Accident Analysis and Prevention 90 (2016) 82–94

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Methodological development for selection of significant predictors explaining fatal road accidents

Bahar Dadashova [a,*], Blanca Arenas-Ramírez [b], José Mira-McWilliams [b], Francisco Aparicio-Izquierdo [b]

[a] Texas Transportation Institute, Texas A&M University, College Station, TX 77843-3135, USA
[b] University Institute of Automobile Research (INSIA), Universidad Politécnica de Madrid, 28006 Madrid, Spain

ARTICLE INFO

Article history:
Received 18 May 2015
Received in revised form
22 December 2015
Accepted 2 February 2016
Available online 27 February 2016

Keywords:
Variable/predictor selection
Structural model
Neural networks
Markov Chain Monte Carlo
Dynamic regression
Accident data
Accident prediction

ABSTRACT

Identification of the most relevant factors for explaining road accident occurrence is an important issue in road safety research, particularly for future decision-making processes in transport policy. However model selection for this particular purpose is still an ongoing research. In this paper we propose a methodological development for model selection which addresses both explanatory variable and adequate model selection issues. A variable selection procedure, TIM (two-input model) method is carried out by combining neural network design and statistical approaches. The error structure of the fitted model is assumed to follow an autoregressive process. All models are estimated using Markov Chain Monte Carlo method where the model parameters are assigned non-informative prior distributions. The final model is built using the results of the variable selection. For the application of the proposed methodology the number of fatal accidents in Spain during 2000–2011 was used. This indicator has experienced the maximum reduction internationally during the indicated years thus making it an interesting time series from a road safety policy perspective. Hence the identification of the variables that have affected this reduction is of particular interest for future decision making. The results of the variable selection process show that the selected variables are main subjects of road safety policy measures.

Published by Elsevier Ltd.

1. Introduction

Road accidents are very complex processes which can be explained as a consequence of set of different influential factors. Impact of the external factors on road accidents is studied through statistical and time series models. Parsimonious models are conventionally preferred over more complex ones for their ability to interpret the data by the simplest way available, known as principle of parsimony. Using an adequate model selection method to find such a model is therefore very important step during model construction and significantly affects to the performance of the model (e.g. interpretation of the accident data, optimal prediction accuracy, etc.). The conventional methods used for the model and hence variable selection stage are stepwise elimination (backward or forward) and subset selection methods. Stepwise elimination is based on sequential introduction/elimination of the variables into/from the model one by one. The elimination of the variable from the model is based on its goodness of fit measures (*p*-values). Selection based on *p*-values can result in misleading results, since the interpretation of a certain variable, when there is a large set of variables in the model, is not the same when there are only a few included in the model (Raftery, 1995). Subset selection is based on Akaike Information Criteria of each candidate model. However with more variables this method can become unattractive since the number of candidate models increase thus becoming hard to handle.

One of the popular methods for the model selection is the Lasso (least squares shrinkage and selection operator) method, a tool developed by Tibshirani (1996). The method is based on minimizing the residual sum of squares subject to the absolute value of coefficients being less than a constant, *s*. The method has been demonstrated to be more stable and accurate than the other model selection methods such as subset selection. The methodology is designed for selection of individual input variables, not for general factor selection. When directly applied to model they tend to make selection based on the strength of individual derived input variables

* Corresponding author.
E-mail address: bahar.d@tamu.edu (B. Dadashova).

http://dx.doi.org/10.1016/j.aap.2016.02.003
0001-4575/Published by Elsevier Ltd.

rather than the strength of groups of input variables (Yuan and Lin, 2006).

In this paper a neural network (NN) based model selection method is proposed. The methodology proposed in this article follows a similar reasoning as the Lasso, in a sense that predictors are selected considering their individual variable importance. For the selection of variables a set of candidate models (TIM – Two Input Models) are built with combinations of variables based on a NN design. Given their flexibility and the ability to combine different inputs to explain the output, NN have proven to be very efficient tools for the variable (predictor) selection process (Moody, 1992; Finnoff et al., 1993; Swanson and White, 1997; Anders and Korn, 1999; Curry et al., 2000; May et al., 2011; Delen et al., 2006). By combining only two variables and estimating their joint effect on the response variable we can observe their near-individual effect on the response variable. Also, since all the combination are considered, the collinearity between a given variable and the other predictors can be detected by examining the goodness of fit measures of the TIM models. TIMs are estimated through Markov Chain Monte Carlo (MCMC) sampling techniques. The results of the model estimation are compared to those of dynamic regression (DR) which is able to model the stochasticity observed in the data through sophisticated functional forms. Both the results of the model selection and DR model are cross-validated through the prediction analysis.

The model selection method is applied to explain the fatal accident frequency in Spain, during 2000–2011. This series has very interesting properties from a road safety research perspective. Compared to 2000, in 2011 the number of fatal accidents has decreased by 63% which is the highest reduction achieved internationally (IRTAD, 2013). This period in Spain is characterized by significant road safety policies inducted by the safety-concerned authorities such as Penalty Point System (2006), Penal Code Reform (2007), media campaigns, different public plans to encourage drivers to renew vehicle fleets, etc (Castillo-Manzano et al., 2010, 2011; Aparicio et al., 2011; Dadashova et al., 2014). Although these factors had a positive impact on road safety however the impact of other factors is also accounted for, particularly the significant effect of the declining economic situation. The short term effect of a declining economy is assumed to be positive on road safety, due to the decreasing exposure and mobility (Fridstrøm, 2000; Scuffham and Langley, 2002; Garcí-Ferrer et al., 2006; Loeb and Clarke, 2007). Identifying the predictors that may have played an important role in decreasing the number of fatal accidents is not merely an academic question but also would have important policy implications for future decision making in road safety research.

The rest of this paper is organized as follows. In Section 2 the data are discussed and literature on the effects of the factors is reviewed. Section 3 presents the statistical methodology and introduces the stages of the methodological development. In Section 4 the estimation results and the discussions are presented. The paper ends with the conclusions and suggestions for future work.

2. Data

2.1. Fatal accidents

Empirical analysis is carried out using the monthly data on number of *fatal accidents* in Spain during 2000–2011. The data were obtained from the General Directorate of Traffic (DGT) of Spanish Government. The number of fatal accidents and the number of deaths in Spain have generally decreased during 1990–2011 and more so after 2000. Although the indicators representing "road network intensity use" factors such as drivers, vehicles and the distance traveled, have increased during this time, however we can observe a certain decline from 2007 on (Fig. 1). From the road safety policy perspective the first decade after 2000 have been the most effective years from road safety policy perspective in Spain compared to previous periods. In July, 2006 and December, 2007, the Penalty Point System and the Penal Code Reform respectively, were enacted. Other preventive and corrective strategies were conducted in the framework of the 2005–2008 Road Safety Strategic Plan. Another factor that has significantly affected the road safety indicators in Spain during this period has been the declining economic situation which, as mentioned previously, has a negative effect on the transport related socioeconomic factors such as mobility, investment on infrastructure, vehicle fleets, etc. These factors, in turn, affect the road safety indicators in different directions.

Table 1

The preliminary descriptive data analysis of the dependent variable was carried out through the TRAMO (Time series Regression with ARIMA noise, Missing values and Outliers) and SEATS (Signal Extraction in ARIMA Time Series) (Gómez and Maravall, 1996) program in order to filter the time-varying component and detect the outliers that might be present in the data (Fig. 2). The modified

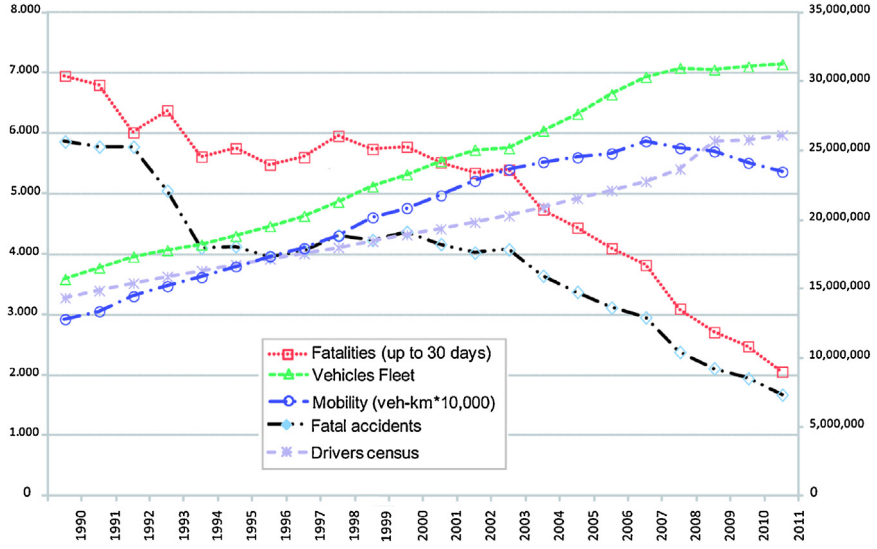

Fig. 1. Road use indicators: drivers, vehicles and total distance traveled vs. road traffic accident frequency and severity in Spain 1990–2011.