Meteorological Variables and Prediction of Road Traffic Accident Severity in Suzhou city of Anhui Province of China

Mingming Liang
Anhui Medical University
Yun Zhang
Anhui Medical University
ZhenHai Yao
anhui public meteorological serice cnter
Guagbo Qu
Anhui Medical University
Tingting Shi
Anhui Medical University
Min Min
Anhui Medical University
Pengpeng Ye
National Center for Chronic and noncommunicable disease control and prevention
Leilei Duan
center for chronic and noncommunicable disease control and prevention
Peng Bi
The University of Adelaide
Yehuan Sun (yhsun_ahmu_edu@yeah.net)
Anhui Medical University https://orcid.org/0000-0002-8651-8059

Research

Keywords: Meteorological factors; Machine learning; Traffic Accident Severity; Forecast model

Posted Date: January 1st, 2020

DOI: https://doi.org/10.21203/rs.2.19867/v1

Abstract

Background: The prediction of the severity of traffic accidents is concerned by researchers and law enforcement. In order to simulate the relationship between road severity results and meteorological factors, a large number of models have been proposed. This study purpose is to conduct a machine learning model to investigate the impact of meteorological variables on the severity of road traffic accidents. Methods: Using data from the 2007 and 2008 -2017 the Traffic Police Detachment of the Public Security Bureau of Suzhou, 7,795 traffic accidentswere included in this study. We attempted to use a random forest model to convey the nonlinear relationship between meteorological variables and the severity of traffic accidents, and to compare the prediction accuracy of the neural network model. The model is constructed by the randomForest package and the neuralnet package in the R software. 75% of the training samples were divided from the data to establish a prediction model, and the remaining 25% of the test samples were used for testing. In addition, in order to understand the accuracy of the model prediction, the predicted results were calculated and compared with the actual results. Results: In the random forest model, the most optimal mtry parameter value was 5, the number of decision trees is 400. The weight of wind direction, atmospheric pressure and temperature might be higher than other variables. The OOB (out of bag) estimate of error rate was 51.09%, and the error rate for general traffic accident prediction is the lowest (45.97%). Similarly, in the neural network model, the calculated error rate is 61.01%, with the lowest error rate for minor traffic accidents (35.84%). Conclusions: The results of this study show that how using meteorological data predicts the severity of a traffic accident with relative accuracy, and the random forest model may be more suitable than the neural network model. Research and application of machine learning algorithms in the field of traffic accidents should be further explored.

Background

Recently, evidence from a large number of epidemiological studies suggests that the meteorological factors, including temperature, precipitation, wind speed, and visibility, may be related to the occurrence of road traffic accidents and the severity of accidents [1].

A large number of models have been proposed in order to simulate the relationship between traffic severity results and meteorological factors. For example, Zeng using a bayesian spatial generalized ordered logit model found weather conditions have a significant impact on the level of severe collisions [2]. Bailey and Hewson [3] analyzed the incidence of fatal and severe road traffic accident among different types of road users through generalized linear mixed model. Zhai et al. used a geographical information system approach to integrate high-resolution weather data with crash data established a mixed logit model to determine the cause of pedestrian crash severity [4].

Although the above issues have been extensively studied in the previous literatures, there are still some gaps that need to be further addressed. In fact, when traffic accidents occur, the combinations of different meteorological factors could result in different traffic results. In addition, there are few studies have report the accuracy of model predictions. The general linear regression and nonlinear regression are difficult to

meet these requirements. Therefore, the forecast model established by the machine learning method may be more suitable [5].

The purpose of this study is to conduct a comprehensive study based on real-time meteorological and traffic accident data from Suzhou of Anhui Province, try using a random forest model to convey the non-linear relationship between meteorological variables and traffic accident severity, and compare the prediction accuracy with that of a neural network model.

Methods

Study site

Suzhou is a prefecture-level city in Anhui Province, China. It governs four counties and one district, namely Dangshan, Xiaoxian, Lingbi, Sixian county and Yongqiao district. Suzhou has a total area of 9,939 square kilometers, of which the urban construction land area is 195.89 square kilometers. The annual average temperature is 16℃, and the annual precipitation is 975 mm.

The total population of Suzhou was 6.55 million until end of 2017. The total length of highways was 16,471 kilometers, the expressway was 359 kilometers, the length of urban roads was 1,737.61 kilometers, and the mileage of the first-class road had reached 350 kilometers. By the end of 2017, and the total automobiles of Suzhou was 440,900, the road freight volume was 252.27 million tons, and the average car ownership per thousand people was 67.32. In addition, the volume of passenger traffic of Suzhou highway was 36.71 million per year [6].

Real-time meteorological data

We got 10 years of meteorological data extracted and accumulated from the National Meteorological Information Center (http://data.cma.cn). These data are collected and stored in the database by the meteorological observatories of five counties or districts in Suzhou. Data include atmospheric pressure (hPa), temperature (°C), relative humidity (%), precipitation (mm), wind direction (°), wind speed (m/s), visibility (m), snow depth (cm), evaporation(mm), total cloud amount(%), etc.

In meteorological observatories, atmospheric pressure, temperature, humidity, precipitation, wind direction, and wind speed are all recorded by electronically controlled mechanical equipment. These equipment are equipped with an embedded chip that automatically collects surrounding meteorological data on time (every 2 minutes, 10 minutes, 1 hour or 1 day,). Then, the collected data is automatically encoded into a binary data stream that is sent to the database. For variables such as visibility and total cloud amount, they are manually recorded by observers and stored in the database.

Accident data

The cases included all recorded road traffic accidents in the city of Suzhou from January 1, 2008 through December 31, 2017. The traffic accident data were obtained from the Traffic Police Detachment of the Public Security Bureau of Suzhou. And the traffic accident data mainly included the number of accident, the administrative division, the accident time, the total number of deaths, the number of injured, the accident identification cause, the direct property loss, the location of accident, and the highway administrative grade. Other vehicle, population and road information are derived from the other information comes from the Suzhou Municipal Bureau of Statistics website.

According to the loss caused by the accident and “the Notice of the Ministry of Public Security on Revising the Classification Standard”, the severity of the traffic accident is divided into four different levels:

The minor road traffic accident (level □): It means that one or two people are slightly injured at one time, and the property loss is less than 1,000 yuan (RMB, the same below);

The general road traffic accident (level □): It means that the number of injured person is less than 10, and the property loss is less than 30,000 yuan;

The serious road traffic accident (level □): It means that an accident that caused one or two deaths and the number of injured persons was less than 10; or more than 10 people were injured; or the property loss was more than 30,000 yuan and less than 60,000 yuan;

The particularly serious road traffic accident (level □): It means that an accident that caused one or two deaths at a time but the number of injured more than 10 people; or caused more than two deaths at a time; or a property loss of more than 60,000 yuan.

Data treatment

All data were sort out by SPSS 23.0. The accident data information and meteorological data were matched according to time (hours). For missing values of meteorological data, if the previous hour and the last hour were available, then their average regarded as the substitute value. If the adjacent data were also missing, the data would be deleted. For traffic accident information, the data would be deleted if the related detail was missing.

Before data treatment, all variables are converted into numerical types according to *R* package requirements: the ordered categorical variables, such as road grades, converted from national roads, provincial roads, county roads, and rural roads to “1”, “2”, “3”, “4”; the categorical variables converted to dumb variables such as “1” indicated that the accident occurred on the highway, “2” indicated not. And different variables may have different dimensions, which could lead to large differences between the

data. Failure to process may affect the results of the data analysis. In order to eliminate the influence of the dimension and the range of values between the indicators on the results of the data analysis, the data needs to be standardized:

$$y_i = \frac{x_i - \bar{x}}{s}$$

$\bar{x}$ is the mean of the data, s is the standard deviation of the data, and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,,\ s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Land use random forest model

Random forest model have no requirements for data types and can solve regression and classification problems. The advantages are not only widely applicable, but also suitable for situations with many variables or large sample sizes. However, the disadvantage is that sometimes the model is difficult to interpret and the amount of calculation is complicated.

Random forest models have already been used in contributions dealing with the problem of traffic accident and severity prediction [7]. In the model, random forest is designed to produce accurate predictions that do not overfit the data. Random forests are similar to bagging trees in that bootstrap samples are drawn to construct multiple trees, the difference is that each tree is grown with a randomized subset of predictors, hence the name “random” forests. A large number of trees are grown, hence a “forest” of trees. Random forests are more like a "black box" approach because each individual tree cannot be inspected separately. However, it provides some indicators that help explain. The results table can be used to compare the relative importance between predictors. Therefore, this process is easier to interpret than a method such as a neural network.

Out-of-bag samples can be used to calculate an unbiased error rates and variable importance without the need for test sets or cross-validation. Because a large number of trees are grown, the generalization error is limited, which means that over-fitting is not possible, which is a very useful feature of prediction. And another advantage of RF is that the predicted output depends only on one user-selected parameter, the number of predictors to be chosen randomly at each node [7].

The randomForest package (version 4.6-14) in the R software (version 3.5.1) implements a random forest model. We randomly chose 75% of the data as training data. The remaining 25% of the data were treated as testing data. According to the methods previously studied, this study determines the variables method is assess how output changes by varying input variable values one by one [9-10].

To select model parameters, the appropriate number of variables "*mtry*" and decision trees "*ntree*" were chosen for better model fitting. The *mtry* is the number of variables randomly sampled as candidates at each split, the *"for()"* function in R software can traverse all variables and select the *mtry* value with the lowest error rate. The *ntree* is the number of trees to grow, it indicates the number of decision trees when modeling. Too high will increase the complexity, and too low will increase the error rate. Random forest modeling is performed with the *mtry* of the minimum mean error described above, and the relationship between the model error rate and the decision tree is visualized. Select the optimal *ntree* parameter. Establish a random forest model to obtain the importance of various variables in the model. Compute an out-of-bag (OOB) error rate by using the data not in the bootstrap sample, and the predicted results were calculated and compared with the actual results.

The back-propagation neural network model

The back-propagation neural network (BPNN) is one of the artificial neural networks, and has a classical multilayer topology with feed-forward connections[11]. A BPNN does not need any a priori assumptions on relationships between linear or non-linear variables, and offers the opportunity to investigate and create the first discriminant analysis in problems where the phenomena (the relationships between input and output) are not well known [12].

In this investigation, the neural network model could be implemented in the R software using the neuralnet package (1.44.2). The data was randomly divided into training and test sets by 3:1. The R software will output the neural network structure diagram through the "plot" function. The parameters were selected according to the method of Mussone et al, and 10 neurons in the hidden layer were set [10]. And by using the algorithm, the weight matrix of the model and the visualization of the importance of each variable are obtained. Sensitivity analysis was performed by changing the neurons for parameters (7-14).

In addition, in order to understand the accuracy of the model prediction, the predicted results were calculated and compared with the actual results.

Results

In 2008-2017, there were a total of 7,795 traffic accidents, 2,659 minor accidents, 2817 general accidents, 2,264 serious accidents, and 55 particularly serious accidents. Table 1 details the meteorological variables and traffic parameters, including abbreviations, units, types and data sources. As shown in Figure 1, the trend of increasing or decreasing traffic accidents of different severity was similar.

As seen in the Table 1, all variables were confirmed in the algorithm selection. Through the variable selection, exclude the candidate independent variables one by one to ensure that the variables in the model are significantly correlated with the outcome variables. Finally, real-time atmospheric pressure, temperature, relative humidity, precipitation, wind direction, wind speed, visibility, and administrative division, expressway or not , highway grade, non-motor vehicle, year, month and hour of occurrence are included in the model.

In the random forest model, all the variables were loop through by the "*for()*" function and the most optimal *mtry* parameter value was 5. Modeling visualizes the relationship between model error rates and decision trees. As can be seen from Figure 2, when the number of decision trees is greater than 400, the error rate tends to be stable, so the *ntree* is set to 400. The importances of the variables were shown in Figure 3. The measure is the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index. It is found that in the construction of this model, the weight of wind direction, atmospheric pressure and temperature might be higher than other variables. As shown in Table 2, the OOB (out of bag) estimate of error rate was 51.09%, and the error rate for general traffic accident prediction is the lowest (45.97%).

Similarly, the variables are incorporated into the neural network model. The structure diagram of the neural network model is shown in Figure 4. In the neural network diagram, the line from the input node showed the connection between each layer and the weight on each connection. The line starting from the "1" node showed the deviation added in each step, which could be considered as the intercept of the linear model. As shown in table 2, the calculated error rate is 61.01%, with the lowest error rate for minor traffic accidents (35.84%). In the model, the variables with the highest weight are visibility, month and wind speed (figure 5). Sensitivity analysis showed the effects were slightly decreased when altering the number of neurons (7-9□11-14) in the model.

Discussion

There has been a lot of road safety research on the modeling of road traffic accident severity. A common method is to use the collision severity as a dependent variable, the driver, road, weather, meteorology and other characteristics as independent variables. They usually use a binary logit or probit model with two levels of severity traffic accidents, or use the multinomial logit model to predict multiple levels of severity traffic accidents [10]. Unlike these modeling methods that rely on fixed rules, machine learning training computers find the logic inherent in data from the data through "training" [14]. In this paper, we used two different machine learning modeling approaches, random forests and neural networks, to establish predictive models for the severity of road traffic accidents. And the meteorological and traffic accident time included in the model could be accurate to the hours. Of course, the neural networks models have already been used extensively in contributions dealing with the problem of crash prediction or severity. Wang et al. investigated the location and timing of potential secondary accidents after the initial traffic accident and used back-propagation neural networks to predict potential secondary traffic accidents. From the results, BPNN is more adequate in describing the impact of most variables, and its goodness

and prediction accuracy are relatively good. It is believed that the BPNN model can be used to predict the time gap between initial and secondary incidents, and policymakers and event managers can use it to prevent or reduce secondary collisions [14]. Mussone et al. used traffic flow and meteorological data to predict traffic accident ratings, using back-propagation neural network models and generalized linear mixed models for analysis, demonstrating that these variables play a role in predicting severity levels[10].

While not very frequent, the random forest was used in crash severity modeling and prediction and its performance was reported satisfying. Iranitalab et al. compared the performance of four statistical and machine learning methods, and nearest neighbor classification, random forest and support vector machines have better performance, while the multinomial logit is the weakest method [7]. Das et al. using random forest model to identify the factors affecting crash severity on arterial corridors, and they use the model identify roadway locations where severe crashes tend to occur [15].

To the best of our knowledge, this is the first study that has applied the random forest model and neural network model to predict the traffic accident severity in China. This investigation used randomForest and neuralnet two packages in the R software. A random forest predictive model and a neural network predictive model were established by using 7,795 traffic accidents and local real-time meteorological variables in five regions of the same city within ten years. In the modeling process, the model was optimized by traversing the method of comparing and observing the visual graphics by selecting appropriate parameters.

The study defined input and output variables and normalizes the data to eliminate dimensional relationships between variables, making the data comparable. Then, 75% of the training samples were divided from the data to establish a prediction model, and the remaining 25% of the test samples were used for testing. The results of our study found that the random forest model has a better predictive effect than the neural network model, which is similar to the results of other predictive models studies [7, 10]. And the accuracy of our random forest model prediction is about 50%. The accuracy for general traffic accident prediction is 54%. It should be noted that although the neural network is relatively low in overall prediction, the accuracy of prediction for minor traffic accidents is 65%. The prediction probability of this study is slightly higher than some studies [16]. And the results have reached the prediction level of some of the same machine learning studies[2, 15].

This study used the real-time meteorological and traffic factors to predict road safety and reduces the deficiencies of previous studies using daily average data. And these models do not need to include drivers and vehicle information such as driver age, blood alcohol content and vehicle type. In fact, these data are often difficult to obtain (before a traffic accident occurs. But real-time weather data is easy to obtain in advance and is very accurate.

Of course, the higher weight of the variables in the model does not mean that the correlation is higher. In machine learning, the variable with the highest weight means that the variable is more important in the model construction, and the relationship between the input and output variables does not necessarily

have the same correlation. The weights in machine learning model do not directly explain the quantitative relationship between variables. Although that, it is undoubted that these meteorological factors play an important role in causing serious consequences of road traffic accidents. Some of these interrelationships have been verify in early studies. A Southern California study using real-time meteorological and flow data found that collisions are more likely to occur on wet roads[17]. Abdel-Aty et al. [18] found that the damage severity was high under low visibility conditions, while front and back collisions were the main types of collision accidents. Hermans suggested that increased gusts and longer duration of precipitation are associated with increased numbers of crashes [19]. And while the correlation between other relevant meteorological variables and traffic accidents is not fully understood, our results might provide direction for further research.

It needs to be emphasized that there were also several limitations in our study. First, this paper used only one city's data. Different regions may have different economic, transportation and climate conditions. The promotion of models requires more urban data to compare. Second, there are too few particularly serious road traffic accidents to effectively modeling. In fact, in the Suzhou traffic accident database, particularly serious road traffic accidents accounted for only 7% of all accidents. And when we built the model, we randomly selected 25% of the accidents to verify the accuracy of the predictions, which directly affects the overall effectiveness of the model and cause the part of the particularly serious traffic accident model cannot apply. Third, the phenomena (the relationship between input and output) in the machine learning model were not well known and did not provide an analytical formula between input and input. The quantitative relationship between meteorological variables and road traffic accidents requires further research. Moreover, this study is lacks some information about roads, drivers, and vehicles information, including them in the model may effectively improve the predictive performance.

Conclusions

This study described how to use real-time meteorological factors to predict traffic accidents based on the R software, and to provide reliable information for researchers in the public health and public safety fields. The results of this study show that for the prediction of the severity of traffic accidents using meteorological data, the random forest model may be more suitable than the neural network model. Future research will include more comprehensive variables and different city data to compare, and the possibility of other models should also be considered.

Declarations

Ethics approval and consent to participate

Prior to data collection, this study was approved by the ethics committee from the Chinese Center for Disease Control and Prevention Institute for Environmental Health and Related Product Safety (201606).

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by special foundation of basic science and technology resources survey of ministry of science and technology of China (No.2017FY101200).

Authors' contributions

Conceptualization: ML; Methodology: ML, YP; Formal Analysis: ML; Resources: PY, YW, ZY, LD, YS; Writing – Original Draft Preparation: MM; Writing – Review and Editing: PY, YS, PB and LD; Supervision: TS; ST; YZ; ML;GG; Project Administration: PY, YW, YS, LD; Funding Acquisition: YS, LD. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Authors' information

Pengpeng Ye3, E-mail yepengpeng@ncncd.chinacdc.cn

Leilei Duan 3, E-mail：duanleilei@ncncd.chinacdc.cn.

Peng Bi4, E-mail：peng.bi@adelaide.edu.au

Yehuan Sun * 1, 5：E-mail：yhsun_ahmu_edu@yeah.net

1.Department of Epidemiology and Health Statistics, School of Public Health, Anhui Medical University, No. 81 Meishan Road, Hefei 230032, Anhui, P.R.China

2.Anhui Public Meteorological Service Center, No.16 Shihe Road , Hefei 230011, Anhui, P.R. China

3.Division of Injury Prevention and Mental Health, National Center for Chronic and Noncommunicable Disease Control and Prevention, Chinese Room 916, No.27 Nanwei Road, Xicheng District, Beijing, P.R. China, 100050

4.School of Public Health, The University of Adelaide, Level 8, Hughes Building, North Terrace Campus, Adelaide, SA 5005, Australia

5.Center for Evidence-Based Practice, Anhui Medical University, No. 81 Meishan Road, Hefei 230032, Anhui, P.R. China

References

1. Theofilatos A and Yannis G. A review of the effect of traffic and weather characteristics on road safety. Accid Anal Prev; 2014; (72): 244-256.
2. Zeng Q, Gu W, Zhang X, Wen H, Lee J and Hao W. Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors. Accid Anal Prev; 2019; (127): 87-95.
3. Bailey TC and Hewson PJ. Simultaneous modelling of multiple traffic safety performance indicators by using a multivariate generalized linear mixed model. J. R. Stat. Soc. A Stat.; 2004; (167): 501-517.
4. Zhai X, Huang H, Sze NN, Song Z and Hon KK. Diagnostic analysis of the effects of weather condition on pedestrian crash severity. Accident; analysis and prevention; 2019; (122): 318-324.
5. Wainberg M, Alipanahi B and Frey BJ. Are Random Forests Truly the Best Classifiers? J. Mach. Learn. Res.; 2016; (17).
6. Statistics SBO. Suzhou Statistical Yearbook 2017, 2018.
7. Iranitalab A and Khattak A. Comparison of four statistical and machine learning methods for crash severity prediction. Accident; analysis and prevention; 2017; (108): 27-36.
8. Chen X and Ishwaran H. Random forests for genomic data analysis. Genomics; 2012; (99): 323-329.

9. Wang Y, Song Q and Du Y, et al. A random forest model to predict heatstroke occurrence for heatwave in China. Sci. Total Environ.; 2019; (650): 3048-3053.
10. Mussone L, Bassani M and Masci P. Analysis of factors affecting the severity of crashes in urban road intersections. Accident; analysis and prevention; 2017; (103): 112-122.
11. Lo SC, Freedman MT, Lin JS and Mun SK. Automatic lung nodule detection using profile matching and back-propagation neural network techniques. J. Digit. Imaging; 1993; (6): 48-54.
12. Abdelwahab H and Abdel-Aty M. Investigating the effect of light truck vehicle percentages on rear-end fatal traffic crashes. J. Transp. Eng.; 2004; (130): 419-428.
13. Malyshkina NV and Mannering FL. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accident; analysis and prevention; 2010; (42): 131-139.
14. Wang J, Liu B, Fu T, Liu S and Stipancic J. Modeling when and where a secondary accident occurs. Accident; analysis and prevention; 2018.
15. Das A, Abdel-Aty M and Pande A. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. J. Safety Res.; 2009; (40): 317-327.
16. Gebers MA and Peck RC. Using traffic conviction correlates to identify high accident-risk drivers. Accid Anal Prev; 2003; (35): 903-912.
17. Golob TF and Recker WW. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions[J]. Journal of transportation engineering. J. Transp. Eng.; 2003; (129): 342-353.
18. Abdel-Aty M, Ekram AA, Huang H and Choi K. A study on crashes related to visibility obstruction due to fog and smoke. Accid Anal Prev; 2011; (43): 1730-1737.
19. Hermans E, Brijs T, Stiers T and Offermans C, The Impact of Weather Conditions on Road Safety Investigated on an Hourly Basis, Transportation Research Board 85th Annual Meeting, 2006.

Tables

Table 1 Description of the model variables and data sources.

Variables	Units	Abbreviation	Type	Data source
Basic information				
City		city	Character	Statistical Yearbook
Highway mileage(km)	km	mile	Numerical	
Car ownership		car	Numerical	
Total population		population	Numerical	
Meteorological variables				
Observation code		code1	Numerical	National Meteorological Information Center
Year of observation		year1	Date	
Month of obsevation		month1	Date	
Specific time of observation		hour1	Date	
Real-time atmospheric pressure	hPa	pressure	Numerical	
Real-time temperature	℃	tem	Numerical	
Real-time relative humidity	%	humid	Numerical	
Real-time precipitation	mm	precipitation	Numerical	
Real-time wind direction	°	direction	Numerical	
Real-time wind speed	m/s	speed	Numerical	
Real-time visibility	m	visibilty	Numerical	
Real-time evaporation amount (mm)	mm	eva	Numerical	
Real-time snow depth (cm)	cm	dep	Numerical	
Total cloud amount	%	tc	Numerical	
Low cloud amount	%	lc	Numerical	
Road traffic accident variables				
Administrative division code		code2	Numerical	Traffic police detachment
Expressway or not		exprss	Categorical	
Highway grade		highway	Categorical	
Non-motor vehicle or not		non-motor	Categorical	
Highway location	km	highloc	Numerical	
Year		year2	Date	
Month		month2	Date	
Accident specific time		hour2	Time	
Road number		road	Numerical	
Accident severity				
Total death		death	Numerical	National Meteorological Information Center
Total injured		injured	Numerical	
Direct property loss	yuan	loss	Numerical	
Accident severity		severity	Categorical	

Table 2 Machine learning models prediction proportion

	RF model	BPNN model
Error rate	51.09%	61.01%
Minor	48.80%	35.84%
General	45.97%	59.33%
Serious	58.88%	93.63%
Particularly serious	100.00%	100.00%

Figures

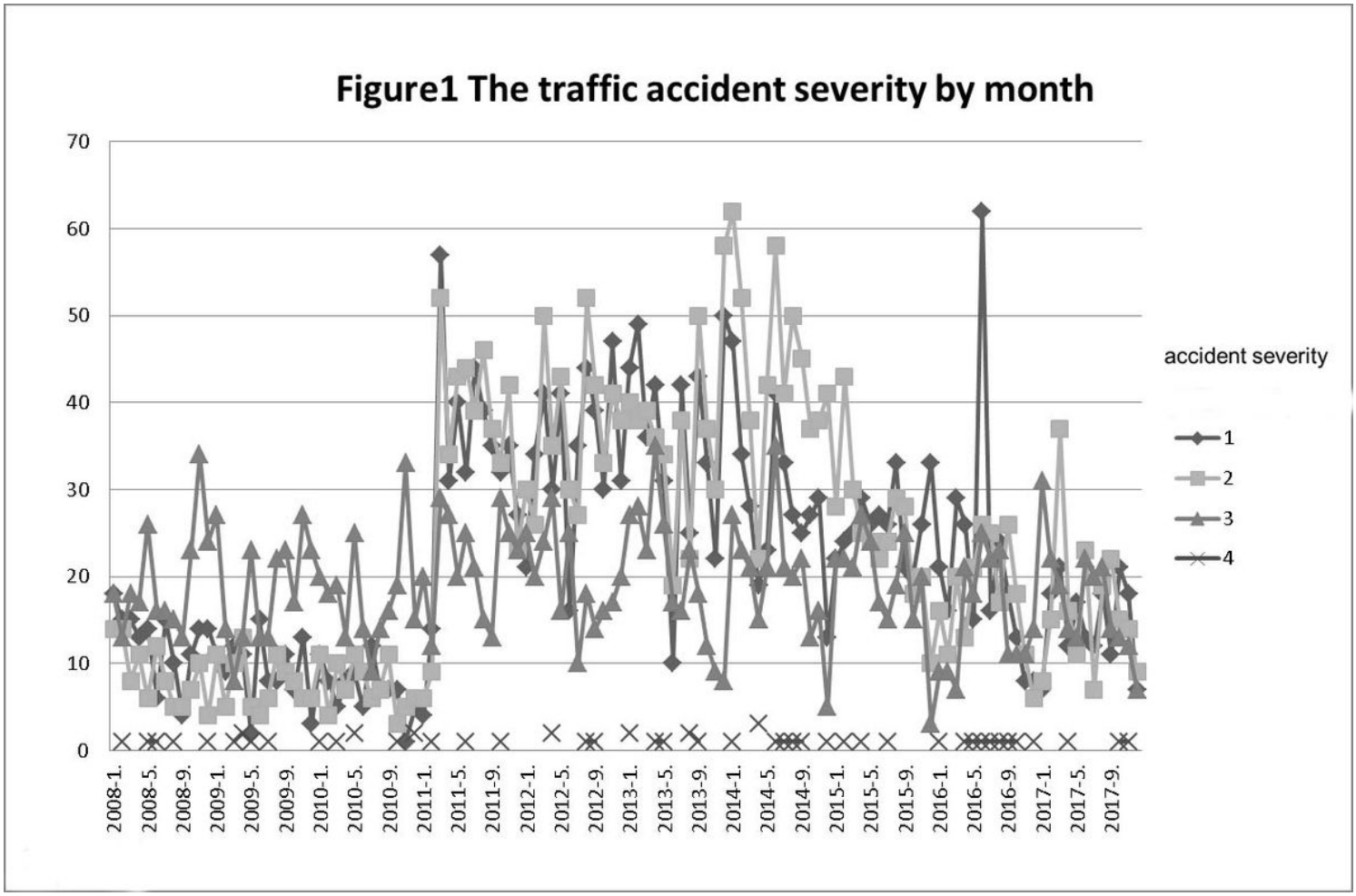
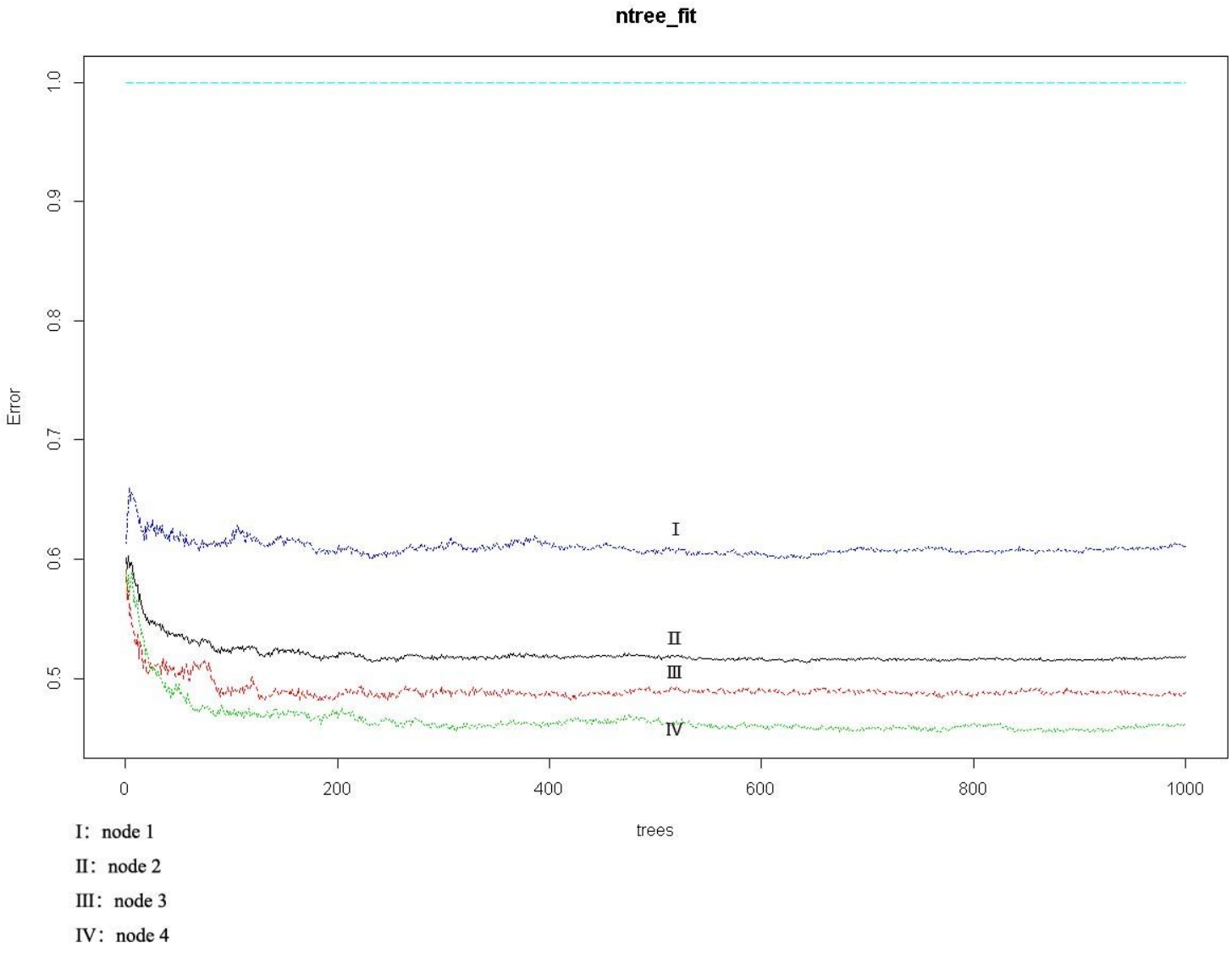


Figure 2



I: node 1

II: node 2

III: node 3

IV: node 4

Figure 3

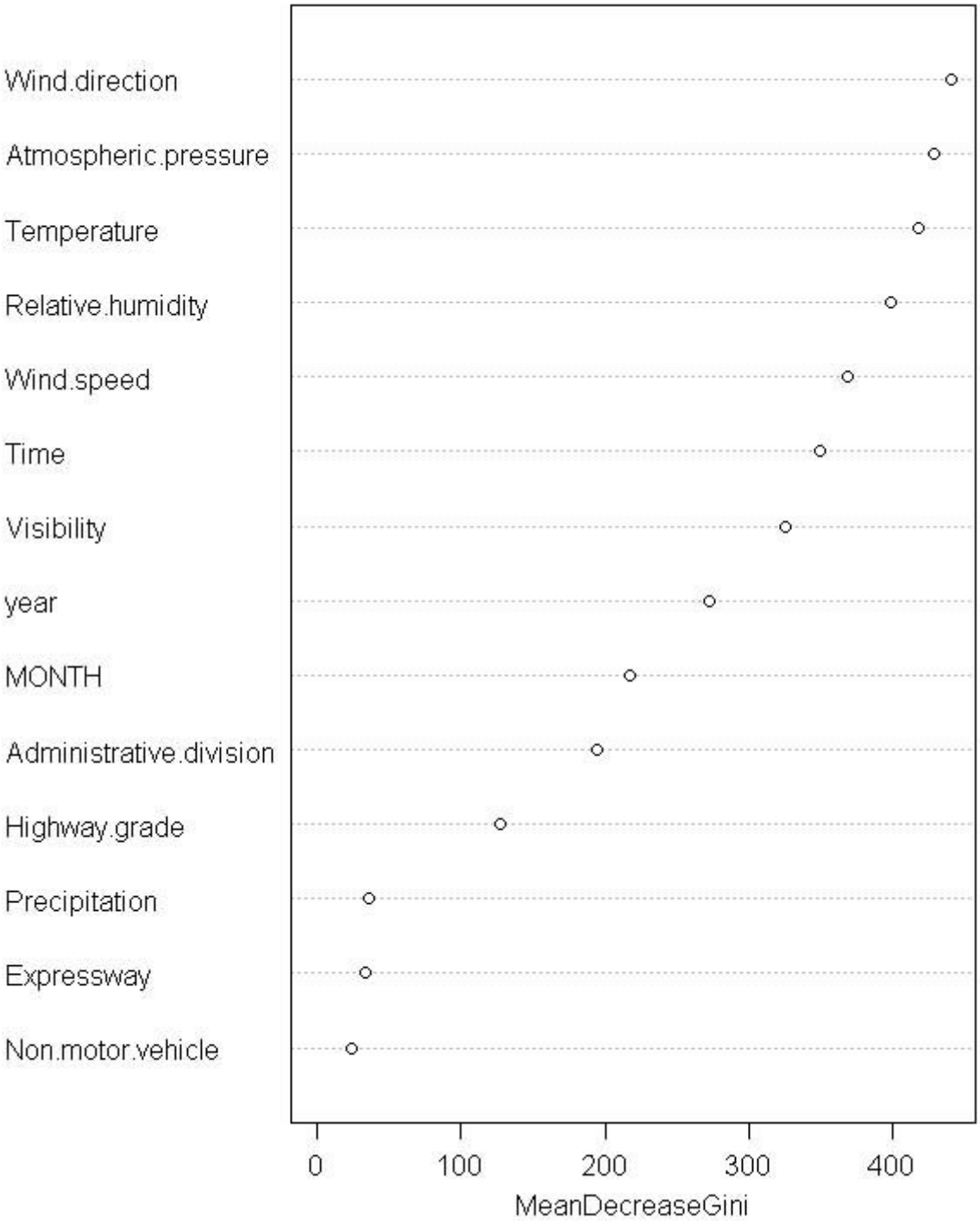

Wind.direction
Atmospheric.pressure
Temperature
Relative.humidity
Wind.speed
Time
Visibility
year
MONTH
Administrative.division
Highway.grade
Precipitation
Expressway
Non.motor.vehicle
0
100
200
300
400
MeanDecreaseGini

Figure 5

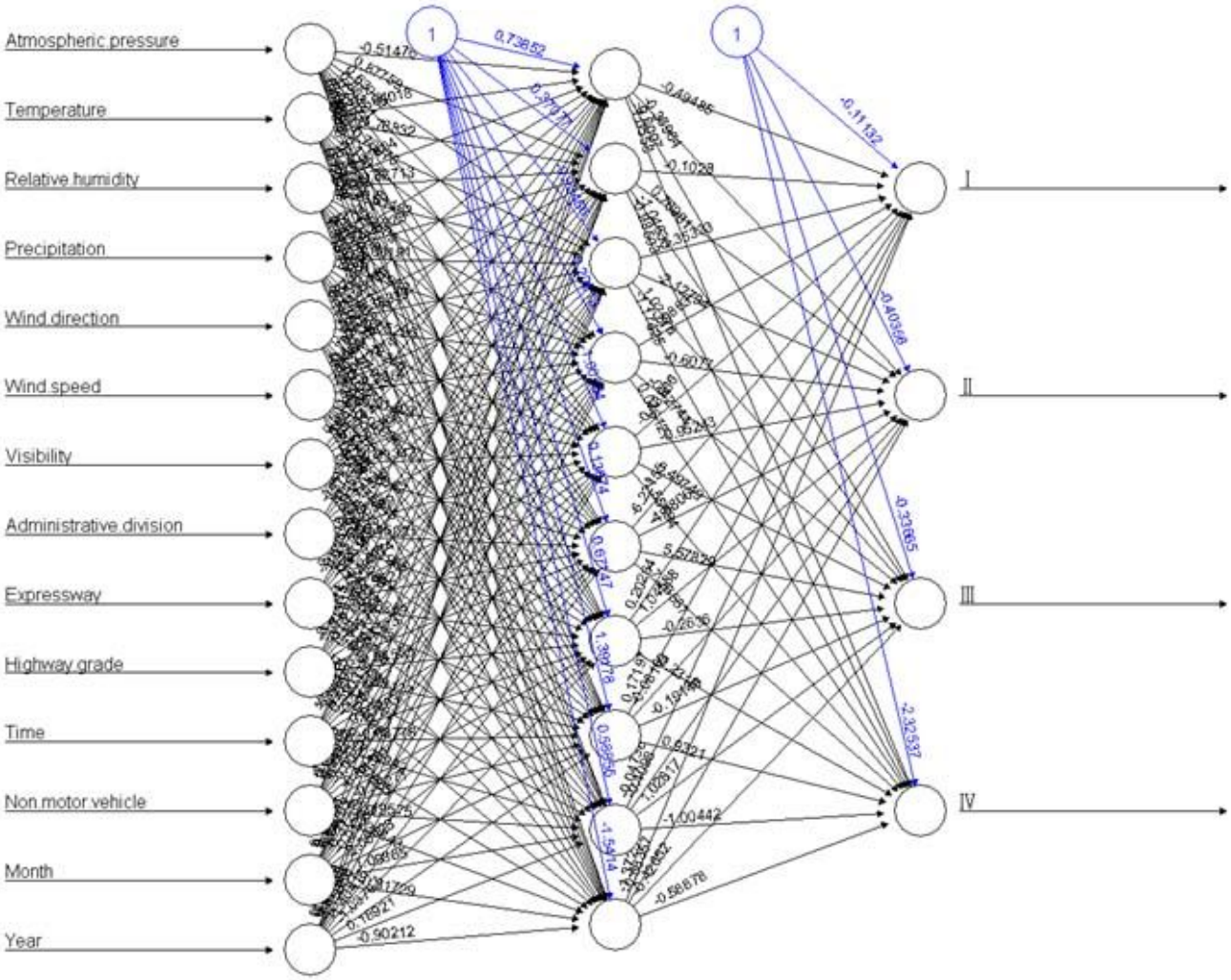

Atmospheric pressure
Temperature
Relative humidity
Precipitation
Wind direction
Wind speed
Visibility
Administrative division
Expressway
Highway grade
Time
Non motor vehicle
Month
Year
1
1
I
II
III
IV

Figure 8

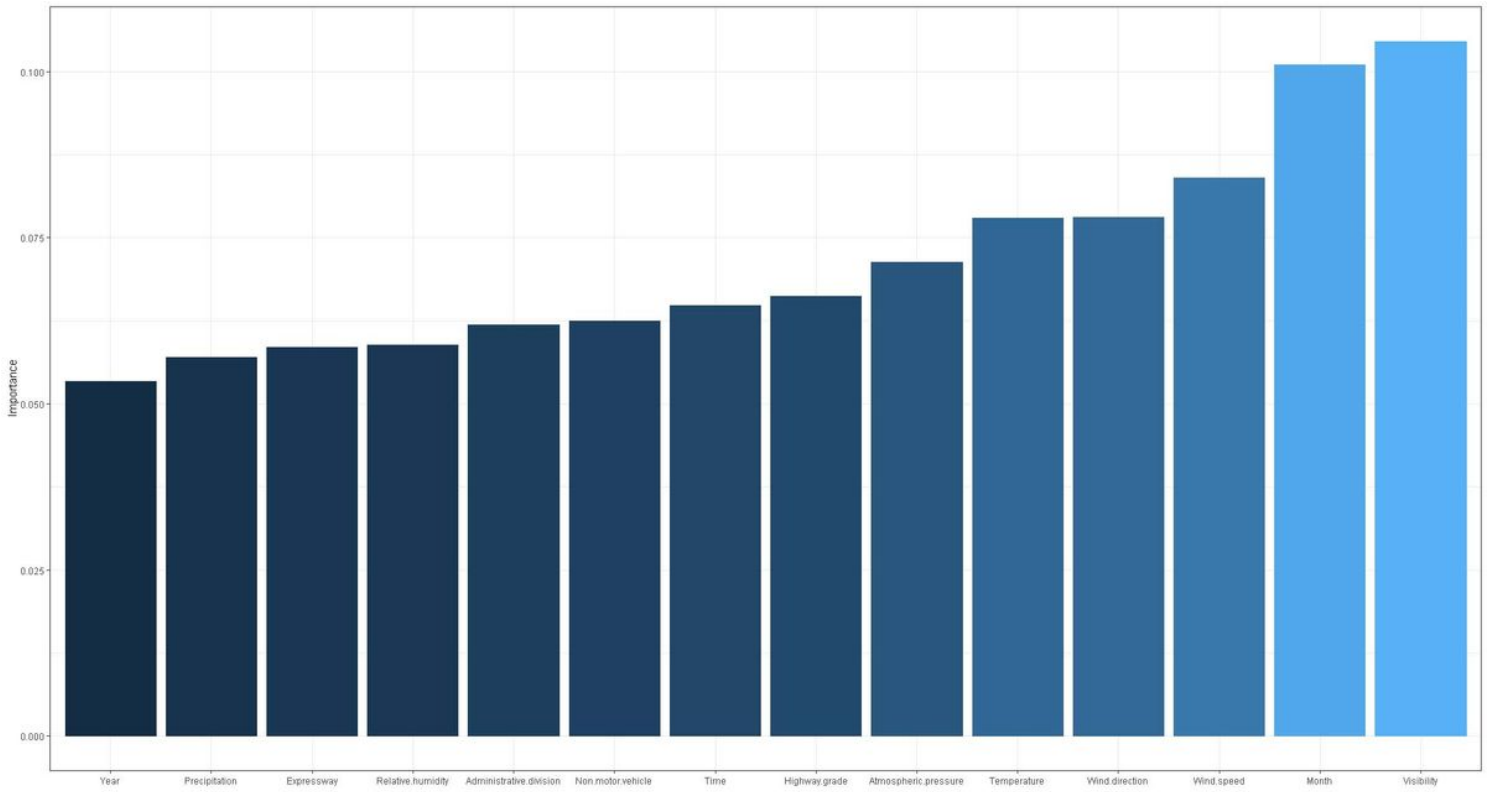

Importance
0.100
0.075
0.050
0.025
0.000
Year
Precipitation
Expressway
Relative.humidity
Administrative.division
Non.motor.vehicle
Time
Highway.grade
Atmospheric.pressure
Temperature
Wind.direction
Wind.speed
Month
Visibility

Figure 10