TRANSPORT
ISSN 1648-4142 / eISSN 1648-3480
2017 Volume 32(4): 379–385
doi:10.3846/16484142.2014.995702

A COMPARISON BETWEEN PREDICTION POWER OF ARTIFICIAL NEURAL NETWORKS AND MULTIVARIATE ANALYSIS IN ROAD SAFETY MANAGEMENT

Mario De Luca

Dept of Civil, Construction and Environmental Engineering,
University of Naples Federico II, Naples, Italy

Submitted 11 March 2014; resubmitted 24 June 2014; accepted 9 September 2014;
first published online 14 January 2015

Abstract. This paper presents a methodology for the management of road safety on two-lane highways. The methodology is based on an experimental investigation carried out on a stretch of road located in southern Italy (the two-lane highway SS106). The study analyses accidents occurring between 2000 and 2005 and the data concerning the accidents that were acquired from police reports. The geometric data were acquired from the official cartography, while the traffic and environmental data were provided by the regional agency for roadway management. The data, organized and stored in a specific designed Geographic Information System (GIS), were processed using a series of statistical procedures, in particular, the results took out the following two models: Model 1 was produced by MultiVariate Analysis (MVA) and the Model 2 was obtained using the Artificial Neural Network (ANN) technique. Comparing the two models, it emerged that Model 2 is better than Model 1 because the total sum of the residual is lower. However, Model 1 is more efficient in estimating the more dangerous black spots.

Keywords: artificial neural network; GIS; non-linear model; cluster analysis; road safety.

Introduction

In recent years, artificial-computational intelligence has found increasing applications in management of transportation infrastructures. Many researchers, in particular, have employed Artificial Neural Network (ANN) procedures to analyse factors related to data processing. In the scientific literature, many research works have dealt with the road safety issues (Vujanić *et al.* 2013) driver speed behavior and traffic flow.

Chiou (2006) employs ANN to develop an accident appraisal expert system. The results show that the ANN model can achieve a high correctness rate of 85.72% in training and 77.91% in validation and a low Schwarz's Bayesian information criterion of –0.82 in training and 0.13 in validation, which indicates that the ANN model is suitable for accident appraisal. Furthermore, in order to measure the importance of each explanatory variable, a general influence index is computed based on the trained weights of ANN. It is found that the most influential variable is right-of-way, followed by location and alcoholic use. This finding concurs with the prior knowledge in accident appraisal. Thus, for the fair assessment of accident liabilities the correctness of these three key variables is very important for police investigation reports.

Chang (2005) shows that the Poisson or negative binomial regression model can be used to analyse vehicle accident frequency for many years. However, these models have a pre-defined underlying relationship between dependent and independent variables. If this assumption is violated, the model could lead to erroneous estimation of accident likelihood. On the contrary, the ANN, which does not require any pre-defined underlying relationship between dependent and independent variables, has been shown to be a powerful tool in dealing with prediction and classification problems. Thus, this study employs a negative binomial regression model and an ANN model to analyse 1997–1998 accident data from the National Freeway 1 in Taiwan. Comparing the prediction performance between the negative binomial regression model and ANN model, this study demonstrates that ANN is a consistent alternative method for analysing freeway accident frequency.

Delen *et al.* (2006) uses a series of artificial neural networks to model the potentially non-linear relation-

Corresponding author: Mario De Luca
E-mail: *mario.deluca@unina.it*

ships between the injury severity levels and crash-related factors. He then conducts a sensitive analysis on the trained neural network models to identify the importance of crash-related factors applied to different injury severity levels. In the process, the problem of five-class prediction is decomposed into a set of binary prediction models (using a national representative sample of 30358 police-recorded crash reports) to obtain the information useful to identify the 'true' cause and effect relationships between the crash-related factors and different levels of injury severity. The results, validated by previous studies, show the changing importance of crash factors with the changing injury severity levels.

Abdel-Aty and Abdelwahab (2004) investigated the use of two well-known ANN paradigms: the MultiLayer Perceptron (MLP) and fuzzy adaptive resonance theory neural networks for analysing the severity of driver injury. The objective of this study was to investigate the viability and potential benefits of using ANN to predict driver injury severity if a crash occurs. ANN performance was compared with a calibrated ordered probit-model. Modelling results showed that testing classification accuracy was 73.5% for MLP, 70.6% for fuzzy adaptive resonance theory map, and 61.7% for the ordered probit-model. This result indicates more accurate injury severity prediction capability for the ANN (particularly MLP) compared with other traditional methods.

1. Techniques Used in Data Analysis

Three different types of techniques are used for the analysis in this study: the cluster analysis, applied by the algorithm 'hard c-means', groups accidents with the same characteristics, while the ANN technique and the MultiVariate Analysis (MVA) technique are used to construct, after aggregating the data, two models of road accidents (Model 1 with MVA and Model 2 with ANN). The next paragraphs will describe the basic principles that characterized the two models.

1.1. Hard C-Means Technique

The principles of this technique are as follows. The aim of the group analysis consists in identifying a specific *U* partition, in c groups ($2 \leq c \leq$ n) of the *U* collection space constituted by *n*-elements. The hypothesis upon which this method is based is the following: the elements of the *X* space, which belong to a group, are characterized by a mathematical affinity and this affinity is greater than the elements of the different groups. Each element in the sample can be schematized, as a point identified by *m*-coordinates, and each coordinate constitutes an attribute of the same element. One of the simpler measures of affinity is represented by the distance measured between two points, and these belong to the data-space. Author define an appropriate measurement for distance, and author measure this between each unit of observation and all the units as a whole. Of course, the distance between points belonging to the same group is smaller than the distance between points contained in different groups. Let $X = \{x_1, x_2, x_3, \dots x_n\}$, the set of *n* data to be divided into *c* groups. Each element x_i, is defined by m characteristics ($x_i = \{x_1, x_2, x_3, \dots x_{im}\}$). For this reason x_i (where x_i represents the accident *i*) can be represented by a point on the *Rm* space.

This method is based on the use of a *J* objective function that tends to create 'spherical' groups for successive approximations. The objective function follows two results simultaneously: firstly it minimizes the Euclidean distance between the points of each group and the center of the same group (which generally does not coincide with any of the collection points), and in the second place, it maximizes the Euclidean distance between the centers of all the groups, *U* indicates the generic partition and U^* is the optimum that belongs to the *Mc* space of the possible partition of *X*. The $J = J(U)$ value, assumed by the objective function for each *U* partition, constitutes a relative measure of how close it is to the optimum.

The objective function is to minimize the square addition of the Euclidean distances measured between all points and the center of each group. It is difficult to find the *U* partition because the cardinality of the *Mc* space of *X*'s possible partitions tends rapidly to infinity. The search for the global optimum in problems of significant dimensions is not possible without laborious computation so the problem is resolved using an iterative optimization algorithm. Hypothesizing the first attempt with a $U(r = 0)$ partition, number *c* groups and an iteration tolerance value ε (accuracy required for the solution) the position of the group center can be determined. Starting from these, author calculate again the attribution of each point to the different groups, and author obtain a new calculation for the matrix $U(r = 1)$. Then author compare the two successive determinations of the *U* matrix and author repeat the process until the difference between the partitions, obtained over two successive cycles exceeds the predefined level of tolerance.

This technique presupposes that the number (Dell'Acqua *et al.* 2012) of clusters is known beforehand, but as the optimum number of clusters with which to make the definitive classification is not known (this is due to the substantial lack of initial information on the structure of the clusters within which the units of observation are to be placed), author proceeded at random. Author hypothesized different divisions of the database and then chose a value for an *S* index, defined as the best grouping index (Čokorilo *et al.* 2014).

1.2. The ANN Multilayer Approach

Inspiration for the structure of the ANN is taken from the structure and operating principles of the human brain. It is made of interconnected artificial neurons that mimic some properties of biological neurons. The function of a biological neuron is to add its input and produce an output.

This output is transmitted to subsequent neurons, through the synoptic joints, only if the transmitted signal is high (i.e., greater than a predetermined value), otherwise, the signal is not transmitted to the next neuron. In the network, therefore, a neuron calculates the weighted sum, using Eq. (1) (considering the input x_i

and weights w_i) and compares it with a threshold value; if the sum is greater than the threshold value, the neuron lights up and the signal is transmitted. Otherwise, the neuron does not turn on and the flow stops:

$$I = \sum_{i=1}^{n} w_i \cdot x_i, \quad (1)$$

where: I is the weighted sum [dimensionless]; w_i is the weight [dimensionless]; x_i is the input [dimensionless].

The activation value u_i rather than u_j, connected to weight W_{ij}, is a function of the weighted sum of the input. This function may take various forms. In this study, a function of type (Eq. (2)) was used:

$$u_j = \frac{1}{1 + e^{-\left(\sum(i) w_{ij} \cdot u_i + \theta_j\right)}}, \quad (2)$$

where: θ_j is the bias unit [dimensionless]; u_i is the degree of sensitivity of u_j when it receives an input signal from u_j [dimensionless]; w_{ij} is the weight between the connection of the neuron i with the neuron j [dimensionless].

The algorithm used for ANN application is the MLP – back propagation (Žilionienė *et al.* 2014).

2. Data Collection in a GIS Environment

The segments analysed belong to the SS 106 situated in southern Italy (Fig. 1).

The analysed stretches are shown in Table 1. For each stretch, data concerning geometry, traffic and accidents (Dell'Acqua *et al.* 2011) were collected in the period between 01/01/2000 and 12/31/2005 (see data-table in Fig. 2).

The data were stored (Carrion *et al.* 2009) in a suitable trained Geographic Information System (GIS). The road axis was constructed in a CAD environment with an appropriate layer, and each planimetric element of the road was separated. Cross sections were inserted into another layer (*step* = 50 m and at known points, i.e., curve center, curve end, etc.). A file in dxf-format, containing the road axis and sections, was imported into ARCGIS, geo-referenced according to UTM coordinates, and was converted into the format used in ARCGIS (i.e. shp-file format). Finally, the characteristics of each element of the alignment (Fig. 2, the data-table) were loaded into the GIS environment.

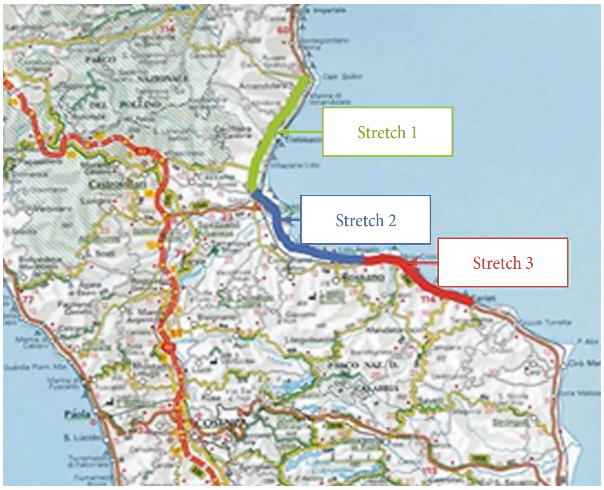


Fig. 1. Analysed segments

Table 1. Analysed stretches

Stretch	Start distance [km]	End distance [km]
1	0.00	28.00
2	28.00	59.00
3	59.00	90.00

All other information regarding the characteristics of accidents, traffic and vertical alignment was included in cross-section layers. Fig. 2 shows an extract of the procedure. In particular, 'the road axis' is schematically shown in the GIS environment where each object (planimetric element, cross section, etc.) is shown in the attribute table (one row for each object).

3. Cluster Analysis Application

Cluster analysis was applied to the data matrix loaded into the GIS system shown in Fig. 2 (in particular to the variables shown in Table 2). This technique allows to aggregate accidents in groups (i.e. cluster) with a high level of affinity (e.g. the accidents take place in the same geometric situation, the same environment situation, etc.). The best aggregation was obtained in 19 groups.

Each of the 19 groups shown in Table 3, obtained by calculating the mean value of the group, can be considered as a black spot.

In particular, the length of the black spot (indicated by the acronym *Li*) was calculated as an area of influence (Dell'Acqua *et al.* 2013) for each accident that occurred at the same distance. The sum of the areas of influence was assumed as the length (*Li*) of the black spots (where two or more accidents happen with the same distance, only the area of influence was considered).

4. MVA Application (Model 1)

The Model 1 was obtained using a MVA; the structure of the model (Highway Safety Manual 2009) and the variables used are listed:

$$Ni \cdot Sev = \left(365 \cdot AADT \cdot Li\right) \cdot b1 + e^{\left(Curv \cdot b2 + LG \cdot b3 + SP \cdot b4 + WR \cdot b5 + UorNU \cdot b6\right)},$$

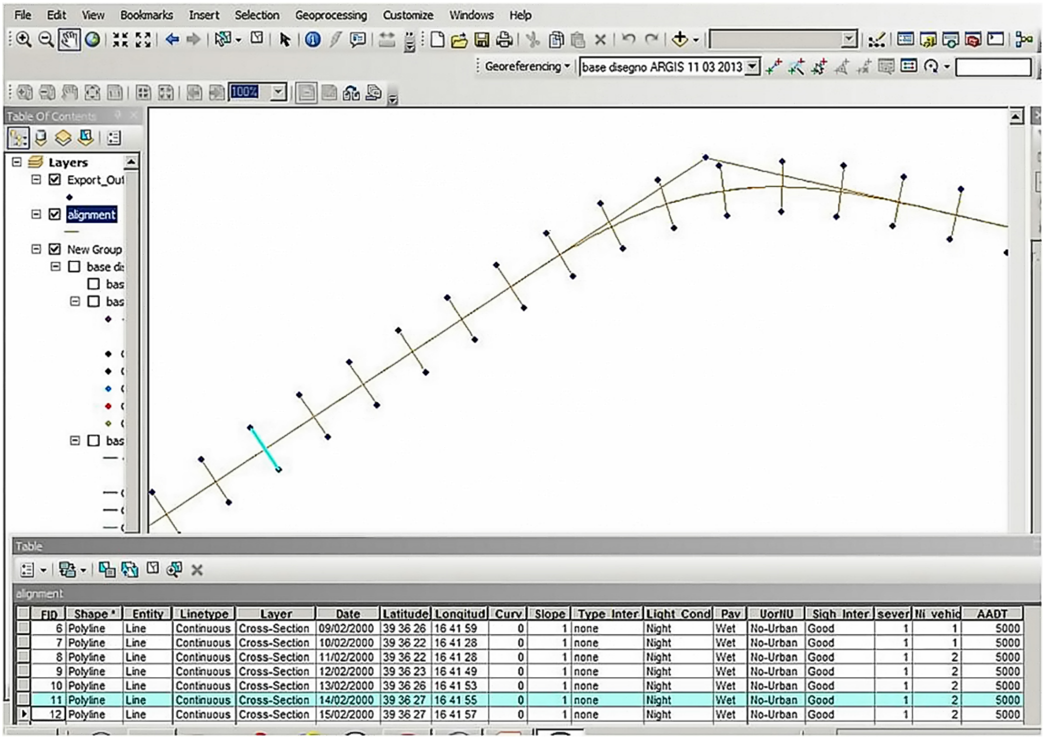
$$r^2 = 0.758, \quad (3)$$

where: *Li*, *Curv*, *LG*, *SP*, *WR*, *UorNU* – predictors; *Ni·Sev* – dependent variable.

The results of the multiple non-linear regression (De Luca, Dell'Acqua 2012) are shown in Tables 4 and 5.

5. ANN Application (Model 2)

The Model 2 was obtained using the ANN technique (Jin *et al.* 2002) shown in Chapter 1 using the same variables used in the Chapter 4. In particular, Model 2 was obtained using 70% of dataset for training and 30% for testing. Fig. 3 shows the ANN architecture while the parameters of the ANN estimates are shown in Table 6.



FID	Shape *	Entity	Linetype	Layer	Date	Latitude	Longitud	Curv	Slope	Type Inter	Light Cond	Pav	UorNU	Sigh Inter	sever	Ni vehic	AADT
6	Polyline	Line	Continuous	Cross-Section	09/02/2000	39 36 26	16 41 59	0	1	none	Night	Wet	No-Urban	Good	1	1	5000
7	Polyline	Line	Continuous	Cross-Section	10/02/2000	39 36 22	16 41 28	0	1	none	Night	Wet	No-Urban	Good	1	1	5000
8	Polyline	Line	Continuous	Cross-Section	11/02/2000	39 36 22	16 41 28	0	1	none	Night	Wet	No-Urban	Good	1	2	5000
9	Polyline	Line	Continuous	Cross-Section	12/02/2000	39 36 23	16 41 49	0	1	none	Night	Wet	No-Urban	Good	1	2	5000
10	Polyline	Line	Continuous	Cross-Section	13/02/2000	39 36 26	16 41 53	0	1	none	Night	Wet	No-Urban	Good	1	2	5000
11	Polyline	Line	Continuous	Cross-Section	14/02/2000	39 36 27	16 41 55	0	1	none	Night	Wet	No-Urban	Good	1	2	5000
12	Polyline	Line	Continuous	Cross-Section	15/02/2000	39 36 27	16 41 57	0	1	none	Night	Wet	No-Urban	Good	1	2	5000

Fig. 2. Data collection in the GIS

Table 2. Variables used in the cluster analysis

Variables	Label	Variable types	Variable codes	Range	Units
Curvature (1/R)	*Curv*	numeric	–	[0.04, 0.25]	1/m
Longitudinal grade [%]	*LG*	numeric	–	[–0.04, 0.67]	%
State of paving	*SP*	non-numeric	1.0 = dry 2.0 = wet	[1.0, 2.0]	dimensionless
Urban or Non-urban	*UorNU*	non-numeric	1.0 = Non-urban 2.0 = Urban	[1.0, 2.0]	dimensionless
Width road	*WR*	numeric		[6.80, 12.80]	m

Table 3. Results obtained with the 'hard c-means'

Cluster	Average distance	*Curv*	*LG*	*SP*	*WR*	*UorNU*	Average Severity	*Ni* [No of accidents]	*Li* [km]
R	23.01	0.0003	0.20	1.00	12.80	1.00	1.38	108	1.30
I	20.83	0.0003	0.11	1.00	12.20	1.00	1.35	34	1.30
B	6.99	0.0000	–0.28	1.00	10.60	2.00	1.16	55	1.20
D	9.68	0.0001	–0.18	1.27	8.50	2.00	1.36	23	1.30
J	11.47	0.0003	–0.32	1.00	7.90	2.00	1.41	63	1.40
P	16.43	0.0002	0.02	2.00	11.20	1.00	1.32	38	1.40
M	322.68	0.0001	0.00	1.11	6.80	2.00	1.17	19	1.30
F	14.49	0.0004	0.75	1.08	7.50	1.25	1.29	23	0.60
A	321.99	0.0001	0.00	2.00	6.80	2.00	1.40	27	1.30
L	264.95	0.0001	0.10	2.00	7.40	1.07	1.47	26	0.80
G	320.31	0.0001	0.00	1.00	7.00	2.00	1.40	107	1.40
E	339.15	0.0000	0.16	1.00	8.70	1.00	1.33	50	1.40
S	12.39	0.0000	–0.67	1.00	8.90	1.00	1.31	57	1.40
T	356.08	0.0001	–0.04	1.00	6.90	1.00	1.33	78	0.60
H	13.01	0.0003	0.43	1.00	7.50	1.00	1.42	79	0.60
K	8.73	0.0003	0.64	1.00	7.50	1.00	1.43	16	1.40
C	357.91	0.0001	–0.01	1.27	6.90	1.20	1.40	25	0.60
Q	114.13	0.0000	0.23	1.00	7.30	2.00	1.50	18	0.60

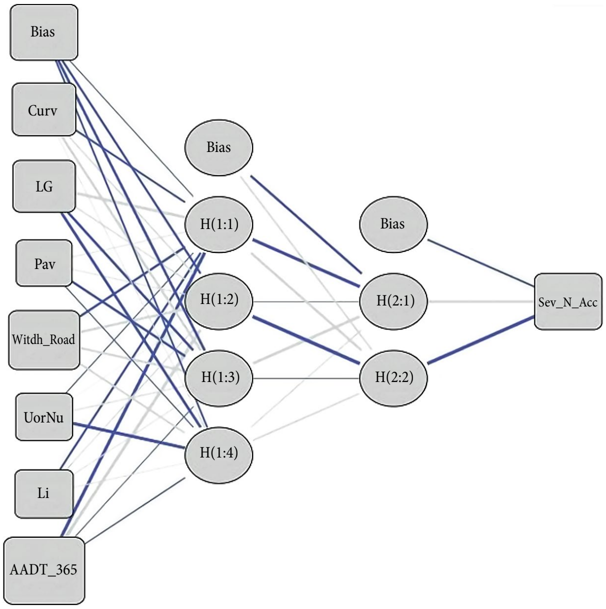


Fig. 3. ANN architecture

Table 4. Parameter estimates for the 'non-linear model'

Parameter	Estimate	Std. error	95% Confidence interval	
			Lower bound	Upper bound
*b*1 (*Li·AADT*)	1.680E–5	0.000	1.15E–5	2.20E–5
*b*2 (*Curv*)	3090.750	2326.067	–1977.31	8158.81
*b*3 (*LG*)	–0.202	1.066	–2.52	2.12
*b*4 (*SP*)	–0.630	0.952	–2.70	1.44
*b*5 (*WR*)	0.274	0.077	0.10	0.44
*b*6 (*UorNU*)	0.066	0.437	–0.29	1.62

Table 5. ANOVA test for 'non-linear model'

Source	Sum of squares	*df*	Mean squares
Regression	95285.31	6	15880.88
Residual	7015.91	12	584.65
Uncorrected total	102301.22	18	–
Corrected total	28995.02	17	–
Dependent variable: *Ni·Sev*; *R*-squared (r^2) = 1 – (*Residual sum of squares*) / (*Corrected sum of squares*) = 0.758.			

Table 6. Parameters of the ANN Estimates

Predictor		Predicted						
		Hidden layer 1				Hidden layer 2		Output layer
		H(1:1)	H(1:2)	H(1:3)	H(1:4)	H(2:1)	H(2:2)	Ni·Sev
Input layer	(Bias)	–0.112	–0.339	–0.423	–0.230			
	Curv	–0.278	0.215	0.475	0.140			
	LG	0.806	0.212	–0.388	–0.462			
	Pav	0.157	0.162	–0.378	–0.114			
	Road width	–0.363	0.482	0.449	0.343			
	UorNU	–0.112	0.014	0.193	–0.686			
	Li	–0.299	0.183	0.153	0.143			
	AADT_365	–0.721	1.443	–0.043	–0.145			
Hidden layer 1	(Bias)					–0.381	0.290	
	H(1:1)					–0.687	0.604	
	H(1:2)					–0.007	–1.267	
	H(1:3)					0.927	–0.002	
	H(1:4)					0.200	0.286	
Hidden layer 2	(Bias)							–0.247
	H(2:1)							0.665
	H(2:2)							–1.142

6. ANN Versus the MVA

Table 7 shows the comparison between the two models and it denotes that Model 2 is better than Model 1 because the residual has a lower total sum (see last two columns of Table 7). However Model 1 gives a better estimate for the more dangerous black spots (see Table 7, lines from 1 to 5). In Table 7 it is possible to see that ANN technique in the cluster *R*, *B* and *I* has higher residuals than MVA technique. This represents a limit for ANN technique, which is valid for a net global analysis, but does not result confident for the most dangerous clusters.

Table 7. Comparison model

Cluster ID	Col. 1 *Ni·Sev* observed	Col. 2 *Ni·Sev* estimated R.N.L (Mod 1)	Col. 3 *Ni·Sev* estimated ANN (Mod 2)	Residual 1 (Col. 1 – Col. 2)	Residual 2 (Col. 1 – Col. 3)
R	149.04	147.26	134.89	1.78	14.15
I	45.9	53.34	61.83	7.44	15.93
B	63.8	58.39	50.03	5.41	13.77
D	31.2	27.9	34.43	3.3	3.23
J	88.8	90.33	70.52	1.53	18.28
P	50.1	51.88	51.35	1.78	1.25
M	22.23	17.72	26.02	4.51	3.79
F	29.67	19.63	24.81	10.04	4.86
A	37.8	25.49	25.22	12.31	12.58
L	38.22	16.74	24.64	21.48	13.58
G	149.8	151.9	124.83	2.1	24.97
E	66.5	84.22	60.47	17.72	6.03
S	74.6	98.71	110.2	24.11	35.6
T	103.7	60.61	107.1	43.09	3.4
H	112.1	85.76	112.06	26.34	0.04
K	22.8	46.87	25.68	24.07	2.88
C	35	35.56	27.32	0.56	7.68
Q	27	40.62	25.16	13.62	1.84
			Residual sum:	221.19	183.86

Conclusions

This study shows the results of prediction road accidents comparing Artificial Neural Network (ANN) technique and MultiVariate Analysis technique (MVA). This study considers the accidents occurring on a two-lane highway (SS 106 located in southern Italy) between 01/01/2001 and 31/12/2005.

The data, aggregated by cluster analysis (using the algorithm binary partition 'hard c-mean') are worked out with MVA and ANN techniques and two models have been obtained: Model 1 (MVA model) and Model 2 (ANN model).

Comparing two models, it is evident that Model 2 is better than Model 1 because it has a lower total residual, although Model 1 seems better to characterize the most dangerous black spots (see row 1 to 5 in Table 7).

Tests are currently trying to transfer this methodology to other roads with similar characteristics.

Although in the initial phase, the assessment is providing very interesting results.

Acknowledgments

This study was financed by the University of Naples Federico II as a part of the Agreement of Cooperation with the University of Belgrade (21 September 2012).

References

Abdel-Aty, M.; Abdelwahab, H. 2004. Predicting injury severity levels in traffic crashes: a modeling comparison, *Journal of Transportation Engineering* 130(2): 204–210. http://dx.doi.org/10.1061/(ASCE)0733-947X(2004)130:2(204)

Carrion, D.; Maffeis, A.; Migliaccio, F. 2009. A database-oriented approach to GIS designing, *Applied Geomatics* 1(3): 75–84. http://dx.doi.org/10.1007/s12518-009-0008-y

Chang, L.-Y. 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network, *Safety Science* 43(8): 541–557. http://dx.doi.org/10.1016/j.ssci.2005.04.004

Chiou, Y.-C. 2006. An artificial neural network-based expert system for the appraisal of two-car crash accidents, *Accident Analysis & Prevention* 38(4): 777–785. http://dx.doi.org/10.1016/j.aap.2006.02.006

Čokorilo, O.; De Luca, M.; Dell'Acqua, G. 2014. Aircraft safety analysis using clustering algorithms, *Journal of Risk Research* 17(10): 1325–1340. http://dx.doi.org/10.1080/13669877.2013.879493

Delen, D.; Sharda, R.; Bessonov, M. 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks, *Accident Analysis & Prevention* 38(3): 434–444. http://dx.doi.org/10.1016/j.aap.2005.06.024

Dell'Acqua, G.; De Luca, M.; Mauro, R. 2011. Road safety knowledge-based decision support system, *Procedia – Social and Behavioral Sciences* 20: 973–983. http://dx.doi.org/10.1016/j.sbspro.2011.08.106

Dell'Acqua, G.; De Luca, M.; Russo, F. 2012. Procedure for making paving decisions with cluster and multicriteria analysis, *Transportation Research Record* 2282: 57–66. http://dx.doi.org/10.3141/2282-07

Dell'Acqua, G.; Russo, F.; Biancardo, S. A. 2013. Risk-type density diagrams by crash type on two-lane rural roads, *Journal of Risk Research* 16(10): 1297–1314. http://dx.doi.org/10.1080/13669877.2013.788547

De Luca, M.; Dell'Acqua, G. 2012. Freeway safety management: case studies in Italy, *Transport* 27(3): 320–326. http://dx.doi.org/10.3846/16484142.2012.724447

Highway Safety Manual. 2009. 1st edition. 1057 p. Available from Internet: http://www.este.civ.uth.gr/apodeltiosi/HSM.pdf

Jin, X.; Cheu, R. L.; Srinivasan, D. 2002. Development and adaptation of constructive probabilistic neural network in freeway incident detection, *Transportation Research Part C: Emerging Technologies* 10(2): 121–147. http://dx.doi.org/10.1016/S0968-090X(01)00007-9

Vujanić, M.; Lipovac, K.; Jovanović, D.; Pešić, D.; Antić, B. 2013. 'Bottom-up' and 'top-down' approach for defining road safety strategy – case study: city of Belgrade, *International Journal for Traffic and Transport Engineering* 3(2): 185–203. http://dx.doi.org/10.7708/ijtte.2013.3(2).07

Žilionienė, D.; De Luca, M.; Dell'Acqua, G.; Lamberti, R.; Biancardo, S. A.; Russo, F. 2014. Evaluating freeway traffic noise using artificial neural network, in *9th International Conference on Environmental Engineering: Selected Papers*, 22–24 May 2014, Vilnius, Lithuania, 1–9. http://dx.doi.org/10.3846/enviro.2014.182