Heuristic Modelling of traffic accident characteristics

Article in Transportation Letters The International Journal of Transportation Research - February 2020
DOI: 10.1080/19427867.2020.1734273

3 authors, including:

Serkan Tapkin
Antalya Bilim University
52 PUBLICATIONS
546 CITATIONS
SEE PROFILE
Heuristic Modelling of traffic accident characteristics

Emre Tercan, Erkan Beşdok & Serkan Tapkı

To cite this article: Emre Tercan, Erkan Beşdok & Serkan Tapkı (2020): Heuristic Modelling of traffic accident characteristics, Transportation Letters, DOI: 10.1080/19427867.2020.1734273

To link to this article: https://doi.org/10.1080/19427867.2020.1734273

Published online: 29 Feb 2020.
Heuristic Modelling of traffic accident characteristics

Emre Tercan\textsuperscript{a,}\textsuperscript{*}, Erkan Beşdök\textsuperscript{b} and Serkan Tapkin\textsuperscript{c}

\textsuperscript{a}Department of Survey, Project and Environment, 13th Region, General Directorate of Highways, Antalya, Turkey; \textsuperscript{b}Department of Geomatics Engineering, Erciyes University, Kayseri, Turkey; \textsuperscript{c}Department of Civil Engineering, Antalya Bilim University, Antalya, Turkey

ABSTRACT
Due to the complex structure of observation based traffic accident data and the absence of an analytic model to define their characteristics, different models are required. Accident characteristics have been modeled for different road segments with two different methods: evolutionary data clustering method and resilient neural networks. In the first method, observation data was clustered using an evolutionary search-based clustering algorithm. The first method is based on determining whether observation based test data have the conditions of a possible death or injury accident based on the distance to the cluster centers obtained. The second one is a regression method that predicts whether an accident will cause death or injury according to observation based traffic data in test road segments by using resilient neural networks. Experiment results show that data analysis methods are very effective in determining the existence of the conditions that may cause accidents resulting in death or injury.

KEYWORDS
Traffic accident; traffic safety; data clustering; resilient neural networks; evolutionary computation; optimization; differential search algorithm; Calinski-Harabasz index

Introduction

More than 1.2 million people die and about 50 million people suffer temporary injuries or become permanently disabled every year from traffic accidents around the world. In case of failure in ensuring traffic safety and taking necessary precautions, it is predicted that the number of people who will die or get injured in traffic accidents will increase by 65% around the world until 2020 (WHO 2009; Van Beeck, Borsboom, and Mackenbach 2000). Traffic accidents still cause catastrophic socio-economic losses because of injuries and permanent loss of bodily functions in many countries every year (WHO 2009). The differences in local and cultural traditions in transportation, topographical and seasonal conditions, development levels of transportation infrastructures and levels of traffic habitus in societies make it difficult to put forward global conclusions about the parameters causing traffic accidents. Therefore, in order to model the relations between the parameters causing traffic accidents on bias basis, working on a scale of the country, region, and highway separately provides more accurate conclusions. Traffic accidents have the potential to cause catastrophic socio-economic losses. Therefore, the development of new technologies is still essential to take protective and preventive actions against traffic accidents. Apprehending the relations between the parameters that cause traffic accidents are very beneficial in developing new preventive technologies against traffic accidents. Human behavior (i.e. driver (Teran-Santos, Jimenez-Gomez, and Cordero-Guevara 1999; Wang et al. 2014), passenger, and pedestrian behaviors (i.e. alcohol, hand-held cell phone, sleepiness) (Sutlovic et al. 2014; Nikolaev, Robbins, and Jacobson 2010; Burger, Kaffine, and Yu 2014; Moradi, Nazari, and Khaled 2019), vehicle properties (i.e. type, age, structure of vehicles) (Almeida et al. 2013; Hsu et al. 2015; Yau, Lo, and Fung 2006), road and environmental factors (Xu et al. 2018) are among the most important parameters collectively causing traffic accidents. Complex relations between relevant traffic parameters make it difficult to model the relations between such parameters with analytic methods. Therefore, utilizing modern data processing methods is a frequently used approach (Thakur 2014; Liu, Solomon, and Hardy 2015; Fan et al. 2019).

Turkey located in the southwest of Europe has a population of about 82 million. According to the statistics of the end of 2017, Turkey has a total of 6733 km national main highways and 12532 km national railways. The number of motor vehicles per 1000 people is 275. Fifty-four percent of 22218945 vehicles in traffic consist of personal automobiles. According to observation-based traffic accident data of 2017, the number of daily average traffic accidents is 3295 and 20 people die and 823 people get injured because of these accidents. A total of 49656 people have died and 2568996 people have sustained injuries or become permanently disabled in the 1498160 traffic accidents happening in the last ten years in Turkey (GDH 2017). The statistical results given above show that traffic accidents are very important socio-economic disasters for Turkey. Traffic and transportation problems have evolved in Turkey because of the rapid increase in population and the number of motor vehicles in traffic; therefore, these problems require developing new traffic-calming techniques and making massive transportation infrastructure investments. Road transport is more frequently used in comparison to other modes of transport in Turkey. The traffic density in the highway networks has increased because of the continual improvement of the nation. Due to the inadequacy of highway networks, irregularities in maintenance and operation conditions and partial modernization of traffic control systems in Turkey, relatively new public policies such as proliferation of mass transportation, improvement of railways and proliferation of smart transportation systems remain inadequate. In addition to, the increase in load and passenger demand, environmental conditions, negative results of extreme seasonal differences, highway design problems, and adverse behaviors of drivers and pedestrians increase the number of traffic accidents.

In this paper, the traffic accidents happened between 2012 and 2013 on 7 divided roads in Turkey have been investigated. The total length of related divided roads is 1408 km and 506 people have died and 18441 people have sustained injuries or have become permanently disabled on the accidents studied in this paper. This paper aims to study the relations between the traffic accident parameters
causing traffic accidents resulting in death or injury by analyzing observation-based traffic accident data sets consisting of 23 parameters compiled from official reports of traffic accidents happened in the above mentioned divided road segments. The data produced on the basis of simulations or observations is used to analyze the relations between the parameters causing traffic accidents scientifically. Generally, obtaining observation data is more difficult and expensive. However, observation data provides more realistic results than simulation data. This study contributes to filling the gap due to the lack of an analytical method to identify the factors that cause traffic accidents. Our main contributions to scientific research for traffic accident characteristics modelling are summarized as follows:

1. The relationship between death and injury accidents is modeled with observation data consisting of 23 different parameters that cause traffic accidents.
2. A novel differential search algorithm (DSA) based evolutionary clustering method is proposed.
3. Resilient neural networks (Rprop) Based Regression method is used in order to modelling traffic accident characteristics.
4. It is analyzed whether there is a consistency between the observation-based traffic data and the estimated traffic accident data.
5. Data model can be provided to decision support mechanisms for traffic flow management in the prevention of accidents.

The rest of the paper is organized as follows: Section 2 describes the Literature review, Section 3 introduces Structure of data set, Section 4 describes the Data analyzing methods for modelling traffic accident characteristics, and Sections 5 and 6 present the Experiments and Conclusions, respectively.

Literature review

In the literature, by using various methods, lots of studies have been undertaken to examine traffic accident data (Shafabakhsh, Famili, and Akbari 2016; Ryder et al. 2017; Briz-Redon, Martínez-Ruiz, and Montes 2019a, 2019b). Decision trees (Jung, Qin, and Oh 2016) are frequently used in literature because they are easy to construct and comprehend. When traffic accident data were analyzed on the basis of classification and regression trees, vehicle type was found to be the most important parameter affecting traffic accidents (Arenas-Ramírez et al. 2009; Chang and Wang 2006). When traffic accident data were analyzed with classification and regression trees, improperly passing and not fastening a seatbelt were found to be the most important parameters increasing the severity of injury (Thomas 1999; Kashani and Mohayman 2011). When the traffic accidents of trucks were analyzed by using classification and regression trees, alcohol intake by drivers, not-fastening seatbelts, vehicle and collision type and a number of vehicles were found to be relatively the most effective parameters affecting the severity of injury (Chang and Chien 2013). When regression methods (Lord and Manning 2010) were used to determine more effective parameters in traffic accidents causing injury, the reason and location of accidents were found to be relatively the most effective factors (Al-Ghamdi 2002). Logistic regression method is another method that enables the analysis of injuries and fatal accidents. In this study (Bedard et al. 2002), they analyzed that the fastening of a seatbelt seriously decreases accidents end up with injury by using logistic regression method. When ordered probit model was used to analyze observation-based traffic data; sex and age of driver, fastening seatbelt, driving speed, vehicle type, and collision location were found to be relatively the most important parameters on injury severity (Gomei et al. 2013). In Briz-Redon et al. (2019a, 2019b), statistical and conditional autoregressive modelling-based methods were used to analyze the traffic accident risks around schools. Bayesian networks are one of the methods used to determine traffic accidents without any assumptions (De Ona, Mujalli, and Calvo 2011; Mujalli and De Ona 2011; De Ona et al. 2013; Karimnezhad and Moradi 2017). In (De Ona et al. 2013), latent class clustering method and Bayesian networks were used together in the analysis of traffic accidents. Bayesian networks were used to determine the relation between 18 parameters including driver, vehicle, road, and environmental properties causing traffic accidents. New methods were developed to decrease the number of variables to obtain more successful data analyses and optimize the structure of Bayesian Networks used in the analysis of traffic accidents (Mujalli and De Ona 2011). Karimnezhad and Moradi (2017) analyzed strongly related traffic parameters using a bayesian network-based model for traffic accident analysis. Haridwar and Uttarakhand in India have utilized latent class clustering (LCC) and k-modes clustering techniques to analyze traffic accident data verification. The obtained results were found to be verifying heterogeneity in the data set and the utilized clustering techniques were deemed to lessen this heterogeneity in an effective manner (Kumar, Toshniwal, and Parida 2017). Artificial neural networks were also used to model the relation between the severity of injury and parameters causing traffic accidents (Sliupas and Bazaras 2013; Delen, Sharda, and Bessonov 2006; Zeng et al. 2016). Artificial neural networks were used to prioritize the factors with regard to traffic accidents (Delen, Sharda, and Bessonov 2006) and to detect the parameters causing traffic accidents (Moghaddam, Ziyadi, and Afandizadeh 2011). The variables such as highway width, head-on collision, vehicle defects, following distance, failure in adequate auditing of physical and technical qualifications of vehicles, speeding violations, and running-off-road are other parameters that increase the severity of traffic accidents. Genetic algorithm, pattern search, and artificial neural networks were used to analyze the data set total of 1000 traffic accidents happening on Tehran-Ghom highway in 2007 and it was observed that relatively the most effective traffic accident model could be established by using artificial neural networks (Kunt, Aghayan, and Noii 2012). Such methods as artificial neural networks and ANFIS (adaptive network-based fuzzy inference system) were frequently used to classify the severity of injury in traffic accidents happening on highways (Alikhani, Nedaie, and Ahmadvand 2013). The methods using k-means clustering and self-organizing maps as hybrid were developed to increase the classification accuracy of injury severity. In (Jafari et al. 2015), highway traffic death rate was predicted according to World Health Organization data by using artificial neural network optimized with a genetic algorithm.

Structure of data set

In this paper, 23 different parameters affecting traffic accidents in Turkey between 2012 and 2013 were used. These parameters were obtained from seven different divided road segments belonging to different geographical regions of Turkey. This data was provided by the Republic of Turkey General Directorate of Security. Thirty different control section points on total of 1408 km length roads have been chosen amongst 7 axis for research. The divided roads have been constructed from asphalt concrete. Geographical positions of road segments, which are the main objective of this paper, are shown in Figure 1.

A number of investments have been made in recent years to improve the highway transportation network and meet the constantly increasing needs of the dynamic Turkish economy. Table 1 includes the changes in various traffic parameters in Turkey year by year. The number of accidents including death or injury tends to
increase rapidly as the number of vehicles increase in a considerable trend in Turkey (Table 1). The relevant statistics given in Table 1 have been published by the Turkish Statistical Institute, TUIK (Traffic Statistics 2018). The number of death cases in traffic accidents prior to 2015 given in Table 1 does not include the number of death cases in hospitals after the accident due to accident-related causes. In the experiments carried out in this paper, the numbers of death cases, which occurred at the time of the accident and officially reported in the field, were used.

When Table 1 is reviewed, it can be easily observed that new methods should be developed to understand the relations between the parameters causing traffic accidents in Turkey. Considerable number of traffic accidents have occurred in these segments. In addition, these segments differ from each other in the north-south and east-west directions. Therefore, they have different geographical characteristics. The observation-based traffic data consists of 23 different parameters compiled from official reports of 8078 traffic accidents happened on national highways between 2012 and 2013. While determining the parameters affecting the traffic accidents, 23 different parameters which are included in the accident reports were preferred. These are listed in Table 2.

Data analyzing methods for modelling traffic accident characteristics

In this paper, observation data of traffic accidents including death and injury were used to determine traffic accident characteristics of relevant seven road segments. The numerical methods and analysis used for the determination of traffic accident characteristics are introduced in this chapter.

Preliminary analysis

When piecewise Pearson correlation product values (r) of the observation of fatal accidents are analyzed statistically, these observations were calculated as random variables in normal distribution defined with $N(\mu, \sigma) = N(0.3863, 0.2869)$. For 34.59% of observations: $r \geq 0.5$. For r values of the observations of accidents including injury $r \sim N(0.3271,$

---

[Table 1. The changes in various traffic parameters in Turkey between 2008–2017.]

<table>
<thead>
<tr>
<th>Year</th>
<th>Population (x1000)</th>
<th>Number of drivers (x1000)</th>
<th>Number of motor vehicles (x1000)</th>
<th>Number of accidents</th>
<th>Accidents with fatality or injury</th>
<th>Number of fatalities</th>
<th>Number of injured ones</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>71.517</td>
<td>19.377</td>
<td>13.765</td>
<td>950120</td>
<td>104212</td>
<td>4236</td>
<td>184468</td>
</tr>
<tr>
<td>2009</td>
<td>72.561</td>
<td>20.460</td>
<td>14.316</td>
<td>1053346</td>
<td>111121</td>
<td>4324</td>
<td>201380</td>
</tr>
<tr>
<td>2010</td>
<td>73.723</td>
<td>21.548</td>
<td>15.095</td>
<td>1106201</td>
<td>116804</td>
<td>4045</td>
<td>211496</td>
</tr>
<tr>
<td>2011</td>
<td>74.724</td>
<td>22.798</td>
<td>16.089</td>
<td>1228928</td>
<td>131845</td>
<td>3835</td>
<td>238074</td>
</tr>
<tr>
<td>2012</td>
<td>75.627</td>
<td>23.760</td>
<td>17.033</td>
<td>1296634</td>
<td>153552</td>
<td>3750</td>
<td>268079</td>
</tr>
<tr>
<td>2013</td>
<td>76.668</td>
<td>24.778</td>
<td>17.939</td>
<td>1207354</td>
<td>161306</td>
<td>3685</td>
<td>274829</td>
</tr>
<tr>
<td>2014</td>
<td>77.696</td>
<td>25.972</td>
<td>18.828</td>
<td>1199010</td>
<td>168512</td>
<td>3524</td>
<td>285059</td>
</tr>
<tr>
<td>2015</td>
<td>78.741</td>
<td>27.489</td>
<td>19.994</td>
<td>1313359</td>
<td>183011</td>
<td>7530</td>
<td>304421</td>
</tr>
<tr>
<td>2016</td>
<td>79.815</td>
<td>28.223</td>
<td>21.090</td>
<td>1182491</td>
<td>185128</td>
<td>7300</td>
<td>300812</td>
</tr>
<tr>
<td>2017</td>
<td>80.811</td>
<td>28.181</td>
<td>22.218</td>
<td>1202716</td>
<td>182669</td>
<td>7427</td>
<td>300383</td>
</tr>
</tbody>
</table>

[Table 2. Parameters affecting traffic accidents.]

<table>
<thead>
<tr>
<th>ID</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>In settlement?</td>
</tr>
<tr>
<td>2</td>
<td>Out of settlement?</td>
</tr>
<tr>
<td>3</td>
<td>The number of vehicle = 1?</td>
</tr>
<tr>
<td>4</td>
<td>The number of vehicle&gt;1?</td>
</tr>
<tr>
<td>5</td>
<td>Is weather clear?</td>
</tr>
<tr>
<td>6</td>
<td>Is weather rainy?</td>
</tr>
<tr>
<td>7</td>
<td>Is weather foggy or snowy?</td>
</tr>
<tr>
<td>8</td>
<td>Is it daytime?</td>
</tr>
<tr>
<td>9</td>
<td>Is it night or twilight?</td>
</tr>
<tr>
<td>10</td>
<td>Is the pavement dry?</td>
</tr>
<tr>
<td>11</td>
<td>Is the pavement wet?</td>
</tr>
<tr>
<td>12</td>
<td>Is the pavement icy or snowy?</td>
</tr>
<tr>
<td>13</td>
<td>Is there any alignment made?</td>
</tr>
<tr>
<td>14</td>
<td>Are there any horizontal curves?</td>
</tr>
<tr>
<td>15</td>
<td>Is the road curved vertically?</td>
</tr>
<tr>
<td>16</td>
<td>Is the road not curved vertically?</td>
</tr>
<tr>
<td>17</td>
<td>Is there a junction on the road?</td>
</tr>
<tr>
<td>18</td>
<td>Is there no junction on the road?</td>
</tr>
<tr>
<td>19</td>
<td>Is it a rear-end collision?</td>
</tr>
<tr>
<td>20</td>
<td>Is it a rollover?</td>
</tr>
<tr>
<td>21</td>
<td>Is it a run-off-road?</td>
</tr>
<tr>
<td>22</td>
<td>Is it a side collision?</td>
</tr>
<tr>
<td>23</td>
<td>Is it a collision with a pedestrian or fixed object?</td>
</tr>
</tbody>
</table>
0.3117) was calculated. In this case, for 28.95% of observations in the accidents including injury: \( r \geq 0.5 \). These simple statistical results show that the parameters chosen for the examination of traffic accidents can be used to examine the traffic accidents including death and injury. Total observation values of the parameters in the traffic accidents including death and injury are given in Figures 2 and 3.

**Evolutionary search-based data clustering**

Clustering techniques (Jain, Murty, and Flynn 1999) can be used to understand whether there is a relation between the observation data of traffic accidents and estimated results. In accordance with previously defined rules, data clustering techniques allow researchers to allocate data set elements to k-means clustering (Kanungo et al. 2002) consisting of data set elements sharing similarity with each other most. It is perhaps the most frequently used unsupervised learning technique in various fields such as clustering machine learning, marketing, pattern recognition, spatial database applications, medical diagnostics, image processing, computational biology, bioinformatics, and observation-based traffic data analysis. Frequently used clustering algorithms (i.e. k-means (KM) (Kanungo et al. 2002), Fuzzy C-Means (FCM) (Bezdek, Ehrlich, and Full 1984), Self-Organizing Map (SOM) (Kohonen 1990)) are based on minimizing the sum of the distance between cluster elements and cluster center (i.e. centroids). In addition, probability-based clustering techniques (i.e. Gaussian Mixture Models (Maugis, Celeux, and Martin-Magniette 2009), Expect maximization algorithm (Ruxanda and Smeureanu 2012)) connectivity-based clustering models (i.e. hierarchical clustering (Chen et al. 2015)) and graph-based clustering methods (i.e. Minimal spanning tree-based clustering (Chowdhury and Murthy 1997)) are also frequently used for data partition. KM, which is a partitioning clustering method, determines the initial values of centroid locations randomly. This affects the performance of KM radically as the number of clusters increase. If the performances of clustering algorithms are over-sensitive to initial values of their parameters, finding the best solution among the multiple runs is the most frequently preferred solution. Data clustering achievements of KM, FCM, and SOM are generally more sensitive to the number of clusters and this achievement decreases as the number of clusters increase.

Classical optimization methods are fast, deterministic, and stable. While classical methods produce the expected accuracy in low dimensional, simple geometry and linear problems, they can produce rough results in complex and multivariate problems. Evolutionary search based are used in the solution of restricted-unrestricted, continuous-discrete, linear-nonlinear, single-multivariate or differentiable-non differentiable problems. Traffic accident observations are complex and difficult to model. Evolutionary search-based data clustering methods are frequently used in various engineering problems (Günen, Atasever, and Beşdok 2017; Çivicioğlu et al. 2018). Using evolutionary computing algorithms (EA) for data clustering has becomes an important field of research (Wang, Yuan, and Cheng 2015; Hong, Chen, and Lin 2015). Generally, EAs need longer periods of time than KM, FCM, and SOM to calculate the location of the centroids. Additionally, EAs avoid local solutions better than classic clustering methods. In this paper, DSA (Çivicioğlu 2012) has been used to compute more reliable centroid values. Since DSA randomly controls the internal parameters, it does not need optimal adjustment of the initial values of the internal parameters. Therefore, DSA’s numerical problem-solving success is not significantly sensitive to the initial values of its internal parameters. DSA’s strategy for crossover and generating scale-values gives it the capacity to avoid local solutions. Also, the structure of DSA is simple and this makes it easy to adapt to different numerical problems. The Bartlett’s test (Ma, Lin, and Zhao 2015) with significance level 0.05 was used to understand whether each of 23 observation parameters affects estimated results separately. The results showed that all of 23 parameters are necessary to explain nonrandom variation in observation-based traffic data. It is difficult to determine the number of optimum k-cluster in clustering applications. Various clustering validation indexes are frequently used to determine clustering validity and optimum k value. Frequently used clustering validation indexes (Gunter and Hunke, 2003) are Silhouette index, Davies-Bouldin index, Calinski-Harabasz index, Dunn index, C index, Krzanowski-Lai index, Hartigan index, Weighted Inter-Intra index and gap-statistics (Tibshirani, Walther, and Hastie 2001). Quality indexes are widely

![Figure 2](image2.png)

**Figure 2.** Total observations of variables for fatal traffic accidents.

![Figure 3](image3.png)

**Figure 3.** Total observations of variables for injury traffic accidents.
used to determine the optimal number of clusters in unsupervised classification. Calinski-Harabasz index uses only euclidean-based attributes when investigating the optimal number of clusters. Therefore, it works very fast and its structure is simple. The most obvious differences between Calinski-Harabasz index and other quality indexes are that Calinski-Harabasz index is simple in structured, can provide high accuracy, and works fast when processing big data. Calinski-Harabasz clustering index (Calinski and Harabasz 1974) was used in this paper to determine the number of optimum cluster (i.e. k) in clustering analysis used to analyze whether traffic observation sets verify traffic estimated results. Calinski-Harabasz index is defined by using Equation 1.

\[ VRC_k = \frac{SS_B}{SS_W} \frac{(N - k)}{(k - 1)} \]  

(1)

where \(SS_B, SS_W\) and \(k\) denotes overall variance between-clusters, the overall variance within clusters and number of clusters, respectively. \(N\) denotes the number of observations. The \(SS_B\) is defined by using Equation 2;

\[ SS_B = \sum_{i=1}^{k} n_i \cdot ||m_i - \mu||^2 \]  

(2)

where \(m_i\) and \(\mu\) denote centroid of \(i^{th}\) cluster and overall mean of the sample data, respectively. \(SS_W\) is defined by using Equation 3;

\[ SS_W = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \]  

(3)

\(x\) is an observation point and \(C_i\) is the data appointed to \(i^{th}\) cluster. Well-defined clusters have relatively higher \(SS_B\) value and relatively lower \(SS_W\) value. The \(k\) value that provides the highest \(VRC_k\) value gives the best \(k\) value to be used for partition of relevant data to clusters. Calinski-Harabasz index values of observation-based traffic data set used in this paper are given in Figure 4 for \(k = 2, 3, 4, \ldots, 30\). As seen in Figure 4, \(k = 2\) is the most suitable option for the observation-based traffic data set used in this paper. Therefore, \(k = 2\) value has been used to examine the relations between traffic parameters.

If the clustering results verify the estimated results, the estimated results of the test observations are generated using the clustering parameters. Clustering observation-based traffic data according to estimated results make it possible to remove estimated results from outliers in the observation data. In this paper, coherence of observation data with estimated results was examined by using a conditioned-clustering method (Günen, Atasever, and Beş dok 2017) defined with objective function given in Equation 4.

\[ \arg \min_c \left( \sum_{i=1}^{u} \sum_{x \in C_i} ||x - \mu_i|| - t_i \right) \]  

(4)

where \(u\) and \(t\) denote the number of total clusters and estimated results of data appointed to \(i^{th}\) cluster, respectively. \(t\) shows the traffic accidents including death and/or injury with a binary valued vector. Defined problem in Equation 4 has been solved by using DSA. DSA is a population based, iterative evolutionary search-based numerical optimization algorithm. DSA was used for the solution of various optimization problem types (Çivicioğlu 2012; Alhalabi and Dragoi, 2017; Günen, Çivicioğlu, and Beş dok 2016; Guha, Roy, and Banerjee 2017). DSA has the phases such as initialization, selection of parents, mutation, and selection of new population members like other evolutionary algorithms. In DSA, population is defined as a super-organism consisting of clans including elements as much as the extent of the problem. DSA has four different mutation strategies. The first mutation strategy is a bijective strategy which ensures that each clan evolves into another clan in each iteration step. The second strategy is a surjective strategy which allows clans to evolve into the clans providing relatively better solutions with a random probability. In this strategy, more than one clan can evolve into a clan proving relatively better solutions. The third mutation strategy forces all clans to evolve into the clans which provide relatively better solutions only in limited numbers. The fourth mutation strategy evolves clans into the clans which provide the best solutions among all clans. Therefore, mutation strategies of DSA are elitist except for the first mutation strategy. DSA can use different random number generators to scale the direction matrix. This allows DSA to benefit from different scaling strategies. Generally, DSA searches for the numerical solution of which search space is looked for, with a random walk-based strategy. The clan members to be mutated in an iteration are chosen with a random probability. DSA is a simple and robust algorithm (Çivicioğlu 2012; Günen, Çivicioğlu, and Beş dok 2016). The structure of bijective DSA is given in Figure 5.

**Resilient neural networks (rprop) based regression**

Rprop is a local-adaptive learning algorithm developed for Multilayer Perceptron (MLP) neural networks (Saini 2008; Beş dok, Çivicioğlu, and Alci 2004; Dutta, Chatterjee, and Rakshit 2006). Backpropagation learning algorithm is the supervised learning algorithm which is most frequently used for MLP neural networks. Fundamental system equation of backpropagation learning is given in Equation 5.

\[ \frac{\partial E}{\partial w_{lj}} = \frac{\partial E}{\partial s_i} = \frac{\partial s_i}{\partial net_i} = \frac{\partial net_i}{\partial w_{lj}} \]  

(5)
where \( \omega_{ij} \) denotes the weight value between \( i \)th and \( j \)th neurons. \( s_i \) is the output value of \( i \)th neuron. \( net_i \) denotes the sum of neurons of \( i \)th inputs. \( \omega_{ij} \) is updated by using Equation 6.

\[
\omega_{ij}^{(t+1)} = \frac{\omega_{ij}^{(t)} + \Delta \omega_{ij}^{(t)}}{C_0} \tag{6}
\]

In classic backpropagation learning algorithm, \( \Delta \omega_{ij} \) value is generally defined by using Equation 7.

\[
\Delta \omega_{ij}^{(t)} = -\varepsilon \cdot \frac{DE}{\omega_{ij}^{(t)}} + \mu \cdot \omega_{ij}^{(t-1)} \tag{7}
\]

Here, \( \varepsilon \) is the learning rate or step size value that ensures partial derivatives to be scaled. If \( \varepsilon \) has a very low or very high value, the period of detecting the problem by neural network may increase extremely. Momentum value (i.e. \( \mu \)) is used to scale the effect of \( \Delta \omega_{ij} \) on previous value. There is no analytic method providing optimum values of \( \mu \) for a problem. Therefore, various adaptive learning algorithms using different methods, such as Rprop, were developed for the calculation of \( \Delta \omega_{ij}^{(t)} \). Rprop updates the weights of MLP network by handling only signs of partial derivatives as given in Equation 8;

\[
\Delta \omega_{ij}^{(t)} = \begin{cases} 
-\Delta \omega_{ij}^{(t)} & \text{if } \frac{DE}{\omega_{ij}^{(t)}} > 0 \\
+\Delta \omega_{ij}^{(t)} & \text{if } \frac{DE}{\omega_{ij}^{(t)}} < 0 \\
0 & \text{else}
\end{cases} \tag{8}
\]

Here, \( \varepsilon \) is the learning rate or step size value that ensures partial derivatives to be scaled. In the second phase of Rprop, \( \Delta \omega_{ij}^{(t)} \) updates are calculated by using Equation 9;
\[ \Delta_{ij}^{(t)} = \begin{cases} 
\eta^- \cdot \Delta_{ij}^{(t-1)} & \text{if } \frac{\partial g_{(t)}^{(ij)}}{\partial w_{ij}} \cdot \frac{\partial g_{(t)}^{(ij)}}{\partial w_{ijn}} > 0 \\
\eta^- \cdot \Delta_{ij}^{(t-1)} & \text{if } \frac{\partial g_{(t)}^{(ij)}}{\partial w_{ij}} \cdot \frac{\partial g_{(t)}^{(ij)}}{\partial w_{ijn}} < 0 \\
\Delta_{ij}^{(t-1)} & \text{else} 
\end{cases} \quad (9) \]

It has been defined with \(0 < \eta^- < 1 < \eta^+\) here. The partial derivative corresponding to \(w_{ij}\) always changes its sign. Therefore, if the algorithm is hindered by a local minimum in the last update, \(\Delta_{ij}^{(t)}\) value is decreased by using \(\eta^-\) factor. If the derivative maintains its sign, update value is increased slightly to expedite convergence to result. Additionally, if derivative sign changes, no adjustment will be made in the current learning phase. Practically, this is ensured by using \(\frac{\partial g_{(t)}^{(ij)}}{\partial w_{ij}} \leftarrow 0\) rule. In order to decrease the number of parameters to be adjusted, values of increasing and decreasing factors (i.e. \(\eta\)) are fixed. In this paper \(\eta^- = 0.50, \eta^+ = 1.20\) and learning rate was fixed as \(\varepsilon = 0.01\) and \([\Delta_0 \Delta_{\max}] = [0.07 50.00]\).

**Experiments**

The details of experiments conducted in this paper are given in this section. Accuracy, sensitivity, and specificity statistics, which are frequently used in binary clustering, were used to compare the clustering of data. Since our experimental study is a binary clustering (injury/death) problem, these statistical tools are used.

**Evolutionary clustering-based modelling of traffic accident characteristics**

Eight thousand seventy-eight observation-based traffic data were obtained in total seven different road segments by Turkey General Directorate of Security. Each observation data includes 23 different parameters given in Table 2. All the observation-based data belonging to seven road segments were used when creating a traffic accident characteristics model with the evolutionary-based clustering method. One thousand data were chosen randomly among 8078 data as test data. The rest 7078 data were used as training data while solving Equation 4 with DSA. Training data set consists of 118 accidents including death (1.67%) and 6960 accidents including injury (98.33%). Testing data set consists of 36 accidents (3.6%) including death, 964 accidents (96.4%). Population size, \(N\), and the parameters of DSA while solving Equation 4 determined experimentally 50 and \(p_1 = p_2 = 0.3 \cdot \kappa \cdot U(0, 1)\), respectively. In Equation 4, \(t\) value was designed as a binary-valued vector that can address each accident including death and/or injury separately; \([1 0]\): Accident including death, \([0 1]\): Accident including injury. The number of searched parameters, \(D, 23 \times 2 = 46\) for Equation 4. MaxCycle = 1e6, Low\(y_{D+1} = 0\) and Up\(y_{D+1} = 1\) for DSA. Obtaining clustering results are summarized at Table 3. As a result of the experiment, 94.90% of test data were clustered correctly and 95.08% of training data were clustered correctly. Experiment results show that observation data can provide information in reliable levels to understand whether risky conditions have occurred for accidents including death or injury by using traffic accident characteristics model shown in Equation 4.

When Table 3 is analyzed, it can be clearly seen that the presented method provides high sensitivity and specificity values. Hence, evolutionary clustering-based modelling by using DSA method has a technical potential on traffic accident data analyzing.

| Table 3. Statistical results for evolutionary clustering-based modelling of observation-based traffic accident data by using DSA.
<table>
<thead>
<tr>
<th></th>
<th>Testing/Training Data</th>
<th>Accuracy (%)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Number of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing</td>
<td>94.90</td>
<td>0.98</td>
<td>0.93</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>95.08</td>
<td>0.99</td>
<td>0.94</td>
<td>7088</td>
<td></td>
</tr>
</tbody>
</table>

| Table 4. Basic statistics for Rprop-based regression of traffic accident data.
<table>
<thead>
<tr>
<th></th>
<th>Road Segment</th>
<th>Testing/Training Data</th>
<th>Accuracy (%)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Testing</td>
<td>62.00</td>
<td>0.71</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Testing</td>
<td>71.33</td>
<td>0.79</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Testing</td>
<td>72.00</td>
<td>0.74</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Testing</td>
<td>84.45</td>
<td>0.92</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Testing</td>
<td>79.00</td>
<td>0.86</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Testing</td>
<td>96.48</td>
<td>0.98</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Testing</td>
<td>83.00</td>
<td>0.87</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Testing</td>
<td>92.17</td>
<td>0.96</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Testing</td>
<td>84.00</td>
<td>0.88</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Testing</td>
<td>89.10</td>
<td>0.92</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Testing</td>
<td>85.00</td>
<td>0.91</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Testing</td>
<td>91.51</td>
<td>0.98</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Testing</td>
<td>90.00</td>
<td>0.95</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Testing</td>
<td>88.45</td>
<td>0.92</td>
<td>0.87</td>
<td></td>
</tr>
</tbody>
</table>

**Resilient neural networks (rprop) based regression of traffic accident characteristics**

There are 10 neurons in the hidden layer and 2 neurons in the output layer of MLP network used in the experiments performed. Tangent-Sigmoid (tansig) activation function in the input layer and the linear activation function in the output layer was used. MLP networks were trained through 100000 epochs by using Rprop. Tests were made for each road segment separately. Each test data includes 100 patterns chosen randomly. The neurons on output layer produce output data corresponding to the cases of death (i.e. \([1 0]\)) and injury (i.e. \([0 1]\)). Only \([0.95]\) and higher output values were interpreted in order to avoid the problems that may be caused by misleading post-processing interpretations and evaluate test results with bias. Test results were listed according to test successes in Table 4. According to experimental results, MLP Rprop traffic accident characteristics model can show whether a random observation data have necessary conditions for the occurrence of traffic accidents including death or injury with 62-90% accuracy, 0.71–0.95 specificity and 0.58–0.88 sensitivity for testing data and 71-97% accuracy, 0.79–0.98 sensitivity and 0.68–0.95 specificity for training data.

**Conclusions**

Traffic accidents have deep socio-economic effects because they cause death, labor force loss and financial loss. It is very important to determine traffic accident characteristics to understand whether conditions necessary for the occurrence of traffic accidents are present. Traffic accident characteristics depends on the analysis of whether observation data are in accordance with estimated results. Therefore, traffic accident characteristics require establishing analytical relations between observation data and estimated results. However, because of the complex nature of the parameters of traffic accidents, it is practically difficult to develop analytical models between observation data and estimated results. Two different traffic accident characteristics analyses have been performed by using two different methods: evolutionary search-based data clustering and artificial neural networks.
based methods. Experiment results show that these methods can compute whether there is a risk of traffic accident including death or injury with a high level of accuracy according to observation data for accident characteristics. The results obtained by using the observation-based traffic data used in this paper are listed below:

(1) A new DSA-based conditional clustering has been developed to the modelling of traffic accident characteristics.

(2) MLP-Rprop traffic accident characteristics model is very successful in understanding whether observation data have necessary conditions for the occurrence of a traffic accident with death or injury.

(3) Modelling parameters that cause death and injury is very important for reducing and preventing traffic accidents.

(4) In order to reduce traffic accidents, it is necessary to examine the parameters such as road geometry, vehicle and human factor, road design, and road operation conditions separately.

(5) Observation data are very effective in modelling traffic accident characteristics.

(6) In order to better understand the relationships between the parameters affecting traffic accidents, the evaluation of all proposed parameters and rearranging the traffic accident report according to these parameters will prevent the traffic accidents.

In the traffic accidents that occur every year, many people lose their lives, are injured; are partially or completely left outside of the labor market. It is necessary to develop new analytical models in order to understand the factors affecting the occurrence of traffic accidents and to determine what changes in the instantaneous traffic parameters observed pose a traffic accident risk. Future scientific efforts will focus on developing analytical or heuristic methods that can identify risks that threaten to manage traffic networks more efficiently.

Disclosure Statement

No potential conflict of interest was reported by the authors.

ORCID

Emre Tercan http://orcid.org/0000-0001-6309-1083
Erkan Beşdock http://orcid.org/0000-0001-9309-375X
Serkan Tankan http://orcid.org/0000-0003-1417-9972

References


GDH. 2017. Summary of traffic accidents in Turkey, General Directorate of Highways.


