Accident Analysis and Prevention 96 (2016) 108–117

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

M5 model tree based predictive modeling of road accidents on non-urban sections of highways in India

Gyanendra Singh (Assistant Professor)[a,*], S.N. Sachdeva (Professor)[b], Mahesh Pal (Professor)[b]

[a] Civil Engineering Department, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonepat, Haryana, India
[b] Civil Engineering Department, National Institute of Technology, Kurukshetra, Haryana, India

ARTICLE INFO

Article history:
Received 22 March 2016
Received in revised form 1 August 2016
Accepted 4 August 2016

Keywords:
Predictive modeling
Road safety
M5 model tree
Random effect negative binomial model

ABSTRACT

This work examines the application of M5 model tree and conventionally used fixed/random effect negative binomial (FENB/RENB) regression models for accident prediction on non-urban sections of highway in Haryana (India). Road accident data for a period of 2–6 years on different sections of 8 National and State Highways in Haryana was collected from police records. Data related to road geometry, traffic and road environment related variables was collected through field studies. Total two hundred and twenty two data points were gathered by dividing highways into sections with certain uniform geometric characteristics. For prediction of accident frequencies using fifteen input parameters, two modeling approaches: FENB/RENB regression and M5 model tree were used. Results suggest that both models perform comparably well in terms of correlation coefficient and root mean square error values. M5 model tree provides simple linear equations that are easy to interpret and provide better insight, indicating that this approach can effectively be used as an alternative to RENB approach if the sole purpose is to predict motor vehicle crashes. Sensitivity analysis using M5 model tree also suggests that its results reflect the physical conditions. Both models clearly indicate that to improve safety on Indian highways minor accesses to the highways need to be properly designed and controlled, the service roads to be made functional and dispersion of speeds is to be brought down.

1. Introduction

Indian road network of 4.96 million km is second largest in the world (NHAI, 2014). It has been experiencing a very fast and unprecedented growth. Registered motor vehicle growth rate in the last five decades has been above 10% (MORTH, 2013). But this growth has not resulted in safe travel. The widening and upgradation of roads have not been able to contain the fatal road accidents. Road accidents alone accounted for 1, 41,526 accidental deaths in 2014 (National Informatics Centre (NIC, 2014). The proportion of fatal to total accidents has consistently increased from 18.1% in 2003 to 25.2% in 2013 (MORTH, 2013). Studies suggest that conversion of roads from two to four lanes has resulted in an increase of fatality rate from 41% to 51% on the high-crash-rate sections (Shaheem and Das Gupta, 2005; Shaheem et al., 2006). On 4-lane divided roads, head-on collisions comprise 19% of the crashes due to wrong side traffic alone (MOST, 2000).

Accident prediction models are being used worldwide by engineers and planners as a useful tool to understand the causal factors of road accidents and to suggest measures for improvement. Accident prediction models try to understand the factors associated with accident occurrence by developing statistical relationships correlating various risk factors with the number of accidents occurring on a road section over a period of time. Accident occurrence being a complex phenomenon presents serious challenges to the modeller. The possibility of making causal inferences based on accident prediction models depends strongly on how well the assumptions reflect the reality, what functional relationship was chosen and which method was adopted to overcome disturbing factors. These data and methodological issues have been thoroughly discussed in the literature by various researchers (Miaou, 1996; Persaud, 2001; Hauer, 2010; Lord and Mannering, 2010; Elvik, 2011; Savolainen et al., 2011; Mannering and Bhat, 2014).

Indian scenario has added complexities in accident prediction due to heterogeneity of traffic (Landge et al., 2006; Mohan et al., 2009), under-reporting of accidents (Varghese and Mohan, 1991;

* Corresponding author.
E-mail addresses: singhgyan27@yahoo.in, singhgyan27@gmail.com (G. Singh), snsachdeva@yahoo.co.in (S.N. Sachdeva), mpce_pal@yahoo.co.uk (M. Pal).

http://dx.doi.org/10.1016/j.aap.2016.08.004
0001-4575/

Mohan, 2002; MORTH, 2007) and poor quality of available accident data (Sharma and Landge 2012, 2013; Sharma et al., 2013; Jacob and Anjaneyulu, 2013). Studies have also suggested that separate models must be developed according to the conditions of the individual countries as the traffic flow and its composition, road conditions, driver and road user behaviour are very different in different countries (Jacobs et al., 2000; Fletcher et al., 2006).

The present study was carried out in Haryana, a state in Northern India, lying on the border of the national capital region, New Delhi. In the last two decades (1991–2011) the road accidents in Haryana have increased six times, while the fatalities in road accidents have gone up by 11 times. With 3.5% and 2.2% share in road fatalities and road accidents respectively, Haryana is among top thirteen unsafe states on road safety indicators in India although its geographical area is around 1.34% and population is 2.09% of India (MORTH, 2013). Fig. 1 indicates that all road safety indicators in Haryana are worse than the national average.

Keeping above issues related to road safety in mind, this study focuses on development of predictive model for traffic accidents on non-urban sections of highways in Haryana which are home to about two-third of road accidents and three-fourth road fatalities in the state (MORTH, 2013), and examines the effectiveness of M5 model tree based regression model and conventionally used fixed/random effect negative binomial (FENB/RENB) regression models for prediction of road accidents and identification of the key risk factors.

2. Literature review

2.1. Predictive modeling of road accidents

The occurrence of accidents is a rare and random event and number of accidents is a non-negative discrete variable. Therefore the probability of occurrence of accidents could be better represented by Poisson distribution. As Poisson regression assumes that mean is equal to variance, which is not true for most of the highway safety problems, various variants of Poisson model have been in use for accident prediction according to the data specifications. Negative Binomial regression (NB) models (Lord et al., 2005; Fletcher et al., 2006; Robert and Veeraragavan, 2007; Cafiso et al., 2010; Jacob and Anjaneyulu, 2013; Divakaran and Sreelatha, 2013; Sharma et al., 2014) are best suitable to model over-dispersed data. When accident data has a large number of zero accident sites, zero-inflated Poisson or/and NB models (Sharma et al., 2013; Sharma and Landge, 2013; Jacob and Anjaneyulu, 2013) are employed. Poisson-Weibull models (Maher and Mountain, 2009; Chikkakrishna et al., 2013) provide flexibility in choosing the distribution of error terms. Conway-Maxwell-Poisson models (Lord et al., 2008) also provide flexibility in choosing the distribution of error terms and can handle over- as well as under-dispersion.

Accident data are generally produced by repeated measurements in time over road sections. Thus the data may have spatial and temporal correlations due to some unobserved effects like regional correlation in the data, variation in traffic and driver related effects which are particularly significant in mixed traffic conditions. Poisson and NB distributions cannot handle these unobserved heterogeneities arising from spatial and temporal effects, as the accident distributions for the sites with similar observed characteristics are considered the same in these models. Furthermore, accident counts for a specific location at different time periods are also assumed to be independent of each other. Without appropriately accounting for the location-specific effects and potential temporal correlations, the estimates of the standard error in the regression coefficients may be underestimated. To tackle this problem RENB model was proposed by Shankar et al. (1998). Generalised estimation equations were used by Lord and Persaud (2000) to model crash data with serially correlated repeated measurements. Two-state Markov switching NB models (Malyshkina et al., 2009) which assume switching of roadway segment between two unobserved states of roadway safety to account for unobserved effects, also resulted in a better fit as compared to regular NB models. Anastasopoulos and Mannering (2009) and Dinu and Veeraragavan (2011) applied Random Parameter model by employing a normally distributed error term in the coefficients to allow them to vary across observations. Finite mixture NB regression models with fixed and varying weight parameters were also employed to address the unobserved heterogeneity problem in accident data (Zou et al., 2014). Although these models provide a statistical fit that is significantly better than traditional NB model but they are very complex, may not necessarily improve predictive capability, and model results may not be transferable to other data sets because the results are observation specific (Shugan, 2006; Washington et al., 2010).

The major advantage of applying these statistical models is their ability to identify a broad range of risk factors that can contribute significantly to accidents. However, most of the statistical models have their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, the model could lead to erroneous estimation of accident likelihood (Chang, 2005; Savolainen et al., 2011). NB models, though most widely used in predictive modeling of accidents, are unable to handle under-dispersed data and the low sample mean problem (Lord, 2006; Lord et al., 2008). These models can easily and significantly be influenced by outliers, cannot handle discrete independent variables with more than two levels, and can be adversely affected by multi-colinearity among independent variables (Karlaftis and Golias, 2002).

Another class of predictive algorithms, which does not require any pre-defined underlying relationship between dependent and independent variables, has also been reported in the literature. The main algorithms belonging to this category are Hierarchical Tree based regression (Karlaftis and Golias, 2002; Fletcher et al., 2006), Artificial Neural Network (ANN) (Chang, 2005; Riviere et al., 2006; Xie et al., 2007; Sikka, 2014) and Support Vector Machine (SVM) (Li et al., 2008). Applications of these algorithms are new to the field of highway safety but found to be performing well in comparison to most widely used NB regression approach.

Few applications of tree based regression are reported in highway safety analysis literature. Hierarchical Tree based regression (Karlaftis and Golias, 2002) was tested in Indian conditions by Fletcher et al. (2006) but due to the availability of limited amount of data, the results were found inferior to generalised linear models. In a comparative study (Chang and Chen, 2005), Classification and Regression Tree (CART) was proposed as a good alternative method to NB regression to analyse freeway accident frequencies. Emerson et al. (2011) used M5 model tree based regression (Quinlan 1992; Wang and Witten, 1997) and other data mining approaches to predict crash counts based upon skid resistance values and suggested that M5 regression tree produced high classification rates of instances with a low rule count. M5 model tree can tackle tasks with very high dimensionality (up to hundreds of attributes) and can learn efficiently from large datasets with a small computational cost. The advantage of M5 over CART (Breiman et al., 1994) is that model trees are generally much smaller than regression trees and have proven to be more accurate, due to their ability to exploit local linearity in the data. M5 model tree can have multivariate linear models at its terminal nodes; and thus analogous to piecewise linear functions. Regression trees never give a predicted value lying outside the range observed in the training cases, whereas model trees are found to extrapolate well (Quinlan, 1992). Keeping in view the encouraging performance of M5 model tree, this study