Accident Analysis and Prevention 111 (2018) 354–363

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network

Yunjie Li[a,b], Dongfang Ma[c,b], Mengtao Zhu[a], Ziqiang Zeng[d,b], Yinhai Wang[b,*]

[a] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, PR China
[b] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195, USA
[c] Institute of Marine information science and technology, Zhejiang University, Zhoushan 316021, PR China
[d] Uncertainty Decision-Making Laboratory, Sichuan University, Chengdu 610064, PR China

ARTICLE INFO

Keywords:
Significant factor
Highway crash
Genetic algorithm
Neural network
Traffic safety

ABSTRACT

Identification of the significant factors of traffic crashes has been a primary concern of the transportation safety research community for many years. A fatal-injury crash is a comprehensive result influenced by multiple variables involved at the moment of the crash scenario, the main idea of this paper is to explore the process of significant factors identification from a multi-objective optimization (MOP) standpoint. It proposes a data-driven model which combines the Non-dominated Sorting Genetic Algorithm (NSGA-II) with the Neural Network (NN) architecture to efficiently search for optimal solutions. This paper also defines the index of Factor Significance (F_s) for quantitative evaluation of the significance of each factor. Based on a set of three year data of crash records collected from three main interstate highways in the Washington State, the proposed method reveals that the top five significant factors for a better Fatal-injury crash identification are 1) Driver Conduct, 2) Vehicle Action, 3) Roadway Surface Condition, 4) Driver Restraint and 5) Driver Age. The most sensitive factors from a spatiotemporal perspective are the Hour of Day, Most Severe Sobriety, and Roadway Characteristics. The method and results in this paper provide new insights into the injury pattern of highway crashes and may be used to improve the understanding of, prevention of, and other enforcement efforts related to injury crashes in the future.

1. Introduction

According to statistics from the National Highway Traffic Safety Administration (NHTSA, 2017), which stores road data for the 50 US states, in 2015 alone, nationwide, there were a total of 22,441 deaths and 2.18 million injuries due to automobile accidents. Since fatal highway crashes are a major cause of injury, death, and economic loss, the identification of the significant factors associated with such crashes has become a major interest in transportation safety research. One challenge with such studies is that potential redundant information and correlations among different candidate factors must be addressed properly and effectively. As the most common crash data set, the police crash report contains almost all the related variables describing a crash. For example, there are more than one hundred factors in the crash report recorded by the Washington State Police (WSDOT, 2014). The academic community has made continuous efforts to develop more robust and efficient methods for the exploration and analysis of such data.

In literature, statistical models have been the primary method for the analysis of such data. From an early stage, methods such as Logistic regression (Singleton et al., 2004; Dissanayake and Lu, 2002; Hanrahan et al., 2009) have been commonly used for such analysis. Subsequently many researchers have applied novel methods to broaden the scope of applicability of the statistical models. Due to the ordinal nature of the injury outcomes (for example, ranging from no injury, to injury to fatal), ordered choice models have also been popularly applied to the analysis and severity modeling of the crash injury data (Kockelman and Kweon, 2002; Kaplan and Prato, 2012; Mohamed et al., 2013). More recently, to take into account the limitation of the assumption that all parameters estimated in the models were constant across observations and to address the heterogeneity of the crash outcomes, some multinomial logit models (Hu et al., 2010; Hu and Donnell, 2011; Shankar and Mannering, 1996) and mixed logit models (Milton et al., 2008; Malyshkina and Mannering, 2010; Zeng et al., 2017) have been developed to analyze the crash injury severities. However, the mass of complicated data on crashes nowadays will still make it difficult to use statistical models to investigate the factors related to injury severity efficiently. One restriction of such analysis is the requirement that the

* Corresponding author.
E-mail address: yinhai@edu.edu (Y. Wang).

https://doi.org/10.1016/j.aap.2017.11.028
Received 19 September 2017; Received in revised form 20 October 2017; Accepted 21 November 2017
Available online 22 December 2017
0001-4575/

data must meet some statistical assumptions (Harrell, 2001; Cohen et al., 2003; Tabachnick and Fidell, 2012) while such assumptions are hard to be valid in most crash circumstances. Another drawback of such an approach is due to their poor performance in handling several discrete variables or variables with a high number of categories (Cohen et al., 2003; Tabachnick and Fidell, 2012).

To overcome the shortcomings of statistical models, researchers have proposed many non-parametric models and artificial intelligence models for the study of crash injury patterns. The Classification and Regression Tree (CART), is a non-parametric model without any predefined underlying relationship between the dependent and independent variables. It has been widely employed for the study of crash outcomes (Chang and Wang, 2006; Yan and Radwan, 2006; Pande and Abdel-Aty, 2006; Chen et al., 2016a). The Support Vector Machine (SVM) model is also a relatively new method to solve classification problems, and has been utilized for the classification of crash injury severity (Li et al., 2012; Yu and Abdel-Aty, 2014; Chen et al., 2016b). Li et al. applied the SVM model for crash injury severity analyses, concluding that SVM models outperform the popular ordered probit model for the prediction of injury severity and factor impact assessment (Li et al., 2012). Yu and Abdel Aty compared the performance of the SVM model, random parameter models, and fixed parameter models to predict the severity of crash injuries. They concluded that SVM and random parameter models outperform fixed parameter models (Yu and Abdel-Aty, 2014). Artificial neural network (ANN), have also been applied for a long time for the classification of crash severity analysis; and recently the applications of this method have grown immensely (Abdelwahab and Abdel-Aty, 2003; Lu et al., 2012; Ali and Tayfour, 2012). All these methods have demonstrated powerful and adaptive analysis capabilities, and researchers have drawn several useful conclusions using these methods. With the coming of a big-data era, optimization efficiency has become a primary concern during the analysis procedure. For the identification of significant factors related to crash severity, performance of CART models is highly dependent on the values of the parameters and the generated model ends up with a weak generalization capability (Harrell, 2001; Chang and Chen, 2005). Also, the use of a greedy searching strategy in the CART method requires an exhaustive search procedure which can be very time consuming (Kashani et al., 2011; Prati et al., 2017). For SVMs and ANNs, the models themselves lack the capability of automatically selecting significant factors contributing to the target variable (Chen et al., 2016b).

To summarize, analysis methods used in previous studies still face a few challenges or have a limited efficiency. There is a need to develop novel algorithms to efficiently handle the traffic crash records. Considering the process of identification of significant factors as a multi-objective optimization (MOP) problem, genetic algorithm techniques may be applied as new optimal factors searching algorithms to improve the performance of the analysis (Li et al., 2012). The Non-dominated Sorting Genetic Algorithm (NSGA-II) (Deb et al., 2002) is a fast elitist multi-objective genetic algorithm. It has already found many applications in different fields, such as spectrum assignment in a spectrum sharing networks (Martínez-Vargas et al., 2016), modeling and control of output fiber length distribution in paper-making (Zhou et al., 2017), improvement of dynamic cellular manufacturing systems (Azadeh et al., 2017), and traffic signal optimization (Branke et al., 2007). However, to our knowledge, no research has used the NSGA-II algorithm for the study of injury patterns in highway crashes. The main aim of this paper is to propose a new hybrid model combining the NSGA-II algorithm and the Neural Network (NN) model for significant factors identification in highway fatal-injury crashes. Results of the study have demonstrated its capability for the searching and evaluation of optimal solutions in the global or sub-global space. The rest of this paper is organized as follows: the data description and preprocessing procedure are provided in Section 2. Sections 3 and 4 present the NSGA-NN hybrid model and discuss the analysis results. Finally, in Section 5 we present a summary about the findings.

Table 1
Summary statistics of crash data for three highways in Washington between 2011 and 2013.

		Total Crashes	Fatality-Injury Crashes		PDO Crashes	
Data in Different Year	D2011	11621	3669	31.6%	7952	68.4%
	D2012	11878	3654	30.8%	8224	69.2%
	D2013	11670	3302	28.3%	8368	71.7%
Data in Different Route	D-I5	23264	6869	29.5%	16395	70.5%
	D-I90	5988	1897	31.7%	4091	68.3%
	D-I405	5917	1859	31.4%	4058	68.6%

2. Data description

This study was performed with data from police crash reports provided by the Washington State Department of Transportation (WSDOT). The data consists of records of three years from January 2011 to December 2013 for three main Interstate highways (including I-5, I-90, and I-405), collected in the Washington State (Table 1). The total number of crashes during this period was 35,169. All the data are divided into three sets, by year, denoted as D2011, D2012 and D2013 from a temporal perspective. Additionally, they are divided from a spatial perspective (i.e., by interstate highway), into three other sets called D-I5, D-I90 and D-I405.

The severity of a crash severities is usually the target variable for the analysis of most injury patterns. According to the WSDOT, crash severity levels are often referred to using the KABCO scale, which uses the parameters: fatal (K), incapacitating-injury (A), non-incapacitating injury (B), minor injury (C), and property damage only (PDO or O)). Most of previous literatures aggregated five KABCO levels into 3 levels or 2 levels. Different divisions of two level include PDO and Injury/Fatal (Ma et al., 2017), Minor Injury and Severe/Fatal (Theofilatos, 2017; Abellán et al., 2013; Zhang et al., 2013; Mujalli et al., 2016), Death and not Death (Abu-Zidan and Eid, 2014). This study regards people are the most important part to be protected in an accident even they suffer minor injuries only, so it have broken down all the collected crash data into two categories labeled as Fatal-injury crashes (including KABC) and PDO crashes.

Over one hundred items describing the characteristics of a crash are included in the police crash report in Washington. Weather/Time, Road, Vehicle, and Driver are four categories of such items that may affect traffic safety. Based on the strong correlation among the factors, this study selects 14 factors which fall into one of these four categories. Before discussing the methodology, we describe the data cleaning and preprocessing steps. First, all incomplete records containing an “unknown” or “none” field were eliminated from the original data set. Next, for those factors that include many detailed values, a combination operation was applied to combine these small values into a single new larger value. The principle behind the combination of these different values mainly focuses on the similarity of their effects on the traffic safety. For example, “Driver Conduct” has 35 different values in the original report. “Apparently Asleep”, “Apparently Fatigued,” and “Apparently Ill” were combined into a new grouped value since they all represent cases of driver ailment. Another operation was converting the numerical factors into enumerated values. For example, all the twenty four hours in a day of the Hour of Day field were replaced with four categorical values as “midnight,” “morning rush hours,” “daytime” and “afternoon rush hours”. Table 2 shows the grouping and summaries of the value definitions for all the selected candidate factors.

3. Methodology

Fig. 1 presents the research outline of this paper. The methodology consists of two parts: a) A factors optimization model constructed on the multi-objective optimization (MOP) idea and b) a quantifying index

Table 2
Definitions and values of all the candidate factors from four categories.

Category	Factor Num.	Candidate Factor	Factor Values Definition
Weather/Time	F-1	Month	(1) Feb.–Apr. (2) May.–Jul. (3) Aug.–Oct. (4) Nov.–Jan.
	F-2	Day of Week	(1) Sun. (2) Mon. (3) Tue. (4) Wed. (5) Thu. (6) Fri. (7) Sat.
	F-3	Hour of Day	(1) Midnight. (2) Morning rush hours. (3) Afternoon rush hours. (4) Daytime.
	F-4	Weather	(1) Clear or Partly Cloudy (2)Overcast (3)Severe Crosswind (4)Sleet or Hail or Freezing Rain (5)Snowing/Raining (6) Fog or Smog or Smoke/Blowing Sand or Dirt or Snow
Road	F-5	Roadway Surface Condition	(1) Dry (2)Wet/Standing Water (3)Sand/Mud/Dirt (4)Oil (5) Ice/Snow/Slush
	F-6	Lighting Conditions	(1)Daylight (2)Dusk/Dawn (3)Dark-Street Lights On (4)Dark-No Street Lights/Dark-Street Lights Off
	F-7	Roadway Characteristics	(1)Straight & Level/Straight at Hillcrest/Straight in Sag (2)Curve & Level/Curve at Hillcrest/Curve in Sag (3) Straight & Grade (4)Curve & Grade
Vehicle	F-8	Vehicle Type	(1) Passenger Car/Taxi (2)Bus or Motor Stage/School Bus (3)Truck with Trailer/Truck/Pickup (4)Truck Tractor/ Farm Tractor and/or Farm equipment (5)Motorcycle (6)Moped/Scooter Bike
	F-9	Vehicle Action	(1)Going Straight Ahead (2)Changing Lanes (3)Making Left Turn/Making Right Turn/Making U-Turn (4)Merging (Entering Traffic) (5)Overtaking and Passing (6)Starting From Parked Position/Starting in Traffic Lane (7)Stopped at Signal or Stop Sign/for Traffic/in Roadway (8)Legally Parked, Occupied/Unoccupied (9)Illegally Parked, Occupied/Unoccupied (10)Going Wrong Way on Divided Hwy/on One-Way Street or Road/on Ramp (11)Backing
Driver	F-10	Driver Age	(1)Teenage (2)Youth (3)Middle Aged (4)Senior
	F-11	Driver Gender	(1)Male (2)Female
	F-12	Driver Conduct	(1)Operating Defective Equipment (2)Inattention/Distractions Outside Vehicle/Unknown Driver Distraction/ Operating Handheld Telecommunication/Eating or Drinking/Other Driver Distractions Inside Vehicle/Operating Other Electronic Device/Adjusting Audio or Entertainment/Grooming/Smoking/Reading or Writing/Operating Hands-free Wireless Tel/Interacting with Passengers, Animal (3)Apparently Asleep/Apparently Fatigued/ Apparently Ill (4)Improper Passing/U-Turn/Backing/Turn/Signal/Follow Too Closely/Over Center Line/Did Not Grant RW to Vehicle/Fail to Yield Row to Pedestrian (5)Exceeding Reas. Safe Speed/Stated Speed Limit, Disregard Stop Sign – Flashing Red/Yield Sign – Flashing Yellow/Flagger – Officer/Stop and Go Light/Failing to Signal (6) Had Taken Medication/Under Influence of Drugs/Alcohol
	F-13	Driver Restraint	(1)No Restraints Used (2)Lap Belt Used (3)Shoulder Belt Used (4)Lap & Shoulder Used
	F-14	Most Severe Sobriety	(1)Had NOT Been Drinking/Had NOT Been Drinking (tox test) (2)HBD – Ability Not Impaired/HBD – Ability Not Impaired (tox test) (3)HBD – Ability Impaired (tox test)/HBD – Ability Impaired (4)HBD – Sobriety Unknown

defined from the concept of Schema in genetic algorithm. The set of all the optimal solutions found by the model is called as the Pareto set. This set is used to determine the optimal factor number for the identification of crashes causing injuries. The corresponding F_s values are calculated based on the important gene analysis technique of the Pareto set. Next, the performance of selected factors is used to create a basic injury pattern and their temporal and spatial characteristics are discussed in detail. Finally, comparisons of the findings with the conclusions of the Washington's SHSP and the outputs of a popular CART method are used to prove the correctness and efficiency of the proposed method in this paper.

3.1. Optimization model mapped into multi-objective optimization (MOP)

MOP is a method for multiple criteria decision making. It is used to solve mathematical optimization problems involving two or more objective functions that need to be optimized. It has been applied in many fields of science, including engineering, economics and logistics. This study proposed a hybrid model for the optimization of safety factors by selecting fewer safety factors (g_1) and higher identifying accuracy (g_2) as the two optimizing objectives.

The main procedure of a MOP problem can be described as an iterative loop between a decision space X and objective space Y. A function f transfers the decision space into the objective space as f: X → Y. Next, an evaluation and search of the results in the objective space is

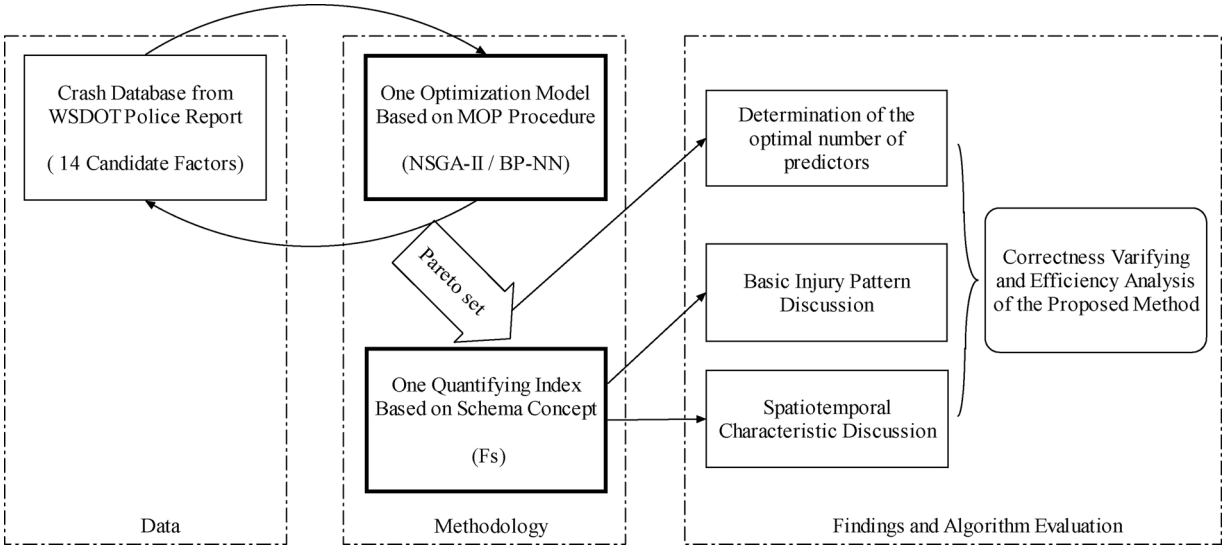


Fig. 1. Main contents and research outline of this study.

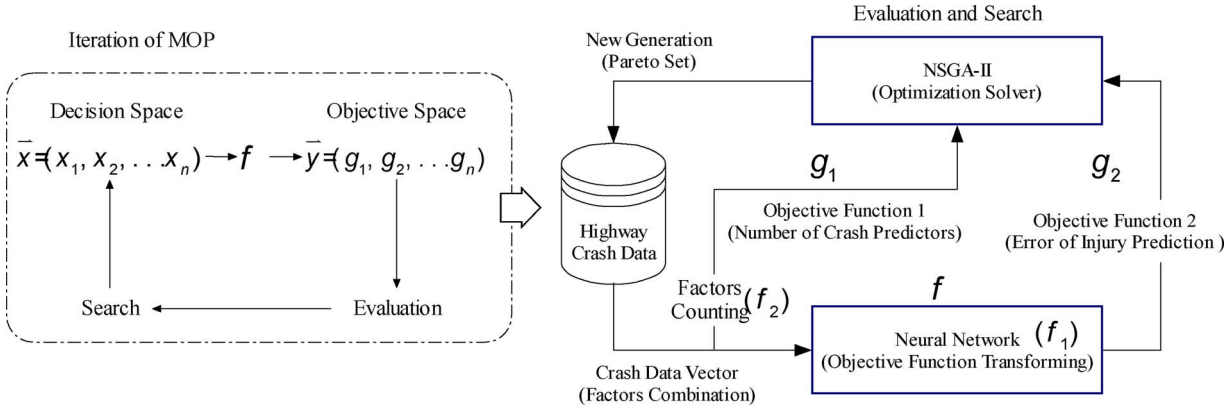


Fig. 2. Principle of Multi-objective Optimization (MOP) and its mapping in the proposed NSGA-II/NN model.

performed to provide feedback to the decision space. Fig. 2 shows the mapping of the MOP iterative loop with the proposed model, in which the NSGA-II genetic algorithm is the optimization and search solver, while a neural network (f_1) and a factor counting function (f_2) act as the transforming functions f.

To formulate the identification of significant factors in a MOP standpoint, the "decision vector" is set as $\vec{x} = (x_1, x_2, \ldots x_{14})$, where:

$$x_i = \begin{cases} 1, & i^{th}\text{variableselected}, \\ 0, & i^{th}\text{variableunselected}, \end{cases} \quad i=1, 2, \cdots, 14.$$

The first objective function is $f_1(x) = g(D(x))$, where $D(x)$ is the MSE of the neural network output from a selected sub-dataset with given factor combination. The second objective function is the total number of the selected variables as $f_2(x) = \sum_{i=1}^{14} x_i$. Thus, the MOP problem in this paper can be formulated as follows:

$$\min_{x \in X} f(x) = (f_1(x), f_2(x))^T \tag{1}$$

$$s.t. \quad f_1(x) = g(D(x)) \tag{2}$$

$$f_2(x) = \sum_{i=1}^{14} x_i \tag{3}$$

$$X = \{x \in \mathcal{R}^n | x_i(x_i - 1) = 0, \ i = 1, 2, \ldots, n\} \tag{4}$$

3.2. Implementation of the optimization model

In this study we use the NSGA-II which is a generic multi-objective algorithm used to search for non-dominating solutions. This model has better convergence properties because it stores all non-dominated solutions (NDS). It also adapts a suitable automatic mechanism based on the crowding distance (CD) to guarantee the diversity and spread of its solutions. To implement the proposed model, the chromosome of the NSGA-II is encoded as a binary sequence of fourteen bits. Every bit stands for a corresponding factor and the whole bit-string represents a certain combination of the candidate safety factors. For the initialization of the population, a parent generation that includes a size of nPOP chromosomes is randomly generated. In each loop of the following iterations, two offspring populations are generated from corresponding parent generation through genetic operators of crossover and mutation. The size of these two offspring populations are cPOP and mPOP respective. For each individual in the simulation population, all the "1" bit fields in the chromosome will be retrieved from the original data set and connected to the NN input. All the crash records are divided into training set, validation set and test set with the proportion ratio of 70%, 15% and 15%, which can prevent the model from over fit during the training period. The function used to measure the performance of the neural network is the mean squared error (MSE) which is calculated by adding the training and the testing error with weighting coefficients of 0.8 and 0.2 respectively. Then based on the output of two objective functions, better qualified chromosomes are chosen according the NDS and CD through the NSGA-II algorithm. At the end of the simulation of the iterations, the model converges to the best chromosomes that represent the optimal or sub-optimal solutions. Fig. 3 shows the main steps in the implementation of the proposed optimization model.

For the proposed hybrid model, the important parameters in the NSGA-II and NN are listed in Table 3.

3.3. Quantitative index of the significance of factors

For quantitative analysis, this study defines an index of factor significance (Fs) based on the Schema Theorem in genetic algorithm. This theorem is introduced by John Holland and has been widely used as the go to analysis tool of a GA process. From the Schema Theorem, a gene space S is composed of individuals of length l which can be thought of as a vertex set in an l-dimensional cube.

$$S = \{0,1\}^l \tag{5}$$

Considering an example case with $l = 3$, all the possible 3-bit encode can be represented by all vertices in a 3-dimensional cube. Each vertex has one corresponding bitstring value, as shown in Fig. 4 (Whitley, 1994).

The labels for each adjacent pair of vertices differ by exactly 1-bit in Fig. 4. The frontmost face has corners labeled with a "0" in the first bit. This face can be expressed with a format of "0**" where "*" is the wildcard notation. Bitstrings containing "*"s are termed schemata. Each schema corresponds to a particular hyperplane in the search space. The number of fixed bits in a schema is referred to as that hyperplane's order K. More generally,Set $a_{i_k} \in \{0,1\}$, $1 \le i_k < i_{k+1} \le l$, the schema can be expressed more generally as:

$$\mathcal{L} = \left[(i_k, a_{i_k}); 0 \le k \le K\right] = \left\{X = (x_1, \cdots, x_l) \in S; x_{i_k} = a_{i_k} (k \le K)\right\} \tag{6}$$

Eq. (6) can simply be written as $\mathcal{L}[(\{i_k\}, a_{i_k}); K]$, where $\{i_k\}$ is the list of all gene loci where $x_{i_k} = a_{i_k}$. $a_{i_k} = 0$ *or* 1 is the value of superior gene selection. K is the order of the schema with $O(\mathcal{L}) = K$.

Considering the ith gene, if the following equation (Eq. (7)) holds correct for any$x_k = 0,1 (k \ne i)$, the ith gene is referred as an important gene and the "1" value represents a superior selection (i.e. $a_{i_k} = 1$).

$$f((x_1, \cdots x_{i-1}, 1, x_{i+1}, \cdots x_l)) \ge f((x_1, \cdots x_{i-1}, 0, x_{i+1}, \cdots x_l)) \tag{7}$$

Once an important gene is identified, more important genes can be found on the basis of the acquired "hyper-plane". This reduces the search space of the optimization process by half and makes the genetic algorithm an efficient optimization engine. Also for a given factor, the first schema in which it occurred for the first time and the total number it occurred in different schemata can be used to calculate its significance quantitatively.

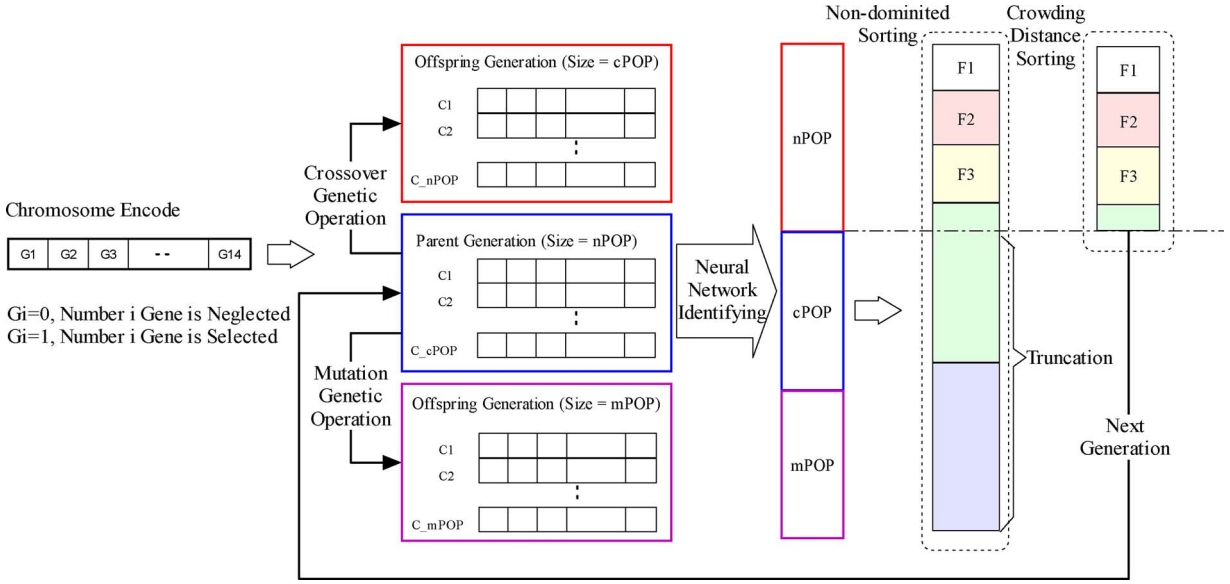


Fig. 3. Description of the main steps of the proposed NSGA-II/NN model implementation.

Table 3
Summary of parameters in implementation of the hybrid model with NSGA-II and NN.

No.	Parameter Name.	Parameter Value.
NSGA-II algorithm		
1	number of decision variables	14
2	maximum number of iterations	60
3	size of parent population	200
4	crossover percentage	0.7
5	single point crossover ratio	0.1
6	double point crossover ratio	0.2
7	uniform crossover ratio	0.7
8	mutation percentage	0.4
9	mutation rate	0.1
10	mutation bits	2
Neural Network algorithm		
1	network structure (layer number)	3
2	neurons in input/hidden/output layer	[14,10, 2]
3	Ratio of training/validation/test data set	[70%, 15%, 15%]
4	active function	Tansig
5	training algorithm	trainlm
6	Maximum number of epochs to train	1000
7	Maximum validation failures	6
8	normalization method in input/output	mapminmax
9	Performance function	MSE
10	Minimum performance gradient	$1e^{-7}$
11	Maximum mu	$1e^{10}$

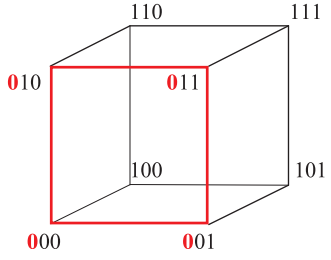


Fig. 4. Illustration of the 3-dimensional space encoded with a 3-bit bitstring.

Based on the conceptual schema explained above, the schema in this study is a set of binary strings including fourteen bits. The optimized solutions of the Pareto set for D2011 with different number of input factors are shown in Table 4.

The last column of the table shows the schema expressions of the significant factors selected using the NSGA-II algorithm. There are k important factors (genes) used to represent the expression L_k in each row. The bold numbers in parentheses in the list for L_k emphasize the newly added factor compared to the factors in L_{k-1}. For example, the "schema" expression of $L_1 = \mathcal{L}[(\{12\}, 1); 1]$shows the 12th factor of the Driver Conduct is selected as the important and elite gene. While $L_2 = \mathcal{L}[(\{9, 12\}, 1); 2]$ shows that the 9th factor of Vehicle Action is added into the optimization result of L_1. In each of the row of Table 3 we add a new important gene. Hence the serial number k is used to compare the significance of the newly added factor. For $L_1 = \mathcal{L}[(\{12\}], 1); 1]$ and $L_2 = \mathcal{L}[(\{9, 12]\}, 1); 2]$, it is reasonable to say that Driver Conduct is more important than Vehicle Action for crash injury identification.

Suppose there are M different data sets and the optimization process independently runs N times on each data set. The Pareto set is acquired from the n^{th} simulation on the j^{th} data set. The number k of the schema L_k, where the i^{th} candidate factor appears for the first time is extracted as k_{inj}. The final significance index of F_s for the i^{th} factor can be calculated as follows:

$$F_{si} = \frac{1}{M}\frac{1}{N}\sum_{j=1}^{M}\sum_{n=1}^{M}(\omega_{inj}) = \frac{1}{M}\frac{1}{N}\sum_{j=1}^{M}\sum_{n=1}^{M}((14 - k_{inj} + 1)/15 \quad (i=1, 2, ...14, j= 1, 2, 3, ...M, n= 1, 2, ...N) \tag{8}$$

4. Results analysis and discussion

4.1. Optimal factor number for injury identification

A function $f = randi$ ([01], [1,m]) is used to generate binary sequences of m bits. This function is used to perform the initialization of the population. Simulations with different initialized input factors are carried out by changing the value of m from one to fourteen.

In Fig. 5 we record and plot all the Pareto solutions from the simulations of the D2011 data. The number of initialized input factors and the MSE errors of the Pareto solutions are the two coordinate axes used in Fig. 4. The Pareto set is the minimum error of each input combination on the solid line.

Fig. 5(a) shows that input combinations with fewer factors generate more Pareto set than input combinations with more factors. At first, the errors in the Pareto Set line monotonically decrease as the number of predicting factors is increased. An optimized number of about seven or eight brings out the best identification performance, after which the

Table 4
Schema expressions of the solutions in Pareto Set.

Input Factors Number	Solution of Pareto Set	Identification Error of the Optimal Solution	Schema Expression of Significant Factors Selected
1	00000000000100	0.214	$L_1 = \mathcal{L}[(\{12], 1); 1]$
2	00000000100100	0.212	$L_2 = \mathcal{L}[(\{\mathbf{9}, 12], 1); 2]$
3	00001000100100	0.211	$L_3 = \mathcal{L}[(\{\mathbf{5}, 9, 12], 1); 3]$
4	00001000100110	0.2105	$L_4 = \mathcal{L}[(\{5, 9, 12, \mathbf{13}], 1); 4]$
5	00001001100110	0.2101	$L_5 = \mathcal{L}[(\{5, \mathbf{8}, 9, 12, 13], 1); 5]$
6	00001011100110	0.21	$L_6 = \mathcal{L}[(\{5, \mathbf{7}, 8, 9, 12, 13], 1); 6]$
7	00001001111110	0.2098	$L_7 = \mathcal{L}[(\{5, 8, 9, \mathbf{10}, \mathbf{11}, 12, 13], 1); 7]$
8	00001011111110	0.2101	$L_8 = \mathcal{L}[(\{5, \mathbf{7}, 8, 9, 10, 11, 12, 13], 1); 8]$
9	01001011111110	0.2106	$L_9 = \mathcal{L}[(\{\mathbf{2}, 5, 7, 8, 9, 10, 11, 12, 13], 1); 9]$
10	01001011111111	0.2104	$L_{10} = \mathcal{L}[(\{2, 5, 7, 8, 9, 10, 11, 12, 13, \mathbf{14}], 1); 10]$
11	00101111111111	0.2112	$L_{11} = \mathcal{L}[(\{\mathbf{3}, 5, \mathbf{6}, 7, 8, 9, 10, 11, 12, 13, 14], 1); 11]$
12	01111110111111	0.21125	$L_{12} = \mathcal{L}[(\{\mathbf{2}, 3, \mathbf{4}, 5, 6, 7, 9, 10, 11, 12, 13, 14], 1); 12]$
13	11111111111110	0.2118	$L_{13} = \mathcal{L}[(\{\mathbf{1}, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13], 1); 13]$

identification accuracy might get worse.

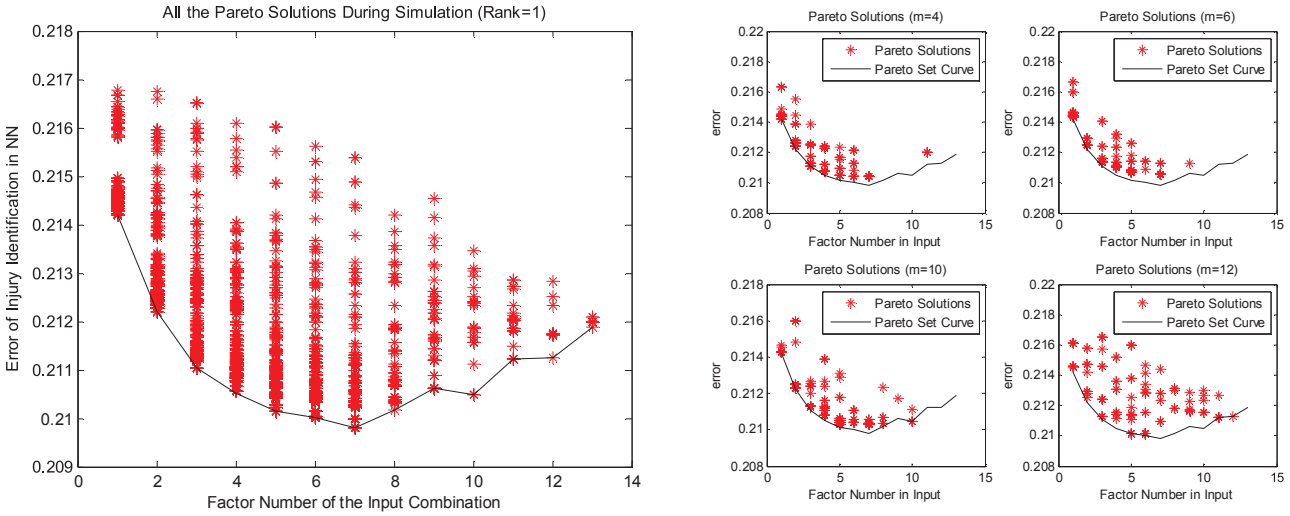
The elite optimal property of the NSGA-II can explain the above distribution of the Pareto solutions. From Fig. 5(b), we can observe from the subplots with m = 4 and m = 6, that there are only a few solutions dispersed in the number area of the larger factor. When the initial number of factors is less than seven, most of the search scope is restrained around a small combination size from the beginning of the iterations. It is very likely that at first the elite positions are occupied by solutions with fewer factors. Although some chromosomes with more factors may appear during the mutation operations in later iterations, the increasing correlations among the larger group of factors will make the errors due to misidentification worse than the errors of the previous Pareto solutions. Thus, a solution with a larger number of factors and a worse prediction error can hardly be accepted and sustained as a Pareto solution during the simulation. The strong correlation between candidate factors can be analyzed from former research and empirical knowledge. On the other hand, if the initialized number of factors is greater than seven, some Pareto solutions with a larger number of factors can be sustained from the early iteration loops. There are more recorded solutions in the larger factor number areas, shown in the subplots of in Fig. 5(b) with m = 10 and m = 12.

The curves in Fig. 6 are the Pareto sets generated from all the six available data sets for additional simulations. They all agree with the conclusion that the optimized number of factors for Fatal-injury crash identification should be around seven.

4.2. Factor significance in a basic injury pattern

This study conducted five times of simulations on each of the six data sets. So, after implementing the optimization analysis, we use M = 6 and N = 5 in formula (4) to calculate the F_s index for each factor. From Fig. 7, we observe all the fourteen factors sorted according to their F_s values.

Investigating the basic injury pattern, it shows that the top five factors of the sorted F_s list are the 1) driver conduct (F-12), 2) vehicle action (F-9), 3) road surface condition (F-5), 4) driver restraint (F-13) and 5) driver age (F-10). Driver conduct, vehicle action, and driver age are the explicit driver related factors, while vehicle action belongs to the vehicle category and describes the traffic behavior of the crashed vehicles. This factor is implicitly connected to the driver to reflect the driver's knowledge of driving, awareness of traffic situation or proper operation of the vehicle. So for factor analysis, the driver is definitely the most important element for traffic safety. Road surface condition is taken from the road category and occupies the third place in terms of importance. It is also a key influential factor for injury identification. None of the factors in the Weather/Time category appears at the top of the list. Month, day of week and hour of day are located at the bottom of the list. So, the influence of the weather/time factor on the injury pattern is shown to be of the least importance. Summing up the corresponding F_s values for each category, the results obtained are 3.02 for Driver, 1.26 for Vehicle, 1.25 for Road and 1 for Weather/Time. This shows that the Driver related factors are the most significant among all



(a) Simulations with m varying from 1 to 14 (b) Simulations with specific *m* number

Fig. 5. Distribution of Pareto solutions from optimization simulations of the data set D2011.

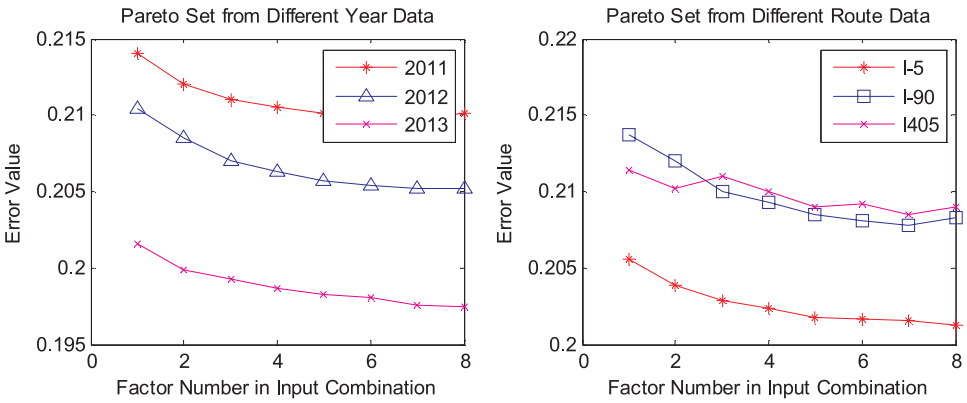


Fig. 6. Curves of Pareto set simulated from different year and route data.

the factors, while Weather/Time factors are of the least importance among all the categories. This conclusion agrees with the previous discussions of each individual factor.

In the US, each state has a Strategic Highway Safety Plan (SHSP). The SHSSP of the Washington State is called Target Zero. It is a data-driven strategic plan which focuses on the identification of the primary factors for fatal-injury accidents. The new Target Zero report in 2016 (WSDOT, 2016) worked on the data from 2012 to 2014 and the results indicate that the top three factors for high traffic safety are 1) impairment, 2) lane departure, and 3) speeding. Since the data collection for the Target Zero report only differs by a period of one year from the data collected for this study, conclusions in the Target Zero report can be used to verify the analysis result in this study to a great extent. Comparing the scope of the different categories of factors considered in this study, impairment and speeding belong to the Driver Conduct category, while lane departure belongs to the Vehicle Action category. Indeed, these two are the top two factors in the final list of this study. For a more detailed comparative analysis, Table 5 shows the statistics regarding the high risk behavior of road users from the report.

The Target Zero plan grouped the primary factors into three priority levels based on the percentage of traffic fatalities and the serious injuries associated with each factor. Priority level one includes the factors which were involved in at least 30% of the traffic fatalities or caused serious injuries due to accidents. For the calculation of the High Risk Behavioral metrics, impairment and speeding fall into priority level one and Distraction takes the first place in priority level two. All of these three behaviors fall under the Driver Conduct factors explained in this paper. The next important factor in the High Risk Behavior group is Unrestrained Occupants. It is also a very important factor considered in this paper to generate the ranked list belonging to Driver Restraint. When it is comes to Road Users, the only factor labeled as level one is when a young driver is involved. It is consistent with the conclusion of this paper that Driver Age is an important factor. The comparison of the results above demonstrates the correctness of the proposed method in

Table 5
Statistics of primary factors in the 2016 report of Washington SHSP (Target Zero Plan).

Washington State 2012–2014		Fatalities		Serious Injuries	
Priority Level	Factors Description	Number 1336	%Tatal 100%	Number 6123	%Tatal 100%
High Risk Behavior					
1	Impairment Involved	756	56.6%	1366	22.3%
1	Speeding Involved	508	38.0%	1622	26.5%
2	Distraction Involved	395	29.6%	1403	22.9%
2	Unrestrained Occupants	296	22.2%	627	10.2%
2	Unlicensed Driver Involved	248	18.6%	**	**
3	Drowsy Driver Involved	39	2.9%	194	3.2%
Road Users					
1	Young Drivers Involved	423	31.7%	2057	33.6%
2	Motorcyclists	224	16.8%	1110	18.1%
2	Pedestrians	204	15.3%	906	14.8%
2	Older Driver Involved	162	12.1%	524	8.6%
3	Heavy Truck Involved	122	9.1%	318	5.2%
3	Bicyclists	29	2.2%	294	4.8%

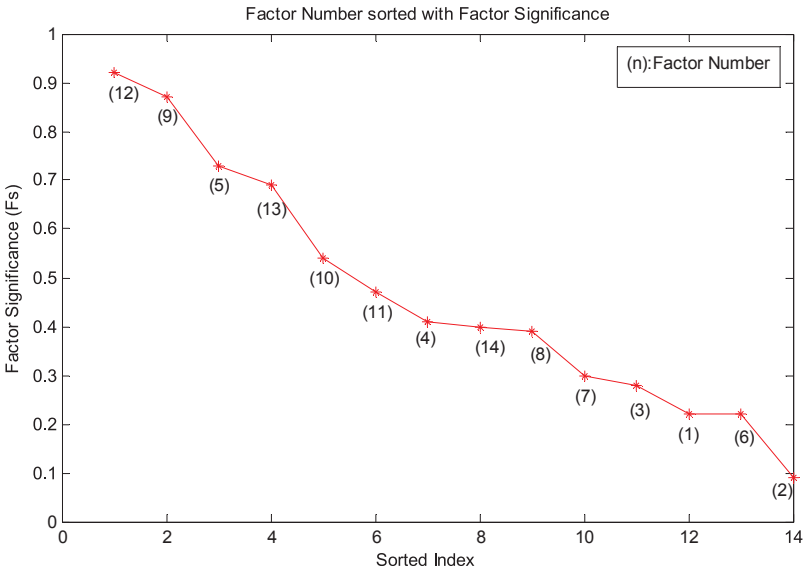
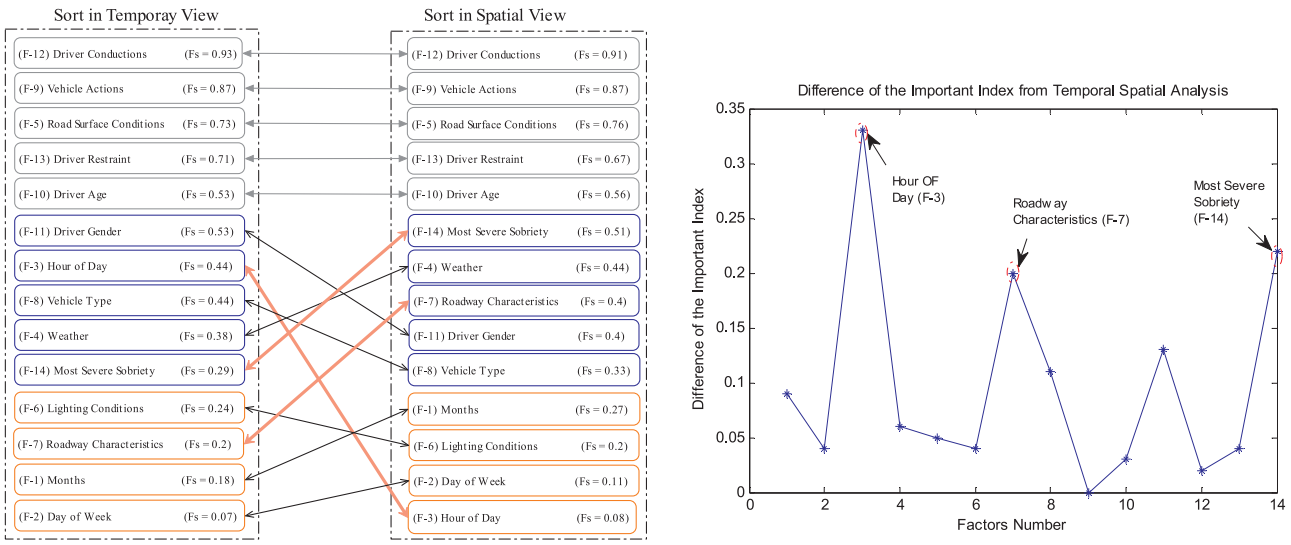


Fig. 7. Sorting result for the significance of all the factors according to their F_s index value.



(a) Comparison of sorted factors (b) Sensitive factors extraction

Fig. 8. Comparison chart of all the factors in spatiotemporal characteristic discussion and the different values of each factor for sensitive factors extraction.

factor significance analysis.

4.3. Sensitivity of factors from a spatiotemporal perspective

Using data sets from different years and different routes from the division discussed in Section 2, two injury patterns are obtained from the perspective of spatial and temporal views. The lists of the F_s indices are plotted together in Fig. 8(a). A variance analysis considering the differences between the spatial and temporal F_s arrays shows that $F < Fc$ with a P-value of 98.86%. Thus, there is no significant difference between these two arrays.

A double-arrowed line connects the same factor in each of the two lists (Fig. 8(a)). As the result of the discussion on the basic injury pattern, the top five factors from both sides are same and appear in the same order. Also, Lighting Condition, Month, and Day of Week are characteristics that stay in the lower ranking positions.

For performing sensitivity analysis, we calculate the F_s difference for each factor (Fig. 8(b)). The Hour of the Day (F-3), Most Severe Sobriety (F-14), and Roadway Characteristics (F-7) are the most sensitive factors in terms of spatiotemporal characteristics. Features from the Roadway Characteristic category inherently have spatial attributes. The geometric characteristics, including the curve, grade, or elevation parameters, may vary a lot from road to road. The more complex the said characteristics are in a roadway, the greater is the likelihood of a high number of crashes. Also, lack of sobriety may degrade the driver's performance in a more complex roadway. Thus, Roadway Characteristics and Most Severe Sobriety have higher positions in the spatial result. On the other hand, the Hour of the Day is a variable of predominantly temporal influence as it demonstrates stronger influence for the analysis of different years of data than for the analysis of the different roadways.

On the other hand, all F_s values of each category factor in Table 1 can be summed up into Table 6. Besides the conclusion that Driver Related factors have the most impact and Weather/Time has the least on the injury pattern, Vehicle factors show more significant contributions for the analysis of data from different years; Road factors contribute more significantly while comparing the different road conditions. These results agree with former studies and empirical understandings.

Table 6
Summation of the important indices in different factor category.

Category Name	All data analysis	different year analysis	different road analysis
Weather/Time	1	1.07	0.93
Road	1.25	1.15	1.36
Vehicle	1.26	1.31	1.2
Driver	3.02	2.99	3.05

4.4. Performance analysis of the algorithm

The section discusses the performance of the algorithm in three aspects. Firstly, its robust is discussed with the capacities dealing with outlier samples and correlations among factors. Then, its effectiveness is validated with a comparable analysis with the popular CART method. Finally, its efficiency is demonstrated through the fast convergence of the simulation procedure.

As one major kind of machine learning algorithm, the artificial neural network used in the proposed model can tackle the possible outliers from following aspects: 1) Large size of sample data can benefit the network training and mitigate the impact of minor outliers; 2) The activation function in the ANN is a Tansig function. It generates intermediary outputs -1 and $+1$ which can enhance the capability to deal with non-linear events and outliers. 3) The normalization operation on the input/output can also assist in reducing the impact of outliers. As to the possible correlation between the candidate factors, our multi-objective optimization based model itself has the capability to figure out the most important independent factors even there are correlation problem. In fact, the optimal factor number determined from our model has reflected the model's capability in correlation processing. The prediction errors increase after the optimal factor number which mainly because the effect of increasing correlations among factors. Initial correlation analysis shows Hour of Day and Weather with a correlation coefficient greater than 0.5 in the correlation matrix, so Hour of Day is not included in the final optimal schema of L_7. Also, Lighting and Month has a correlation coefficient as 0.23 and they are all excluded from L_7. So the proposed model can automatically figure out and discard the factors contributing less to the overall performance and the factors may make negative effect on the overall performance because of the correlation.

Table 7
Top seven factors selected from CART analysis and their corresponding importance index values (the shaded top four rows has same factors and order compared to the result from the proposed hybrid model).

Sorted Number	Selected Factor Name (Number)	Training Importance (I_t)	Validation Importance (I_v)	Weighted Importance ($I_w = 0.7 * I_t + 0.3 * I_v$)
1	Driver Conduct (F-12)	1	1	1
2	Vehicle Action (F-9)	0.9636	0.7861	0.91035
3	Roadway Surface Condition (F-5)	0.6175	0.7193	0.64804
4	driver restraint (F-13)	0.6392	0.4565	0.58439
5	Most Severe Sobriety (F-14)	0.4006	0.2632	0.35938
6	Hour of Day (F-3)	0.4033	0.1584	0.32983
7	Driver Gender (F-11)	0.3181	0.2143	0.28696

There are several popular methods, such as discrete choice models, CART, random forest, etc, used to sort factors according to their relative importance. CART has proven itself as an effective method for the analysis of traffic crashes (Chang and Wang, 2006; Yan and Radwan, 2006; Pande and Abdel-Aty, 2006; Chen et al., 2016b). This paper also applies the CART algorithm on the same crash data used in the study as aforementioned. All the data is randomly divided into a training set and a validation set in a proportion 70% and 30%. The main parameters of the CART analysis are a set based on the characteristics of the data set and the comparison of the results of several different trials. Based on the optimal number of factors determined before, Table 7 lists the top seven factors in the CART output as sorted by a weighted sum (I_w) of the Training Importance Index (I_t) and the Validation Importance Index (I_v). The weighted coefficient of I_t and I_v are set as 0.7 and 0.3 according to the ratio of sample size. The top four factors selected by the CART are Driver Conduct, Vehicle Action, Roadway Surface Condition, and Driver Restraint. They are same factors we have seen before and appear in the same order compared with the output of the proposed genetic method in this paper. Also, Driver Gender falls into the top seven factors in the output list of both methods.

As to the efficiency of the searching procedure of the optimal factors, the NSGA-II protocol used in this study shows a comprehensive and parallel performance capability. Firstly, the proposed NSGA-II model belongs to a non-parametric model while the CART model is highly dependent on the parameters. Different parameters can result in different tree structures (14). Secondly, the NSGA-NN explores factor combinations in the form of a chromosome while CART needs to split and calculate the significance of each individual factor during the search procedure. This ensures that more correlated relationships between the crash injury and corresponding factors can be explored from an overall perspective. Last but not the least, NSGA-II runs a heuristic search strategy rather than a greedy one. The implicit parallel procedure gives it a good performance for optimization efficiency. Fig. 9 describes the change in the number of non-dominant fronts as the NSGA-II/NN iteration goes on performing the simulations. It shows NSGA-II can converge to the optimal solution or the approximate optimal solutions very fast.

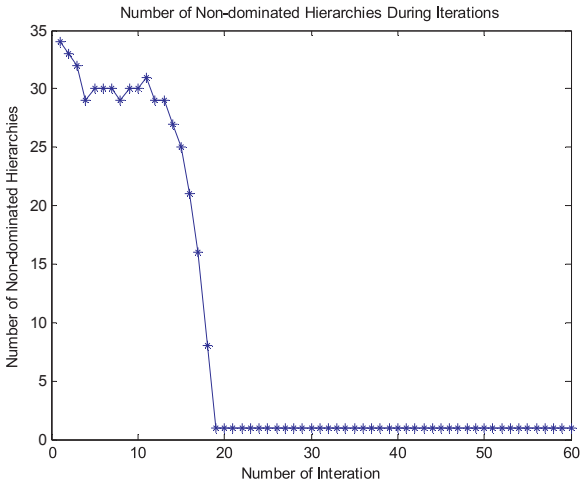


Fig. 9. Convergence process of the NSGA-II (Iteration Number = 60, data set used is D2011).

5. Conclusions

The exploration and analysis of the significant factors leading to fatalities and injuries in highway crashes can guide the traffic safety improvements. This paper addresses this problem from a multiple-factor optimization perspective and employs a hybrid model composed of the NSGA-II and a NN to conduct the analysis. The crash data were collected from three interstate highways in the Washington State from the period between 2011 and 2013. A total of 14 parameters are selected as inputs from the four categories of Weather/Time, Road, Vehicle and Driver. The two crash severities serving as model outputs are classified as Fatal-injury and PDO.

The analysis shows that when the number of input factors is less than seven, using more factors leads to higher accuracy in the crash severity identification task. Because of the strong correlation among the factors, the prediction error will decrease accordingly as more input factors are added. Based on the elitism nature of genetic algorithms and the concept of a schema, the index F_s is proposed for the measurement of the factor significance. Based on the F_s value of each factor and their variations in their positions in the sorted lists, the basic injury pattern and the factor sensitivity with respect to spatiotemporal characteristics have been discussed. Driver Conduct, Vehicle Action, Roadway Surface Condition, Driver Restraint, and Driver Age are the five factors at the top of the sorted list. Four of these factors are either explicitly or implicitly driver related. As such, this emphasizes and enforces the importance of human responsibility for traffic safety, especially for the avoidance of injury in a highway crash. Lighting Condition, the Month, and the Day of Week do not have any significant impact on the fatal-injury outcome of a crash. From a spatial and temporal perspective the Hour of the Day, Most Severe Sobriety, and Roadway Characteristics are the most sensitive factors. The corresponding conclusions from our method comply with the report of the Target Zero plan of the Washington State. The comparison of the factor selection output of the NSGA-II and CART verifies the accuracy of the proposed method and its fast convergence speed. Further, the insights of the injury pattern of highway crashes from the data-driven study in this paper helps to improve our understanding of, prevention of, and enforcement activities related to injury crashes in the future.

In this study there are some aspects which can be further improved in the future.

First, based on the global optimal solution searching strategy and fast convergence speed, the genetic algorithm based hybrid model has the potential to deal with dataset with larger size and higher variable dimensions. Collecting a larger crash dataset from different data sources and forming higher dimensional decision vectors with more candidate factors would benefit the model's improvement. Also, more data can support more independent data divisions and make further analysis possible for all of the injury levels on the KABCO scale.

Second, this paper has made a comparison between the outcomes of the proposed model and the published governmental report to validate its findings of factor significance. The fourteen variables used in this paper have a larger granularity than the factors in the SHSP report. Take Road User for example, the highest risk factor is Young Driver in SHSP report while our study only reached the Driver Age level. It is reasonable to carry out further studies to explore the significance of each possible value in the selected significant factor from this paper. For example, contribution of each kind of risky behavior in Driver Conduct can be analyzed. As such, we can get more convincing conclusions about the significance of each factor.

Third, the candidate factors in this study come from the police crash report only. Therefore, some other important traffic variables, such as the real time traffic speed and volume, are not available. Although the traditional variables like ADDT for traffic volume or average traffic speed from LOOP detectors can be extracted from other data source, their precision can't describe the traffic situation of the crash moment exactly. So some new data collection methods should be explored to collect more precise traffic related variables, which can increase the accuracy and the reliability of the analysis.

Last but not least, we put more focus on the newly employed algorithm's validation and correctness in transportation research than its accuracy in this paper. Since the proposed model has been proved to be usable in the crash data analysis, more efforts can be conducted to improve its overall accuracy performance for application in reality.

Acknowledgements

This research was supported by the State Scholarship Fund of China Scholarship Council (Grant No. 201606035008), in part by the National Natural Science Foundation of China (Grant Nos. 51329801; 61773337; 71501137), the Shenzhen Science and Technology Planning Project (Grant No. GJHZ20150316154158400), the International Postdoctoral Exchange Fellowship Program of China Postdoctoral Council (Grant No. 20150028). The authors would like to give our great appreciates to the editors and anonymous referees for their helpful and constructive comments and suggestions, which have helped to improve this paper.

References

Abdelwahab, H., Abdel-Aty, M., 2003. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transp. Res. Rec.: J. Transp. Res. Board 1746, 6–13. http://dx.doi.org/10.3141/1746-02.

Abellán, J., López, G., Oña, J.D., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Syst. Appl. 40 (15), 6047–6054.

Abu-Zidan, F.M., Eid, H.O., 2014. Factors affecting injury severity of vehicle occupants following road traffic collisions. Injury-Int. J. Care Injured 46 (1), 136–141.

Ali, G.A., Tayfour, A., 2012. Characteristics and prediction of traffic accident casualties in Sudan using statistical modeling and artificial neural networks. Int. J. Transp. Sci. Technol. 1 (4), 305–317.

Azadeh, A., Ravanbakhsh, M., Rezaei-Malek, M., Sheikhalishahi, M., et al., 2017. Unique NSGA-II and MOPSO algorithms for improved dynamic cellular manufacturing systems considering human factors. Appl. Math. Modell. 48, 655–672.

Branke, J., Goldate, P., Prothmann, H., 2007. Actuated traffic signal optimization using evolutionary algorithm. Proceedings of 6th European Congress and Exhibition on Intelligent Transport Systems and Services.

Chang, L.Y., Chen, W.C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Saf. Res. 36 (4), 365–375.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 38 (5), 1019–1027.

Chen, C., Zhang, G.H., Yang, J.F., John, C., Milton, etc., 2016a. An explanatory analysis of driver injury severity in rear-end crashes using a decision Table/Naïve bayes (DTNB) hybrid classifier. Accid. Anal. Prev. 90, 95–107.

Chen, C., Zhang, G.H., Qian, Z., Tarefder, R.A., et al., 2016b. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accid. Anal. Prev. 90, 128–139.

Cohen, J., Cohen, P., West, S.G., Aiken, L.S., 2003. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast elitist multi-objective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6 (2), 182–197.

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.

Hanrahan, R.B., Layde, P.M., Zhu, S., Guse, C.E., Hargarten, S.W., 2009. The association of driver age with traffic injury severity in wisconsin. Traffic Inj. Prev. 10 (4), 361–367.

Harrell, F., 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer, New York.

Hu, W., Donnell, E.T., 2011. Severity models of cross-median and rollover crashes on rural divided highways in Pennsylvania. J. Saf. Res. 42 (5), 375–382.

Hu, S.-R., Li, C.-S., Lee, C.-K., 2010. Investigation of key factors for accident severity at railroad grade crossings by using a logit model. Saf. Sci. 48 (2), 186–194.

Kaplan, S., Prato, C.G., 2012. Risk factors associated with bus accident severity in the United States: a generalized ordered logit model. J. Saf. Res. 43 (3), 171–180.

Kashani, A.T., Shariat-Mohaymany, A., Ranjbari, A., 2011. Data mining approach to identify key factors of traffic injury severity. Traffic Transp. 23 (1), 11–17.

Kockelman, K., Kweon, Y., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prev. 34 (3), 313–322.

Li, Z.B., Liu, P., Wang, W., Xu, C.C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486.

Lu, J., Chen, S.Y., Wang, W., Zuylen, H.V., 2012. A hybrid model of partial least squares and neural network for traffic incident detection. Expert Syst. Appl. 39 (5), 4775–4784.

Ma, X., Chen, S., Chen, F., 2017. Multivariate space-time modeling of crash frequencies by injury severity levels. Anal. Methods Acc. Res. 15, 29–40.

Malyshkina, N.V., Mannering, F.L., 2010. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accident. Accid. Anal. Prev. 42 (1), 131–139.

Martínez-Vargas, J., Domínguez-Guerrero, Á.G., Andrade, R., et al., 2016. Application of NSGA-II algorithm to the spectrum assignment problem in spectrum sharing networks. Appl. Soft Comput. 39, 188–198.

Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40 (1), 260–266.

Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., Ukkusuri, S.V., 2013. A clustering regression approach: a comprehensive injury severity analysis of pedestrian-vehicle crashes in new York, US and montreal, Canada. Saf. Sci. 54, 27–37.

Mujalli, R.O., López, G., Garach, L., 2016. Bayes classifiers for imbalanced traffic accidents datasets. Accid. Anal. Prev. 88, 37–51.

NHTSA, 2017. 2015 Passenger Vehicles Traffic Safety Fact Sheet. U.S. Department of Transportation.

Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prev. 38 (5), 936–948.

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

Shankar, V., Mannering, F., 1996. An exploratory multionmial logit analysis of single-vehicle motorcycle accident severity. J. Saf. Res. 27 (3), 183–194.

Singleton, M., Qin, H., Luan, J., 2004. Factors associated with higher levels of injury severity in occupants of motor vehicles that were severely damaged in traffic crashes in Kentucky, 2000–2001. Traffic Inj. Prev. 5 (2), 144–150.

Tabachnick, B.G., Fidell, L.S., 2012. Using Multivariate Statistics, 6th edition. Pearson, Boston.

Theofilatos, A., 2017. Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. J. Saf. Res. 61, 9–21.

WSDOT, 2014. Police Traffic Collision Report Instructions Manual. 9th edition. Law Enforcement Officers of Washington State.

WSDOT, 2016. Target Zero Washington State Strategic Highway Safety Plan. Department of Transportation, Washington (accessed 2016). http://www.targetzero.com.

Whitley, D., 1994. A genetic algorithm tutorial. Stat. Comput. 4 (2), 65–85.

Yan, X., Radwan, E., 2006. Analyses of rear-end crashes based on classification tree models. Traffic Inj. Prev. 7 (3), 276–282.

Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Saf. Sci. 63, 50–56.

Zeng, Ziqiang, Zhu, Wenbo, Ke, Ruimin, Ash, John, 2017. Yinhai wang a generalized nonlinear model-based mixed multinomial logit approach for crash data analysis. Accid. Anal. Prev. 99, 51–65.

Zhang, G., Yau, K.K.W., Chen, G., 2013. Risk factors associated with traffic violations and accident severity in China. Accid. Anal. Prev. 59, 18–25.

Zhou, P., Li, M.J., Guo, D.W., Wang, H., Chai, T.Y., 2017. Modeling for output fiber length distribution of refining process using wavelet neural networks trained by NSGA II and gradient based two-stage hybrid algorithm. Neurocomputing 238, 24–32.