Please cite the paper as:

Wu, W., Jiang, S, Liu, R, Jin, W and Ma, C (2019) Economic development, demographic characteristics, road network and traffic accidents in Zhongshan, China: Gradient boosting decision tree model. **Transportmetica A: Transport Science**. In press.

Economic development, demographic characteristics, road network and traffic accidents in Zhongshan, China: Gradient boosting decision tree model

Weitiao Wu[a1], Shuyan Jiang[a], Ronghui Liu[b], Wenzhou Jin[a], Changxi Ma[c]

a. School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510641, China

b. Institute for Transport Studies, University of Leeds, Leeds LS29JT, United Kingdom

c. School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China

ABSTRACT

This paper explores the joint effect of economic development, demographic characteristics and the road network on regional road safety. Although extensive efforts have been undertaken to model and predict the safety effects of different influential factors using statistical regression or machine learning models, little evidence is provided on the relative importance of explanatory variables by accounting for their mutual interactions and non-linear effects on traffic accidents. We present an innovative gradient boosting decision tree (GBDT) model to explore the joint effects of these comprehensive factors on four traffic accident indicators (i.e., the number of traffic accidents, injuries, deaths, and the economic loss). A total of 27 elaborated influential factors associated with the economic, demographic and road network conditions in Zhongshan, China for the period of 2000-2016 are collected. The results show that, compared to other traditional machine learning methods, the GBDT not only presents a higher prediction accuracy, but can also better handle the multicollinearity between the explanatory variables; more importantly, it can rank the influential factors on traffic accident prediction. The results also show that there are both similarities and differences in the key influential factors for the four traffic accident indicators. In particular, we also investigate the partial effects of the key influential factors. Based on the key findings, we highlight the practical insights for planning practice.

Keywords: Traffic accidents; Socio-economic; Demographics; Relative importance; Gradient boosting decision tree; Partial effect

[1] Corresponding Author: Tel: (+86)13763327125; Email: ctwtwu@scut.edu.cn

1 Introduction

Traffic accidents result in significant economic losses, which account for a certain proportion of a country's gross national product (WHO, 2017). The crash occurrence causes enormous human, economic and social losses. Given this fact, road safety has long been a serious issue that is closely related to health and development, which requires agencies to take effective comprehensive measures. Thus, both transportation and economic researchers have attached great importance to the risk factors influencing road safety.

In this context, extensive efforts have been undertaken to investigate the influential factors of traffic accidents. Among them, human factors (e.g., driving behaviour) (Dong et al., 2014; Hong et al., 2015; Huang et al., 2014; Zheng et al., 2017, 2018; Ma et al., 2018), vehicle characteristics (e.g., motorcycle, trunk) (Manan et al., 2013; Law, 2015; Moomen et al., 2018), and the road environment (Ha and Thill., 2011; Chen et al., 2011) have been profoundly proven to have a strong correlation with traffic accidents. Apart from studies that aim to explore influential factors at a microscopic level, some other researchers have investigated how macroscopic socio-economic conditions affect traffic accidents, such as the Gross Domestic Product (GDP) growth (Bener et al., 2011; Yannis et al., 2014), the population and vehicle ownership (Van et al., 2004; Kopits et al., 2005; Garg et al., 2006; Ziyab et al., 2012).

Most of the existing literature on the analysis of traffic accidents, however, only considers limited number or sides of influential factors (see also the detailed analysis in Section 2). Traffic accidents happen as a result of intertwined factors; economic development, demographic characteristics, and the road network are key elements for understanding the regional social conditions. Their mutual interaction is prevalent and important for various fields, including urban planning, transport, and social science. For example, socio-demographics is closely correlated to mobility patterns and economic development (Wissen and Huisman, 2002). Socio-economic factors are highly related to vehicle ownership and potential crash risks. With the expectation that a variety of socio-economic, demographic and road network factors will macroscopically exert an influence on traffic safety, there is an imminent need to identify how the revolution of these variables jointly affect road safety. However, to date the understanding of the joint effects of the comprehensive factors on road safety is still limited. In this study, we select 27 underlying influential factors, comprehensively covering the aspects of economic development, demographic characteristics, and the road network. Using the case of Zhongshan, China over the period of 2000-2016, we examine the contributions of explanatory variables to four traffic accident indicators, including the number of traffic accidents, injuries, deaths, and the economic loss. This study will help planners and decision-makers evaluate the wider socio-economic impacts of traffic accidents in combination with conventional impacts, and take effective countermeasures to improve road safety in a holistic manner. However, as the number of influential factors increases, the mutual interaction effects become remarkable, which we term as a "multicollinearity phenomenon". This poses challenges in modelling the complex non-linear relationships while disentangling the interaction effects between the influential factors.

Methodologically, there are mainly two groups of approaches: statistical regression models and machine learning. The former includes the negative binomial regression model (Milton and Mannering, 1998; Li et al., 2014), the vector auto regression model (Wiebe et al., 2016), the cubic regression model (Moomen et al., 2018), the logistic regression model (Pugachev et al., 2017), multiple logistic regression (Hong et al., 2015), etc. In contrast to statistical methods based on prior assumptions of the input data, the principle of machine learning methods is to construct a nonlinear relationship between the input and output variables without any prior knowledge. Among them, artificial neural network (ANN) models have been widely applied in many fields of transportation since they are able to handle the complex relationship in datasets (Huang et al., 2017). The drawback of ANNs, however, is the local minima and parameterization problem, as well as the potential occurrence of over-fitting. In addition, a large number of training samples are needed in order to give a good generalization performance. Support vector machine (SVM) models are another type of machine learning method. SVM models can overcome the shortcomings of ANNs, and deal with the problems of nonlinearity, high dimensions, local minima and over-fitting (Dong et al., 2015).

Unfortunately, the relative importance of the influential factors cannot be explicitly evaluated for the models mentioned above. Understanding the contributions of the influential factors to traffic accidents facilitates the prediction and future improvement of traffic safety. Moreover, identifying and ranking the influential factors could help cost savings, since data acquisition and maintenance are usually expensive. Although a sensitivity analysis could be used for identification, its drawback is that it exclusively evaluates one variable each time, and assumes that the other variables remain constant. As such, the complex mutual interaction effect among the influential factors is missing.

To solve this conundrum, this study presents a new data mining method, namely, the gradient boosting decision tree (GBDT) model, to examine the effects of these factors on traffic safety. The results show that the GBDT model can effectively fit complex non-linear relationships among variables, and that it outperforms traditional machine learning methods featuring a “black-box” in both prediction accuracy and when handling multicollinearity. More importantly, the GBDT can identify and prioritize the influential factors on traffic accident prediction. In particular, we investigate the marginal effects of key factors. Based on the key findings, we highlight the practical insights for planning practice.

The remainder of this paper is organized as follows. In Section 2, we review the relevant literature. In Section 3, the modelling approach is introduced. In Section 4, the data and variables are presented. Section 5 presents the results and analyses the important influential factors. The final section concludes the key results and highlights the implications for planning practice.

2 Literature review

Since the 1990s, a number of researchers have started to pay attention to traffic accidents. Many efforts have been made to identify and analyse the factors influencing road crashes and their severity. In addition to microscopic factors, such as road environment (e.g., Ha and Thill., 2011) and driving behaviour (e.g., Huang

et al., 2014; Zheng et al., 2017, 2018), some researchers have explored factors affecting road safety at the macro-level. Vasconcellos (1995) concluded that traffic accidents are comprehensively affected by a wide variety of socio-economic factors. There are mainly three groups of macroscopic factors in the literature: economic development, demographic characteristics, and road network conditions.

2.1 Effects of economic development

Among various factors, economic factors have always been a prominent research area in the analysis of traffic accidents (Delmelle et al., 2012; Wiebe et al., 2016; Sakhapov et al., 2017). García-ferrer et al. (2007) investigated the relationships between traffic accidents and economic activities over several years in Spain. The most commonly used economic factor is GDP per capita (Garg et al., 2006; Iwata, 2010). Bener et al. (2011) analysed the relationships between the GDP per capita, population and collisions using a linear regression model. Bougueroua et al. (2016) explored the relationships between GDP per capita, fuel consumption and collisions in Algeria.

In addition to GDP per capita, some literature also exists that considers the effect of the overall GDP (Zhao, 2009; He et al., 2015; Law et al., 2015). Yannis et al. (2014) built up a relationship between the GDP and the number of traffic deaths. Hughes et al. (2014) analysed the effects of the GDP and freight volume on traffic accidents. Lu et al. (2016) suggested that both freight volume and passenger volume play important roles in road safety.

Vehicle ownership is another indication of economic development. Chang et al. (2013) investigated how factors associated with the driver, vehicle and environmental characteristics affect traffic accidents. Goel et al. (2016) found that heavy vehicles (e.g., trucks, canter and buses), which represent only a small proportion of the overall traffic, are involved in as high as 42% of traffic accidents. Assemi et al. (2018) reported that traffic accidents and serious economic loss are more likely to arise in the presence of heavy vehicles.

Apart from car accidents, a number of works have focused on the behavioural characteristics of trucks and the influential factors of truck accidents (Chang et al., 2013; Hao et al., 2016; Moomen et al., 2018). To improve the accuracy of different collision prediction models, Juneyoung et al. (2018) classified traffic accidents while considering truck crashes. In addition to trucks, the effect of motorcycles has also been considered in the analysis of traffic accidents (Dapilah et al., 2016).

In reality, vehicle ownership also contributes to the traffic accident frequency (Van et al., 2004; Bishai et al., 2006; Garg et al., 2006; Pugachev et al., 2017). For example, Ziyab et al. (2012) reported that vehicle ownership contributed differently to traffic accidents. In view of this, in this study the effects of various kinds of vehicle ownership are considered, such as truck, motorcycle, and car ownership. Moreover, we make a further classification for both cars and trucks.

2.2 Effects of demographic characteristics

Delmelle et al. (2012) made the prediction for crash frequency considering not only economic factors but also demographic characteristics. Albalate et al. (2010) analysed the relationships between demographic characteristics, the environment, traffic conditions and motorcycle traffic accidents via a multinomial logistic

regression model. Unfortunately, to date there is little literature that systematically assesses the impact of various vehicle types and demographic characteristics on road safety.

In the literature, another influential factor is population (Albalate et al., 2010; Bener et al., 2011; Ziyab et al., 2012; Wiebe et al., 2016). Law et al. (2015) studied the relationships between the population, GDP, motorcycle ownership and the number of traffic injuries. In their study, the population was further divided into the total population and urban population. In reality, since the resident population and mobile population can reflect the regional mobility characteristics, their provisions may exert great influences on the effects of enforcement and educational measures. Therefore, in this study we also considered the impact of resident population and mobile population, in addition to the total population and urban population.

2.3 Effects of road networks

Ren et al. (2013) took both the total mileage of the standard highway and the substandard highway into account. Later, Iwata (2010) analysed the effect of economic development on traffic accidents, which proved that highway mileage was an important factor. In contrast to previous studies, we enhance the influential factors by additionally considering the impact of different levels of standard highways in addition to the total mileage on traffic accidents.

2.4 Summary and contributions

In summary, the aforementioned factors could be divided into three categories, namely, economic development, demographic characteristic, and the road network. The economic development factors include GDP per capita, GDP, freight volume, passenger volume, vehicle types and ownership. In this regard, two main research gaps are identified: First, different industries present their own characteristics in terms of regional development, such that their effects on road safety are different. Therefore, it would be more rational to select primary industry, secondary industry, and the tertiary industry as the influential factors. Second, each of the grouped vehicle categories is endowed with several vehicle types. The crash risks of these vehicle types may also differ. More elaborated classification may provide deeper insight into the relationship between the traffic composition and road safety.

In terms of demographic characteristic factors, the existing literature primarily focuses on the total population. However, the population generally presents a complex composition, such as a registration population, urban population and suburban population. Meanwhile, different types of population present different characteristics in terms of travel patterns and traffic activities. Thus, it is imperative to further refine the population composition in the analysis of traffic accidents. More importantly, in this study the number of drivers is considered as another influential factor since it is closely linked to the number of traffic activities.

Regarding the road network factors, the most commonly used indicators are the total mileage and highway mileage. According to the functionality and traffic demand levels, the roadways can be divided into five classes: expressway, arterial highway, secondary highway, tertiary highway, and standard highway. Any roadways besides these five standard classes are called substandard highways, such as a variety of rural trails with poor transportation facilities. Distinctive driving behaviour and traffic accident characteristics usually

involve different classes of roads. Therefore, in this study we enhance the previous works by further considering the impact of various road classes on traffic accidents.

Generally, the evaluation indicators of road traffic accidents are the number of traffic accidents (García-ferrer et al., 2007; Iwata et al., 2010; Bougueroua et al., 2016), the number of injuries (Garg et al., 2006; Ziyab et al., 2012; Yannis et al., 2014) and the number of deaths (Bishai et al., 2006; Sheng et al., 2018), whereas the indicator of economic loss is rarely considered in the literature. Since there is a direct relationship between economic loss and regional economic development, the economic loss could be utilized to access situations involving traffic accidents (Connelly et al., 2006). These indicators mentioned above can not only be used to access road safety in the same area, but also for a comparison in different periods. As such, this paper also includes the economic loss as a new indicator.

3 Methodology

This study adopts a new machine learning method called the gradient boosting decision tree (GBDT) model. The GBDT model was first proposed by Friedman (2001) and it is based on the integration of statistical and machine learning methods. More specifically, the model is a combination of the gradient algorithm, boosting algorithm and decision trees algorithm. In what follows, we introduce the principle of each component.

3.1 Gradient algorithm

To achieve an accurate prediction result, in the GBDT model the loss function will be minimized step by step. To this end, the Gradient algorithm is used to calculate the gradients of numerical descent. The negative direction of the gradient refers to the direction where the loss function decreases the most. The negative gradient of the loss function is used as an approximation of the residual in the regression problem algorithm. The main steps are described as follows:

Step 1: randomly choose an initial point W_0.

Step 2: calculate the gradient d_i at the current point, according to the loss function and the optimized parameter W_i.

Step 3: select the size of one step ρ.

Step 4: update the parameter $W_{i+1} = W_i + \rho d_i$.

Step 5: repeat steps 3 and 4 until the stop condition is met. The stop condition is generally that the gradient is equal to zero, or the reduction in the loss function is less than a threshold value.

3.2 Bagging and boosting methods

Bagging and boosting are two population ensemble techniques. The boosting method is used to improve the accuracy of weak classification. It constructs a series of prediction functions and combines them in a certain way. There are some differences between the boosting and bagging methods. The bagging method obtains a number of learners via multiple sampling and training, and then obtains the final learning results

by averaging a series of learners. During the sampling process, samples can be selected repeatedly.

Fig. 1 shows the training process of bagging. The training set is randomly extracted from the original sample set $\{(X^{(t)}, y^{(t)})\}_{t=1}^{m}$, which is composed of m samples. Using the bootstrapping method, t training samples are extracted from the original sample set at each time, and then the selected training samples will be put back after this training. The probability of extracting each sample at each training is $\frac{1}{m}$. Thus, some samples in the training set may be extracted multiple times, while others may not be drawn at all. After training b times, b sample sets are selected. The regression coefficients φ_j of each characteristic variable can be calculated during the training process. Subsequently, the final function $f(x)$ can be obtained by adding and averaging.

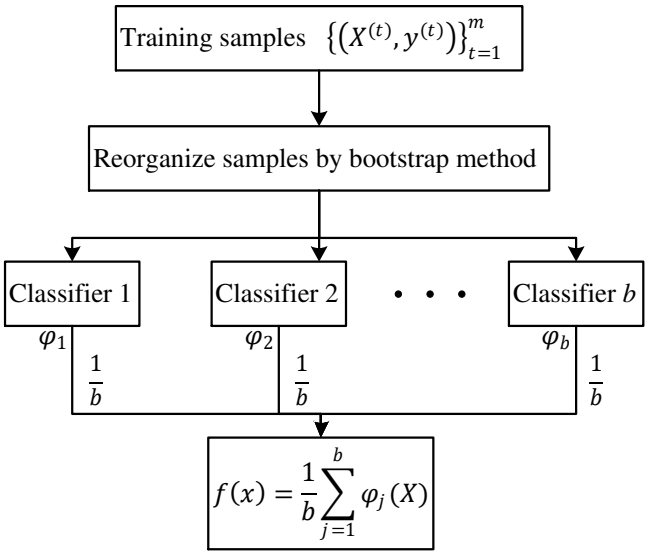


Fig. 1. The training process of bagging

Fig. 2 shows the training process for boosting. Boosting also generates a number of learning machines, but the training process depends more on previous calculations. Different from bagging, the boosting method does not necessarily undertake the process of random sampling. After each calculation iteration, the weight of each feature will be adjusted according to the calculation result, and the residual will be used for the next iteration. Before the final addition calculation, each feature is multiplied by the regression coefficient φ_j and the corresponding weight coefficient θ_j.

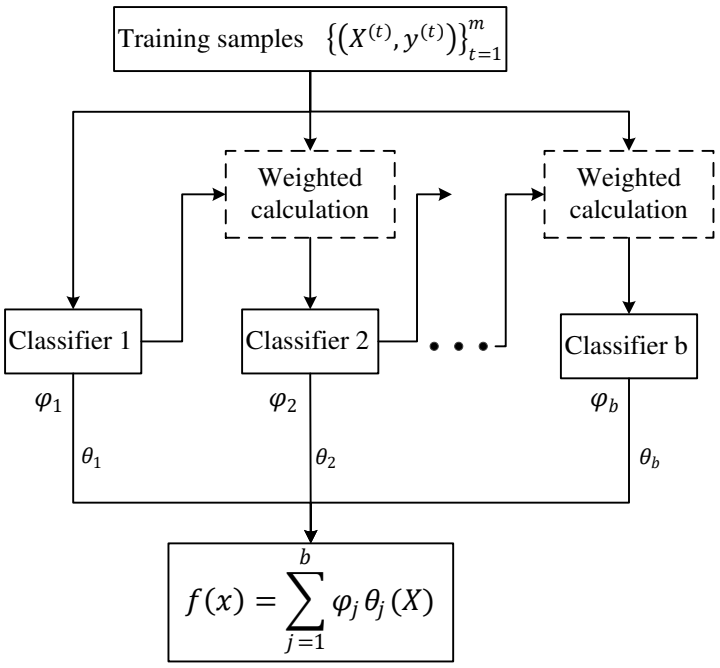


Fig. 2. The training process of boosting

3.3 Decision tree algorithm

The decision tree algorithm, which displays a tree structure, is a kind of function space. Each internal node in the tree represents a test on an attribute. Each branch the represents a test output, and each leaf node represents a category. The single decision tree algorithm is one of the most important components of the GBDT model, since both the gradient and boosting methods are realized at the tree-based structure.

The principle of a single decision tree is illustrated in Fig. 3. The function space can be defined as the dependent variable Y, while X_1 and X_2 can be defined as independent variables. To begin, the overall space can be divided into two parts according to the limitation factor k_1. Then the splitting process will be continued until the stopping rule is reached. As we can see, there are five regions (R_1, R_2, R_3, R_4, and R_5) and four split-points (k_1, k_2, k_3, and k_4) in the final space.

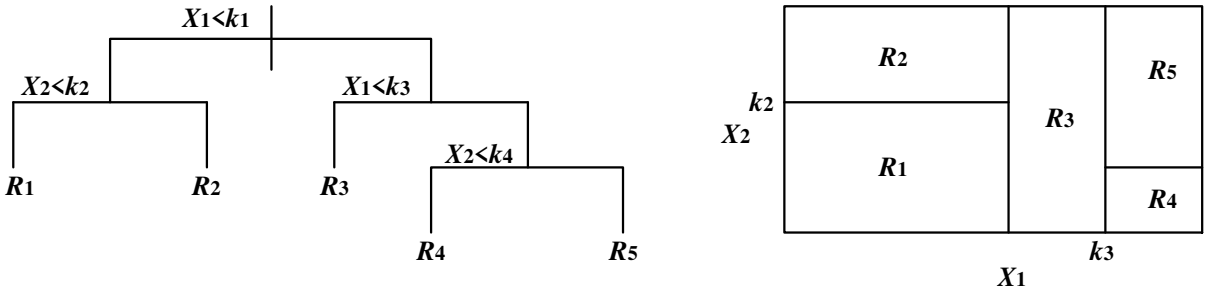


Fig. 3. A single decision tree

3.4 Gradient boosting decision trees (GBDT)

The gradient boosting decision tree model is a combination of the algorithms mentioned above in a linear manner. Classification and regression are achieved by continuously reducing the residuals during the training process. Fig. 4 shows the training process of the GBDT with multiple iterations. In each iteration the model will produce a weak classifier, and each classifier will be trained based on the residuals of the previous classifier. The weak classifiers are required to be sufficiently simple with a low variance and a high deviation. The forecasting accuracy of the classifier is then improved through a variance reduction. The classification and regression tree (CART) algorithm is one of the most widely used decision tree regression models, and is generally used as a weak classifier. Due to the requirements of having a simple structure and high deviations, the depth of each classification regression tree should be limited. The final ensemble model is obtained by the weighted summation of all the weak classifiers.

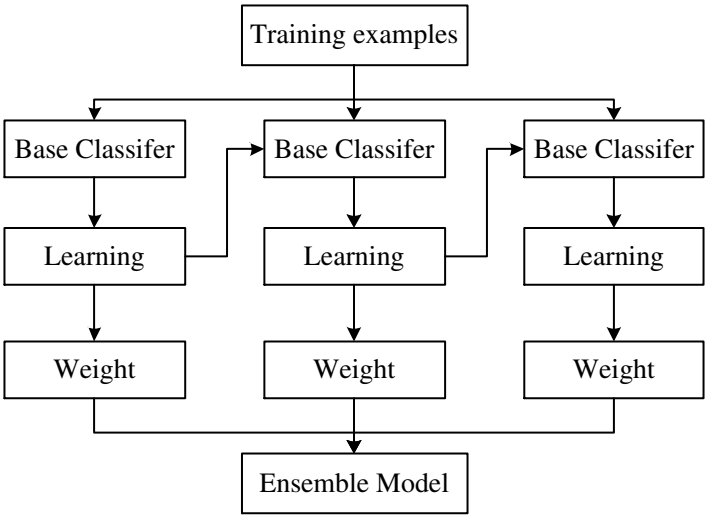


Fig. 4. The training process of the GBDT

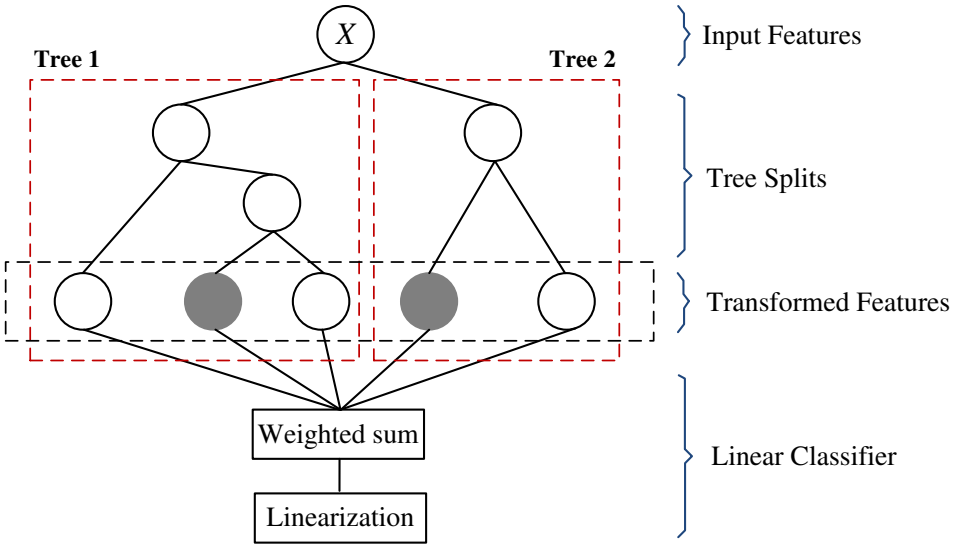


Fig. 5. The feature construction process of the GBDT

To illustrate the feature construction process of the GBDT, as shown in Fig. 5, let us consider two trees with five leaf nodes generated by the GBDT. Let the input feature X fall into the second leaf node of the first tree and the first leaf node of the second tree. Then a five-dimensional feature vector can be constructed,

where each dimension represents a leaf node. If the feature falls into the leaf node, the value of the leaf node is 1; otherwise, the value is 0. Thus, as for this example, we can obtain a vector [0, 1, 0, 1, 0] as the combined feature, which will be used in the logistic regression training process. It has been proven that these combined features contribute to improving the prediction accuracy.

Let X denote a set of explanatory variables (i.e., economic development, demographic characteristic and the road network) and $F(x)$ be an approximation function of the response variable y (i.e., traffic accidents). The main steps of the GBDT can be expressed as follows (Ding et al., 2016; Ding et al., 2017; Ding et al., 2018; Hastie et al., 2008; Zhang et al., 2015):

First, initialize the learning machine by the following equation:

$$F_0(x) = argmin_\rho \sum_{I=1}^{N} L(y_i, \rho) \quad (1)$$

where ρ is the estimated constant value that minimizes the loss function, and $L(y_i, \rho)$ is the mean square error loss function.

Second, use the negative gradient of the loss function in the current model as an approximation of the residual. The calculation process of the residual is shown as follows:

$$For\ \ m = 1\ \ to\ \ M\ \ do: \qquad \overline{y_i} = -\left[\frac{\partial L(y, F(x_i))}{\partial F(x_i)}\right] F(x) = F_{m-1}(x), i = 1, N \quad (2)$$

Third, generate a regression tree with J leaf nodes, which can be described as follows:

$$\{R_{jm}\}_1^J = J - terminal\ \ node\ \ tree\left(\{\overline{y_i}, x_i\}_i^N\right) \quad (3)$$

Fourth, estimate the value of the leaf nodes in the regression tree. The value can be estimated by the following equation:

$$\gamma_{jm} = argmin_\gamma \sum_{xi \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma) \quad (4)$$

Then, the learning machine of this iteration can be obtained, as shown in Eq. (5):

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}) \quad (5)$$

Finally, after a number of iterations, the final decision model can be described as follows:

$$F(x) = F_m(x) = \gamma + \sum_{m=1}^{M} \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm}), where\ \ I(x \in R_{jm}) = \begin{cases} 1, & x \in R_{jm} \\ 0, & x \notin R_{jm} \end{cases} \quad (6)$$

3.5 Relative importance of influential factors

Unlike other machine learning methods (e.g., random forest (RF), support vector regression (SVR), and long short-term memory networks (LSTM)), the GBDT is able to evaluate the relative importance of the independent variables on traffic accidents. For a collection of decision trees $\{T_m\}_1^M$, the importance of feature k can be obtained by its average over all additive trees. The relative importance of each influential factor can be formulated as follows:

$$I_k^2 = \frac{1}{M}\sum_{m=1}^{M} I_k^2(T_m) \quad (7)$$

where M represents the number of trees, and the importance of the feature k in a single tree can be described as follows:

$$\widehat{I_k^2}(T_m) = \sum_{m=1}^{L-1} i_t^2 I(v_t = k) \quad (8)$$

where L represents the number of leaf nodes, $L-1$ represents the number of non-leaf nodes, v_t represents the feature associated with node t, and $\widehat{I_k^2}$ represents the reduction of the squared loss after splitting.

3.6 Performance measures

In this study, the mean absolute percentage error (MAPE) is used as the evaluation indicator of the prediction accuracy. MAPE is defined as the mean value of the average percentage error of various predicted values. The formulation is expressed as follows:

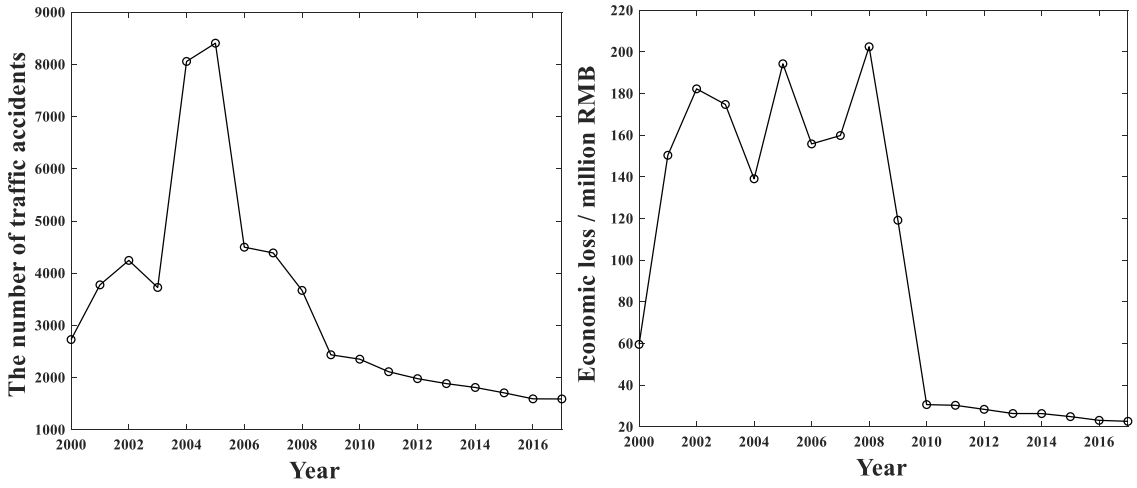
$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{O_i - \hat{O}_i}{O_i} \right| \quad (9)$$

where n represents the total number of samples, O_i represents the actual value of the dependent variable of the i th sample, and $\hat{O}_i$ represents the predicted value of the corresponding dependent variable.

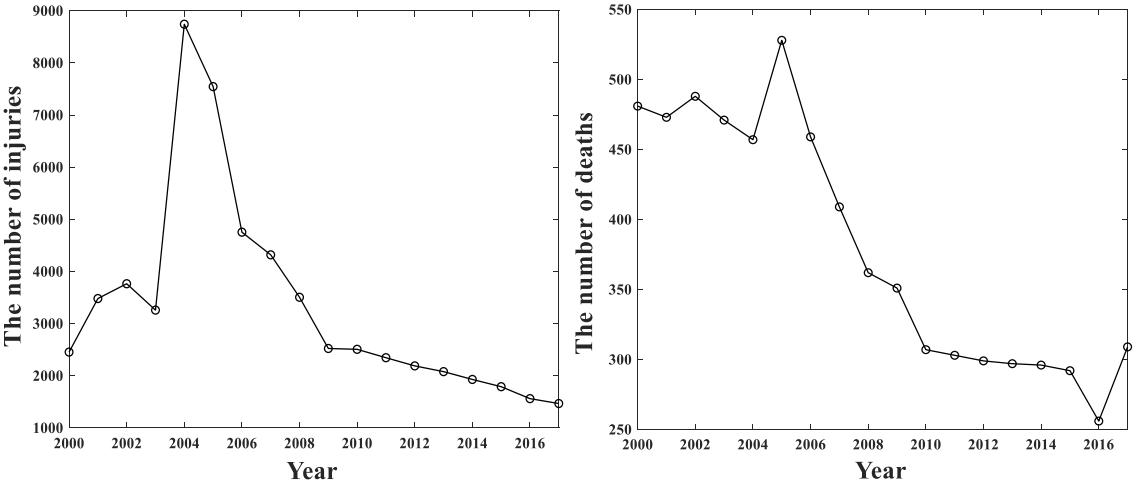
4 Data and variables

As a prefecture-level city governed by Guangdong Province, Zhongshan has a population of over 3 million. It is situated in the central south region of the Pearl River Delta, and borders Guangzhou to the north, Zhuhai to the south and Jiangmen to the west, in addition to adjoining with Hong Kong and Macao. In recent years, the local road network expansion has made great achievements. In 2016 there was more than 2600 km of roads, while the mileage of expressway reached 166.6 kilometers in total. In 2016, the number of traffic accidents decreased by 6.73%, from 1707 in the previous year to 1592. The number of injuries, the number of deaths, and the economic loss also are decreased by 12.79%, 12.33% and 7.17%, respectively. The four traffic accident indicators have decreased significantly since the year 2008, which suggests a continuous improvement of traffic safety due to traffic regulation actions.

In this study, the data of four traffic accident indicators and 27 underlying influential factors in Zhongshan during the period of 2000-2016 are collected to build the models. Most data (e.g., vehicle ownership) is collected from the local police office, while the other data is from the Statistical Yearbook. The number of traffic accidents, economic loss, number of injuries, and number of deaths over the studied period are presented in Fig. 6. These variables will be used as the dependent variables in the models, while the selection of influential factors covers economic development, demographic characteristic and road network factors. The key information is shown in Table 1.



(a) The number of traffic accidents (b) Economic loss



(c) The number of injuries (d) The number of deaths

Fig. 6. The development trends of the four traffic accident indicators

Table 1 Descriptions and statistics of variables

	Variables	Variable Description	Mean	Std.
Traffic accident indicators	The number of traffic accidents	Traffic accident count per year	3492.18	1984.48
	Economic loss	The direct economic loss per year	101.66	69.311
	The number of injuries	The number of injuries caused by traffic accidents per year (10^5 RMB)	3455.88	1932
	The number of deaths	The number of traffic deaths per year	384.06	87.27
Economic development	Gross Domestic Product (GDP)	Gross domestic product per year (10^4 RMB)	15523986	9523069.4
	GDP Per capita	Gross domestic product per capita (RMB)	56404	24415.67
	Primary industry	The output value of the primary industry (10^4 RMB)	447495.24	161997.23

	Secondary industry	The output value of the secondary industry (10^4 RMB)	8772262.3	5058220.1
	Tertiary industry	The output value of the tertiary industry (10^4 RMB)	6302473.7	4328825.7
	Passenger volume	Passenger traffic count per year (10^4)	10228.3	7083.09
	Freight volume	Freight traffic count per year (10^4 ton)	9348.98	5542
	Car ownership	The number of registered cars	251010.76	203406.87
	Medium bus ownership	The number of registered medium bus	4869.12	1741.48
	Bus ownership	The number of registered buses	3500.88	1486.22
	Van ownership	The number of registered vans	67789.06	18351.05
	Medium truck ownership	The number of registered medium trucks	8444.12	1694.15
	Truck ownership	The number of registered trucks	4283.82	2753.44
	Motorcycle ownership	The number of registered motorcycles	337251.29	16446.7
Demographic characteristic	Total population	Total population in the city count (10^4)	318.99	146.62
	Urban population	Urban population count (10^4)	106.44	82.63
	Resident population	The number of permanent population (10^4)	146.35	8.16
	Mobile population	The number of temporary population (10^4)	121.81	15.86
	The number of drivers	The number of registered drivers	704725.35	276626.57
Road network	Expressway mileage	Total mileage of expressway (km)	96.62	44.94
	Arterial highway mileage	Total mileage of arterial highway (km)	269.74	93.41
	Secondary highway mileage	Total mileage of secondary road (km)	495.95	79.46
	Tertiary highway mileage	Total mileage of tertiary highway (km)	270.67	92.58
	Standard highway mileage	Total mileage of standard highway (km)	531.84	292.69
	Substandard highway mileage	Total mileage of substandard highway (km)	66.98	35.29
	Total mileage	Total mileage of road network (km)	1731.82	567.79
	Density of road network	Total mileage of road network / total area (km / 100 km^2)	95.56	31.43

5 Model results and discussion

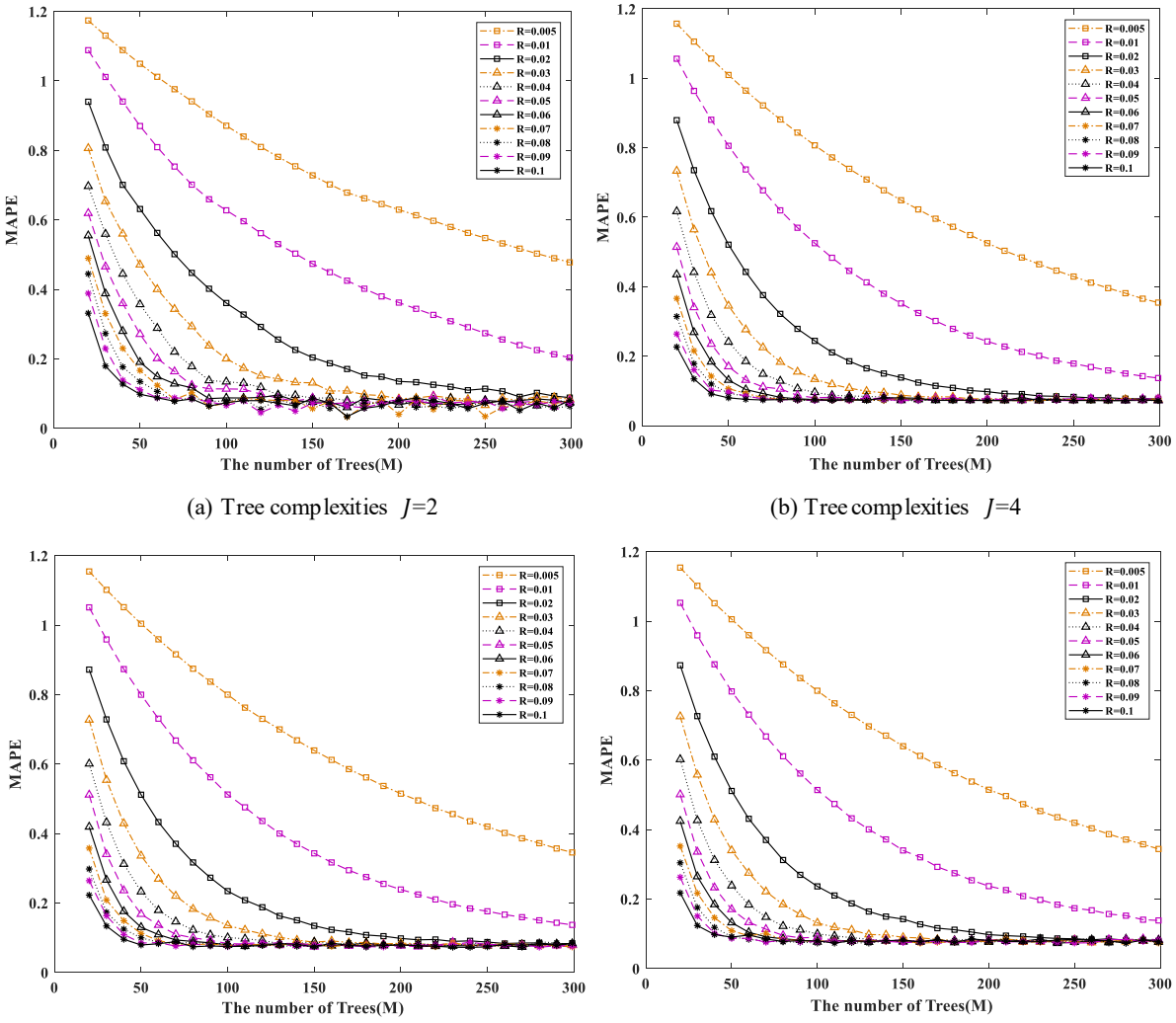
5.1 Model optimization

In this study, the data of the predicted year is used for the test, while the data before the predicted year is used for training. After tuning the models, the four traffic accident indicators from 2013 to 2016 (i.e., the number of traffic accidents, economic loss, the number of injuries, and the number of deaths) are predicted separately.

The performance of the GBDT model is determined by the number of decision trees M, a shrinkage parameter R and the number of leaves in one single decision tree J (namely, the tree complexity). The number of trees M represents the number of base models in the GBDT model, and the shrinkage represents the learning rates. To determine the optimal parameter combination, multiple GBDT models were established

with various numbers of trees (M=20-500), learning rates (R=0.005-0.1), and tree complexities (J=2-10).

The performance of the prediction models depends on a good set of hyper-parameters. In practice, traditional methods (e.g., manual search, grid search, and random search) are usually used to tune the hyper-parameters, but are computationally expensive (Lin et al., 2018; Zhang et al., 2019a; 2019b). Although Bayesian optimization method is reported to outperform other strategies in computational time, such method requires a prior distribution over the objective function and an acquisition function (Zhang et al., 2019a). Therefore, in this study the grid search method is adopted to search the optimal combination of parameters with minimum error. The optimal parameter combinations of the GBDT model for the four indicators are given in Table 2.



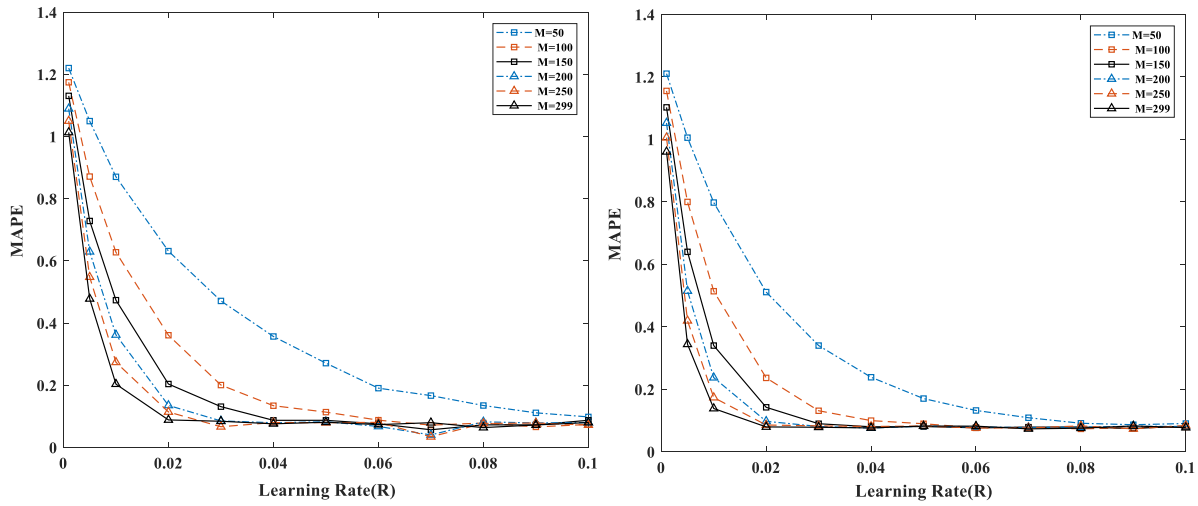
(a) Tree complexities J=2

(b) Tree complexities J=4

(c) Tree complexities J=6

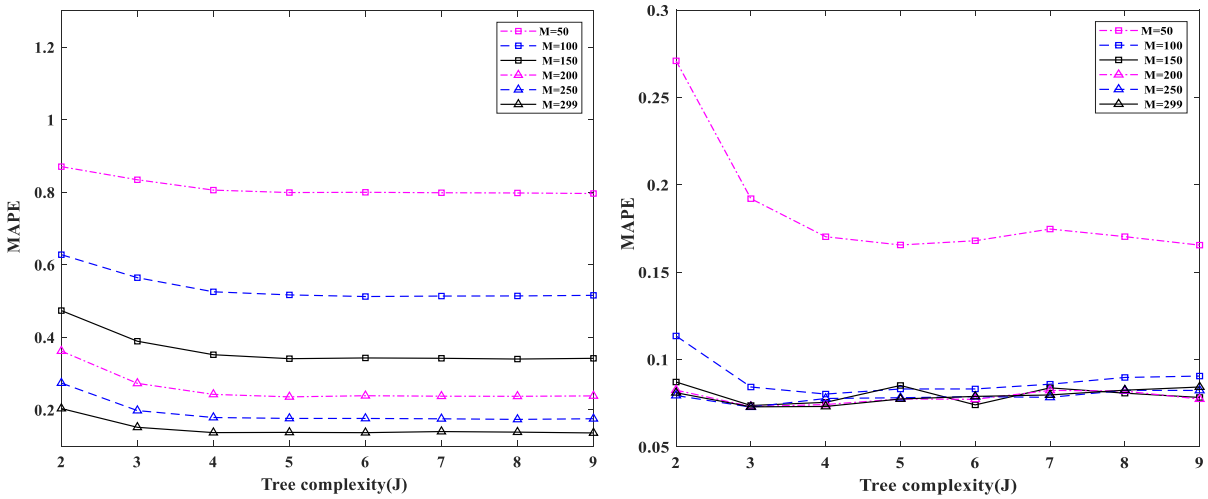
(d) Tree complexities J=8

Fig. 7. The relationship between MAPE and M with different combinations of R and J



(a) Tree complexities J=2 (b) Tree complexities J=8

Fig. 8. The relationships between MAPE and R with different combinations of M and J.



(a) Learning rate R=0.01 (b) Learning rate R=0.05

Fig. 9. The relationships between MAPE and J with different combinations of M and R

Fig. 7 presents the relationships between MAPE and the number of trees (M) under different values of R and J. As we can see, the prediction accuracy is higher as M increases. However, when the value of M increases to a certain value, the improvement in the prediction accuracy becomes trivial. As the learning rate (R) increases, the prediction accuracy also increases, and the slope of the line becomes steeper. However, when the number of trees is sufficiently large, using a higher learning rate value does not contribute to any improvement in the prediction accuracy.

Fig. 8 shows the relationships between MAPE and the learning rate R with different combinations of M and J. One can see that as the learning rate (R) increases, the prediction accuracy becomes higher and the curves tend to converge to a horizontal line. This suggests that when the learning rate is sufficiently high,

the contribution of adding trees is limited with respect to improving the prediction accuracy. In other words, with a higher learning rate, fewer trees are required to achieve better performance.

Fig. 9 shows the relationships between MAPE and J with different combinations of M and R. As we can see, with a small learning rate (R=0.01), the value of MAPE decreases continuously as the tree complexity increases before reaching a threshold. A marginal increase of M contributes less to an improvement in the prediction accuracy. On the other hand, with a large learning rate (R=0.05), the marginal diminishing effect of adding trees is more obvious. More importantly, an increase in the tree complexity may in turn lead to an increase in the prediction error, where over-fitting occurs.

Table 2 Optimal parameter combinations of the GBDT models

	Year	GDBT				
		The number of trees	Tree Complexity	Shrinkage	MAPE	Mean
The number of traffic accidents	2013	103	3	0.001	0.0109	0.0147
	2014	123	2	0.001	0.0104	
	2015	157	5	0.001	0.0086	
	2016	278	2	0.07	0.0288	
Economic loss	2013	132	3	0.1	0.2947	0.2267
	2014	64	3	0.1	0.2996	
	2015	273	3	0.06	0.3064	
	2016	247	2	0.1	0.0060	
The number of injuries	2013	20	2	0.001	0.0693	0.0448
	2014	20	2	0.001	0.0059	
	2015	49	4	0.001	0.0124	
	2016	123	2	0.09	0.0914	
The number of deaths	2013	141	2	0.1	0.0308	0.0583
	2014	136	2	0.1	0.0308	
	2015	235	2	0.1	0.0309	
	2016	279	2	0.09	0.1406	

5.2 Model comparisons

To validate the effectiveness of the GBDT model for traffic accident prediction, three methods are applied for comparisons using the same dataset, i.e., RF, SVR, and LSTM. The optimal parameters and corresponding prediction outcomes are given in Table 3.

Table 3 Comparison of different models' prediction accuracy

	Year	GDBT		RF		SVR		LSTM	
		MAPE	Mean	MAPE	Mean	MAPE	Mean	MAPE	Mean
The number of traffic accidents	2013	0.0109	0.0147	0.5117	0.3880	0.4516	0.4730	0.1199	0.0892
	2014	0.0104		0.4143		0.4595		0.0758	
	2015	0.0086		0.6030		0.4809		0.0802	
	2016	0.0288		0.0230		0.5001		0.0810	
Economic loss	2013	0.2947	0.2267	0.5188	0.4490	1.2365	1.3937	0.2207	0.2400
	2014	0.2996		0.6069		1.2790		0.3112	
	2015	0.3064		0.5762		1.4413		0.2474	
	2016	0.0060		0.0939		1.6181		0.1807	
The number of injuries	2013	0.0693	0.0448	0.6777	0.6197	0.4115	0.3819	0.1082	0.1036
	2014	0.0059		0.7679		0.3752		0.1362	
	2015	0.0124		0.8737		0.3749		0.0945	
	2016	0.0914		0.1594		0.3660		0.0753	
The number of deaths	2013	0.0308	0.0583	0.0603	0.0815	0.5546	0.5247	0.1128	0.0910
	2014	0.0308		0.0545		0.5546		0.0798	
	2015	0.0309		0.0688		0.5084		0.0824	
	2016	0.1406		0.1427		0.4814		0.0888	

In the random forest model, the number of decision trees M and the number of leaves J in one single decision tree also need to be determined. Therefore, in order to obtain the optimal numbers of M and J, multiple random forest models were established with a variety of decision trees (M= 20-500), and tree complexities (J=2-10). Then, the combination of the optimal parameters can be obtained via grid search.

A support vector machine is usually used for data classification and regression by introducing a dummy variable for each categorical attribute. The support vector regression (SVR), which uses the support vector to achieve the regression function, is an important application branch of the support vector machine. The kernel function is selected as the radial basis function in this study.

Long short-term memory networks (LSTM) is a popular deep learning model, which is a variant of the recurrent neural network (RNN). It can learn both the short-term and long-term dependent information. To achieve the best predictive effect of the LSTM, multiple models were established with a variety of hidden units (hidden units =20-80), learning rates (learning rates=0.005-0.1), epochs (epochs=50-120) and batch-sizes (batch-size=2-50), then a grid search is conducted to optimize these hyper-parameters, while Adam is selected as the optimizer in this study.

As we can see, the average MAPE of the GBDT models for the four traffic accident indicators from

2013 to 2016 are 0.2267, 0.0147, 0.0448 and 0.0583, respectively. For the number of traffic accidents and the number of injuries, the average MAPE for the RF and SVR are more than 10 times larger than that of the GBDT model. SVR exhibits the worst performance when predicting the economic loss and the number of deaths with the largest MAPE values of 1.3937 and 0.5247, respectively. LSTM performs better than RF and SVR. Obviously, the GBDT model outperforms the other models in four indicators. This suggests the advantage of the GBDT model in modelling complex non-linear relationships between influential factors.

5.3 Relative importance of influential factors

To study the distinct impact of each influential factors, the relative contributions of dependent variables (the number of traffic accidents, economic loss, the number of injuries and the number of deaths) are calculated using the optimal models, with the results shown in Table 4. A higher value of relative importance represents a greater impact on traffic accidents. The top six key influential factors of the four traffic accident indicators are summarized in Table 5.

As we can see, each potential influential factor reveals a distinct effect on the four traffic accident indicators. For the economic development factors, motorcycle ownership is the greatest contributor to all indicators except for economic loss. Medium bus ownership ranks as the second most influential factor for economic loss, as well as the number of traffic accidents and injuries. GDP per capita is the primary important influential factor of economic loss. The GDP per capita has a greater impact on road safety than the GDP, of which the relative importance of GDP on the four indicators is less than 15%.

Among the various industries, the influence of the tertiary industry on traffic accidents is more remarkable than that of other industries. The contribution of the tertiary industry on the number of traffic accidents, injuries and deaths as well as economic loss are 13.74%, 14%, 6.72% and 16.36%, respectively. This is because transportation mostly belongs to the tertiary industry. The contribution of the primary industry to the number of deaths is 11.42%, which is greater than the secondary and tertiary industries. The secondary industry has a low impact on the number of traffic accidents, economic loss and the number of injuries, with relative importance values of 4.16%, 6.03% and 6.67%, respectively.

The impacts of passenger volume on the four traffic accident indicators are all higher than that of the freight volume. Passenger volume plays an important role in the economic loss and the number of deaths. In particular, it contributes 79.48% to the economic loss. However, the freight volume contributes only 1.69% to the number of injuries.

Among the different types of vehicles, motorcycle ownership is the greatest contributor to all indicators except for economic loss. This is consistent with the fact that a large number of casualties have been induced due to the frequent occurrence of motorcycle accidents in Zhongshan. Medium bus ownership ranks as the second most contributory factor for economic loss, as well as the number of traffic accidents and injuries, with relative importance values of 98.05%, 94.13% and 86.19%, respectively. Medium trucks also contributes more than 60% to the above three indicators. However, medium bus ownership and truck

ownership contribute only 2.74% and 1.36% to the number of deaths, respectively.

Table 4 Relative importance of predictor variables in traffic accidents

	The number of traffic accidents		Economic loss		The number of injuries		The number of deaths	
	R.Imp (%)	Rank	R.Imp (%)	Rank	R.Imp (%)	Rank	R.Imp (%)	Rank
Gross Domestic Product (GDP)	14.64	12	5.01	25	9.02	20	6.63	17
Primary industry	9.82	21	12.2	17	12.78	17	11.42	9
Secondary industry	4.16	26	6.03	23	6.17	23	10.79	10
Tertiary industry	13.74	14	16.36	14	14	15	6.72	16
GDP per capita	40.67	**5**	100	**1**	37.42	**6**	16.59	**3**
Passenger volume	11.25	18	79.48	**4**	28.02	9	14.07	**5**
Freight volume	9.84	20	12.27	16	1.69	26	10.06	12
Car ownership	14.06	13	11.29	18	7.39	22	8.38	14
Medium bus ownership	94.13	**2**	98.05	**2**	86.19	**2**	2.74	22
Bus ownership	11.36	17	27	13	18.96	13	7.7	15
Van ownership	19.56	10	14.01	15	21.18	12	11.62	7
Medium truck ownership	69.36	**3**	62.42	7	76.96	**3**	1.36	25
Truck ownership	11.47	16	6.85	21	12.94	16	8.4	13
Motorcycle ownership	100	**1**	65.58	**6**	100	**1**	100	**1**
Total population	5.45	23	7.82	20	12.73	18	16.72	**2**
Urban population	40.42	**6**	67.19	**5**	29.7	8	15.19	**4**
Resident population	5.28	25	5.07	24	14.24	14	11.75	**6**
Mobile population	33.28	9	60.84	8	27.99	10	1.52	24
The number of drivers	11.65	15	8.76	19	8.97	21	11.45	8
Expressway mileage	0.79	27	0.15	27	1.08	27	5.79	19
Arterial highway mileage	5.4	24	32.82	12	10.45	19	4.74	21
Secondary highway mileage	7.26	22	49.79	10	26.27	11	0.78	27
Tertiary highway mileage	33.6	8	32.97	11	36.31	7	4.93	20
Standard highway mileage	36.27	7	54.96	9	38.9	**5**	1.1	26
Substandard highway mileage	44.07	**4**	94.29	**3**	47.89	**4**	1.73	23
Total mileage	9.85	19	6.23	22	4.22	24	6.05	18
Density of road network	15.49	11	4.84	26	3.21	25	10.16	11

Note: R.Imp = Relative importance.

Regarding socio-demographic factors, the urban population contributes the most to the four indicators, which is also an important factor influencing the number of traffic accidents and deaths, and the economic loss. Among the demographic characteristics, the total population is the second contributory factor of the number of deaths, whereas it has a minor impact on the other indicators. The relative importance values on the number of traffic accidents and deaths, and the economic loss is 5.45%, 12.73%, and 7.82%, respectively. The mobile population has a great impact on the number of traffic accidents and deaths, as well as economic

loss with relative importance values of 33.28%, 60.84% and 27.99%, respectively. A possible reason for this is that a larger number of mobile population indicates a stronger population mobility, which directly affects the implementation of traffic safety education and enforcement. The number of drivers only contributes 11.65%, 8.76%, 8.97%, and 11.45% to the four indicators, which may be due to the good education for registered drivers provided by local authorities.

Table 5 The top six most important influential factors of the four traffic accident indicators

Variable importance	The number of traffic accidents	Economic loss	The number of injuries	The number of deaths
First	Motorcycle ownership	GDP per capita	Motorcycle ownership	Motorcycle ownership
Second	Medium bus ownership	Medium bus ownership	Medium bus ownership	Total population
Third	Medium truck ownership	Substandard highway mileage	Medium truck ownership	GDP per capita
Fourth	Substandard highway mileage	Passenger volume	Substandard highway mileage	Urban population
Fifth	GDP per capita	Urban population	Standard highway mileage	Passenger volume
Sixth	Urban population	Motorcycle ownership	GDP per capita	Resident population

In terms of the road network factors, the impact of road mileage on traffic accidents will generally increase when the class of the road declines. The impact of road network factors on the number of deaths is generally low, with a relative importance of less than 6%. However, the impact of substandard highway mileage on the number of traffic accidents and injuries, and the economic loss is relatively large, with relative importance values of 44.07%, 47.89% and 94.29%, respectively. Standard highway mileage and substandard highway mileage contribute more than 30% to the number of traffic accidents and deaths, and the economic loss, where the impact of the latter is slightly greater than that of the former. In addition, the contribution of expressway mileage to each indicator is the lowest, with relative importance values of only 0.79%, 0.15%, 1.08% and 5.79%, respectively.

5.4 Partial effects of key influential factors

The GBDT model can not only calculate the relative importance on the independent variables to the dependent variables, but in can also illustrate the partial effects of independent variables. As opposed to the sensitivity analysis where only one variable is exclusively evaluated at a time and assuming the other variables remain constant, the advantage of the partial effect is that the association between the individual variables can be captured, while representing the influence of one single variable on the dependent variable.

To further investigate how the potential influential factors affect traffic accidents, we provide partial dependence plots that show the relationships between the four traffic accident indicators and a set of important influential factors. The partial effect helps to understand the impacts from the changes of one single influential factor when integrated across all other influential factors.

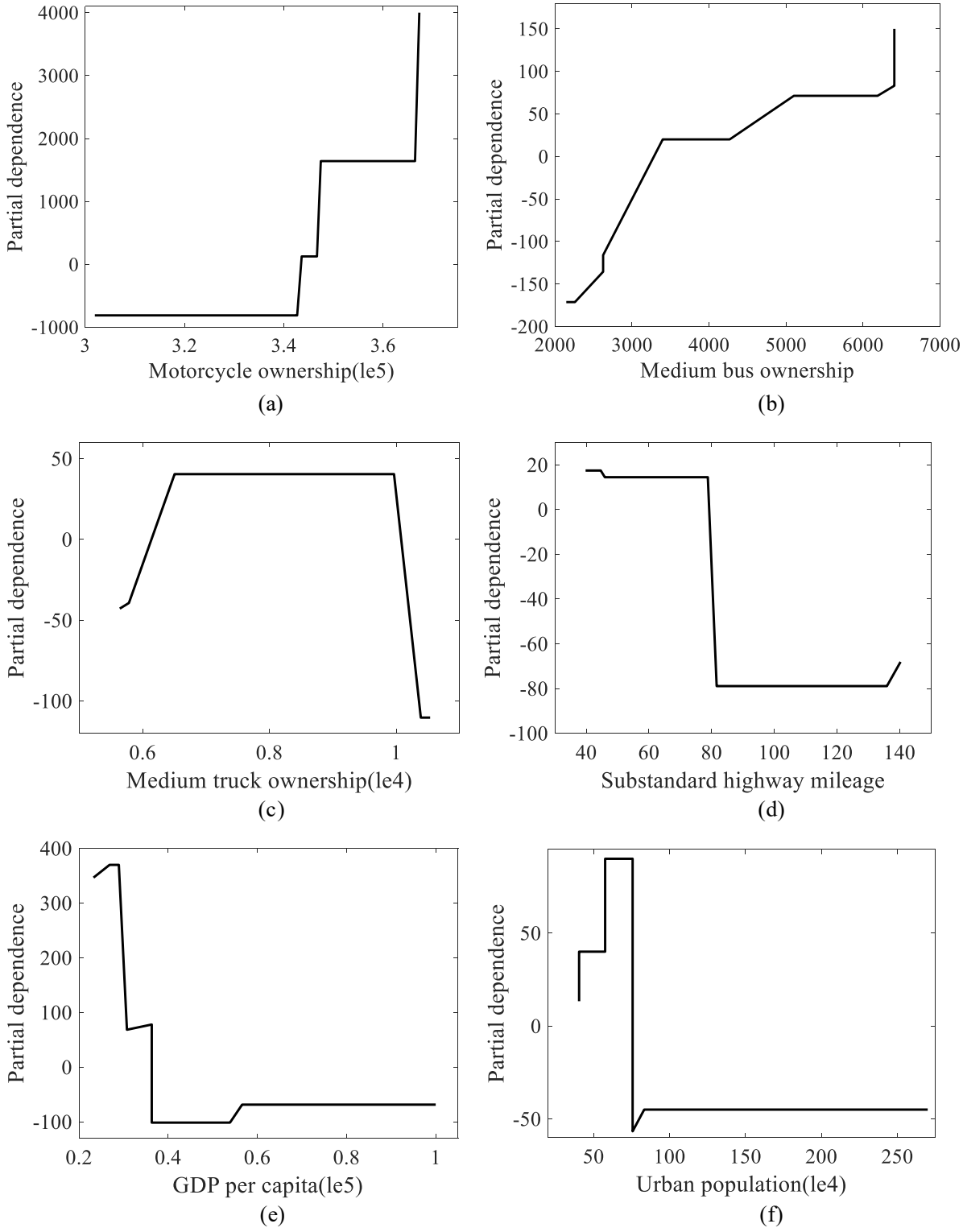
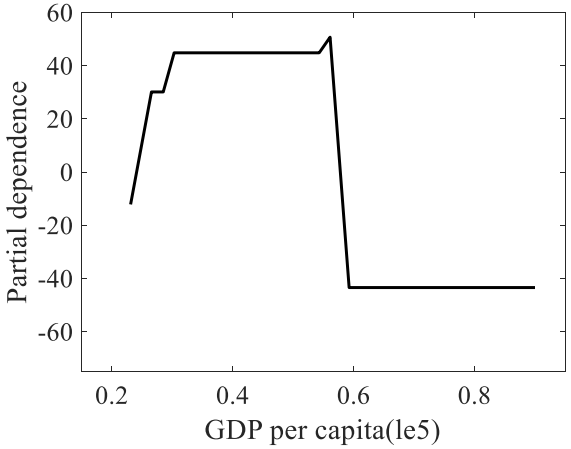
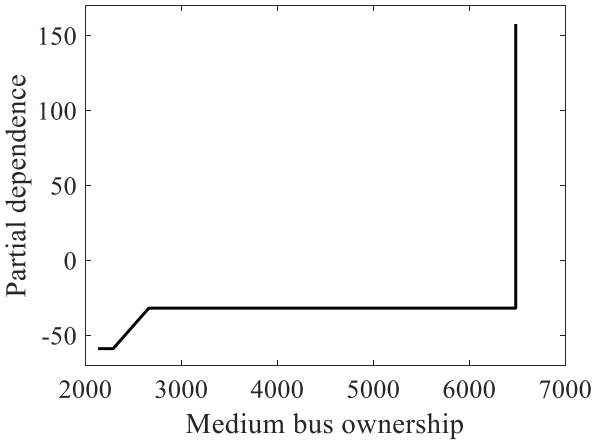


Fig. 10. Partial effects of important influential factors on the number of traffic accidents

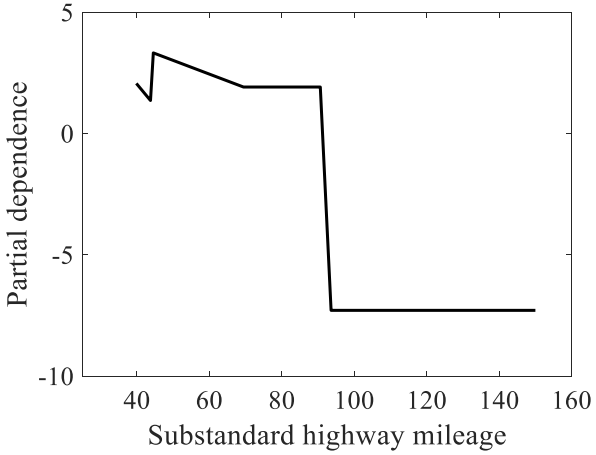
The partial effect goes from a negative value to a positive value as the medium truck ownership increases to 6,200. The number of traffic accidents increases substantially when it moves to the range of 6,500-10,000, but within the range the difference is trivial. Beyond this range, the partial effect jumps to a negative value. When the substandard highway mileage increases to approximately 90 km, the partial effect drops from 13 to approximately -75. For the GDP per capita, when it increases from 35,000 RMB to 40,000 RMB, the partial effect drops from 350 to approximately -80, followed by a long plateau. Within the range between 50,000 RMB and 90,000 RMB, a 15,000 RMB increase in the GDP per capita may contribute to an 80 increase in the number of traffic accidents. The partial effect first increases and then decreases with the growth of the urban population. A jump is observed when the urban population increases to 900,000. The boundary where jumps occur is mainly due to the traffic regulation actions in Zhongshan in the period of 2008-2009, such as the special enforcement for medium and heavy trucks, which effectively reduces the truck accidents (See Section 5.5).



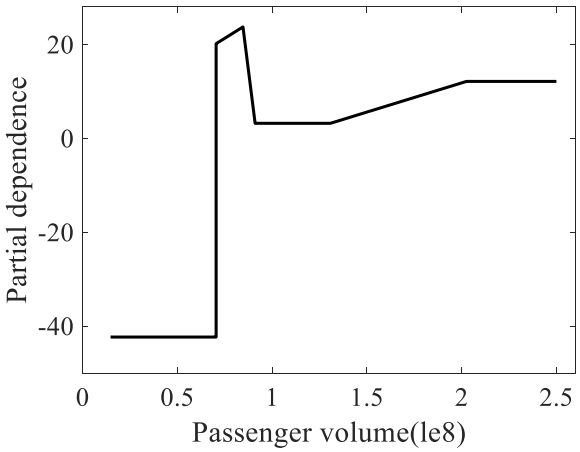
(a)



(b)



(c)



(d)

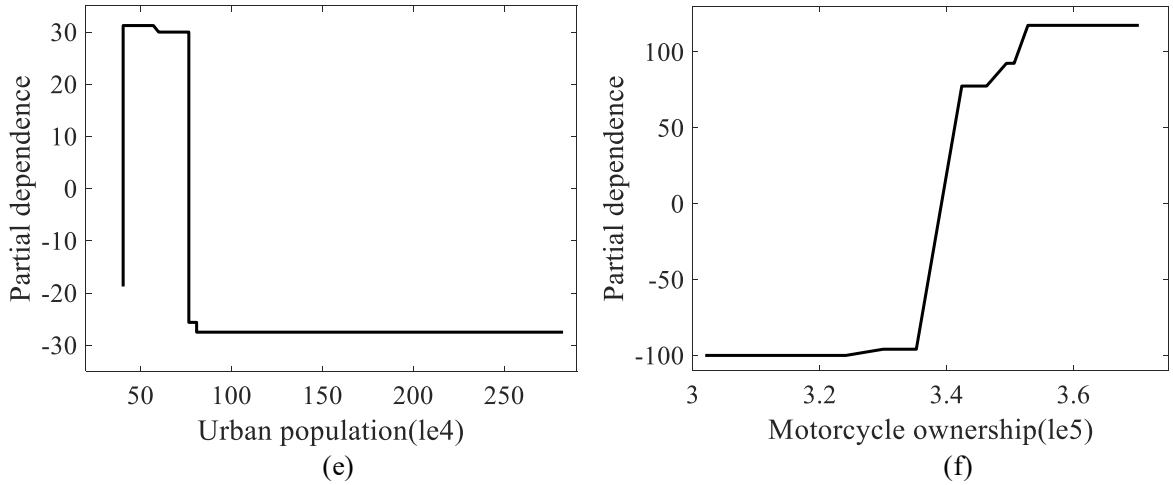
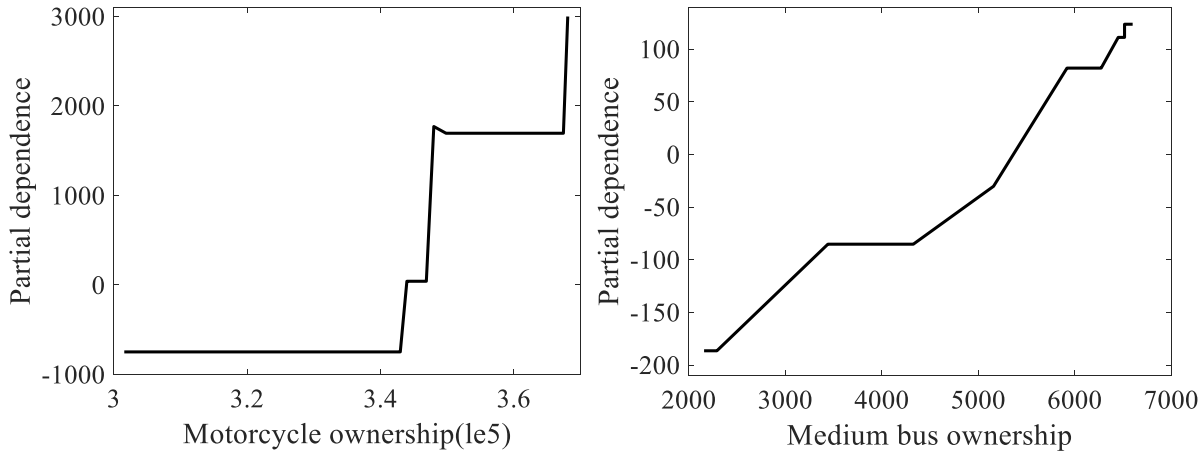


Fig. 11. Partial effects of important influential factors on economic loss

Fig. 11 shows the partial effects of six key influential factors on economic loss. The partial effect increases linearly from a negative value to 70 when the GDP per capita is within 30,000 RMB. The difference in the economic loss is trivial when the GDP per capita is within the range of 30,000-50,000 RMB. After this range, the economic loss drops to -50 and remains stable thereafter. The partial effect grows in a linear manner when the medium bus ownership increases to 2600. The difference in the economic loss is small when it lies in the range of 2600-6500. The effect of medium bus ownership jumps up to a higher level after that range. When the substandard highway mileage is within 80 km, it appears to have a small impact on the economic loss. Beyond this range, the partial effect jumps to a negative value and then remains constant. In contrast, when the passenger volume increases to approximately 7,000,000, an inflection point exists where the partial effect changes from a negative value to a positive one. This partial effect does not contribute to the economic loss when the motorcycle ownership is with 340,000. The economic loss increases in an almost linear way when it moves to the range of 340,000-350,000. However, its effect does not change after 350,000. The urban population can contribute to the economic loss either positively or negatively, depending on its level. It has a negative effect in the range of 500,000-800,000, but a positive effect after this range. The positive effect may also be thanks to the traffic regulation action in Zhongshan in 2008-2009.



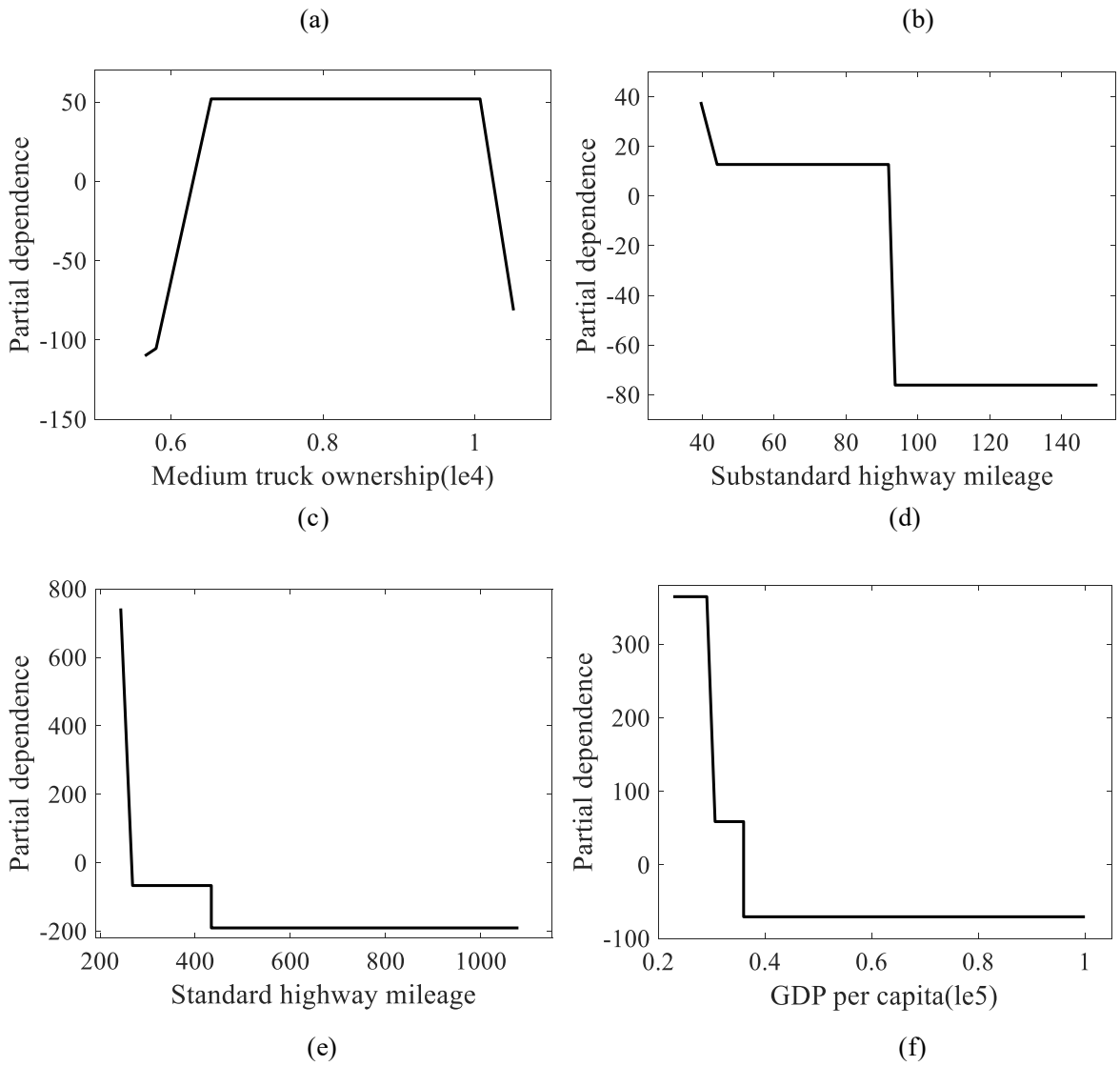


Fig. 12. Partial effects of the important influential factors on the number of injuries

Fig. 12 shows the partial effects of the six most important influential factors on the number of injuries. In line with the number of traffic accidents (Fig.10), the top four influential factors are motorcycle ownership, medium bus ownership, medium trunk ownership and substandard highway mileage with similar change patterns. The growth of substandard and standard highway mileage contributes to reducing traffic injuries when they are larger than 90 km and 250 km. An inflection point can also be found when the GDP per capita reaches a certain level (more than 30,000 RMB).

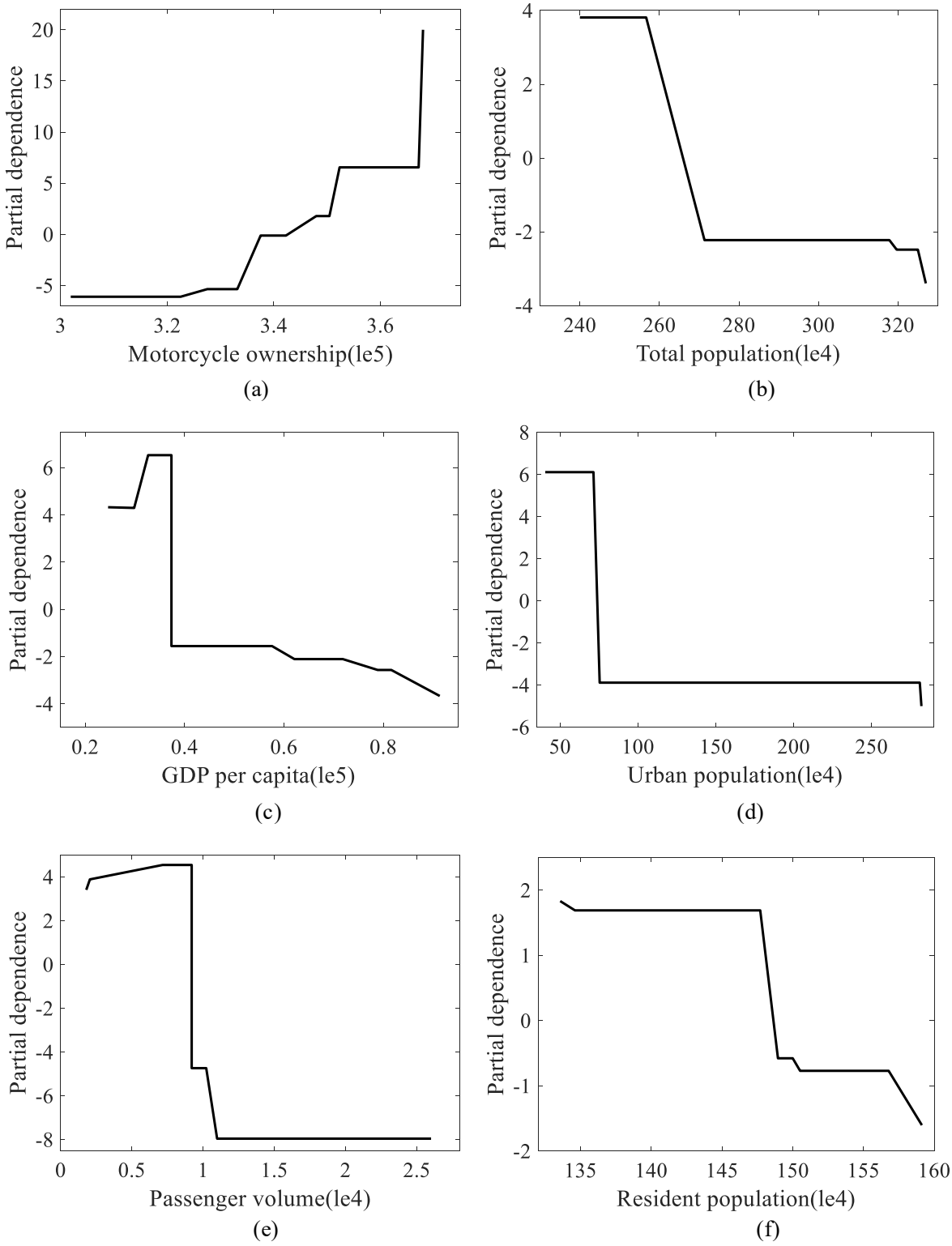


Fig. 13. Partial effects of important influential factors on the number of deaths

Fig. 13 shows the partial effects of six important influential factors on the number of deaths. As with the indicators of the number of traffic accidents and injuries, the most contributory factor is motorcycle ownership, of which the effect presents a sequence of scales. In contrast, the effects of other factors generally exhibit a downward trend. Inflection points exist where the effect of these factors become negative. In other

words, after the inflection points, increases in the GDP per capita, urban population and passenger volume contribute to reducing the number of deaths. Interestingly, the inflection points where jumps occur correspond to the period of traffic regulation actions in the period of 2008-2009.

5.5 Discussion and suggestions

In this section, we highlight the key findings from the rankings of the relative importance and the marginal effects of important influential factors, and discuss their implications for Zhongshan, China. As has been implied, since many factors could influence road safety, comprehensive countermeasures should be adopted.

First, motorcycle ownership contributes most to the number of traffic accidents, injuries, and deaths. Table 6 presents the trends in motorcycle accidents in Zhongshan from 2011 to 2015, which reinforces the message that the traffic accident rate of motorcycles is relatively high. This mainly results from the poor protection measures and the drivers' carelessness of traffic regulations. In addition, the illegal operation of motorcycles in Zhongshan has become increasingly rampant in recent years, which has worsened the traffic situation. Therefore, to improve local traffic safety performance, it is imperative to strengthen the management of motorcycles. For example, more intensive education on traffic safety should be focused on motorcycle drivers and stricter enforcement should be conducted regarding their common risky driving behaviours, such as speeding and not using helmets. Meanwhile, countermeasures should be adopted to control the increment of motorcycle ownership.

Table 6 Trends in motorcycle accidents in Zhongshan from 2011 to 2015

Year	The number of traffic accidents		Economic loss (million RMB)		The number of injuries		The number of deaths	
	Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion
2011	738	31.38%	69.29	22.58%	988	39.41%	87	28.34%
2012	778	36.84%	82.50	27.19%	1051	44.80%	108	35.64%
2013	720	38.22%	82.36	31.19%	940	45.21%	92	30.98%
2014	753	41.60%	90.69	34.39%	967	50.13%	113	38.18%
2015	711	41.65%	76.26	30.61%	856	47.79%	113	38.70%

Second, medium bus ownership and medium truck ownership play important roles in road safety. Over multiple years, efforts have been made to transform the industrial structure and attract investment to improve the local economic development. As a result, the demand for vans, medium trucks, and trucks in the industry has increased over the years. The growth of medium bus ownership and medium trunk ownership results in higher mobility and potential crash risks. During the local traffic regulation actions in the period of 2008-2009, the enforcement for medium and heavy trucks was strengthened. Traffic violations such as illegal assembly, overloading and speeding were strictly prohibited. As a result, the number of truck accidents in this period were greatly reduced. Therefore, strengthening the inspection and supervision of trucks and buses,

and improving the inspection system is suggested, particularly for tourist passenger vehicles and carriage of hazardous materials.

Third, although the ownership of motorcycles, buses and trunks contribute differently to road safety, their partial effects reveal similar upward trends with a sequence of scales and cut-off points for each traffic accident indicator (Figs. 10-13). Therefore, according to the effective ranges, policy makers and urban planners could draft differential safety countermeasures for different levels of vehicle ownership.

Fourth, the growth of the GDP per capita of Zhongshan contributes to an increase in traffic accidents before growing to a critical value. Afterwards, it contributes to reducing traffic accidents (e.g., Fig. 10(e), Fig. 11(a)). The reason for this is that vehicle ownership and the resulting traffic accident rate tend to increase with the growth of GDP initially. However, with the improvement of economic and living standards, the government has continuously raised the importance of traffic safety, and introduced more comprehensive legal provisions and the corresponding education, which significantly improves traffic safety. Specifically, during the period of 2008-2009, the local government invested 546 million RMB to improve road facilities, such as converting unsignalized intersections to signalized intersections and building/rebuilding more roads. A series of enforcement had also been employed considering local characteristics, such as on-road licenses for tricycles and the periodic inspections of heavy vehicles. Meanwhile, the government has launched a campaign with leaflets and manuals. More than 7 million RMB was spent in publicity and education in this period. According to our survey, the GDP per capita of Zhongshan was approximately 60,000 RMB in 2010. This explains the drop of partial effect when the GDP per capita increased from 50,000 to 60,000 RMB (Fig.11a).

Fifth, the effects of the urban population and mobile population on road safety should not be ignored (see Table 4). Table 7 presents the trends in the local population involved from 2011 to 2015. By 2014, the urban population jumped up to approximately 28 million. The mobile population accounts for more than 40% of the total population. The ever-increasing mobile population has put more pressures and potential risks on transportation systems, such as more poorly educated pedestrians. Therefore, more education and enforcement efforts for the mobile population can significantly improve road safety.

Table 7 Trends in the population in Zhongshan from 2011 to 2015

	Total population (10^4)	Urban population (10^4)		Resident population (10^4)		Mobile population (10^4)	
	Count	Count	Proportion	Count	Proportion	Count	Proportion
2011	314.23	80.09	25.49%	150.73	47.97%	126.63	40.30%
2012	315.5	80.89	25.64%	152.01	48.18%	130.45	41.35%
2013	317.39	82.5	25.99%	154.09	48.55%	132.3	41.68%
2014	319.27	281.2	88.08%	156.06	48.88%	133.9	41.94%
2015	320.96	282.83	88.12%	158.68	49.44%	134.5	41.91%

Finally, in regard to road network factors, the number of traffic accidents with higher classes is relatively

small, particularly for the expressway (see Table 4). Generally, a higher road class indicates better countermeasures. A substandard highway includes rural roadways with few road markings and traffic safety facilities. Moreover, motorcycles, which are the main source of traffic accidents, are more likely to appear on low-class roads, particularly on substandard highways. As a result, the crash risks in substandard highways would be relatively large. Therefore, it is suggested to continuously increase the investment in improving the classes of roads and safety facilities and reduce the safety hazards of low-class roads.

6 Concluding remarks

The study contributes to understanding the joint effect of economic development, demographic characteristics and the road network on regional traffic accidents, which could help decision-makers adopt comprehensive countermeasures to improve road safety. A total of 27 elaborated influential factors in Zhongshan, China over the period 2000-2016 are collected. Then this study proposes a new data mining method termed the gradient boosting decision tree (GBDT) model to examine the effects of these factors on four traffic accident indicators (the number of traffic accidents, injuries, deaths, and the economic loss). The impacts of the learning rate, tree complexities, and the number of leaves in one single decision tree on the performance of the GBDT are also investigated. The optimal GBDT for each indicator is achieved by comparing the model's performance with different parameter combinations.

The results show that, compared to other traditional machine learning methods that use 'black-box' procedures, the GBDT not only presents a higher prediction accuracy, but can also better handle the multicollinearity between the explanatory variables. Moreover, the GBDT has a stronger interpretation power of identifying and prioritizing influential factors on the traffic accident prediction. This is beneficial for data acquisition and maintenance, and provides a better understanding of the safety effects of economic development, demographic characteristics and the road network. The results also show that there are both similarities and differences in the important influential factors for the four traffic accident indicators. For example, motorcycle ownership and the GDP per capita are great contributors to all indicators, while the total population and resident population are only important for the number of deaths. In particular, we also investigate the partial effects of the key influential factors, from which subtle and sudden changes in each indicator can be accurately identified. These findings, particularly the effective ranges of the influential factors, could provide evidence to support transport planning, policy guidance and road safety programs.

The machine learning model presented in this paper can be applied to other regions to help planners gain a clear understanding of how changes in economics, socio-demographics and the road network jointly affect road safety. As a first exploration of the modelling technique in safety analysis, this paper examines the safety effects of a series of influential factors at aggregated city-level data, and has not considered the underlying spatial correlation and the potential heterogeneity in the four indicators and independent variables within the county. Therefore, it would be beneficial for future research to collect more detailed geographic dimensional data and incorporate a spatial correlation effect into the model. It would also be interesting in

future work to consider the factors of the built environment in tandem with socio-economic characteristics.

Acknowledgements

This work is jointly supported by Science and Technology Program of Guangzhou, China (Project No. 201904010202), and the National Natural Science Foundation of China (Project No. 61703165; 71861023; 71890972; 71890970).

References

Albalate, D., Fernández-Villadangos, L., 2010. Motorcycle injury severity in Barcelona: the role of vehicle type and congestion. Journal of Crash Prevention & Injury Control, 11(6), 623-631.

Assemi, B., Mark, H., 2018. Relationship between heavy vehicle periodic inspections, crash contributing factor and crash severity. Transportation Research Part A, 113, 441-459.

Bener, A., Yousif, A., Al-Malki, M. A., El-Jack, I., Bener, M., 2011. Is road traffic fatalities affected by economic growth and urbanization development? Advances in Transportation Studies, 23, 89-104.

Bougueroua, M., Carnis, L., 2016, Economic development, mobility and traffic accidents in Algeria. Accident Analysis & Prevention, 92, 168-174.

Chang, L. Y., Chien, J. T., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Safety Science, 51(1), 17-22.

Connelly, L.B., Supangan, R., 2006. The economic costs of road traffic crashes: Australia, states and territories. Accident Analysis & Prevention, 38(6), 1087-1093.

Chen, F., Chen, S., 2011. Injury severities of trunk drivers in single- and multi-vehicle accidents on rural highways. Accident Analysis & Prevention, 43(5), 1677-1688.

Dapilah, F., Guba, B. Y., Owusu-Sekyere, E., 2016. Motorcyclist characteristics and traffic behaviour in urban northern ghana: implications for road traffic accidents. Journal of Transport & Health, 4, 237-245.

Delmelle, E. C., Thill, J. C., & Ha, H. H., 2012. Spatial epidemiologic analysis of relative collision risk factors among urban bicyclists and pedestrians. Transportation, 39(2), 433-448.

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

Ding, C., Wang, D., Liu, C., Zhang, Y., Yang, J., 2017. Exploring the influence of built environment on travel model choice considering the mediating effects of car ownership and travel distance. Transportation Research Part A, 100, 65-80.

Ding, C., Cao, X., Næss, P, 2018. Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. Transportation Research Part A, 100, 65-80.

Dong, C., Clarke, D. B., Nambisan, S.S., & Huang, B., 2016. Analyzing injury crashes using random-parameter bivariate regression models. Transportmetrica A, 1-17.

Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects. Accident Analysis & Prevention, 82, 192-198.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

García-ferrer, A., De Juan, A., Poncela, P., 2007. The relationship between road traffic accidents and real economic activity in Spain: common cycles and health issues. Health Econ. 16 (6), 603–626.

Garg, N., Hyder, A.A., 2006. Exploring the relationship between development and road traffic injuries: a case study from India. Eur. J. Public Health 16 (5), 487–491.

Goel, G., Sachdeva, S.N., 2016. Analysis of road accidents on NH-1 between RD 98 km to 148 km. Perspectives in Science, 8(C), 392-394.

Hao, W., Kamga, C., Yang, X., et al. 2016, Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States. Transportation Research Part F, 43, 379-386.

Hastie, T., Tibshirani, R., Friedman, J., 2008. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Second Edition. Springer.

Ha, H., and Thill, J-C., 2011. Analysis of traffic hazard intensity: A spatial epidemiology case study of urban pedestrians. Computers, Environment and Urban Systems, 35(3), 230-240.

He, H., Paichadze, N., Hyder, A. A., et al. 2015, Economic development and road traffic fatalities in Russia: analysis of federal regions 2004–2011. Injury Epidemiology, 2(1), 19-35.

Hong, K., Lee, K. M., & Jang, S. N., 2015. Incidence and related factors of traffic accidents among the older population in a rapidly aging society. Archives of Gerontology & Geriatrics, 60(3), 471-477.

Huang, H. L., Zeng, Q., Pei, X., Wong, S. C., Xu, P. P., 2016. Predicting crash frequency using an optimised radial basis function neural network model. Transportmetrica A: Transport Science, 12(4), 330-345.

Huang, H., Wang, D., Zheng, L., Li X., 2014. Evaluating time-reminder strategies before amber: Common signal, green flashing and green countdown. Accident Analysis and Prevention, 71, 248-260.

Hughes, B.P., Newstead, S., Shafiei, S., et al. 2014, Data foundations for relationships between economic and transport factors with road safety outcomes. Journal of the Australasian College of Road Safety, 25(3), 41-49.

Iwata, K., 2010. The relationship between traffic accidents and economic growth in China. Econ. Bull. 30 (4), 3306–3314.

Juneyoung, P., Mohamed, A. A., & Jaeyoung, L., 2018. School zone safety modeling in countermeasure evaluation and decision. Transportmetrica A: Transport Science, 1-25.

Law, T. H., 2015. Factors associated with the relationship between non-fatal road injuries and economic growth. Transport Policy, 42, 166-172.

Li, R., Zhao, L., 2014. Factor analysis of traffic accidents, Journal of Mathematical Medicine, 6, 634-636.

Lin, L., He, Z., Peeta, S., 2018. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. Transportation Research Part C, 97, 258-276.

Lu, H.E., Lin, Z., Mathematics, S. O., 2016. Stepwise regression analysis of the influencing factors of traffic

accident. Journal of Yili Normal University, 04, 20-24.

Ma, C., Hao, W, Xiang, W., Yan, W., 2018. The impact of aggressive driving behavior on driver-injury severity at highway-rail grade crossings accidents. Journal of Advanced Transportation, Vol. 2018, Article ID 9841498, 10 pages.

Milton, J., & Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation, 25(4), 395-413.

Moomen, M., Mahdi, R., Khaled, K., 2018. An investigation of influential factors of downgrade truck crashes: A logistic regression approach. Journal of Traffic and Transportation Engineering (English Edition). In press.

Pugachev, I., Kulikov, Y., Markelov, G., Sheshera, N., 2017. Factor analysis of traffic organization and safety systems. Transportation Research Procedia, 20, 529–535.

Ren, Y., Peng, H., 2013. Factors affecting China traffic accident casualties: an empirical study, Forecasting, 32(3), 1-7.Sakhapov, R., Nikolaeva, R., 2017, Economic aspects of traffic safety administration. Transportation Research Procedia, 20, 578-583.

Sheng, R., Zhong, S., Barnett, A. G., Weiner, B. J., Xu, J., Li, H., et al., 2018. Effect of traffic legislation on road traffic deaths in Ningbo, China. Annuals of Epidemiology, 28(8), 576-581.

Van den Bossche, F., Wets, G., Brijs, T., 2004. A regression model with ARIMA errors to investigate the frequency and severity of road traffic accidents. Proceedings of the 83rd Annual Meeting of the Transportation Research Board, Washington.

Vasconcellos, E.A., 1995. Reassessing traffic accidents in developing countries. Transport Policy, 2(4), 263-269.

Wiebe, D.J., Branas, C.C., 2016. Economic development and road traffic fatalities in two neighbouring African nations. African Journal of Emergency Medicine, 6(2), 80-86.

Wissen L.V., Huisman C., 2002. Simulating the interplay between regional demographic and economic change in two scenarios. Networks & Spatial Economics, 2(2), 127-150.

World Health Organization (WHO), 2017. Road traffic injuries. Retrieved from <http://www.who.int/mediacentre/factsheets/fs358/en/>[accessed on April 17th, 2017].

Yannis, G., Papadimitriou, E., Folla, K., 2014. Effect of GDP changes on road traffic fatalities. Safety Science, 63(4), 42-49.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transportation Research Part C, 58, 308-324.

Zhang, K., Zheng, L., Liu, Z. Jia, N., 2019a. A deep learning based multitask model for network-wide traffic speed prediction. Neurocomputing, 10.1016/j.neucom.2018.10.097.

Zhang, K., Liu, Z., Zheng, L., 2019b. Short-term prediction of passenger demand in multi-zone level: Temporal convolutional neural network with multi-task learning. IEEE Transactions on Intelligent Transportation Systems, 10.1109/TITS.2019.2909571.

Zheng, L., Ran, B., Huang, H., 2017. Safety evaluation for driving behaviors under bi-directional looking context. Journal of Intelligent Transportation Systems, 21(4), 255-270.

Zheng, L., Zhu, C., He, Z., He, T., 2018. Safety rule-based cellular automaton modeling and simulation under V2V environment. Transportmetrica A: Transport Science, 10.1080/23249935.2018.1517135.

Ziyab, A. H., Akhtar, S., 2012. Incidence and trend of road traffic injuries and related deaths in Kuwait: 2000-2009. Injury-international Journal of the Care of the Injured, 43(12), 2018-2022.