Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 58 (2021) 277–284

Transportation Research Procedia

www.elsevier.com/locate/procedia

14th Conference on Transport Engineering: 6th – 8th July 2021

Comparison of Multivariate Regression Models and Artificial Neural Networks for Prediction Highway Traffic Accidents in Spain: A Case Study

Ali Alqatawna[a,*], Ana María Rivas Álvarez[a], Santos Sánchez-Cambronero García-Moreno[a]

[a]*University of Castilla-La Mancha, Department of Civil and Building Engineering, Ciudad Real, 13071, Spain*

Abstract

In recent years Spain shows the great reduction in the accident rate that has been achieved and the improvement of the behavior of road users, despite this, there is still a need to improve many areas. In 2016 for the first time since the last 13 years, the number of deaths increased by 7% concerning to the previous year. In this paper, analysis, and prediction of road traffic accidents (RTAs) of high accident locations highways in Spain, were undertaken using Artificial Neural Networks (ANNs), which can be used for policymakers, this paper contributes to the area of transportation safety and researchers. ANN is a powerful technique that has demonstrated considerable success in analyzing historical data to forecast future trends.There are many ANN models for predicting the number of accidents on highways that were developed using 4 years of data for accident counts on the Spain freeway roads from 2014 to 2017. The best ANN model was selected for this task and the model variables involved highway sections, years, section length (km), annual average daily traffic (AADT), the average horizontal curve radius, Slope gradient, traffic accidents with the number of heavy vehicles. In the ANN model development, the sigmoid activation function was employed with the Levenberg-Marquardt algorithm and the different number of neurons.The model results indicate the estimated traffic accidents, based on appropriate data are close enough to actual traffic accidents and so are dependable to forecast traffic accidents in Spain. However, it demonstrates that ANNs provide a potentially powerful tool in analyzing and predicting traffic accidents. The performance of the model was in comparison to the multivariate regression model developed for the same purpose. The results prove that the ANN model stronger forecasted model which produced estimates fairly close to forecast future highway traffic accidents with Spanish conditions.

* Corresponding author. Tel.: 0034-632227026
E-mail address: AliIssaabkarim.Alqatawna@alu.uclm.es

2352-1465
10.1016/j.trpro.2021.11.038

Keywords: Artifical Neural Netoweks ; Forecast ; Multiple linear regression ; high accident locations highways in Spain

1. Introduction

Traffic accidents in Spain show the great reduction in the accident rate, that has been achieved in recent years as shown in Figure 1 and the improvement of the behaviour of road users. This reduction is mainly due to the increase in the use of the helmet and the belt, the downward trend in the consumption of alcoholic beverages, better user behaviour, improvement in infrastructure and the updating of the security systems of the vehicle fleet according to (DGT,2011-2013). In spite of this, many areas such as the conservation and signalling of roads, speeding or distractions due to the use of mobile phones are still to be improved.

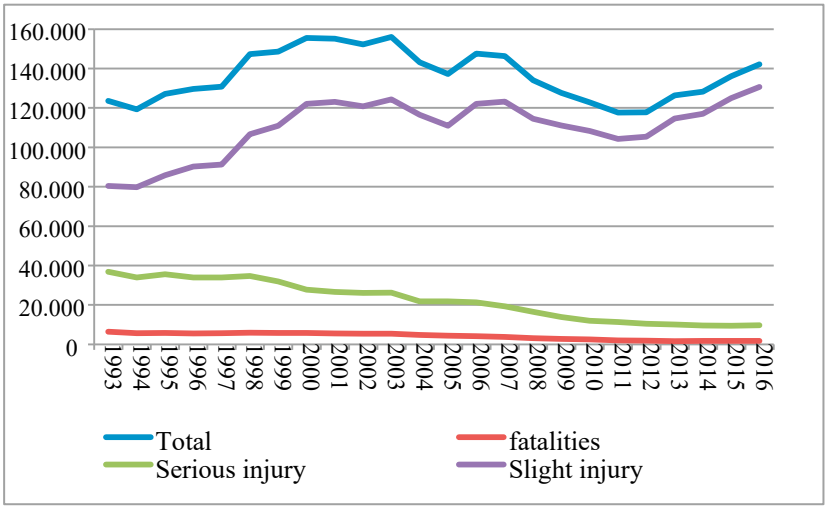

Fig. 1. Number of Accidents (interurban and urban roads) in Spain. (DGT, 2016) [4].

However, although the accident rate has been reduced in recent years, in 2016 for the first time since the last 13 years the number of deaths increased by 7% compared to the previous year (Figure 2). The cost in human lives of these traffic accidents requires the implementation of road safety policies as well as studies and methodologies that allow preventing accidents by identifying potential causes and improving infrastructure.

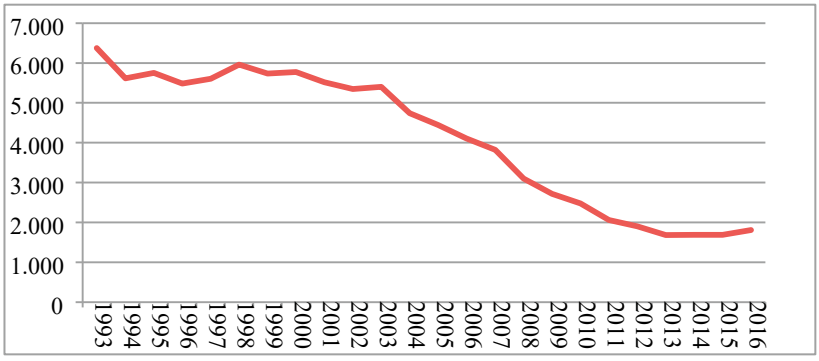

Fig. 2. Number of fatalities (interurban and urban roads) in Spain. (DGT, 2016)

Even with relatively low risks, the General Direction of Traffic tries to further increase road safety. From the perspective of road owners and road safety managers, the understanding of the factors that contribute to a higher frequency of traffic accidents is of paramount importance. In addition to the three Es; engineering, enforcement and education used to manage road safety (DFT, 2011), the planning, construction and maintenance of safe roads require in general the consideration of different fields influencing the occurrence of road accidents (e.g. human behaviour, improvements in automotive manufacturing and weather).

Many research has been carried out on the prediction of traffic accidents in countries using various statistical techniques. However, the numerous variables and complex relationships between the characteristics of the various traffic elements require different analytical techniques than traditional ones. A recent approach to analyze these relationships is the artificial neural networks (ANN) that many scientists have proposed and used successfully as an alternative to conventional ones.

Regression method in predicting time series associated to complex atmospheric and environmental phenomena. This paper presents and discusses the development of a prediction model to estimate future traffic accidents in Spain using the ANN method.

2. Multivariate Regression Model

Regression analysis approaches implement ancient accident statistics to relate accidents to the most contributing factors. Multiple linear regression model is planned to have the following form:

$$Y= B_0+ B_1(X1) +B_2(X2) +B_3(X3) +B_4(X4) +B_5(X5) +B_6(X6) \quad (1)$$

Where:

Y: the predicted number of accidents.
B0: the constant coefficient of the regression line.
B1, B2, B3, B4, B5, B6: the regression coefficients. Were,
B1: Segment length
B2: Slop gradient %
B3: Horizontal curve.
B4: Annual Average Daily Traffic (AADT)
B5: Heavy vehicle percentage .
B6: Speed.

The analysis was carried out using Statistical Package for Social Science (SPSS) version24

The performance of the model is estimated based on the R-Squared (R2) value of the regression line which is presented in the results (Section 5)

3. Artificial neural networks and applications

Artificial Neural Network is a subdomain of Artificial Intelligence (AI) system which has been used recently to solve many variety of civil engineering problems. A neural network is an information processing prototype and a data-modelling tool that represents complex relationships with similar to the human brain. ANNs are known to be universal function approaches and are capable of exploiting nonlinear relationships between variables. Neural networks are a class of flexible nonlinear regression, data reduction models, and nonlinear dynamical systems. They consist of a large number of neurons, i.e. simple linear or nonlinear computing elements, interconnected in often complex ways and organized into layers according to (Sarle, April 1994).

The fundamental element of this model is the novel structure of the information processing system. It is composed of extremely interconnected processing elements termed neurons. Every neuron has a value, weight and bias (constant) where the neuron's net input is the value of the neuron multiply by the weight plus the bias.

Layers composed of an input layer which contains the data to be classified by the network (independent variables),

one or more hidden layers which do the processing, and an output layer which contains the desired output (dependent variable). Every layer contains of neurons connected to each additional neuron in the preceding layer by a connection that represents the weight. An example of an ANN with its several layers is shown in Figure 3. (S. D. Balkina, et al., 2000)

Activation functions: These are also called transfer functions that define the mappings from inputs to hidden nodes and from hidden nodes to outputs, respectively. (Demuth, H., & Beale, M.,2000). ANNs have been applied effectively in solving in many civil engineering problems related to classification, prediction, and function approximation. In the transportation area, ANN has many applications and when applied to predict speed, for example, (J. McFadden, W. T. Yang, and R. Durrans,2000) found it to offer predictive power superior to those of regression models. This is mainly because of their ability to model non-linearity, and flexibility with large complex data sets.

Further applications include the work of (Shoukry ,2005) who used the ANNs in classification of severity levels of accidents and reported several applications of ANN in the transportation field particularly in the traffic safety area.

(Chiou,2006) employed ANN to develop an expert system for the appraisal of two-car accidents (Xiangzheng Xu ,2009) applied the ANNs technique to estimate traffic safety in China, and (Wenhui,2009) researched the evaluation of safety in traffic accident scene based on ANN.

(Borja et al., 2018)was presented an approach for founding an accident risk prediction model, which can be used as a policymaking tool in infrastructure supervision. The method allows for an appropriate handling of the existing data, study show it can be used to develop models using artificial neural networks (ANNs) and creates a systematic optimisation process to determine the optimal architecture of the ANN model. Which was executed using data for accident counts for the Swiss national roads (2009 -2012). It was found that ANNs can be used as a practicable method to predict the frequency of road accidents. As accident rates are quite exceptional events, the data were categorized through a large portion of zero observations.

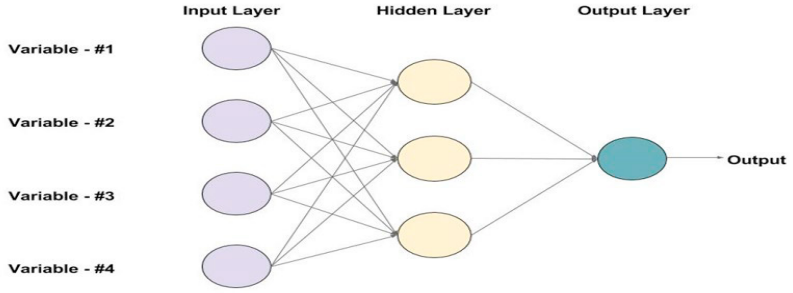

Fig. 3. Typical layers in neural networks

4. Methodology and Analysis

4.1. Case study

The collected data of the number of accidents covered a period of eight years from 2014 to 2017 and relate to the road network of the many province of Spain as Segovia, Burgos, Madrid, Zaragoza, Soria, Guadalajara, Cuenca, Valencia, Toledo, Sevilla, Cordoba. The traffic accident data were collected from the General Traffic Directorate. Each accident data has several information such as the date, accident location, vehicle type, driver's gender, driver's age, accident type, the day and time, the number of fatalities, the number of injured persons, the number of involved vehicles, and the number of damaged vehicles. In addition to these data, geometric characteristics of the highway such as 85th percentile speed, Annual Average Daily Traffic (AADT), the degree of horizontal and slope gradient in each section, were collected from ministry of public work and housing of Spain. After removing the absent and incorrect data, these data were categorized with 9 variables as shown in Table 1 (Çodur MY, 2012).

Table 1. Input variables

Variable	Variable Name	(Numerical /Binary) code
X1	Year (2014-2017)	Categorical Value (2012-2013-2014-2015-2016-2017)
X2	Segment length (m)	Numerical Value
X3	Slope gradiant %	Numerical Value
X4	Radius of horizontal curve	Numerical Value
X5	AADT	Numerical Value
X6	Heavy vehicle %	Numerical Value
X7	85^{th} percentile speed (kph)	Numerical Value
X8	Highway Sections	Categorical Value (A-1, A-2, A-3, A-4)
X9	Freeway Segment	Categorial Value (A-1(Segovia, Burgos, Madrid), A-2(Madrid, Zaragoza, Soria, Guadalajara), A-3(Madrid, Cuenca, Valencia) , A-4(Toledo, Sevilla, Córdoba))
Y	Number of Accidents	Numerical Value

In order to analyze the traffic accidents on the highways, one needs to select highways that possess a wide variety of geometric and traffic characteristics. The goal of this data collection exercise is to divide these highways into segments with homogenous characteristics. After reviewing numerous highways around Spain, it was decided that A-1, A-2, A-3, and A-4 median-divided highways were most proper for this task (Çodur MY, 2012).

The highways with total of 655 km of freeway that connect around of Spain. These freeways are long enough to produce a satisfactory number of segments to develop the model. The information on highways includes geometric characteristics such as horizontal curve, slope gradient, and traffic characteristics such as AADT. A-1, A-2, A-3, and A-4 were divided into 3, 4, 3, and 3 highway segments, respectively and defined by any change in the geometric and highway variables. So, each highway segment is uniform with respect to all the possible geometric and traffic features .The freeway routes are as follows:

- (A-1) Segovia, Burgos, Madrid (101 km);
- (A-2) Madrid, Zaragoza, Soria, Guadalajara (231 km);
- (A-3) Madrid, Cuenca, Valencia (163 km); and
- (A-4) Toledo, Sevilla, Córdoba (160 km).

4.2. Development of ANN models

The ANN models are multilayer feedforward ANNs (i.e. no loops in the network) with between two and ten neurons in a single hidden layer and the equal number of neurons in the output layer as there are output variables. the number of neurons in the hidden layer (h_n) are different from 2 to 10, as advised by (Blum, 1992).

In the ANN model, independent variables are named as the input, and dependent variables are named as the output. Correlation analysis was performed to access the linear association between the variable. Result of the correlation analysis are shown in table 2. From the parameters applied in modeling, eight significant parameters were found based on those criteria. Highlights are years, highway segments, section length (m), annual average daily traffic (AADT), the radius of the horizontal curve, slope gradient (percentage), heavy vehicle (percentage), and 85^{th} percentile speed (km/hr).

Data sets were divided into three sets: the training set (70% of the total data), the validation set (15% of the total data) and the test set (15% of the total data), these phases were performed using MATLAB. Training algorithms don't use the validation or test sets to adjust network weights.

The validation set may optionally be used to track the network's error performance, to identify the best network and to stop training if over-learning occurs. The test set is not used in training at all, and it is designed to give an independent assessment of the network's performance when an entire network design procedure is completed. Selected activation functions were used for input hidden layer; Tan-sigmoid transfer function and for output hidden layer; Linear transfer function.

Table 2. Importance of traffic accidents variables

Variable name	Importance %
AADT	77 %
Segment length	53.4 %
Heavy vehicle percentage	50.3 %
Radius of horizantal curve	41 %
Slope gradiant percnetage	35.5 %
85^{th} percentile speed	22.7 %

5. Results

The Neural Networks allow the development of different alternatives by changing the number of hidden layers. Nine alternative models, with different number of hidden layers, were considered and Table 3 summarizes the results. Model 8 was found to be the best model with the highest coefficient of determination (R^2 =0.9992). A comparison between the actual and the predicted values using model 8 produced the results shown in Table 4. The results were found to be very satisfactory with relatively small residuals especially in recent years where more reliable data bases are available through using more advanced data compilation techniques. The ANN model outputs are exhibited in Figure 6.

Table 3. ANN models alternatives with different number of neurons

Model No.	Number of hidden neurons (h_n)	Correlation Coefficient (r)
1	2	0.9668
2	3	0.9930
3	4	0.9772
4	5	0.9639
5	6	0.9900
6	7	0.9950
7	8	0.9871
8	**9**	**0.9992**
9	10	0.9927

R^2 is used to measure the nearness of fit. A perfect fit would result in R^2 approximately equal 1 , a very good fit near 1, and a poor fit would be near 0. In the ANN model R^2 is 0.9992 . This shows that the ANN model is an suitable method for analysing road traffic accidents.

Table 4. Actual and forecasted number of accident by ANN during four years (2014-2017) for model No.8.

Segment Name	Actual	Forecasted	Residual
Segovia(A-1)	9	40.783	-31.783
Burgos (A-1)	36	28.390	7.609
Madrid(A-1)	735	735	0
Madrid(A-2)	704	704	$-2.27e^{-13}$
Zaragoza(A-2)	325	325	$5.68e^{-14}$
Soria(A-2)	38	38	$-1.27e^{-13}$
Guadalajara(A-2)	183	183	$-5.68e^{-14}$
Madrid(A-3)	543	534.093	8.906
Cuenca(A-3)	42	42	$2.842e^{-14}$
Valencia(A-3)	481	481	0
Toledo(A-4)	209	209	$-8.526e^{-14}$
Sevilla(A-4)	62	53.445	8.554
Córdoba(A-4)	228	228	$-5.68e^{-14}$

A summary of the results obtained from the regression model is shown in table 5. it reported the strength of the relationship between the dependant and independent variables.

Table 5. Model Summary

R	R^2	Adjusted R Square	Std. Error of the estimate
0.976[a]	0.952	0.903	80.851

a. Predictors: (Constant), Speed, Segment length, Slop, Heavy vehicle, AADT, Horizontal Curve

The multiple–correlation coefficient(R) reflects the linear correlation between the observed value and the expected value, it is large value reflects a strong relationship of how the independent variables can affect the predicted value of accidents.

The high value of R^2 indicates that 0.95 of the variation in accidents is explained by the independent variables in cooperated in the model. Table 6 shows the results obtained from the regression model

The final regression model developed using the available data as the following form:

$$Y= (B1*0.004) +(B2*7081.988) +(B3*-11.383) +(B4*0.008) +(B5*3.07) +(114*B6) -6770.128 \quad (2)$$

Table 6. Actual and predicted number of accidents using Regression

Segment Name	Actual	Forecasted	Residual
Segovia(A-1)	9	-49.2	58.2
Burgos (A-1)	36	29.1	6.9
Madrid(A-1)	735	573.1	161.9
Madrid(A-2)	704	677.9	26.1
Zaragoza(A-2)	325	339.8	-14.8
Soria(A-2)	38	-58.4	96.4
Guadalajara(A-2)	183	121.6	61.4
Madrid(A-3)	543	605.5	-62.5
Cuenca(A-3)	42	141.8	-99.8
Valencia(A-3)	481	394.2	86.8
Toledo(A-4)	209	191.1	17.9
Sevilla(A-4)	62	28.7	33.3
Córdoba(A-4)	228	119.9	108.1

A comparison between multiple linear regression model and the model produced by the ANN is presented in tables above ,it can be seen that the ANN model has better predictive power than regression model.

6. Conclusion

In this study, the factors that cause accidents are investigated, for providing road safety, and accident prediction models that include relations between these factors have been established. For the geometrical features of highways sections and traffic accident reports the years from 2014 to 2017 were used to form the database. The obtained data from the database in this study have been investigated with ANN as a tool of forecasting techniques. Since ANN method is a more flexible methodology also, capable of evaluating all the traffic accident characteristics, it is selected for modelling the traffic accidents data.

The model results show that the traffic volume (AADT) with the high percentage (77%) is the most significant factor affecting the number of accidents on the highways. Heavy vehicle percentage and segment length have almost the same effect as the second important parameter. The degree of horizontal curvature is the third one and Slope gradient %, percentile speed have small influence on the output factor.

These results have suggestions for policy makers, transportation system designers, and researchers. Transportation safety designers cannot easily identify factors, make recommendations for incremental changes in the factor, and hope

to achieve major differences in accident levels. The problems must be evaluated from a multidimensional perception: a wide variety of geometric and traffic characteristics. Researchers similarly may adopt techniques such as neural networks for analysis of such variables.

References

Alkheder, S., Taamneh, M. & Taamneh, S. (2017) Severity prediction of traffic accident using an artificial neural network. Journal of Forecasting 36.1, 100-108.

Balkina. S. D, et al. (2000), "Automatic neural network modelling for univariate time series," International Journal of Forecasting, 16.4, 509-515.

Blum A (1992) Neural Networks in C++: An Object-Oriented Framework for Building Connectionist Systems. Wiley, Hoboken, NJ, USA.

Borja, G. D. S., Andreas, B., Markus, D. & T, A. B. (2018) Predicting road traffic accidents using artificial neural network models. Infrastructure Asset Management 5.4, 132-144.

Chiou, Y. C, (July 2006). An artificial neural network- based expert system for the appraisal of two-car crash accidents, Department of Traffic and Transportation Engineering and Management, Feng Chia University, Taiwan.

Çodur. MY, (2012). Traffic Accident Prediction Models: Applications for Surrounding Highways of Erzurum [PhD Thesis] [in Turkish]. Graduate School of Natural and Applied Sciences, Department of Civil Engineering, Ataturk University.

Demuth, H., & Beale, M. (2000). Neural network toolbox user's guide.

Deublein, M., Schubert, M., Adey, B. T. & García De Soto, B. (2015) A Bayesian network model to predict accidents on Swiss highways. Infrastructure Asset Management 2.4,145-158.

DFT (Department for Transport), (2011) Strategic Framework for Road Safety DFT, London, UK. See https://assets.publishing.service.gov.uk / government / uploads / system / uploads / attachment_data / file / 8146 / strategicframework.pdf (accessed 04/24/2017).

DGT (Dirección General de Tráfico), (2011-2020), Road Safety Strategy, Annual report

DGT (Dirección General de Tráfico), (2017), Road Safety Strategy, Annual report.

McFadden. J, Yang.W. T, and Durrans.R, (2000). "Application of ANN to predict speeds on two-lane rural highways," Transportation Research Record 1751, 9-17.

Sarle. W. S., (April 1994). "Neural networks and statistical models," in Proc. Nineteenth Annual SAS Users Group International Conference, NC, USA.

Shoukry, F.N. (2005), "Artificial neural network in classification of severity levels in crashes with guardrail," Master of Science Thesis, West Virginia University, USA.

Wenhui. Z, (2009), "Safety evaluation of traffic accident scene based on artificial neural network," ICICTA, China, vol. 1, Oct. 408-410.

Xu.X, B. Chen, and Gan.F, (2009), "Traffic safety evaluations based on gray systems theory and neural networks," in Proc. WRI World Congress on Computer Science and Information Engineering, 5, 603-607.