of analysis because it matches the existing travel demand output model and quantifies built environment factors at a relative small geospatial scale. There are 863 TAZs in Seattle. The process of built environment feature quantification is done in ArcGIS using overlay functions. Table 2 defines selected variables and includes a data summary. It is worth noting that MAPRT model can handle different types of independent variables as model input, and hence requires very little data preprocessing.

4. Results

4.1. Building the model

In the model building process, it is necessary to select parameters of learning rate and tree complexity. A greater value of learning rate (i.e., 0.1) is too fast for both tree complexity and the number of trees to minimize the error (Elith et al., 2008). According to previous studies (Chung, 2013; Guelman, 2012), a learning rate of 0.001 generally generates a final model with lower predictive deviance and a reasonable tree size (at least 1000 trees). Therefore, this study fixes the learning rate as 0.001. Another important model parameter is tree complexity. Tree complexity should reflect the true interaction among the explanatory variables (Friedman et al., 2001). However, this information is always unknown. To obtain reliable results, a series of models are tested by sequentially increasing the interaction depth level of trees from one to ten. A five-fold cross-validation procedure has been used to obtain the model performance with different levels of tree complexity and different numbers of trees. The dataset is equally split into five distinct subsets. Each subset (20%) is used as the testing data while the remaining subsets (80%) are used to train the model.

When the model uses a ten-way interaction, it gets the lowest predictive deviation with an ensemble of 3141 trees. The relationship between the predictive deviation and the number of trees is shown in Fig. 2. The training error (the black line) continues to decrease as the number of trees increases. However, the test error (the red line) increases when the number of trees reaches certain values. This finding is indicative of the over-fitting problem. To solve this problem, the optimal number of trees is automatically set at the point where the cross-validation error (the green line) ceases to decrease.

4.2. Relative contribution of explanatory variables

To explore the relationship between various explanatory variables and automobile-involved pedestrian crash frequency, the relative contribution of each explanatory variable is calculated with different levels of tree complexity. Fig. 3 shows that changes in the relative contribution are larger at a low level of tree complexity and the relative contribution reaches the relatively stable state when the tree complexity exceeds eight. It is worth noticing that the vertical scale to the right is for the number of trips since its effect is much larger than others. A higher value of relative contribution indicates a stronger effect. In this study, effects are measured in a relative form, and the total effects of all explanatory variables are 100%.

Using the MAPRT model with a tree complexity of ten, Table 3 presents the relative contribution and the rank of explanatory variables on automobile-involved pedestrian crash frequency. There are obviously differential effects among variable categories. The number of trips is the most important variable that contributes to pedestrian crash frequency with a value of 23.78%. This finding indicates that traffic volume plays an important role in impacting pedestrian crash frequency, which is consistent with a prior studies (Miranda-Moreno et al., 2011). Household density and commercial land use are the second and third most important variables in predicting pedestrian crash frequency, with a contribution of 14.38% and 13.16%, respectively. A few studies suggest that pedestrian crash frequency is closely correlated with density and commercial land use (Narayanamoorthy et al., 2013; Wier et al., 2009), and such a relationship is clearly confirmed with Seattle’s data. Land use mixture, with a contribution of 11.39%, is the fourth influential variable. A similar significant relationship between land use mixture and pedestrian crash frequency has been also identified in prior studies (Chen and Zhou, 2016; Wang and Kockelman, 2013).

Collectively, the number of trips, household density, commercial land use, and land use mixture account for approximately 66% of the total effects on pedestrian crash frequency, indicating the key role of density and diversity of the built environment. If we use the traditional Poisson regression, household density is insignificant because of its high correlation with other variables. This problem points out the superiority of the MAPRT model, which can capture interactions among different explanatory variables.

In terms of road network factors, the density of interactions accounts for about 8% of the total effects on pedestrian crash frequency. Especially, the 4-way intersection density contributes the most to with a value of 5.17%. Unexpectedly, sidewalk density is not a particularly important factor with a contribution of less than 1%. All street elements play important roles in relation to pedestrian crash frequency with more than 1% contribution. Among them, zonal speed limits have a nearly 6% contribution which is the fifth most influential variable, followed by the steep area proportion, bus stop density, and stop sign density with about 1.5%–2.5% contributions. Previous studies have shown that there are close connections between street elements and pedestrian crash frequency. For example, a higher zonal speed limit, steep area proportion, and bus stop density are generally associated with more pedestrian crashes (Chen and Zhou, 2016; Miranda-Moreno et al., 2011; Wang and Kockelman, 2013). These findings have policy implications as they can help guide street redesign processes.

As to land use variables, the factors of employment density, industrial land density, and activity center density, each have about a 3%–4% contribution in predicting pedestrian crash frequency. Generally, dense employment areas are mostly friendly environments for pedestrians (Geyer et al., 2006). This study suggests that industrial land use plays an important role in relation to pedestrian crash frequency (Chen and Zhou, 2016; Ukkusuri et al., 2012). Activity center density, including public libraries, education centers, churches, and community centers, show a clear effect on the pedestrian crash frequency with a contribution of 4.11%. Regarding traffic demand factors, walking mode share contributes about 3.32% in the model.

4.3. Non-linear effects of key explanatory variables

To further investigate how built environment variables affect pedestrian crash frequencies, we use partial dependence plots to illustrate these associations. It is worth noting that partial dependence plots show the dependence between the response variable and a set of features after accounting for the average joint effects of all other variables in the model (Chung, 2013; Guelman, 2012; Saha et al., 2015). Compared with traditional regression models, the relationship between the response variable and explanatory variables is no longer assumed to be linear. This method is useful for understanding the impact of changes in a single variable when integrating across all other explanatory variables.

The effects of key land use variables on pedestrian crash frequency are shown in Fig. 4. As noted, when household density is within 500 households per hectare, the likelihood of a pedestrian crash occurrence increases with a higher household density in an almost exponential way. Beyond this range, its effect remains stable. As to commercial land use, Narayanamoorthy et al. (2013) and Miranda-Moreno et al. (2011) have found that commercial land use has a positive effect on pedestrian crash frequency. This study shows a non-linear relationship between them. When the commercial land use value it between 0–8%, pedestrian crash frequency increases by about 1.2 crashes. Beyond this range, commercial land use has a negative effect on the pedestrian crash frequency with a low decreasing rate. Regarding land use mixture, mixed results are identified. For example, land use mixture shows a positive association with pedestrian crash frequency in the study conducted by Chen and Zhou (2016), whereas a negative association has been found in Wang and Kockelman’s study (2013). A complex non-linear relationship is identified in this study. When the land use mixture value is below 0.6, it seems to have a small effect on pedestrian crash frequency. However, this effect increases substantially when land use mixture moves into the range of 0.6–0.7. After that, its effect is stable. The effects of two variables, activity center density and employment density, follow a similar trend, but in slightly different ways. The effect of activity center density is in an exponential way, and the effect of employment density shows a linear pattern. Previous studies provide mixed conclusions on the effect of industrial land use on pedestrian crash frequency. A significant negative relationship between industrial land use and pedestrian crash frequency has been identified in studies conducted by Chen and Zhou (2016) in Seattle and Miranda-Moreno et al. (2011) in Montreal, whereas this relationship is positive in a study conducted by Ukkusuri et al. (2012). As noted in this study, the effect of industrial land use is erratic when it is below 3%, and is very small. From 3% to 10%, this effect decreases in a linear way, followed by a long plateau. To summarize, the effects of land use factors are quite complex in nature, and most of them appear to be non-linearly correlated with pedestrian crash frequency.

Fig. 5 shows the effects of street elements on pedestrian crash frequency. When it comes to zonal speed limit, the plot shows a non-linear relationship with pedestrian crash frequency. Specifically, it has a mostly linear effect on pedestrian crash frequency between 20~25 miles per hour, indicating that a higher zonal speed limit is associated with more pedestrian crashes. Beyond this range, it has a slightly negative effect on pedestrian crash frequency, and remains stable at the speed of 28 mph. Although previous studies suggest that slowing down the driving speed limit is an effective strategy to reduce automobile-involved crashes (Chen, 2015; Zahabi et al., 2011), in this study such a strategy is only effective when the speed limit is 25 mph or below. Beyond 25 mph, a higher speed limit does not have any expected effect on pedestrian crashes. The plot indicates a logarithmic relationship between steep area proportion and pedestrian crash frequency. From 0 to 6%, the pedestrian crash frequency decreases as the steep area proportion increases. This fact may be that many people do not walk in the steep areas. For bus stop density, there is a clearly non-linear effect on pedestrian crash frequency. Pedestrian crashes would be affected by a sharply increasing rate for the 0.15 bus stops per hectare, and then by a substantially decreasing rate for a density between 0.15 and 0.28 per hectare. After a slight increase between 0.28 and 0.45 per hectare, the effect of bus stop density reaches a stable state. This implies that areas with low bus stop density should be more carefully designed. Stop sign density has an expectedly negative effect on pedestrian crash frequency. Within its effective ranges (i.e., 0 to 1.5 stop signs per hectare), a higher stop sign density is more likely associated with fewer pedestrian crashes. Beyond this range, increasing the stop sign density does not have any effect on pedestrian crashes. In summary, these findings provide helpful policy guidance since effective ranges are clearly identified by the model.

Traffic demand factors (i.e., the number of trips, walking mode share) and road network factors (i.e., densities of 4-way intersections and 3-way intersections) show important effects on pedestrian crash frequency, and Fig. 6 displays their relationship to pedestrian crash frequency. Within the range of 5–17 thousand trips, the number of trips seems to have a positive effect on pedestrian crash frequency in a nearly linear way. After that value, its effect does not change. Walking mode share shows a clearly non-linear effect on pedestrian crash frequency. Its effect gradually decreases from 7% to 16% walking mode share and followed by an increasing trend for a share between 16% and 26%. Finally, it remains stable after 26%. This finding implies that policymakers and urban planners should pay more attention to areas with a walking mode share between 16% and 26%. Regarding road network features, the densities of 4-way intersections and 3-way intersections have similarly positive effects on pedestrian crash frequency. These findings are consistent with a previous study (Ukkusuri et al., 2012). Pedestrian crashes are more likely to occur at complicated intersection areas.

5. Conclusions

Using the dataset collected from the city of Seattle, Washington, a series of MAPRT models are built with various levels of tree complexity. This study contributes to identify the non-linear relationship between the built environment and automobile-involved pedestrian crash frequency. In addition, this model used in this study deepens our understanding through fitting complex relationships and accounting for interaction effects between explanatory variables. As observed, the performance and results of MAPRT models are affected by interactions among explanatory variables. By employing this method, subtle and sudden changes in pedestrian crash frequency are accurately captured. Third and most important, different from other machine learning algorithms (i.e., support vector machine, neural networks) as 'black-box' procedures, this approach provides an alternative way of sorting factors and identifies their non-linear effects. These results can provide helpful policy recommendations and safety improvement recommendations.

As noted from above modeling outcomes, built environment attributes tested in this study are identified to have a substantial effect in explaining pedestrian crash frequency. The collective contribution of household density, commercial land use, and land use mixture roughly accounts for 40% of all effects. Overall, their effects become stable after passing certain cut off points. This suggests that urban planners should pay more attention to pedestrian crashes in compact developed areas especially under the principles of smart growth and new urbanism. Additionally, the results also highlight the important role of zonal speed limits and 4-way intersection density in reducing pedestrian crashes. Thus, urging drivers to slow down may reduce pedestrian injuries and deaths. It is worth noting that this is only effective when speed limit is below 25 mph. Finally, since it is typically unreasonable to decrease the intersection density of cities, pedestrian safety may be enhanced by strategically building footbridges and underpasses.

Although the MAPRT model does not provide any significance for inferences, its output provides many helpful results. Policy makers can propose policy recommendations based upon the relative importance of these factors. What is more, recommended strategies could be varied by geo-spatial location. For example, the cutoff point between pedestrian crash frequency and land use mixture is identified. Policy makers and urban planners could draft different safety countermeasures in areas with different levels of mixing land use. Downtown Seattle and other communities could enhance safety by promoting different levels of mixed land use.

An assumption of this model is that all the included factors suffice to explain the response variable. The machine learning technique requires researchers to build a complete modeling framework. In addition, factor sorting is built upon an agreed modeling framework. Therefore, early works of conventional statistical models are the cornerstone in building factor sorting models. In such a context, statistical models are indispensable, and their results serve to improve the quality of machine learning results. Both approaches are mutually compatible. When facing a large data set and expecting computational efficiency, or in the demand of learning non-linear relationships among explanatory factors, the MAPRT approach is a reliable approach to explore data.

References

Abay KA, 2013. Examining pedestrian-injury severity using alternative disaggregate models. Res. Transp. Econ. 43 (1), 123–136.

Alliance for Biking & Walking, A, 2016. Bicycling and Walking in the United States – 2016 Benchmarking Report.

Aziz HMA, Ukkusuri SV, Hasan S, 2013. Exploring the determinants of pedestrian–vehicle crash severity in New York City. Accid. Anal. Prev. 50 (0), 1298–1309. [PubMed: 23122781]

Breiman L, Friedman J, Stone CJ, Olshen RA, 1984. Classification and Regression Trees. CRC press.

Bureau of Transportation Statistics, B, 2016. Passenger Travel Facts and Figures.

Cai Q, Lee J, Eluru N, Abdel-Aty M, 2016. Macro-level pedestrian and bicycle crash analysis: incorporating spatial spillover effects in dual state count models. Accid. Anal. Prev. 93, 14–22. [PubMed: 27153525]

Centers for Disease Control and Prevention, C, 2012. More People Walk to Better Health.

Chen P, 2015. Built environment factors in explaining the automobile-involved bicycle crash frequencies: A spatial statistic approach. Saf. Sci. 79, 336–343.

Chen P, Zhou J, 2016. Effects of the built environment on automobile-involved pedestrian crash frequency and risk. J. Transp. Health 3 (4), 448–456.

Chung Y-S, 2013. Factor complexity of crash occurrence: an empirical demonstration using boosted regression trees. Accid. Anal. Prev. 61, 107–118. [PubMed: 22975365]

Clifton KJ, Burnier CV, Akar G, 2009. Severity of injury resulting from pedestrian -vehicle crashes: what can we learn from examining the built environment? Transportation Res. Part. D: Transp. Environ. 14 (6), 425–436.

Dai D, 2012. Identifying clusters and risk factors of injuries in pedestrian–vehicle crashes in a GIS environment. J. Transp. Geogr. 24, 206–214.

Dai D, Taquechel EP, Steward J, Strasser S, 2010. The impact of built environment on pedestrian crashes and the identification of crash clusters on an urban university campus. West. J. Emerg. Med. 11 (3).

Ding C, Wu X, Yu G, Wang Y, 2016. A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data. Transp. Res. Part. C: Emerg. Technol. 72, 225–238.

Elith J, Leathwick JR, Hastie T, 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77 (4), 802–813. [PubMed: 18397250]

Friedman J, Hastie T, Tibshirani R, 2001. The Elements of Statistical Learning. Springer series in statistics New York.

Friedman JH, 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Geyer J, Raford N, Pham T, Ragland DR, 2006. Safety in numbers: data from Oakland, California. Transp. Res. Rec.: J. Transp. Res. Board 1982 (1), 150–154.

Guelman L, 2012. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst. Appl. 39 (3), 3659–3667.

Hastie T, Tibshirani R, Friedman J, 2009. The elements of statistical learning: Data mining, inference and prediction, second edition. Springer, New York.

Islam S, Hossain AB, 2015. Comparative analysis of injury severity resulting from pedestrian–motor vehicle and bicycle–motor vehicle crashes on roadways in Alabama. Transp. Res. Rec.: J. Transp. Res. Board 2514, 79–87.

Kim J-K, Ulfarsson GF, Shankar VN, Kim S, 2008. Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. Accid. Anal. Prev. 40 (5), 1695–1702. [PubMed: 18760098]

Miranda-Moreno LF, Morency P, El-Geneidy AM, 2011. The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. Accid. Anal. Prev. 43 (5), 1624–1634. [PubMed: 21658488]

Mohamed MG, Saunier N, Miranda-Moreno LF, Ukkusuri SV, 2013. A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Saf. Sci. 54 (0), 27–37.

Moudon A, Lin L, Hurvitz P, Reeves P, 2008. Risk of pedestrian collision occurrence: case control study of collision locations on state routes in King County and Seattle, Washington. Transp. Res. Rec.: J. Transp. Res. Board 2073 (−1), 25–38.

Moudon AV, Lin L, Jiao J, Hurvitz P, Reeves P, 2011. The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. Accid. Anal. Prev. 43 (1), 11–24. [PubMed: 21094292]

Narayanamoorthy S, Paleti R, Bhat CR, 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transp. Res. Rec.: J. Transp. Res. Board 55 (0), 245–264.

Prati G, Pietrantoni L, Fraboni F, 2017. Using data mining techniques to predict the severity of bicycle crashes. Accid. Anal. Prev. 101, 44–54. [PubMed: 28189058]

Puget Sound Regional Council, P, 2014. Activity-Based Travel Model: SoundCast. http://www.psrc.org/data/models/abmodel/.

Pulugurtha SS, Krishnakumar VK, Nambisan SS, 2007. New methods to identify and rank high pedestrian crash zones: an illustration. Accid. Anal. Prev. 39 (4), 800–811. [PubMed: 17227666]

Pulugurtha SS, Sambhara VR, 2011. Pedestrian crash estimation models for signalized intersections. Accid. Anal. Prev. 43 (1), 439–446. [PubMed: 21094342]

Saha D, Alluri P, Gan A, 2015. Prioritizing highway safety Manual's crash prediction variables using boosted regression trees. Accid. Anal. Prev. 79, 133–144. [PubMed: 25823903]

Schneider R, Diogenes M, Arnold L, Attaset V, Griswold J, Ragland D, 2010. Association between roadway intersection characteristics and pedestrian crash risk in Alameda County, California. Transp. Res. Rec.: J. Transp. Res. Board 2198, 41–51.

Schneider RJ, Grembek O, Braughton M, 2013. Pedestrian crash risk on boundary roadways. Transp. Res. Rec.: J. Transp. Res. Board 2393 (1), 164–173.

Siddiqui C, Abdel-Aty M, Choi K, 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes. Accid. Anal. Prev. 45 (0), 382–391. [PubMed: 22269522]

Smart Growth America & National Complete Streets Coalition, S.N, 2017. Dangerous by Design.

Tarko A, Azam MS, 2011. Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data. Accid. Anal. Prev. 43 (5), 1689–1695. [PubMed: 21658495]

The National Highway Traffic Safety Administration, N, 2015. Pedestrians.

Ukkusuri S, Hasan S, Aziz H, 2011. Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Transp. Res. Rec.: J. Transp. Res. Board 2237 (−1), 98–106.

Ukkusuri S, Miranda-Moreno LF, Ramadurai G, Isa-Tavarez J, 2012. The role of built environment on pedestrian crash frequency. Saf. Sci. 50 (4), 1141–1151.

Wang J, Huang H, Zeng Q, 2017. The effect of zonal factors in estimating crash risks by transportation modes: motor vehicle, bicycle and pedestrian. Accid. Anal. Prev. 98, 223–231. [PubMed: 27770688]

Wang Y, Kockelman KM, 2013. A poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accid. Anal. Prev. 60, 71–84. [PubMed: 24036167]

Wier M, Weintraub J, Humphreys EH, Seto E, Bhatia R, 2009. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. Accid. Anal. Prev. 41 (1), 137–145. [PubMed: 19114148]

Zahabi SAH, Strauss J, Manaugh K, Miranda-Moreno LF, 2011. Estimating potential effect of speed limits, built environment, and other factors on severity of pedestrian and cyclist injuries in crashes. Transp. Res. Rec.: J. Transp. Res. Board 2247 (1), 81–90.

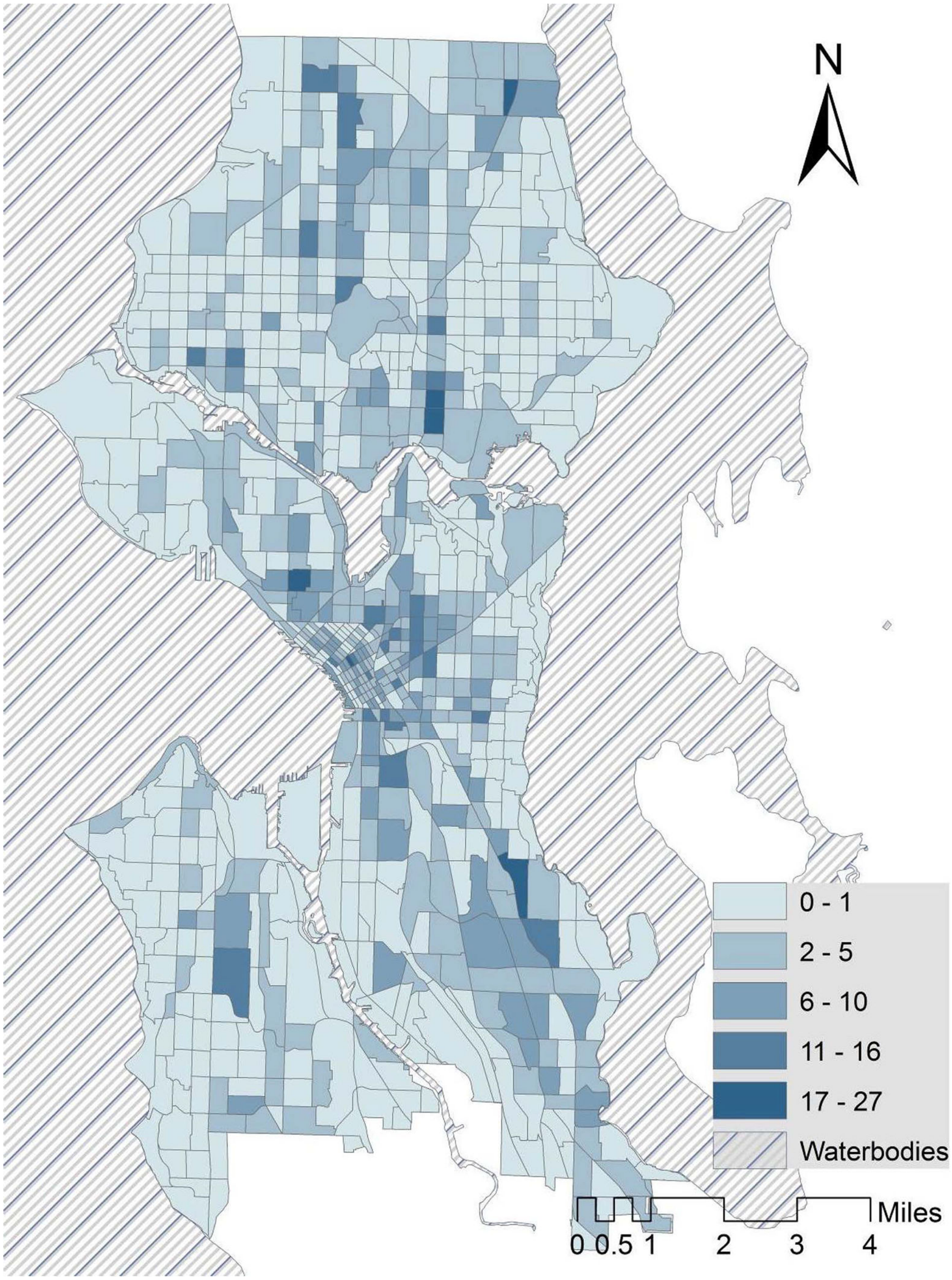

Fig. 1.

Pedestrian crash frequency in Seattle TAZs, 2008–2012.

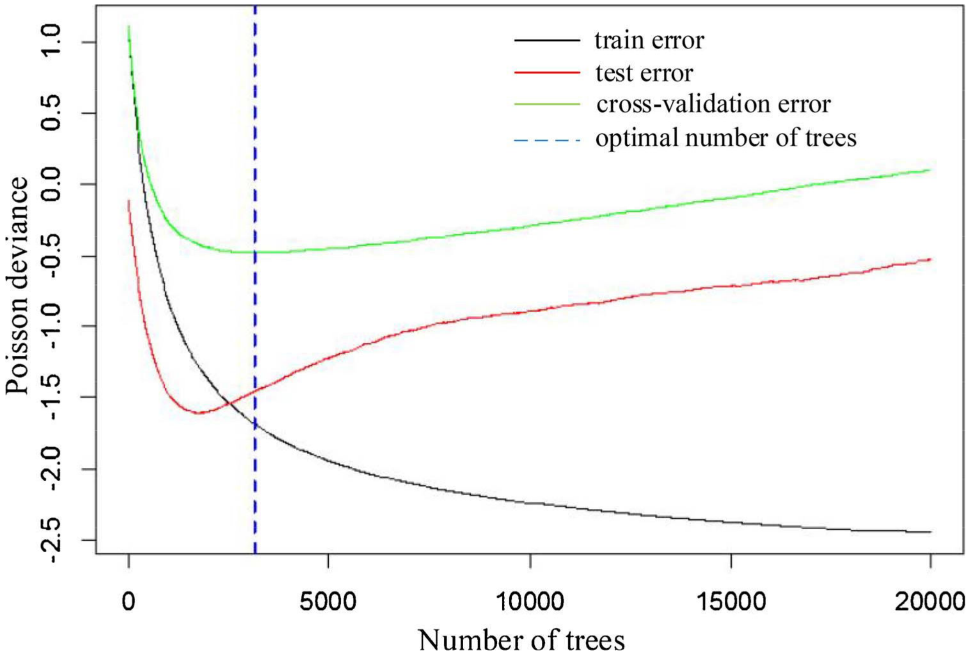

Fig. 2.

Relationship between predictive deviation and number of trees.

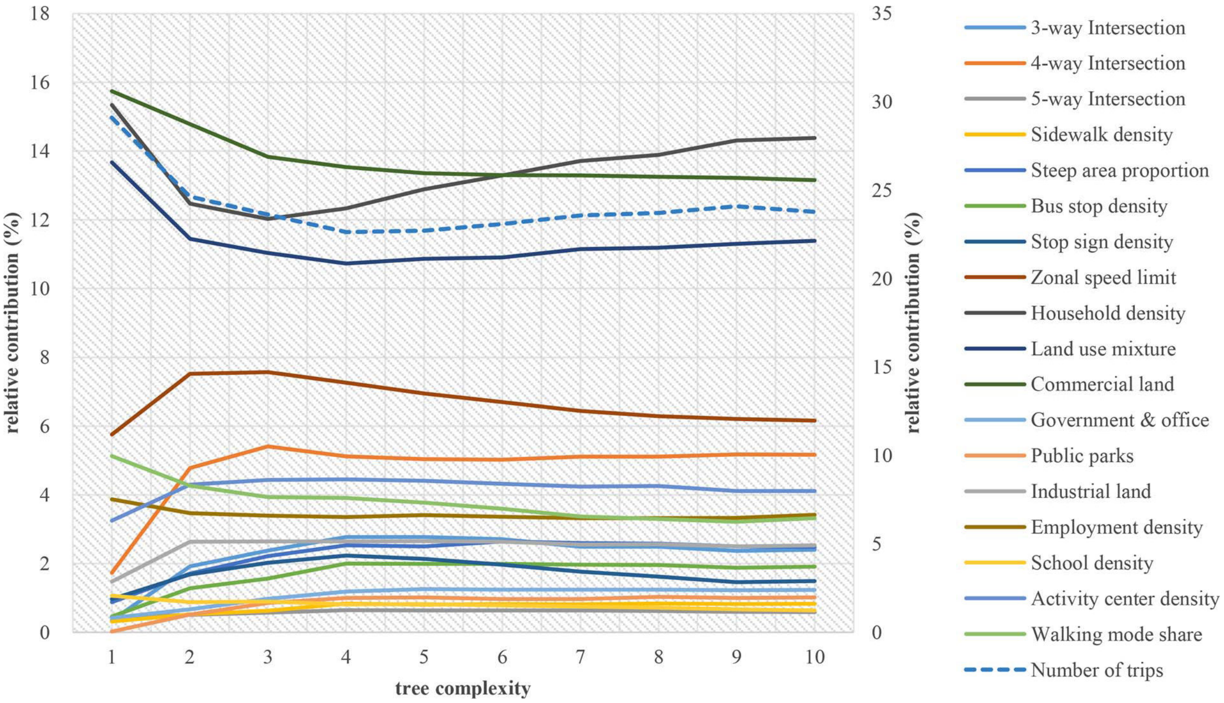

Fig. 3.

Relative contributions of explanatory variables with different levels of tree complexity.

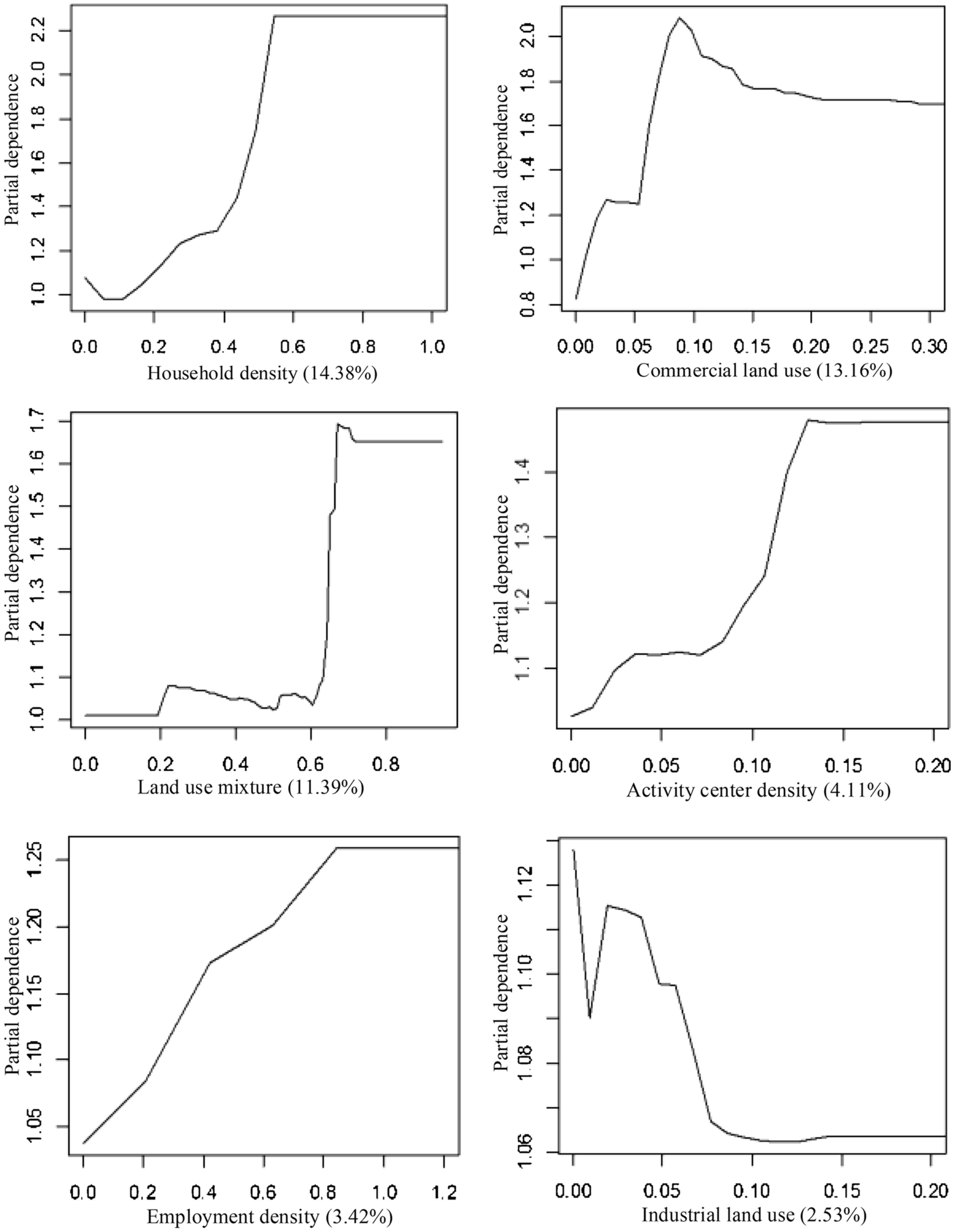

Fig. 4.

Non-linear effects of key land use variables on pedestrian crash frequency.

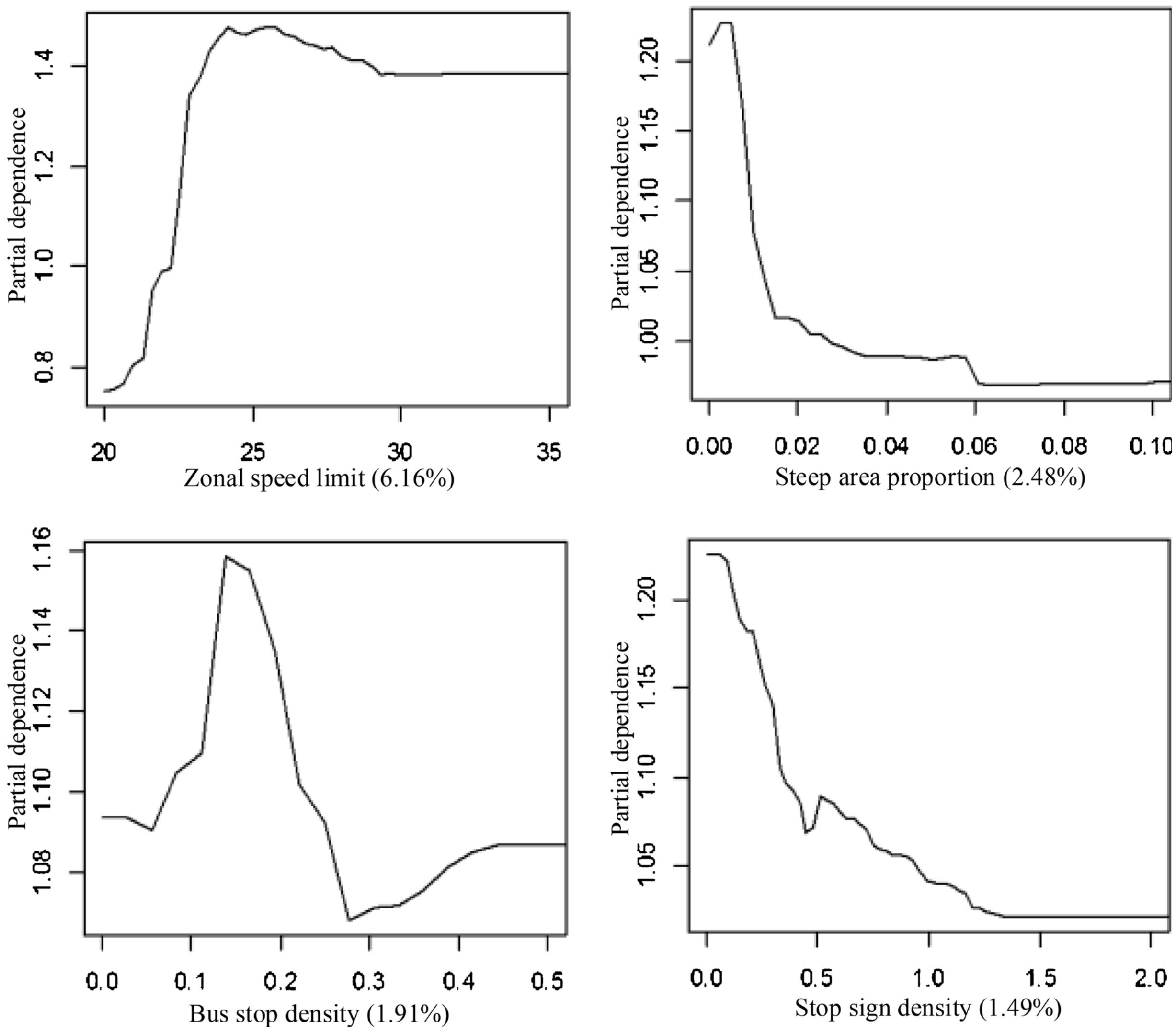

Fig. 5.

Non-linear effects of street elements variables on pedestrian crash frequency.

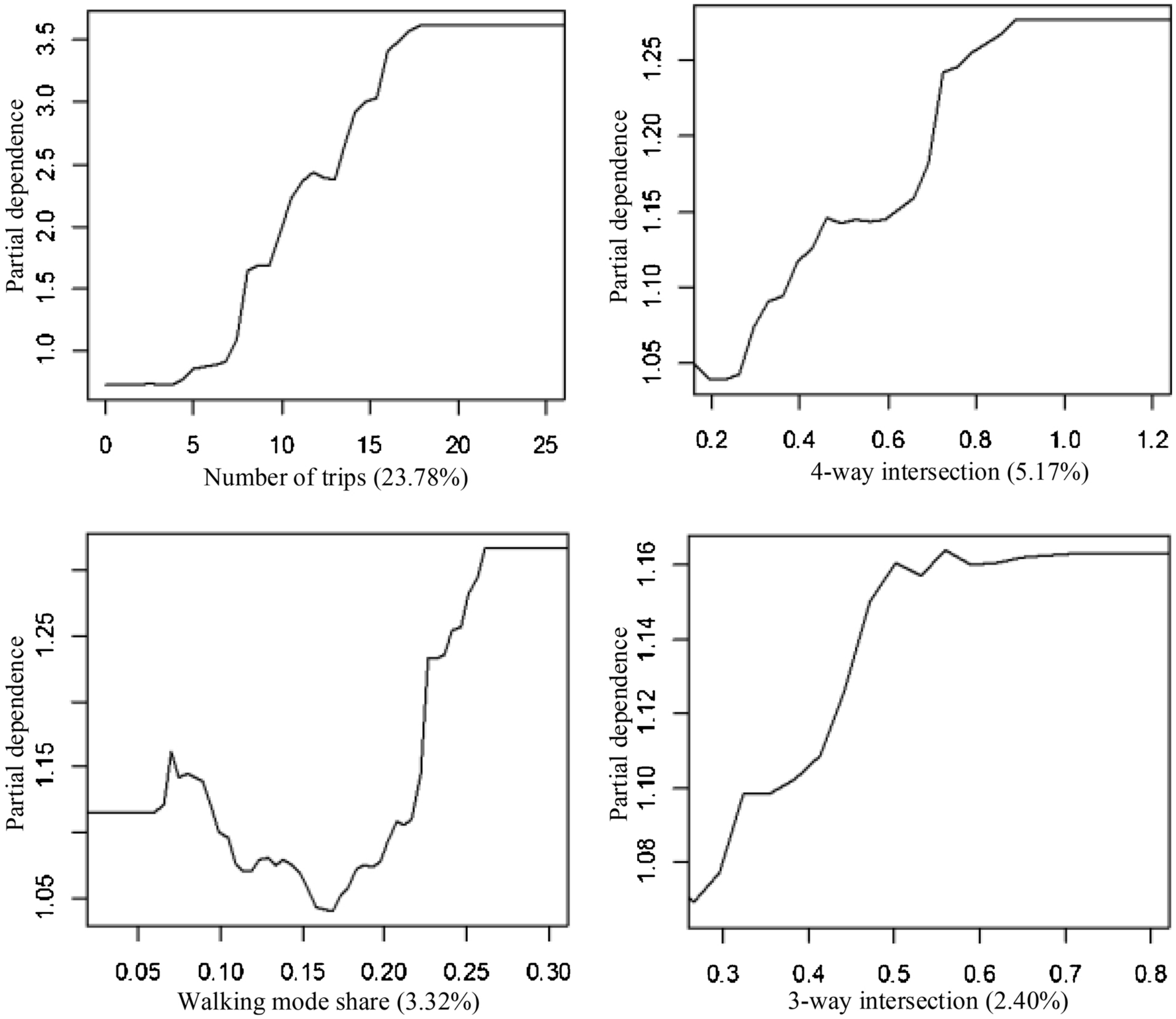

Fig. 6.

Non-linear effects of key explanatory variables on pedestrian crash frequency.

Table 1

Variable selection in existing pedestrian crash frequency studies.

Group	Variable	Studies
Roadway design	Road length or density of different road (sidewalks, local streets, and arterials) Intersections/crosswalks Bus stops/stations	(Cai et al., 2016; Chen and Zhou, 2016; Moudon et al., 2011; Narayanamoorthy et al., 2013; Siddiqui et al., 2012; Ukkusuri et al., 2011; Ukkusuri et al., 2012; Wang and Kockelman, 2013; Wier et al., 2009) (Cai et al., 2016; Chen and Zhou, 2016; Miranda-Moreno et al., 2011; Moudon et al., 2011; Pulugurtha and Sambhara, 2011; Siddiqui et al., 2012; Ukkusuri et al., 2011; Ukkusuri et al., 2012) (Chen and Zhou, 2016; Miranda-Moreno et al., 2011; Moudon et al., 2011; Pulugurtha and Sambhara, 2011; Ukkusuri et al., 2011; Ukkusuri et al., 2012)
Land use	Land use mix Number of parcels or percentages of different types of land use Zone size Number of schools / universities	(Chen and Zhou, 2016; Ukkusuri et al., 2011; Wang and Kockelman, 2013) (Chen and Zhou, 2016; Moudon et al., 2011; Narayanamoorthy et al., 2013; Pulugurtha and Sambhara, 2011; Siddiqui et al., 2012; Ukkusuri et al., 2012; Wang and Kockelman, 2013; Wier et al., 2009) (Moudon et al., 2011; Wang and Kockelman, 2013; Wier et al., 2009) (Cai et al., 2016; Miranda-Moreno et al., 2011; Moudon et al., 2011; Narayanamoorthy et al., 2013; Ukkusuri et al., 2011; Ukkusuri et al., 2012)
Demographics	Population/employment density Poverty level / median household income Senior citizen/ teenager/ children Race	(Cai et al., 2016; Chen and Zhou, 2016; Moudon et al., 2011; Narayanamoorthy et al., 2013; Pulugurtha et al., 2007; Ukkusuri et al., 2011; Ukkusuri et al., 2012; Wang and Kockelman, 2013; Wier et al., 2009) (Moudon et al., 2011; Narayanamoorthy et al., 2013; Siddiqui et al., 2012; Wier et al., 2009) (Narayanamoorthy et al., 2013; Wier et al., 2009) (Ukkusuri et al., 2011; Ukkusuri et al., 2012)
Travel demand	Walk miles travelled Pedestrian/ traffic volume Walking mode share / Number of commuters for a mode	(Wang and Kockelman, 2013) (Miranda-Moreno et al., 2011; Moudon et al., 2011; Pulugurtha et al., 2007; Pulugurtha and Sambhara, 2011; Wier et al., 2009) (Cai et al., 2016; Chen and Zhou, 2016; Narayanamoorthy et al., 2013)
Traffic control	Speed limit	(Cai et al., 2016; Chen and Zhou, 2016; Miranda-Moreno et al., 2011; Moudon et al., 2008; Moudon et al., 2011; Narayanamoorthy et al., 2013; Pulugurtha and Sambhara, 2011; Siddiqui et al., 2012; Ukkusuri et al., 2011; Ukkusuri et al., 2012; Wang and Kockelman, 2013; Wier et al., 2009)

Table 2

Description of variables (*N*=863).

Category	Variable	Description	Source
Crash	Pedestrian crash frequency	Number of pedestrian crashes	SDOT
Road network	3-way intersection	Number of 3-way intersections/zonal area (1/ha)	SDOT
	4-way intersection	Number of 4-way intersections/zonal area (1/ha)	SDOT
	5-way intersection	Number of more than 5-way intersections/zonal area (1/ha)	SDOT
	Sidewalk density	Sum of sidewalk length/zonal area (1km/km^2)	SDOT
Street elements	Steep area proportion	Proportion of steep areas (0~ 1)	PSRC
	Bus stop density	Number of bus stops/zonal area (1/ha)	King County
	Stop sign density	Number of stop signs/zonal area (1/ha)	SDOT
	Zonal Speed Limit	Zonal-mean posted driving speed limit (mph)	SDOT
Land use	Household density	Number of households/zonal area (1k/ha)	PSRC
	Employment density	Number of employments/zonal area (1k/ha)	PSRC
	Land use mixture	Entropy of five types of land use (0~1)	PSRC
	Commercial land use	Proportion of commercial and mixed land use	PSRC
	Government & office	Proportion of office and government land use	PSRC
	Public parks	Proportion of public parks	SDOT
	Industrial land use	Proportion of industrial land use	SDOT
	School density	Number of schools/zonal area (1/ha)	SDOT
	Activity center density	Number of activity centers/zonal area (1/ha)	SDOT
Traffic demand	Walking mode share	Proportion of pedestrian trips divided by total number of trips (0~ 1)	PSRC
	Number of trips	Total number of trips (1k)	PSRC

Table 3

Relative contributions of explanatory variables on pedestrian crash frequency.

Category	Variable	Automobile-involved pedestrian crash frequency	
		Rank	Relative importance (%)
Road network	3-way intersection	12	2.40
	4-way intersection	6	5.17
	5-way intersection	19	0.59
	Sidewalk density	17	0.83
Street elements	Steep area proportion	11	2.48
	Bus stop density	13	1.91
	Stop sign density	14	1.49
	Zonal speed limit	5	6.16
Land use	Household density	2	14.38
	Employment density	8	3.42
	Land use mixture	4	11.39
	Commercial land use	3	13.16
	Government & office	15	1.23
	Public parks	16	1.02
	Industrial land use	10	2.53
	School density	18	0.64
	Activity center density	7	4.11
Traffic demand	Walking mode share	9	3.32
	Number of trips	1	23.78

Note: sample size = 863.