

# Journal of Transportation and Logistics

Research Article

 Open Access

## Modeling mode choice behaviors of commuters in car-dependent small country discrete choice models: A case study of Bahrain



Marwa Jazi Ghareibeh<sup>1</sup>  & Uneb Gazder<sup>2</sup>  

<sup>1</sup> University of Bahrain, Department of Civil Engineering, Sakhir, Bahrain

<sup>2</sup> University of Bahrain, Department of Civil Engineering, Sakhir, Bahrain

### Abstract

The aim of this study is to determine the factor affecting the mode choice of travelers in Bahrain, which presents a unique case due to its smaller area size and current dependence on cars. Hence, the need for promoting sustainable modes of transportation is critical for the country. The study used 3864 diverse data records extracted from traveler surveys. This data comprised of revealed preference responses. The variables considered in the modelling included traveler characteristics, and trip information. The logit model and the classification tree models were used to predict the mode choice, considering the currently available modes of transportation currently available (Car and Bus). The accuracy of the models was ascertained through a validation sample collected independently from the initial sample. Trip cost was the most influential factor on mode choice. Other important variables included direct and quick travel, accessibility, and convenience. In terms of model performance, the logit model demonstrated higher accuracy than the classification tree when modeling binary responses. The models and results of this study provide important conclusions for the transportation authorities, which can be utilized for developing and promoting sustainable transportation modes in Bahrain.

### Keywords

Mode choice · Prediction · Models · Logit Model · Classification Tree · Behavior



“ Citation: Ghareibeh, M. J. & Gazder, U. (2025). Modeling mode choice behaviors of commuters in car-dependent small country discrete choice models: A case study of Bahrain. *Journal of Transportation and Logistics*, 10(1), 59-74. <https://doi.org/10.26650/JTL.2025.1527557>

Ⓞ This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License. 

© 2025. Ghareibeh, M. J. & Gazder, U.

✉ Corresponding author: Uneb Gazder [ugazder@uob.edu.bh](mailto:ugazder@uob.edu.bh)



## Modeling mode choice behaviors of commuters in car-dependent small country discrete choice models: A case study of Bahrain

Transportation systems in most countries are currently facing serious challenges such as congestion, accidents, and air and noise pollution (Ortúzar and Willumsen, 2011). These challenges arise due to the rapid population growth and increase in car ownership levels, which have consequently led the demand for these systems to surpass the available supply of transportation facilities and services. Therefore, it becomes necessary to establish efficient policies and to have a thorough comprehension of travel demands, travel patterns, and driver characteristics for the development of projects that can facilitate and promote the advancement of transportation systems (Mwale et al., 2022).

Travel forecasting models are a crucial aspect of transportation planning and serve as a measure to identify the travel needs of cities (Sowjanya et al., 2014). These models consist of a set of mathematical equations and algorithms which are employed in a stage-wise or activity-based manner to simulate travel patterns and behaviors (Waghmare et al., 2022). One of the most notable models used for demand forecasting in transportation planning is the modal split model. This modeling operation, also known as mode choice modeling, focuses on predicting and managing the travel demands of different transport modes in a specific system by determining the key variables affecting the mode choice decision-making process (Pineda-Jaramillo, 2019).

In the Kingdom of Bahrain, urban development and population growth have resulted in a substantial increase in car ownership, leading to further congestion on Bahrain's road network. The number of cars in Bahrain increased from approximately 400,000 in 2009 to approximately 700,000 at the end of 2019 (Waleed, 2019). This increase, accompanied by scarcity and the near absence of a public transportation system, made it improbable to navigate through the country without traffic delays, reaching several hours during peak flow conditions.

Most of the efforts related to transportation modeling and innovative planning ventures are directed toward large countries and metropolitan cities. Therefore, this research aims to analyse the mode choice behavior of travelers and determine the factors influencing it in Bahrain. The results of this study are expected to highlight the uniqueness of travel choices in a country such as Bahrain, which has heavy car-dependent travel choices and a smaller size and population, where public transport does not seem to be an economically feasible option. The results could be applied to other countries facing similar challenges.

### Literature review

The economic growth of cities, which consequently boosts income and population growth, generates an increase in both passenger and freight demands, which necessitates the improvement of the current transportation system to satisfy the new travel demand while maintaining equilibrium (Modi et al., 2011; Waghmare et al., 2022). Travel demand is defined as the number of persons or vehicles per unit of time that can be predicted to use a certain segment of a transportation system under specific conditions, such as availability, quality, and cost (Ortúzar and Willumsen, 2011). Forecasting travel demand is a vital step in the transportation planning process, and because it is undergoing a continuous increase, more emphasis is given to understanding its relationship with factors affecting it, including mode choice (Hoel et al., 2011).

## Factors affecting mode choice

Mode choice and its modeling are considered the most prominent aspects of travel demand. In developing cities, the most commonly available transport modes are private cars, public busses, trains, taxis, walking, and cycling. Commonly, travelers tend to favor transport modes that fit their preferences and traveling habits best; thus, the factors that control their decision can be arranged into three categories (Chen et al., 2013).

These categories are as follows (Ratrou et al., 2014):

- Characteristics of trip makers: These are the socioeconomic and cultural aspects of a traveler, such as automobile ownership, income, employment, gender, age, and personal desires.
- Characteristics of the trip: Mainly concerning the purpose of the trip, the time of the day at which it is made, and the land use of the area in which it is made.
- Characteristics of the transport mode or service: travel time, cost, comfort, safety and availability.

Automobile ownership, public-transport availability and land-use are considered to have a larger influence on travelers' choice than the other factors mentioned above (Convery and Williams, 2019). Considering that the accurate prediction of travelers' mode choice behavior depends on several variables, it is crucial to select the right and most suitable approach to modeling (Chen et al., 2013).

## Distinct mode choice modeling

Transportation planning generally relies on discrete choice models to model travelers' mode choice behavior. These models employ the theory of utility maximization and assume that a traveler is most likely to choose the mode that presents the most benefits to its users (Hillel et al., 2019). The logit model, which uses simple mathematical techniques, is considered the most widely used distinct approach for mode choice modeling (Sekhar, 2014). According to this model, the utility function of a mode is demonstrated by a linear equation that links the factors influencing mode choice as independent variables and the utility of a particular mode as the dependent variable.

Equation (1) shows the basic form of the utility function for a traveler's mode choice while Equation (2) is used to calculate the probability of his/her choice (Puan et al., 2019):

$$U_m = C + A_1X_1 + A_2X_2 + \dots + A_iX_i \tag{1}$$

Where:

C=Constant

Ai= vector of coefficients

Xi=Independent Variables

$$P_{mi} = e^{U_{mi}} / (\sum_j e^{U_{mj}}) \tag{2}$$

Generally, based on the number of alternatives included in the model, a logit model can be either of two forms: The Binary Logit model or the Multinomial Logit model.

The binary logit model is used when commuters have two options of travel modes to choose from. The probability of choosing mode "A" in a binary logit model can be obtained through Equation 3, which is a simpler form of Equation 2, while Equation 4 represents the probability of choosing the other alternative (B) (Puan et al., 2019):

$$P_A = e^{U_A} / (1 + e^{U_A}) \tag{3}$$

$$P_B = 1 - P_A \tag{4}$$



Where:

$P_{(A,B)}$  = Probability of choosing mode A, B

$U_A$  = usefulness of alternative A

When commuters have more than two modes to choose from, the multinomial logit model is used to build the utility functions and estimate probabilities as per Equations (1-2) (Puan et al., 2019). The multinomial logit model is based on the theoretical assumptions that the error elements follow Gumbel distribution rather than normal distribution, that the error elements for each alternative are comparable and independently distributed, and that the error elements are similar and independently distributed for each observation (Hussain et al., 2017).

## Machine learning mode choice models

Due to the introduction of new transportation technologies, which allowed a wider range of data collection and remarkable developments in machine learning research, transportation planners began to explore more advanced alternative approaches (Hillel et al., 2019). Machine learning is a term used to refer to a group of algorithms that support computers in mechanizing data and using it to build models. These models overcome the challenges posed by growing travel demands by detecting statistical patterns (Bhavsar et al., 2017).

A key machine-learning approach is Decision trees. These nonparametric methods classify data into groups and model the relationships between features and possible outputs by following a structure identical to a flowchart or a tree (Pineda-Jaramillo, 2019). The most frequently used decision trees are those that use binary splits. To compute each split, the data is classified depending on the specified features, and then for each feature, the possible binary split points were tested (Hillel et al., 2019).

## Literature gaps

Most of the above-mentioned literature deals with countries/cities that have various mode choices that are used by a significant proportion of the population. However, Bahrain is a unique case due to its small size and dependence on car mode. Understanding mode choice behavior in such conditions is expected to provide a unique perspective to the research topic. This may also pave the way for the promotion and development of alternative sustainable transportation systems in Bahrain, as well as other car-dependent countries.

## Data collection

The previous literature on mode choice shows that information concerning gender, age, nationality, occupation, salary, driver's license, and car ownership appears in many of the previous studies as predictors of mode choice (McCarthy et al., 2017; Geng et al., 2016). Additionally, Origin-Destination, current mode, the purpose of the trip, travel time, and total cost of the trip were also observed to be common features in the previous studies as well as some recent studies such as by Mohamed and Oke (2023) and Ha et al. (2020). Hence, these were the factors considered in this study, and their data were collected in the survey. The details of these variables are given in Table 1. Note that some of the categories for questions needed to be merged due to the lack of responses, for example, responses for age categories above 45 years.

**Table 1**

*Standardization of data items.*

| <i>Data Item</i> | <i>Description</i>     |
|------------------|------------------------|
| Gender           | Binary: Male or Female |

| <i>Data Item</i>    | <i>Description</i>   |
|---------------------|--|
| Age                 | Categorical; "Under 18", "18 – 25", "26 – 35", "36 – 45" or "Above 45".  |
| Nationality         | This data item has two options: Bahraini and Non-Bahraini.   |
| Origin-Destination  | Categorical: Capital, Muharraq, Northern and Southern  |
| Occupation          | Categorical: employee, student, retired, unemployed, and others  |
| Average Salary      | Continuous   |
| Driving License     | Binary: Yes or No  |
| Car Ownership       | Categorical: 0, 1, 2, 3, and 3+.   |
| Purpose of the Trip | Categorical: work, education, shopping/leisure, and others.  |
| Travel Time         | Categorical: "0 – 10", "11 – 20", "21 – 30", "31 – 40". "41 – 50", "51 – 60", "61 – 120" and "120+".   |
| Total trip cost     | Continuous   |
| Current Mode        | Categorical: private cars, car-sharing, bus (public busses, school busses and private bus) and non-motorized transportation (walking and cycling). |

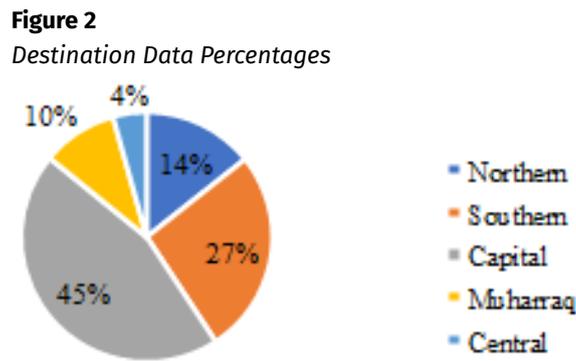
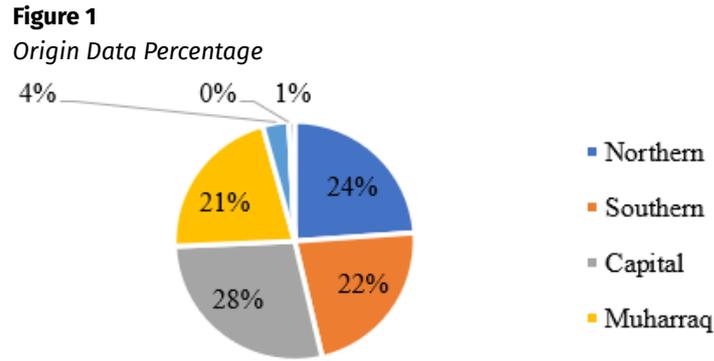
The data for this data was collected through online surveys which resulted in 3864 samples which could be considered for modelling. From these responses, filtering process was applied to eliminate the responses without the mode choice or those who had missing data for more than one of the variables. This step improved data performance by reducing data insufficiencies and streamlining the process. Table 2 and Figure 1 and Figure 2 highlight the key statistical metrics of the final dataset. The gender representation was almost equal for males and females in the final dataset. The dataset had higher responses from Bahrainis from the age group of 18-35 which could be considered an active population for any country. The dominance of car use among the responses is a true representation of travel mode choice behavior in Bahrain as well as in other neighboring countries (Mahmood et al., 2022). The responses for other modes were merged into one category to avoid model bias due to the large dominance of one mode in the choice set.

**Table 2**

*Key statistical metrics of the final dataset.*

| <i>Data Item</i> | <i>Statistical Metrics</i>         |
|------------------|------------------------------------|
| Gender           | M = 51%, F = 49%                   |
| Nationality      | Bahraini = 61%; non-Bahraini = 39% |
| Age              | Under 18 = 5%                      |
|                  | 18–25 = 37%                        |
|                  | 26–35 = 29%                        |
|                  | 36–45 = 14%                        |
|                  | Above 45 = 15%                     |
| Travel Mode      | Car = 71%                          |
|                  | Sharing Car = 9%                   |
|                  | Bus = 19%                          |
|                  | Non-motorized Transportation = 1%  |





## Model building

### Data preparation

Table 3 summarizes the scales used for each category. The categorical variables had to be coded as Minitab was used for modelling which does not take text variables as input for models.

**Table 3**  
Model Variables

| Category    | Variable        | Scale  |
|-------------|-----------------|--|
| Continuous  | Age             | Continuous Number*                                       |
|             | Travel Time     | Continuous Number  |
|             | Trip Cost       | Continuous Number  |
|             | Salary          | Continuous Number  |
| Categorical | Gender          | 1 for males and 0 for females                            |
|             | Nationality     | 1 for Bahraini, 0 for non-Bahraini                       |
|             | Origin          | Distinct Numbers (1 - 4) each representing a governorate |
|             | Destination     | Distinct Numbers (1 - 4) each representing a governorate |
|             | Current Mode    | 1 for a car, 2 for Others                                |
|             | Trip Purpose    | 1 for Work, 2 for Education, 3 for Shopping, 4 for Other |
|             | Occupation      | 1 for Employees, 2 for students, and 3 for others        |
|             | Driving License | 1 for Yes and 0 for No.                                  |
|             | Car Ownership   | Distinct Numbers (1 - 4)                                 |

\*The age categories were replaced with the middle or extreme values in the case of the highest or lowest category, respectively.



## The logit model

To develop a binary logit model, several models were built using various variables until an adequate goodness of fit and accuracy was achieved. The optimal model was selected by evaluating and testing different variables until the best combination was identified by the software. The proceeding sections provide more details about this model.

### Utility equations

The selected model uses five variables, three of which are of a continuous nature—travel time, trip cost, and salary—and two categorical variables: trip purpose and occupation. Table 4 presents the resulting utility functions for Car and Bus. Note that C12 and C19 refer respectively to the trip purpose and occupation. More details on these variables are presented in Table 3.

**Table 4**

Utility functions

| C12         | C19 | Car Utility  | Other Mode Utility  |
|-------------|-----|--|---|
| 1           | 1   | $Y = 0.4280 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$ | $Y' = 0.4280 + 0.01267 \text{ Average Travel Time} - 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$ |
| 1           | 2   | $Y = 2.111 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$  | $Y' = 2.111 + 0.01267 \text{ Average Travel Time} - 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$  |
| 1           | 3   | $Y = 1.389 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$  | $Y' = 1.389 + 0.01267 \text{ Average Travel Time} - 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$  |
| 2           | 1   | $Y = 1.379 - 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$  | $Y' = 1.379 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$  |
| 2           | 2   | $Y = 0.3040 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$ | $Y' = 0.3040 + 0.01267 \text{ Average Travel Time} - 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$ |
| 2           | 3   | $Y = 0.4178 - 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$ | $Y' = 0.4178 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$ |
| 3           | 1   | $Y = 2.448 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$  | $Y' = 2.448 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$  |
| 3           | 2   | $Y = 0.7652 - 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$ | $Y' = 0.7652 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$ |
| 3           | 3   | $Y = 1.487 - 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$  | $Y' = 1.487 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$  |
| 4           | 1   | $Y = 3.513 - 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$  | $Y' = 3.513 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$  |
| 4           | 2   | $Y = 1.830 - 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$  | $Y' = 1.830 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$  |
| 4           | 3   | $Y = 2.552 - 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} + 0.001054 \text{ Average Salary}$  | $Y' = 2.552 + 0.01267 \text{ Average Travel Time} + 0.9388 \text{ Total Cost of Trip} - 0.001054 \text{ Average Salary}$  |
| Probability |     | $P(\text{Car}) = \exp(Y) / (1 + \exp(Y))$  | $P(\text{Bus}) = \exp(Y') / (1 + \exp(Y'))$   |

### Coefficients analysis

By thoughtfully examining the coefficients' analysis, the importance of each variable in the model can be assessed. This is accomplished using Wald approximation tests to calculate the Z- and P-values of each variable. The Z-value computes the ratio between the coefficient and its standard error, indicating whether its estimate is large and precise (far from 0) or small and too imprecise (close to 0). In Table 5, it can be concluded that all coefficients have precise estimates that are not statistically equal to 0, since the Z-



values calculated are sufficiently far from 0. Successively, Minitab uses the Z-value to calculate the P-value, which is a probability that suggests that there is no association between the independent variable and the dependent variable. As commonly used in analysis, a 95% confidence level implies that if the P-value is less than or equal to 0.05, the variable is considered statistically significant. Accordingly, based on the P-values shown in Table 5, it can be deduced that there is no evidence for the null hypothesis; therefore, all variables used in the model are significant (Mendenhall et al., 2013).

**Table 5**  
*Binary Logit Model Coefficient Analysis*

| Term            | Coefficient | Standard Error | Z-Value | P-Value |
|-----------------|-------------|----------------|---------|---------|
| Constant        | -0.428000   | 0.204          | -2.100  | 0.035   |
| Travel Time     | -0.012670   | 0.004          | -3.030  | 0.002   |
| Total trip cost | 0.938800    | 0.082          | 11.420  | 0.000   |
| Salary          | 0.001054    | 0.000          | 4.330   | 0.000   |
| C12             |             |                |         |         |
| 2               | 1.807000    | 0.674          | 2.680   | 0.007   |
| 3               | 2.876000    | 0.714          | 4.030   | 0.000   |
| 4               | 3.940000    | 1.120          | 3.520   | 0.000   |
| C19             |             |                |         |         |
| 2               | -1.683000   | 0.680          | -2.480  | 0.013   |
| 3               | -0.961000   | 0.411          | -2.340  | 0.019   |

**Odds ratios**

Odds ratios are prominent in statistical analysis because they can be used to measure the odds of an outcome of interest, occurring with an association to a specific variable. In general, the odds ratios for continuous variables demonstrate the change in the odds of the dependent variable with a unit increase in the independent variable, whereas the odds ratios of categorical predictors compare the odds of the outcome of non-reference categories to that of the reference category. This helps identify the extent of each variable’s effect on the outcome and assists informed decision-making. Table 6 and Table 7 display the odd ratios for both continuous and categorical predictors used in the logit model (Kleinbaum et al., 1988). It can be observed that the greatest impact on the mode choice is due to the change in the trip purpose, especially when it changes from traveling to work.

**Table 6**  
*Odd Ratios for Continuous Predictors*

| Term                | Odds Ratio | 95% CI       |
|---------------------|------------|--------------|
| Average Travel Time | 0.98       | (0.97, 0.99) |
| Total trip cost     | 2.55       | (2.17, 3.00) |
| Average Salary      | 1.00       | (1.00, 1.00) |

**Table 7**  
*Odd Ratios for Continuous Predictors*

| Reference Level | Comparison Level | Odds Ratio | 95% CI        |
|-----------------|------------------|------------|---------------|
| C12             |                  |            |               |
| 2               | 1                | 6.09       | (1.62, 22.84) |
| 3               | 1                | 17.74      | (4.37, 71.94) |

| Reference Level | Comparison Level | Odds Ratio | 95% CI         |
|-----------------|------------------|------------|----------------|
| 4               | 1                | 51.46      | (5.75, 460.47) |
| 3               | 2                | 2.91       | (0.57, 14.68)  |
| 4               | 2                | 8.45       | (0.95, 74.80)  |
| 4               | 3                | 2.90       | (0.26, 31.36)  |
| C19             |                  |            |                |
| 2               | 1                | 0.18       | (0.04, 0.70)   |
| 3               | 1                | 0.38       | (0.17, 0.85)   |
| 3               | 2                | 2.05       | (0.61, 6.84)   |

**Goodness of fit**

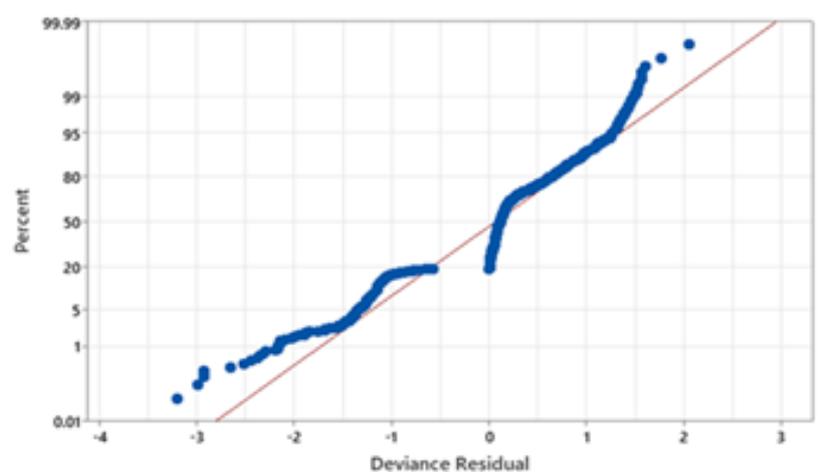
Predominantly, the goodness-of-fit tests’ results (Table 8) suggest that the model fits the data well, as both deviance and Pearson’s chi-square tests have a high P-value of 1. This implies that the model’s estimates are not significantly different from the observed data. However, since the Hosmer-Lemeshow test has a low P-value of 0, suggesting that there is some evidence of a lack of fit, possibly due to the large sample size used in the analysis (Kutner et al., 2004; Hadi and Chatterjee, 2015).

**Table 8**  
*Binary Logit Model Goodness-of-fit*

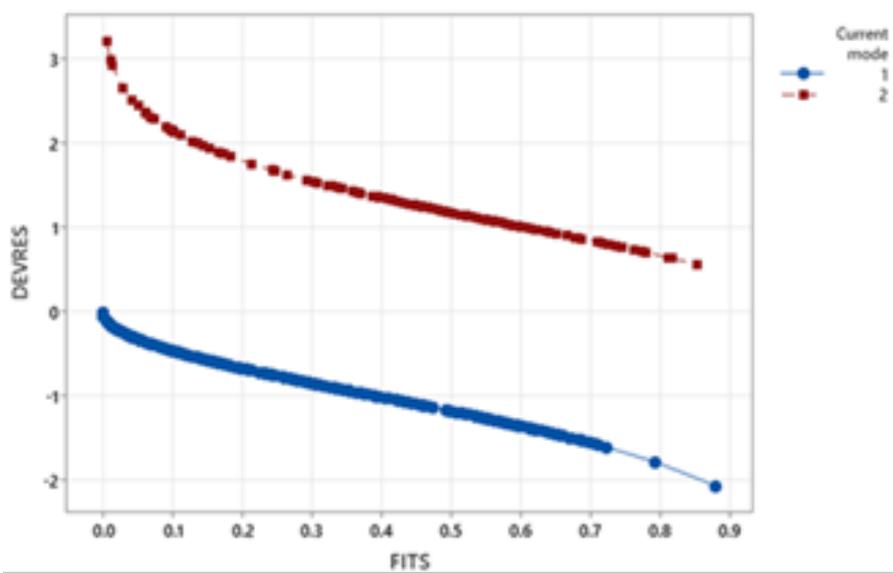
| Test                | Degree of Freedom | Chi square | P-Value |
|---------------------|-------------------|------------|---------|
| Deviance            | 1387              | 842        | 1.00    |
| Pearson             | 1387              | 1162       | 1.00    |
| Hosmer and Lemeshow | 8                 | 50         | 0.00    |

Therefore, to better evaluate the model’s goodness-of-fit, the normal probability plot and standard errors vs. fits were generated (Figure 3 and Figure 4). Based on the statistical principles and interpretation of diagnostic plots, the approximate linear form of both plots suggests that the model has an acceptable fit to the data and that the linearity assumption between the dependent and independent variables within the utility functions has been met (Kutner et al., 2004; Hadi and Chatterjee, 2015). Subsequently, despite the lack of fit provided by Hosmer-Lemeshow test, the other tests and diagnostic plots do not reject the model generated in this study.

**Figure 3**  
*Normal Probability Plot (percentage vs. Residual)*



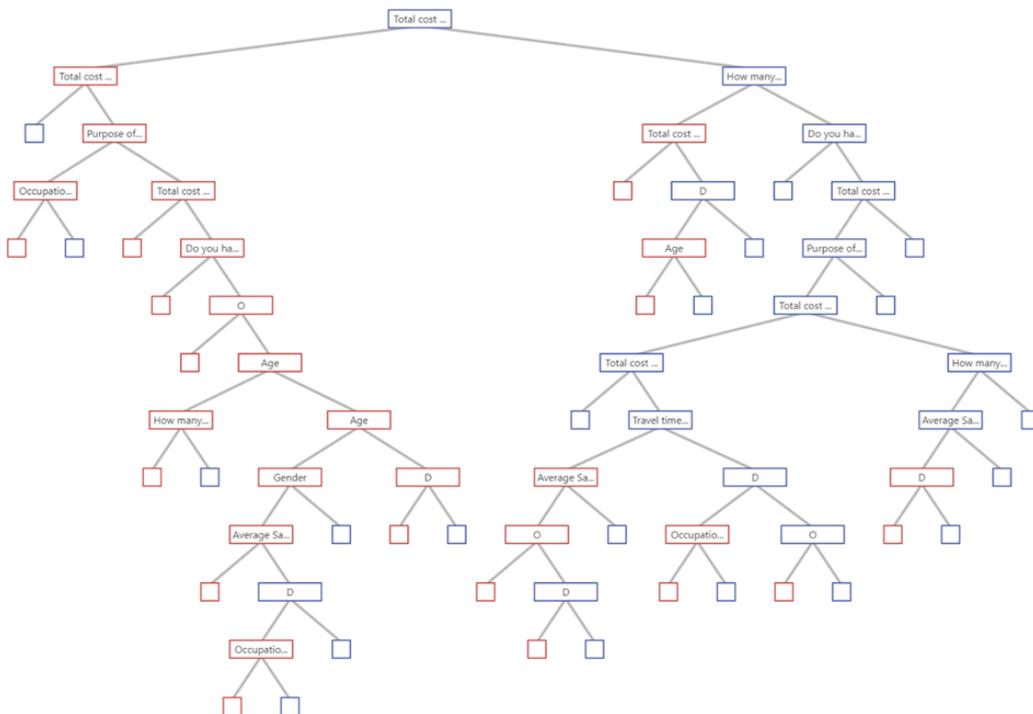
**Figure 4**  
Residuals vs. Fitted Value Plot (Devers vs. Fits)



### Classification tree model

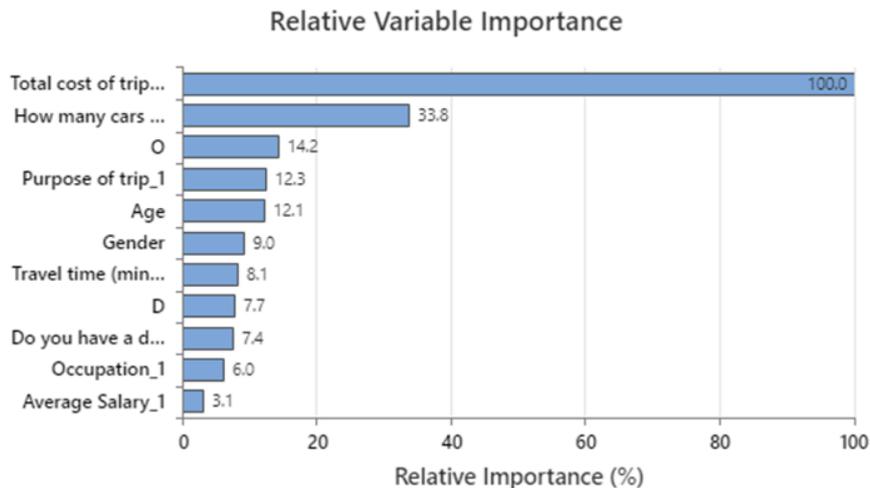
Similar to the logit model, Minitab was used to develop classification tree models. Various trees were generated using the software algorithm while calculating the misclassification cost (error) of each tree to determine the number of terminal nodes necessary to achieve the optimal model. Finally, the optimal model (35 terminal nodes) was identified to account for a misclassification cost of 0.2472, which is within 1 standard error of the minimum misclassification cost recorded. A simplified view of the classification tree is shown in Figure 5. More details about the extracted tree are provided in the subsequent figures and text.

**Figure 5**  
Simplified View of Classification Tree



The model was developed using 3804 responses, 3079 (80.94%) of which belong to Car and the rest for other mode users, respectively. It consists of 35 decisive nodes (out of which 20 represent Car mode and 15 represent other modes). The following variables were found to be affecting the model predictions, namely trip cost, car ownership, Origin-Destination, age, gender, purpose of the trip, travel time, driving license, occupation, and salary. Figure 6 demonstrates the relative variable importance, which is the percentage of improvement achieved when splits are made on a predictor concerning the top predictor (trip cost).

**Figure 6**  
*Relative Variable Importance*



*Variable importance measures model improvement when splits are made on a predictor. Relative importance is defined as % improvement with respect to the top predictor.*

### Split-half reliability test

The internal consistency of the data is an important issue which could greatly influence the model developed from it. One of the methods to check the internal consistency of the data is Split-half reliability test. In this test, the available dataset is randomly divided into two subsets and the correlation between each subset is taken as the measure of consistency. This test also ascertains if the data can be utilized for drawing generalized conclusions about the larger population (Nunnally and Bernstein, 1994).

This analysis was conducted by dividing the data into two samples. The correlation coefficients for each item—current mode, travel time, trip cost, and salary—were computed and found to be 0.63, indicating a moderate to high level of consistency (Nunnally and Bernstein, 1994).

### Initial model validation

To evaluate the predictive performance of a model, validation methods are crucial. Therefore, in this study, to ensure that both the logit and classification tree models were generated with adequate accuracy, a 10-fold cross-validation technique was used. This method involves dividing the dataset into 10 equal samples (folds), in which the model is trained on nine and tested on the remaining samples. The process is then repeated 10 times so that each fold serves as the test sample exactly once. As a result, ten varying estimates of the model's performance are generated, and by finding its average for the 10 folds, a more accurate estimate can be computed (Hastie et al., 2009).

The 10-fold cross-validation analysis conducted on the Logit model resulted in a 10-fold deviance R-sq of 34.24% and a 10-fold area under the receiver operating characteristics (ROC) curve of 0.8837. The area under

the curve (AUC) receiver operating characteristic curve (ROC) is commonly used to evaluate the performance of binary classifier models. It measures the model’s ability to distinguish between classes by plotting the true positive rate against the false positive rate at each threshold value. The area under this plot ranges from 0 to 1, where a value of 0.5 indicates a model with no predictive power and a value of 1 is considered perfect (Fawcett, 2006). Therefore, despite the relatively low deviance R-sq, the AUC ROC results suggest that the binary model has a good predictive power and can accurately distinguish between the two modes (car and other). In fact, the low deviance R-sq should not necessarily be interpreted as poor performance of the model, as it is a measure that is highly affected by the sample size and is generally used to compare different models alongside other measures rather than evaluating the performance solely on its basis (Fawcett, 2006).

For the classification tree, the 10-fold cross-validation analysis demonstrated that the model can accurately distinguish between positive and negative classes for both the training and testing datasets. The AUC ROC was calculated as 0.9505 for training data and as 0.9236 for testing data. In addition, the percentage of misclassified cases, which is the overall error rate, is largely insignificant, being as low as 11% for training and 13.4% for testing.

### Validation survey

The validity of the models was enforced by taking a small survey sample and applying the models to the samples of that survey. 49 diverse responses were collected in this validation survey.

Inputting the data collected from the validation survey into the models, we found that the logit model surpassed the classification tree with an accuracy of 92% compared to an accuracy of 81% for the latter. However, it should be noted that the validation sample only consisted of 49 samples; hence, the superiority of the logit model prediction can be enforced by applying it to a larger sample. Tablo veya Table 9 and Tablo veya Table 10 summarize the findings of the analysis. The accuracy of both models is satisfactory for evaluating the relationships between different variables and mode choices.

**Tablo veya Table 9**

*Logit Model Confusion Matrix (Validation Survey)*

| Actual/Predict | Car | Other |
|----------------|-----|-------|
| Car            | 41  | 4     |
| Public Bus     | 0   | 3     |
| Total Positive |     | 44    |
| Accuracy       |     | 92%   |

**Tablo veya Table 10**

*Classification Tree Model Confusion Matrix (Validation Survey)*

| Actual/Predict | Car | Other |
|----------------|-----|-------|
| Car            | 37  | 8     |
| Public Bus     | 1   | 2     |
| Total Positive |     | 39    |
| Accuracy       |     | 81%   |

## Results and Discussion

### Logit model results

After reviewing the odds ratios in Tables 6 and 7, the following summarizes the effects of both continuous and categorical variables on the transportation decisions. An increase in the travel time reduces the proba-



bility of choosing car mode over other modes. On the other hand, increased travel costs and salaries increase the odds of choosing a car over other modes. Trips related to education, shopping, and other purposes lead to higher odds of choosing a car over other modes compared to work-related trips. Moreover, shopping and other-purpose trips result in higher odds of choosing a car over other modes than education-related trips. Employees are more likely to choose a car over other modes than students and Others (unemployed, retired and others). Students are more likely to choose other modes when compared to others.

According to these observations, it can be concluded that trip cost is the most significant trip-related factor influencing the mode choice behavior of commuters in Bahrain. Similar findings were noted by (Kumar et al., 2004) in a study that focused on intercity bus services in India, where travel expenses were found to harm bus travelers' utility. Additionally, students have a higher likelihood of traveling by other modes compared to other occupations, and interestingly, trips that are produced for purposes other than work, education, and shopping are 51 times more likely to be done by car than work trips. Similar findings have been reported in previous studies, such as those by Daisy et al. (2018).

### **Classification tree model results**

The node rules extracted from the binary classification tree provided beneficial insights into the attributes of commuters in Bahrain and their trips, which can be used to improve the effectiveness of current transportation policies and systems.

Upon reviewing the rules, the following was noticed. Education and work are the primary purposes for car trips, which are mostly undertaken by commuters who own 1, 2 or 3 cars. Most car trips originate from the capital city. Other modes are mainly used for educational purposes. Surprisingly, most bus users have a driving license. This can be attributed to several reasons:

- o Having a driver's license gives individuals the flexibility to choose between different modes of transportation. Some may use a bus to avoid parking problems or high fuel costs but may possess a driving license as an alternative.
- o Young individuals who recently obtained their driving license may still prefer to travel by bus, especially if they have limited financial resources.
- o They may possess a driving license as a job requirement (taxi drivers, bus drivers, truck drivers, etc.) but cannot financially afford a car for personal use.

### **Policy implications**

These findings lead to some important policy implications that can be used for reducing the dominance of car use, consequently car ownership, and promoting other sustainable modes of transportation in Bahrain. More efforts are required to plan and provide cheap modes of transport that can serve longer distances. In terms of area, the focus should be more on the capital governorate, which also contains the diplomatic and other business-related areas of Bahrain. The results show that young people, especially students, and those with lower salaries have a higher likelihood of using other modes. Hence, future plans should strive to attract high-income people, consequently being older, to use other modes for their work-related trips. These trips eventually cause recurring congestion on highways (Roy et al. 2020). Hence, contemporary solutions, such as metro, autonomous mobility-on-demand, and mobility-as-a-service models, could be more appealing for these types of travelers instead of traditional bus or existing modes of transportation.

## Conclusions

This research focused on presenting the demanding need for thorough modeling of the current mode choice in Bahrain. Two types of models used in this research—the logit model and the classification tree model. The findings of this research provide important contributions to construct solid recommendations for promotion of sustainable transportation systems in Bahrain and other neighboring countries who face similar issues of car dominance.

The primary outcomes derived from this research are as follows:

- Trip production is distributed in almost equal proportions among the four governorates.
- The capital governorates most trips for various purposes, including business and diplomatic activities.
- The majority of trips are primarily undertaken for work, followed by shopping and education.
- Trip cost was identified as the most influential factor for mode choice in Bahrain.
- Car is likely to be preferred by commuters with high socioeconomic status.
- The travel time and speed of travel play a considerable role in shaping mode choice decisions; trips by bus account for longer travel times than car trips.
- Other modes of transport are primarily used by students for education-related trips.
- The accuracy of the logit model in modeling binary responses was proven to outperform that of the classification tree in predicting the choices in the validation survey.

The findings of this study identified the traveler groups and areas that are the major causes of the dominance of cars in mode choice. The focus for future planning should be on providing convenient and contemporary solutions, such as metro, autonomous mobility-on-demand, and mobility-as-a-service models, for minimizing recurring congestion on roads due to work trips. The findings also show that young students and people from low-income categories are likely to be more attracted to future alternative modes of transport. The findings of this study are consistent with those of previous studies. Hence, the recommendations of this study can be applied to a wider scale, especially including the neighboring countries.

On the basis of these findings, several future revenues are recommended to be explored. One recommendation is to extend the analysis to incorporate more qualitative variables, such as comfort, safety, and network characteristics. Additionally, studying the effects of transport measures on mode choice, such as public-transport subscriptions, road pricing, parking fees and congestion charges, could provide valuable insights. Another avenue worth investigating is the connection between sustainable transportation and other sustainable objectives such as health, climate change, and energy.



|                      |   |
|----------------------|---|
| Peer Review          | Externally peer-reviewed.   |
| Author Contributions | Conception/Design of Study- U.G.; Data Acquisition- M.J.G.; Data Analysis/Interpretation- M.J.G., U.G.; Drafting Manuscript- M.J.G., U.G.; Critical Revision of Manuscript- U.G.; Final Approval and Accountability- M.J.G., U.G. |
| Conflict of Interest | The authors have no conflict of interest to declare.  |
| Grant Support        | The authors declared that this study has received no financial support.   |

|                |  |
|----------------|--|
| Author Details | <b>Marwa Jazi Ghareibeh (MSc Student)</b><br><sup>1</sup> University of Bahrain, Department of Civil Engineering, Sakhir, Bahrain<br> 0009-0000-2671-8086  marwah.jazi@gmail.com |
|                | <b>Uneb Gazder (Asst. Prof.)</b><br><sup>2</sup> University of Bahrain, Department of Civil Engineering, Sakhir, Bahrain<br> 0000-0002-9445-9570  ugazder@uob.edu.bh             |

## References

- Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R., & Dera, D. (2017). Machine learning in transportation data analytics. In *Data Analytics for Intelligent Transportation Systems* (pp. 283-307). Elsevier.
- Chen, X., Liu, X., & Li, F. (2013). Comparative study on mode split discrete choice models. *Journal of Modern Transportation*, 21(4), 266-272.
- Convery, S., & Williams, B. (2019). Determinants of transport mode choice for non-commuting trips: the roles of transport, land use and socio-demographic characteristics. *Urban Science*, 3(3), 82.
- Daisy, N. S., Millward, H., & Liu, L. (2018). Trip chaining and tour mode choice of non-workers grouped by daily activity patterns. *Journal of Transport Geography* 69: 150-162.
- Fawcett, T. (2006). Introduction to receiver operator curves. *Pattern Recognit. Lett*, 27, 861-874.
- Geng, J., Long, R., & Chen, H. (2016). A review of the influencing factors of residents' travel mode choice. *Journal of Beijing Institute of Technology (Social Sciences Edition)*, (5), pp. 1-9.
- Ha, J., Lee, S., & Ko, J. (2020). Unraveling the impact of travel time, cost, and transit burdens on commute mode choice for different income and age groups. *Transportation Research Part A: Policy and Practice*, 141, 147-166.
- Hadi, A. S., & Chatterjee, S. (2015). *Regression Analysis by Example*. John Wiley & Sons.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: Springer.
- Hillel, T., Bierlaire, M., & Jin, Y. (2019). *A systematic review of machine learning methodologies for modeling passenger mode choice*. Technical Report TRANSP-OR 191025. EPFL.
- Hoel, L. A., Garber, Nicholas J., & Adek, A. W. S. (2011). *Transportation infrastructure engineering a multimodal integration* SI edition.
- Hussain, H. D., Mohammed, A. M., Salman, A. D., Rahmat, R. A. B. O. K., & Borhan, M. N. (2017). Analysis of transportation mode choice using a comparison of artificial neural network and multinomial logit models. *ARPN Journal of Engineering and Applied Sciences*, 12(5), 1483-1493.
- Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). Odds ratio. In *Applied regression analysis and other multivariate methods* (pp. 104-118). PWS-Kent.
- Kumar, C. P., Basu, D., & Maitra, B. (2004). Modeling generalized cost of travel for rural bus users: a case study. *Journal of Public Transportation*, 7(2), 59-72.
- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin.
- Mahmood, H., Asadov, A., Tanveer, M., Furqan, M., & Yu, Z. (2022). Impact of oil price, economic growth and urbanization on CO2 emissions in GCC countries: asymmetry analysis. *Sustainability*, 14(8), 4562.
- McCarthy, L., Delbosc, A., Currie, G., & Molloy, A. (2017). Factors influencing travel mode choice among families with young children (aged 0-4): a review of the literature. *Transport Reviews*, 37(6), 767-781.
- Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2013). *Introduction to probability and statistics*. Cengage Learning.



- Modi, K. B., Zala, L. B., Umrigar, F. S., & Desai, T. A. (2011, May). Transportation planning models: a review. In *National Conference on Recent Trends in Engineering and Technology*, Gujarat India.
- Mohammed, M., & Oke, J. (2023). Origin-destination inference in public transportation systems: A comprehensive review. *International Journal of Transportation Science and Technology*, 12(1), 315-328.
- Mwale, M., Luke, R. & Pisa, N. (2022). Factors that affect travel behavior in developing cities: A methodological review. *Transportation Research Interdisciplinary Perspectives*, 16, 100683.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.
- Ortúzar, Juan de Dios, & Willumsen, L. G. (2011). *Modeling Transport* (fourth edition), John Wiley & Sons, Ltd.
- Pineda-Jaramillo, J. D. (2019). A review of Machine Learning (ML) algorithms used for modeling travel mode choice. *Dyna*, 86(211), 32-41.
- Puan, O. C., Hassan, Y. A. H., Mashros, N., Idham, M. K., Hassan, N. A., Warid, M. N. M., & Hainin, M. R. (2019, May). Transportation mode choice binary logit model: A case study for Johor Bahru city. In *IOP Conference Series: Materials Science and Engineering* (Vol. 527, No. 1, p. 012066). IOP Publishing.
- Ratrou, N. T., Gazder, U., & Al-Madani, H. M. (2014). A review of mode choice modeling techniques for intra-city and border transport. *World Review of Intermodal Transportation Research*, 5(1), 39-58.
- Roy, S., Cooper, D., Mucci, A., Sana, B., Chen, M., Castiglione, J., & Erhardt, G. D. (2020). Why is traffic congestion getting worse? Decomposition of Contributors to Growing congestion in San Francisco: Determining the Role of TNCs. *Case Studies on Transport Policy*, 8(4), 1371-1382.
- Sekhar, C. (2014). Mode choice analysis: the data, the models and future ahead. *International Journal for Traffic & Transport Engineering*, 4(3).
- Sowjanya, D., Tahlyan, D., & Sekhar, C. R. (2014). Travel demand modeling for a metropolitan city. In *International Conference on Recent Trends and Challenges in Civil Engineering* (pp. 19-40).
- Waghmare, A., Yadav, G., & Tiwari, K. (2022). Four step travel demand modeling for urban transportation planning. *Sci. Eng. Technol.*, 5, 1254.
- Waleed, S. (2019). The Ministry of Works is implementing a package of projects to reduce traffic congestion. *Al Bilad News*, Kingdom of Bahrain. Retrieved from <http://www.albiladpress.com/news/2019/4016/bahrain/602413.html>. Accessed on October 19, 2019.

