Decision tree analysis of commuter mode choice in Baguio City, Philippines

I H V Gue$^{1,2}$, J Soliman$^1$, M De Guzman$^3$, R Cabredo$^4$, A Fillone$^5$, N S Lopez$^{1*}$ and J B M Biona$^{1,6}$

$^1$Mechanical Engineering Department, De La Salle University, Manila, Philippines
$^2$Mechanical Engineering Department, Far Eastern University – Institute of Technology, Manila, Philippines
$^3$Civil Engineering Department, Saint Louis University, Baguio City, Philippines
$^4$Software Technology Department, De La Salle University, Manila, Philippines
$^5$Civil Engineering Department, De La Salle University, Manila, Philippines
$^6$Enrique K. Razon Logistics Institute, De La Salle University, Manila, Philippines

*neil.lopez@dlsu.edu.ph

Abstract. Transportation is a multidisciplinary system. Solving its issues would require the knowledge of social, economic, engineering, environmental, and technological disciplines. Emerging techniques used in problem-solving involve the use of machine learning techniques. In this study, a machine learning technique, decision tree, is used for mode choice analysis in Baguio City, Philippines. Using data from a household survey, the developed model uncovers the most significant factors affecting mode choice of residents in the city. The results highlight the role of income, which is related to the individuals' career level and stage in life. Interestingly, a mid-level income group seems to be highly inclined towards private vehicle use. To conclude, the authors note that the primary advantage of a decision tree is its simplicity and straightforward results interpretation, which is paramount in policymaking. For future work, the authors recommend exploring larger decision tree models for mode choice and conducting a validation interview of the insights obtained from the study.

1. Introduction

Recent research works are gaining interest in using machine learning (ML) techniques for modeling and pattern recognition. ML has applications in sustainability, including urban transportation [1]. ML algorithms such as Artificial Neural Network (ANN), Support Vector Machines (SVM), Cluster Analysis (CA), and Decision Trees (DT) are among the considered alternatives for choice modeling. Among the enumerated algorithms, DT shows significant potential as existing studies have indicated its superior performance. Paredes et al. [2] compared the different ML algorithms and discrete choice models. Their findings indicated better performance of DT after proper data treatment. Meanwhile, the review of Pineda-Jaramillo [3] identified that the random forest (RF) algorithm, a subset of DT, is better than the other algorithms and the multinomial logit model (MNL).

The application of DT for choice modeling focuses on the locality of transportation mode selection. Chapelau et al. [4] applied RF for commuter choice in Montreal, Canada. Their work attained accurate estimations on commuter preference for private cars, public vehicles, school buses, and walking.
Additionally, Pawar et al. [5] assessed transportation mode selection under restrictions caused by the COVID-19 pandemic. They used DT in modeling commuter preference in India. DT models developed from prior studies are limited to constraints in the locality due to differing socio-cultural characteristics. The models are not universally applicable; therefore, utilization of DTs requires data reflecting local socio-cultural characteristics. Ranosa et al. [6] studied the willingness of Baguio City residents to walk and their understanding of proper public transport stops. On the other hand, Estrella et al. [7] assessed the acceptability of a new transport mode, the Aerial Ropeway Transport, in Baguio City. It also looked into the current transport characteristics of the city to determine a suitable route for the new mode. From a more general perspective, Rith et al. [8] analyzed the relationship between household vehicle ownership and energy use, while on [9], the authors looked into the effects of individual type choice on energy consumption in Metro Manila.

Contributing to the present literature on transport mode choice modeling and machine learning, this study develops a DT model revealing commuter preference between public utility vehicles (PUVs) and private vehicles. The authors considered DT because of its high accuracy prediction and its ability to generate linguistic rules. The linguistic rules provide insights into the local commuter preference. Findings and insights generated by this work may guide the development of action plans for commuter switching behavior. The results also provide comparative preferences among different socio-cultural systems.

2. Methodology

2.1. Study Area

A household survey was developed to build a transport database for Baguio City. Participants (N = 3216) were randomly selected, while the number of participants per barangay was proportional to the ratio of barangay population relative to city population. The survey consisted of questions on socio-economic, origin-destination, and mode share parameters. However, the present study focuses on the socio-economic (income, number of household cars, etc.) and mode share (private or public) parameters only. It is important to take note that taxi services are also considered private modes in this study. Participant profile was compared to the 2015 Family Income and Expenditure Survey (FIES) of Baguio City [10]. A notable difference was observed in the percentage of lower age intervals, mainly because the present survey design only considered participants who are at least seven years old. Otherwise, participant profiles of other parameters such as monthly household income, number of cars in the household, and mode share were similar for both surveys. Hence, the data used in the present study was deemed sufficient to represent the population of Baguio City. According to Kotsiantis [11], rule-based algorithms perform better under discretized data; therefore, the data have been discretized as shown in Table 1.

2.2. Decision Tree

Morgan & Sonquist [12] was the first to present DT’s initial variant, the Automatic Interaction Detection (AID) algorithm. The family of DT then grew and included algorithms such as Classification and Regression Trees (CART) [13] and C4.5 [14].

DT’s general concept is partitioning a dataset to form a top-down structure of if-then rules describing recognized patterns. Pruning is then applied to the tree to avoid overfitting. Kotsiantis [15] provides a detailed pseudo code on the iterative partitioning of DT’s growth phase. The initial step is in determining whether the current dataset has observations of the same category. If so, then a node is assigned according to the category. If not, the dataset is partitioned based on an identified feature. The feature is chosen based on its contribution to further partitioning. The algorithm creates a branch for each of the feature’s values. From the branch, the process recursively iterates from the start of forming subtrees.

This study utilized the Statistics and Machine Learning Toolbox of MATLAB in the development of the DT model. The command fitctree of MATLAB uses the CART algorithm in growing the tree. We determined the required number of observations per node with the least classification loss from k-fold
cross-validation for better generalization. The minimum number of observations per node was 75 with 0.215 classification loss. A 70:30 ratio was then applied to allocate the training and testing dataset.

Table 1. Participant Profile.

| Demographic                        | No. of Participants | % of Participants | % Baguio* | Source |
|------------------------------------|---------------------|-------------------|-----------|--------|
| **Age**                            |                     |                   |           |        |
| < 25                               | 293                 | 9                 | 53        | PSA [10] |
| 25 to 29                           | 675                 | 22                | 9         |        |
| 30 to 34                           | 530                 | 17                | 7         |        |
| 35 to 39                           | 495                 | 16                | 6         |        |
| 40 to 44                           | 282                 | 9                 | 5         |        |
| 45 to 49                           | 279                 | 9                 | 5         |        |
| > 49                               | 572                 | 18                | 14        |        |
| **Monthly Household income**       |                     |                   |           |        |
| < 10,000                           | 100                 | 3                 | 4         | PSA [10] |
| 10,000 to 20,000                   | 529                 | 17                | 25        |        |
| 20,000 to 30,000                   | 727                 | 23                | 25        |        |
| 30,000 to 40,000                   | 591                 | 19                | 17        |        |
| 40,000 to 50,000                   | 351                 | 11                | 9         |        |
| > 50,000                           | 828                 | 26                | 20        |        |
| **Number of cars in the household**|                     |                   |           |        |
| 0                                  | 2145                | 69                | 83        | PSA [10] |
| 1                                  | 730                 | 23                | 13        |        |
| > 2                                | 251                 | 8                 | 4         |        |
| **Main mode of transportation**    |                     |                   |           |        |
| Private                            | 941                 | 30                | 27        | PSA [10] |
| Public                             | 2185                | 70                | 73        |        |

*Baguio City data is from 2015 FIES (N = 362).

3. Results and Discussion

The DT model attained accuracies of 78.53% during training and 75.53% during testing; therefore, the model achieved generalization of the datasets. Overall, the model has an accuracy of 77.63%. Figure 1 shows the confusion matrices of the DT model.

Figure 2 depicts the DT for the transportation mode selection. The values indicated below each node represent its classification counts, e.g., Node 1 correctly classified 444 observations as ‘private’ from the training dataset. The accuracy in the classification of each node differs from one another. Nodes 1 and 2 have the highest accuracy rating as Node 1 correctly classified 82.75% of its observations from both datasets, while node 2 attained 79.43% accuracy. Notably, node 4 attained an accuracy of 61.46%
during training and 53.90% for both datasets. The result indicates that datasets within the node are nearly homogenous.

![Decision Tree Model](image)

**Figure 2.** Decision tree model of transportation mode selection

To further understand the decision tree model, the age distribution of each classification of participants with no family car is illustrated in Fig. 3. It is assumed that the personal income level may explain an individual’s decision to either use public or private transport. Participants with a personal income of less than Php 15,000 consist of individuals who are in the early stage of their career. This group tends to use public transportation due to their limited personal income. Likewise, participants with a personal income of Php 25,001 to Php 40,000 consist of individuals who are in the middle stage of their career. Although this group is more financially capable to use private transport, it is also susceptible to increasing financial responsibilities since individuals tend to start a family at this stage.

An interesting result in Fig. 2 is the inclusion of participants with a personal income of Php 15,001 to Php 25,001 together with those with a personal income of Php 40,001 to Php 90,000. Based on Fig. 3, this group is mostly composed of individuals who are between the early-stage and middle stages of their careers. This group tends to accumulate wealth as they gain more personal income without any significant financial burden. It is also observed that the age distribution of participants with a personal income of Php 15,001 to Php 25,001 is similar regardless of family income level. This is a possible explanation for the homogeneity of datasets in Node 4. Lastly, participants with a personal income of Php 40,001 to Php 90,000 consist of individuals who are in the late stage of their career. This group tends to accumulate more wealth due to higher personal and family incomes. Hence, the group may tend to use private transport more than public transport.

4. Conclusion

In this brief paper, the authors had demonstrated the potential application of DT models in mode choice modeling. While traditional mode choice models based on regression techniques can offer more insights, the advantage of DT is its simplicity and straightforward interpretation of results, which is paramount
in policymaking. The results of this study provide interesting insights into the role of income to private or public mode choice selection, which seems to be related to an individual’s career level and life stage. As previous studies have shown, having children and starting a family can affect a person’s mobility choices.

For future work, larger DT models can be developed to consider other potential factors affecting this decision. Moreover, personal interviews may also be carried out to validate the insights obtained from the results.

![Figure 3](image-url)  
**Figure 3.** Age distribution of participants with no family car

**Acknowledgments**

The authors would like to express their gratitude to the Commission on Higher Education, Philippines for their generous funding in the conduct of this work, through the Grants-in-Aid program. The authors also received support from the University Research Coordination Office of De La Salle University, Philippines, through its Interdisciplinary Research Grants.
References

[1] Gue, I H V, Ubando A T, Tseng M-L and Tan R 2020 Artificial neural networks for sustainable development: a critical review. Clean Technologies and Environmental Policy, (0123456789). https://doi.org/10.1007/s10098-020-01883-2

[2] Paredes M, Hemberg E, O’Reilly U-M and Zegras C 2017 Machine learning or discrete choice models for car ownership demand estimation and prediction? 2017 5th IEEE Int. Conf. on Models and Technologies for Intelligent Transportation Systems (MT-ITS) pp 780–785 IEEE. https://doi.org/10.1109/MTITS.2017.8005618

[3] Pineda-Jaramillo J D 2019 A review of Machine Learning (ML) algorithms used for modeling travel mode choice DYN A 86(211) 32–41 https://doi.org/10.15446/dyna.v86n211.79743

[4] Chapleau R, Gaudette P and Spurr T 2019 Application of machine learning to two large-sample household travel surveys: A characterization of travel modes. Transportation Research Record: Journal of the Transportation Research Board, 2673(4), 173–183. https://doi.org/10.1177/0361198119839339

[5] Pawar D S, Yadav A K, Akolekar N and Velaga N R 2020 Impact of physical distancing due to novel coronavirus (SARS-CoV-2) on daily travel for work during transition to lockdown Transportation Research Interdisciplinary Perspectives 7 100203 https://doi.org/10.1016/j.trip.2020.100203

[6] Ranosa L, Villaraza, C M, Fillone A, Guillon M D and Biona J B M 2017 Providers’ and users’ assessment of public transport and facilities: Baguio City, Philippines 2017 Proceedings of the Eastern Asia Society for Transportation Studies 11

[7] Estrella A F, Mendoza A K, Mendoza M A, Sanchez K, Fillone A, Paringit M C and Roxas N 2017 Impact analysis of aerial ropeway transport system as a form of mass transportation in Baguio City Proceedings of the 24th Annual Conference of the Transportation Science Society of the Philippines

[8] Rith M, Roquel K I D Z, Lopez N S A, Fillone A M and Biona J B M 2020a Towards more sustainable transport in Metro Manila: A case study of household vehicle ownership and energy consumption Transportation Research Interdisciplinary Perspectives 6 100163.

[9] Rith M, Lopez N S, Fillone A M and Biona J B M 2020b Determinants of individual vehicle type choice and energy consumption in a heavy traffic metropolis of Southeast Asia featuring the case of Metro Manila Engineering and Applied Science Research 47(1) 56-65.

[10] Philippine Statistics Authority (PSA) (2017). 2015 Family Income and Expenditure Survey (FIES). https://psa.gov.ph/sites/default/files/FIES%202015%20Final%20Report.pdf

[11] Kotsiantis S B 2007 Supervised machine learning: A review of classification techniques Informatica 31 pp 249–268.

[12] Morgan J N and Sonquist J A 1963 Problems in the analysis of survey data, and a proposal Journal of the American Statistical Association 58(302) p 415 https://doi.org/10.2307/2283276

[13] Breiman L, Friedman J, Stone C J and Olshen R A 1984 Classification and regression trees CRC press

[14] Quinlan J R 1992 C4.5 Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

[15] Kotsiantis S B 2013 Decision trees: a recent overview Artificial Intelligence Review 39(4) pp 261–283. https://doi.org/10.1007/s10462-011-9272-4