Combining data mining techniques to investigate crash severity in urban motorways

Theofilatos A.*, Ziakopoulos A., Yannis G.

**National Technical University of Athens, Department of Transportation Planning and Engineering, Iroon Polytechneiou str. 5, Zografou Campus, Athens, Greece*

Abstract - The impact of real-time traffic parameters on crash severity has only recently been explored. The current research adds to knowledge by examining data mining techniques to investigate the determinants behind injury severity of occupants involved in crashes in the urban motorway Attica Tollway Athens, Greece. The proposed methodological approach involves a three-level process. Firstly, a two-step cluster analysis is performed to classify data into different groups (clusters). Secondly, a factor analysis is applied to group a number of relevant traffic variables into appropriate factors. Lastly, binary logit regression models are used to correlate clusters and factors with crash injury severity. The required crash data were extracted from the database SANTRA of National Technical University of Athens, consisting of 387 injury cases, for the period 2006 to 2011. The findings of this study demonstrate that hazardous traffic conditions are identified and real-time safety management policies can be promoted and improved.

Keywords: Injury severity, real-time; traffic data, data mining

1. INTRODUCTION

The effect of traffic characteristics on road safety has been extensively investigated. During the past decade, increased attention has been given to developing relationships between real-time traffic parameters and road safety indicators [1-3]. However, limited knowledge is available regarding the impact of traffic states on road safety. Another gap of knowledge identified is that European countries are underrepresented in the literature as very few relevant studies were identified [4].

In that context, Abdel-Aty et al. divided freeway traffic flow in high and low speed states for crash severity analyses [5]. Xu et al. gave emphasis on the need to divide traffic in states and explore their effect on safety [6]. More specifically, the authors used traffic occupancy measured from nearby loop detectors and classified traffic flow into traffic states and was found that the impact of traffic flow parameters on crash risk is not the same across different traffic flow state. Golob and Recker and Golob et al. investigated the impact of traffic by dividing traffic flow into different traffic states by using cluster analysis [7,8].

Yeo et al. defined traffic states (free flow, back of queue, bottleneck front, congestion) according to their distinctive patterns and modelled crash involvement rate for each traffic state [9]. Studies using real-time traffic data to investigate crash severity are also relatively few [10-13].

Thus, the main objective of the paper is to add to current knowledge by addressing the research gaps identified and divide traffic parameters in freeways into different states and investigate their impact on crash injury severity of occupants.

2. DATA

Empirical data have been collected for the period 2006-2011 to investigate the relationship between traffic crash injury severities in Attica Tollway ("Attiki Odos") in Athens, Greece. The unit of analysis was any vehicle occupant involved in a crash (rider, driver or passenger) resulting in at least one person being slightly injured – damage only crashes are not included in the data. Each of the total cases in the dataset is a record of the severity level sustained by each vehicle occupant involved in the crash. Therefore, a single crash would correspond to various observations that are equal to the number of all injured persons involved in the crash.

Traffic data for the Attica Tollway were extracted after a close collaboration with the Traffic Management and Motorway Maintenance Centre of the tollway. Inductive loops (sensors) are placed every 500 meters, inside the asphalt pavement of the open sections of the motorway, and every 60 meters for the tunnel segments. The loops are continuously providing information regarding the volume, speed and density of traffic. The data that were collected for the Attica Tollway consisted of the traffic parameters of traffic flow and occupancy.

For each crash, a 1-hour time series of traffic flow, speed, occupancy and truck proportion were extracted from the closest upstream loop detector measured in 5-minutes intervals. Then, the 5-minutes were aggregated in order to obtain averages, standard deviations etc. It is noted that a 5-min time lag was used to avoid the impact of the crash itself on the traffic variables and also to anticipate any possible inaccuracy in the reporting of the exact time of the crash. For example, if a crash occurred at 21:00, the traffic data considered were obtained from the 19:55-20:55 period.

Similar approaches have been applied in other real-time data analyses [5, 10]. The raw 5-min traffic data before the time of the crash occurrence (considering the 5-min time lag that was described earlier) were extracted from the closest upstream loop detector and then were further aggregated in

60-min batches in order to obtain average values. Traffic flow was divided by the number of lanes in order to be consistent throughout all road segments. At times, loop detectors suffered from operational problems that might have resulted in unreasonable values for speed, volume, and occupancy. Such values (e.g. occupancy>100%, speed>200 km/h or speed>0 coinciding with traffic flow=0) were discarded from the database. Crashes with traffic data unavailability were also discarded.

Descriptive Statistics		
Traffic Variables	**Mean**	**Std. Deviation**
Q_avg_1h	63.077	39.132
Q_stdev_1h	7.047	5.020
Q_median_1h	63.020	39.396
Q_cv_1h	0.142	0.12555
V_avg_1h	100.968	17.192
V_stdev_1h	4.437	4.9976
V_cv_1h	0.053	0.0830
Occ_avg_1h	0.043	0.0340
Occ_stdev_1h	0.005	0.0095
Occ_cv_1h	0.162	0.1302
TrProp_avg_1h	5.299	4.3130
TrProp_stdev_1h	2.267	2.1493
TrProp_cv_1h	0.639	0.7259

Table 1: Descriptive statistics of considered traffic variables.

3. METHODS

3.1 Two-step cluster analysis

The first step is to group observations together on the basis of the traffic parameters. The method of cluster analysis that was chosen was the two-step cluster analysis. This method of clustering is most appropriate for very large data files and it can produce solutions based on both continuous and categorical variables. The clustering algorithm is based on a distance measure that gives the best results if all variables are independent, continuous variables have a normal distribution, and categorical variables have a multinomial distribution, and these assumptions are adequately met in the present analysis.

The first step of the two-step procedure is the formation of pre-clusters. The goal of pre-clustering is to reduce the size of the matrix that contains distances between all possible pairs of cases. In the second step, the standard hierarchical clustering algorithm is applied on the pre-clusters. The clustering criterion (in this case the BIC - Bayesian Information Criterion) is computed for each potential number of clusters. Smaller values of the BIC indicate better clustering outcome. Also, a satisfactory solution should have a large Ratio of BIC Changes and a large ratio of distance measures.

3.2 Factor analysis

Another approach was subsequently followed. More specifically, it was aimed to find meaningful groups of variables reflecting traffic parameters. In order to achieve this objective, a factor analysis was performed. Factor analyses assist in understanding the structure of a large set of independent variables and reduce this data set to a more easily managed one. The Kaiser-Meyer-Olkin measure of

sampling adequacy was used, with values above 0.7 considered to be very satisfactory. The optimal number of components retained can be defined through a combination of more than one criterion and more specifically, the criteria of Kaiser-Meyer-Olkin and the variance explained (as much variance explained as possible is sought). Factor scores are produced and saved in order to be used for further analysis after performing the current one. Rotation, namely the orthogonal rotation (Varimax) was selected to improve the interpretability of the factors.

3.3 Binary logistic regression

Linear regression models that have been used in transport applications have the assumption that the response (dependent) variable is continuous. However, in cases where the response variable is not continuous, discrete outcome models should be applied instead. When there are two discrete outcomes, logistic regression models can be applied. It is obvious that the goal is the same as in simple linear regression. The best fitting model which describes the linear relationship between a binary (dichotomous) dependent variable and a number of explanatory variables (predictors) is pursued.

The goodness-of-fit of the model can be assessed with the likelihood ratio test. The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_f) over the maximized value of the likelihood function for the simpler model (L_0). Another indicator of the goodness of fit of the model is the Hosmer-Lemeshow statistic [14]. A non-significant value in the chi square suggests a good fit.

4. RESULTS

4.1 Cluster analysis

The two-step cluster analysis revealed 2 groups (clusters). This was defined on the basis of the BIC criterion and it was the optimum number of clusters. The model is characterized as having a good fit. Table 2 that follows below illustrates the distribution of observations in the two clusters. Cluster 1 consists of 265 injury cases and cluster 2 consists of 122 injury cases.

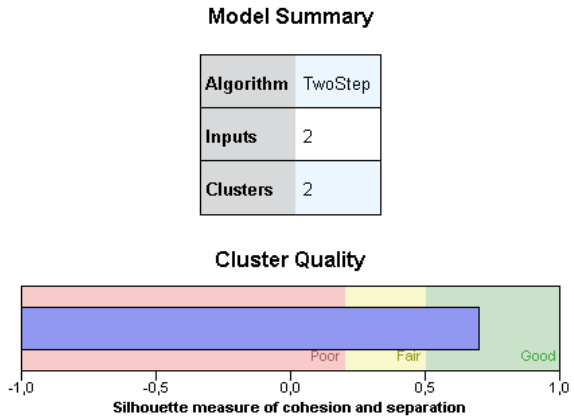

Figure 1: Model fit of the two step cluster analysis

Cluster	Observations
1	265
2	122
Total	387

Table 2: Distribution of observations in the two clusters.

Table 3 illustrates the centroids of traffic flow and occupancy in the two clusters. More specifically, it can be observed that cluster 1 has a mean value of 41.63 vehicles per lane per 5 minutes and low occupancy (3%). Thus it can be assumed to represent less congested conditions. On the other hand, cluster 2 has a mean value of 110.43 vehicles per lane per 5 minutes and higher occupancy levels (8%), representing more congested conditions compared to cluster 1.

Cluster	1-h average flow		1-h average occupancy	
	Mean	Std. Deviation	Mean	Std. Deviation
1	41.63	21.153	0.03	0.013
2	110.43	25.453	0.08	0.032

Table 3: Centroids of flow and occupancy in the two clusters.

Figures 2 and 3 illustrate the kernel density estimation regarding traffic clusters in respect to each independent traffic variable at a time: traffic flow and occupancy. The kernel density estimation (KDE), which is a non-parametric way to estimate the probability density function of a variable, was used here to provide a first visual inspection of the data. It is noted that the y-axes are unit-less, as they represent the probability density.

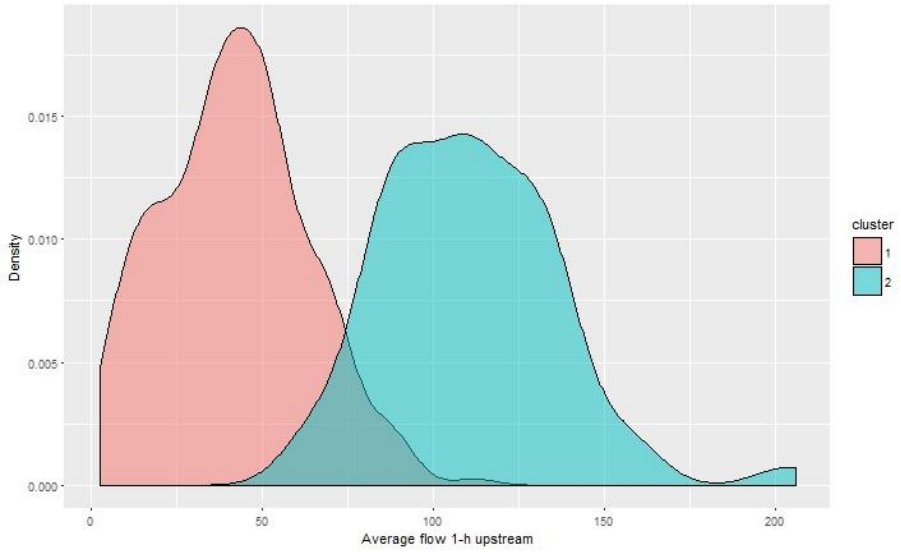

Figure 1: Probability density function for average flow.

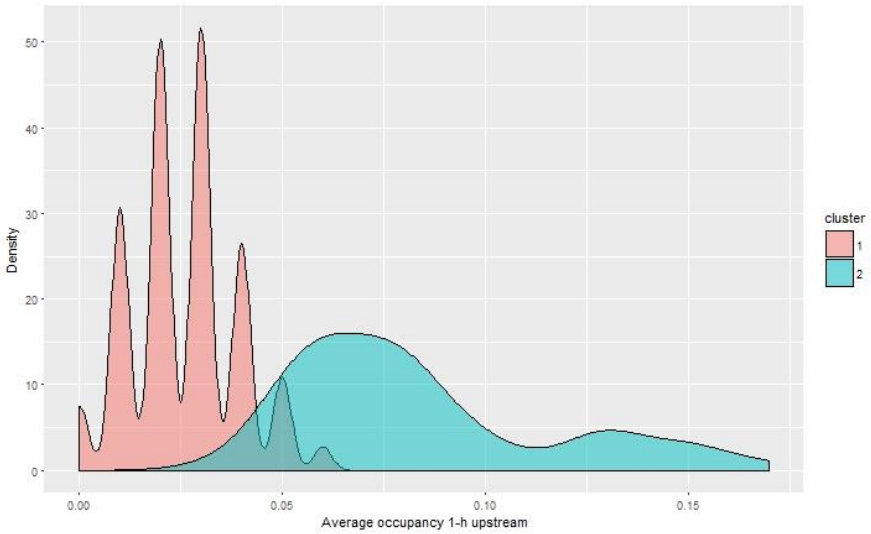

Figure 2: Probability density function for average occupancy.

4.2 Logistic regression to unveil the impact of traffic states

The cluster membership was used as independent variable to find the effects of traffic states on crash injury severity. Model results are presented on Table 4.

Variable	Beta coefficient	Std.error	p-value
Constant term	-1.740	0.171	<0.001
Cluster_2	-1.412	0.488	0.004

Table 4: Summary of the logistic regression models on clusters.

The AIC was 261.3 and the log-likelihood ratio was 10.39 (>3.84 for 1 degree of freedom). The Hosmer and Lemeshow test was also non-significant, indicating an adequate fit. Results indicate the cluster 2 is associated with lower severities of involved occupants (beta coefficient=-1.3654, p-value=0.005), suggesting that more congesting traffic conditions result in less severe injuries. This finding is in line with previous literature in the field [6, 10, 15] may be attributed to the fact that driving speeds are lower in congestion and therefore more slight injury crashes may occur.

4.3 Factor analysis results

The factor analysis was performed on the following traffic variables (Q: Traffic flow, V: Speed, Occ: Occupancy, TrProp: Truck Proportion in traffic. All are referring to the in the 1-hour time slice prior to crash): Q_avg_1h (average flow), V_avg_1h (average speed), Occ_avg_1h (average occupancy), TrProp_avg_1h (average truck proportion), V_cv_1h (coefficient of variation of speed), Occ_stdev_1h (standard deviation of occupancy), , V_stdev_1h (standard deviation of speed), Q_stdev_1h (standard deviation of flow), Occ_cv_1h (coefficient of variation of occupancy), Q_cv_1h (coefficient of variation of flow), TrProp_stdev_1h (standard deviation of truck proportion), TrProp_cv_1h (coefficient of variation of truck proportion) and Q_median_1h (median of flow). Four (4) factors were determined to be the optimal number, explaining 83.82% of the total variance. The KMO test was 0.651 and considered adequate.

Rotated Factor Matrix				
Variables considered	**Factors**			
	1	**2**	**3**	**4**
V_cv_1h	0.891			
Occ_stdev_1h	0.888			
Occ_avg_1h	0.850			
V_avg_1h	-0.826			
V_stdev_1h	0.810			
Q_stdev_1h	0.631	0.451		
Occ_cv_1h		0.928		
Q_cv_1h		0.916		
TrProp_stdev_1h			-0.812	
TrProp_cv_1h			-0.671	-0.592
Q_avg_1h	0.579		0.630	
Q_median_1h	0.583		0.622	
TrProp_avg_1h				0.934

Table 5: Factor analysis results.

Results show that the Factor 1 consists of average occupancy, average flow, average speed and median of flow as well as variations in speed, flow and occupancy. Factor 2 is characterized by variations in flow and occupancy (Q_stdev_1h, Occ_cv_1h and Q_cv_1h). Factor 3 involves variations in truck proportion (TrProp_stdev_1h, TrProp_cv_1h), average flow (Q_avg_1h) and median of flow (Q_median_1h). Lastly, Factor 4 consists of truck proportion parameters, namely, average truck proportion (TrProp_avg_1h) and coefficient of variation of truck proportion (TrProp_cv_1h).

4.4 Logistic regression on factors

Based on the results of the factor analysis, the four produced factors were used as independent variables (Table 6):

Variables in the Equation			
Variables	**Beta coefficient**	**Std. error**	**p-value**
Factor 1	-0.385	0.219	0.079
Factor 2	0.358	0.136	0.008
Constant term	-2.132	0.172	0.000

Table 6: Variables in the Equation

The Hosmer and Lemeshow test was non-significant (p-value = 0.219), indicating an adequate fit. Overall, the average and variations of truck proportion were not found to have any impact on resulted injuries of occupants. Results of the beta coefficient of Factor 1 indicate that increased values of flow and occupancy lead to less severe crashes. Moreover, increased average speed is associated with more severe injuries. Similarly, Factor 1 interpretation means that V_cv_1h (coefficient of variation of speed), Occ_stdev_1h (standard deviation of occupancy), V_stdev_1h (standard deviation of speed) and Q_stdev_1h (standard deviation of flow) are associated with lower injury severity. However, this is in contrast to the interpretation of the beta coefficient of Factor 2, which suggest the opposite.

Consequently, the impact of traffic variations on crash injury severity is not clear and needs further investigation in dedicated studies.

5. CONCLUSIONS

This study extended previous literature in the field and endeavored to add to current knowledge by combining data mining techniques to model the influence of traffic parameters on crash severity of a motorway in Greece, namely Attica Tollway. For that reason traffic data from loop detectors upstream to crash locations were used. Firstly, two step cluster analysis was performed on the basis of average flow and occupancy measured at the in order to classify motorway traffic conditions into meaningful groups (traffic states). Secondly, a different approach was selected. More specifically, all traffic variables were used as input to the factor analysis. Afterwards, logistic regression models were applied to unveil the influence of different traffic conditions on crash injury of occupants.

The results revealed 2 critical traffic states for estimating crash injury severity. When analyzing crash severity, low occupancy and high traffic flow were found to be associated with lower crash severities. Regarding the grouping of variables together, results are not very straightforward as it is suggested that the traffic variations have contradictory effects. However, some results were very clear. For instance, congested conditions lead to slight injuries, whilst increased mean speeds lead to severe injuries and fatalities. Interestingly, truck proportions did not seem to have a singular effect at all.

The research demonstrated in the paper can be considered a supplement to previous studies which could assist transportation professionals to better understand the impact of traffic parameters on road safety and develop proactive safety management strategies. If the crash prone traffic parameters are identified then specific actions should be taken towards that direction to improve road safety (e.g. variable messages signs) to warn drivers and increase road safety levels.

6. ACKNOWLEDGEMENTS

This research is implemented through IKY scholarships programme and co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the action entitled ”Reinforcement of Postdoctoral Researchers”, in the framework of the Operational Programme ”Human Resources Development Program, Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) 2014 – 2020. The authors would like to thank the Attica Tollway Operations Authority (“Attiki Odos”) for providing the data, especially officers Pantelis Kopelias and Fanis Papadimitriou.

7. 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. 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.
3. Zheng, Z., Ahn, S., Monsere, C.M., Impact of traffic oscillations on freeway crash occurrences. Accident Analysis and Prevention 42 (2010) 626– 636.
4. 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.

5. 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.
6. 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.
7. Golob, T., Recker, W., A method for relating type of crash to traffic flow characteristics on urban freeways. Transportation Research Part A 38 (1) (2004) 53-80.
8. Golob, T., Recker, W., Alvarez, V., Freeway safety as a function of traffic flow. Accident Analysis and Prevention 36 (6) (2004) 933-946.
9. 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.
10. Christoforou, Z., Cohen, S., Karlaftis, M., Vehicle occupant injury severity on highways: An empirical investigation. Accident Analysis and Prevention 42 (2010) 1606-1620.
11. 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.
12. Yu, R., Abdel-Aty, M., Using hierarchical Bayesian binary logit models to analyze crash injury severity on high speed facilities with real-time traffic data. Accident Analysis and Prevention 62 (2014) 161-167.
13. Theofilatos A., (2017). Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. Journal of Safety Research, 61, 9-21.
14. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. Communications in Statistics - Theory and Methods. 1980; 10, 1043–1069.
15. Xu, C., Tarko, A.P., Wang, W., Liu, P. (2013). Predicting crash likelihood and severity on freeways with real-time loop detector data. Accident Analysis and Prevention 57, 30–39.