MAEN QASEEM GHADI, Ph.D. Candidate[1]
(Corresponding author)
E-mail: ghadi.maen@mail.bme.hu
ÁRPÁD TÖRÖK, Ph.D.[1]
E-mail: arpad.torok@auto.bme.hu,
[1] Faculty of Transport Engineering and Vehicle Engineering
Budapest University of Technology and Economics
H-1521 Budapest, P.O.B. 91, Hungary

Traffic Infrastructure
Preliminary Communication
Submitted: 2 Mar. 2018
Accepted: 25 Jan. 2019

COMPARISON OF DIFFERENT ROAD SEGMENTATION METHODS

ABSTRACT

In road safety, the process of organizing road infrastructure network data into homogenous entities is called segmentation. Segmenting a road network is considered the first and most important step in developing a safety performance function (SPF). This article aims to study the benefit of a newly developed network segmentation method which is based on the generation of accident groups applying K-means clustering approach. K-means algorithm has been used to identify the structure of homogeneous accident groups. According to the main assumption of the proposed clustering method, the risk of accidents is strongly influenced by the spatial interdependence and traffic attributes of the accidents. The performance of K-means clustering was compared with four other segmentation methods applying constant average annual daily traffic segments, constant length segments, related curvature characteristics and a multivariable method suggested by the Highway Safety Manual (HSM). The SPF was used to evaluate the performance of the five segmentation methods in predicting accident frequency. K-means clustering-based segmentation method has been proved to be more flexible and accurate than other models in identifying homogeneous infrastructure segments with similar safety characteristics.

KEY WORDS

K-means clustering; road segmentation; safety performance function; road accidents; homogeneous segment;

1. INTRODUCTION AND LITERATURE REVIEW

An appropriate safety performance function (SPF) is considered to be one of the basic methods of road safety analysis [1]. The SPF represents a mathematical relationship between accident frequency and other related explanatory variables for different road segments. In most cases, the reliability of the SPF depends fundamentally on the validity of the applied statistical methods and the way data is organized, i.e., clustered into specific homogeneous sets or groups of similar entities. Most often, road segmentation is based on researchers' experiences, methodological decisions or objectives. A variety of popular segmentation methods [2] exists. Black and Thomas (1998) [3], observed a positive level of network autocorrelation of contiguous road segments that have segments with constant lengths. While, trying to search for more homogeneous segments [4], other scholars have introduced other segmentation processes based on different road infrastructure attributes, i.e., speed, number of lanes, average annual daily traffic (AADT). However, some scholars [5, 6] have argued that applying different lengths and start points for segmenting road network can result in different definitions of hazardous locations which in turn affect the stability of results. Koorey (2009) [7] discussed the benefit of applying variable length segments and their effect on locating high-risk road sites. Cafiso et al. (2013) [8] compared the efficiency of different SPFs created from five segmentation approaches, which segmented the road based on geometric and/or traffic related attributes. It was concluded that segmentation methods based on design parameters (i.e., curvature characteristics) are better in developing the SPF than others since the set of high-risk sections provided by them is deemed to be well correlated with the set of locations characterized by high accident density.

Generally, segmenting road sections into homogeneous groups based on too many variables can result in very short average segment lengths [9] which can eventuate in many zero sections. In contrast, increasing the length of the segment would scarify homogeneity. Besides this, most segmentation approaches apply only traffic conditions and attributes regarding road geometrics in identifying homogeneous segments and fail to consider accident data, which can in some cases significantly improve the reliability of the model. Considering that traffic accident data is heterogeneous, in general, on the one hand, traffic accident analysis using a clustering approach can be a useful technique to find hidden relationships and patterns for a large number of accidents or data [10], and on the other hand, clustering can also be an efficient way to generate accident groups according to some similarity

measures in their attributes and spatial distributions. Depaire et al. (2008) [11] used latent class clustering for identifying homogeneous groups of traffic accidents. Luca et al. (2012) [12] applied C-mean clustering to identify accident sets from which subsequently an empirical Bayesian (EB) model was constructed. Ghadi et al. (2018) [13] used K-means clustering techniques to classify accidents into clusters based on the spatial factor, followed by applying the EB method to identify high-risk accident segments. Kumar & Toshniwal (2015) [14] applied K-mode clustering and the association rule of mining to identify the main circumstances associated with accident occurrences. The result revealed different trends in different clusters and helped detect hidden patterns of accidents.

In this article, a road network segmentation method based on accident clustering is being introduced, which classifies accidents and road dataset into homogeneous clusters based on spatial interdependence of accidents, and traffic characteristics and geometric attributes of the road network. To evaluate the performance of the presented method, it is compared with four other segmentation approaches. The attributes applied in the four segmentation approaches are based on the Highway Safety Manual (HSM) procedures, constant AADT segments, constant length segments and segments characterized by curvature. The SPF was used to evaluate the performance of the five segmentation methods in predicting accident frequency.

2. SEGMENTATION APPROACHES

2.1 Segmentation variables

In order to segment a road into homogeneous sections, firstly it has to be decided which attributes of the clustered section should be homogenous. Of course, the more infrastructure-related variables the processes involves, the more attributes of the clustered sections become homogenous. On the other hand, involving many variables in a segmentation process would increase its complexity and reduce average segment length. In accordance with this, there are different ways to segment the road network. The AADT is considered a major variable in road segmentation and its value plays an important role in predicting the number of accidents. Recently, many other variables have started to be used (other than the AADT) in defining homogeneous segments on a road. The main variables included in this study, apart from the AADT, are described as follows:

- Speed limit: assuming the driver will not break the speed limit, it can directly affect accident occurrence, where low-speed vehicles have a higher perception-reaction time to avoid an accident than high-speed vehicles. Also, the speed limit can affect accident severity. For example, a low-speed vehicles accident could result only in property damages (POD), but a high-speed vehicles accident is more likely to include injuries or even fatalities [15].
- Percent of trucks and percent of small vehicles: in this study, small vehicles include a passenger car, a small van and a light truck (under 3.5 t); while trucks include medium and large vehicles and trailers (over 3.5 t).
- Horizontal curve: instead of straight roads with long stopping sight distance, horizontal curves can have an unfavorable effect on stopping sight distance, which can be considered an important risk factor influencing accident probability. Besides this, the speed limits of infrastructure elements characterized by outstanding curvature values could also have a negative effect on the probability of run-off-road accidents.

A detailed description of the segmentation methods is presented in the following two sections.

2.2 Proposed segmentation method (K-means clustering)

Clustering analysis could have several definitions, depending on the specialties of the discussed application field. However, in general, its major objective is to organize a large dataset into a small number of homogeneous groups in which the degree of association between the objects of the same group is maximal. K-means clustering is generally based on the definition of the cluster structure that minimizes a specific error criterion. During the iteration process of the method, each object (e.g. accident) is represented by a geographical location, and each location has different attributes or coordinates. Thus, a good way to measure the affinity between any two points is the distance. The K-means clustering algorithm starts with an initial set of cluster centers chosen randomly for a predetermined number of clusters (k). In the iteration process, each data object is assigned to its nearest center, according to the Euclidean distance. In the next step, cluster centers are recalculated in accordance with *Equation 1* [16]. The iteration stops when no more cluster centers need to be relocated.

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} q_k \tag{1}$$

where μ_k is the mean of clusters k and N_k is the number of objects (accidents) belonging to cluster k.

However, the greater the number of attributes (coordinates) used in measuring the distance, the more homogeneous the object of one group is. While complexity of the identification process is increasing, the average size of homogeneous segments is decreasing. In this analysis, three main attributes were

chosen to obtain a homogeneous cluster: (1) spatial distribution of accidents, (2) average AADT and (3) road geometrics.

The main idea of the applied clustering method is based on the assumption that the spatial convergence of the accidents can frequently be explained by common accident causes [17]. Accordingly, a linear referencing method has been applied where a road has been represented as a line with a starting point. All accidents have been located on the line by measuring their distances from the start point, as presented in *Figure 1*. This single-variable model seems to fit better to the K-means clustering method than the dual-variable model based on geographical coordinates (the longitude and latitude) (presented in *Figure 1*). This approach allows us to avoid clustering accidents which are closely located but have occurred on different roads. *Figure 1* presents a visual explanation of how accident clusters are generated, according to the proposed methodology, to define homogeneous segments based on the spatial distribution and other traffic attributes (i.e., the AADT). For instance, if the search for homogeneous clusters is related to the AADT and accident spatial distribution, then clusters c1, c2, c3 and c4 will be generated (see *Figure 1*). Each of these clusters contains accidents that occur at relatively close distances and have the same AADT conditions.

The application of the K-means clustering method in road segmentation makes it possible to link the length of the segment to the length of the identified cluster. The length of each cluster may be influenced by the number of accidents the cluster contains and their spatial distribution. Therefore, empty sections between clusters with no accident history can be easily separated from the other clusters. On the other hand, in the case of the application of the K-means clustering method, the number of clusters (k) must be determined before the clustering procedure. This factor can cause uncertainty if the analyst has no prior knowledge related to the investigated road. To determine the optimal estimated number of clusters (k_0), it is required to test different k values and analyze the changes of the variances depending on the number of clusters. The adapted method for minimizing the error is often called minimum variance partition [18]. The partitioning method aims to create clusters where the variation within a cluster is minimized. The quality of the clustering method can be measured by the sum of squared errors (SSE) parameter of the distances between each cluster object and its center, by using *Equations 2* and *3*.

$$SSE = \frac{1}{2}\sum_{k=1}^{K} kS_k \tag{2}$$

where,

$$S_k = \frac{1}{k^2}\sum_{K} |x_i - x_j|^2 \tag{3}$$

where: $|x_i - x_j|$ is the absolute distance between the objects (accidents) j and their cluster centers i.

The minimum k_0 value of the SSE error function is obtained when the number of clusters (k) is equal to the number of objects (accidents), since in this case each centroid migrates to an individual object, which leads to each object belonging to a single cluster at the end of the process, and each distance between the cluster centroids and the objects being zero, which is the lowest possible SSE value. To handle the problem of over-fragmentation, an additional methodological step needs to be introduced to the process. The process of segmentation can be characterized by the changes of the SSE value depending on the changes of cluster numbers. In this case, obviously, the velocity of decreasing the SSE is investigated by projecting the changes of the SEE value to the unit change of the number of clusters (if the analyzed functions were continuous, the mentioned operation would result in a differential of the SSE function). In light of the above-mentioned aspects, when the slope of the SSE function curve's tangent changes, an order of magnitude (e.g. when its absolute value is less than 1) seems to be critical. Based on our previously performed experimental investigations, at this point the number of clusters starts to grow increasingly faster and, on the other hand, at this point the generated clusters fit reasonably well to the location of black spots. Accordingly, this point seems appropriate to be applied as a constraint of the number of clusters.

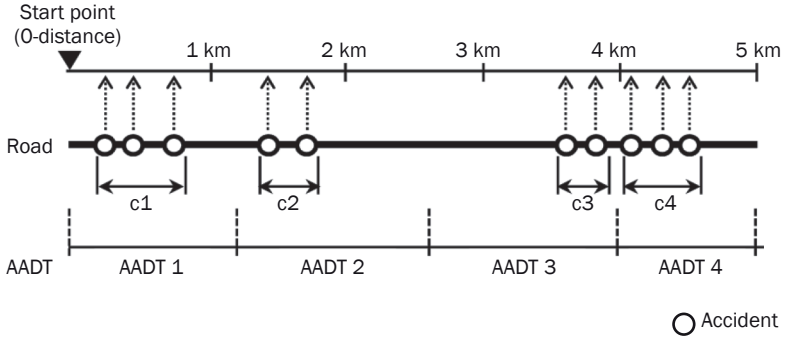

Figure 1 – A visual illustration of K-mean clustering considering spatial and AADT attributes [13]

2.3 Other segmentation approaches

In order to assess the influence of the K-means clustering segmentation method (Seg-1) in organizing road accident data into segments, its effect on the SPF has been compared with other segmentation approaches. Each of the investigated methods differs in the criteria applied to determine homogeneous road segments. The other segmentation methods investigated in the article represent the most used approaches, which are described as follows:

- HSM method (Seg-2): The generated segments are homogenous in the AADT, roadside hazards and presence of curves, as recommended by HSM specifications for highway roads.
- Constant AADT (Seg-3): The generated segments are homogenous in the AADT while other variables do not affect the segmentation structure.
- Constant length (Seg-4): The road network is split in equidistant intervals, while other variables do not affect the segmentation structure. The segmentation length was chosen to be 750 meters. It is an average length that is longer than the recommended minimum value of the HSM but short enough to provide homogeneity, so the characteristics of the road do not change too much within the length.
- Curves (Seg-5): Segments are generated considering the endpoints of the curves and straight lines. Each different curve and straight line is ordered to a different segment. Curves and straight lines are identified according to HSM specifications.

The below presented descriptive statistics table (*Table 1*) represents the five segmentation methods describing the minimum, maximum, mean and standard deviation values related to the different approaches.

It can be noted that all segmentation methods consider segments with zero accidents as well, except the proposed K-means clustering method (Seg-1), which needs at least two accidents to form a segment. Its basic concept assumes that single accidents (spatially and temporally distant from other accidents) or distinct accidents (which do not have any common characteristics with their neighboring accidents) are considered a randomly occurred noise, therefore they are not considered during the segmentation method. It is obvious that this consideration is a heavy simplification; however, it can help to reduce negative effects of over-fragmentation.

3. DATA DESCRIPTION

The used data is from Hungary and originates from expressway numbers 25, 35, 36, 49 and 82. The investigated data has been generated from 2013 to

Table 1 – Descriptive statistics of length and accident numbers per segment

Method of segmentation	Descriptive statistics (per segment)	Minimum	Maximum	Mean	Standard deviation
Seg-1	Number of accidents	2.00	18.00	4.86	3.23
	Length [km]	0.04	10.59	1.68	3.35
Seg-2	Number of accidents	0.00	6.00	0.43	0.81
	Length [km]	0.02	2.61	0.47	1.63
Seg-3	Number of accidents	0.00	19.00	3.29	3.19
	Length [km]	0.77	16.34	4.01	2.17
Seg-4	Number of accidents	0.00	6.00	0.47	0.87
	Length [km]	0.75	0.75	0.75	0.00
Seg-5	Number of accidents	0.00	10.00	0.68	1.25
	Length [km]	0.01	10.63	1.10	1.14

Table 2 – Description of the data used in developing the models

Road Ref.	Accidents (fatal and injury)				AADT		Length [km]
	2013	2014	2015	2016	min	max	
25	57	56	49	46	827	19,771	83.6
35	30	43	38	33	2,985	23,600	82.5
36	26	39	36	31	2,392	19,884	53.8
49	24	37	39	35	2,879	11,556	62.5
82	59	60	68	64	3,890	18,029	76.5
Total	870						358.9

2016, and it includes accident data, traffic characteristics and road design parameters. There are two lanes in both directions on each road. The total length of the investigated network is about 359 km (in one direction). Road segments without intersections have only been considered in the analysis. During the analysis period, 870 fatal and injury accidents occurred. The data has been divided into two parts; the data of the first three years (2013–2015) has been applied in developing the models, while the data of the last year (2016) has been used for checking the performance of the developed models. *Table 2* presents the basic statistical characteristics of the used dataset.

4. MODEL

4.1 Model variables

The selected segmentation methodology significantly determines the output of the following analytical steps (e.g. the results of the SPF), hence, the segmentation model must be chosen in accordance with the following research. Since the main objective of this research is to evaluate safety-related characteristics of the infrastructure network, it is important to choose segmentation variables that affect the safety level of an infrastructure component (e.g. the AADT, speed, curvature). Potential explanatory variables are described as follows:

- AADT: the AADT is considered a major factor in predicting the number of accidents.
- Speed limit: speed is also an important factor of accident risk.
- Percent of trucks and percent of small vehicles: the percentage is measured per total traffic in case of each specific segment.
- Degree of curvature (DOC): in a mathematical sense, the curvature is the reciprocal of the radius. A small curve is easily laid out by using the radius. But, if the radius is as large as a mile or a kilometer, the DOC is more convenient for describing the horizontal curve. The DOC is defined as a central angle to the ends of a chord of agreed length, and it is mathematically calculated as follows:

$$\mathrm{DOC}(degree\ per\ unit\ length[\mathrm{feet}]) = \frac{5279}{\mathrm{Radius\ of\ curve}[\mathrm{feet}]} \quad (4)$$

For each specific segment, curves have been determined by their transition points, their dimensions were measured and the total DOC in degrees was calculated per unit length of the segment.

When a segment does not have a constant AADT, speed or traffic value, its length value (as a linear length-weight) is used to estimate average values.

However, the variables used to determine the segmentation structure should be carefully chosen to ensure their correlation with the dependent variable (i.e., accident frequency) and independence among themselves. For that purpose, correlation analysis and stepwise method have been applied. *Table 3* contains the resulting correlation coefficients (Seg-1).

The strongest correlation has been detected between the AADT, percent of trucks and percent of small vehicles. On these grounds, only one variable related to traffic flow was selected. The same situation was found with the other segmentation methods (their correlation data is not described in this article). The stepwise forward approach was also used to check the significance of inserting or removing different explanatory variables for each SPF every time. Finally, the AADT, speed, and the DOC were selected to be the input variables of the segmentation models.

4.2 Model description

The model development process was implemented in an R software package and the maximum likelihood method was used to estimate the model parameters. The negative binomial regression was assumed to fit well to the number of accidents. Two types of SPF models were developed. The first type includes only the AADT, while the other includes the AADT, speed and the DOC. For all the models, the segment length is included as an offset variable. The general form of the applied SPF is represented by the following *Equation 5* [19]:

$$SPF = \exp\left[\alpha + \left(\sum_n \beta_n \cdot \ln(X_n)\right) + \ln(length)\right] \quad (5)$$

where α is the intercept of the ordinate axis, and β_n is a regression coefficient of the corresponding explanatory variable X_n (i.e., β_1= AADT, β_2= speed and β_3= DOC).

Table 3 – Correlations coefficient parameter of cluster segments

Explanatory variables	Percent of trucks	Speed	DOC	AADT	Percent of small vehicles
Percent of trucks	1.000				
Speed	-0.251	1.000			
DOC	-0.082	0.002	1.000		
AADT	0.466	0.076	0.120	1.000	
Percent of small vehicles	0.805	-0.153	-0.170	-0.314	1.000

To evaluate and compare the efficiency of the developed SPFs, two different statistical methods were applied: the Akaike information criterion (AIC) [20] and the Pearson's correlation coefficient (PCC) [21]. The AIC is a statistical method based on the likelihood value and represents the possibility of over-fitting by describing the trade-off between the goodness of fit and the simplicity of the model (the number of explanatory variables). The AIC value is calculated as follows:

$$AIC = -2 \cdot ML + 2 \cdot p \quad (6)$$

where ML is the maximum log-likelihood of the fitted model and p is the number of model parameters. The first term in the AIC equation measures the bias of fit when the variables' maximum likelihood estimation is used. The second term measures the complexity of the model by actually penalizing the model for using more variables. The goal of the AIC is to choose the best-fitting model with the least complexity. However, the AIC offers an estimate of the relative information lost, therefore, the lower the AIC value the better the model.

Generally, the AIC does not provide information about the absolute quality of the models, it only reflects the relative model quality compared to other models. Thus, Pearson's correlation coefficient (PCC) test is used to measure the efficiency of the developed models in predicting accident data. The PCC measures the linear correlation between any two variables X and Y. It can also be defined as a covariance of the two variables divided by the product of their standard deviations. The PCC has a value between +1 and -1. A value between 0 and 1 implies a positive linear correlation between X and Y. A value between 0 and 1 implies a negative linear correlation between X and Y. Correlations equal to 1 or -1 correspond to a perfect correlation. The PCC has been represented by the following *Equation 7*:

$$PCC = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}} \quad (7)$$

where: x is the observed number of accidents that occurred in a segment, y is the predicted number of accidents, n is the sample size; and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i (\bar{y} \text{ can be expressed similarly}) \quad (8)$$

5. RESULTS AND DISCUSSIONS

The proposed K-means clustering has resulted in 126 segments in the case of the investigated roads regarding the analyzed time period (2013–2015). *Table 4* summarizes the results of the applied K-means clustering.

Data of the segments resulting from the other segmentation methods have been also used to develop new SPF models. The model calibration results of different explanatory variables are shown in *Table 5*.

Most of the variables (i.e., α, β_1, β_2 and β_3) considered in the stepwise procedure in *Table 5* are statistically significant to the 0.05 significance level (95% confidence level). The intercept coefficient α of Seg-1 (*Table 5b*) is significant to the 0.1 level. This can be explained by the relatively small sample size used in a cluster-based segmentation model (Seg-1). Generally, the role of α is not crucial, so a small deficiency does not critically affect the efficiency of the whole model since it rather acts as a calibration factor for the model. Nevertheless, the whole model is statistically significant for all of its parameters.

It can be concluded by evaluating *Table 5* that the models (*Table 5a*) are slightly improved by involving new variables. The AIC values have been the highest in the case of the developed segmentation method (Seg-1) (AIC_{1a}=656 and AIC_{1b}=634 in *Table 5*), while the segmentation method developed by the HSM shows the worst AIC results (AIC_{2a}=2,457 and AIC_{2b}=2,420).

Pearson correlation coefficient (PCC) has been applied to evaluate and compare the developed models by measuring their efficiency in predicting accident data for another year in the future (2016). The PCC has been used to describe the strength of the relationship between prediction and observation (*Table 5*). In accordance with this, *Figure 2* present the scatterplots of relationships between the observed accidents (x-axis) and the predicted accidents (y-axis) in 2016 for all the segmentation models presented in *Table 5*. *Figures 2a–e* represent the models contained in *Table 5a*, while *Figures 2f–j* represent the models contained in

Table 4 – Summary of the results of the application of K-means clustering

Road Ref.	Total number of segments (for a period of 3 years)	Average segment length [km]	Average accident frequency (per segment per year)		
			2013	2014	2015
25	27	4.41	6.5	5.5	5.9
35	27	2.83	4.1	4.0	3.8
36	17	4.72	6.3	8.0	4.6
49	20	4.13	3.7	5.8	6.3
82	35	2.76	5.3	5.2	5.4
Total number of segments = 126					

Table 5 – Values of the model parameters, (p-value), AIC, PCC and over-dispersion (k)

a) Model 1 with one explanatory variable; AADT					
	α (Intercept) [p-value]	β_1 (AADT) [p-value]	k	AIC	PCC [R-square]
Seg-1 (K-means clustering)	- 5.329 [>0.001]	0.664 [>0.001]	1.228	656	0.749 [0.56]
Seg-2 (HSM)	- 7.319 [>0.001]	0.789 [>0.001]	2.151	2457	0.378 [0.23]
Seg-3 (Constant AADT)	- 8.424 [>0.001]	0.912 [>0.001]	1.222	810	0.583 [0.36]
Seg-4 (Constant length)	- 11.858 [>0.001]	1.258 [>0.001]	1.029	1733	0.358 [0.13]
Seg-5 (Curvature)	- 8.646 [>0.001]	0.927 [>0.001]	1.222	1503	0.716 [0.51]

b) Model 2 with three explanatory variables: AADT, speed, DOC							
	α (Intercept) [p-value]	β_1 (AADT) [p-value]	β_2 (Speed) [p-value]	β_3 (DOC) [p-value]	k	AIC	PCC [R-square]
Seg-1 (K-means clustering)	- 3.577 [0.063]	0.674 [>0.001]	- 0.465 [0.041]	1.059 [>0.001]	1.179	634	0.700 [0.49]
Seg-2 (HSM)	- 4.812 [>0.001]	0.770 [>0.001]	- 0.583 [>0.001]	1.199 [0.001]	1.981	2420	0.443 [0.30]
Seg-3 (Constant AADT)	- 9.212 [>0.001]	0.975 [>0.001]	- 0.120 [0.088]	4.408 [>0.001]	1.176	802	0.598 [0.36]
Seg-4 (Constant length)	- 11.585 [>0.001]	1.305 [>0.001]	- 0.202 [0.001]	1.771 [>0.001]	1.100	1708	0.427 [0.14]
Seg-5 (Curvature)	- 8.267 [>0.001]	0.934 [>0.001]	- 0.131 [0.010]	1.505 [>0.001]	1.122	1476	0.688 [0.47]

Table 5b. *Figures 2a–j* also present how linear regression lines (solid lines) fit the data and their R-squared values. The dashed lines indicate a perfect prediction of the accident data. Solid regression lines above and below the dashed lines indicate that model prediction is overestimated or underestimated compared to the actual number of accidents.

It is evident from *Figure 2* that all models tend to underestimate the accident numbers at lower frequencies and overestimate it at higher frequencies, except the cluster regression model (Seg-1) (*Figure 2a*). In the case of Seg-2 and Seg-3, the prediction quality is slightly improved by increasing the number of explanatory variables (*Table 5*: regarding the PCC and R-squared values). However, the differences in the PCC values between model 1 and model 2 (*Table 5*) do not seem to be decisive. Generally, based on the results of *Table 5*, it can be concluded that the capability of the models to predict accident frequencies is between weak and moderate (PCC: 0.358–0.749, R-squared: 0.13–0.56). This can be explained by the relatively small dataset analyzed in this evaluation. The poorly fitted regression lines in Seg-2 to Seg-5 (*Figures 2b–e* and *2g–j*) are caused by the strictly ordered discrete data which makes it difficult to provide a well-fitted regression line. In addition, zero-accident segments can also weaken the models, and this is obvious in the case of Seg-4 which gives the lowest PCC around 0.4 and it almost failed in describing the linear regression line with R-squared values slightly above 10%. Even when the empty segments were eliminated from the model, the results did not show much improvement compared to the proposed model. Despite this, Seg-5 provides an efficient prediction model with relatively reasonable PCC; PCC_{5a}=0.716 and PCC_{5b}=0.688. In general, the developed cluster-based segmentation models (Seg-1) provide the best prediction efficiency, with the highest PCC and R-squared values (PCC_{5a}=0.749, R^2=0.56 and PCC_{5b}=0.700, R^2=0.49) as well as the lowest AIC values (AIC_{1a}=656 and AIC_{1b}=634).

6. CONCLUSION

This paper summarized the benefit of applying K-means clustering to segmenting road accident data. K-means clustering was applied to the identification of the structure of homogeneous segments by spatially defining a group of closely located points (accidents). The distance was represented in terms of a traffic condition, road geometric and a distribution of accidents

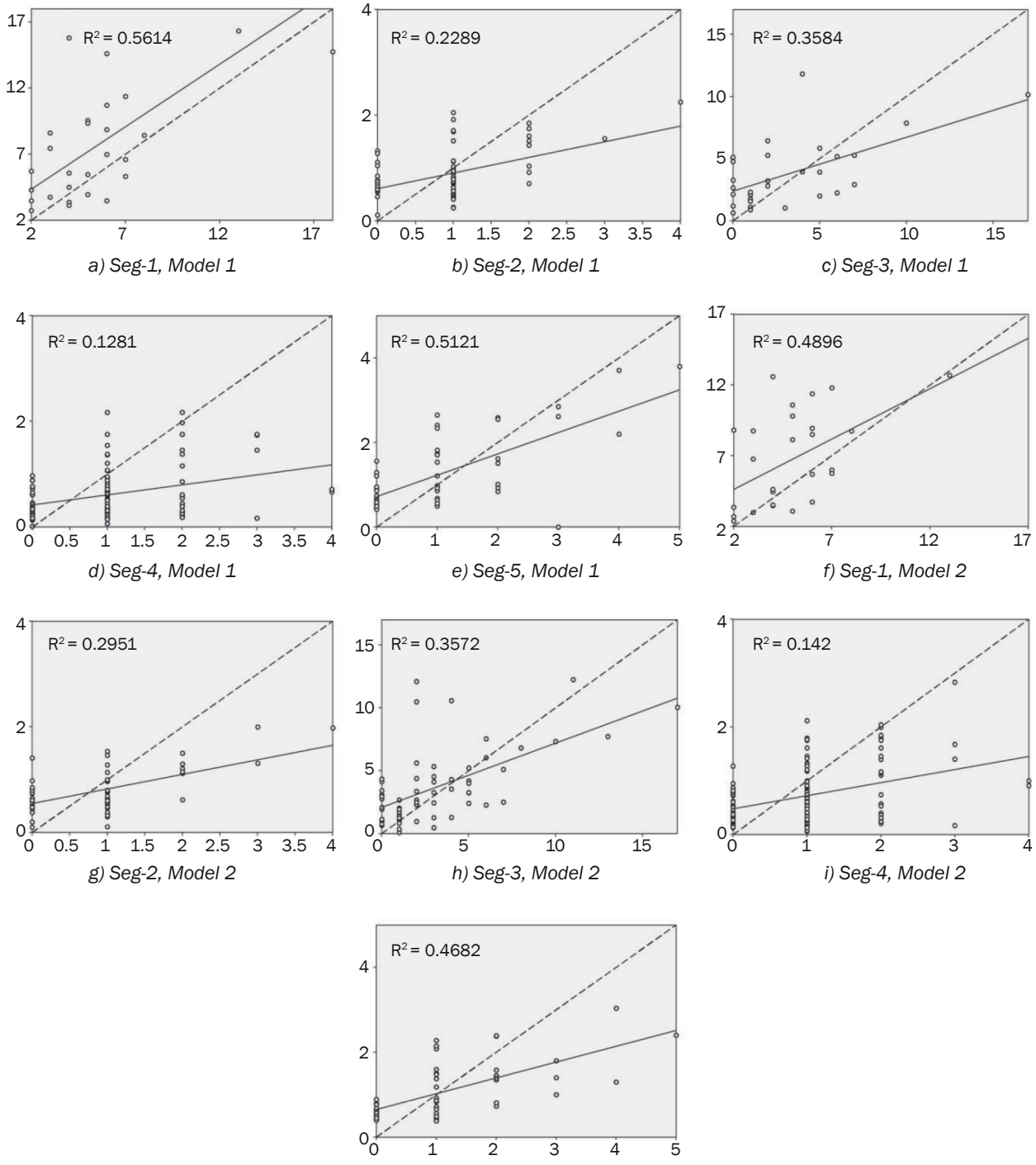

Figure 2 – Graph of the linear correlation between the observed accidents (x-axis) and the predicted accidents (y-axis) of the year (2016)

in space and time. The linear referencing method was used to count accident distributions on the road, from a reference point, in a single attribute.

The performance of K-means clustering was compared with four other segmentation methods. The second segmentation method is based on the specifications of the Highway Safety Manual (HSM), using curvature and the AADT. The third method is based on constant length segments, whilst the fourth method is based on constant AADT segments. And in the case of the fifth method, curvature characteristic has been applied to separating segments. The performance of the five segmentation methods has been analyzed using two evaluation models in the case of the Hungarian highway. To develop SPFs, a negative binomial model has been used. The goodness of fit of the models has been evaluated with the AIC method and the Pearson correlation coefficient (PCC).

The significance of the models, as reported in *Tables 5a–b,* is good. The best goodness of fit results have been obtained by the proposed segmentation models which are based on K-means clustering. This is likely because the segment length in this method is influenced by data availability, quality and other variables to optimize the SPF calibration. However, segmentation approach based on clustering seems to be promising and should be further improved by considering other clustering variables.

MAEN QASEEM GHADI, Ph.D. Candidate[1]
E-mail: ghadi.maen@mail.bme.hu
ÁRPÁD TÖRÖK, Ph.D.[1]
E-mail: arpad.torok@auto.bme.hu,
[1] Faculty of Transport Engineering and Vehicle Engineering
Budapest University of Technology and Economics
H-1521 Budapest, P.O.B. 91, Hungary

A KÖZÚTHÁLÓZAT EGYES SZEGMENTÁCIÓS MÓDSZEREINEK ÖSSZEHASONLÍTÁSA

ABSZTRAKT

A közúti közlekedésbiztonság területén a közúti infrastruktúra hálózati adatainak szakaszokra bontásának folyamatát szegmentálásnak nevezik. Az úthálózat szegmentálása az első és legfontosabb lépés a biztonságteljesítmény-függvények (BTF) kifejlesztésében. A cikk célja az újonnan kifejlesztett hálózati szegmentációs módszer előnyeinek vizsgálata, amely a K-közép klaszterezési megközelítést alkalmazó baleseti csoport generáláson alapul. A homogeny baleseti csoportok azonosítására a K-közép algoritmust használjuk. A javasolt módszer fő koncepciója szerint a baleset-bekövetkezési kockázatot erősen befolyásolja a balesetek térbeli függősége és forgalmi jellemzői. A K-közép osztályozás hatékonyságát négy további szegmentálási módszerrel hasonlítottuk össze. Ezzel összhangban megvizsgáltuk az átlagos napi forgalom alapú szakaszolást, az állandó hosszúságú szakaszolást, a hosszirányú vonalvezetésen alapuló szakaszolást és egy többváltozós módszert igazodva a Highway Safety Manual (HSM) által azonosított módszertani keretekhez. Az öt szegmentációs módszer teljesítményének értékelésére a BTF-et használták a baleseti gyakoriság előrejelzésére. A hasonló biztonsági jellemzőkkel rendelkező homogeny infrastruktúra-szegmensek azonosításakor a K-közép klaszter alapú alapú szegmentálási módszere rugalmasabbnak és pontosabbnak bizonyult, mint a többi modell.

KULCSSZAVAK

K-közép osztályozás; Közúti szegmentálás; biztonságteljesítmény-függvény; Közúti balesetek; Homogén szegmens;

REFERENCES

[1] Federal Highway Administration. Federal Highway Administration. *Safety Analyst Overview.* 2009a [Internet]. 2010 [cited 2010 Feb 16]. Available from: http://www.safetyanalyst.org

[2] Gupta M, Solanki VK, Singh VK. Analysis of Datamining Technique for Traffic Accident Severity Problem: A Review. In: *Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering* [Internet]. ACSIS; 2017. p. 197-9. Available from: https://fedcsis.org/proceedings/rice2017/drp/121.html

[3] Black WR, Thomas I. Accidents on belgium's motorways: a network autocorrelation analysis. *J Transp Geogr.* 1998 Mar;6(1): 23-31. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0031813561&partnerID=tZ0tx3y1

[4] Sadeghi A, Ayati E, Neghab MP. Identification and prioritization of hazardous road locations by segmentation and data envelopment analysis approach. *Promet - Traffic - Traffico.* 2013;25(2): 127-36. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84937346591&partnerID=40&md5=6c1fa7c191213cfa7ae5bee1df78305e

[5] Kwon OH, Park MJ, Yeo H, Chung K. Evaluating the performance of network screening methods for detecting high collision concentration locations on highways. *Accid Anal Prev.* 2013;51: 141-9.

[6] Thomas I. Spatial data aggregation: exploratory analysis of road accidents. *Accid Anal Prev*. 1996;28(2): 251-64.

[7] Koorey G. Road Data Aggregation and Sectioning Considerations for Crash Analysis. *Transp Res Rec J Transp Res Board.* 2009;2103(1):61–8. Available from: doi:10.3141/2103-08

[8] Cafiso S, D'Agostino C, Persaud B. Investigating the influence of segmentation in estimating safety performance functions for roadway sections. *TRB 92nd Annu Meet.* 2013;15.

[9] P. Resende RB. Effect of roadway section length on accident modeling traffic congestion and traffic safety. In: *The 21st Century Conference, II. Chicago.* ASCE; 1997.

[10] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Mag*. 1996; 37-54. Available from: http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230

[11] Depaire B, Wets G, Vanhoof K. Traffic accident segmentation by means of latent class clustering. *Accid Anal Prev.* 2008;40(4): 1257-66.

[12] De Luca M, Mauro R, Lamberti R, Dell'Acqua G. Road Safety Management Using Bayesian and Cluster analysis. *Procedia - Soc Behav Sci.* 2012;54: 1260-9. Available from: http://www.sciencedirect.com/science/article/pii/S1877042812043029

[13] Ghadi M, Török Á, Tánczos K. Integration of Probability and Clustering Based Approaches in the Field of Black Spot Identification. *Period Polytech Civ Eng.* 2018 Oct 19; Available from: https://pp.bme.hu/ci/article/view/11753

[14] Kumar S, Toshniwal D. A data mining framework to analyze road accident data. *J Big Data.* 2015;2(1).

[15] Ghadi M, Török Á. Comparison Different Black Spot Identification Methods. In: *Transportation Research Procedia.* 2017;27: 1105-12.

[16] Everitt BS, Landau S, Leese M, Stahl D. *Cluster Analysis. 5th* Edition [Internet]. Wiley Series in Probability and Statistics; 2011. Available from: http://onlinelibrary.wiley.com/book/10.1002/9780470977811

[17] Flahaut B, Mouchart M, San Martin E, Thomas I. The

local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach. *Accid Anal Prev.* 2003;35(6): 991-1004.

[18] Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat - Theory Methods.* 1974;3(1): 1-27. Available from: doi:abs/10.1080/03610927408827101

[19] American Association of State Highway and Transportation Officials. *Highway Safety Manual.* 1st Edition; 2010.

[20] Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics.* 2001;57(1): 120-5.

[21] Smyth GK. Pearson's Goodness of Fit Statistic as a Score Test Statistic. *Sci Stat A Festschrift Terry Speed.* 2003;40(March): 1-12. Available from: http://www.statsci.org/webguide/smyth/pubs/goodness.pdf