COMPARISON OF REGRESSION AND ARTIFICIAL NEURAL NETWORK MODELS FOR VEHICULAR SPEED PREDICTION WITH LIMITED OBSERVATION DATA

Shibaura Institute of Technology	Ph.D. Student	Student Member	Nitesh Acharya
Shibaura Institute of Technology	Associate Professor	Regular Member	Michael Henry

1. INTRODUCTION

Speeding can make vehicles errant and difficult to control, which can result in severe road crashes, leading to fatality or severe injury. Mountainous roads have some distinctive characteristics like narrow lanes and shoulders, steep grades, sharp turns, and inadequate sight distance that make them more prone to traffic crashes, and speeding there is more hazardous.

Road traffic crashes can be reduced by lowering vehicular speed. Speed prediction models can be used to predict the speed at spots where a speed survey was not conducted and would help in predicting vehicle speeds once the geometric features of the highway like carriageway width, shoulder width, and radius of the curve, have been upgraded. Robust prediction models can be developed using plentiful vehicle speed observations and appropriate machine-learning tools. However, it is difficult to gather huge amounts of data because of time and budget constraints and hence becomes necessary to explore the appropriate statistical models and technique that can provide reliable and accurate prediction results. Therefore, this study aims to compare the results of speed prediction models developed using regression and artificial neural networks and analyze their effectiveness when there are limited field observations.

2. METHODOLOGY

2.1. Field inspection and data collection

The study examined the Dhulikhel-Sindhuli-Bardibas (H13) highway, a 160-kilometer highway that traverses through the mountainous terrain of central Nepal. Sixty locations along the highway were chosen for spot speed measurements by considering the general road and traffic conditions. In addition to the geometric characteristics of the road, traffic features were evaluated using the average daily traffic (ADT), and the International Roughness Index (IRI) was used to measure the pavement condition. The spots were also chosen on sections of the road with speed limits, approaches to intersections, and pre-detected crash hotspots. The speed was measured with a speed radar gun and, at each location, about 100 vehicles were sampled to measure their speed at free flow conditions. The speed observations were made in August 2022 and all the speed measurements were conducted during the daytime. In total, 6233 vehicle speed observations were made.

2.2. Data Analysis

The 85th percentile speed (V_{85}) for each of the sixty spots was determined using the spot's speed observations. V_{85} is the speed at or below which 85 percent of all vehicles are observed to travel under free-flow conditions at a certain observation spot. Correlation analysis was carried out to assess the relationship between V_{85} and different variables, and it was found that speed restriction, crash blackspots, and approach of intersection did not show any strong correlation with V_{85}, and hence were omitted in further analyses. The resulting dataset was 60 observations of six independent variables (ADT, IRI, radius of curve, longitudinal grade, width of carriageway, and shoulder) and V_{85} as the dependent variable. Of the total dataset, 80% was used to train the model, and the remaining 20% was used to test the model.

First, multivariate linear regression (MLR) and backward step regression (BSR) were used to develop regression models for predicting V_{85}. BSR is an advanced type of linear regression where gradually step-by-step variables are eliminated from the full model to find a reduced model that best explains the data. Next, another prediction model was constructed using an artificial neural network (ANN). ANN is a computational model which mimics the work of the nerve cells in our brain. It is a type of supervised machine learning, where the analyst feeds the data into the input layer and these data are passed and processed in the hidden neural layers, where the actual data processing occurs, and finally, the processed data is fed into the output layer. Compared to regression models, ANN can discover non-linear and complex relationships between the variables. Prediction of V_{85} was carried out using a sigmoid type of activation function to construct a forward-fed back propagation type ANN model with a single hidden layer having five neurons.

ANN models constructed with smaller training dataset often leads to overfitting and yield unreliable results [1]. Since there are not any alternatives to collect further data or to increase the dataset, the bootstrap sampling technique is used to create several sample datasets. Bootstrap sampling is a method that involves drawing sample data repeatedly with replacement from a data source to estimate any population parameter. In this study, 50 bootstrap samples containing 60 observations were prepared and fed into the ANN model to calculate the 95% confidence interval for RMSE. Open-source statistical analysis tool R (https://cran.r-project.org/) was used to develop and analyze prediction models.

3. RESULTS AND DISCUSSION

Tables 1 and 2 show the results of the MLR and BSR models, respectively. It can be seen that radius of curve, grade, and shoulder width are significant ($p<0.05$) in the BSR model, but only radius of curve is significant in the MLR model. In addition, compared to the results from the standard MLR model, the BSR model showed better RMSE and R^2 values.

Keywords: 85th percentile speed, multivariate linear regression, artificial neural network, bootstrapping
Address: Toyosu 3-7-5, Koto-ku, Tokyo, Japan 135-8548; TEL +81-03-5859-8363; email: na21504@shibaura-it.ac.jp

Table 1: Results of the MLR model

Coeff.:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	61.834	7.432	8.320	2.01e-10***
ADT	0.0001	0.0006	0.297	0.768
grade	-0.582	0.366	-1.590	0.119
rad_cur	-4.215	0.714	-5.901	5.51e-07***
car_wd	-0.774	0.882	-0.878	0.385
shd_wd	-1.958	1.578	-1.240	0.222
IRI	0.794	1.090	0.728	0.471
F-stat:	6.36 on 6 and 42 DF, p-value: 8.072e-05			
R-square:	0.438	RMSE:	7.487	

Table 2: Results of the BSR model

Coeff.:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	55.639	5.351	10.398	1.35e-14***
grade	-0.690	0.335	-2.060	0.0441*
rad_cur	-4.189	0.651	-6.436	3.16e-08***
shd_wd	-2.448	0.996	-2.457	0.0172*
IRI	1.690	0.956	1.768	0.0826.
F-stat:	11.68 on 4 and 55 DF, p-value: 6.16e-07			
R-square:	0.464	RMSE:	7.239	

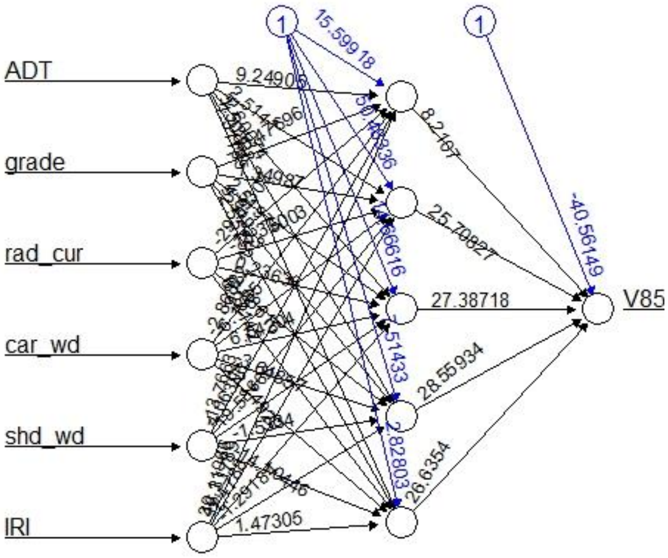

Figure 1: ANN architecture with weights and bias

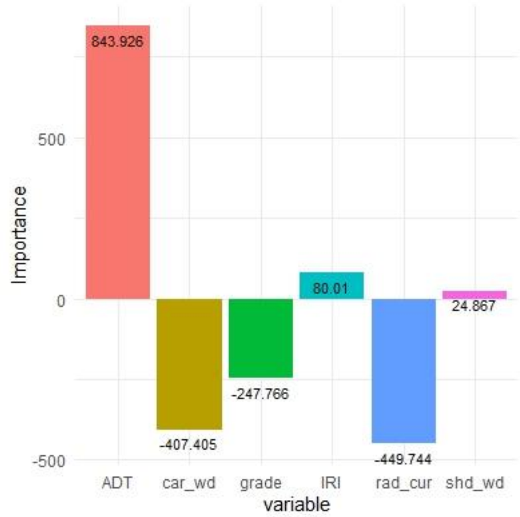

Figure 2: Variable importance plot using the Olden algorithm

The architecture of the ANN model is shown in Figure 1. The developed model showed R^2 value of 0.478 and RMSE value of 6.942. Both metrics show improvement from the above-developed regression models. In Figure 2 the importance of the features in the ANN model is plotted using the Olden algorithm [2], which is based on the sum of the products between the connection weights of the network for any variables. In MLR and BSR models, radius of curvature was the most significant variable, but the ANN model shows ADT as the most important one followed by radius of curvature.

Since models developed using smaller datasets are prone to overfitting, the same dataset was used to prepare fifty bootstrapped samples. ANN models using these samples showed a mean R^2 of 0.556 and mean RMSE of 8.01, with a 95% confidence interval of 8.01 ± 0.86. In the case of bootstrapped samples, even though the R^2 value has improved the RMSE has declined when compared to the ANN model built using the original dataset. The reason for this model to provide better RMSE can be the inability of the ANN model to learn the relationship between the variables due to limited observations. Using the bootstrapped samples, it was possible to analyze these relationships and achieve more realistic and reliable results. Although ANN models were expected to provide better prediction than the regression models, the results from the bootstrapped samples and the confidence interval showed that ANN may not provide accurate prediction due to overfitting, and hence it would be appropriate to conduct regression analysis when the sample size is small.

4. CONCLUSIONS

Through this study, it can be concluded that between the two regression models, BSR showed better predictive ability than MLR. Furthermore, although ANN is a robust technique for prediction, a smaller dataset can cause overfitting and yield unreliable results. Therefore, bootstrapping and developing confidence intervals for any statistical parameter can be a good technique to ascertain the results of prediction models. Also, it can be noted that there exist uncertainties in the results when the observation data is limited, and hence to overcome this issue, prediction models can be developed by gathering abundant field data such that the developed robust models can be effectively used in highway safety management.

ACKNOWLEDGEMENT

This research was supported by a scholarship for road asset management from the Japan International Cooperation Agency.

REFERENCES

[1] Pasini. A., “Artificial neural network for small dataset analysis”, Journal of Thoracic disease (2015), 7(5): 953-960

[2] Olden, J.D., Joy, M.K., Death, R.G., “An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data”, Ecological Modelling 178 (2004), 389-397