Accident Analysis and Prevention 146 (2020) 105744

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: A sequential approach of multivariate imputation by chained equations and random forest classifier

Linchao Li [a], Carlo G. Prato [b,*], Yonggang Wang [c]

[a] College of Civil and Transportation Engineering, Shenzhen University, Shenzhen, Guangdong, 518060 People's Republic of China
[b] School of Civil Engineering, The University of Queensland, St. Lucia 4072, Brisbane, Australia
[c] School of Highway Chang'an University Xi'an, Shann'xi, 710064 People's Republic of China

ARTICLE INFO

Keywords:
Missing values
Multiple imputation
Machine learning
Traffic safety
Mountainous roads

ABSTRACT

The estimation of the effect of contributors to crash injury severity and the prediction of crash injury severity outcomes suffer often from biases related to missing data in crash datasets that contain incomplete records. As both estimation and prediction would greatly improve if the missing values were recovered, this study proposes a sequential approach to handle incomplete crash datasets and rank contributors to the injury severity of crashes on mountainous freeways in China. The sequential approach consists of two parts: (i) multivariate imputation by chained equations imputes the missing values of independent variables; (ii) a random forest classifier analyses the correlation between the dependent and the independent variables. The first part considers different imputation methods in light of the independent variables being either binary, categorical or continuous, whereas the second part classifies the correlations according to the random forest classifier. The proposed method was applied to the case-study about mountainous freeways in China and compared to the analysis of the raw dataset to evaluate its effectiveness, and the results illustrate that the method improves significantly the classification accuracy when compared with existing methods. Moreover, the classifier ranked the contributors to the injury severity of traffic crashes on mountainous freeways: in order of importance vehicle type, crash type, road longitudinal gradient, crash cause, curve radius, and deflection angles. Interestingly, a lower importance was found for environmental factors.

1. Introduction

Road crashes cause a large amount of economic losses and a large number of injuries in both developed and developing countries. For example, automobile crashes in China caused in 2016 an economic loss of about 1.21 billion Chinese yuan, as well as 226,430 injuries and 63,093 fatalities (National Bureau of Statistics of China, 2017). Recently, road safety of mountainous regions in China has attracted increasing attention in light of the rapid development of mountainous freeways in the western regions. The frequently adverse weather and the complex environmental conditions increase the probability of crash occurrence as well as the likelihood of more severe crash outcomes (Ma et al., 2015; Peel et al., 2017; Huang et al., 2018; Wang et al., 2019a). In fact, more than 22 % of the total Chinese traffic fatalities occur on mountainous freeways, even though they account for only 2.2 % of the kilometres on the total Chinese freeways (Huang et al., 2018). Moreover, the mountainous location makes rescue operations more difficult, which implies that the congestion is severe and the duration of the impact on traffic is long (Ahmed et al., 2011). Given these premises, it is necessary to understand the major contributors to crash injury severity to provide benefit to the mountainous freeway management and operation, their design, and the possible regulations to have in place to curtail this major road safety problem.

When studying crash injury severity, statistical methods have been traditionally applied. Initially, binary methods such as binary logit or probit models have been widely used to study factors contributing to crash injury severity (see, e.g., Abdel-Aty, 2001; Dissanayake and Lu, 2002; Yu and Abdel-Aty, 2014). Then, models have been extended to multinomial methods to consider multiple injury severity categories (see, e.g., Wang et al., 2019a; Wang and Prato, 2019b) and also to

* Corresponding author.
E-mail address: c.prato@uq.edu.au (C.G. Prato).

https://doi.org/10.1016/j.aap.2020.105744
Received 3 January 2020; Received in revised form 24 July 2020; Accepted 14 August 2020
Available online 27 August 2020

accommodate for heterogeneity (see, e.g., Çelik and Oktay, 2014; Kaplan and Prato, 2012a; Linchao and Fratrović, 2016). Considering the nature of severity outcomes, ordered methods such as ordered logit and generalized ordered logit models have been estimated to account for the inherent relations within the dependent variable (see, e.g., Abdel-Aty, 2001; Quddus et al., 2002; Kaplan and Prato, 2012b; Ye and Lord, 2014; Chen et al., 2016).

In recent years, the realization that data about traffic crashes are quite large has suggested that perhaps traditional statistical methods may not be suitable to mine these large datasets and find contributors to crash injury severity. One limitation that is often mentioned is that statistical regression methods rely on restrictive assumptions and predefined relationships between the dependent and the independent variables that are difficult to validate in real-world crash datasets. Another limitation is that they are limited in handling variables with a large number of categories (see, e.g., Delen et al., 2006; Deka and Quddus, 2014; Scott-Parker et al., 2015; Li et al., 2017). With the aim of overcoming the drawbacks of statistical methods, non-parametric and artificial intelligence methods can be applied to rank the importance of injury severity contributors. Artificial neural networks (ANNs) have been applied for the classification of crash injury severity and the ranking of key contributors (see, e.g., Chen et al., 2012; Bai et al., 2017), and support vector machine (SVM) methods have also been implemented to solve the crash injury severity classification problem (see, e.g., Li et al., 2012; Yu and Abdel-Aty, 2013). Decision tree and imputed value sampling based on a measure of correlation was used to impute missing values in crash databases (Deb and Liew, 2016). Random forest (RF) classifiers, non-parametric methods without any predefined relationship, have been also employed, and it has been proved that they can explore effectively detailed datasets describing road characteristics, driver characteristics, vehicle characteristics, and environmental conditions (see, e.g., Abdel-Aty and Haleem, 2011; Lin et al., 2015; Shi and Abdel-Aty, 2015).

All the aforementioned methods have been demonstrated to have powerful and adaptive analysis capabilities to draw valuable conclusions on the factors that aggravate or mitigate crash injury severity. All the methods, whether statistical or artificial intelligence ones, require high quality datasets that are unfortunately often incomplete in the real world because of large quantities of missing values that are related to human errors, equipment malfunctioning, or faulty data transmission. This problem is more serious in developing countries where policemen who record the data might not be properly trained and the documentation and retrieval system might not be technically advanced (Al-Madani, 2018). Missing values in crash datasets cause loss of information as sample sizes become small and consequently standard errors become larger in statistical methods and overfitting becomes common in artificial intelligent methods.

To deal with missing values in crash datasets, different methods have been compared recently for crash frequency analysis (see, e.g., Afghari et al., 2019). The most common methods to handle missing values in crash injury severity analysis are case-wise deletion, mean imputation, class mean imputation and k-nearest neighbour. Case-wise deletion directly deletes incomplete records, at the price though of reducing sample size and increasing standard errors of the estimates (Li et al., 2017). The limitation of the method was noticeable when the missing values were not necessarily at random (Lord, 2006; Lord and Miranda-Moreno, 2008). Mean imputation, where the missing values of one variable are imputed by mean (continuous variable) or mode (categorical variable) of non-missing values of the same variable (Bhat, 1994). The method is popular because of its easy implementation and it is currently used (see, e.g., Ding et al., 2019; Shen et al., 2020), even though correlation among variables is ignored and can hinder its performance. Class imputation overcomes this limitation by using means or modes over all cases within the same class and it is quite popular (see, e.g., Ouimet et al., 2010; Parsa et al., 2019). Last, k-nearest neighbour replaces the missing values with the average of the corresponding variables of its k nearest neighbours where the cases were complete (see, e.g., Alrefaie et al., 2019; Qi and Guan, 2019; Jiang et al., 2020).

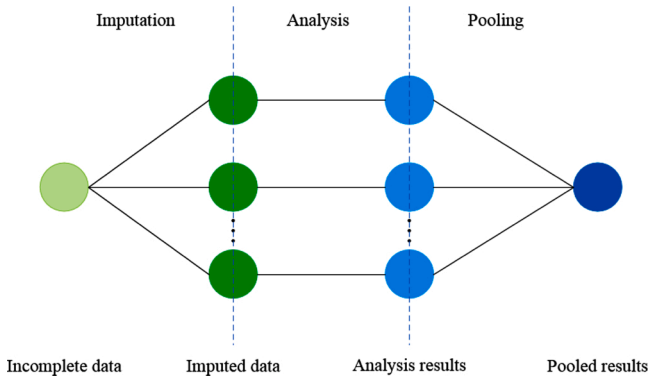

Fig. 1. Architecture of multiple imputation.

While the issue of missing values has been recognised and treated extensively in the traffic safety literature, it is obvious that the same methods are used in the last couple of decades and some of the studies have been published very recently. Accordingly, an answer has not been given to the need for exploration of advanced models for data imputation that has been suggested in the past to improve the performance of analytical methods (Mitra and Washington, 2012; Janstrup et al., 2016). Moreover, the methods for dealing with missing values are often paired with statistical analyses, thus ignoring the contribution that machine learning methods have made in classifying the most significant contributors to crash injury severity.

With the aim of answering the need for an advanced model for data imputation that allows to draw proper conclusions about the relevance of contributors to crash injury severity, this study proposes a sequential approach that combines multivariate imputation by chained equations (MICE) and random forests (RF) classifiers. The rationale for choosing MICE was to utilise a method that allows to impute different variable types with proper model specifications, thus answering the need for advanced models for data imputation. The rationale for choosing RF was threefold: (i) RF is a collection of tree classifiers that inherits the idea of ensemble in machine learning, and so RF can achieve high accuracy; (ii) the growing procedures of RF can offer an experimental method to rank contributors of traffic crashes; (iii) when sample size is limited, the out-of-bag strategy of RF can help obtain unbiased error estimates. In the sequential method, the implementation of MICE allows to impute missing values in the original dataset, while the implementation of RF allows to identify and rank the most significant contributors to crash injury severity. The sequential approach answers the need for fusion of statistical and machine learning methods to mine traffic crash datasets containing some missing values. This seems essential when considering that the way forward in crash safety analysis should attempt to use the advantages from both approaches rather than ascertain the dominance of one over the other. The effectiveness of the approach is demonstrated by studying crash injury severity on mountainous freeways in China while recognizing the limitations in the existing crash datasets. Also, the effectiveness of the approach is compared to the most common and the most performing methods for dealing with missing data.

The remainder of this paper is organized as follows. The following section introduces the methods (MICE and RF). Then, the case-study is presented and the results of the implementation to the original and the imputed datasets are illustrated. Last, a summary about findings is provided alongside further research directions.

2. Methods

The proposed approach consists of MICE and RF, where the former allows to deal with the incomplete records, and the latter allows to

provide the ranking of the independent variables. The framework of the method is shown in Fig. 1, where it can be seen that the approach contains three steps: imputation, analysis, and pooling. Their details are presented in the following subsections.

2.1. Imputation

The first step of the proposed approach is the imputation, where MICE is applied because of its wide implementation in many fields where it was requested to impute values missing at random in multiple variables within a crash dataset.

Consider the n -dimensional vector $\boldsymbol{x}$ combining the n_o -dimensional vector $\boldsymbol{x_o}$ of observed values and the n_m -dimensional vector $\boldsymbol{x_m}$ of missing values. For each variable $\boldsymbol{x}$ with missing values, in the first step the missing values are initialized with a simple imputation where they are replaced by the mean of the observed values. In the second step, consider one variable $\boldsymbol{y}$ and regress the observed values in vector $\boldsymbol{y_o}$ on the other variables in an imputation model. In the third step, the missing values in vector $\boldsymbol{y_m}$ are replaced with predictions from the estimated imputation model and, when $\boldsymbol{y}$ is subsequently used as an independent variable in the imputation model for another variable, both the observed and these imputed values will be used. In the fourth step, the previous two steps are repeated for each variable with missing data to complete a cycle, and multiple cycles are then performed with the imputations being updated at each one until convergence. At the end of these cycles, the final imputations are retained in an imputed dataset. In this study, the process is repeated for L times to remove the effect of uncertainty and variability of variables within the raw dataset, and hence L imputed datasets are created.

It should be noted that, as discussed in the remainder of this section, the imputation models can be defined for different types of variables. In fact, another difficulty of an incomplete crash dataset is how to handle different types of variables such as continuous, binary and categorical variables. MICE can deal with the different types because each variable can have its own imputation model. In the following subsections, we introduce different types of imputation models applied in this study: linear regression, logistic regression, and multinomial logistic regression are implemented to estimate missing values of continuous variables, binary variables, and categorical variables (more than two classes), respectively.

Table 1
Training of the RF classifier.

Input:
T: Training dataset,
F: Input features,
B: Number of trees in the forest.
Output: MDIA
function *RandomSample* (T, F)
  $rf \leftarrow \emptyset$
  for $i \in 1, \ldots B$ **do**
    $T^i \leftarrow$ A bootstrap sample from T
    $rt_i \leftarrow RandomTree\ (T^i, F)$
    $rf \leftarrow rf \cup \{rt_i\}$
  end for
  return rf
end function
function *RandomTree* (T, F)
  At each node
    $f \leftarrow$ very small subset of F
    Split on best feature in f
  return the trained tree
end function

2.1.1. Linear regression model

In the case of continuous variables, the imputation model is estimated as a linear regression model:

$$\boldsymbol{y_o}|\boldsymbol{x_o};\ \boldsymbol{y_o} \sim N(\boldsymbol{\beta x_o},\ \sigma^2) \tag{1}$$

where $\boldsymbol{\beta}$ is a j -dimensional vector of parameters estimated by regressing the observed values in vector $\boldsymbol{y_o}$ on the other variables in vectors $\boldsymbol{x_o}$ in the dataset, and σ is the residual error. Given this imputation model, the parameters $\boldsymbol{\beta}^*$ of the linear regression model and the residual error σ^* are estimated from the joint posterior distribution of σ and $\boldsymbol{\beta}$:

$$\sigma^* = \sigma\sqrt{(n_o - j)/g} \tag{2}$$

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} + \sigma^* / \sigma(\boldsymbol{u}_1\sqrt{\Sigma}) \tag{3}$$

where $\sqrt{\Sigma}$ is the Cholesky decomposition of the covariance matrix Σ of $\boldsymbol{\beta}$, g is a scalar that is randomly drawn from a χ^2 distribution with $(n_o - j)$ degrees of freedom, and $\boldsymbol{u}_1$ is a j -dimensional vector whose elements are randomly drawn from a standard normal distribution. Accordingly, the missing values are obtained as follows:

$$\boldsymbol{y_m} = \boldsymbol{\beta}^*\boldsymbol{x_m} + \ \boldsymbol{u}_2\sigma^* \tag{4}$$

where and $\boldsymbol{u}_2$ is a n_m -dimensional vector whose elements are randomly drawn from a standard normal distribution.

2.1.2. Logistic regression model

In the case of binary variables, the imputation model is estimated as a logistic regression:

$$logit\ p(\boldsymbol{y_o} = 1\boldsymbol{X_o};\boldsymbol{\beta}) = \ \boldsymbol{\beta X_o} \tag{5}$$

where $\boldsymbol{\beta}^*$ is obtained from the posterior distribution of $\boldsymbol{\beta}$ that is approximated by a multivariate normal distribution MVN$(q,\ \Sigma)$. Given the imputation model, each element y_m^i in the vector $\boldsymbol{y_m}$ can be estimated as follows:

$$y_m^i = \begin{cases} 1\ if\ u_i < p_i \\ 0\ otherwise \end{cases} \tag{6}$$

where the probability p_i for the i -th element is equal to $p_i = [1 + \exp(-\boldsymbol{\beta}\boldsymbol{x}_m^i)]^{-1}$, $\boldsymbol{x}_{mi}^i$ is the i -th vector of the matrix $\boldsymbol{X_m}$ of variables related to y_{mi}^i, and u_i is randomly drawn from the uniform distribution $U(0,\ 1)$.

2.1.3. Multinomial logistic regression model

In the case that a variable has more than two categories, the imputation model is a multinomial logistic regression where each category is compared to the reference one by logistic regression:

$$P(\boldsymbol{y_o} = \boldsymbol{c}\boldsymbol{X_o};\boldsymbol{\beta}) = \left[\sum_{c_i}^{c_n} \exp(\boldsymbol{\beta}_{c_i}\boldsymbol{X_o})\right]^{-1} \exp(\boldsymbol{\beta}_{c_i}\boldsymbol{X_o}) \tag{7}$$

where $\boldsymbol{c}$ is a vector of elements representing categories c_i for the i -th observed value y_o^i of vector $\boldsymbol{y_o}$, c_n is the number of categories, $\boldsymbol{\beta}_{c_i}$ are n_o -dimensional vectors containing parameters associated with each category c_i, and $\boldsymbol{\beta}_{c_1} = 0$ as the first category is the reference one. Consider $p_{i,c_i} = p(y_o^i = c_i\boldsymbol{X_o};\ \boldsymbol{\beta}^*)$, where $\boldsymbol{\beta}^*$ is a vector of length $n_o \times (n_c - 1)$ that is randomly drawn from a normal approximation to the posterior distribution of $\boldsymbol{\beta} = (\boldsymbol{\beta}_{c_2}, \ldots, \boldsymbol{\beta}_{c_n})$. The i -th missing value is imputed as follows:

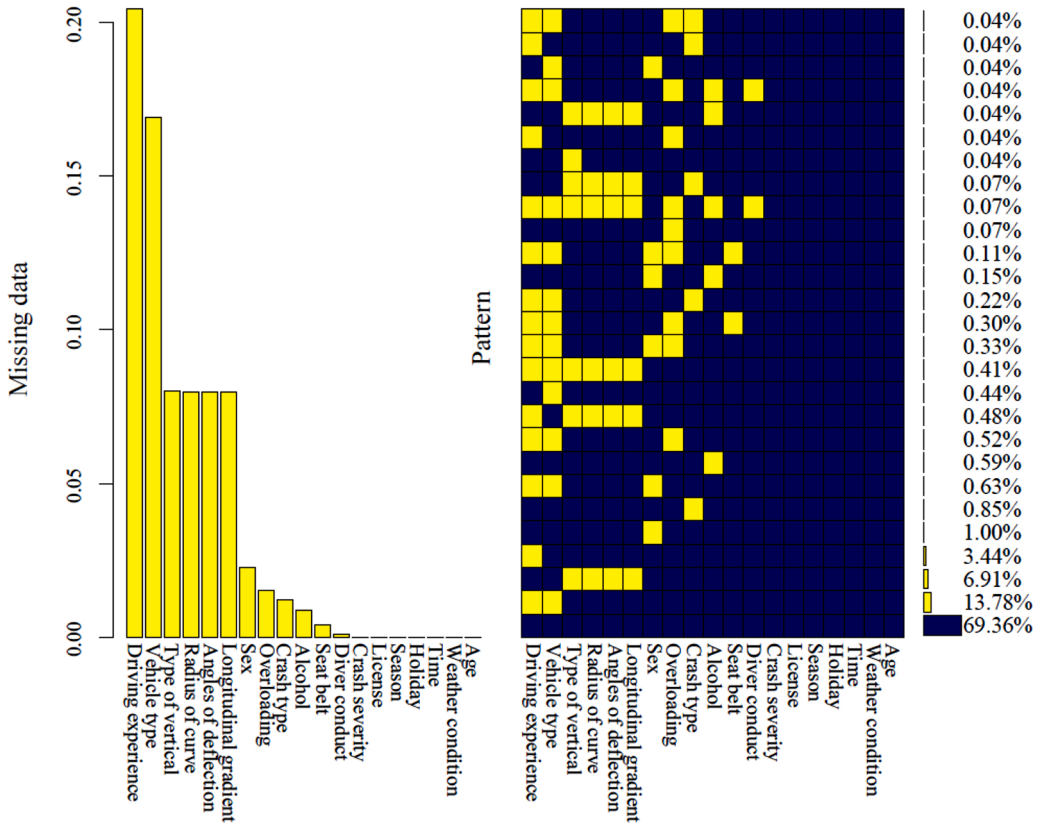

Fig. 2. Missing patterns of the raw dataset.

$$y_m^i = 1 + \sum_{c_i=1}^{c_n-1} I(u_i > c_{i,c_i}) \tag{8}$$

$$I(u_i > c_{i,c_i}) = \begin{cases} 1 \ if \ u_i > c_{i,c_i} \\ 0 \ otherwise \end{cases} \tag{9}$$

where $c_{i,c_i} = \sum_{c_i=1}^{c_n} p_{i,c_i}$ and u_i is randomly drawn from the uniform distribution $U(0,\ 1)$.

2.2. Analysis

The imputation produced L complete datasets that are analysed via the RF classifier. RF is a popular and promising model that is commonly used to solve the classification problem and rank the importance of variables. This model combines the bagging idea and the random subspace method to develop a collection of classification trees with control variations. In the model, some trees are grown by bootstrapping samples and searching a selected subset randomly of inputs at each split. Then, these trees are integrated in an appropriate way. The process to train RF to classify the crash injury severity is summarized in Table 1.

With the aim of ranking contributors to crash injury severity, the variable importance should be calculated. Let T and T^i denote respectively all the complete samples and the i -th bootstrap samples. $(T - T^i)$ represents out-of-bag samples containing all observations in T but not selected into T^i. After a tree rt_i is constructed by using T^i, out-of-bag samples $(T - T^i)$ are sent down to rt_i and a misclassification error e_b can be calculated. Then, the j -th variable x^j of out-of-bag samples is randomly permuted and rt_i is run again with misclassification error e_b^i being also recorded. This process is done for every tree and repeated for B times. Finally, the importance of the variable is represented by the mean of decrease in accuracy ($MDIA$) that can be calculated by averaging relative difference between e_b^i and e_b.

$$MDIA(x^j) = \frac{1}{B}\sum_{i=1}^{B}(e_b - e_b^i) \tag{10}$$

2.3. Pooling

Given the L complete datasets, different results are obtained from the implementation of the RF classifier. This last step of the implementation of the proposed method pools all the estimation according to the Rubin's rules (Rubin, 2004). Consider $MDIA_l$ as the estimated variable importance from the RF implementation for the l -th imputed dataset, it can then be calculated the average MDIA (minimum discriminant information adjustment) as follows (Rubin, 2004):

$$MDIA = \frac{1}{L}\sum_{l=1}^{L} MDIA_l \tag{11}$$

Moreover, a variance can also be computed to consider the variation across datasets (Rubin, 2004):

$$\theta = \frac{1}{L}\sum_{l=1}^{L}(MDIA_l - MDIA)^2 \tag{12}$$

2.4. Evaluation criteria

According to common practice in machine learning, a confusion matrix is used to visualize the performance of the classification algorithm (Washington et al., 2011). In this study, the confusion matrix is applied to evaluate the RF for classifying the crash injury severity. After obtaining the confusion matrix, the accuracy of the classification is calculated as the proportion of the total number of crashes that are correctly predicted, similar to recent studies focusing on crash safety analysis (see, e.g., Zhao et al., 2010; Olutayo and Eludire, 2014; Chen

Table 2
Description of categorical variables.

Variable	Type	Categories	Percentage in the raw dataset	Percentage in the fully observed dataset
Crash severity	Categorical	Property	61.35 %	71.71 %
		Injury	27.79 %	21.68 %
		Fatal	10.86 %	6.61 %
Crash type	Categorical	Head-on	28.88 %	28.72 %
		Rear-end	62.48 %	64.84 %
		Side-wipe	6.40 %	4.79 %
		Angle	1.57 %	0.85 %
		Other	0.67 %	0.80 %
Sex	Binary	Male	92.21 %	97.12 %
		Female	7.79 %	2.88 %
License	Binary	Invalid	1.37 %	1.44 %
		Valid	98.63 %	98.56 %
Alcohol	Binary	No	99.74 %	99.79 %
		Yes	0.26 %	0.21 %
Seat belt	Binary	No	0.96 %	0.80 %
		Yes	99.04 %	99.20 %
Vehicle type	Binary	Car	76.65 %	75.65 %
		Truck	23.35 %	24.35 %
Overloading	Binary	No	99.21 %	99.25 %
		Yes	0.79 %	0.75 %
Driver conduct	Categorical	Brake failure	31.56 %	29.94 %
		Speeding	22.68 %	21.74 %
		Risky following distance	43.80 %	47.04 %
		Other	1.96 %	1.28 %
Type of vertical	Categorical	Straight line	44.92 %	43.63 %
		Concave curve	33.59 %	35.38 %
		Convex curve	21.49 %	20.99 %
Season	Categorical	Spring	35.37 %	36.92 %
		Summer	21.25 %	21.36 %
		Autumn	24.06 %	22.54 %
		Winter	19.33 %	19.18 %
Holiday	Binary	No	69.07 %	69.74 %
		Yes	30.93 %	30.26 %
Time	Binary	06:00–18:00	77.79 %	79.22 %
		18:00–06:00	22.21 %	20.78 %
Weather condition	Categorical	Sunny	38.25 %	37.29 %
		Cloudy	38.51 %	38.41 %
		Rainy, snowy or foggy	23.24 %	24.29 %

et al., 2015; Iranitalab and Khattak, 2017; Amiri et al., 2020).

3. Case-study

3.1. Data

This study applied the proposed method to a section of 168 km of the Chang-Jing freeway and a section of 128 km of the Tai-Gan freeway located in Jiangxi Province. The selected sections are typical mountainous freeways with design speed equal to 100 km/h. The crash data were extracted from the Traffic Accident Database System published by Jiangxi Provincial Public Security Ministry and maintained by the Jiangxi Transport Policy Bureau. The information related to each road crash was recorded and reported via on-scene assessments by traffic policemen and sent to the traffic management department within 24 h. Previous studies suggested that the geometry of freeways such as vertical type, curve radius and section grade were important factors affecting crash severity, especially in mountainous regions (Huang et al., 2018). Therefore, we also extracted detailed information about the freeway geometry for the selected sections from the freeway management bureau of the Jiangxi Province. Geographic information system (GIS) analysis was performed to identify and match the information from the geometry to the corresponding samples, and the fusion of GIS and crash data allowed to identify a total of 2706 crashes over the period from January 2008 to February 2013.

After the data fusion, the crash dataset contained several variables including crash injury severity, crash type, driver demographic characteristics, crash reasons, vehicle types, time of the crash, environmental factors such as weather conditions, and freeway geometry such as curve radius. Consistently with the literature, the crash injury severity was defined as the most severe outcome for a road user involved in the crash. The reports categorized three levels of injury: property damage only (61.4 % of the crashes), injury (25.9 %) where at least one road user required hospitalisation, and fatality (12.7 %) where at least one road user died within 30 days of the crash. The geometry presented variation, as the horizontal curve radius ranges from 0.7 to 6 km, the curve angle ranges from 10 to 67 degrees, and the longitudinal gradient ranges from -3.8 % to 3.9 % (where a negative value indicates a downhill segment and a positive value indicates an uphill segment). It should be noted that the dataset with all the 2706 crashes includes missing values for different variables and it is named the raw dataset (the missingness patterns are described in Fig. 2). The dataset with the 1877 crashes without any missing value in any variables is named the fully observed dataset.

Tables 2 and 3 show respectively the description of the categorical and continuous variables in the raw and the fully observed dataset. Interestingly, crash severity, driving license, holiday, time of day and weather conditions do not present have missing values and hence do not require imputation. Also, it should be noted that crash injury severity is fully observed, which implies that only independent variables require the implementation of the proposed approach.

4. Results and discussion

The MICE-RF approach was applied to study the relationship between risk factors and crash injury severity in mountainous freeways while using the raw dataset. This dataset contained continuous, binary and categorical variables that required imputation by linear, logistic and multinomial logistic regression to obtain an imputed dataset. Then, each imputed dataset was analysed via an RF classifier to obtain an explicit understanding of factors contributing to crash injury severity. All the calculations were performed in the statistical package R 3.4.3.

4.1. Analysis of the imputation

Given the existence of missing values in the raw dataset, MICE was

Table 3
Description of continuous variables.

Variable	Raw dataset			Fully observed dataset	
	Mean	Standard deviation	Missing ratio	Mean	Standard deviation
Age	36.23	9.15	0.44%	36.28	9.31
Driving experience	8.38	6.10	0.52%	8.42	6.15
Radius of curve	1390.38	1536.56	8.02%	1433.21	1567.21
Angles of deflection	22.95	21.41	8.02%	23.27	20.60
Longitudinal gradient (%)	0.48	2.04	8.02%	0.41	2.21

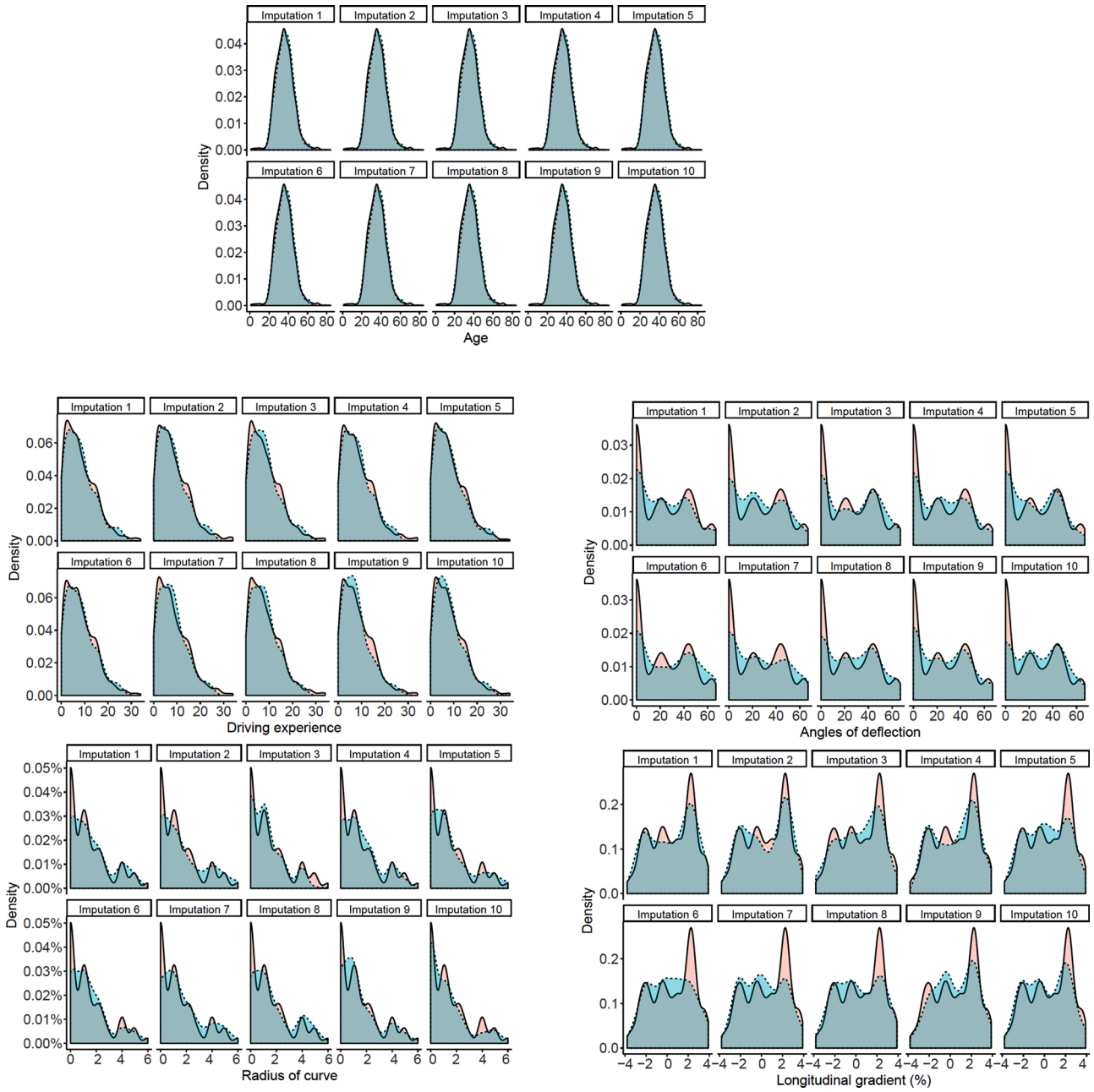

Fig. 3. Comparison of the imputed (TRUE) and observed (FALSE) values for continuous variables.
a)Confusion matrix for the fully observed dataset
b) Confusion matrix for the imputed datasets

utilised to impute missing values. In the imputation, the number of cycles h was set to 10 and the number of imputed datasets L was also set to 10 (i.e., 10 imputed datasets were created). Fig. 3 presents the densities of both imputed (dotted lines) and observed (continuous lines) values of all continuous variables in order to evaluate the effect of the proposed imputation method. As 10 imputed datasets were obtained, 10 subfigures were drawn for each continuous variable.

Table 4 presents the average frequency of each category of the categorical variables over the 10 imputed datasets and compares the fully observed dataset with the imputed one. It should be noted that the values in the imputed dataset are the same as the values in the raw dataset for the variables where imputation was not needed. As suggested in Buuren and Groothuis-Oudshoorn (2010), a big difference in densities or frequencies between imputed and observed values may indicate a problem. It can be seen that densities and frequencies between our imputed value and observed value are similar, suggesting that the imputation method performs efficiently.

4.2. Accuracy of the classification

The RF classifier was implemented on the 10 imputed datasets as well as the fully observed one (i.e., the one where the missing values were removed) in order to evaluate the performance of the imputation model and examine the effect of the missingness. Then, a ten-fold cross-validation technique was applied to evaluate the accuracy rather than dividing the dataset into training and testing datasets. In the cross-validation, the dataset was randomly divided into 10 blocks with equal size and each time nine blocks were used to train the RF and the one remaining block was used to test the RF. This process was repeated ten times by using different blocks for testing purposes, and then the

Table 4
Frequency of the imputed and observed values for categorical variables.

Variable	Categories	Percentage in the fully observed dataset	Percentage in the imputed dataset
Crash injury severity	Property	71.71 %	61.35 %
	Injury	21.68 %	27.79 %
	Fatal	6.61 %	10.86 %
Crash type	Head-on	28.72 %	29.06 %*
	Rear-end	64.84 %	57.15 %*
	Side-wipe	4.79 %	10.10 %*
	Angle	0.85 %	3.32 %*
	Other	0.80 %	0.37 %*
Sex	Male	97.12 %	80.68 %*
	Female	2.88 %	19.32 %*
License	Invalid	1.44 %	1.37 %
	Valid	98.56 %	98.63 %
Alcohol	No	99.79 %	99.61 %*
	Yes	0.21 %	0.39 %*
Seat belt	No	0.80 %	1.40 %*
	Yes	99.20 %	98.60 %*
Vehicle type	Car	75.65 %	63.70 %*
	Truck	24.35 %	36.30 %*
Overloading	No	99.25 %	99.10 %*
	Yes	0.75 %	0.90 %*
Driver conduct	Brake failure	29.94 %	35.28 %*
	Speeding	21.74 %	24.74 %*
	Risky following distance	47.04 %	36.41 %*
	Other	1.28 %	3.57 %*
Type of vertical	Straight line	43.63 %	47.41 %*
	Concave curve	35.38 %	30.81 %*
	Convex curve	20.99 %	21.79 %*
Season	Spring	36.92 %	35.37 %
	Summer	21.36 %	21.25 %
	Autumn	22.54 %	24.06 %
	Winter	19.18 %	19.33 %
Holiday	No	69.74 %	69.07 %
	Yes	30.26 %	30.93 %
Time	06:00–18:00	79.22 %	77.79 %
	18:00–06:00	20.78 %	22.21 %
Weather condition	Sunny	37.29 %	38.25 %
	Cloudy	38.41 %	38.51 %
	Rainy, snowy or foggy	24.29 %	23.24 %

* Variables contain missing values.

results were averaged.

The RF algorithm was implemented to classify the crash injury severity by growing 100 trees for all datasets as suggested in previous studies (Harb et al., 2009; Abdel-Aty and Haleem, 2011). Also, the algorithm set the number of variables randomly selected as candidates at each split as the square root of the number of explanatory variables $\sqrt{18} \approx 4$ as suggested in previous studies (Iranitalab and Khattak, 2017). Once the algorithm was implemented, 10 accuracies were obtained for each dataset by using the ten-fold cross-validation. Fig. 4a shows the confusion matrix for the RF algorithm trained with the fully observed dataset, while Fig. 4b shows the same matrix for the average over the confusion matrices from the RF algorithm trained with the 10 imputed datasets. The predicted accuracy for each class is presented in the secondary diagonal of the figure, and the comparison shows that the accuracy from the imputed datasets for the three injury severity outcomes (0.97 for property damage, 0.76 for injury, 0.66 for fatality) has improved on the fully observed dataset (0.96, 0.53, 0.50). Accordingly, the proposed imputation method was preferable than the deletion of cases with missing values that is the most common practice in crash analysis. In relative terms, the imputation method has improved the classification accuracy of injury crashes by 43.40 % and the one of fatal crashes by 32.00 %, a very important result given the major focus of countermeasures being on the most severe crashes.

Fig. 5 presents the classification accuracy of the RF algorithm trained by the 10 imputed datasets as well as the fully observed dataset, where the accuracy of each of the 10 imputed datasets (red line for the average) is higher than the one from the fully observed dataset (blue line). On average, there is an improvement of overall accuracy by about 6.02 %.

Last, a sensitivity analysis was performed to verify whether the order of presentation of the variables for imputation had an effect on the imputation results, since the method requires for the variables with missing values to be imputed one by one, where the imputed variables are used in the subsequent imputation of the next variables. Randomisation of the order of the variables revealed that the accuracy of the imputation was not affected by the order of the variables being considered for imputation.

4.3. Comparison of the accuracy of the classification

Given the need for advanced methods for imputation, a comparison with the most common and the most performing methods used currently for imputation was performed. As presented in the introduction, mean imputation, class imputation and k-nearest neighbours are the most common methods for the task at hand, while decision tree and sampling based missing value imputation (DSMI, Deb and Liew, 2016) is the most performing method for the same task. Accordingly, they were applied to the dataset via an experiment where missing values were randomly extracted from the fully observed dataset in order to allow the comparison between the imputed values and the actual ones (a comparison that the actual missing values would have not allowed to be performed). Four levels of missingness were considered (5, 10, 15, 20 %) and the comparison between the four methods was performed by measuring the mean absolute percentage error (MAPE, namely the average error between imputed and observed cases) and the correct imputation rate (CIR, namely the ratio between correctly imputed values and total number of values).

Mean imputation was implemented by replacing the missing values with the means of the continuous variables and the modes of the categorical ones over the entire dataset. Class imputation was applied by using the means and modes over all the cases within the same class as replacement. K-nearest neighbours was performed by replacing the missing values with the average of the corresponding variables of its k nearest negihbours with complete cases for k equal to 10. DSMI was utilised with the first level similarity weight equal to 0.6 and the second level similarity weight equal to 0.4 (Deb and Liew, 2016), after modifying the continuous variables in Table 3 into categorical variables.

Table 5 presents the measures for the five applied methods and shows that the MICE outperforms significantly the three most common methods even when varying the degree of missingness. It also shows that MICE outperforms slightly DSMI, but it should also be noted that the proposed method does not require the manipulation of variables as its flexibility allows for treating the missingness of both categorical and continuous variables.

4.4. Importance of the contributing factors

After the implementation of the RF classifier, the importance of the various variables in classifying the crash injury severity was ranked. Fig. 6 summarizes the MDIA and its standard error for the variables that required imputation. It can be observed that the standard errors of all variables are low, which implies that the MDIA of the 10 imputed datasets do not present much variation. In turn, this indicates that the proposed imputation method is robust. It can also be observed that the importance rank of the variables is as follows: vehicle type, crash type, longitudinal gradient, driver conduct, curve radius, angle of deflection, season, weather condition, type of vertical, holiday, time, sex, age, driving experience, license, alcohol, seat belt, and overloading.

From the driver perspective, driver conduct ranked fourth in MDIA among all variables. Driver conduct is an explicit factor that can reflect how well a driver knows the rules of mountainous freeways, which may be different from other roads where then drivers might misjudge

a) Confusion matrix for the fully observed dataset

b) Confusion matrix for the imputed datasets

Fig. 4. Comparison of confusion matrices.

situations related to the severity of the crashes (Öz et al., 2010; Scott--Parker et al., 2015). The characteristics of the driver including sex (MDIA = 1.68 %), age (MDIA = 1.15 %) and driving experience (MDIA = 0.59 %) are all ranked rather low, a finding different from previous studies. A possible explanation is that the drivers in the sample are more experienced than the one in other crash datasets, as the average driving experience is over 8 years, and it is in line with the fact that new drivers in China rarely drive on freeways in mountainous area.

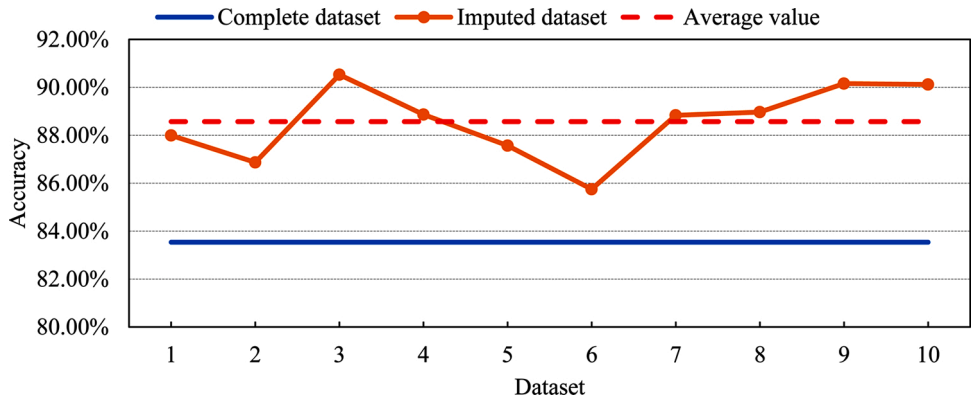

Fig. 5. Comparison of overall classification accuracy.

Table 5
Comparison of the performance of the proposed method and the most frequently used and most performing methods for imputation.

Missing rate	MAE (%)					CIR (%)				
	AI	CI	KNN	DSMI	MICE	AI	CI	KNN	DSMI	MICE
5%	32.24	36.98	37.24	29.85	27.33	69.15	71.28	65.96	84.34	87.21
10%	36.39	41.07	40.11	28.59	28.23	68.09	66.49	65.43	86.47	87.75
15%	40.61	42.86	40.35	30.54	29.41	68.44	67.02	65.96	85.37	86.86
20 %	40.03	40.15	44.25	32.86	31.10	64.53	65.87	66.93	85.12	85.35

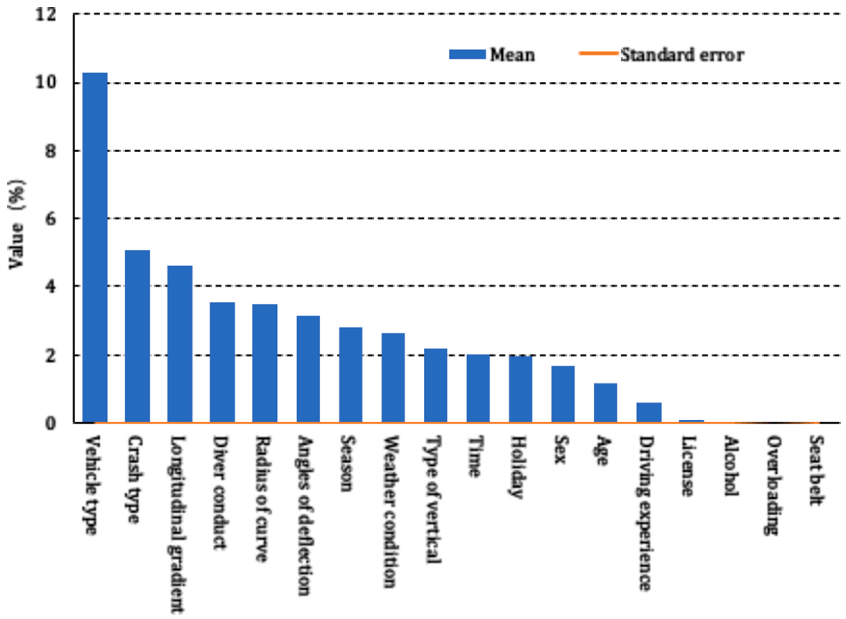

Fig. 6. Variable importance and its standard error of contributed factor.

The states of the driver including license (MDIA = 0.11 %), alcohol (MDIA = 0.04 %), and seat belt use (MDIA = 0.02 %) are ranked even lower and are not significantly related to crash injury severity. Unlike findings in the urban context, the relation between crash injury severity and a driver not having a valid license in particular suggests that licensing is not a major issue on mountainous freeways. Another interesting finding is the less importance of alcohol and seat belt contributing to more severe crashes, possibly because there are professional drivers that are likely not to drink and not to fail to use seat belts on mountainous freeways. It can be in fact seen that the MDIA of these two aberrant driving behaviours in our raw dataset are 0.21 % and 0.80 %.

From the vehicle perspective, the MDIA of vehicle type is the highest (>10.00 %). Compared to other types of vehicles, trucks display higher risk of being involved in severe crashes, as they have unique operating characteristics such as width and high gross weight that can result in poor acceleration/deceleration performance (Chang and Chien, 2013). Given the disparity of speed and the complex traffic condition, crashes including trucks in mountainous regions have an increasing likelihood to be severe, in line with findings from previous studies (Dong et al., 2015; Iranitalab and Khattak, 2017). Overloading is ranked penultimate among all variables, most likely because of strict laws enforced by the Chinese government, especially in mountainous areas, that makes the percentage of crashes involving overloading low in the first place.

From the road perspective, it can be observed that its geometrical characteristics are also important factors in relation to crash injury severity. The radius of curve, angle of deflection and longitudinal

gradient have MDIA between 3.00 and 5.00 %. As found in previous studies, these factors contribute positively to the occurrence of crashes (Li et al., 1994; Fu et al., 2011)and, on mountainous freeways, also to the crash injury severity (Ma et al., 2015; Huang et al., 2018). Due to the complex topographic conditions in mountainous areas, some adverse curves and grades cannot be avoided, and this is a major safety issue for Chinese mountainous freeways. This finding suggests that traffic management measures and safety facilities should be built on these freeways to limit the outcome of crashes in the case that they would occur.

Interestingly, none of the environmental characteristics appears at the top of the ranking of the importance of variables. It can be seen that the MDIA of season, holiday, time of day, and weather conditions are between 1.90 % and 3.00 %, therefore less relevant than most road characteristics, but more relevant than most driver characteristics.

5. Conclusions

This study presented a sequential MICE-RF method to impute missing values in a dataset of crashes occurring on mountainous freeways in China and then analyse the factors contributing to crash injury severity. The implementation of the method showed a very good classification accuracy leading then to the ranking of contributors to crash injury severity on mountainous freeways. Also, the classification accuracy was an improvement over methods currently and efficiently used for data imputation in crash safety analysis, thus answering the need for advanced methods as noted in the literature for almost a decade (Mitra and Washington, 2012; Janstrup et al., 2016).

The first contribution of this study is the proposal of the sequential MICE-RF method. It is flexible, as it allows for different imputation methods to be used for continuous, binary and categorical variables, and it is promising, as the case-study showed an improvement in classification accuracy. Multiple imputation appeared recently to be effective for count models (Afghari et al., 2019), but for this crash injury severity study the proposed method is more flexible not only for the imputation method but also for the analysis of all the variables with missing values. It should be also noted that the proposed method mixes machine learning and statistical modelling to exploit the strengths of both types of methods. Interestingly, the improvement of classification accuracy with respect to widely use methods shows that this could be a promising method to deal with the often encountered, and generally neglected, issue of missing data.

The second contribution of this study is the analysis of contributors to the injury severity of crashes on mountainous freeways in China. Notably, the method allowed to use the entire information collected in the dataset rather than removing the cases with missing data. Firstly, findings confirmed the fact that trucks are a relevant issue when it comes to crash injury severity, and traffic management measures should be implemented to mitigate the heavy traffic with the aim of reducing its consequences in the case that a crash occurred. Secondly, findings suggested that geometric characteristics of mountainous freeways play a role in increasing the probability of severe outcomes for crashes. In particular, the design of curves should avoid sharp traits that are related to more severe injuries and, if the topography presents restrictions, the limitation of speed should be adapted to different values for the curve radius. Lastly, findings suggested a relation between driver conduct and crash injury severity, possibly for a mismatch between the understanding of the context and the mountainous freeways themselves. Future mobility solutions, in particularly connection between vehicles, could help recognize possible hazards earlier and help drivers adapt their behaviour.

While the study has contributions, limitations should be acknowledged. The accuracy of the imputation was evaluated only for this case-study, although the dataset is quite generic in that it is extracted from a crash database and matched with GIS information as it could be available in several other countries. Possibly, further research could consider generating missing values in a complete dataset to compare the imputed values with the real ones and hence provide a more comprehensive evaluation of the imputation model. However, the promise from this hybrid machine learning and statistical method is certainly the most relevant finding.

Author statement

The authors confirm contribution to the paper as follows: study conception and design: Linchao Li; data collection: Yonggang Wang; analysis and interpretation of results: Linchao Li and Carlo Prato; draft manuscript preparation: Linchao Li and Carlo Prato. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The first author expresses their gratitude to the School of Transportation at Southeast University for supporting this research. All the authors are grateful to two anonymous reviewers for the insightful comments that helped improve significantly an earlier version of the manuscript.

References

Abdel-Aty, M.A., 2001. Using ordered probit modeling to study the effect of ATIS on transit ridership. Transp. Res. Part C Emerg. Technol. 9 (4), 265–277.
Abdel-Aty, M., Haleem, K., 2011. Analyzing angle crashes at unsignalized intersections using machine learning techniques. Accid. Anal. Prev. 43 (1), 461–470.
Afghari, A.P., Washington, S., Prato, C., Haque, M.M., 2019. Contrasting case-wise deletion with multiple imputation and latent variable approaches to dealing with missing observations in count regression models. Anal. Methods Accid. Res., 100104
Ahmed, M., Huang, H., Abdel-Aty, M., Guevara, B., 2011. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. Accid. Anal. Prev. 43 (4), 1581–1589.
Al-Madani, H.M.N., 2018. Global road fatality trends' estimations based on country-wise micro level data. Accid. Anal. Prev. 111, 297–310.
Alrefaie, M.T., Summerskill, S., Jackon, T.W., 2019. In a heart beat: using driver's physiological changes to determine the quality of a takeover in highly automated vehicles. Accid. Anal. Prev. 131, 180–190.
Amiri, A.M., Sadri, A., Nadimi, N., Shams, M., 2020. A comparison between artificial neural network and hybrid intelligent genetic algorithm in predicting the severity of fixed object crashes among elderly drivers. Accid. Anal. Prev. 138, 105468.
Bai, Y., Sun, Z., Zeng, B., Deng, J., Li, C., 2017. A multi-pattern deep fusion model for short-term bus passenger flow forecasting. Appl. Soft Comput. 58, 669–680.
Bhat, C.R., 1994. Imputing a continuous income variable from grouped and missing income observations. Econ. Lett. 46 (4), 311–319.
Buuren, Svan, Groothuis-Oudshoorn, K., 2010. Mice : Multivariate Imputation by Chained Equations in R. J. Stat. Softw, pp. 1–68.
Çelik, A.K., Oktay, E., 2014. A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey. Accid. Anal. Prev. 72, 66–77.
Chang, L.Y., Chien, J.T., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Saf. Sci. 51 (1), 17–22.
Chen, C., Wang, Y., Li, L., Hu, J., Zhang, Z., 2012. The retrieval of intra-day trend and its influence on traffic prediction. Transp. Res. Part C Emerg. Technol. 22, 103–118.
Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., Guan, H., 2015. A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. Accid. Anal. Prev. 80, 76–88.
Chen, C., Zhang, G., Huang, H., Wang, J., Tarefder, R.A., 2016. Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. Accid. Anal. Prev. 96, 79–87.
Deb, R., Liew, A.W.C., 2016. Missing value imputation for the analysis of incomplete traffic accident data. Inf. Sci. (Ny) 339, 274–289.
Deka, L., Quddus, M., 2014. Network-level accident-mapping: distance based pattern matching using artificial neural network. Accid. Anal. Prev. 65, 105–113.
Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38 (3), 434–444.
Ding, C., Rizzi, M., Strandroth, J., Sander, U., Lubbe, N., 2019. Motorcyclist injury risk as a function of real-life crash speed and other contributing factors. Accid. Anal. Prev. 123, 374–386.

Dissanayake, S., Lu, J.J., 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes. Accid. Anal. Prev. 34 (5), 609–618.
Dong, C., Richards, S.H., Huang, B., Jiang, X., 2015. Identifying the factors contributing to the severity of truck-involved crashes. Int. J. Inj. Contr. Saf. Promot. 22 (2), 116–126.
Fu, R., Guo, Y., Yuan, W., Feng, H., Ma, Y., 2011. The correlation between gradients of descending roads and accident rates. Saf. Sci. 49 (3), 416–423.
Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests. Accid. Anal. Prev. 41 (1), 98–107.
Huang, H., Peng, Y., Wang, J., Luo, Q., Li, X., 2018. Interactive risk analysis on crash injury severity at a mountainous freeway with tunnel groups in China. Accid. Anal. Prev. 111, 56–62.
Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108, 27–36.
Janstrup, K.H., Kaplan, S., Hels, T., Lauritsen, J., Prato, C.G., 2016. Understanding traffic crash under-reporting: linking police and medical records to individual and crash characteristics. Traffic Inj. Prev. 17 (6), 580–584.
Jiang, F., Yuen, K.K.R., Lee, E.W.M., 2020. A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. Accid. Anal. Prev. 141, 105520.
Kaplan, S., Prato, C.G., 2012a. Associating crash avoidance maneuvers with driver attributes and accident characteristics: a mixed logit model approach. Traffic Inj. Prev. 13 (3), 315–326.
Kaplan, S., Prato, C.G., 2012b. Risk factors associated with bus accident severity in the United States: a generalized ordered logit model. J. Safety Res. 43 (3), 171–180.
Li, J., Abdelwahab, W., Brown, G., 1994. Joint effects of access and geometry on two-lane rural highway safety in British Columbia. Am. J. Civ. Eng. Archit. 21 (6), 1012–1024.
Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486.
Li, L., Zhang, J., Wang, Y., Ran, B., 2017. Multiple imputation for incomplete traffic accident data using chained equations. 2017 IEEE 20th Conference on Intelligent Transportation Systems (ITSC) 1–5.
Lin, L., Wang, Q., Sadek, A.W., 2015. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transp. Res. Part C Emerg. Technol. 55, 444–459.
Linchao, L., Fratrović, T., 2016. Analysis of factors influencing the vehicle damage level in fatal truck-related accidents and differences in rural and urban areas. Promet - Traffic - Traffico 28 (4), 331–340.
Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accid. Anal. Prev. 38 (6), 751–766.
Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Saf. Sci. 46 (5), 751–770.
Ma, X., Chen, F., Chen, S., 2015. Empirical analysis of crash injury severity on mountainous and nonmountainous interstate highways. Traffic Inj. Prev. 16 (7), 715–723.
Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accid. Anal. Prev. 46, 439–448.
National Bureau of Statistics of China, 2017. China Road Traffic Accident Statistics 2016. Wuxi, China..
Olutayo, V.A., Eludire, A.A., 2014. Traffic accident analysis using decision trees and neural networks. Int. J. Inf. Tech. Comp. Sci. 2, 22–28.
Ouimet, M.C., Simons-Morton, B.G., Zador, P.L., Lerner, N.D., Freedman, M., Duncan, G. D., Wang, J., 2010. Using the US National Household Travel Survey to estimate the impact of passenger characteristics on young drivers' relative risk of fatal crash involvement. Accid. Anal. Prev. 42 (2), 689–694.
Öz, B., Özkan, T., Lajunen, T., 2010. Professional and non-professional drivers' stress reactions and risky driving. Transp. Res. Part F Traffic Psychol. Behav. 13 (1), 32–40.
Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A.K., 2019. Real-time accident detection: coping with imbalanced data. Accid. Anal. Prev. 129, 202–210.
Peel, T., Ahmed, M., Ohara, N., 2017. Investigating safety effectiveness of Wyoming snow fence installations along a rural mountainous freeway. Transp. Res. Rec. 2613, 8–15.
Qi, G., Guan, W., 2019. Quantitatively mining and distinguishing situational discomfort grading patterns of drivers from car-following data. Accid. Anal. Prev. 123, 282–290.
Quddus, M.A., Noland, R.B., Chin, H.C., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. J. Safety Res. 33 (4), 445–462.
Rubin, D.B., 2004. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Hoboken, NJ.
Scott-Parker, B., Goode, N., Salmon, P., 2015. The driver, the road, the rules… and the rest? A systems-based approach to young driver road safety. Accid. Anal. Prev. 74, 297–305.
Shen, Y., Hermans, E., Bao, Q., Brijs, T., Wets, G., 2020. Towards better road safety management: lessons learned from inter-national benchmarking. Accid. Anal. Prev. 138, 105484.
Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C Emerg. Technol. 58, 380–394.
Wang, Y., Prato, C.G., 2019b. Determinants of injury severity for truck crashes on mountain expressways in China: a case-study with a partial proportional odds model. Saf. Sci. 117, 100–107.
Wang, Y., Li, L., Prato, C.G., 2019a. The relation between working conditions, aberrant driving behaviour and crash propensity among taxi drivers in China. Accid. Anal. Prev. 126, 17–24.
Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2011. Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall/CRC, Boca Raton, FL.
Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Anal. Methods Accid. Res. 1, 72–85.
Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.
Yu, R., Abdel-Aty, M., 2014. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. Accid. Anal. Prev. 62, 161–167.
Zhao, Z., Jin, X., Cao, Y., Wang, J., 2010. Data mining application on crash simulation data of occupant restraint system. Expert Syst. Appl. 37 (8), 5788–5794.