 https://doi.org/10.3311/PPtr.14692

Periodica Polytechnica Transportation Engineering, 49(2), pp. 146–155, 2021

Evaluation of the Impact of Spatial and Environmental Accident Factors on Severity Patterns of Road Segments

Maen Qaseem Ghadi[1]*, Árpád Török[1]

[1] Department of Transport Technology and Economics, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, H-1111 Budapest, Stoczek st. 2, Hungary
* Corresponding author, e-mail: ghadi.maen@mail.bme.hu

Received: 16 July 2019, Accepted: 19 November 2019, Published online: 06 April 2020

Abstract

Several studies have examined a wide range of accident risk factors affecting road safety. The purpose of this study is to examine the main traffic accident factors that affect the severity of road segments. The practical objective of the article is to assist specialists in identifying risk patterns both from a spatial and casualty point of view. To achieve the desired goals, accidents of a road network have been analyzed through three major steps; segmentation, black spot identification, and decision analysis. A new spatial clustering methodology has been used to divide accidents into smaller groups (or clusters) based on their spatial aggregations. The spatial characteristics are argued to be an important factor, in revealing the heterogeneity between accident data. Then, the empirical Bayesian has been applied to rank the resulted segments by severity level. During this step, the technique of decision rules has been applied to identify the main contributors to accidents in certain segments. The result shows that there is a significant relationship between the accident severity level and the traffic and geometrical characteristics (i.e. speed limits, average daily traffic, path shape) of road segments. The results also revealed that the closer the road to secure and non-hazardous road environmental conditions, the lower the risk level of the road segment.

Keywords

spatial clustering, road accidents segmentation, black spots identification, decision tree

1 Introduction

Every year, traffic accidents cause large economic and social losses to countries, families, and individuals (Ghadi et al., 2018b). Road accident is an unpredicted event which can occur under any circumstances. In most cases, accidents are influenced by human, traffic, and geometric factors of a given road segment. In other words, when certain circumstances present in a particular place and time, an accident could happen.

Several literature focused on developing or application of different black spot (BS) identification methods. Cheng and Washington (2005) defined the objective of BS identification is to "identify locations in a transportation system that have problems and their effects will be revealed by assessing their accident frequencies related to other similar locations". In more simplified words, accidents could be observed in both safe and unsafe sites, and the challenge is to avoid the false negative (sites not included in safety investigations while it is needed) and the false positive (site included in safety investigations while it isn't needed) in identifying the real hazardous spots. Elvik (2008) evaluated the applied BS methods in a number of European countries. He proved the weakness of most of the applied BS methods in these countries and recommended using the Empirical Bayesian (EB) method, as did by a number of researchers (Ghadi and Török, 2019b; Montella, 2010; Qu and Meng, 2014). The EB is a state of the art BS method that benefits from both predicted and observed number of accidents.

The heterogeneity between accident data plays an important role in the analysis process. Typically, methods of data segmentation are used to reduce heterogeneity and reveal hidden relationships between accident data. Data segmentation is considered to be a good way to classify a large data set into smaller homogeneous groups in which similarities between objects of the same group are maximal. The result of the segmentation could vary significantly depending on the applied method (Cafiso et al., 2018). Methods of data segmentation should be fitted to the

Cite this article as: Ghadi, M. Q., Török, Á. (2021) "Evaluation of the Impact of Spatial and Environmental Accident Factors on Severity Patterns of Road Segments", Periodica Polytechnica Transportation Engineering, 49(2), pp. 146–155. https://doi.org/10.3311/PPtr.14692

characteristics of input variables. Most of the current studies focus only on the definition of factors related to accident circumstance attributes as input variables for the segmentation process. Few studies focus on the spatial dependency factors. In both cases, a data mining framework has been applied as a tool to explore the required information.

Data Mining is a process for revealing hidden patterns from a large dataset. Its main goal is to collect useful information from a heterogeneous data set after classifying it into homogeneous structures. Accordingly, data mining methods can be used to segment and investigate road accident data (Barai, 2003). Decision Tree (DT) is a popular supervised data mining technique to analyze and classify accident related factors. The reason why this technique is attractive is that it does not need any certain assumption or predefined hypothesis regarding the relationship between dependent and explanatory variables, which are needed in case of the traditional statistical methods. DT is based on Decision Rule (DR) to map the most important influencing factors related to the investigated process. Abellán et al. (2013) applied a methodology to define DR from more than one DTs to enable more effective interaction between accident attributes, focusing on accident severity as a label variable. Kashani et al. (2011) studied factors affecting the injury severity of drivers involved in traffic accidents in Iran using the classification DT.

Clustering analysis is a widely used unsupervised data mining technique used to classify a large dataset into homogeneous entities based on the input variables. Many researchers applied variably related to accidents causes, results and circumstances as inputs for the segmentation process (Abellán et al., 2013; De Oña et al., 2013). Kumar and Toshniwal (2015) applied K-mode clustering and the association rule of mining to study the most contributed factors associated with the accidents' occurrence. The result showed different trends in different clusters and helped identify hidden patterns of accidents. De Luca et al. (2012) used fifteen environmental characteristics associated with accidents as input variables to help classifying accidents into clusters using the C-means technique, from which the Empirical Bayesian model was constructed. It is important to mention here that, causes of accidents in the macro-analysis are not usually the main causes of occurrence, but factors that contribute to or surrounding by the accident - such as environmental, geometric or traffic conditions -. However, the classification of accidents based only on factors related to their causes, their results or the surrounding environmental conditions can have two major disadvantages. Firstly, applying only accident result (i.e. severity, number of injuries) as an input variable may be misleading since it can happen that two accidents have the same results, but totally different causes. Secondly, accidents in a cluster generated in the above-mentioned methods - can share similar circumstances but significantly differ in their spatial distribution, as seen in some studies (Depaire et al., 2008), which can support neither the determination of risky road segments nor the definition of the required improvements.

Methods which explain accident occurrence rarely consider a spatial dependency as an efficient factor for the classification process. Some researchers draw spatial or temporal maps showing the difference in accidents densities for regions or cities (Bíl et al., 2013; Kumar and Toshniwal, 2016b). These methods did not give any specific spatial location for accidents but an overview of their distribution on the map at a city or district level. A number of traditional techniques (i.e. kernel, spatial-autocorrelation) have been more specific in identifying hazardous areas for certain parts of a road, like junctions, curves, or specific road segments (Flahaut et al., 2003; Kumar and Toshniwal, 2016c). These methods are limited to detecting only accidents within hazardous areas. On the other hand, Ghadi et al. (2018a) and De Luca et al. (2012) applied both the K-means clustering and the EB methods in order to divide accidents into specific road segments, depending on their spatial dependence, and finally identify the most hazardous segments. Although some of these methods are able to identify spatially-dependent accident groups, usually they do not consider any other characteristics of an accident itself (i.e. accidents environmental factors) in the analysis process.

Until now, the resources allocated to both spatial and environmental characteristics based road accident classification were reasonably limited. In light of the presented shortage, the main idea behind the study is to link the spatial factor and other accident environmental attributes during the classification process of road accident data, which result can strongly assist specialists in identifying risk patterns on roads both spatially and causally.

2 Data description

The data of the Hungarian network have been analyzed in this study; the analysis has covered approximately 1965 km long of Hungarian expressway. The investigated data describes the main traffic, road geometric and accident parameters from the year 2013 to 2015. Individual accident data includes information on severity level, type of accident (e.g. pedestrian accident, run-off-road, etc.),

time of occurrence (e.g. month, day, hour) weather condition (e.g. clear, foggy, rainy, snowy), visibility (e.g. daylight, restricted visibility, night with or without public light), and location of accident (e.g. x-y coordinates). Roadway data includes information on path shape (e.g. vertical and horizontal curvature), roadside hazard, median type, and pavement conditions. Traffic data includes information on the speed limit, AADT, and percent of trucks. The analysis focuses only on the road segments without intersections. During the investigated interval, 2155 fatal, serious and light injury accidents occurred on roadway segments.

3 Methodology

Traffic accidents usually occurred as a result of integrating human, traffic and road geometric factors (hereafter referred to accident environmental factors) in a particular road segment. The methodology followed, in this research, is divided into three main processes, as shown in Fig. 1. The first method has been used to segment a road network into homogeneous segments of accidents using a spatial clustering technique. The second part has been applied to rank and classify each segment by severity level using the EB method. The last part tries to identify the main contributors of accidents for different segment severity level. More details and explanation of Fig. 1 are presented in Subsections 3.1-3.3.

3.1 Method of road segmentation

Accidents are usually represented as a geographical object (e.g. as points on a map) since their spatial location are quite well defined and the spatial aspects of accident analysis open new investigation orientations in the field of accident analysis. A variety of spatial techniques have been applied to help to understand the geographical differences of point (i.e. accident) patterns. Most of these techniques, i.e. kernel, spatial autocorrelation, have concentrated on analyzing the accident's situation, based on the spatial dependences, focusing on specific high-risk sites called "BS". Moreover, the approach of classifying all accidents in a specific road into segments based on their spatial distribution is not yet researched in a detailed form, although the deep investigation could open up many new directions in the field of accident and data analysis.

One of the main assumptions of this article is that accidents occurred in a nearby area can be spatially dependent, especially considering the increased accidents density in the area (Flahaut et al., 2003). This assumption has been extended to include the spatial locations of all accident groups. In other words, all accidents will be classified into spatial groups (into specific road segments with different lengths and different accident contents), based on their spatial distribution along roads. This task is accomplished by the application of K-mean clustering and linear referencing methods.

K-mean clustering is a useful tool to classify similar objects in groups based on the input variables. The dual representation of accidents geographical location has been solved by applying a linear referencing method. The linear referencing method refers to the accident location, along a one-dimensional roadway, using a single reference value measured from a Reference Point (RP), as shown in Fig.1 (b). For example, in Figs. 1 (a) and(b), the first accident (represented by a point calculated from left) is located

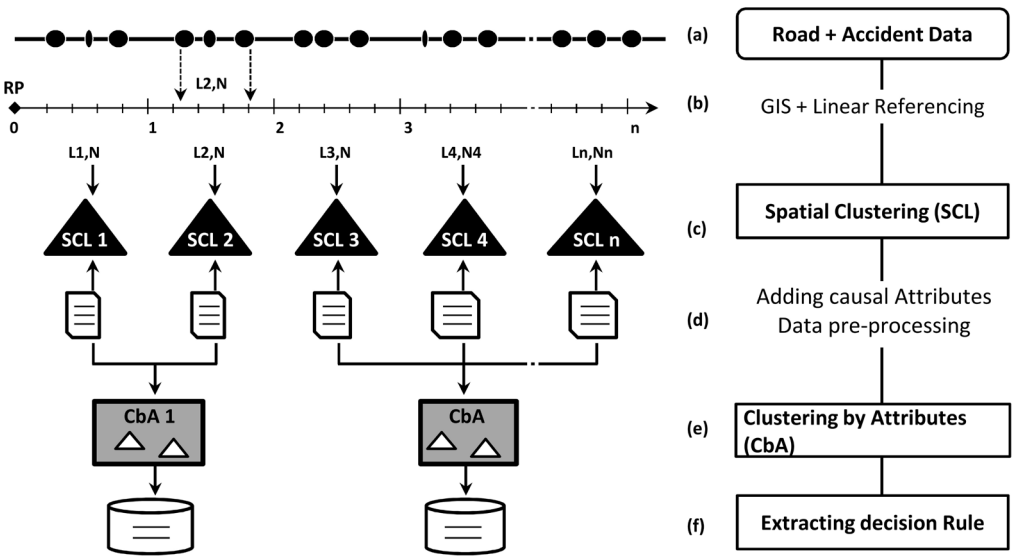

Fig. 1 Proposed framework for the analysis. L is the segment length. N is the number of accident for any road segments RSG

at a distance of 0.2 measured from the RP. This linearized uni-variable model appears to be better suited to the K-means algorithm than the binary coordinate's variable model and allows avoiding accident classifications in the same group, which are closely observed but occurred in different ways. The segmentation of accidents into K SCLs is summarized as follows (see Figs. 1 (a)-(c)):

- Identifying accident distributions on map.
- Converting the geographically referred roadways into linearly referred one-dimensional lines.
- Locating accident objects along the line road, measured from the RP.
- Applying the measured distances (in the previous step) as one input to the clustering algorithm.
- Determine the optimal number of clusters.
- Applying the K-means to locate all accident clusters.

The application of K-mean clustering and linear referencing can help in identifying the optimum length of the study segment based on the number and spatial distribution of its accidents. The longest distance between any two accidents on the linear road represents the length of the segment. In Figs. 1 (a)-(c), for instance, the spatial cluster SCL2 has a segment length equal to 0.6 (i.e. $L2 = 1.8 - 1.2$) and accidents number equal to 3 (i.e. $N2 = 3$). Therefore, at least two accidents are required to form a segment, according to this method. Accordingly, clusters of single accidents which are far enough from any neighbor accidents spatially and temporally, or characterized by significantly different types or causes, are considered to be randomly occurred noise, and should not be further investigated.

However, the proposed spatial clustering method has proved its effectiveness compared to other commonly used segmentation methods, according to Ghadi and Török (Ghadi and Török, 2019a; Ghadi et al., 2018a)

3.2 Ranking road segments by severity

To ranking road segments by severity level, the EB method has been applied. Empirical Bayesian (EB) method is used to estimate the expected average accident frequency (N_E) combining the benefit of both the observed and predicted number of accidents (N_O and N_P, respectively) in one statistical model (see Eq. (1)).

$$N_E = w \times N_P + (1 - w) \times N_O \tag{1}$$

Where: w is the weight factor represented the reliability level in predicting accident frequency. However, N_P appears to be the most influential factor, to some extent, in Eq. (1). Safety Performance Function (SPF) is usually used to measure NP. The SPF applies a regression equation to estimates the average accident frequency for a given site. The following equation (Eq. (2)) represents the general form of the SPF (Cafiso et al., 2013).

$$\mathrm{SPF} = \exp\left[\alpha + \left(\sum_n \beta_n \times \ln\left(X_n\right)\right) + \ln\left(length\right)\right] \tag{2}$$

The statistical goodness-of-fit has been used to examine the validity of the SPF using two methods: the Akaike Information Criterion (AIC) (Flahaut et al., 2003) and the Pearson Correlation Coefficient (PCC). AIC is a method that provides a mean for model selection. It is accounts for the possibility of over-fitting the data by trading-off between the simplicity and the goodness-of-fit of the model. AIC aims for selecting the model that is best fit the data with the least complexity. The best model is the one with the least AIC value. The PCC works as a complementary test to examine the data fitting. The PCC measures the linear correlation between the predicted and observed accidents using a value between +1 and −1. As the absolute value of the PCC increases, the correlation between the expected data and the observed data increases.

The explanatory variables of the SPF were selected using the stepwise methodology by testing the effect and significance of inserting or removing the various SPF variables. Possible explanatory variables are described as follows:

- Average Annual Daily Traffic (AADT): AADT is one important factor in predicting accident frequencies (Hiselius, 2004).
- Speed limit: It is also an important factor in accident risk (Ghadi and Török, 2017).
- Percent of trucks and small vehicles.
- Horizontal deflection angle: It measures the rate of curvature change from a straight line to stay on course. For any single SCL based segment (n) the road is divided into M smaller sections ($m1$, $m2$, … M) and the summation deflection curves (ADF) is measured with the help of the ArcGIS software and Eq. (3).

$$ADF_n = \frac{\sum_{m=1}^{M} \left| DF_m \right|}{SL_n} \tag{3}$$

Where: DF is the deflection angle measured for horizontal curves, for each m section, SL is the total length for segment n.

3.3 Decision analysis

Decision Rule (DR) is a well-known data mining method used in supervised learning. DR helps in recognizing useful information and relationships between different attributes in a large dataset. Safety engineers usually apply DR to discover behaviors of road accident datasets (De Oña et al., 2013). An especially useful DR tool is the Decision Tree (DT) method. DR and DT have been widely used to analyze road accident data (Abellán et al., 2013; Chang and Chen, 2005; Kumar and Toshniwal, 2016a). DT includes a set of useful methods which are applicable to uncertain decision processes. DT can be used in the field of data classification and regression without any requirement to specify parameters or previous relationships between dependent and independent variables. These characteristics make it a useful tool for classifying accident data based on the most influential factors.

The process of extracting DR's from DT is constrained by the structure of the tree. Classification And Regression Tree (CART) and C4.5 methods are usually used to construct the DT (Aljofey and Alwagih, 2018; Kashani et al., 2011). The main differences between the two methods are originated from the structure of the tree and the splitting criteria. CART method applies the Gini index to measure the dispersion among the variables, and it results in a binary tree, while the C4.5 method applies the Gain ratio (Quinlan, 1993). In this paper, the CART-based DT method is used to define DR's from road accident data.

To better understand the difference in severity, the road segments are divided into four groups of equal size but different severity level. Each group took a particular code; B1, B2, B3 and B4, and severity level ranked from higher risk (B1) to lower risk (B4) in ascending order. DT is then applied to derive the DR from the resulted groups.

DT is a technique used to predict the class label. The four road segments groups (B1, B2, B3, and B4) are classified into the target field of the class label. To verify the quality of the classification process, the data is divided into two groups: a training group and a test group. Two-thirds of the cases has been taken by the training group, while the test group took the rest of the cases. Because accident data was collected sequentially in time, the R software package was used to shuffle the data to ensure random selection of training and testing groups. The training group is used to create the DT model while the test group to ensure the accuracy of the model. The CART DT was formed by some important variables related to individual accident attributes. DR's were derived from the constructed CART tree.

4 Results and discussions

4.1 Road segmentation results

The linear reference technique has been used to locate all accident along the different roadways. The reference locations have been considered as a single-input variable in the K-means algorithm to divide accidents into clusters according to their spatial aggregations. Applying the proposed segmentation method for the given case study has resulted in 181 segments regarding the analyzed period (2013-2015). The description of the resulted segments is summarized in Table 1.

As mentioned in Subsection 3.1, the K-means clustering has an advantage of relating the identified road segment length to the cluster length. Therefore, empty road segments between the resulted clusters with no accident history are excluded from any further analysis. And, this explains why lower total segments length (1224 km) is involved in the clustering analysis (see Table 1) than the total road length (1965 km). It can also be noted from Table 1 that at least two accidents are required to form a cluster. Since it is assumed that a cluster of a single accident that is far enough from any neighboring accident, spatially and temporally, is more likely to occur randomly. This explains why lower total accident frequency (1926) included in the analysis (see Table 2) than the real total number (2155).

Table 1 Description statistics of the spatial segmentation outcome

Parameter	Total	Min.	Max	Mean	Standard deviation
Segment length (km)	1225	0.23	12.32	6.76	4.57
Accident frequency	1926	2*	13*	4*	2.54

*The values are averaged per the three study years (2013-2015)

Table 2 Parameters values and goodness-of-fit of the developed SPF

α (Intercept)		-3.577
[p-value]		[0.070]
β1 (AADT)		0.674
[p-value]		[>0.001]
β2 (Speed)		-0.465
[p-value]		[0.041]
β3 (HDA)		1.059
[p-value]		[>0.001]
Over-dispersion (k)		1.179
Goodness of fit	AIC	634
	PCC	0.768

4.2 Black spots identification results

After segmenting the road accident data spatially using the clustering method, the EB method has been applied to identify and rank BS segments. Because the EB method combines both observable and predicted accident frequencies, the predicted values have been estimated using the SPF. Accident data of the year 2013 has been used to develop the SPF for the case study dataset. Applying the stepwise forward approach has finally resulted in the selected three explanatory variables: AADT, speed, and degree of the horizontal curve as an input for the SPF.

The dataset of the dependent variable (accident frequency) and independent variables (AADT, speed, and horizontal curve) have been identified for each road segment. When a segment has not included a constant AADT, traffic, or speed values the length is used to estimate their average values (as a linear length-weight). The SPF has been developed using the data of the 181 road segment. The model calibration results are presented in Table 2.

Most of the resulted parameters (i.e. α, $\beta 1$, $\beta 2$ and $\beta 3$), presented in Table 2, are statistically significant to 95 % confidence level. The intercept (α) is statistically significant to a 90% confidence level. This might be explained due to applying a relatively small sample size in developing the model. However, α is not much concern, for a small deficiency, because it just acts as a calibration for the model.

The Akaike Information Criterion (AIC) and the Parson Correlation Coefficient (PCC) have been used to evaluate the goodness-of-fit of the model. The road accident dataset almost fitted with the model with a low acceptable error as presented in the AIC value (676). Moreover, the study of the relationship between the observed and predicted accident data for the year 2015 gives a good correlation presented with the higher positive PCC value (0.768) approaching one, as shown in Table 2.

However, the goodness of fit result could also approve the efficiency of the applied segmentation method. Details of comparing the applied spatial segmentation method with the other well-known methods are presented in Ghadi and Török (2019a) work.

The developed SPF has been used in the EB method to rank the 181 road segments based on accident severity. To describe the severity differences, the set of road segments are divided equally into four groups. Groups are code as B1, B2, B3, and B4. Segments are classified into the groups based on their severity levels ranked from higher risk (B1) to lower risk (B4) in ascending order.

4.3 Extracting decision rule from decision tree

Twelve predictor variables have been used with the four qualitative target variables (i.e. BS1, BS2, BS3, and BS4 from highest to lowest severity in ascending order) to identify important accident patterns along the road. The predictors include environmental characteristics (e.g. AADT, trucks percent, speed limit, path shape, vertical alignments, road surface conditions), temporal characteristics (e.g. time, weekday, month), and accident variables (e.g. severity, visibility, weather condition). A descriptive statistics of accidents' predicted variables are presented in Table 3 differentiating the data according to BS severity levels (BS1, BS2, BS3, and BS4).

To define the adequate decision rules, CART decision tree and the Gini splitting criterion have been used. Fig. 2 shows the produced classification tree. The tree has five terminal nodes with five DR's. It can be easily distinguished that AADT, path shape, and the speed limit are the primary selected splitter attributes for the given phenomenon described by the tree. The splitter AADT appeared in the top of the tree as the most influential variable in identifying BS accidents by their spatial locations. This indicates that these variables are the most critical in classifying clusters of BS accidents by severity.

All of the selected variables are related to environmental factors and all accidents within a certain cluster can take the same value of these environmental attributes. In other words, the accident density of certain segments and the associated environmental characteristics, have the main role in classifying BS segments by severity level.

This is rational since most of the well-known BS identification methods use these environmental factors to identify BS sites. For instance, the EB method relies mainly on the AADT to predict future changes in accident frequency (Elvik, 2008; Ghadi et al., 2018a). Similarly, the accident ratio method identifies BS segments based on the value of accidents per segment length or accidents per AADT (Borsos et al., 2016).

The interpretation of the results is straightforward. Segments of BS1 are ranked to be the riskiest BS sites, according to the applied EB method. These segments include about one-third of the total sample of accidents and have the highest accidents density per unit length (2.98 accidents per km). It has to emphasize that about 40 % of BS1 accidents occurred under high traffic volume conditions (AADT >10000). This type of accidents constitutes 26 % of total accidents included in the training CART tree, as shown in the terminal node 1 (from the left) of Fig. 2.

Table 3 Description of accident attributes

Attribute	Description	Accident frequency	% of accidents per BS severity level			
			BS1	BS2	BS3	BS4
Accident Severity	Fata	143	0.29	0.33	0.21	0.17
	Sever	622	0.32	0.30	0.20	0.18
	Slight	1161	0.32	0.29	0.22	0.16
AADT	<3000 (AADT1)	140	0.32	0.11	0.16	0.40
	3000-10000 (AADT2)	1296	0.26	0.29	0.24	0.20
	>10000 (AADT3)	490	0.47	0.37	0.15	0.02
Speed Limit	<=60	758	0.31	0.31	0.24	0.15
	>60	1168	0.33	0.30	0.20	0.18
The path shape	Straight	1237	0.27	0.31	0.24	0.18
	Curve	416	0.42	0.28	0.15	0.16
	Multi-curve	40	0.35	0.15	0.30	0.20
	Others	233	0.38	0.30	0.19	0.13
Type of accident	Face to face	306	0.33	0.32	0.20	0.15
	Vehicles in the same direction	353	0.35	0.29	0.20	0.16
	Standing vehicle, solid object collision	451	0.28	0.28	0.23	0.20
	Slipping, overturning	52	0.21	0.37	0.21	0.21
	Run Off the Road	200	0.28	0.29	0.25	0.18
	Hitting a pedestrian	149	0.30	0.34	0.21	0.16
	Others	415	0.36	0.29	0.20	0.14
Weather conditions	Clear	1167	0.31	0.30	0.21	0.18
	Cloudy	459	0.31	0.32	0.22	0.16
	Foggy	65	0.22	0.14	0.35	0.29
	Rainy	184	0.40	0.32	0.20	0.09
	Snow	51	0.43	0.22	0.22	0.14
Visibility	Daylight, natural light	1396	0.32	0.30	0.21	0.17
	Night, with active lighting	470	0.31	0.30	0.24	0.15
	Night without public lighting	59	0.32	0.29	0.19	0.20
Ground cover	Flawless	1252	0.33	0.29	0.21	0.16
	Fractured, uneven, wavy	518	0.33	0.30	0.19	0.17
	pits	156	0.17	0.34	0.28	0.22
Vertical alignments	Flat	1741	0.31	0.30	0.22	0.17
	Slope	109	0.38	0.31	0.14	0.17
	Ascending	76	0.45	0.28	0.13	0.14
Time of day	Mid-night (0-6)	180	0.34	0.27	0.24	0.15
	Morning (6-12)	599	0.31	0.28	0.23	0.17
	Afternoon (12-18)	770	0.32	0.32	0.18	0.18
	Night (18-0)	376	0.32	0.29	0.23	0.15
Day of week	Work day	1337	0.32	0.29	0.21	0.17
	Weekend	589	0.31	0.31	0.21	0.16
Month	Winter (12-2)	382	0.33	0.28	0.24	0.15
	Spring (3-5)	392	0.28	0.31	0.23	0.18
	Summer (6-8)	633	0.33	0.31	0.20	0.16
	Autumn (9-11)	519	0.33	0.29	0.20	0.18

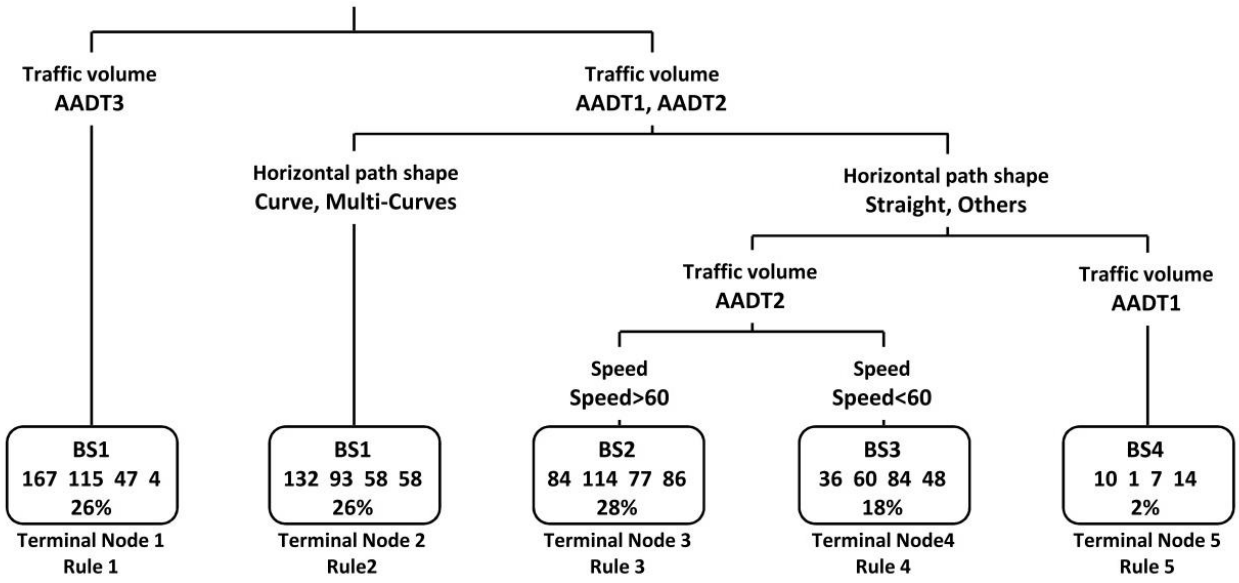

Fig. 2 The output of CART tree

This can be explained due to the strong correlation between the AADT and accident frequency. Benedek et al. (2016) proved that AADT is the most influential variable in accident prediction models, and accident frequency is directly proportional to the AADT and the intensity of the linear dependency between the two variables (accidents and traffic) is strong.

At lower traffic volume (AADT < 10000) the CART tree is branched again but in this case by the horizontal path shape. At the first branch, accidents follow a curved path shape (horizontal curve) rule ended up by terminal node 2, which is also dominated by accidents of BS1 segments. Generally, accidents occurred on curved segments constitute about 24 % of all accidents, of the case study, and included 20 % of fatal accidents. This is supported by previous studies which have shown that the rate of accidents on horizontal curves is higher than 1.5 to 4 times than on straight roadways (Polus, 1982).

At the second tree branch of the horizontal path shape joint point, road segments become straighter and less dangerous from a horizontal curve point of view. In this case, high-speed limits (speed≥60 km/hour) are the distinguishing factor in the rule of the terminal node 3 (i.e. rule 3). This is not surprising as driver speed and speed limits have been recognized as one of the main contributing factors of road accidents. Speed limits can affect accident likelihood, severity, and density per road segment (Ghadi and Török, 2017). This rule is almost dominated by accidents of the second hazardous road segments (i.e. BS2). However, a good per cents of BS2 accidents follow the same BS1 rules (rule 1 and rule 2, as described in Fig. 2).

The last less hazardous terminal nodes; node 4 and node 5, are almost dominated by accidents of BS3 and BS4 respectively. The typical properties of these nodes can be characterized by average environmental factors that follow their rules. For instance, accidents of node 4 mainly occurred in case of lower speed limits (<60 km/hour), moderate traffic volume (10000<AADT<3000), and straight road section. While accidents of node 5 mostly occurred under low traffic volume conditions (AADT <3000) and a straight road section. This can be explained by bad weather conditions that accompanied most BS3 and BS4 accidents (see Table 3).

It can be also concluded from the tree that as the rules of the nodes are more dominated by lower severity segments (BS1, BS2, BS3, and BS4 respectively), accidents contributing factors (AADT, speed, horizontal path shape) fit better to the safe and non-hazardous road scheme conditions. This is logical since accidents within the low severity BS segments are more likely to be caused by human errors or special events rather than specific environmental factors like in the case of BS1 and BS2 segments.

5 Conclusion

Classification of road accident data still remains an important road safety issue. This paper has presented a new methodology for examining the main contributing variables on accident risk severity of road segments. The idea behind

this research is to link the spatial factor and other accident environmental attributes in the classification process of road accident data, which result can strongly support decision making processes.

To achieve the desired goals, the accident data of the analyzed Hungarian expressway has been manipulated in three major steps (see Fig. 1): spatial segmentation, black spots identification, and decision analysis. In the first step, the K-means clustering algorithm has been used in an innovative way, with the assist of the linear referencing approach, to classify all accidents in the road network into segments based on their spatial dependency. This resulted in homogeneous segments with more flexible length pattern. Secondly, the Empirical Bayesian (EB) method has been used to classify the resulted segments into four groups according to severity level. In the next step, the decision tree has been applied to identify the main accident contributors for each segment group.

Spatial clustering method has resulted in 181 segments with an average accident content equal to 3.5 accidents per segment. The strength of this method has been demonstrated from the robustness of the developed accident prediction model, which is used by the EB method in order to rank road segments by severity. The resulted road segments have been divided into four groups with a different severity level. The application of decision tree on the resulted groups has revealed a significant relationship between three environmental factors (traffic volume, speed limits, and horizontal path shape) and the typical accident severity of the analyzed segments. The results have also proved that the less special the characteristic of the road environment is, the lower the risk level the road segment has.

In this study, the key factors characterizing road environment have been identified for homogeneous accident groups instead of individual accidents to avoid errors originating from randomly occurred accidents resulted by events independent of infrastructure (e.g. individual human errors). Accordingly, the proposed methodology can help in deriving useful information from similar road segments in terms of their accident characteristics as well as their location on the network. The information can help road safety experts in understanding the accident profiles of the clusters and their pattern along with the road network and proposing uniform safety improvements that could fit better to the defined segments classes. In the next step of the research, the effect of accident contributors for different data samples should be analyzed.

References

Abellán, J., López, G., De Oña, J. (2013) "Analysis of traffic accident severity using Decision Rules via Decision Trees", Expert Systems with Applications, 40(15), pp. 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027

Aljofey, A. M., Alwagih, K. (2018) "Analysis of Accident Times for Highway Locations Using K-Means Clustering and Decision Rules Extracted from Decision Trees", International Journal of Computer Applications Technology and Research, 7(1), pp. 001–011. https://doi.org/10.7753/IJCATR0701.1001

Barai, S. K. (2003) "Data mining applications in transportation engineering", Transport, 18(5), pp. 216–223. https://doi.org/10.1080/16483840.2003.10414100

Benedek, J., Ciobanu, S. M., Man, T. C. (2016) "Hotspots and social background of urban traffic crashes: A case study in Cluj-Napoca (Romania)", Accident Analysis & Prevention, 87, pp. 117–126. https://doi.org/10.1016/j.aap.2015.11.026

Bíl, M., Andrášik, R., Janoška, Z. (2013) "Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation", Accident Analysis & Prevention, 55, pp. 265–273. https://doi.org/10.1016/j.aap.2013.03.003

Borsos, A., Cafiso, S., D'Agostino, C., Miletics, D. (2016) "Comparison of Italian and Hungarian Black Spot Ranking", Transportation Research Procedia, 14, pp. 2148–2157. https://doi.org/10.1016/j.trpro.2016.05.230

Cafiso, S., D'Agostino, C., Persaud, B. (2013) "Investigating the influence of segmentation in estimating safety performance functions for roadway sections", In: Transportation Research Board 92nd Annual Meeting, Washington, DC, USA, Article number: 13–4372.

Cafiso, S., D'Agostino, C., Persaud, B. (2018) "Investigating the influence of segmentation in estimating safety performance functions for roadway sections", Journal of Traffic and Transportation Engineering (English Edition), 5(2), pp. 129–136. https://doi.org/10.1016/j.jtte.2017.10.001

Chang, L. Y., Chen, W. C. (2005) "Data mining of tree-based models to analyze freeway accident frequency", Journal of Safety Research, 36(4), pp. 365–375. https://doi.org/10.1016/j.jsr.2005.06.013

Cheng, W., Washington, S. P. (2005) "Experimental evaluation of hotspot identification methods", Accident Analysis & Prevention, 37(5), pp. 870–881. https://doi.org/10.1016/j.aap.2005.04.015

De Luca, M., Mauro, R., Lamberti, R., Dell'Acqua, G. (2012) "Road Safety Management Using Bayesian and Cluster analysis", Procedia - Social and Behavioral Sciences, 54, pp. 1260–1269. https://doi.org/10.1016/j.sbspro.2012.09.840

De Oña, J., López, G., Abellán, J. (2013) "Extracting decision rules from police accident reports through decision trees", Accident Analysis & Prevention, 50, pp. 1151–1160. https://doi.org/10.1016/j.aap.2012.09.006

Depaire, B., Wets, G., Vanhoof, K. (2008) "Traffic accident segmentation by means of latent class clustering", Accident Analysis & Prevention, 40(4), pp. 1257–1266. https://doi.org/10.1016/j.aap.2008.01.007

Elvik, R. (2008) "The predictive validity of empirical Bayes estimates of road safety", Accident Analysis & Prevention, 40(6), pp. 1964–1969. https://doi.org/10.1016/j.aap.2008.07.007

Flahaut, B., Mouchart, M., San Martin, E., Thomas, I. (2003) "The local spatial autocorrelation and the kernel method for identifying black zones. A comparative approach", Accident Analysis & Prevention, 35(6), pp. 991–1004. https://doi.org/10.1016/S0001-4575(02)00107-0

Ghadi, M., Török, Á. (2017) "Comparison Different Black Spot Identification Methods", Transportation Research Procedia, 27, pp. 1105–1112. https://doi.org/10.1016/j.trpro.2017.12.104

Ghadi, M. Q., Török, Á. (2019a) "Comparison of Different Road Segmentation Methods", PROMET – Traffic & Transportation, 31(2), pp. 163–172. https://doi.org/10.7307/ptt.v31i2.2937

Ghadi, M., Török, Á. (2019b) "A comparative analysis of black spot identification methods and road accident segmentation methods", Accident Analysis & Prevention, 128, pp. 1–7. https://doi.org/10.1016/j.aap.2019.03.002

Ghadi, M., Török, Á., Tánczos, K. (2018a) "Integration of Probability and Clustering Based Approaches in the Field of Black Spot Identification", Periodica Polytechnica Civil Engineering, 63(1), pp. 46–52. https://doi.org/10.3311/PPci.11753

Ghadi, M., Török, Á., Tánczos, K. (2018b) "Study of the Economic Cost of Road Accidents in Jordan", Periodica Polytechnica Transportation Engineering, 46(3), pp. 129–134. https://doi.org/10.3311/PPtr.10392

Hiselius, L. W. (2004) "Estimating the relationship between accident frequency and homogeneous and inhomogeneous traffic flows", Accident Analysis & Prevention, 36(6), pp. 985–992. https://doi.org/10.1016/j.aap.2003.11.002

Kashani, A. T., Shariat-Mohaymany, A., Ranjbari, A. (2011) "A Data Mining Approach To Identify Key Factors of Traffic Injury Severity", Promet – Traffic & Transportation, 23(1), pp. 11–17. https://doi.org/10.7307/ptt.v23i1.144

Kumar, S., Toshniwal, D. (2015) "A data mining framework to analyze road accident data", Journal of Big Data, 2, Article number: 26. https://doi.org/10.1186/s40537-015-0035-y

Kumar, S., Toshniwal, D. (2016a) "A data mining approach to characterize road accident locations", Journal of Modern Transportation, 24(1), pp. 62–72. https://doi.org/10.1007/s40534-016-0095-5

Kumar, S., Toshniwal, D. (2016b) "A novel framework to analyze road accident time series data", Journal of Big Data, 3, Article number: 8. https://doi.org/10.1186/s40537-016-0044-5

Kumar, S., Toshniwal, D. (2016c) "Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC)", Journal of Big Data, 3, Article number: 13. https://doi.org/10.1186/s40537-016-0046-3

Montella, A. (2010) "A comparative analysis of hotspot identification methods", Accident Analysis & Prevention, 42(2), pp. 571–581. https://doi.org/10.1016/j.aap.2009.09.025

Polus, A. (1982) "Relationship of overall geometric characteristics to the safety level of rural highways", Traffic Quarterly, 34(4), pp. 555–585. [online] Available at: https://trid.trb.org/view/166981 [Accessed: 24 June 2019]

Qu, X., Meng, Q. (2014) "A note on hotspot identification for urban expressways", Safety Science, 66, pp. 87–91. https://doi.org/10.1016/j.ssci.2014.02.006

Quinlan, J. R. (1993) "C4. 5: programs for machine learning", Morgan Kaufmann Publishers, Inc., San Mateo, CA, USA.