Modelling the effect of traffic regimes on safety of urban arterials: the case study of Athens

Abstract

This study aims to divide traffic into meaningful clusters (regimes) and to investigate their impact on accident likelihood and accident severity. Furthermore, the likelihood of Powered-Two-Wheelers (PTWs) involvement in an accident is examined. To achieve the aims of the study, traffic and accident data for the period 2006-2011 from two major arterials in Athens were collected and processed. Firstly, a finite mixture cluster analysis was implemented to classify traffic into clusters. Afterwards, discriminant analysis was carried out in order to correctly assign new cases to the existing regimes by using a training and a testing set. Lastly, Bayesian logistic regression models were developed to investigate the impact of traffic regimes on accident likelihood and severity. The findings of this study suggest that urban traffic can be divided into different regimes by using average traffic occupancy and its standard deviation, measured by nearby upstream and downstream loop detectors. The results revealed potential hazardous traffic conditions, which are discussed in the paper. In general, high occupancy values increase accident likelihood but tend to lead to slight accidents, while PTWs are more likely to be involved in an accident when traffic occupancy is high. Transitions from particularly high to low occupancy also increase accident likelihood.

Keywords: Real-time traffic; regime; urban arterials; cluster analysis; Bayesian logistic regression; accidents

1. Introduction

The effect of traffic characteristics on road safety has been investigated for many years. During the past decade, increased attention has been given to developing relationships between real-time traffic characteristics and road safety [1,2,3,4,5,6,7,8,9,10]. The great majority of studies utilize data from freeways, whilst there are some studies which investigate accident likelihood in urban expressways [11].

Although great effort has been made from researchers and practitioners, the impact of traffic states on traffic safety has not been deeply explored although a number of studies have suggested the division of traffic into regimes [12,13,14].

More recent efforts have been found in literature. Abdel-Aty et al. [15] divided freeway traffic flow in high and low speed states and then examined severity and mechanism of multi-vehicle accident occurrence under these two different states, finding different results. Golob et al. [16,17] investigated the safety impact of traffic by dividing traffic flow into different traffic states (traffic regimes) by means of cluster analysis on the basis of traffic flow data collected from the nearest loop detector station from crash locations. The authors attempted to associate traffic regimes with accident type.

A recent study by Xu et al. [18] demonstrates an effort by applying this methodology. The authors emphasized on the need to divide traffic in states and explore their effect on safety, due to the fact that different traffic states may have different influence on the risk of an accident. More specifically, the authors utilized traffic occupancy measured from nearby loop detectors and classified traffic flow into traffic states. Then, each traffic state was associated with a certain safety level. Moreover, it was found that the impact of traffic flow parameters on crash risk is not the same across different traffic flow state.

Yeo et al. [19] proposed a methodology to investigate the relationship between traffic states and crash involvements on freeways. The authors defined the traffic states (free flow, back of queue, bottleneck front, congestion) according to their distinctive patterns and attempted to model the crash involvement rate for each traffic state. It was concluded that crash involvement rate in free flow state is approximately 5 times lower than in other traffic states.

The literature review revealed that a major limitation is data availability, due to the fact that real-time data mainly regard freeways and not major urban arterials. Studies using real-time traffic data to investigate accident severity are relatively few [20,9,10], while only a few studies investigate both accident likelihood and severity [7]. Moreover, European countries are rarely considered, as only one study was found that explored safety of a motorway in Belgium [21]. Although there is a lot of studies investigating Powered-Two-Wheeler (PTW) accident risk

(22,23,24,25), to the best of our knowledge no studies linking PTW accident risk with real-time traffic data were found.

Ensuring safety in major urban roads holds high priority. Consequently the primary objective of this study is to divide urban traffic flow into different regimes and to investigate their effect on accident likelihood and severity. Furthermore, PTW accident risk (involvement of a PTW in an accident) is also explored for the first time using this approach.

The remainder of the paper is organized as follows. Firstly, the proposed methodology is demonstrated (finite mixture cluster analysis, discriminant analysis, Bayesian logistic regression). Then the data description and preparation are provided. Next, the application of the models is explained and the results are presented and discussed. The final section provide the conclusions.

2. Methodology

The statistical methods that applied to achieve the aims of this chapter, are described in the following subsections. Expectation Maximization clustering (EM) (or Finite Mixture) was used to classify traffic into different regimes. In addition, Bayesian logistic regression models were applied in order to correlate traffic regimes with traffic safety.

2.1. Finite mixture cluster analysis

Cluster analysis is a widely used method for grouping observations on the basis of similar data structure. In this study, finite mixture cluster analysis is used to identify homogenous groups of traffic conditions, which can be called “regimes”.

A Gaussian finite mixture model-Based clustering (covariance parameterization and number of cluster selected via the Bayesian Information Criterion) was followed. The models were fitted by Expectation-Maximization algorithm. Fraley et al. [26] and Fraley and Raftery [27] provide a detailed description of Normal Mixture Modeling. All following equations appear in Fraley et al. [26].

A normal or Gaussian mixture model is assumed:

$$\prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \varphi_k (x_i | \mu_k, \Sigma_k), \qquad \text{(Eq. 1)}$$

where x are the data, G is the number of components, τ_k is the probability that a case belongs to the k^{th} component ($\tau_k \geq 0;\ \Sigma_k^{-1}\tau_k = 1$) and

$$\varphi_k(x|\mu_k,\Sigma_k) = (2\pi)^{-\frac{p}{2}}|\Sigma_k|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right\}. \text{(Eq. 2)}$$

The cluster is ellipsoidal, centered at the means μ_k. Their other geometrical features are determined by Σ_k. Banfield and Raftery [28] suggested that each covariance matrix parameterized by eigenvalue decomposition takes the following form:

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \qquad \text{(Eq. 3)}$$

where λ_k is a scalar, D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix where all the elements are proportional to Σ_k. It is suggested that characteristics of distributions such as orientation, volume and shape, are estimated by the data and can either vary between clusters or remain the same for all clusters [28,26,29].

According to Fraley et al. [26], the distribution for Expectation Maximization (EM) algorithm for multidimensional data, can be Spherical, Diagonal or Ellipsoidal. The Volumes and the Shapes of clusters can be equal or variable. The combination of these characteristics, defines each model (namely the covariance matrix Σ_k). For more information, the reader is encouraged to read the report by Fraley et al. [26].

The best model is determined according to the BIC (Bayesian Information Criterion) as initially proposed by Schwarz [29]. The BIC is basically the maximized log-likelihood but in order to avoid overfitting it includes a penalty term for the number of parameters in the model. The optimum number of clusters and the best model are defined by the value of BIC. A larger value of BIC indicates stronger evidence for the best model and number of clusters [27].

2.2. Discriminant analysis

In general, cluster analysis provides the optimum number of clusters as well as their centres. However, as correctly stated by Xu et al. [18], new observations cannot be directly assigned to the defined traffic regimes. In this case, discriminant analysis is needed to be carried out. According to Johnson and Wichern [30], discriminant analysis allocates new cases to the pre-defined cluster groups. Fraley and Raftery [27] state that in discriminant analysis (or supervised classification) known classifications (training set) are used to classify others (testing set).

Several discriminant analysis methods exist. In this study, a Discriminant analysis through Eigenvalue Decomposition was applied. More specifically, the procedure of applying a Gaussian finite mixture modelling for discriminant analysis where each known group (class) is modelled by a single Gaussian term with the same covariance structure among classes, is named as Eigenvalue Decomposition Discriminant Analysis (EDDA) by Bensmail and Celeux [31]. When the model is a normal mixture fitted by model-based clustering, the procedure is known as mclustDA [27].

In the followed approach, a separate mean vector for each class is calculated, but with the same ellipsoidal covariance matrix, which basically is the same with linear discriminant analysis.

2.3. Bayesian logistic regression

When the dependent (response) variable is not continuous, the appropriate statistical models are discrete outcome models (e.g. binary, multinomial etc.). In this study, the response variables are binary, therefore binary logistic regression models are appropriate.

The utility function U is:

$$U = \beta_0 + \Sigma \beta_i x_i \qquad \text{(Eq. 4)}$$

and the probability of an event is P is:

$$P = \frac{e^U}{e^U + 1} \qquad \text{(Eq. 5)}$$

where β_0 is the constant term and β_i is the coefficient for the explanatory variable x_i.

The Bayesian logistic regression approach is different than the traditional frequentist approach, because prior distributions for each parameter are defined and then the data are used to update beliefs about parameters. Therefore, the posterior distributions and the 95% credible intervals are produced. Lunn et al. [32] suggest that it is preferable to use “vague” or “non-informative” priors if little is known about the coefficient values.

It is noted that the likelihood function for Bayesian logistic regression is the same as in the frequentist inference. A rule of thumb is that an independent variable is statistically significant if the 95% credible interval (2.5%-97.5%) of the beta coefficient does not include the value of zero [32]. The DIC as the Bayesian generalization of Akaike information criterion [33] is a measure of model fit and to compare models.

3. Data preparation

Accident and traffic data from the Kifisias and Mesogeion avenues were used to explore the effect of traffic regimes on accident likelihood and severity. Only urban roads were considered in contrary to previous studies which explored data from freeways [18]. Furthermore, Powered-Two-Wheelers (PTWs) were considered as well in order to model the likelihood of accidents including a PTW.

The required accident data for Kifisias and Mesogeion avenues were collected from the Greek accident database SANTRA, which is provided by the Department of Transportation Planning and Engineering of the National Technical University of Athens. Only accidents with injuries were included (slight, severe, fatal). A 6-year period was considered for the analyses of the present study, namely 2006-2011.

Traffic data were extracted from the Traffic Management Centre (TMC) of Athens, which operates on a daily basis from July 2004 covering various major arterials in Athens. TMC collects traffic occupancy (measured in %), traffic flow (measured in number of vehicles per 5min) and mean-time speed (measured in km/h).

Each accident case was matched with traffic data from the closest upstream and downstream loop detectors. If an accident occurred on Wednesday 14 October 2011 at 19:00, traffic data for Wednesday 14 October 2011 18:00-19:00 from the closest upstream and downstream loop detectors are considered. To explore accident likelihood, a sample of non-crash cases was selected as in previous similar studies in international literature. More specifically, 2 non-accident cases for the same location (same time, 1 week before and after the accident case) for each accident case were selected. For example, if an accident occurred on Wednesday 14 October 2011 at 19:00, traffic data for Wednesday 7 October 2011 18:00-19:00 and Wednesday 21 October 2011 18:00-19:00 were extracted. The purpose of this approach was to compare accident-free traffic regimes with traffic regimes just before occurrence of an accident.

Unrealistic traffic data were excluded from datasets and were not considered for further analysis (e.g., values of average occupancy that exceeded 100%, average speed higher than 150-160 km/h and so on), following the same approach as in Christoforou et al. [8] and Xu et al. [18]. This was observed in 4 cases only.

A number of response variables were created, namely accident occurrence (1 for accident occurrence, 0 for non-accident occurrence), accident severity (1 for severe accidents, 0 for slight accidents) and PTW accident involvement (1 if a PTW was involved in an accident, 0 otherwise).

Summing up, two final datasets were considered; the first involved 1434 total cases for accident likelihood (480 accident and 954 non-accident cases), and the second 480 cases for accident severity (56 severe accidents, 424 slight accidents) and PTW accident involvement (298 accidents with PTWs, 182 accidents without PTWs).

This study followed the approach of Xu et al. [18], who used the traffic occupancy of nearby upstream and downstream loop detectors to classify traffic regimes. This approach was extended by considering also the standard deviation of occupancy as well.

4. Results

To achieve the aims of the study, a finite mixture cluster analysis was carried out to identify traffic regimes, followed by a discriminant analysis to correctly assign cases to traffic regimes on the basis of average and standard deviation of occupancy extracted from nearby loop detectors. It is noted that this is one of the first times that this alternative clustering method is applied for this purpose. Lastly, the effects of traffic regimes on accident likelihood and severity by applying Bayesian logistic models were investigated. As mentioned earlier, two distinct datasets were created; the former involves all accident cases and a random selection of non-accident cases and was used for accident likelihood, while the latter involves all accident cases only and was exploited to model accident severity and PTW accident involvement.

4.1. Finite mixture cluster analysis

4.1.1.Accident and non-accident cases

The average and standard deviation of occupancy were used to the first dataset of the 1434 total cases (480 accident and 954 non-accident cases). The finite mixture cluster analysis results revealed nine clusters. This was the optimal number of clusters as determined by the BIC criterion. Similarly, the finite mixture models showed the optimal covariance matrix Σ_k. More specifically, it has the form $\Sigma_k = \lambda_k D_k A_k D_k^T$ having an Ellipsoidal distribution, with a varying volume, shape and orientation between clusters (abbreviated VVV). The optimum number of clusters was determined to be five.

Table 1, presents information about each cluster, as well as the mean value (clustering centre).

Traffic Regimes	Regime 1	Regime 2	Regime 3	Regime 4	Regime 5	Regime 6	Regime 7	Regime 8	Regime 9
Percentage of cases (%)	12.06%	21.41%	17.43%	11.30%	10.25%	1.88%	6.42%	9.48%	9.76%
Cases	173	307	250	162	147	27	92	136	140
Accident cases	54	103	66	47	66	16	34	40	54
Non-accident cases	119	204	184	115	81	11	58	96	86
Average Occupancy upstream (%)	10.81	4.35	10.52	23.52	18.89	30.82	28.59	12.87	29.74
Average Occupancy downstream (%)	14.15	5.45	7.95	11.24	14.89	6.07	34.78	27.9	24.53
St.deviation of Occupancy upstream (%)	1.32	0.67	1.45	3.75	7.86	11.81	10.41	1.57	5.67
St.deviation of Occupancy downstream (%)	2.82	0.76	1.028	1.18	3.99	1.44	9.55	6.69	4.15

Table 1: Clustering centres and information for different traffic regimes (accident and non-accident cases).

The next two figures (Figure 1 and Figure 2) illustrate the scatterplot of occupancy values within the 9 clusters.

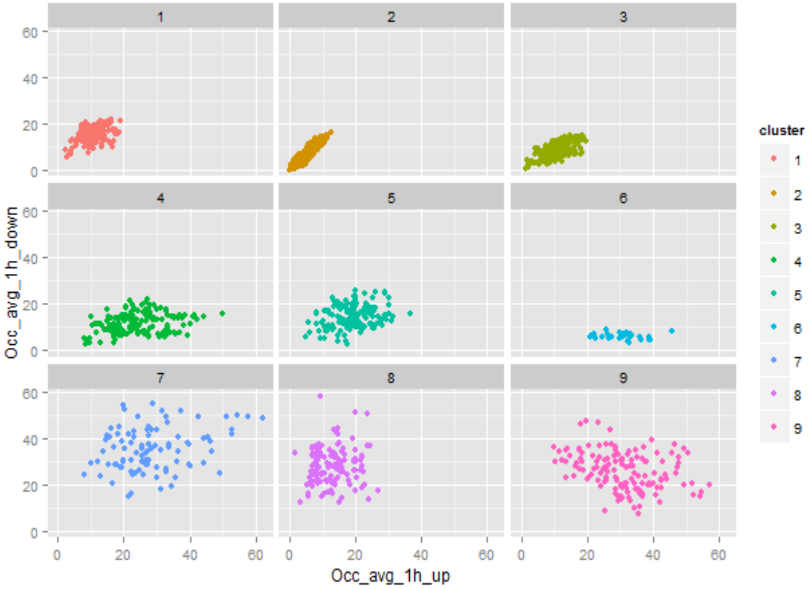


Figure 1: Scatter plots for average occupancy from upstream and downstream loop detectors for different clusters (regimes) regarding accident and non-accident cases.

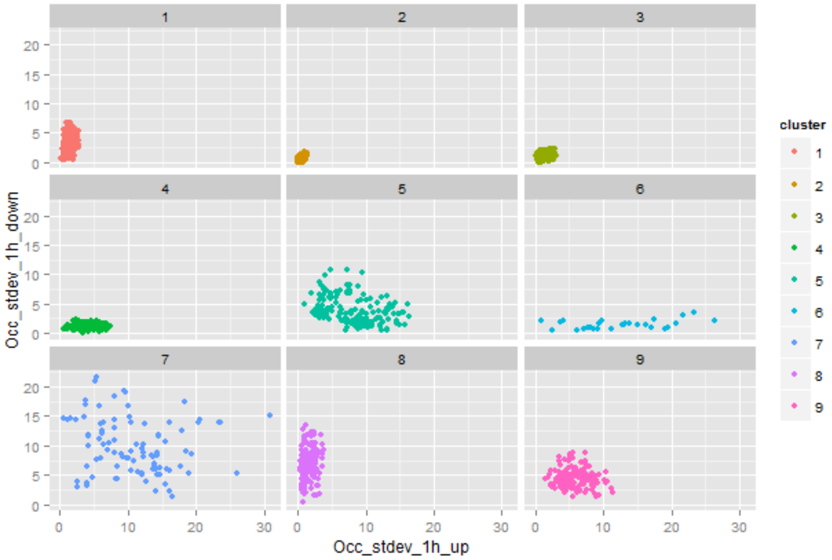


Figure 2: Scatter plots for standard deviation of occupancy from upstream and downstream loop detectors for different clusters (regimes) regarding accident and non-accident cases.

The nine clusters are summarized as follows:

<u>Traffic regime 1:</u> A relatively low percentage of total cases (12.06%) is assigned in cluster 1 (traffic regime 1). In traffic regime 1, there is a difference between upstream and downstream average occupancy (10.81% and 14.15% respectively). The standard deviation is low in both loop detectors, however a difference is observed.

<u>Traffic regime 2:</u> Almost 1/5 of cases (21.41%) were classified in traffic regime 2. In this regime, especially low and homogenous values of occupancy measurements can be observed (both lower than 5.5%). The standard deviation of occupancies is also low.

<u>Traffic regime 3:</u> 17.43% of total cases were assigned to cluster 3. The traffic characteristics of this traffic regime, are quite opposite to those of traffic regime 1. This regime presents a situation where there is a decrease in occupancy from

upstream to downstream. However the occupancy variation is quite low and similar.

Traffic regime 4: 11.30% of total cases belong to cluster 4. As shown in table 6-2, the clustering centre for upstream loop detector is 23.52%, while the respective centre for downstream detector is 11.24%. Thus, there is a great difference in traffic occupancy between upstream and downstream loop detectors, implying a transition from congestion to much better traffic conditions. A difference in occupancy variation can also be observed, where the upstream detector has the higher.

Traffic regime 5: The traffic conditions in this regime could be characterized as opposite to these of traffic regime 1. In this regime, both occupancies are relatively low, however, occupancy upstream is higher (almost 19%) than downstream (almost 15%). It is interesting, that the upstream occupancy faces a relatively high variation, having a high clustering centre for standard deviation (7.86%), while the standard deviation downstream is almost 4%. About 10% of cases belong to this traffic regime.

Traffic regime 6: Despite consisting only of 1.88% of cases, this traffic regime might have some of the most interesting characteristics. There is a very high difference in occupancy between upstream (30.82%) and downstream (only 6.07%) loop detectors, meaning the existence of a transition of very high congestion to a very high level of service. The clustering centre for upstream standard deviation of occupancy is also higher.

Traffic regime 7: Traffic regime 7 consists of less than 7% of the total cases. Downstream and upstream occupancy values are considered high (34.78% and 28.59%) indicating potential congestion. The clustering centres for standard deviations can be considered homogenous as well (9.55% for downstream and 10.41% for upstream).

Traffic regime 8: 9.48% of the sample is assigned to traffic regime 8. As shown in table 6-2, the clustering centres of average occupancy and standard deviation of occupancy upstream are 12.87% and 1.57% respectively, while those at downstream loop detector are 27.9% and 6.69% respectively. Therefore, a transition from low to high congestion can be observed.

Traffic regime 9: Both occupancy measurements of upstream and downstream loop detectors are high (29.74% and 24.53% respectively), indicating dense traffic conditions, however the clustering centres for the standard deviation of occupancy in this regime are significantly lower than those in traffic regime 7.

4.1.1.Accident cases

The average and standard deviation of occupancy were assigned to the dataset of the 480 accidents. The finite mixture cluster analysis results revealed five clusters. This was the optimal number of clusters as determined by the BIC criterion. Moreover, the finite mixture models showed the optimal covariance matrix Σ_k. More specifically, it has the form $\Sigma_k = \lambda_k D_k A_k D_k^T$ having an Ellipsoidal distribution, with a varying volume, shape and orientation between clusters (abbreviated VVV, please see Fraley et al. [8]. The optimum number of clusters was determined to be five. Table 2 illustrates the distribution of cases in each cluster, as well as the mean value (clustering centre) for each cluster.

Regimes	Regime 1	Regime 2	Regime 3	Regime 4	Regime 5
Percentage of total cases (%)	25.83%	25.21%	12.92%	23.75%	12.29%
Cases	124	121	62	114	59
Slight accidents	115	93	55	105	56
Severe accidents	9	28	7	9	3
Accidents with a PTW	78	57	49	72	42
Accidents without a PTW	46	64	13	42	17
Average Occupancy upstream (%)	11.28	4.92	21.1	26.53	24.28
Average Occupancy downstream (%)	11.49	6.01	30.16	12.03	27.06
St.deviation of Occupancy upstream (%)	1.66	0.75	8.52	8.17	4.38
St.deviation of Occupancy downstream (%)	2.45	0.84	6.29	1.94	7.88

Table 2: Clustering centres and information for different traffic regimes (accident cases).

The next two figures (Figure 3 and Figure 4) illustrate the scatterplot of occupancy values within the 5 clusters.

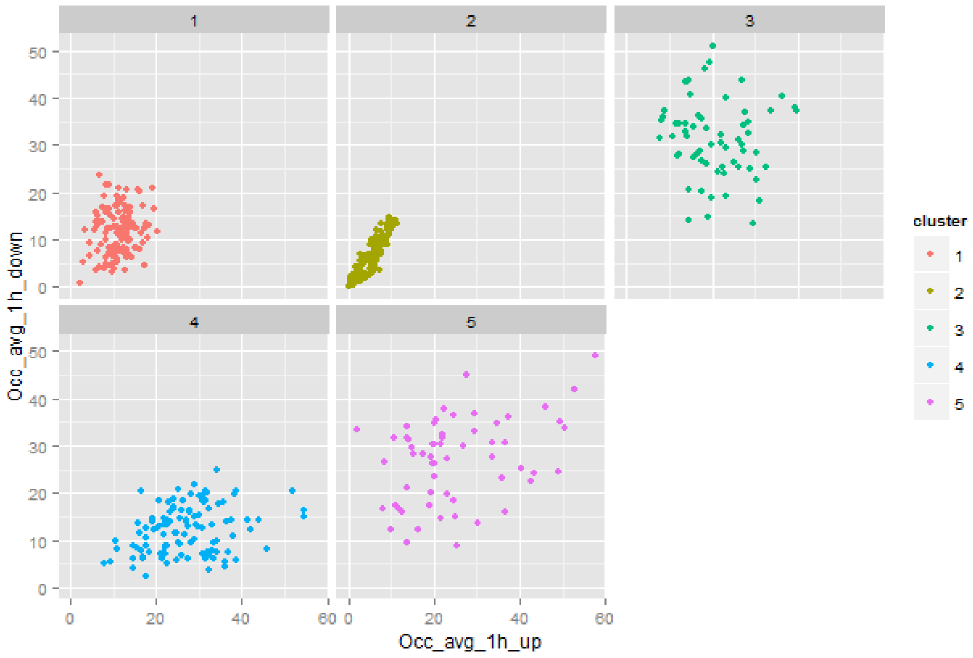


Figure 3: Scatter plots for average occupancy from upstream and downstream loop detectors for different clusters (regimes) regarding accident cases.

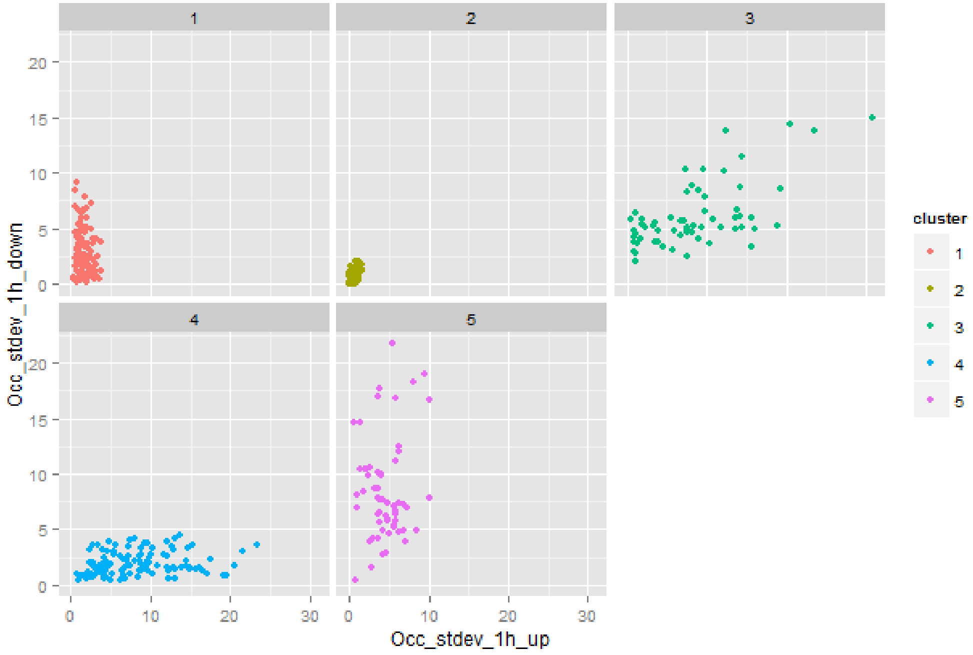


Figure 4: Scatter plots for standard deviation of occupancy from upstream and

downstream loop detectors for different clusters (regimes) regarding accident cases.

The description of the five clusters (regimes) is as follows:

Traffic regime 1: 26% of accident cases belong to traffic regime 1. Traffic regime 1 is characterized by quite identical medium occupancy values in both upstream and downstream loop detectors. More specifically, the clustering centre for average occupancy upstream is 11.28%, while for downstream the respective mean is 11.49%. The fluctuations in occupancy are similar as well. The level of service in that regime could be considered adequate.

Traffic regime 2: 25.24% of accident cases belong to traffic regime 2. This regime is characterized by quite homogenous and low occupancy across the two loop detectors (4.29% and 6.01%). The standard deviations of occupancy upstream and downstream are homogenous and low as well.

Traffic regime 3: 12.71% of cases are assigned to cluster 3. The main characteristic of this cluster, is the great difference observed in occupancy between upstream and downstream loop detectors. More specifically, the clustering centre for average occupancy downstream is 30.16%, while the respective centre upstream is 21.1%. These high values of occupancy indicate potential traffic congestion. Another interesting characteristic of this cluster is that the standard deviation of occupancy upstream is higher (8.52%) than downstream (6.29%).

Traffic regime 4: This traffic regime consists of 23.72% of total accident cases. The main difference from traffic regime 4 is the totally different traffic conditions. More specifically, there is a transition from high occupancy upstream (26.53%) to low occupancy downstream (12.03%). The clustering centre for standard deviation of occupancy is significantly higher upstream (8.17%) than downstream (only 1.94%).

Traffic regime 5: Only 12.33% of accidents belong to this cluster. Cases in this cluster are characterized by homogenous and high occupancy across upstream and downstream loop detectors (24.28% and 27.06% respectively). The cluster’s centre for standard deviation of occupancy downstream is slightly higher than upstream.

4.2. Discriminant analysis

The cluster analysis revealed meaningful groups of traffic regimes for the total cases (accident and non-accident cases) and also for accident cases. However, new observations cannot be directly assigned to previously pre-defined regimes. For that reason, discriminant analysis was carried out. Each dataset was randomly divided into a training and testing set. Each training set accounted for the 80% of each dataset, while each testing set accounted for the rest 20% of each dataset. The training sets were used for calibrating the models, while the testing sets were used to validate the models and test the accuracy for identifying traffic regimes and assigning new observations to the traffic regimes.

4.2.1. Accident and non-accident cases

The discriminant analysis performed on the total cases (accident and non-accident) and managed to classify 76.58% of total validation cases. The accuracy of predicted traffic regime memberships for the 20% testing set is illustrated on Table 3. The lowest classification accuracy was observed for traffic regime 3 (60.71%), while some substantially higher accuracies were observed, for example for traffic regimes 2 (91.04%), 6 (100%) and 8 (92.31%).

Traffic Regimes	Predicted group membership using discriminant analysis								
	Regime 1 (%)	Regime 2 (%)	Regime 3 (%)	Regime 4 (%)	Regime 5 (%)	Regime 6 (%)	Regime 7 (%)	Regime 8 (%)	Regime 9 (%)
Regime 1	**69.23%**	0.00%	30.77%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Regime 2	0.00%	**91.04%**	8.96%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Regime 3	1.79%	37.50%	**60.71%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Regime 4	0.00%	0.00%	20.59%	**76.47%**	2.94%	0.00%	0.00%	0.00%	0.00%
Regime 5	7.14%	0.00%	10.71%	0.00%	**67.86%**	7.14%	0.00%	7.14%	0.00%
Regime 6	0.00%	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%
Regime 7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**76.47%**	11.76%	11.76%
Regime 8	7.69%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**92.31%**	0.00%
Regime 9	0.00%	0.00%	0.00%	7.14%	10.71%	0.00%	3.57%	7.14%	**71.43%**

Table 3: Validation results of discriminant analysis (accident and non-accident cases).

4.2.2. Accident cases

The accuracy of predicted traffic regime memberships for the 20% testing set is presented on Table 4. The 80.21% of validation cases can be correctly classified. The lowest classification accuracy was observed for traffic regime 5 (50%), however all other traffic regimes have significantly higher classification accuracies. Classification accuracy for traffic regimes 1, 3 and 4 was substantially high reaching 82.14%, 100% and 89.4% respectively.

Traffic Regimes	Predicted group membership using discriminant analysis				
	Regime 1 (%)	Regime 2 (%)	Regime 3 (%)	Regime 4 (%)	Regime 5 (%)
Regime 1	**82.14%**	17.86%	0.00%	0.00%	0.00%
Regime 2	25.00%	**75.00%**	0.00%	0.00%	0.00%
Regime 3	0.00%	0.00%	**100.00%**	0.00%	0.00%
Regime 4	10.53%	0.00%	0.00%	**89.47%**	0.00%
Regime 5	25.00%	0.00%	16.67%	8.33%	**50.00%**

Table 4: Validation results of discriminant analysis (accident cases).

4.3. Bayesian logistic regression models

4.3.1. Effect of traffic regimes on accident likelihood

Using finite mixture cluster analysis, traffic data in both accident and non-accident cases were separated into 9 different traffic regimes on the basis of average and standard deviation of occupancy upstream and downstream of the accident location. A Bayesian logistic regression model was then developed to examine the relationship between traffic regimes and accident likelihood. Other variables such as traffic flow and speed were not included in the model because of the potential correlation with the traffic regimes, following the approach of previous studies [28].

The priors for the constant term and for the independent variables were all "vague" (non-informative), assuming to follow a normal distribution with zero mean and very low precision. The prior for the constant term was $alpha\sim dnorm(0, 0.0001)$. All categories of the independent variable "traffic regime", were following the exact same non-informative normal distribution, e.g. for traffic regime 2, $beta1\sim dnorm(0, 0.0001)$. The first 5,000 samples were discarded as adaptation and burn-in. Three chains and 20,000 more samples were used to ensure convergence. In addition, monitoring of the MC errors was performed as previously.

Table 5 summarizes the findings of the Bayesian logistic model for accident probability and provides the estimates of beta coefficients, the standard deviation and the 95% credible interval CI (2.5%-97.5%) and the odds ratios (OR). Only statistical significant parameters are illustrated on the table. The DIC of the model was 1821.37.

Variables	Parameters Estimates			Credible Intervals	
	Mean	St.Deviation	Odds Ratio	2.50%	97.50%
Constant term	-0.7932	0.1643	0.452	-1.121	-0.4779
Traffic regime 1 (ref.)	-	-	-	-	-
Traffic regime 5	0.5862	0.2334	1.797	0.1323	1.047
Traffic regime 6	1.183	0.4332	3.264	0.3482	2.047
DIC	1821.37				

Table 5: Significant parameters estimates, credible intervals and odds ratios for accident probability model.

Traffic regime 1 was the reference category for the dummy variable "traffic regime". The mean value of coefficient on traffic regime 5, provides a credibly nonzero predictiveness for accident probability, because the 95% credible interval of the beta coefficient does not contain zero (beta=0.5862, CI=0.1323-1.047). Furthermore, traffic regime 6 was also found to be significant (beta=1.183, CI=(0.3482-2.047)). All other traffic regimes were not considered significant, because the value zero was included in the credible interval for the posterior distributions. The positive signs of the mean values of the parameters of traffic regimes 5 and 6, indicate that these traffic regimes are associated with higher accident risk than traffic regime 1, and thus being considered being hazardous traffic regimes.

The odds ratio for traffic regime 5 is 1.797, meaning that the odds of accident occurrence for traffic regime 5 is almost twice than for traffic regime 1. Traffic conditions in this regime present a situation quite opposite of that of traffic regime 1. In traffic regime 5, the occupancy is reduced from 18.89% upstream to 14.89% downstream, while in traffic regime 1 a small increase is observed (from 10.81% to 14.15%) and the values of occupancy are relatively lower. However, traffic regime 6 was found to be associated with even higher accident risk (OR=3.264). These results show that the transition from very high occupancy (30.82%) to very low occupancy (6.07%), is associated with high accident likelihood. It is also worth noticing, that in this regime, a significant change in standard deviation of occupancy from upstream (11.81%) to downstream (1.44%) is observed, indicating accident risk.

4.3.2. Effect of traffic regimes on accident severity

A Bayesian logistic regression model was developed to examine the relationship between traffic regimes and accident severity.

The priors for the constant term and for the independent variables were “vague” (non-informative), assuming to follow a normal distribution with zero mean and very low precision. The prior for the constant was $alpha{\sim}dnorm(0, 0.0001)$. All categories of the independent variable “traffic regime”, were assumed to follow the same non-informative normal distribution, e.g. for traffic regime 2, $beta1{\sim}dnorm(0, 0.0001)$. The first 1,000 samples were discarded as adaptation and burn-in. Three chains and 5,000 more samples were used to ensure convergence. Aside from visual inspection of the chains, the Monte Carlo (MC) errors (i.e. the Monte Carlo standard error of the mean values) were also monitored. According to Spiegelhalter et al. [33], MC errors less than 0.05 indicate that convergence may have been achieved. In the model all MC errors were very low (less than 0.005) indicating convergence.

Table 6 summarizes the findings of the Bayesian logistic regression model for accident severity, and provides the estimates of beta coefficients, the standard deviation and the 95% credible interval (2.5%-97.5%) and the odds ratios (OR). Only statistical significant parameters are illustrated on the table. The value of the DIC of the model was 331.

Variables	Parameters Estimates			Credible Intervals	
	Mean	St.Deviation	Odds Ratio	2.50%	97.50%
Constant term	-2.589	0.357	0.075	-3.353	-1.936
Traffic regime 1 (ref.)	-	-	-	-	-
Traffic regime 2	1.377	0.419	3.963	0.5839	2.242
DIC	331				

Table 6: Significant parameters estimates, credible intervals and odds ratios for accident severity model.

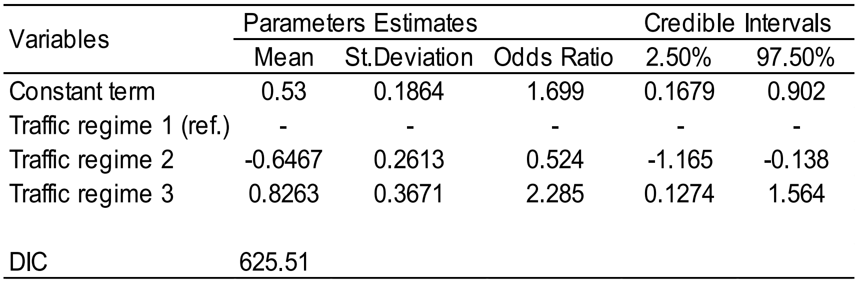
Traffic regime 1 was used as a reference category and only traffic regime 2 was found to be significant, because the 95% credible interval of the beta coefficient does not contain zero (0.5839-2.242). Other traffic regimes do not provide a credibly nonzero predictiveness for accident severity, because zero was well among the 95% credible interval.. The positive sign of the mean value of the parameter of traffic regime 2, means that traffic regime 2 is associated with higher accident severity than traffic regime 1. More specifically, the odds ratio of 3.963 indicates that the odds of an accident being severe or fatal in traffic regime 2 is almost 4 times higher than for traffic regime 1. Both traffic conditions in upstream and downstream loop detectors have a similar and low variation in occupancy. Since traffic regime 1 indicates more congested traffic conditions, less congestion is associated with higher severity levels. This finding might be considered consistent with findings of similar studies in the past [34,35,8], verifying the assumption that under less congestion, drivers tend to drive at higher speeds and therefore more severe accidents may occur.

4.3.3.Effect of traffic regimes on PTW accident involvement

The relationship between traffic regimes and likelihood of PTW accident involvement was examined through the application of binary logistic models.

The priors for the constant term and for the independent variables were "vague" (non-informative), assuming to follow a normal distribution with zero mean and very low precision. The prior for the constant term was $alpha{\sim}dnorm(0, 0.0001)$. All categories of the independent variable "traffic regime", were following the exact same non-informative normal distribution, e.g. for traffic regime 2, $beta1{\sim}dnorm(0, 0.0001)$. The first 1,000 samples were discarded as adaptation and burn-in. Three chains and 20,000 more samples were used to ensure convergence.

Table 7 summarizes the findings of the Bayesian logit model for PTW accident probability, and provides the posterior mean, the standard deviation and the 95% credible interval CI (2.5%-97.5%) and the odds ratios (OR). Only statistical significant parameters are illustrated on the table. The DIC of the model was 625.51.



Variables	Parameters Estimates			Credible Intervals	
	Mean	St.Deviation	Odds Ratio	2.50%	97.50%
Constant term	0.53	0.1864	1.699	0.1679	0.902
Traffic regime 1 (ref.)	-	-	-	-	-
Traffic regime 2	-0.6467	0.2613	0.524	-1.165	-0.138
Traffic regime 3	0.8263	0.3671	2.285	0.1274	1.564
DIC	625.51				

Table 7: Significant parameters estimates, credible intervals and odds ratios for

PTW accident involvement model.

Traffic regime 1 was the reference category. Traffic regimes 2 (beta=-0.6467, CI=(-1.165-0.138)) and 3 (beta =0.8263, CI=(0.1274-1.564)) were found to be significantly associated with PTW accident probability. However, traffic regime 2 had a negative posterior mean value, meaning that at this traffic regime (lower occupancy) the risk of PTW accident involvement is lower. The positive sign of the posterior mean for traffic regime 3 shows that a consistent very high occupancy in upstream (21.1%) and downstream (30.16%) loop detectors, and high variation in occupancy upstream (8.52%) is associated with increased probability of accidents involving a PTW. All other traffic regimes were not statistically significant.

5. Conclusions

This study investigated the influence of traffic regimes on safety of two major urban arterials in Athens. For that reason, real-time traffic data from nearby loop detectors were exploited. This attempt might be considered as one of the first implementations of such approach for urban arterials. More specifically, finite mixture cluster analysis was performed on the basis of average and standard deviation of occupancy measured at the two nearby loop detectors in order to classify urban traffic conditions into meaningful groups (regimes). This clustering method has not been widely used in studies with real-time data and is considered promising. It also has the advantage of defining the optimum number of clusters, in contrast to the k-means method which is a very popular method. The results revealed 9 clusters for accident and non-accident cases (used for accident likelihood) and 5 clusters for accident cases (used for accident severity and Powered-Two-Wheeler likelihood of accident involvement). Then, discriminant analysis was carried out so as to identify and predict cluster membership, on the basis of a validation dataset which was used as a test set (20%). Lastly, Bayesian logistic regression models were applied to unveil the influence of different traffic conditions (regimes) on accident likelihood and accident severity, having an emphasis on PTWs.

The findings of the study demonstrate that this approach is promising when applied on urban networks. Concerning the likelihood of accident occurrence, higher occupancy was found to lead to high accident risk. Consequently, it can be concluded that higher occupancy and potentially congested traffic conditions contribute to higher accident occurrence but also may result to less severe accidents. When analysing accident severity, traffic regime 2 was found to have the greatest influence, correlating high accident severity with lower occupancy levels and therefore less congestion.

The traffic condition that was identified as the most hazardous for accident occurrence was the traffic regime when the transition from very high to very low occupancy took place. This finding is consistent with the results of Hossain and Muromachi [11], who argue that a fast moving uncongested downstream when being followed by slow moving and congested upstream may be more hazardous. One possible reason for this finding may be the fact that drivers may compensate for travel time loss and consequently accelerate [11].

It was also found that PTWs are more likely to be involved in accidents when the occupancy is very high (potential congestion). On the contrary, in low occupancy (traffic regime 2) PTWs are less likely to be involved in accidents, indicating that congestion is correlated with increased probability of PTW involvement in accidents. Consequently, traffic regime 2 is associated with higher accident severity but with lower PTW accident probability.

In previous similar studies in international literature, traffic flow was classified in groups and in few studies the influence of these groups on freeway safety was investigated. This study aimed to add in the current knowledge by using real-time traffic data collected from nearby loop detectors in major urban arterials. Furthermore, alternative modeling methods such as the finite mixture cluster analysis were applied. Therefore, the research demonstrated in the paper can be considered a supplement to previous studies which could assist transportation professionals better understand the impact of traffic on road safety. The results of this study can be applied by transportation professionals to reduce crash risk and severity on urban areas by developing proactive safety management strategies. If the crash prone traffic regimes are identified then considerations should be given to improve road safety. For example, the results showed that the transition from very high occupancy to very low occupancy (traffic regime 6) is associated with high accident risk. Therefore, specific actions could be applied (e.g. variable messages signs) to warn drivers and increase road safety levels.

The authors recognize the limitations of the study. When modeling accident likelihood, it would be desirable to have the whole population of non-accident cases. However, this was not feasible as it would be extremely time consuming to collect and process such huge dataset. However, it is noted the approach of this study was widely followed by relevant past literature in the field and was proved to be efficient.

Future research should focus on finding ways to improve the modeling methods of accident likelihood. In addition, more microscopic traffic data from urban and rural roads could be utilized.

References

[1] Oh, C., Oh, J-S., Ritchie, S.G., Chang, M., Real-time estimation of freeway accident likelihood. Presented at the Annual Meeting of the Transportation Research Board, 8–12 January (2001), Washington DC.

[2] Oh, J., Oh, C., Ritchie, S., Chang, M., Real-Time Estimation of Accident Likelihood for Safety Enhancement. Journal of Transportation Engineering 131(5), (2005) 358-363.

[3] Oh, C, Park, S, Ritchie, SG., A method for identifying rear-end collision risks using inductive loop detectors. Accident Analysis and Prevention 38 (2006) 295–301.

[4] Lee, C., Hellinga, B., Saccomanno, F., Real-time crash prediction model for the application to crash prevention in freeway traffic. Proceedings of the 82nd Annual meeting of the Transportation Research Board, January 12- 16 (2003) Washington, D.C.

[5] Zheng, Z., Ahn, S., Monsere, C.M., Impact of traffic oscillations on freeway crash occurrences. Accident Analysis and Prevention 42 (2010) 626– 636.

[6] Abdel-Aty, M., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., Real-time prediction of visibility related crashes. Transportation Research Part C 24 (2012) 288–298.

[7] Xu, C., Tarko, A.P., Wang, W., Liu, P., Predicting crash likelihood and severity on freeways with real-time loop detector data. Accident Analysis and Prevention 57 (2013) 30– 39.

[8] Christoforou, Z., Cohen, S., Karlaftis, M., Vehicle occupant injury severity on highways: An empirical investigation. Accident Analysis and Prevention 42 (2010) 1606-1620.

[9] Yu, R., Abdel-Aty, M., Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. Accident Analysis and Prevention 62 (2014a) 161-167.

[10] Yu, R., Abdel-Aty, M., Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Safety Science 63 (2014b) 50-56.

[11] Hossain, M., Muromachi, Y., Understanding crash mechanism on urban expressways using high-resolution traffic data. Accident Analysis and Prevention 57 (2013) 17– 29.

[12] Hall, F., Hurdle, V., Banks, J., Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. Transportation Research Record 1365 (1992) 12–18.

[13] Kerner, B.S., Rehborn, H., Experimental properties of complexity in traffic flow. Physical Review E 53 (5) (1996) 4275–4278.

[14] Wu, N., A new approach for modeling of fundamental diagrams. Transportation Research Part A 36 (10), (2002) 867–884.

[15] Abdel-Aty, M., Uddin, N., Pande, A., Split models for predicting multi-vehicle crashes during high-speed and low-speed operating conditions on freeways. Transportation Research Board 1908 (2005) 51–58.

[16] Golob, T., Recker, W., A method for relating type of crash to traffic flow characteristics on urban freeways. Transportation Research Part A 38 (1) (2004a) 53–80.

[17] Golob, T., Recker, W., Alvarez, V., Freeway safety as a function of traffic flow. Accident Analysis and Prevention 36 (6) (2004b) 933–946.

[18] Xu, C., Liu, P., Wang, W., Li, Z., Evaluation of the impacts of traffic states on crash risks on freeways. Accident Analysis and Prevention 47 (2012) 162– 171.

[19] Yeo, H., Jang, K., Skabardonis, A, Kang, S., Impact of traffic regimes on freeway crash involvement rates. Accident Analysis and Prevention 50 (2013) 713-723.

[20] Jung S., Qin X., Noyce D.A., Rainfall effect on single-vehicle crash severities using polychotomous response models. Accident Analysis and Prevention 42 (2010) 213-224.

[21] Pirdavani, A., De Pauw, E., Brijs, T., Daniels, S., Magis, S., Bellemans, T., Wets, G., Application of a Rule-Based Approach in Real-Time Crash Risk Prediction Model Development using Loop Detector Data, Traffic Injury Prevention, (2005) DOI: 10.1080/15389588.2015.1017572.

[22] Montella, A.,, Aria, M., D'Ambrosio, A., and Mauriello, F., Analysis of powered two-wheeler crashes in Italy by classification trees and rules

discovery, Accident Analysis and Prevention 49 (2012) 58-72.

[23] Maestracci, M., Prochasson, F., Geffroy, A., and Peccoud, F., Powered two-wheelers road accidents and their risk perception in dense urban areas: Case of Paris, Accident Analysis and Prevention 49 (2012) 114–123.

[24] Harnen, S., Wong, S.V., Radin Umar, R.S., & Wan Hashim, W.I., Motorcycle crash prediction model for non-signalized intersections. IATSS Research 27(2) (2003) 58-65.

[25] Kasantikul, V., Ouellet, J.V., Smith, T., Sirathranont, J., & Panichabhongse, V., The role of alcohol in Thailand motorcycle crashes, Accident Analysis and Prevention, 37 (2005) 357-366.

[26] Fraley, C., Murphy, T.B., Scrucca, L., Mclust version 4 for R: Normal Mixture modeling for model-based clustering, classification, and density estimation. Technical Report No.597 (2012), Department of Statistics, University of Washington, USA.

[27] Fraley, C., Raftery, A.E., Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association 97 (2002) 611-631.

[28] Banfield, J.D., Raftery, A.E., Model-based Gaussian and non-Gaussian clustering. Biometrics 49 (1993) 803-821.

[29] Celeux, G., Govaert, G., Gaussian parsimonious clustering models. Pattern Recognition 28 (1995) 781-793.

[29] Schwarz, G., Estimating the dimension of a model. The annals of Statistics 6 (1978) 461-464.

[30] Johnson R.A., Wichern D.W., Applied Multivariate Statistical Analysis. Prentice-Hall International, Inc. (1998) Upper Saddle River, NJ.

[31] Bensmail, H., Celeux, G., Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition. Journal of the American Statistical Association 91 (1996) 1743-1748.

[32] Lunn, D.J., Jackson, C., Best, N., Thomas, A., Spiegelhalter, D. The BUGS Book: A practical introduction to Bayesian Analysis, 1st ed. (2012), Boca Raton, FL., Chapman and Hall/CRC.

[33] Akaike, H., A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (1974) 716–723.

[33] Spiegelhalter D., Best N., Carlin B., Linde V., Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society B 64(4), (2003) 583–616.

[34] Martin, J.-L., Relationship between crash rate and hourly traffic flow on interurban motorways. Accident Analysis and Prevention 34 (2002) 619–629.

[35] Quddus, M.A., Wang, C., Ison, S.G., The impact of road traffic congestion on crash severity using ordered response models. Road traffic congestion and crash severity: econometric analysis using ordered response models. Journal of Transportation Engineering 136(5), (2009) 424-435.