Sigma Journal of Engineering and Natural Sciences
Sigma Mühendislik ve Fen Bilimleri Dergisi

Research Article

ARTIFICIAL NEURAL NETWORKS RESTRICTION FOR ROAD ACCIDENTS SEVERITY CLASSIFICATION IN UNBALANCED DATABASE

Maria Lígia CHUERUBIM*[1], Alan VALEJO[2], Barbara Stolte BEZERRA[3], Irineu da SILVA[4]

[1]*Faculty of Civil Engineering, Federal University of Uberlândia, BRAZIL;* ORCID: 0000-0002-2019-9198
[2]*Institute of Mathematical and Computer Sciences, University of São Paulo, BRAZIL;*
ORCID: 0000-0002-9046-9499
[3]*Faculty of Civil Engineering, UNESP São Paulo State University, BRAZIL;* ORCID: 0000-0001-5775-6683
[4]*Department of Transport Engineering, School of Engineering of São Carlos, University of São Paulo, BRAZIL;* ORCID: 0000-0002-8459-4664

Received: 25.10.2018 Revised: 04.02.2019 Accepted: 05.08.2019

ABSTRACT

The objective of this study is to discuss the main constraints in classifying the severity of road accidents using Artificial Neural Networks (ANN). To achieve this, ANN modelling with Multiple Layers Perceptron (MPL) was used. This method is recommended for treating non-linear problems, whose distributions are not normal, which is the case for road accidents. Variables associated with the characteristics of accidents, road infrastructure and environmental conditions were used, with the objective of identifying the influence of these factors in the accident severity. The results indicated that ANN modelling with MPL presents a potential association among the parameters related to road accidents. However, the results are limited, since the classification process provides a low rate of accuracy for accidents with victims. Such accidents correspond to less frequent observations in the database, meaning that the data is less represented, and the database becomes unbalanced. Thus, for further research studies, the use of ANN with MPL associated with data balancing methods is suggested, in order to obtain the best data fit to the model and more consistent and realistic results.

Keywords: Unbalanced data, road accidents, severity, classification, artificial neural networks.

1. INTRODUCTION

In order to carry out measures of prevention and reduction of road accident severity, it is necessary to perform an analysis of the occurrence of road accidents. Such analysis would include associating the record of each accident occurrence with the characteristics of the road, the environment, the vehicular conditions and the driver. These characteristics, when related to accident records, can be used in the prediction and classification of road accident severity [1].

Several methods about road accident analysis can be found in the literature. The principal methods are based on deterministic models and regression analysis, such as logistic regression, negative binominal regression and Poisson regression, which investigate the individual contribution of a specific factor in the variation of road accident severity [2, 3]. Al-Ghamdi [4]

[*] Corresponding Author: e-mail: marialigia@ufu.br, tel: +55 34 3239-4411

applied the logistic regression to analyze the injury severity in 560 accident occurrences involving severe (fatal and non-fatal) victims, in Riyadh, the capital of Saudi Arabia. Farmer *et al.* [5] used logistic regression to analyze the injury severity in shock type accidents involving passenger vehicles and light trucks in the United States for a period of 4 years (1988 to 1992). Lui *et al.* [6] quantified through logistic regression the relationships between the use of safety belts, the direction of impacts in crash accidents, and the weight of vehicles, and the effects of these factors on accidents involving fatalities. Singleton, Qin & Luan [7] used logistic regression to identify driver, vehicle and pathway/environment factors associated with the increased risk of serious injury in Kentucky from 2000 to 2001. Negative binominal regression was used by Pirdavani, Brijs & Bellemans [8] to assess the impact of traffic safety on the basis of fuel price increases for the period 2004-2007 in Flanders (Belgium) in 2,200 traffic zones. Poisson regression was used by Ye *et al.* [9] in the generation of accident severity frequency rate models for highway sections using five years of data for 275 multilane road segments in Washington State. Debrabant *et al.* [10] used the same regression model for the identification of critical areas of road accidents in Funen, Denmark, for the period from 2002 to 2007.

Regression models provide analytical results, based on a predetermined mathematical function. If the function requirements are violated, erroneous probabilities are obtained for accident severity. However, traffic accidents are multicausal events of random nature and cannot be described by simple deterministic models [2]. Therefore, probability models that describe the random behavior of the entities present in the accident database should be used, in order to convey a greater number of parameters associated to the road accident phenomena [2].

In this perspective, stochastic and non-parametric approaches, derived from artificial intelligence, have been employed in the treatment of heterogeneous and multidimensional databases [2, 3, 11, 12, 13, 14]. Among them are those based on data mining, in which there is no need to establish a priori restrictions on the relationships of the database variables.

Data mining techniques used in investigating accident severity fundamentally use tree structures and networks. The techniques based on tree structures, known as decision trees, restrict the results obtained due to the hierarchical and binary structure of the tree. This feature limits the application of tree structure techniques to the analysis of specific categories of a target variable (dependent) belonging to the database, being ineffective for the modelling of multiple problems [11, 15,16].

Data mining techniques based on network structures allow a greater extraction of knowledge from the database. Since a larger number of connections can be made between different types of data, the modelling process can be more flexible [2, 17]. Other studies in this area highlight the use of Artificial Neural Networks (ANN) [3, 12, 18, 19, 20, 21] and Bayesian networks [22, 23, 24, 25] when dealing with non-linear problems in which data do not present a normal distribution [26].The results obtained in the previous studies depict the effectiveness of network modelling in the treatment of large and heterogeneous databases, with emphasis on the prediction and classification process [2].

Considering that ANN have been successfully employed in non-linear problems with a non-normal distribution, the use of ANN in the evaluation of accident database, which has such characteristics, is a feasible hypothesis. However, some limitations emerge regarding the severity classification in unbalanced databases, where fatal and severe accidents are underrepresented in the total number of accident occurrences.

Based on this problem, the scope of this research is to present a discussion about the restriction of ANN for road accident severity classification, using unbalanced traffic accidents database. The most significant variables were selected, in order to correlate the degree of injury with the main risk factors in the road environment.

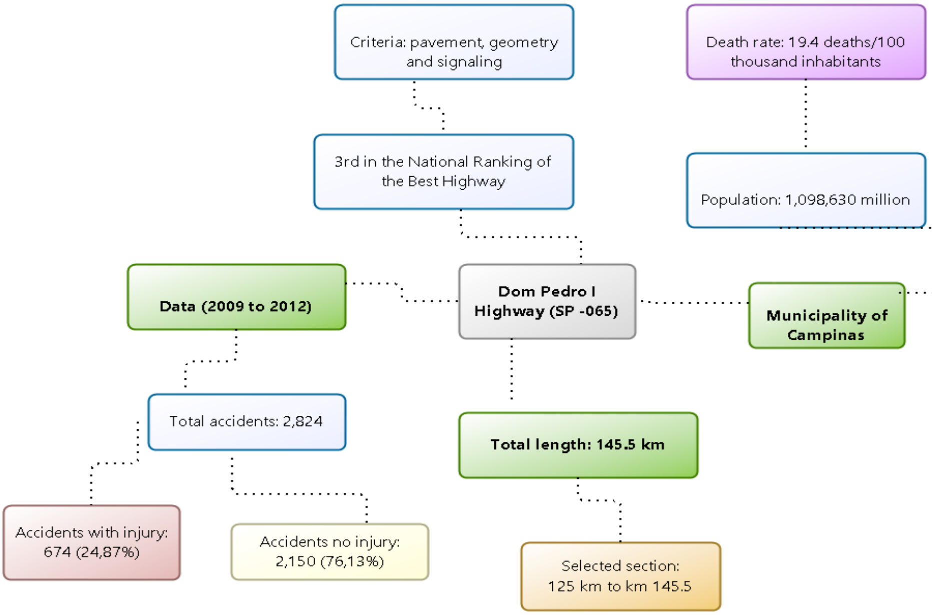
This paper is organized as follows. Section 2 presents the materials and method adopted in the research, with a brief description of variable selection and learning based classification on ANN with multiple layers. Section 3 presents the experiments and discusses the results obtained.

Section 4 presen ts the conclusions about the approach employed, as well as the recommendations for future work.

2. MATERIALS AND METHOD

2.1. Road accident database

The traffic accident database used in this research consists of accidents that occurred between 2009 and 2012 from the km 125 to km 145.5 of Dom Pedro I Highway (SP-065), located in the city of Campinas (Brazil). During the period considered, 2,824 accidents occurred, excluding missing observations, incomplete data and the period of road construction (97.04% of the original data was used). Figure 1 presents a framework with the main characteristics of the study area.



Criteria: pavement, geometry and signaling

Death rate: 19.4 deaths/100 thousand inhabitants

3rd in the National Ranking of the Best Highway

Population: 1,098,630 million

Data (2009 to 2012)

Dom Pedro I Highway (SP -065)

Municipality of Campinas

Total accidents: 2,824

Total length: 145.5 km

Accidents with injury: 674 (24,87%)

Accidents no injury: 2,150 (76,13%)

Selected section: 125 km to km 145.5

Figure 1. Framework with the main characteristics of the study area.

The database contains the following types of information: Accident Type (ACT), Weather Condition (WTC), Visibility (SGC), Road Profile (PFR), Road Geometry (GER), Pavement Condition (PAV), Period of the Day (PER), Accident Cause (ACC), Horizontal Signal (HS), Vertical Signal (VS) and Milestone (km). These eleven variables were selected to classify the accident severity with (non-fatal and fatal) and without victims and are broken down into categories as shown in Figure 2.

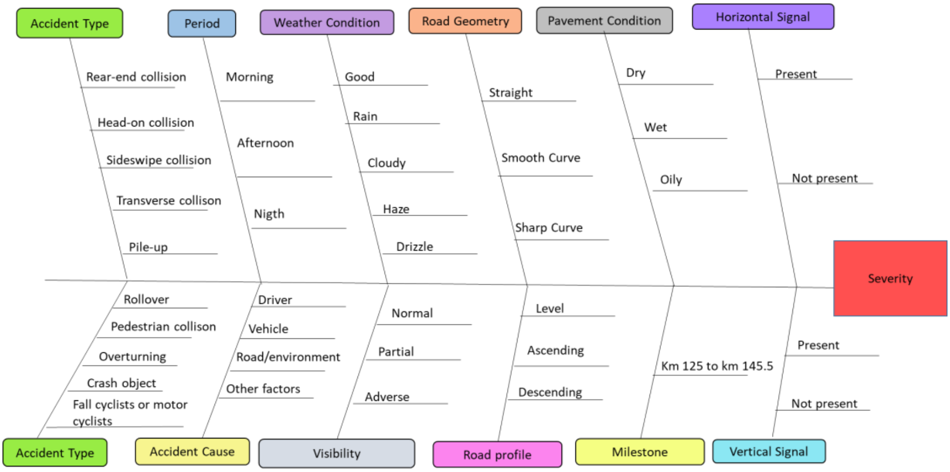


Figure 2. Variables selected for accident severity classification.

Of the 2,824 accidents analyzed, 23.87% represent accidents with victims (fatal and non-fatal) and 76.13% correspond to accidents without victims. Thus, the database is unbalanced, i.e., the number of accidents involving victims will always correspond to minority class of data.

2.2. Artificial neural networks learning

The method used to construct the model is based on Artificial Neural Networks (ANN). This technique is widely used in linear and non-linear problems since it does not require a priori assumptions to be established between the variables. ANN are indicated in the treatment of phenomena that are not well known or that are derived from multiple factors, in which the traditional analytic approach would demand a high computational cost [13].

In practice, ANN basically consists of three layers: the input layer, the hidden layer, and the output layer. Initially, the random extraction of two subsets of data was performed. The principle of ANNs is minimize the Mean Squares Error (MSE) provided by equation:

$$MSE = \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{j=1}^{K} (t_{i,j} - a)^{2}, \tag{1}$$

where t and a are the observed and estimated parameters, respectively; K is the number of neurons of the output layer and N the number of test set data (Chang, 2005).

In this work, the SPSS software was used to construct ANN with Multiple Layers Perceptron (MLP), based on the "back propagation” algorithm, which corrects the weights in all layers, starting from the output layer to the input layer, using the mean square root error, in the two phases: the phase forward and the phase for backward. In the phase forward, each variable of database is stored in a neuron of the network. Na fase *forward*, cada variável da base de dados é armazenada em um neurônio da rede. A ponderação inicial se dá pelas conexões ou relações existentes entre as variáveis da camada de entrada. Cada neurônio nessa camada aplica a função de ativação a sua entrada total e produz um valor de saída, que é utilizado como valor de entrada pelos neurônios da camada seguinte. Esse processo se dá iterativamente até que os neurônios da camada de saída produzam cada um seu valor de saída, que é então comparado ao valor desejado para a saída desse neurônio. A diferença entre os valores produzidos e desejados para cada neurônio da camada de saída expressa o erro do processo de modelagem dos dados (Facelli *et al.* 2011).

O valor do erro de cada neurônio da camada de saída é então utilizado na fase *backward*, no ajuste dos pesos de entrada. O ajuste ocorre da camada de saída até a camada intermediária. O ajuste dos pesos de uma MLP pelo algoritmo *back-propagation* pode ser obtido pela seguinte equação (Facelli *et al.* 2011):

$$w_{j,l}(t+1) = w_{j,l}(t) + \eta x^j \delta_l , \tag{2}$$

onde, $w_{j,l}$ representa o peso entre um neurônio l e o $j-ésimo$ atributo de entrada ou saída do $j-ésimo$ neurônio da camada anterior; δ_l o erro associado ao $l-ésimo$ neurônio; x^j corresponde a entrada recebida por esse neurônio, ou seja, o $j-ésimo$ atributo de entrada ou a saída $j-ésimo$ neurônio da camada anterior; e η a taxa de aprendizado da modelagem.

Nesta modelagem apenas os valores dos erros dos neurônios de saída são conhecidos. Desta forma, deve-se estimar o erro para as camadas intermediárias. No algoritmo *back-propagation*, o erro de um neurônio de uma camada intermediária é estimado como a soma dos erros dos neurônios da camada seguinte, cujos terminais de entrada estão conectados a ele. A ponderação ocorre pelo peso atribuído a estas conexões.

Deste modo, o cálculo do erro dependerá da camada em que se encontra o neurônio, como mostra a equação (3) e a equação (4), (Facelli *et al.* 2011):

$$\delta_l = f_a^{'} e_l , \text{ se } n_l \in camada_{saída} , \tag{3}$$

$$\delta_l = f_a^{'} \sum w_{l,k} \delta_k \text{ , se } n_l \in camada_{\text{int}\,ermediária} , \tag{4}$$

onde, n_l representa o $l-ésimo$ neurônio; $f_a^{'}$ a derivada parcial da função de ativação do neurônio; e e_l o erro quadrático cometido pelo neurônio de saída quando sua resposta (y_q) é comparada a desejada $(\widehat{f}_q)$, definido por:

$$e_l = \frac{1}{2} \sum_{q=1}^{k} (y_q - \widehat{f}_q)^2 . \tag{5}$$

O ajuste dos pesos é definido pela $f_a^{'}$, utilizando o gradiente descente da função de ativação. Essa derivada mede a contribuição de cada peso no erro da rede para cada variável utilizada na análise. Quando o valor da $f_a^{'}$ é positivo para um dado peso, isto indica que o peso atribuído a uma determinada variável está provocando um aumento da diferença entre os valores obtidos e desejados na modelagem. Neste caso, a magnitude do peso deve ser reduzida para minimizar o erro obtido. E, caso contrário, quando o valor da $f_a^{'}$ for negativo, o peso deve ser aumentado, para que os valores obtidos se aproximem dos valores esperados (Facelli *et al.* 2011).

Figure 3 shows the structure of the ANN with MLP used in the modelling of the database.

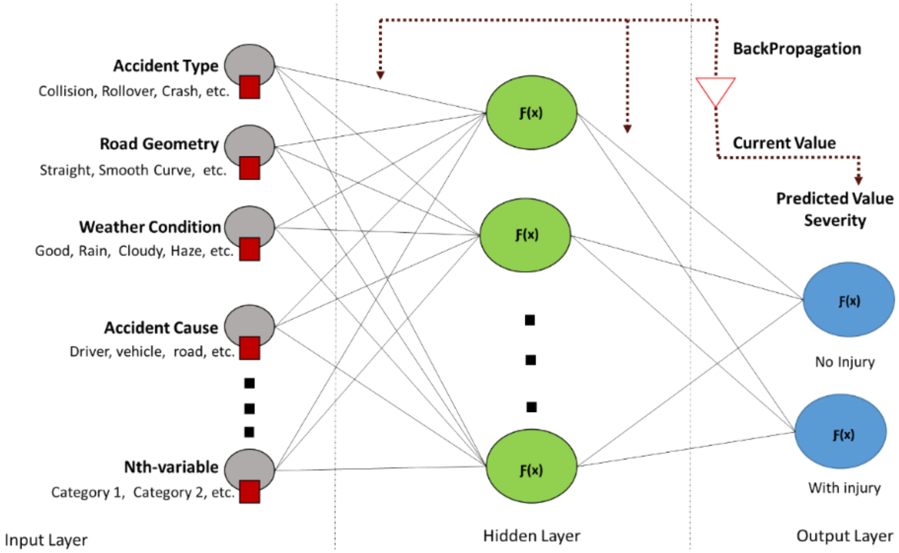


Figure 3. Illustration of ANN model with MLP.

The first subset uses 70% of the data for ANN training and the second subset uses 30% of the data for the ANN test. These subsets are used iteratively for learning, validation and cross-validation without repetition. At the end of the processing, the whole set of data was verified. These subsets should not be too large, because they may hinder the ANN learning process.

In addition, training and test subsets should be as homogeneous as possible for each class of variables, since ANN learning is based on as many combinations as possible. When many variables are used, or the data are not homogeneously distributed, the quality of ANN learning also tends to decline. Therefore, in the input layer of ANN a small number of variables must be selected, in order to obtain best results both in data extraction and in class distribution for the test and training subsets. In addition, the values of the input variables (predictors) must be normalized.

In this study, the input layer presents 39 neurons. The variable type of accident is represented by 10 neurons; the variable weather condition by 5 neurons; the variable accident cause by 4 neurons and the variable mileage by 1 neuron. The other variables, such as visibility, road profile, road geometry, pavement condition, and period are represented with 3 neurons each. Horizontal signal and vertical signal are represented with 2 neurons each.

The hidden layer contains the nodes or unobserved units and, in this study, has 9 neurons. The output layer contains two response neurons (accidents without victims and accidents with fatal and non-fatal victims). Therefore, each neuron belonging to the output layer represents a predicted variable, whereas the hidden layer neurons are directly connected to the values of the independent variables.

For the transfer of information between the input-layer neurons and the hidden layer, the hyperbolic tangent activation function was used, when the selection of the network architecture was performed automatically. In addition, for the transfer of information between the hidden layer and the output layer, the Softmax activation function was used, when all the independent variables are categorical.

The exact relationships between input and output data are not known, which implies that the optimal number of cases is not known for learning and for setting up the best ANN model.

The number of cases used in the network learning process was 2,824 accidents, of which 1,964 (69.5%) were used as a training subset and 860 (30.5%) as a test subset.

3. RESULTS AND DISCUSSION

Table 1 shows the amount of data classified following the ANN model with MPL for each category of injury accidents being accidents with no injury (NI) and accidents with injury (WI).

Table 1. Classification of accidents in relation to victims.

Sub-sets	Injury level*	Predict		
		NI	WI	(%) Correct
Training	NI	1,452	34	97.7%
	WI	460	18	3.8%
	(%) Total	100.00%	0.00%	74.8%
Test	NI	646	8	98.8%
	WI	182	14	7.1%
	(%) Total	97.4%	2.6%	77.9%

*no injury (NI) and with injury (WI).

As shown in Table 1 and in Figure 4, a two-class model (NI and WI) resulted in a general prediction of 77.9%. The results obtained in the classification process – using the ANN and the eleven predictor variables and the variables NI and WI.

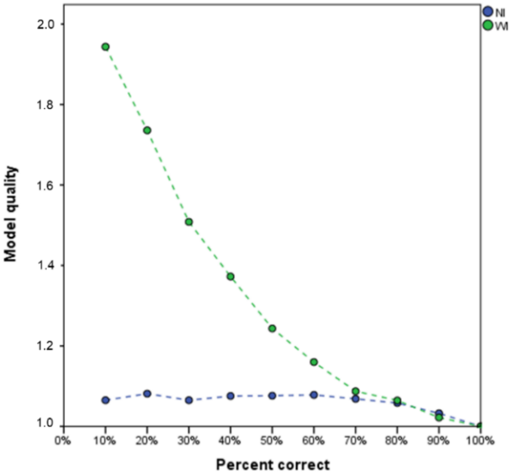


Figure 4. Quality of the model fit for the dependent variable “accident severity”.

However, ANN ranked fatalities (WI) with an accuracy of only 2.6%, while non-fatal accidents (NI) were classified with high accuracy (97.4%). This is a typical particularity of multiple variable problems, in which the database is naturally unbalanced, in relation to the number of observations and the number of variable categories under analysis [19]. Figure 5 shows the Receiver Operating Characteristic (ROC) curve and Figure 6 shows the gain for each dependent variable (NI and WI) as a function of the ROC curve.

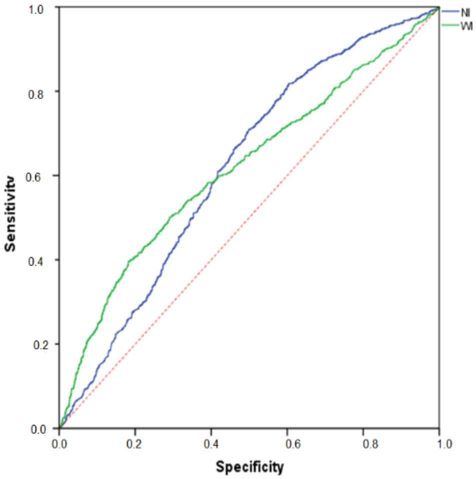


Figure 5. ROC curve obtained for the dependent variable “accident severity”.

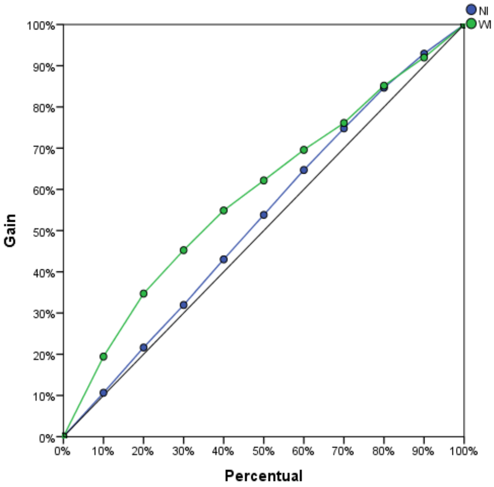


Figure 6. Gain graph for the dependent variable “accident severity”.

The ROC curve corresponds to the graphical representation of sensitivity (true positives) versus specificity (false positives), ranging from zero to one, and provides a measure of the discrimination of the model. When analyzing the ROC curve (Figure 5), the values predicted for the classes of the severity variable NI and WI should be very close to the diagonal, that is, close to one. However, the classification presented mean values of 0.618 for both NI predicted accidents and WI predicted accidents. Figure 6 shows the graph of cumulative gains for categorical variables NI and WI accidents, as a function of the ROC curve. Note that the gain balance between the two variables occurs when there is the best model fit (precision equal to 77.9%).

For this method, the variable accident type was the most important in the prediction process, followed by the variables accident cause, vertical signal, visibility and mileage. The other variables used in the modelling presented importance below 40%, as can be seen in Table 2. The percentage of uncorrected predictions for the training subset was 25.2% and for the test subset was 22.1%.

Table 2. Importance of independent variables.

Variables	Standardized Importance
ACT	100.0%
ACC	65.6%
PER	12.6%
SGC	44.0%
WTC	35.8%
PAV	33.1%
GER	23.5%
PFR	28.9%
HS	16.3%
VS	46.6%
km	42.7%

The Figure 7 shows that the four most important variables accident type, period, visibility, weather condition, road geometry, road profile, pavement condition, milestone, horizontal signal and vertical signal have a nonhomogeneous data distribution.

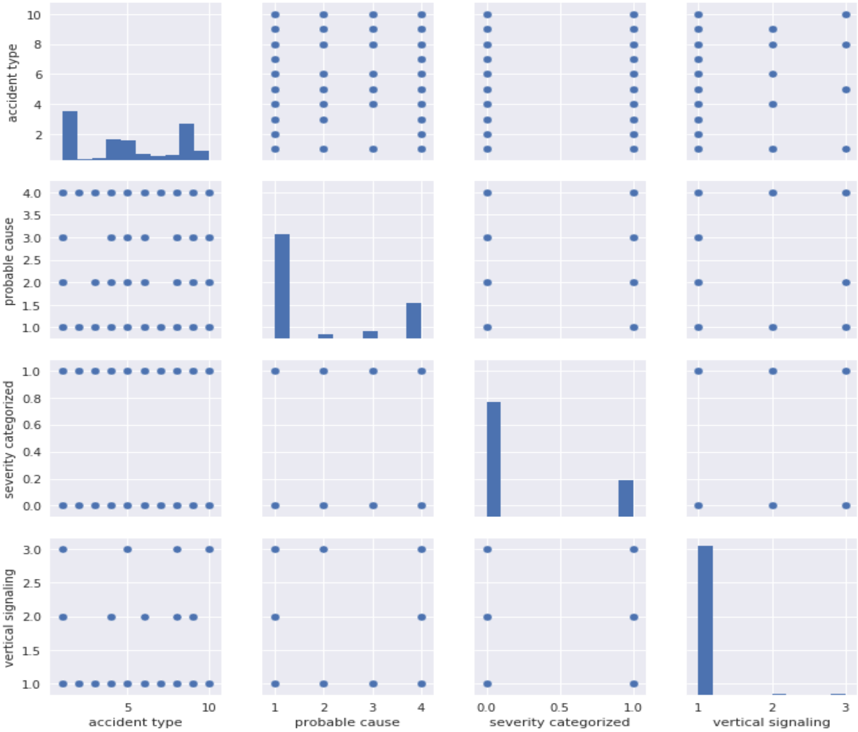


Figure 7. Nonhomogeneous data distribution of the main variables.

Figures 8 and 9 present, respectively, the scores obtained for the predictions of NI and WI accidents.

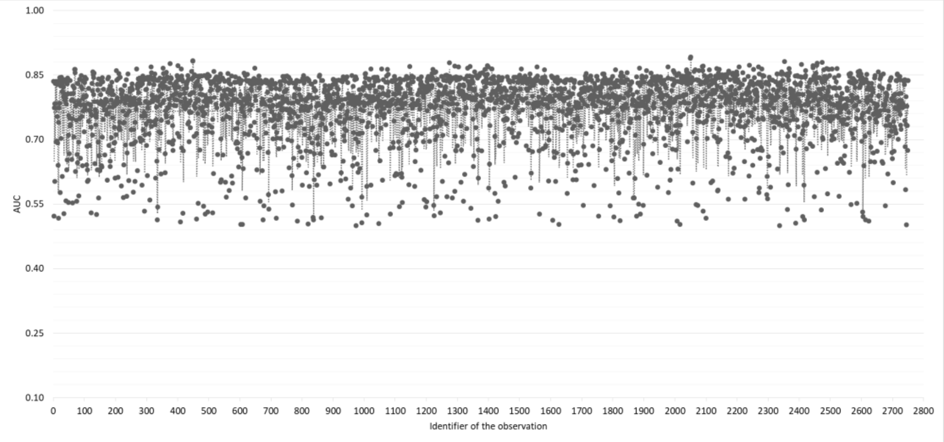


Figure 8. Scores obtained in the prediction of accidents with no injury (NI).

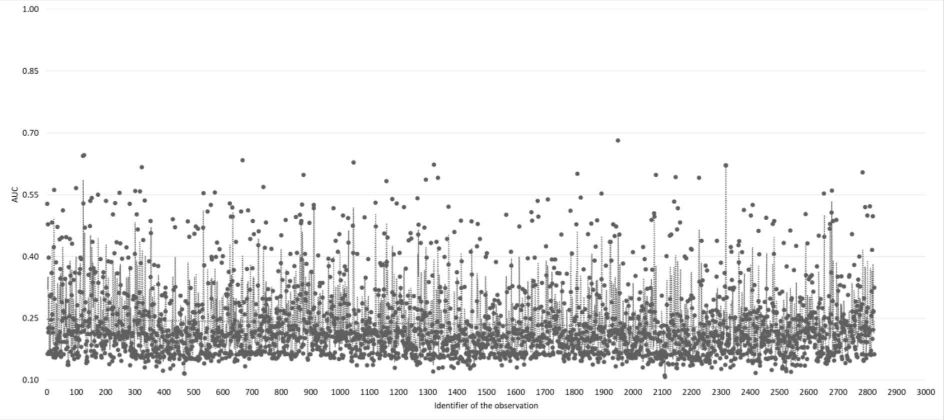


Figure 9. Scores obtained in the prediction of accidents with injury (WI).

The classification for NI accidents resulted in an average score of 0.76, with minimum value of 0.32 and maximum value of 0.89. While the classification for WI accidents resulted in an average score of 0.24, with minimum value of 0.11 and maximum value of 0.68. Therefore, in general terms, the prediction and classification of NI accidents presented the highest scores and, consequently, the highest probabilities of accuracy.

According to Hosmer and Lemeshow [27], values of $0.5 \leq AUC < 0.7$ provide non-discriminatory modeling. Values in the range of $0.7 \leq AUC < 0.8$ provide a model with acceptable discrimination, whereas values between $0.8 \leq AUC < 0.9$ indicate excellent modeling and $AUC \geq 0.9$ a model with extraordinary discrimination potential.

Thus, considering the limitations of ANN for unbalanced data [13, 19] and the predictions that demonstrated larger evidence or probability of occurrence, it was adopted an $AUC \geq 0.6$, aiming only at the exploitation of the results obtained. Table 3 shows the means and standard deviations for the predictive variables, considering the severity dependent variable of NI and WI accidents.

Table 3. Performance of the accident severity classification.

		NI		WI	
Variable	Category	Mean (%)	S. D. (%)	Mean (%)	S. D. (%)
ACT	1.Rear-end collision	29.98	5.58	0.00	0.00
	2. Head-on collision	0.37	0.54	0.00	0.00
	3. Transverse collision	1.53	1.11	0.00	0.00
	4. Lateral collision	15.09	5.93	0.00	0.00
	5. Pile-up	13.57	6.01	0.00	0.00
	6. Rollover	0.36	0.42	25.00	7.50
	7. Pedestrian collision	0.08	0.15	25.00	7.50
	8. Overturning	1.88	1.51	16.67	5.00
	9. Crash with fixed or mobile object	36.24	6.20	0.00	0.00
	10. Fall of motorbikes and motorcycles	0.89	0.86	33.33	8.67
ACC	1. Driver	71.38	3.72	15.63	3.44
	2. Vehicle	3.06	1.49	0.00	0.00
	3. Road and environment	3.87	3.20	59.38	4.53
	4. Other factors	21.69	2.84	25.00	7.50
PER	1. Morning	38.02	8.96	15.63	3.44
	2. Afternoon	43.28	7.76	59.38	4.53
	3. Night	18.70	5.74	25.00	7.50
SGC	1. Normal	62.93	8.83	46.88	6.88
	2. Partial	36.09	8.23	53.13	6.88
	3. Adverse	0.98	1.00	0.00	0.00
WTC	1.Good	80.86	6.31	93.75	10.94
	2. Rain	15.40	6.49	3.13	5.47
	3. Cloudy	2.03	1.34	3.13	5.47
	4. Haze	0.16	0.24	0.00	0.00
	5. Drizzle	1.55	1.32	0.00	0
PAV	1. Dry	81.43	6.75	71.88	5.94
	2. Wet	18.57	6.75	0.00	0.00
	3. Oily	0.00	0.00	28.13	5.94
GER	1. Straight	90.12	8.38	50.00	5.00
	2. Smooth Curve	5.41	5.14	0.00	0.00
	3. Sharp Curve	4.48	4.55	50.00	5.00
PFR	1. Descending	27.36	9.26	30.21	7.76
	2. Level	45.39	7.93	0.00	0.00
	3. Ascending	27.24	3.68	69.79	7.76
HS	1. Present	99.19	0.65	100.00	0.00
	2. Not present	0.81	0.60	0.00	0.00
VS	1 Present	99.44	0.63	100.00	0.00
	2. Not present	0.56	0.55	0.00	0.00

Considering that the AUC is numerically equal to the probability [28], it is seen that the accidents most likely not to cause victims (NI) are the rear-end collision (29.98%), lateral collision (15.09%), pile-up (13.57%) and crash with fixed or mobile object (36.24%). Most of these accidents have as probable cause driver behaviour (71.38%); happen most likely in the morning (38.02%) and in the afternoon (43.28%); with good visibility (62.93%) or partial visibility (36.09%); in good weather condition (80.86%); dry pavement condition (81.43%); in straight segments (90.12%); in level profile (45.39%); and in segments where there is horizontal (99.19%) and vertical (99.44%) signals.

The accidents most likely to cause fatalities (WI) are rollover type (25%), pedestrian collision (25%) and fall of cyclists and motorcyclist (33.33%). Most of these accidents have as probable cause road and environment factor (59.38%); happen most likely in the afternoon (59.38%); with a normal (46.88%) or partial (53.13%) visibility condition; good weather condition (93.75%); dry pavement condition (71.88%); in straight segments (50%) or sharp curve (50%); with descending (30.21%) or ascending (69.79%) road profile; and in segments where there is horizontal (100%) and vertical (100%) signals.

4. CONCLUSIONS

This article presented a discussion of the main limitations encountered when using the ANN approach for road accident severity classification based on unbalanced databases, considering eleven variables that encompass road infrastructure, environment and type of accidents.

The results obtained with the ANN had accuracy of 77.9% and with mean AUC values of 0.618, compatible with the values found in the literature [13,19]. However, an accident rate accuracy of only 2.6% for WI and a high accident rate accuracy for NI (97.4%) were obtained, as observations associated with accidents with victims were more unusual in the data set of traffic accident records (unbalanced dataset).

Because of this limitation, some authors have been exploring data balancing techniques and used series of binary classifiers in ANN modelling, aiming to reduce the training time of the model and increase the accuracy of the general prediction. In this way, it is possible to use resources such as weights of connections in networks, in order to simplify the structure of the model for a better generalization of the results obtained [13, 29].

The ANN architecture allows reconstructing any continuous function independent of the input data sequence, when the iterative learning process uses the entire data set and the actual associations between the categories of the selected variables and the target variable. Thus, the lower the number of predictor variables used, the greater the effectiveness of results generalization obtained in ANN modelling.

Nevertheless, although the ANN is a robust tool for solving complex, multiple-class problems, in cases where data are naturally unbalanced, such as in road accidents, the classification algorithms tend to ignore the less represented categories, in order to optimize the overall precision of the model.

In neglecting these distortions, the ANN algorithm fails to predict and classify the most relevant classes, which are generally less frequent in the database, such as accidents involving victims (fatal and non-fatal). In addition, the size of the database may also influence the results obtained, since the database must be large enough to ensure that the subsets of training and test data are as homogeneous as possible. Conversely, accident databases have a reduced number of observations when compared to databases in the area of transportation for planning and analysis of demand, among others.

Recommendations for future work include the study of database balancing methods, as well as testing different information transfer functions between ANN layers. In addition, it is recommended to explore, where possible, databases of different dimensions, since a larger number of data can increase the generalization power and hence the data adjustment to the ANN model.

Thus, regardless of the structural flexibility of the ANN in the selection of variables and the construction of stochastic models, it is suggested investigating other approaches that are based on network structures, such as complex networks that are suitable for studying random phenomena of complex nature, derived from multiple causes.

Acknowledgments

The authors would like to thank the Coordination of Improvement of Higher Education Personnel (CAPES) and the grant received by the National Council for Scientific and Technological Development (CNPq) process 304683/2015-9 for the financial support to the development of this research.

REFERENCES

[1] Mannering, F. L.; Bhat, C. R. Analytic Methods in Accident Research Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, v. 1, p. 1–22, 2014.

[2] Karlaftis, M. G.; Vlahogianni, E. I. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C*, v. 19, n. 3, p. 387–399, 2011.

[3] Savolainen, P. T.; Mannering, F. L.; Lord, D.; Quddus, M. A. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, v. 43, n. 5, p. 1666–1676, 2011.

[4] Al-Ghamdi, A. S. Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, v. 34, p. 729–741, 2002.

[5] Farmer, C. M.; Braver, E. R.; Mitter, E. L. Two-vehicle side impact crashes: the relationship of vehicle and crash characteristics to injury severity. *Accident Analysis & Prevention*, v. 29, n. 3, p. 399–406, 1997.

[6] Lui, K. J.; McGee, D.; Rhodes, P.; Pollock, D. An Application of a Conditional Logistic Regression to Study the Effects of Safety Belts, Principal Impact Points, and Car Weights on Drivers' Fatalities. *Journal of Safety Research*, Vol. 19, No. 4, 1988, pp. 197–203.

[7] Singleton, M.; Qin, H.; Luan, J. Factors associated with higher levels of injury severity in occupants of motor vehicles that were severely damaged in traffic crashes in kentucky, 2000-2001. *Traffic Injury Prevention*, v. 5, p. 144–150, 2004.

[8] Pirdavani, A.; Brijs, T.; Bellemans, T. Evaluating the Road Safety Effects of a Fuel Cost Increase Measure by means of Zonal Crash Prediction Modeling. *Accident Analysis & Prevention 50*, p. 186–195, 2013.

[9] Ye, X.; Pendyala, R.; Shankar, V.; Konduri, K. A simultaneous model of crash frequency by severity level for freeway sections. *Accident Analysis & Prevention 57*, n. June, p. 140–149, 2008.

[10] Debrabant, B.; Halekoh, U.; Bonat, W. H.; Hansen, D. L.; Hjelmborg, J.; Lauritsen, J. Identifying traffic accident black spots with Poisson-Tweedie models. *Accident Analysis & Prevention*, v. 111, n. September 2017, p. 147–154, 2018.

[11] Chang, L.; Wang, H. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, v. 38, p. 1019–1027, 2006.

[12] Li, Y.; Ma, D.; Zhu, M.; Zeng, Z.; Wang, Y. Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. *Accident Analysis & Prevention*, v. 111, n. October 2017, p. 354–363, 2018.

[13] Mussone, L.; Ferrari, A.; Oneta, M. An analysis of urban collisions using an artificial intelligence model. Accident *Analysis & Prevention 31*, v. 31, p. 705–718, 1999.

[14] Peng, W.; Baowen, X.; Yurong, W.; Xiaoyu, Z. Link Prediction in Social Networks: the state-of-the-art. Sci China *Inf Sci*, v. 58, n. 58, p. 11101–38, 2015.

[15] Chen, C.; Zhang, G.; Qian, Z.; Tarefder, R. A.; Tian, Z. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, v. 90, p. 128–139, 2016.

[16] Pande, A.; Abdel-AtY, M. Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention*, v. 38, p. 936–948, 2006.

[17] Warner, B.; Misra, M. Understanding Neural Networks as Statistical Tools. *The American Statistician*, v. 50 (4), n. February 1970, p. 284–293, 2015.

[18] Abdelwahab, H. T.; Abdel-Aty, M. A. Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record 1746*, n. 1, p. 6–13, 1997.

[19] Delen, D.; Sharda, R.; Bessonov, M. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis & Prevention*, v. 38, p. 434–444, 2006.

[20] Persaud, B.; Retting, R. A.; Lyon, C. Guidelines for Identification of Hazardous Highway Curves. *Transportation Research Record: Journal of the Transportation Research Board*, v. 1717, n. 0, p. 14–18, 2000.

[21] Zeng, Q.; Huang, H. A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis & Prevention*, v. 73, p. 351–358, 2014.

[22] López, G.; Mujalli, R.; Calvo, F. J.; De Oña, J. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis & Prevention*, v. 51, p. 1–10, 2013.

[23] Mujalli, R. O.; Oña, J. De. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. *Journal of Safety Research*, v. 42, n. 5, p. 317–326, 2011.

[24] Oña, J. De; Mujalli, R. O.; Calvo, F. J. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention*, v. 43, n. 1, p. 402–411, 2011.

[25] Xie, Y.; Zhang, Y.; Liang, F. Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. *Journal of Transportation Engineering*, v. 135, n. 1, p. 18–25, 2009.

[26] Detienne, K. B.; Detienne, D. H.; Joshi, S. A. Neural Networks as Statistical. *Organizational Research Methods*, v. 6, n. 2, p. 236–265, 2003.

[27] Hosmer, D.W. & Lemeshow, S. Applied logistic regression, 2nd Ed. John Wiley & Sons, New York, 2000.

[28] Egan, G. The Skilled Helper: A Systematic Approach to Effective Helping. Pacific Grove CA, Brooks/Cole, 1975.

[29] Allwein, E.L., Schapire, R.E., Singer, Y., 2000. Reducing multiclass to binary: a unifying approach for margin classifiers. J. Mach. Learn. Res. 1, 113–141.