Research on Urban Road Traffic Accident Characteristics and Countermeasures: A Case Study of Ningbo City

Xiaoxia Xiong1*, Shuichao Zhang2 and Yujie Shen3
1 School of Automotive and Traffic Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang, China
2 School of Civil and Transportation Engineering, Ningbo University of Technology, 201 Fenghua Road, Ningbo, China
3 Automotive Engineering Research Institute, Jiangsu University, 301 Xuefu Road, Zhenjiang, China
Email: x_xiong623@163.com

Abstract. Current road traffic safety situation is still challenging globally and needs continuous public attention. Exploring the characteristics of road traffic accidents could improve the understanding of accident causation mechanism and help develop countermeasures to reduce the accident occurrences. The paper first explores the general characteristics of urban road traffic accidents from the perspectives of temporal, spatial, and personal distributions. Random Forest (RF) Model is then utilized to identify important features affecting the occurrence of traffic accidents. Accordant countermeasures are finally discussed and proposed.

1. Introduction
Although road safety has been improved over the past years, the number of human deaths caused by road traffic accidents remains as high as 1.35 million per year worldwide (equivalent to one death from a road traffic accident every 25 seconds) according to World Health Organization (WHO) [1]. Therefore, the global road traffic safety situation is still challenging and demands continuous attention. In China, descriptive statistics such as “frequency of accidents”, “number of injured persons”, “number of deaths” and “property loss” are often utilized as main indicators in assessing road traffic safety, which essentially lack in cross-analysis and in-depth mining of accident data [2]. Analysing road traffic accidents from multiple angles, such as time and space characteristics [3-4], could provide insight on rules of accident occurrence and prevention countermeasures. In this context, the paper first explores the general characteristics of urban road traffic accidents from the perspectives of temporal, spatial, and personal distributions, and then Random Forest (RF) Model is introduced to identify important features affecting the occurrence of traffic accidents, and finally potential countermeasures are discussed and proposed accordingly.

2. Analysis of General Characteristics of Urban Road Traffic Accidents
A total of 1,889 records were retrieved from the road traffic accident database of Ningbo City, including accidents with property damages and/or casualties reported to the public security traffic management department, which constitute the dataset for analysis in this paper. These records cover 10 administrative regions (including Beilun, Cixi, Yinzhou, Fenghua, Haishu, Jiangbei, Ninghai, Xiangshan, Yuyao, and Zhenhai) of Ningbo, and span from Jan 1st to Dec 31st of year 2018. The
retrieved accident attributes include accident date and time, accident location, accident type, as well as age and type of person involved. In this section, the analysis of road traffic accidents would be carried out from three aspects, including temporal distribution, spatial distribution and personal characteristics.

2.1. Temporal Distribution
Figure 1 presents the distribution of road traffic accident counts by date and month (divided into early, middle, and late month), respectively. It could be noted that Nov and Dec have significantly higher number of accidents compared to other months, indicating weather could be an important factor affecting accident occurrence.

Figure 2 presents the distribution of road traffic accident counts by week and hour. It could be noted that higher proportion of accidents happened in workdays (which usually feature commuting traffic) compared to weekends. 18:30-21:30 (which usually features dinner trips and entertainment activities) also shows generally higher accident counts than other time slots.

2.2. Spatial Distribution
Figure 3 shows regional distribution of road traffic accidents in Ningbo. It could be seen that Cixi has the highest number of accidents city-wide, but no significant spatial accumulation characteristics are observed across regions. Figure 4 presents the trend in number of road traffic accidents from the 1st to 4th quarter of year by region, which features a similar changing pattern across regions where a rise generally shows up in the 4th quarter, indicating the overall monthly distribution pattern (obtained in section 2.1) generally applies to different regions within Ningbo.
2.3. Person Characteristics

Figure 5 shows the distribution of road traffic accidents by involved person types (which are divided based on person attributes recorded in the accident reports, including retirees and students, employed staff, local workers, farmers, migrant workers, and unemployed person). It could be noted that unemployed and migrant workers constitute the majority of accident-involved persons. Figure 6 shows the distribution of involved person types across different accident types (which are divided into vehicle-and-vehicle accident, pedestrian-and-vehicle accident, and single vehicle accident according to the subject of the accident). It could be seen that migrant workers and farmers have a slightly larger proportion in single vehicle accident than in other types of accidents, indicating possible differential accident occurrence mechanism across types of person.

Figure 7 shows the distribution of road traffic accidents according to the age of involved persons, with (a) presenting original counts and (b) presenting Gaussian kernel density estimation curve (with bandwidth set at one). It could be noted that both single vehicle accidents and vehicle-and-vehicle accidents generally show higher proportion in persons aging around 25 and 50, while pedestrian-and-vehicle accidents show another peak around age 60, indicating more elderly persons have been involved in the pedestrian-and-vehicle accident type.
3. Analysis of Urban Road Traffic Accident Types

Although statistical data in the last section could facilitate general understanding of road traffic accidents in terms of time, place, and person characteristics, no clue could be obtained as to the importance of these features in accident occurrence. In this section, Random Forest (RF) Classifier is employed to examine important features that could impact on accident types. Random Forest was first proposed by Leo Breiman and Adele Cutler, which simultaneously trains multiple decision trees using bootstrapped samples [5]. Random forest can deal with multi-collinearity problems of variables well without feature selection, as well as effectively calculate the importance of model variables [6], and thus was selected for accident type analysis here.

It should be noted that veh-veh type accidents dominated the whole dataset (with proportion at 76.0%), followed by ped-veh type (21.1%) and single vehicle type (2.9%). Such imbalanced data set is likely to lead to classification results biased towards dominated samples (i.e., resulting in higher overall accuracy while the correction rate for small categories is low), as traditional machine learning algorithms generally aim to reduce the overall classification accuracy and the misclassification penalty function for all samples are treated the same [7]. To overcome such limit, class weight was employed in the model training stage to deal with the imbalance in accident type data, where weight value is calculated according to the reciprocal of the occurrence frequency of each category in the data, as in equation (1).

$$w_j = \frac{n}{kn_j}$$  

where $w_j$ is the weight of Class j ($j=1,2,3$ in our case, representing veh-veh type, ped-veh type, and single vehicle type, respectively); n represents the total number of samples, $n_j$ represents the number of samples in Class j, k is the number of categories (classes).

Scikit-learn module was utilized for realizing Random Forest algorithm based on Python platform [7]. As one decision tree in RF is trained using a subset of the whole sample space (bootstrapped samples), there remains a sample subset that could be served as a test set for each decision tree, which is referred to as “Out-of-Bag” sample and is usually used to evaluate the overall performance of RF classifier [5]. Thus, OOB classification error was used here as the performance indicator to determine the optimal collection of variables and number of decision trees (n_estimators) in RF algorithm. Also,
to avoid potential over-fitting problem, the retrieved samples were divided into a training set (75%) and a test set (25%) to validate the classification results.

Based on the training set, {Region, Age, Time, Month} were finally selected as input variables to RF classifier by trial-and-error, and the accordant OOB error curve is presented in figure 8. As a larger number of decision trees does not show significant decrease in OOB error when n_estimators exceeds 40, n_estimators was finally set to 40 where the OOB accuracy rate is 71.1%.

![Figure 8. Out-of-bag classification error curve.](image)

Precision, Recall, and F1-score (see equation 2 to 4) were calculated for the test set to examine the classification performance of RF by categories, with the prediction results presented in table 2.

\[
\text{Precision}(C_j) = \frac{TP_j}{\text{Pred}(C_j)} = \frac{TP_j}{TP_j + \sum_{m \neq j} FP_{mj}}, \quad m, j = 1, 2, 3
\]  
\[
\text{Recall}(C_j) = \frac{TP_j}{\text{True}(C_j)} = \frac{TP_j}{TP_j + \sum_{m \neq j} FP_{jm}}, \quad m, j = 1, 2, 3
\]  
\[
\text{f1-score}_j = \frac{2}{\frac{1}{\text{Precision}_j} + \frac{1}{\text{Recall}_j}}, \quad m, j = 1, 2, 3
\]

where TPj represents the number of correctly predicted records within Class j, FPmj represents the number of records within Class m wrongly predicted to Class j (m \neq j), which could be illustrated by the schematic diagram of prediction results presented in table 1.

| True Class | C1       | C2       | C3       |
|-----------|----------|----------|----------|
| Predicted Class | C1       | FP1(C1→C1) | FP21(C2→C1) | FP31(C3→C1) |
| C2       | FP12(C1→C2) | TP2(C2) | FP32(C3→C2) |
| C3       | FP13(C1→C3) | FP23(C2→C3) | TP3(C3) |

Table 1. Schematic diagram of prediction results.
Table 2. Prediction results based on test set.

| Acc_Type   | Precision | Recall | f1-score |
|------------|-----------|--------|----------|
| Ped-veh    | 0.65      | 0.61   | 0.63     |
| Sin.veh    | 0.50      | 0.43   | 0.46     |
| Veh-veh    | 0.77      | 0.75   | 0.76     |
| Accuracy   |           |        | 0.71     |

It could be noted that except for single vehicle crash type (which features the smallest proportion, i.e., 2.9% of the sample space), f1-score for both Veh-Veh and Ped-Veh exceed 0.60, with the overall accuracy exceeding 70%. Note that the main purpose of establishing the RF model here is not for accurate classification (which could be vital to other application scenarios such as online driving risk state classification), but to examine important factors that could affect different types of accidents, and to further enlighten on probable measures in preventing these accidents. And thus the established RF model is considered acceptable for the subsequent qualitative analysis. More detailed data collection (with more attributes such as age, classes of road, etc.) in the future is expected to improve the classification accuracy. Feature importance list obtained from the established RF classifier is presented in figure 9.

![Figure 9. Important features from RF estimation.](image)

It could be seen from figure 9 that the top 10 important features affecting road accident type include: F1: Time=0:00-5:00, F2: Month=Dec, F3: Age=36~59, F4: Time=18:3~21:30, F5: Month=Nov, F6: Region=Cixi, F7: Age>=60, F8: Region=Beilun, F9: Time=16:30~18:30, F10: Time=14:00~16:30. Given these important features, by examining their distribution within each accident type (as shown in figures 10 to 13), insight could be given on countermeasures to prevent different types of accidents.

![Figure 10. Hourly distribution in different accident type categories.](image)

![Figure 11. Monthly distribution in different accident type categories.](image)
Figure 10 shows that 0:00-5:00 time period has the highest proportion in single vehicle accident type compared to other two types, indicating that F1: Time=0:00-5:00 could be critical in developing single vehicle accident prevention strategies. Possible countermeasures could be stricter traffic monitoring by traffic police department in this time period (such as setting higher penalty on behaviours like speeding and incompliance with traffic signals), as well as realizing more timely emergency rescue when single vehicle accident happens.

Figure 11 shows that December is also significantly higher in single vehicle accident type compared to other two types, indicating that F2: Month=Dec could also be an important factor that should be considered in preventing single vehicle accidents (probably due to the low temperature that could impact on friction of road surface and vehicle performance). More safety promotion activities could be carried out around Dec to improve the drivers’ safety awareness concerning season of winter.

Figure 12 shows that 36-59 group has significantly smaller proportion in ped-veh accident compared to other two types, probably due to their smaller chance falling in distraction compared to younger group (e.g., due to smaller proportion in cellphone use), as well as their more agile moving capabilities compared to older people. This implies that F3: Age=36-59 deserves more research in the future to retrieve meaningful strategies for preventing pedestrian related accidents.

Figure 13 shows that Cixi features higher proportion in veh-veh accident compared to other two types, indicating that F6: Region=Cixi could be a region where special attention should be paid to vehicle interaction management, such as devoting more efforts in improving traffic organization in the area. More analysis could be made for other important features following the same way as discussed above, and is not presented in detail here due to space limit of the paper. Such in-depth analysis could facilitate identification of potential countermeasures at accident-type specific level, which could be beneficial to policy planning and formulation for traffic safety management department.

4. Conclusion
Based on the reported road traffic accident data collected in Ningbo of 2018, characteristics of urban road traffic accidents are explored from temporal, spatial, and personal perspectives. A random forest based on analysis framework is further proposed to identify important features affecting accident occurrences and to develop prevention countermeasures. More research efforts should be devoted to extensive data collection and in-depth data analysis on road traffic accidents in the future.

Acknowledgments
The research is supported by Open Project of Key Laboratory of Ministry of Public Security for Road Traffic Safety with Grant No. 2020ZDSYSKFKT11. The authors also thank Traffic Police Bureau of Ningbo Municipal Public Security Bureau for providing the traffic accident source data.

References
[1] World Health Organization 2018 Global Status Report on Road Safety (Geneva: World Health Organization)
[2] Sun Y X 2014 *Traffic Accident Analysis Based on Data Mining* (Beijing: Beijing Jiaotong University)

[3] Guo L, Zhou J B, Dong S, Zhang S C 2018 Analysis of urban road traffic accidents based on improved K-means algorithm *China Journal of Highway and Transport* **31** (4) 270-279

[4] Guo S, Huang F and H Yun 2014 Analysis of urban traffic accident character: A case study of Urumqi *Traffic & Transportation* (30) 187-191

[5] Breiman L 2001 Random forests *Mach. Learning* **45** (1) 5–32

[6] Fu C Y and Wang J H 2017 A real-time accident risk model on freeways based on monitoring data *Journal of Transport Information and Safety* **5** (35) 11-17+36

[7] Chris A 2018 *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning* (Cambridge: O'Reilly Media)