**FACTA UNIVERSITATIS**
Series: **Electronics and Energetics** Vol. 33, Nº 4, December 2020, pp. 655-668
https://doi.org/10.2298/FUEE2004655B

# APPLICATION OF CLUSTER ANALYSIS IN THE BEHAVIOUR OF TRAFFIC PARTICIPANTS RELATING TO THE USE OF SAFETY SYSTEMS AND MOBILE PHONES

**Marija Blagojević, Stefan Šošić**

University of Kragujevac, Faculty of Technical Sciences Čačak, Serbia

**Abstract**. *This paper presents a cluster analysis related to the behavior of traffic participants in relation to the use of safety systems and mobile phones. The data on traffic behavior were downloaded from an open data portal in Serbia. Three types of cluster analysis have been applied: hierarchical clustering, Bayesian Information Criterion (BIC) clustering and model clustering. The obtained results point to the various possibilities of using these three clustering methods in the field of traffic and suggest further research.*

**Key words**: *cluster analysis, traffic accidents, safety systems*

## 1. INTRODUCTION

Along with the development of society, traffic and traffic communication are developing. The basic factor of development of each society is traffic. The level of traffic development is used to measure the level of development of a particular society.

The traffic in Serbia is of paramount importance because of the country's location at the crossroads of the Balkans. In the area of legislation, the following significant regulations have been adopted in Serbia:

1. The Law on Road Traffic Safety ((“The Official Gazette of the Republic of Serbia”, No. 41/2009, 53/2010, 101/2011, 32/2013 – Constitutional Court decision, 55/2014, 96/2015 – other law, 9/2016 – Constitutional Court decision, 24/2018, 41/2018, 41/2018 – other law, 87/2018 and 23/2019) [1]
2. Strategy on waterborne transport development of the Republic of Serbia, 2015 - 2025 [2]

In the era of open data, the data on traffic in Serbia have also got their space. A separate section of the Serbian open data portal [3] is devoted to public safety and it has been the source of data used for cluster analysis in this paper.

Traffic monitoring and analysis play a key role in raising the level of transport of goods and passengers. Statistics and indicators that characterize traffic are numerous and

Received April 24, 2020; received in revised form June 11, 2020
**Corresponding author:** Marija Blagojević
University of Kragujevac, Faculty of Technical Sciences Čačak, Svetog Save 65, 32102 Čačak, Serbia
E-mail: marija.blagojevic@ftn.kg.ac.rs

often their collection and formation of databases is limited by their availability and efficiency of the system itself. The application of modern statistical and mathematical methods in the evaluation of traffic enables a comprehensive analysis that includes a large number of indicators, as well as a large amount of data. The aim of this paper is to analyze the available open databases on traffic in the domain of use of safety systems and mobile phones, which group the data into clusters whose number was obtained by statistical preprocessing. The goal of the research is to group traffic participants according to the environment in which the highest amount of offenses is committed.

The paper is organized as follows: section 2 gives literature review, section 3 is dedicated to data, and section 4 describes methodology of research. Section 5 presents results while section 6 consists of discussion and conclusion remarks.

## 2. LITERATURE REVIEW

Numerous research studies deal with cluster analysis in traffic. The study presented in [4] shows an analysis of data originating from vehicle trajectories obtained by simulation. Two strategies were implemented: "Two platoon clustering strategies for CACC; an ad hoc coordination strategy and a local coordination strategy". The analysis conducted in [5] provides an overview of the use of the spatial clustering method for macro-level traffic crash analysis. The analysis was based on the open source point-of-interest data. These data were downloaded from an open source Web site. Traffic accidents are discrete and non-negative events and parameters that are used require further evaluation in order to determine the correlation so as to identify the distribution of traffic crash frequency. In [6], while analysing the traffic accidents, the authors sought to identify key factors that influence the severity of the accident. The Latent Class Cluster (LCC) was used as a preliminary analysis tool. The incidents that occurred in Granada (Spain) in 2005-2008 were analysed. The clustering technique (in combination with other techniques) was also used in the research presented in [7]. These techniques were used to predict the Collection of Annual Average Daily Traffic that is relevant to a large number of applications. Cluster analysis and regression analysis were used in [8] to create an algorithm which would be used to "estimate the number of traffic accidents and estimate the risk of traffic accidents in a study area". The authors of [9] paid special attention to younger drivers and their lifestyles in order to make a correlation with traffic accidents. By using cluster analysis they defined the groups of users with similar lifestyles. The research presented in [10] shows the application possibilities and efficiency of latent class clustering with the aim to identify homogeneous types of traffic accidents. The motivation for this research stemmed from the fact that traffic accident data are most often heterogeneous. Some authors, like in [11], created the architecture of a dynamic clustering system using Beowulf class clusters and NoW.

If we compare the research conducted in this paper with the studies given above we can observe both similarities and differences in the approach. The basic concept underlying the clustering method is the same in all studies. Some papers also deal with traffic accidents, but in different contexts. The main difference is reflected in the approach to traffic accident analysis related to the use of safety systems and mobile phones.

## 3. DATA

In order to implement any of data mining techniques, first we need to have a data set which will be analyzed. For our research, a data set of the indicators of traffic participants' behavior has been downloaded from the Serbian open data portal. These indicators are indicators of behavior of road users and are indirect indicators of traffic safety in Serbia. Three Excel documents were available on the open data portal and we have directed our research to the indicators of traffic participants' behavior with regard to the use of safety systems and mobile phones. The document shows the ID of the territory to which the indicator relates, year of measurement of the indicator, type of the vehicle in which the traffic participant was observed, type of indicator, value of the indicator on the roads in the settlement, outside settlements and on highways.

The data from Excel, containing 1932 records, were raw data, which cannot be used in that form for making a data frame. Because of that, in order to implement clustering technique, R Studio software [12] has been used. R Studio is an integrated development environment for the R programming language used for statistical data and graphics.

| | ID Policijska uprava | Policijskauprava | Godina | Vozilo | Indikator | % koriscenja U NASELJU | % koriscenja VAN NASELJA | % koriscenja NA AUTOPUTU | % koriscenja UKUPNO | Klasa NASELJE | Klasa VAN NASELJA | Klasa AUTOPUT | Klasa UKUPNO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | SRBIJA | 2014 | Autobus | Mobilnitelefon | 4.000000 | 3.50000 | 4.1 | 3.800000 | 5 | 4 | 5 | 4 |
| 2 | 1 | BEOGRAD | 2014 | Autobus | Mobilnitelefon | 14.000000 | 5.00000 | 7.8 | 9.000000 | 5 | 5 | 5 | 5 |
| 3 | 2 | KRAGUJEVAC | 2014 | Autobus | Mobilnitelefon | 5.900000 | 17.60000 | 6.3 | 9.000000 | 5 | 5 | 5 | 5 |
| 4 | 3 | JAGODINA | 2014 | Autobus | Mobilnitelefon | 0.000000 | 0.00000 | 0.0 | 0.000000 | 1 | 1 | 1 | 1 |
| 5 | 4 | Nis | 2014 | Autobus | Mobilnitelefon | 2.000000 | 5.30000 | 5.6 | 3.200000 | 3 | 5 | 5 | 4 |

**Fig. 1** Raw data from Excel

The columns representing vehicles and indicators must be serialized first. Data serialization is the process of converting structured data to a format that allows sharing or storage of the data in a form that allows recovery of its original structure. Serialization has been done with "keras" library for R Studio, which has tokenizer method used to accomplish the process.

| | ID Policijska uprava | Godina | Vozilo | Indikator | % koriscenja U NASELJU | % koriscenja VAN NASELJA | % koriscenja NA AUTOPUTU | % koriscenja UKUPNO | Klasa NASELJE | Klasa VAN NASELJA | Klasa AUTOPUT | Klasa UKUPNO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | 2014 | 5 | 3 | 4.000000 | 3.50000 | 4.1 | 3.800000 | 5 | 4 | 5 | 4 |
| 2 | 1 | 2014 | 5 | 3 | 14.000000 | 5.00000 | 7.8 | 9.000000 | 5 | 5 | 5 | 5 |
| 3 | 2 | 2014 | 5 | 3 | 5.900000 | 17.60000 | 6.3 | 9.000000 | 5 | 5 | 5 | 5 |
| 4 | 3 | 2014 | 5 | 3 | 0.000000 | 0.00000 | 0.0 | 0.000000 | 1 | 1 | 1 | 1 |
| 5 | 4 | 2014 | 5 | 3 | 2.000000 | 5.30000 | 5.6 | 3.200000 | 3 | 5 | 5 | 4 |

**Fig. 2** Data frame after serialization

After the serialization and before creating a data frame, scaling of data has been done. In R program language, there is a scale function which places continuous variables on unit scale by subtracting the mean of the variable and dividing the result by the variables standard deviation. As a result, the transformed values have the same relationship but standard deviation 1. Dataset used for research contains values which vary in range and are represented in different units. Clustering algorithms used in this research use Eucledian distance between two data points in their computations. If scaling is not done,

it can affect results, because of using mixed units and ranges in computations. The results would vary between different units. To bypass that issue, data frame needs to be scaled to the same level.

| | ID Policijska uprava | Godina | Vozilo | Indikator | % koriscenja U NASELJU | % koriscenja VAN NASELJA | % koriscenja NA AUTOPUTU | % koriscenja UKUPNO | Klasa NASELJE | Klasa VAN NASELJA | Klasa AUTOPUT | Klasa UKUPNO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.00000000 | 0.25 | 0.8 | 0.2222222 | 0.04000000 | 0.0350000 | 0.041 | 0.03762376 | 1.0 | 0.8 | 1.0 | 0.8 |
| 2 | 0.00000000 | 0.25 | 0.8 | 0.2222222 | 0.14000000 | 0.0500000 | 0.078 | 0.08910891 | 1.0 | 1.0 | 1.0 | 1.0 |
| 3 | 0.01010101 | 0.25 | 0.8 | 0.2222222 | 0.05900000 | 0.1760000 | 0.063 | 0.08910891 | 1.0 | 1.0 | 1.0 | 1.0 |
| 4 | 0.02020202 | 0.25 | 0.8 | 0.2222222 | 0.00000000 | 0.0000000 | 0.000 | 0.00000000 | 0.2 | 0.2 | 0.2 | 0.2 |
| 5 | 0.03030303 | 0.25 | 0.8 | 0.2222222 | 0.02000000 | 0.0530000 | 0.056 | 0.03168317 | 0.6 | 1.0 | 1.0 | 0.8 |

**Fig. 3** Data frame after scaling

Table 1 presents all variables with their type.

**Table 1** Variables and their types

| | Variable name | Type |
|---|---|---|
| 1 | ID | Numerical |
| 2 | Year | Numerical |
| 3 | Vehicle | Categorical |
| 4 | Indicator | Categorical |
| 5 | % of using in colony | Numerical |
| 6 | % of using outside the colony | Numerical |
| 7 | % of using on highway | Numerical |
| 8 | % of using in total | Numerical |
| 9 | Class colony | Numerical |
| 10 | Class outside the colony | Numerical |
| 11 | Class highway | Numerical |
| 12 | Class total | Numerical |

## 4. METHODOLOGY

Data mining technique which was applied to solve the research problem was clustering. To understand clustering technique, the term cluster needs to be explained first. Cluster refers to a group of objects that belong to the same class. That means that similar objects are grouped in one cluster and dissimilar objects in another. Based on that, a cluster of data objects can be presented as one group. The process of making a group of data objects by similarity is called Clustering. According to [13] clustering is unsupervised classification technique in pattern analysis. The main advantage of this technique is that it is adaptable to changes and it helps to separate useful characteristics that distinguish variety of groups. In our study, traffic participants have been clustered according to the most common location where they committed violations - in settlement, outside settlements or on highways.

Important thing to consider when choosing a clustering algorithm is whether the algorithm scales to dataset which is used for clustering. Algorithm should have good performance and efficiency since dataset which is used for clustering can contain huge

amount of data. Complexity notation is used for determining efficiency of the algorithm. Algorithms which have $O(n^2)$ complexity notation are not practical and non-efficient. In proposed research only algorithms with complexity notation lower or equal than $O(n^2)$ are used. For example, k-means algorithm, which is explained in continuation of the paper, has a complexity notation of O(n). Complexity notation O(n), means that the algorithm scales linearly with n.

One of the simpler learning algorithms that solve the clustering problem is K-means and it can be applied to these results. The idea is to define k centers for each cluster. Hofmeyr in [14] noticed that „clusters are associated with compact collections of points arising around a set of cluster centroids". K-means clustering computes the distance between samples and forms clusters by representing a gene as a vector of expression values according to Yang et al. [15]. Different location of k centers gives different result. After that, a loop is created, and as a result k centers change location step by step until they stop moving. This algorithm has a goal of minimizing objective function square error:

$$J(V) = \sum_{i=1}^{C} \sum_{j=1}^{Ci} (||Xi - Vj||)^2 \quad (1)$$

where the function parameters represent the following:

$||Xi – Vj||$ - Euclidean distance between Xi and Vj

Ci – number of data points in cluster on i position

C – number of centers in cluster

When the first cluster center is calculated, the next one must be recalculated using the following function:

$$Vi = (\frac{1}{Ci}) \sum_{j=1}^{Ci} Xi \quad (2)$$

With this function, the distance between each data point and new obtained cluster center is recalculated. If no data point has been reassigned then the process stops.

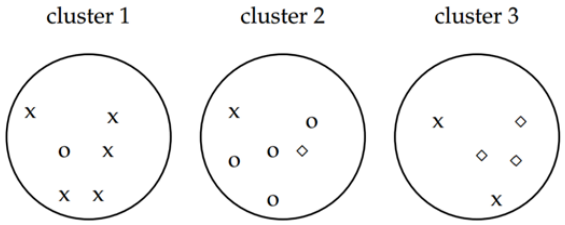


**Fig. 4** Clusters after finishing the process of calculating cluster centers

Clustering techniques which have been applied to the data set of traffic participants' behavior indicators in Serbia are:

- Hierarchical clustering

Hierarchical clustering can be divided into two types of hierarchical cluster analysis strategies, agglomerative and divisive. Hierarchical agglomerative clustering (HAC), also known as bottom-up approach, is more informative than the unstructured set of clusters returned by flat clustering. Algorithms used for HAC treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters

until all clusters have been merged into a single cluster that contains all data. On the other hand, divisive clustering, known also as top-down approach requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton cluster.
Based on the data provided, Euclidean distance must be calculated:
D(Xi,Xj)
Xi – Xj represents the basic distance between any two elements of X, and the minimum distance for defining the sub-set distance:

$$\Delta(Xi,Xj)=\min(X)\in X,Y\in Xj\ D(x,y) \tag{3}$$

According to Abbas [16] hierarchical clustering algorithm „combine or divide existing groups, creating a hierarchical structure that reflects the order in which groups are merged or divided".
In contrast to hierarchical clustering, there is also divisive hierarchical clustering that starts from the root containing all the data-set X, and splits this root node into two children nodes containing respectively X1 and X2 (so that X = X1 ∪ X2 and X1 ∩ X2 = ∅), and so on recursively until we reach the leaves that store the data elements in singletons. The divisive method, according to Wei et al. [17], has a top down style "in which the data objects are initially treated as a unified cluster that is gradually split until the desired number of clusters is obtained".

- Clustering based on Bayesian Information Criterion (BIC), proposed by

Schwarz [18]. According to [19] this model supposes choosing one among a set of candidate models M = $M_1$, $M_2$…$M_m$ to represent a given data set D = $D_1$, $D_2$….$D_N$. BIC of model Mi as:

$$BIC(Mi) = \log P\ (D_1,D_2\ldots D_N \mid Mi)\text{-}1/2d_i \log N \tag{4}$$

where, $d_i$ is the number of independent parameters in model $M_i$ and $P(D_1,D_2,\ldots,D_t|M_i)$ is the maximized likelihood for the model.

- Model based clustering

McLachlan and Peel [20] and Fraley and Raftery [21] gave reviews of the area of model based clustering.
Different clustering algorithms have different objective functions, but the general idea is to minimize the distance between the objects in the same cluster while maximizing the distance between the objects in different clusters. Minimization of the intra-cluster distance can also be viewed as the minimization of the distance between each data Xi and the cluster means Cj. Given a set of clusters, Cj's the expected SSE can be calculated as follows:

$$E\left(\sum_{j=1}^{k}\sum_{j\in C}\left|\left|Cj-Xi\right|\right|^2\right)=\sum_{j=1}^{k}\sum_{i\in Cj}\int||Cj-\ Xj||^2\ f(Xi)dXi \tag{5}$$

where || . || is a distance metric between a data point $x_i$ and a cluster means $c_j$. Cluster means are given by:

$$Cj=E\left(\frac{1}{|Cj|}\sum_{i\in Ci}X\right)=\frac{1}{|Cj|}\sum_{i\in Cj}\int Xi\ f(Xi)dXi \tag{6}$$

For all clustering techniques mentioned above, the process of serialization and normalization of the data frame must be done before applying any of the clustering

algorithms. K-means can be applied to the resulting data frame. It is a simple learning algorithm which has already been mentioned and described in the previous section of the paper.

## 5. RESULTS AND DISCUSSION

The research shows the results of the analysis with different clustering methods. The data set mentioned above was used for the research but with different clustering algorithms. The aim of the research was to group traffic participants according to the environment in which the highest amount of offences was committed.

As it can be seen from the command above, we have taken only the percentages of the offences committed in each environment. After applying the learning algorithm, the clusters can be represented by plotting in R Studio. A dendrogram given in Fig 5 is the result.

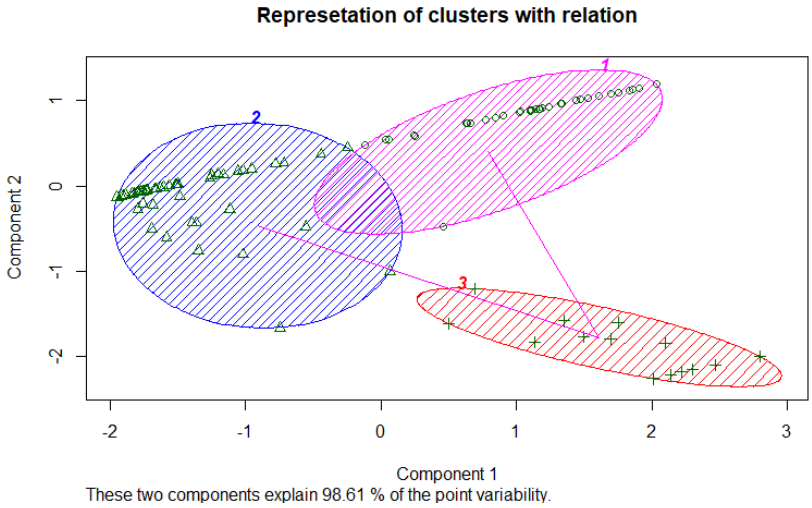


**Fig. 5** Representation of clusters with relation between them

### 5.1. Hierarchical clustering

In proposed research divisive hierarchical clustering was used due the fact that it is more efficient by having lower complexity notation of O ($n^2$). Also, it is more accurate, agglomerative clustering makes decisions by considering the local patterns without considering the global distribution of data. Those early decisions cannot be reversed and that affects result given by hierarchical agglomerative clustering. There can be several closest pairs of subsets, but we have chosen only one pair at each iteration, after which the iteration process is repeated from the beginning. In other words, we have applied a permutation on the elements of X and re-run the algorithm. For numerical data, we can slightly modify the initial data set by adding some small random noise drawn uniformly in (0,1) to bypass this problem. One disadvantage of complete linkage is that it is very sensitive to outliers (that is, artifact data that should have been removed beforehand when possible — the cleaning stage of data sets) (Fig. 6).

Dendrogram has colored cuts (Fig. 6). Each cut represents traffic participants' behavior in different traffic areas. At a given height a flat clustering is obtained. The cut path does not need to be at a constant height. The dendrogram allows one to obtain many flat partitions. Here, three different cuts are shown at a constant height, h = 3. Hierarchical clustering is tightly linked to a class of distances called the class of ultrametrics. A distance is said to be an ultrametric if it is a metric and if it satisfies the following:

$$D(x,y) \leq maxz\ (D(x,z), D(z,y) \tag{7}$$

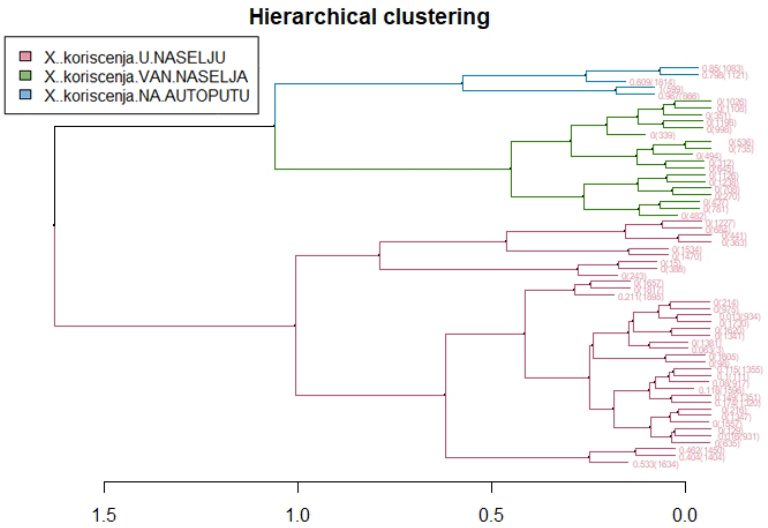


**Fig. 1** Dendrogram representing hierarchical clustering by rows

The same technique can be represented by a dendrogram with clustering by rows down with values (Fig 7).

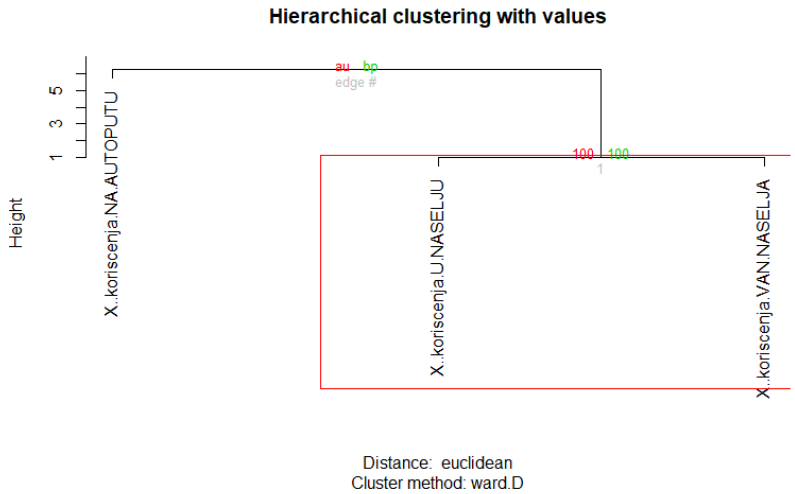


**Fig. 2** Dendrogram representing hierarchical clustering by rows down with values

### 5.2. Clustering based on Bayesian information criterion (BIC)

Bayes factors, approximated by the Bayesian information criterion (BIC), have been successfully applied to the problem of determining the number of components in a model and for deciding which among the three partitions most closely matches the data for a given model.

Partitions are determined by a combination of hierarchical clustering and the expectation-maximization (EM) algorithm. The EM algorithm is an effective approach for performing maximum likelihood estimation in the presence of latent variables. It does this by estimating the values for the latent variables, then optimizing the model and repeating these two steps until convergence. As such, it represents appropriate approach to the use in Bayesian information criterion (BIC) for estimating the parameters of the distributions.

This approach can give much better results than the existing methods. Moreover, the EM result also provides a measure of uncertainty. The model based classification is able to match the traffic classification of a traffic offences data set much more closely than the standard k-means, in the absence of any training data.

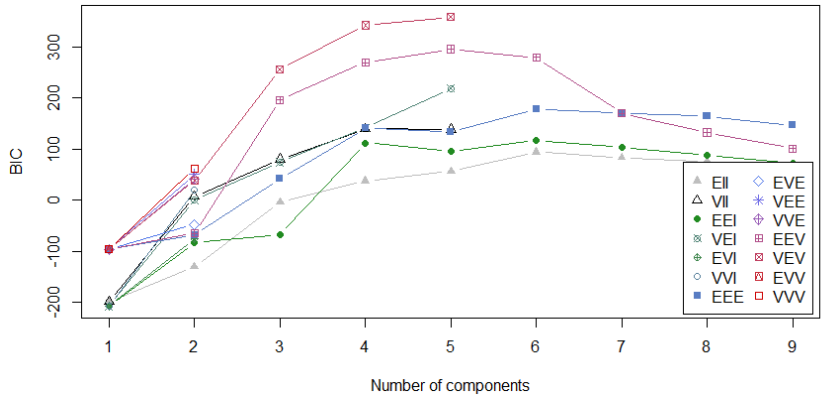


**Fig. 3** Plot representing the Bayesian information criterion (BIC)

The plot shows the Bayesian information criterion (BIC) for the model-based methods applied to the traffic offences data. The first local maximum occurs for the unconstrained model with three clusters.

### 5.3. Clustering based on model

In this section we will illustrate the model-based approach to clustering using a three-dimensional data set involving 1932 observations used for traffic offences in different environments in Serbia.

The plot above (Fig. 9) depicts the uncertainty of the classification produced by the best model (unconstrained, three clusters) indicated by the BIC.

Fig. 10 presents a plot showing the traffic offences classification, which partitions the data into three groups. The variables have the following meanings: in the settlement, outside settlements and on highways. The clusters are overlapping and are far from

spherical in shape. As a result, many clustering procedures would not work well for this application. For example, Figure 3 shows the (1, 3) projection of three-cluster classifications obtained by the single-link (nearest-neighbor) method, standard k-means and the model-based method for an unconstrained Gaussian mixture.

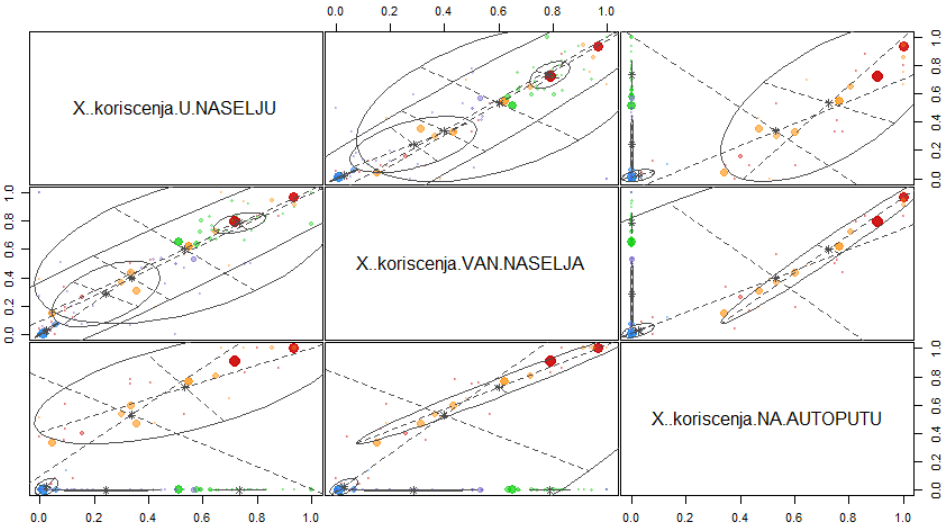


**Fig. 4** Model based on uncertainty clustering

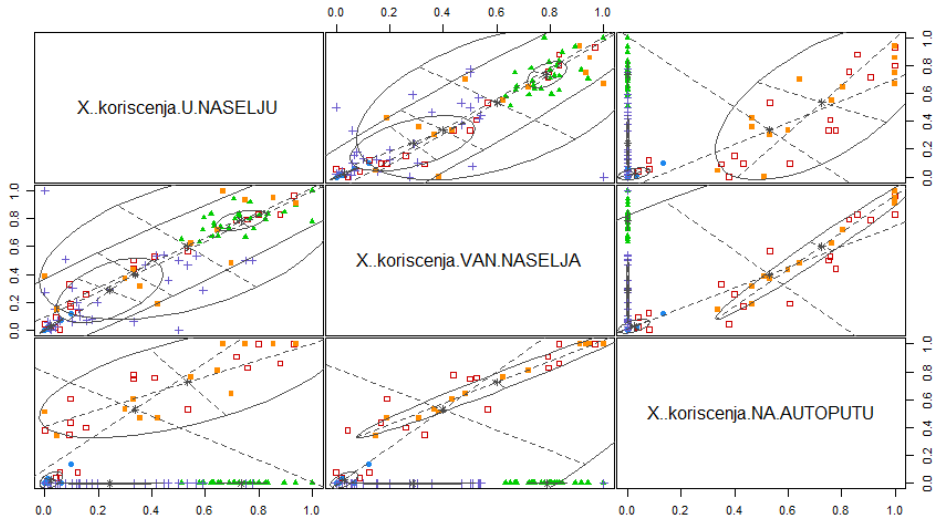


**Fig. 5** Model based on classification clustering

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 5 components:

 log-likelihood   n df      BIC      ICL
       333.7376 100 41 478.6632 474.204

Clustering table:
 1  2  3  4  5
 4 23 41 15 17
```

**Fig. 6** Gaussian finite mixture model fitted by EM algorithm

Fig. 11 shows reciprocal condition estimates for six different Gaussian mixture models for up to nine clusters. It should also be clear that EM started from the partitions obtained by hierarchical clustering and should not be repeated for the following clusters once a traffic offence is encountered.

The assumption of three classes is artificial for the single link and k-means, while for the model-based method the BIC has been used to determine the number of groups. Nearest-neighbor discrimination assigns a data point to the same group as the point in the training set nearest to it. The first local maximum occurs for the unconstrained model with three clusters. For the initial values in EM, *Zik* was used given by the equation for the discrete classification from agglomerative hierarchical clustering for the unconstrained model (λk Dk Ak DT k) in all cases, leaving the model selection to the EM phase (8).

$$Zik = \begin{cases} 1 - & if\ xi\ belongs\ to\ group\ k \\ 0 - & otherwise \end{cases} \tag{8}$$

Clustering algorithm DBSCAN belongs to density-based clustering technique, where the word density refers to the spatial disposition of data points that are dense when forming a group: two data points are in the same cluster if their distance is smaller than the threshold. The result of the DBSCAN algorithm is a set of clusters along their additional merge points (Fig. 12.)

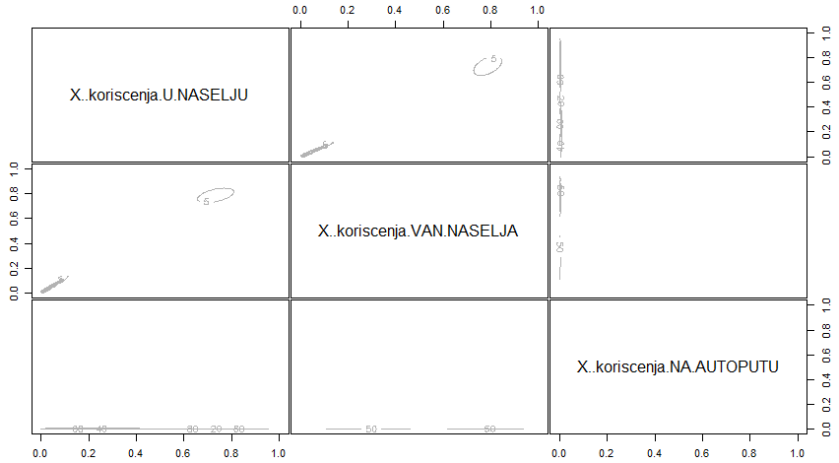


**Fig. 7** Model based on density clustering

Moreover, the number of clusters is not fixed a priori, but it is estimated as a feature of the partition of the observations. To summarize, our model is based on the weakness of the natural clustering rule of species sampling mixture models of parametric densities, by which we mean that two observations Xi and Xj are in the same cluster if, and only if, the latent parameters θi and θj are equal.

## 6. DISCUSSION/CONCLUSIONS

Results represent area where most of traffic offenses are done. This means that special attention should be paid to these areas. All algorithms have the same pattern and it is related to the calculation of the distance between the samples. The advantage could be the complexity of the algorithm, which at the same time can be a disadvantage if the complexity is too large (it supposes longer execution of clustering calculations). Discussion based on presented results could be drawn regarding used data, variables and comparison of used algorithms:

**Data-**The data are collected from the open data portal, so such "raw" data cannot be used directly. Before using, data must be processed with techniques like serialization and scaling. For research purposes in this area, more frequent data updates are desirable. A special contribution would be an open API that would display selected data in real time.

**Variables-**In the chosen data set some variables are more important than others. Variables like percentage of use in specific traffic area are more important than vehicle, type of offense indicator and year of data collection. Clustering results are also related to areas with most and least traffic offenses. Key factor which determine result is percentage of use in specific traffic area. For example, regardless of vehicle in which offense is done it could be focus mainly on area where traffic offense is done.

**Comparison:** Algorithms could be compared regarding complexity:

- K-means – $O(n^2)$
- Hierarchical agglomerative clustering (HAC) – in start $O(n^3)$ is time consumption and demands $O(n^2)$ of memory. In proposed research time consumption is reduced - $O(n^2 \log n)$
- EM – O(n*k*i) where i is the number of iterations (which could be infinite, but it could be set to 0 if there is no need for single iteration)
- BIC – O(i* log n) - where i is the number of parameters. It penalizes the complexity of the model where complexity refers to the number of parameters in model.

Bearing in mind the obtained results several conclusions can be drawn:

- Cluster analysis could be successfully implemented in solving the problems related to traffic accidents and all mentioned techniques of clustering have the advatages and drawbacks;
- There is a significant correlation between the behavior of traffic participants and the use of safety systems and mobile phones
- Each of the types of cluster analysis used is significant and complementary to the cluster information.

In practice, results gained by clustering can be used to determine new road safety laws or changing existing ones. By determining area where most traffic offenses are done, experts can focus on road safety laws only for those area. Focus only on those areas can archive better quality road safety law which can reduce number of traffic offenses.

Proposed research had opened the possibilities for further improvement by clustering in different ways. Clustering could be used to determine area in which most of traffic offenses are done and then cluster again data frame only for that specific area. The proposed improvement could determine details like in which vehicle most traffic offenses are done and which type of traffic offense is mostly done. By gathering those details, experts for creating road safety laws must not only focus on area on which most traffic offenses are done, but also on specific vehicle and type of offenses. That will result more specific laws with focus on most problematic safety factors. Newly created laws will for sure reduce of traffic offenses done. Traffic safety equipment can be also putted to prevent some of offenses.

Future work is concerned with exploring the possibility of using neural networks in order to predict the behavior of traffic participants based on the available open data.

**Acknowledgement**: *This study was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, and these results are parts of the Grant No. 451-03-68/2020-14/200132 with University of Kragujevac - Faculty of Technical Sciences Čačak.*

## REFERENCES

[1] The Law on Road Traffic Safety ((“The Official Gazette of the Republic of Serbia”, No. 41/2009, 53/2010, 101/2011, 32/2013 – Constitutional Court decision, 55/2014, 96/2015 – other law, 9/2016 – Constitutional Court decision, 24/2018, 41/2018, 41/2018 – other law, 87/2018 and 23/2019)

[2] Strategy on waterborne transport development of the Republic of Serbia, 2015 – 2025, available at: http://aler.rs/files/STRATEGIJA_razvoja_vodnog_saobracaja_Republike_Srbije_od_2015_do_2020_godine_Sl_gl_Rs_br_3_2015.pdf, Last access March 23rd 2020.

[3] Open data portal, https://data.gov.rs/sr/, Last access March 23rd 2020.

[4] Z. Zhong, E. Lee, M. Nejad and J. Lee, “Influence of CAV clustering strategies on mixed traffic flow characteristics: An analysis of vehicle trajectory data”, *Transportation Research Part C: Emerging Technologies*, vol. 115, June 2020, 102611

[5] R. Jia, A. Khadka and I. Kim, “Traffic crash analysis with point-of-interest spatial clustering”, *Accident Analysis & Prevention*, vol 121, pp. 223-230. December 2018.

[6] J. Ona, G. Lopez, R. Mujalli and F. Calvo, “Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks”, *Accident Analysis & Prevention*, vol. 51, pp. 1–10. March 2013.

[7] A. Sfyridis and P. Agnolucci, “Annual average daily traffic estimation in England and Wales: An application of clustering and regression modeling”, *Journal of Transport Geography*, vol. 83, 102658, February 2020.

[8] K. Ng, W. Hung and W. Wong, “An algorithm for assessing the risk of traffic accident”, *Journal of Safety Research*, vol. 33, pp. 387–410. October 2002.

[9] N. Gregersen and H. Berg, “Lifestyle and accidents among young drivers”, *Accident Analysis & Prevention*, vol. 26, pp. 297-303. June 1994.

[10] B. Depaire, G.Wets, K. Vanhoof, “Traffic accident segmentation by means of latent class clustering”, *Accident Analysis & Prevention*, vol. 40, pp. 1257–1266. July 2008.

[11] D. Kehagias, M. Grivas, and G. Pantziou, “Using a Hybrid platform for Cluster, NoW and GRID computing”, *Facta Univ. Ser.: Elec. Energ.*, vol. 18, No. 2, pp. 205-218. August 2005.

[12] R Studio, retrieved from: https://rstudio.com/.

[13] A.K. Jain, M.N. Murty, P.J. Flynn, “Data clustering: A review”, *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[14] W. Yang, H. Long, L. Ma, H. Sun, “Research on Clustering Method Based on Weighted Distance Density and K-Means”, *Procedia Computer Science*, vol. 166, pp. 507–511, 2020.

[15] D. Hofmeyr, “Degrees of freedom and model selection for k-means clustering”, Computational Statistics & Data Analysis, vol. 149, 2020.

[16] O.A. Abbas, “Comparisons between data clustering algorithms”, *The International Arab journal of information technology*, vol. 5, no. 3, pp. 320–325. July 2008.

[17] W. Wei, Liang. J, X. Guo, P. Song, Y. Sun, Hierarchical division clustering framework for categorical data, *Neurocomputing*, vol. 341, pp. 118–134, May 2019.
[18] G. Schwarz, "Estimation the Dimension of a Model", *The Annals of Statistics*, vol.6, pp. 461-464, 1978.
[19] B. Zhou, J. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian information criterion", In Proceedings of the 6th International conference on spoken language processing, Bejing, China, 2000.
[20] G. McLachlan, D. Peel, "Robust Cluster Analysis via Mix- tures of Multivariate t-Distributions," *Lecture Notes in Computer Science*, vol. 1451, pp. 658–666, 1998.
[21] C. Fraley, A. Raftery, "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Asso- ciation*, vol. 97, pp. 611–631, 2002.