Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 17 (2016) 644 – 652

Transportation Research Procedia

www.elsevier.com/locate/procedia

11th Transportation Planning and Implementation Methodologies for Developing Countries, TPMDC 2014, 10-12 December 2014, Mumbai, India

Mode Choice Analysis Using Random Forrest Decision Trees

Ch.Ravi Sekhar[1], Minal[2], and E. Madhu[3]

[1] Senior Scientist, Transportation Planning Division, CSIR-Central Road Research Institute, New Delhi 10025, chalumuri.ravisekhar@gmail.com

2 M.Tech Student, Academy of Scientific and Innovative Research (AcSIR), CSIR-Central Road Research Institute New Delhi-110025, minal.crri@gmail.com

3Principal Scientist, Transportation Planning Division, CSIR-Central Road Research Institute, New Delhi 10025, errampalli.madhu@gmail.com

Abstract

Mode choice analysis forms an integral part of transportation planning process as it gives a complete insight to the mode choice preferences of the commuters and is also used as an instrument for evaluation of introduction of new transport systems. Mode choice analysis involves the procedure to study the factors in decision making process of the commuter while choosing the mode that renders highest utility to them. This study aims at modelling the mode choice behaviour of commuters in Delhi by considering Random Forrest (RF) Decision Tree (DT) method. The random forest model is one of the most efficient DT methods for solving classification problems. For the purpose of model development, about 5000 stratified household samples were collected in Delhi through household interview survey. A comparative evaluation has been carried out between traditional Multinomial logit (MNL) model and Decision tree model to demonstrate the suitableness of RF models in mode choice modelling. From the result, it was observed that model developed by Random Forrest based DT model is the superior one with higher prediction accuracy (98.96%) than the Logit model prediction accuracy (77.31%).

Keywords: Mode choice Analysis,;Household survey; MNL Model; Random Forrest Decision Tree

1. Introduction

Dealing with the present bottlenecks as well as creating long lasting and sustainable transport systems has been the greatest challenge of urban transportation planning. Calibrating the present need and forecasting the future demand is the underlying agenda of travel demand forecasting. Mode choice forms an integral part of this process as

doi:10.1016/j.trpro.2016.11.119

it gives a complete insight to the mode choice preferences of the commuters validating the introduction of new transport systems to existing ones. Mode choice analysis is the procedure to study the factors and decision making process of the trip maker and to be able to model it. Trip makers seem to choose the mode that renders highest utility to them. The objective of this study is to model mode choice behaviour of commuters of Delhi by considering Random Forrest (RF) Decision Tree (DT) method. The comparison of the performance of DT model is done with Multinomial Logit (MNL) model. In this study the major modes of transport considered are private Cars, Two wheelers, Bus, Metro, Auto Rickshaw (Three wheeler), and Bicycle.

Nomenclature

RF	Random Forrest
DT	Decision Tree
MNL	Multinomial Logit

2. Literature Review

Discrete choice models based on random utility maximization are widely used in transportation applications. Logit models are the most widely used discrete choice model. Logit model has the ability to model complex travel behaviors of any population with simple mathematical techniques and thus proves to be the most widely used tool for mode choice modeling. Binary Logit, Multinomial Logit, Nested logit and Mixed Logit models have been applied for mode choice analysis in different studies. The above statistical models have certain discrepancies in them especially regarding their accuracy. Due to this reasons researchers have sought out to more recent Artificial Intelligence (AI) methods in recent years have received increased interest especially among transport researchers and practitioners in exploring the feasibility of applying AI methods to address some of the aforementioned problems in transportation engineering

Random Forrest (RF) is a generic principal of classifier combination problem that uses tree structured base classifiers (Breiman, 2002). This method is very unique among popular machine learning methods such as Artificial Neural Networks (ANN). The use of Artificial Intelligence techniques such as Artificial Neural Network (ANN) in travel demand modeling began in 1960. However it wasn’t used for about next three decades in such type of studies due to its limitations, namely the slow response to the modification of inputs despite its extraordinary success at learning or recognizing pattern. Neural networks have been used in the transportation demand forecasting for urban areas as well as intercity flows and has shown advantages in use for traffic behavioral analysis (Nijkamp,1996 and Subba Rao,1998). Forecasting by ANN is done by minimizing an error term indicated as the deviation between input and output through the use of specific training algorithm and random learning rate (Black,1995 and Zhang et al., 1998). The theorem proved by Hornik (1989) and Cybenko (1989) states that a multilayered feed forward neural network with one hidden layer can approximately take any continuous function up to a desired degree of accuracy provided as it contains a sufficient number of nodes in the hidden layer thus they can be considered as universal approximates. Xie, et. al. (2003) considered two data mining methods, namely learning tree algorithm and back propagation neural networks to improve the prediction accuracy of mode choice model. Karlaftis, et. al. (2001) proposed a recursive partitioning methodology for individual mode choice prediction. The methodology is based on tree-structured nonparametric classification technique. In the transportation research, the application of Random Forrest (RF) has been adopted primarily for traffic accident analysis. Haleem, et. al. (2010) and Hossain and Muromachi (2011) used RF to understand the crash mechanism on urban expressways. Pande, et. al. (2011) has used RF to select variables of the crash risk estimation model. Application of RF in modeling commute mode choice was done by Hasegawa, et. al. (2012). Subsequently Hasegawa, et. al. (2013) modeled mode choice preference of the commuters by deploying a hybrid model of Random Forest and Genetic algorithm and compared it to MNL mode choice model demonstrating higher classification potential of this model.

3. Study Area And Data Collection

In this study, Delhi was chosen as the study area which has a population of 16.7 million people (Census of India, 2011) with a population density of 11,297 per square km. The public transport modes in Delhi form a strong network to cater to the needs of the people mobility such dense population in a metropolitan city which houses multiple offices, industries and manufacturing units will be a marathon task. To carryout mode choice analysis, collection of travel behavior data was carried out through the traditional home based personal interview survey. Travel behavior data has been collected through predesigned questionnaire which is aimed at providing the data to meet the objectives of the present study.

The study area was divided into smaller zones and pockets for data collection. The study area was segregated into different survey pockets that were targeted during the survey where data was collected through multistage sampling. A total of 3000 household samples were collected from South Delhi area and a total of 2000 household sample were collected evenly from the North, East, and West Delhi. Eight modes have been considered in this study namely Drive Alone (DA) Car (Private mode), Carpool (Shared mode), Two Wheeler (Private mode), Bus (Public mode), Metro (Public mode), Auto Rickshaw (IPT mode), Bicycle (Personal/ Non-motorized mode) and Walk (Non-motorized mode).

From the data it was observed that the largest share of transport is driven by the purpose of making work trips which are 73%, followed by business, education and recreational trips. From the data, trips are distributed as per the distance as shown in Figure 1and it was observed that the average trip length to work place is 7.83 km. From the total data (4976 sample), it was observed that the mode share for car is 36%, two wheeler is 26%, Bus is 19%, Metro is 2.4%, Cycle is 0.6%, Walk is 15% and Auto Rickshaw is 1%. The effect of age on mode choice is derived from the data and was observed that the largest commuter share comes from the age group of 31 to 50 years. This dominant age group of commuters prefer private vehicle for their mobility with approximately 20%, 12%, 15% of them using drive alone car, two wheelers and bus respectively.

The attributes used for model development are :Household size ,Number of vehicles in household, Household income (Indian Rupees) ,Age of traveller (in years), Gender of traveller ,Education Level, Type of employment ,Possession of Driver’s License ,Trip Purpose, In Vehicle Travel time for MV (Minutes) ,Out of Vehicle Travel time (Minutes), In Vehicle Travel Time (Minutes) and Travel cost (Indian Rupees). Development of mode choice model using Multinomial model and RF based decision tress is discussed in the following section.

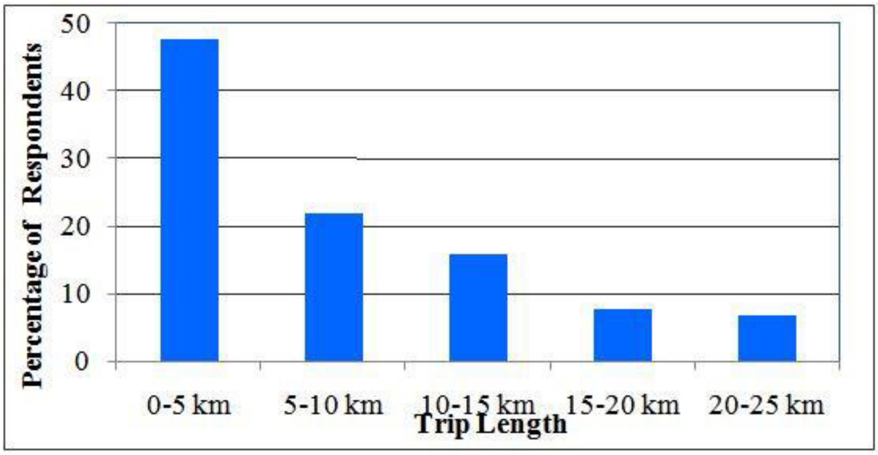

Figure 1 Percentage Distribution of Respondents by the Trip Length

4. Development Of Mode Choice Model

MNL models are most widely used mode choice models are based on the principle of random utility maximization derived from econometric theory while Decision tree is a data mining technique deployed here for classification. A Random Forrest (RF) Decision Tree (DT) method has been considered for modelling the mode choice preference of the commuters in Delhi. The random forest model is one of the most efficient DT methods for solving classification problems. MNL mode choice model was also developed for the same data. A comparative evaluation has been carried out between Multinomial logit (MNL) model and Decision tree mode choice model. A total of 4976 sample were considered for model calibration for both MNL and DT model.

4.1 Multinomial Mode Choice Model

MNL model is widely used disaggregate mode choice model, it estimates the proportion of trip makers who choose available mode types based on given conditions or based on utility criteria. MNL model is often used to compare with other techniques, due to its ability in analyzing the trip maker behavior (Hensher, et. al, 2000). MNL model has been considered in this study to model choice behaviour of commuters in Delhi. The mathematical framework of logit models is based on the theory of utility maximization (Ben-Akiva and Lerman, 1985). Probability of an individual "i" selecting a mode "n", out of "M" number of total available modes, is given in equation (1)

$$\boldsymbol{P_{in}} = \frac{\boldsymbol{e^{V_{in}}}}{\sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \boldsymbol{e^{V_{im}}}} \tag{1}$$

Where, V_{in} is the utility function of mode "n" for individual "i", V_{im} is utility function of any mode "m" in the choice set for an individual "i". P_{in} is the probability of individual "i" selecting mode "n". *M is* the total number of available travelling modes in the choice set for individual "i". However, the Logit model has certain drawbacks like requirement of large sample size and restriction on dependent variable to be of discrete dataset.

4.2 Random Forrest Decision Trees

A Random Forrest Decision Tree is a tree constructed randomly from a set of possible trees with random features at each node. “At random” implies that in the set of trees each tree has an equal chance of being sampled, i.e. trees have a “uniform” distribution. Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models. Random Forest is one of the most efficient methods for classification and regression in data mining. It can classify an object or an instance to a predefined set of classes based on their attributes values such as age or gender. A Decision Tree starts from the root and moves downward. The starting point of the tree is called a root node while where the chain ends is known as the “leaf” node. Different branches can be extended from each internal node, as illustrated in Figure 2. A node represents a certain characteristic while the branches represent a range of values (Ali, et. al. ,2012). The algorithm of RF method is briefly discussed in the following. Let N be the number of trees to build. For each of N iterations the algorithm is briefly explained as follows and the algorithm explained through flow chart in Figure 3.

- *Selection of Sample Data*: Sample data set for model training is to be selected using bootstrap method. For each tree a bootstrap sample of the same size as the training data is created.
- *Growing the tree*: The tree is fully grown on this bootstrap using splitting rules. The tree is left un- pruned.
- *Attribute selection*: Only a random subset of the available features of defined size is considered for each node

- Pruning is not performed and the tree is saved as it is. This tree can be deployed for classifying some other data.
- *Output:* The variable vector is supplied as input to each of the trees in the forest where each tree gives a classification result (referred to as trees 'votes' for a class). The forest chooses the classification having the most votes (over all the trees in the forest). Overall prediction is given as majority vote (classification) from all individually trained trees.

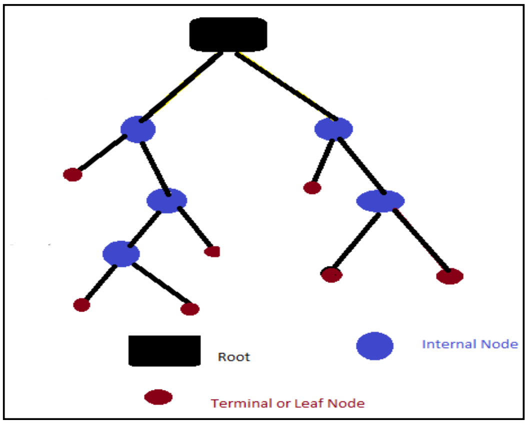

Figure 2.Tree Structure

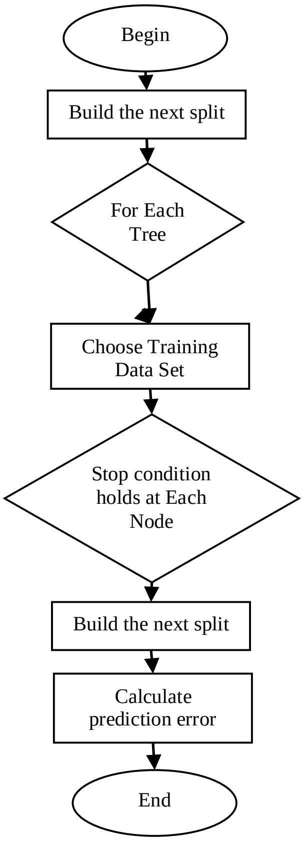

Figure 3.Flow Chart for Random Forrest Decision Trees

In this study, WEKA software (WEKA 6.3.9) has been considered for performing the mode choice analysis using MNL and RF based DT model and the results obtained by these models are discussed in the following section.

5. Results And Discussions

The result of mode choice analysis in terms of prediction accuracy calculated from MNL model and Random forest decision tree model is presented in Table 1. The prediction accuracy of MNL model is 77.31% while that by Random Forest model is 98.96%. The very high prediction rate of Random forest is due to the high data classifying capability due to the tree structured classifiers. The various statistical measures that are employed to measure the statistical significance of the models are Kappa statistic, mean absolute error, Root mean squared error and Relative absolute error The Kappa statistics compares observed accuracy with expected accuracy (random chance). It is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves also taking into account random chance. Landis and Koch proposed kappa values of 0-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as almost perfect. Thus based on this the RFDT model shows almost perfect results with a kappa statistics value of 0.986. Considering the Mean absolute error, it is a quantity which is used to measure how close predictions are to the eventual outcomes. A lower value of mean absolute error implies better predictability on part of model. RFDT model attains a much lower value of 0.0293 as compared to a value of 0.0783 attained by MNL model.

The root mean squared error (RMSE) is a quadratic scoring rule which measures the average magnitude of the error in prediction. A lower score of RMSE is better in RFDT and thus RFDT model depicts a value of 0.0855 compared to a value of 0.1975 of MNL model. The relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor; its value ranges from 0 to infinity, with 0 corresponding to the ideal. Thus RFDT depicts a value of 0.155 which is closer to ideal value of 0 as contrast to a value of 0.414 by MNL model. Table 2 and Table 3 summarize the cross classification of MNL and RFDT models respectively.

Validation of the two models was done by using 507 samples. The validation results are summarized in table 4. The results show better performance of RDFT model over MNL in validation phase as well.

Table 1 Summary of Results of MNL and RFDT Mode Choice Models in training

	MNL Mode Choice Model	RFDT Mode Choice Model
Correctly Classified Instances	(3847 instances) 77.31 %	(4924 instances) 98.96%
Incorrectly Classified Instances	(1129 instances) 22.68 %	(52 instances) 1.05 %
Kappa statistic	0.6961	0.9862
Mean absolute error	0.0783	0.0293
Root mean squared error	0.1975	0.0855
Relative absolute error	0.414	0.155

Table 2 Prediction Accuracy of MNL Mode Choice Model in training

Observed	Predicted Mode Choice							
Mode Choice	Two Wheeler	Bus	DA Car	Walk	Metro	Carpool	Bicycle	Auto Rickshaw
Two	839	42	379	0	4	0	1	0

Wheeler								
Bus	53	870	19	0	15	0	0	0
DA Car	357	14	1321	0	4	0	0	0
Walk	0	0	0	771	0	0	0	0
Metro	3	88	12	0	22	0	0	0
Carpool	21	1	56	0	0	0	0	0
Bicycle	5	1	0	0	0	0	23	0
Auto Rickshaw	3	46	5	0	0	0	0	1

Table 3 Prediction Accuracy of RFDT Mode Choice Model in training

Observed Mode Choice	Predicted Mode Choice							
	Two Wheeler	Bus	DA Car	Walk	Metro	Carpool	Bicycle	Auto Rickshaw
Two Wheeler	1255	1	9	0	0	0	0	0
Bus	1	956	0	0	0	0	0	0
DA Car	22	0	1674	0	0	0	0	0
Walk	0	0	0	771	0	0	0	0
Metro	0	4	0	0	121	0	0	0
Carpool	3	0	4	0	0	71	0	0
Bicycle	0	0	0	0	0	0	29	0
Auto Rickshaw	1	7	0	0	0	0	0	47

Table 4 Summary of Results of MNL and RFDT Mode Choice Models in validation

	MNL Mode Choice Model	RFDT Mode Choice Model
Correctly Classified Instances	(396 instances) 78.01 %	(414 instances) 81.65%
Incorrectly Classified Instances	(111 instances) 21.89 %	(93 instances) 18.35 %
Kappa statistic	0.7058	0.7537
Mean absolute error	0.0772	0.0741
Root mean squared error	0.1929	0.1927
Relative absolute error	0.408	0.392

6. Summary Of Findings

This study focused on the mode choice analysis of Delhi which is subjected to heavy congestion and air pollution due to high number of private vehicles plying on the roads. The data collection was done through a household survey in Delhi. A large household survey sample of 5000 responses was collected. A Random Forrest Decision

Tree (RFDT) mode choice model and a Multinomial Logit (MNL) mode choice model were developed. From the result, it was observed that model developed by Random Forrest based Decision Tree model is superior with higher prediction accuracy (98.96%) than the Multinomial Logit models having prediction accuracy of 77.31%. Model validation was performed and results obtained show better performance of RFDT model over MNL model. RFDT model has a prediction accuracy of 81.65% and MNL model has a prediction accuracy of 78.01% in validation.

The results demonstrate the advantages that Decision Trees have over Logit models. First of all it is one of the most accurate learning algorithms available producing highly accurate classifier. Also it runs efficiently on large databases and can handle thousands of input variables. It generates an internal unbiased estimate of the generalization error as the forest building progresses. There is repeatability in using Random forest as the generated forests can be saved and deployed for future use on other data.

Acknowledgement

Authors are very grateful to the Director, CSIR-Central Road Research Institute (CRRI) for allowing to publish this paper. Also this study acknowledges the resource of data collected for SUSTRANS project, a 12th Five Year Plan Network Project of CRRI.

References

Ali, J., Khan,R., Ahmad,N., and Maqsood,I., (2012), “Random Forests and Decision Trees”, International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3.

Ben-Akiva, M. E. and Lerman, S. R., (1985), “*Discrete Choice Analysis: Theory and Application to Travel Demand”*, The MIT Press, Cambridge, Massachusetts, the USA.

Black, WR 1995, 'Spatial interaction modeling using artificial neural networks', Journal of Transport Geography, vol. 3, no. 3, pp. 159-166.

Breiman, L. (2001) Random forests. *Machine Learning*, Vol. 45, pp. 5–32.

Cybenko, G., (1989) Math. Control Signals Systems, 4, 303–312. Dougherty, M. (1995). A review of neural Networks applied to transport. Transportation Research, 3C, 247-260

Hasegawa, H., Naito, T., Arimura, M., and Tamura, T. (2012) Modal choice analysis using ensemble learning methods. *Journal of Japan Society of Civil Engineering*, Vol. 68, No. 5, pp. 773–780, (in Japanese).

Hasegawa, H., Naito, T., Arimura, M., and Tamura, T. (2013), Hybrid Model of Random Forests and Genetic Algorithms for Commute Mode Choice Analysis, Proceedings of the Eastern Asia Society for Transportation Studies, Vol 9.

Haleem, K., Abdel-Aty, M., and Santos, J. (2010) Multiple Applications of Multivariate Adaptive Regression Splines Technique to Predict Rear-End Crashes at Unsignalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2165,pp. 33–41, December.

Hall, M., Frank,E., Holmes, G., Pfahringer, B.,,Reutemann,P., Witten, I.,H., (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Hensher, D. A. and T. Ton (2000). TRESIS: A transportation, land use and environmental strategy impact simulator for urban areas. Transportation 29(4): 439-457.

Hossain, M. and Muromachi, Y. (2011) Understanding Crash Mechanisms and Selecting Interventions to Mitigate Real-Time Hazards on Urban Expressways. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2213, pp. 53–62, December.

Hornik, K. S. (1989). Multilayer feed forward Networks are Universal Approximators. Neural Networks, 2, 359-366

Karlaftis M.G., Golias I. (2001) An International Comparative Study of Self-Reported Driver Behavior, Transportation Research part F, Vol. 4, Issue 4, pp 243-256.

Nijkamp, P. a. (1996). Modelling inter-urban transport flows in Italy: A comparison between Neural Network analysis and logit analysis. Transportation Research C, (pp. 323-338).

Pande, A., Das, A., Abdel-Aty, M., and Hassan, H. (2011) Estimation of Real-Time Crash Risk. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2237, pp. 60–66, December

Subba Rao, P.V.(1998).Another insight into artificial neural networks through behavioural analysis of access mode choice. Computers, Environment and Urban Systems, Volume 22, Issue 5, 1 September 1998, Pages 485-496.

Xie, C., Lu, J., Parkany, E. (2003) "Work Travel Mode Choice Modeling Using Data Mining: Decision Trees And Neural Networks," Transportation Research Record: Journal of the Transportation Research Board, No. 1854.

Zhang, G, Patuwo, BE & Hu, MY(1998). 'Forecasting with artificial neural networks: The state of the art', International Journal of Forecasting, vol. 14, no. 1, pp. 35-62.