CIRRELT

CIRRELT-2024-18

Long-distance Travel Demand Modeling through Rare Event Modeling Approach

Hamed Ali Zadeh
Catherine Morency
Martin Trépanier

June 2024

Bureau de Montréal
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél : 1-418-656-2073
Télécopie : 1-418-656-2624

Long-distance Travel Demand Modeling through Rare Event Modeling Approach

Hamed Ali Zadeh[1,*], Catherine Morency[1,2], Martin Trépanier[2,3]

1. Department of Civil, Geological and Mining Engineering, Polytechnique Montréal
2. Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)
3. Department of Mathematical and Industrial Engineering, Polytechnique Montréal

Abstract. Intercity travel, or long-distance (LD) travel, is often overlooked by researchers compared to daily trips. Despite its high contribution to the overall distance travelled, there is no clear national, provincial, or inter-regional definition for this type of trip. In this paper, we present a model comparison approach for the LD trip generation model for Canadian residents based on the Travel Survey for Residents in Canada (TSRC) survey. We define LD trips as non-frequent overnight and day trips based on the TSRC survey. To address the imbalanced data issue associated with relatively rare LD trips, we employ three rare event modelling techniques - oversampling, under-sampling, and synthetic oversampling - as part of the data preparation stage. We utilize TSRC data from 2012 to 2017 to estimate the model. We compare the performance of several machine learning models, including random forest, CART, CTree, and logit, and find that random forest has the best prediction performance, while decision tree models have the best overall accuracy. We also identify that income level and educational level play a significant role in intercity trip occurrence. Our study emphasizes the importance of improving intercity travel survey methods and other data collection techniques. By improving the accuracy of intercity travel data collection, we can develop more reliable and robust models to predict and plan for intercity travel demand.

Keywords: long-distance travel, decision tree, travel demand modelling, rare event modelling, machine learning.

Acknowledgements. The authors wish to acknowledge the support and funding provided by the Ministère des Transports du Québec (MTQ).

---

* Corresponding author: alizadeh.hamed@gmail.com

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2024

1 Introduction

Long-distance (LD) or intercity trips are getting more attention in recent years since their contribution is very high in milage, even if their frequency is lower than daily urban trips. A study in Great Britain shows that less than 2% of travel performed by British residents are considered long-distance trips within a country; however, this 2% is responsible for about 30% of distance travelled by British residents (Joyce M. Dargay 2012). Canada covers a broad territory which can probably lead to having a high proportion of kilometers travelled by residents related to long-distance trips. Also, with a growing population, the travel demand for LD trips will probably rise as well. With a potential increase in LD trips, it is crucial to improve knowledge of LD trips and travel demand modelling to correctly assess their environmental impacts and identify opportunities for service improvement. To our knowledge, researchers have paid much less attention to LD trips during the past decades than to within metropolitan areas trips since the latter are much more numerous. Additionally, data collection methods and surveys have continued to improve during recent years, namely for regional trips, while LD still suffers from a lack of datasets to support relevant research and models.

The methodological approach for long-distance travel demand modelling usually follows those of urban trips. In contrast, LD travel differs from daily trips both in terms of frequency and regularity. It is more typical for people to make daily urban trips than to do an LD trip, making it harder to capture over a short observation period. In other countries, survey methods for LD trips are less developed (and lack interest from decision-makers) in Canada than regional surveys covering trips during a typical weekday. Currently, few sources of data are available to analyses LD trips. In Canada, some periodic surveys have been conducted at the federal and provincial levels. Still, they only cover a particular period or a particular mode of transportation (Guillemette 2015). The Travel Survey of Residents of Canada (TSRC) is the only systematic survey conducted in Canada every year with the purpose of tourism study. It is not explicitly designed to model LD trips and needs more in-depth analysis but still provides some relevant insights into travel intensity.

The objective of this study is to identify the most appropriate trip generation model for long-distance (LD) trips, considering the rare event nature of such trips, and to identify the important explanatory variables. The research commences with a comprehensive review of the existing literature on long-distance modelling in the introductory section, which is followed by a detailed presentation of the data and descriptive statistics, as well as the study area in the subsequent section.

The study employs rare event modelling techniques, which are elaborated on in the methodology section, to address the issue of imbalanced data associated with the infrequent occurrence of LD trips. Model selection is then performed, and the results are extensively discussed in the subsequent section.

Finally, the study concludes with a summary of the key findings and contributions to the field of LD trip generation modeling. This paper contributes to the existing literature on rare event modelling and provides valuable insights into the identification of important explanatory variables for predicting the occurrence of LD trips.

2 Literature Review

In long-distance (LD) travel, there are no significant features to distinguish an LD trip from other trips. Previous studies are considering distance as a variable to identify LD and non-LD trips. Some studies are considering overnight trips as LD travel, such as (Aguiléra 2015, J.J. LaMondia 2015) . One study found that various definitions for LD trip can lead to different results (L. e. Aultman-Hall 2018).

Miller's study (Miller 2004) presents several challenges regarding LD travel demand model estimation. It states that higher resolution is needed in both spatial and temporal levels and that data on accessibility should also be collected. The author also mentions that access to private transit company data should be facilitated (Miller 2004). Data collection is one of the critical points of each travel demand model estimation. Several studies mention the need to improve data collection regarding the features of attractions (destinations) or the smartphone-based data collection methodologies such as (Van Nostrand 2013, Outwater 2015).

Many studies stated that variables having a significant effect on LD travel patterns include age, gender, having kids, income, population density, proximity to train station and airport (Rickard 1988). Also, they say that income has a significant positive relation with performing LD trips (Berliner 2018, Czepkiewicz 2020, L. H. Aultman-Hall 2018). The findings of Yao`s analysis reveal that the working population in service industries exhibits a statistically significant and positive impact on trip rates which emphasize employment rate have positive impact on making a trip (Yao 2005).

Two studies mentioned that accessibility to airports plays a crucial role in LD travel behaviours (Enzler 2017, L. H. Aultman-Hall 2018), but fewer studies have dealt with this issue. It can be due to limited data availability and not

having access to individuals' distance and travel time to airports, bus, and train stations. LIorca (Llorca 2018) has conducted a study on LD travel demand modelling with the same data source that we use in this study, and to deal with the challenge of data collection, despite TSRC, they used Foursquare and Rome2rio data. He found that Foursquare check-in data can improve the goodness of fit of models, particularly for leisure trips. Also, Rome2rio data can improve the mode choice model. However, results might be biased since data from both Rome2rio and Foursquare are searches performed by people (not confirmed travels), especially in mode selection related to Rome2rio data since each search results contain all potential options from origin to destination.

Since LD trips getting less attention by researchers with any reason like lack of good dataset for these kinds of the trip, the modelling in the LD trip modelling is treated the same as daily usual trip, (Yao, A study of on integrated intercity travel demand model 2005) developed an integrated intercity travel demand model which considers all component in LD travel demand are interrelated together, they state that the LD travel choice is related to the destination, trip frequency, mode, and route choice etc. Among the four-step of trip modelling for LD trip the mode choice is getting the highest attention by researchers while the trip generation is so important as well, (Hess 2018) developed a hybrid choice model to understand the mode choice of drivers for the intercity trip, the interesting result of their research shows how the travellers attitude like privacy and anti-car attitude can change the mode choice which can be a result of being for a longer time in the travel.

Trip generation is the first step of the four-step travel demand model. Several studies considered different approaches to the trip generation models to find which variables impact LD travel and how these variables affect the LD trips. (LaMondia 2014) used a non-distance-based LD trip threshold to define LD trips by purpose, duration, mode, and destination and used an ordered probit methodology to model trip generation. Some studies used negative binomial regression models to a model annual trip generation, number of trips by purpose, number of domestic and international ground trips (Berliner 2018, Czepkiewicz 2020, L. H. Aultman-Hall 2018).

2.1 Rare Event Model

A classification data set where one class has significantly more observations than the others is defined as an imbalance or rare event data set (Cieslak 2008). To our knowledge, few studies are considering LD trips as a rare event in their models. Accident occurrence is widely modelled in the transportation field using a rare event approach (Theofilatos 2016). Theofilatos study (Theofilatos 2016) uses a rare event approach to predict accidents on the road. It mentions that the condition of non-event data collection might vary by event condition; hence, they used a rare event logit model package in R software to estimate the model. (Vilaça 2019) used the rare event method for modelling injury severity risk of vulnerable road users; he used three methods on data preparation under, over, and synthetically oversampling method to deal with imbalanced data and decision trees and logistic regression model were used for model estimation.

A rare event dataset can be considered a dataset when one class's occurrences are significantly lower than those of another class. In this case, data can be defined as a rare event or an imbalanced dataset (Chawla 2008).

The problem with rare event datasets arises when we are interested in the rare class. The majority class biases the decision tree. It leads to a reasonable model accuracy for the majority class. In contrast, it results in poor performance for the minority class (Chang 2005, Zheng 2016). To overrepresent the rare event class, it is possible to set up the prior probability for both categories (Enterprise 2018). "Increasing the prior probability of the rare event class increases the posterior probability of the class, which moves the classification boundary for that class so that more observations are classified into the class." (Zheng 2016).

Given the rare occurrence of long-distance (LD) trips and the need to improve data collection and survey methods, traditional trip generation modeling approaches may be subject to bias and may not provide an accurate representation of the population of interest. To address this issue, the present study adopts a rare event modeling approach using machine learning methods for model estimation. Specifically, the study explores the use of various machine learning algorithms to determine the most appropriate method for handling data-related issues in LD trip generation modeling.

The results of this study suggest that the use of machine learning methods with a rare event approach provides a more accurate representation of the population of interest compared to traditional trip generation modeling approaches. This approach offers valuable insights into the identification of important explanatory variables for predicting the occurrence of LD trips, which could inform the development of more effective data collection and survey methods in the future.

3 Study area and descriptive analysis

This study covers intercity or long-distance (LD) trips all over Canada. This country is composed of ten provinces and three territories, covering over 9 million square kilometers and a population of over 35 million people, according to the 2016 census. In this study, the intercity trip generation model considers all non-frequent day trips and overnight trips as reported in the Travel Survey for Residents in Canada (TSRC). The TSRC survey is designed with the purpose of supporting domestic and international tourism studies. The TSRC data collection is performed by phone for domestic trips and in-person for international travel; daily commute LD trips are excluded from the survey.

This survey evaluates the volume of domestic and international travel in Canada made by Canadian residents; it includes data on the trip's origin and destination, trip characteristics, duration in case of overnight trips, activities conducted during the trip, expenditures, and socio-demographic characteristics of the respondent. The highest spatial resolution of the origin and destination points is the Census Division (CD), which corresponds to small regions and metropolitan areas. The TSRC survey is the main data source used for the LD trip generation model in this study. The main variables used for estimation of the model are shown in **Error! Reference source not found.**. The survey of 2012 to 2017 was employed for the model. The samples of the TSRC surveys are 88813, 75753, 74391, 67138, 65225, 58361 people, respectively, for the 2012 to 2017 surveys.

TABLE 1 Demographic variable in the TSRC survey

Variable	Description	Coding of Input Value	Abbreviation (in the model)	Variable
Respondent gender	Male	1	SEX	Categorical
	Female	2		
Respondent educational level	Less than high school	1	EDLEVGR	Categorical
	High school certificate	2		
	Some post-secondary diploma	3		
	University degree	4		
Respondent age	18-24	1	AGE_GR2	Categorical
	24-34	2		
	35-44	3		
	45-54	4		
	55-64	5		
	65+	6		
Household size	Number of children	N/A	G_KIDS	Categorical
Income (CAD)	Less than 50k	1	INCOMGR2	Categorical
	50k to 70k	2		
	70k to 100k	3		
	100k and over	4		
	N/A	5		
Respondent employment	Employed	1	LFSSTATG	Categorical
	Unemployed	2		
CMA Level		1-34	RESCMA2	Ordinal

In the TSRC survey, all the respondent characteristics are categorical variables. In this study, for model estimation, all the variables are transformed into ordinal variables. To include the home location of the respondent in the model, the population of the census metropolitan area (CMA) of residence is included. The geography includes 33 census areas (CA) and CMA and one level representing the rest of Canada (all that is not included in any CA or CMA).

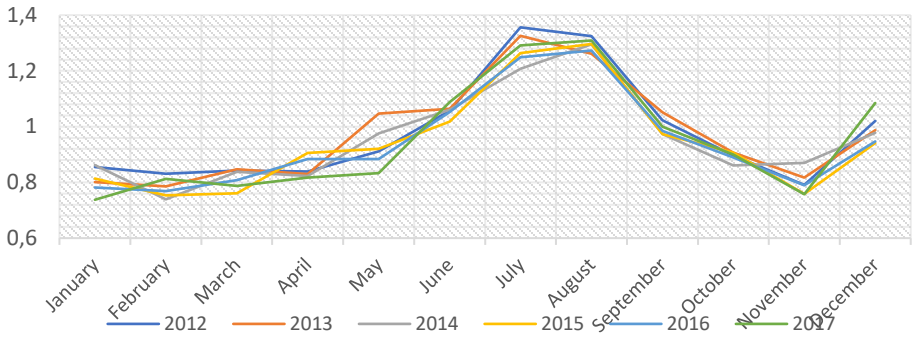


Figure 1 Number of trips per person per month for all trips performed in Canada, 2012-2017

Figure 1 illustrates the LD trip rate per month spanning the duration of 2012 to 2017. It can be inferred from the figure that the LD trip rate remains consistent throughout the aforementioned period. However, it is worth noting that the trip rate experiences a discernible peak during the summer months and the month of December, as anticipated.

In Figure 2, we see the percentage of people who went on long-distance (LD) trips and those who did not, during each month and year, based on a survey. During the summer, when people tend to travel more, only around 35% of the survey respondents went on at least one LD trip, while for other months, it was less than 20%. This indicates that most people did not take LD trips.

The dataset is considered imbalance because there were very few LD trips compared to non-LD trip events. This means that the data requires special statistical methods to analyze and draw meaningful conclusions. A classification data set where one class has a significantly higher number of observations than the others is defined as an imbalance or rare event data set (Cieslak 2008).

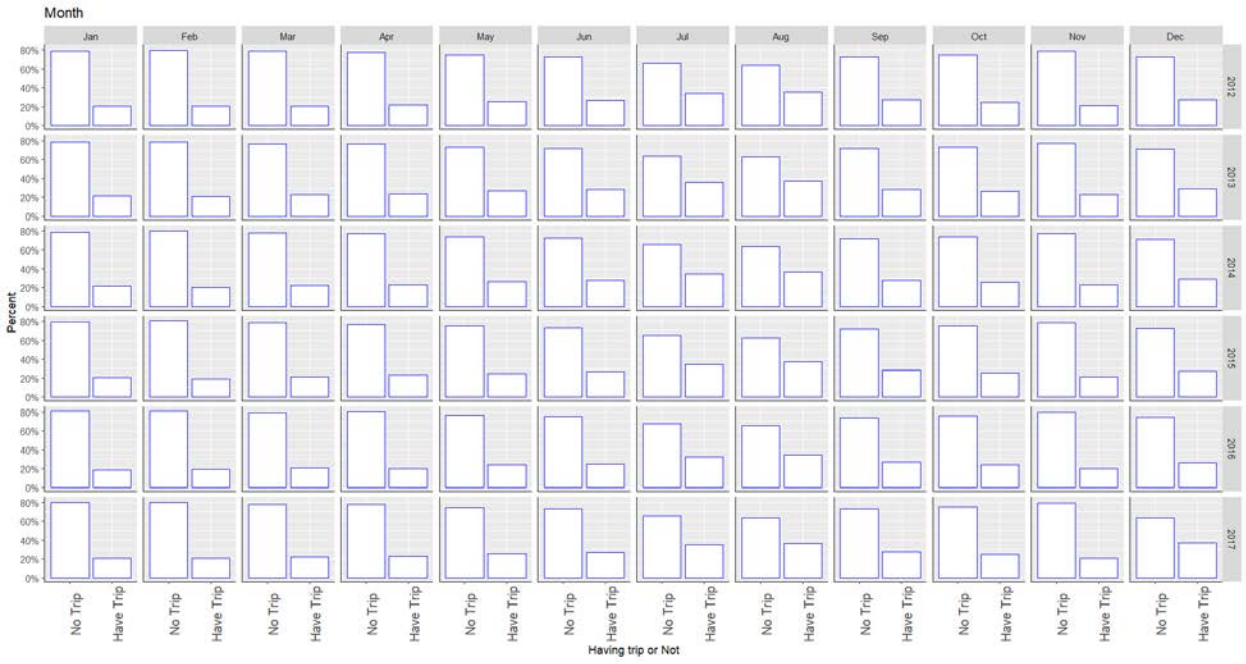


Figure 2 Proportion of people who did at least one LD trip during the month (event) and of those who did not (non-event) during the study period

4 Methodology

4.1 Data Preparation

As previously mentioned, the occurrence of LD trips is considered a rare event, and consequently, the dataset contains imbalanced data with the minority (or positive group) of individuals who make at least one LD trip and the majority (or negative group) of individuals who do not make any LD trip during the month of observation. Resampling of training data set is a frequently used method to tackle the issue of imbalanced data set (He 2009). Three resampling methods are commonly used: under-sampling, oversampling, and synthetically oversampling. These methods take into consideration having a more balanced dataset on the training level. The under-sampling technique aims to reconstruct a more balanced dataset by randomly removing cases from the majority class to reach the desired ratio of class distribution (Haixiang 2017).

In contrast, the oversampling technique aims to reconstruct a more balanced dataset by randomly duplicating cases from the minority class to reach the anticipated ratio of class distribution (Haixiang 2017). The oversampling technique's challenge could be an overfitting model by improving recognition of the minority class (He 2009). The synthetically oversampling method is another technique to tackle the issue of overfitting caused by oversampling (Menardi 2014). In this technique, synthetic cases with a feature of minority class according to the smoothed-bootstrapping method are generated.

In this paper, the three techniques mentioned earlier are employed. We found that the under-sampling technique led to better model performance. So, the cases were randomly removed from the majority class, individuals with no trip cases, to have a more normally distributed classification dataset.

4.2 A Rare Event Approach

The first step in the trip generation modelling is to evaluate the proportion of people who made at least one LD trip against those who did not during the study period. It was found that each month, around 25% of people are making at least one LD trip. Since those who do LD trips are the minority group in the dataset, traditional statistical models underestimate the probability of this minor class. A rare event dataset can be considered a dataset when one class's occurrences are significantly lower than those of another class. in this case, data can be defined as a rare event or imbalanced dataset (Chawla 2008).

The problem with rare event datasets arises when we are interested in the rare class. The majority class biases the decision tree. It leads to a reasonable model accuracy for the majority class. In contrast, it results in poor performance for the minority class (Chang 2005, Zheng 2016). To overrepresent the rare event class, it is possible to set up the prior probability for both categories (Enterprise 2018). "Increasing the prior probability of the rare event class increases the posterior probability of the class, which moves the classification boundary for that class so that more observations are classified into the class." (Zheng 2016).

In general, a dataset with binary classification can be considered imbalanced if the distribution of the classes is not equal. However, determining whether a dataset is truly imbalanced can be a relative issue. If the minority class represents 40% to 60% of the dataset, balancing techniques may not be necessary. However, if the minority class represents less than 30% of the data and the characteristics of the minority class are significantly different from the majority class, balancing the dataset may be necessary. On the other hand, if there is significant difference between the features of the two classes, a minority class of 30% may not be considered imbalanced. Ultimately, it is up to the researchers to determine whether a dataset is imbalanced and whether to use balancing techniques to address any potential issues.

Rebalancing a dataset can be a useful technique to improve the performance of a binary classification model when the original dataset is imbalanced. However, rebalancing may lead to biased results towards the minority class, which can overestimate the model's predictive power for that class. Additionally, using a rebalanced dataset to predict future demand may result in a different distribution of outcomes, which may not reflect the true distribution of the binary variable in the population.

To address this problem, alternative evaluation metrics such as precision, sensitivity, and F1 score have been used to evaluate the performance of the model on imbalanced datasets. These metrics consider both true positive and false positive rates, which are important for evaluating the model's performance on imbalanced data. By using these metrics, one can better understand the model's performance and make more accurate predictions, without the need for rebalancing the dataset.

Rare event modelling has received less attention in long-distance trip generation modelling. In this study, a priori mentioned methods that are common in rare event data preparation were assessed. It was found that under-sampling the majority class leads to better model results. Different machine learning metrics are used to find the best-fitted model. They are described hereafter.

4.3 Logistic Regression Model

Binary models are particularly useful for analyzing data where the dependent variable is binary or dichotomous, which means that it can only take on one of two possible values, such as "yes" or "no." In the context of trip generation modeling, the dependent variable could represent whether or not an individual made a trip to a certain location or for a specific purpose. Therefore, in this study several binary model were employed as follow.

Generalized linear models (GLMs) are an extension of the traditional linear models. An ordinary linear model requires that the error term be normally distributed, while in GLM, this assumption is relaxed. In the GLM model, the response variable can be binomial, Poisson, and other kinds of distributions from the exponential family. In this study, because we have two classes, the response variable is set up as binomial. In the linear model, the predictions estimate the response variable, while in GLM, it is the function of the response variable prediction.

In general, a generalized linear model (GLM) for binary data is the inverse of the standard logistic function:

$$logit(p) = \log(\frac{p}{1-p}) \quad (1)$$

Where p is the probability, and logit(p) is the logarithm of the corresponding odds. The logistic models can lead to biased results on classification problems with imbalanced datasets. Ma and Lukas compared the performance of several machine learning methods and GLM on rare event datasets (Ma, Lukas 2021). They found almost the same performance for classification problems in both ways. The GLM method has better sensitivity for the classification method on rare events. However, some studies are stating that machine learning method have better result in general, (Lu 2021) compared different approach to analyze the hesitancy in choice of transfer airport and results shown that the random forest and deep reinforcement learning models have more accurate and structured result.

In this study, generating an LD trip is assumed as a rare event phenomenon, and "Have Trip" corresponds to a trip made by a respondent and is assumed to be a positive class in the model.

4.4 Decision Tree

Decision trees (DT) are a valuable method to identify homogeneous subgroups distinct by individual characteristics. This study utilized the Classification and Regression tree (CART) technique and the Conditional Inference tree (CTree) technique. These DT approaches have the advantage of being easy to explain and interpret. Results can be represented graphically, and, for qualitative variables, there is no need to create dummy variables. The performance of these techniques is compared with other models in the process of model selection.

One of the most used methods for building a decision tree, "CART", was developed by Breiman (Breiman 2017). A split in CART aims to minimize the relative sum of squared errors in the two partitions of a split. The splitting process consists of two steps: 1) the best split will be found across all covariates; then, 2) the point will be split up for those covariates.

CART searches across all splits generated by predictor variables for split selection. Then the split with the most significant criterion is selected to transfer samples into corresponding sub-nodes. It has been mentioned by studies that when there is a numerous split point for each variable, this method might be biased with variable selection. This issue of variable selection with many possible splits is widely discussed in previous studies (Breiman 2017, Shih 2004, Loh 1997).

The CART approach basically consists of three main steps: 1) growing the three, 2) pruning, and 3) selecting an optimal tree. In the first step, based on the values of a set of covariates, it recursively executes univariate splits of the dependent variable. This splitting of a feature results in two splits by choosing one variable and its split value. Child nodes are then treated like parent nodes, and this process continues until some criterion is met (Mishra 2003). In the pruning process, CART aims to reduce the tree's complexity by replacing nodes and subtrees with leaves. This process can reduce the size of the tree and, in some cases, improve the classification accuracy (Patil 2010).

A conditional inference tree (CTree) algorithm was proposed by (Hothorn 2006). CTree avoids variable selection bias in the CART algorithm; instead of selecting a variable that maximizes the Gini index, it uses a significance test method to choose that variable. CTree selects the predictor variable for split by statistical testing between response and covariate. CTree, in each step, uses traditional statistical procedures; in the case where both response variables and possible split variables are categorical, it uses the Chi-squared test (χ^2), in case where one variable is categorical and one is numeric, one-way ANOVA (analysis of variance) is performed, and for both numerical variables, Pearson correlation test is employed (Schlosser 2019).

To assess the association of each covariate and outcome, the mentioned test is employed in CART. If sufficient evidence is found to reject the global null hypothesis, the node will be selected to be split. The covariate which has the strongest association with the outcome of interest is chosen as a candidate for splitting.

In this study, the CART algorithm was performed from the Rpart R package, the party R package was employed for the CTree algorithm, and the caret R package was used to achieve the predictive performance of both algorithms.

4.5 Random Forest

Random forest (RF) is one of the classifications and regression models widely used for binary class datasets. The random forest model uses many decisions, tree-like models, bootstrapped data, and the model decision based on the average prediction of all decision trees. The random forest model, by reducing the correlation between decision trees, aims to improve variance reduction. The study by (Kabir 2018) expresses how the random forest method improves the predictive performance by deciding based on the growth of numerous trees. Random forest allows each tree to individually sample from the data set randomly with replacement with same sample size and different variables, which results on different trees. Hence, these decision trees are susceptible to "train" data sample that changes in the training set, resulting in various structures in decision trees. This process is called Bootstrap Aggregation (Bagging).

Generally, the RF development process consists of four steps. Firstly, it uses bootstrapping to choose a sample from a training set with the same sample size. It then uses a subspace method to select different dependent variables from the total set of variables. Having a new sample will build a decision tree, and finally, the RF model performs the tree steps repeatedly to make many trees. The number of trees is determined by an error called OOB (Out-of-Bag) error.

Having more trees in the model leads to a lower OOB error rate in an RF tree, which is desired because a lower error rate results in better accuracy in the RF model. On the other hand, having more trees will increase the possibility of similar trees. This issue is tackled in the random forest by restricting the number of variable selections with the subspace method. In this process, adding more trees does not result in overfitting, but there is not much benefit to growing more trees (Friedman 2009).

To measure the importance of variables, random forest utilizes the MeanDecreaseGini index. This index is calculated based on the Gini impurity index used to calculate splits (Atkinson 1970). This study uses the MeanDecreaseGini to identify the importance of explanatory variables that contribute to the model. MeanDecreaseGini is considered as the average decrease of the Gini impurity index over all trees in the model. In this study, the random forest algorithm was used from the Random Forest R package, and the caret R package was used to achieve the predictive performance of the random forest.

4.6 Comparison Indicators

To compare the results of models, classification accuracy is employed. Table 2 presents the agreement of observed and predicted conditions of the test dataset for "having trip" and "no trip." Having an LD trip refers to a positive or event class, and not having an LD trip refers to a negative or non-event class.

Table 2 Confusion matrix components

	Predicted Condition	
Actual Condition	True positive (TP), Hit	False-negative (FN), Miss, underestimation
	False-positive (FP) Overestimation	True negative (TN) Correct rejection

As it can be deduced from Table 2, the components of the confusion matrix are represented as below:
True positive (TP) = the number of instances correctly identified as "have trip."
False-positive (FP) = the number of instances incorrectly identified as "have trip."
True negative (TN) = the number of instances correctly identified as "no trip."
False-negative (FN) = the number of instances incorrectly identified as "no trip."

To compare the models, result accuracy is the most common value found in the literature (Cieslak 2008, Rachman 2019). It is defined as the ability to differentiate the event and non-event correctly; the mathematical calculation of the accuracy is presented in equation 2.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2)$$

The accuracy represents an overall performance of the model, but it does not measure the accuracy of prediction on each class of events and non-events. The model's overall accuracy can be biased by the majority class because most of the observations are in the "no trip" class. To tackle this issue and avoid misinterpretation of prediction accuracy with overall accuracy, sensitivity, specificity, F1 score, and precision are considered prediction performances of the negative and positive classes of the model.

The sensitivity of a test states its ability to determine the "have trip" or positive class cases correctly, and the specificity of a test is its ability to determine the "no trip" or negative cases correctly. Mathematically these two parameters can be stated as equations 3 and 4 consecutively.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

Precision is another factor that describes how well a model predicts the "have trip" or positive class. It evaluates the proportion of correct positive predictions to the overall wrong and correct positive class. This score is mathematically stated as equation 5.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

A harmonic mean of precision and sensitivity called the F1-score is used. It allows having a better measure of accuracy on negative classified cases. There are several advantages of using an F1-score instead of accuracy. In the F1-score, the importance of false positive and false negative is also taken into consideration, while in the accuracy indicator, the focus is just on true positive and true negative classes. Also, it is more relevant for an imbalanced dataset to use F1-score, while for a normally distributed dataset, accuracy can be employed. In this study, the "have trip" class is considered rare, so the F1-score is a better metric for evaluating the model performance. The F1-score is estimated using equation 6.

$$\text{F1} - \text{score} = 2 * \frac{sensitivity * precision}{sensitivity + precision} \quad (6)$$

5 Results

In this study, four machine learning techniques are implemented for trip generation model estimation, and the performance of these models is compared. These models are widely used for rare events or imbalanced data modelling in the literature (Ma, Lukas 2021, Zhou 2020). CART decision tree algorithm, CTree decision tree algorithm, random forest, and generalized linear model are used.

For all techniques, 75% of the dataset is considered as training, and the rest is used for validation. The dependent variable in the model is set as a binominal choice whether an individual performs any long-distance trip during the study period or not, and socio-demographic features are used as independent variables. The "party" and "Rpart" package in R software is used for the estimation of DTs.

Several mentioned factors used to evaluate the performance of the model's estimation and the result of analysis are presented in the models prediction ability.

To tackle the issue of imbalanced data, three different data balancing approaches are implemented before model estimation; since the data set is imbalanced, the F-1 score is used to evaluate the performance of data preparation methods, as is shown in (Table 3), the under-sampling method has the better result. So, this method was employed during the study for model performance evaluation.

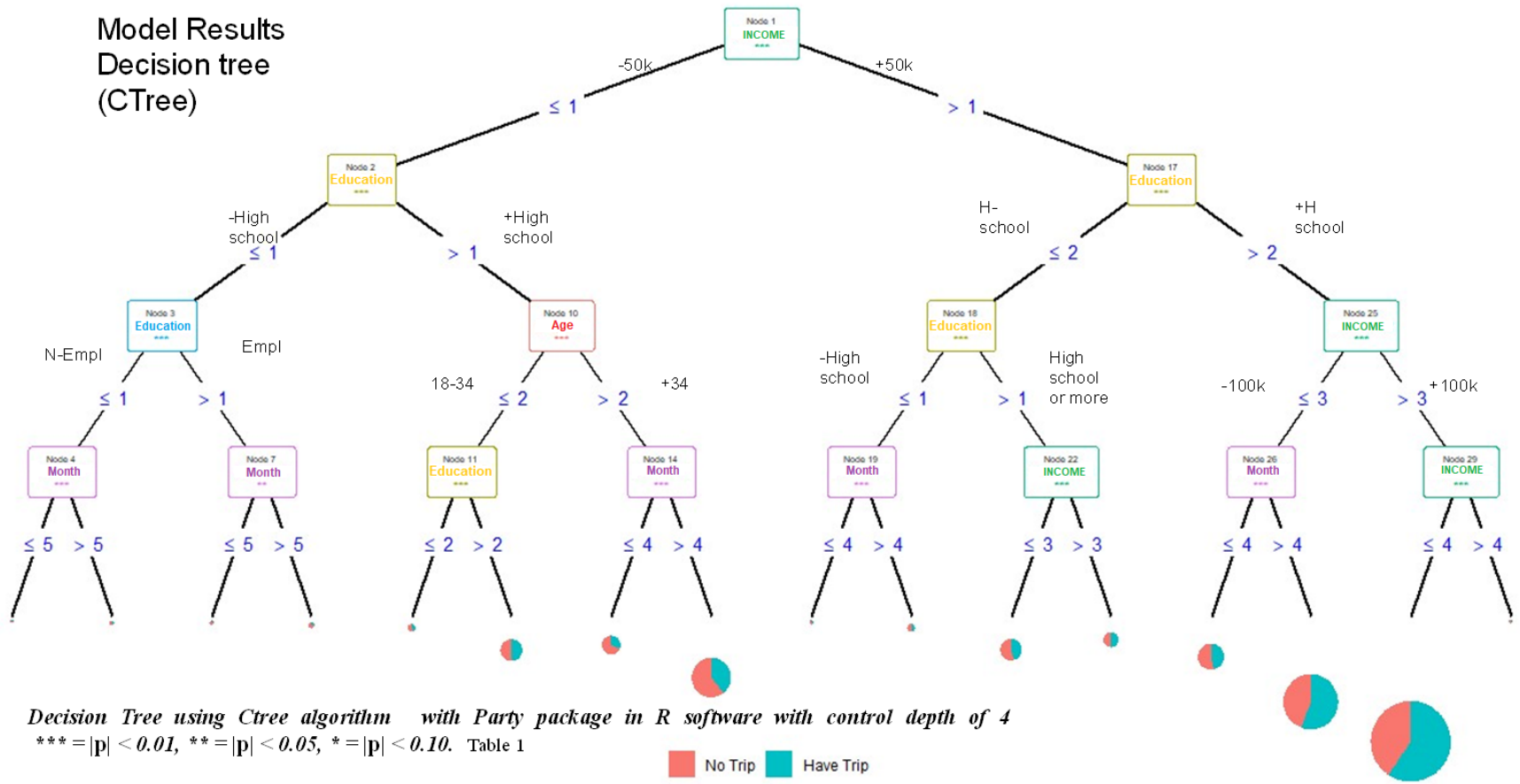
Table 3 F-1 score for different methods of data preparation

F1-score	Under-sampling	Oversampling	Synthetically oversampling
Random Forest	0.457	0.413	0.426

Table 4 Results demonstrate that random forest gives the highest performance in terms of research focus on positive class and CTree in terms of research interest on negative class. The overall accuracy is a factor stating the model's overall performance, which says that the CART model has the best performance overall but the least accuracy on prediction of the positive class. However, in the case of a rare event data set, other scores play an essential role in the model's prediction ability.

Table 4 Model performance score

	CART	CTree	Random Forest	Logistic
Accuracy	0.642	0.612	0.588	0.597
Sensitivity	0.485	0.565	0.656	0.618
Specificity	0.704	0.630	0.563	0.542
Precision	0.392	0.375	0.351	0.358
F1-score	0.433	0.450	0.457	0.453



Decision Tree using Ctree algorithm with Party package in R software with control depth of 4
*** = |p| < 0.01, ** = |p| < 0.05, * = |p| < 0.10. Table 1

Figure 3 Decision Tree using CTree algorithm with Party package in R software with control depth of 4
*** = |p| < 0.01, ** = |p| < 0.05, * = |p| < 0.10.

In the case of this study, since the positive class is a crucial event to predict the factors that affect people's LD trip, sensitivity plays a vital role in the determination of model performance. The random forest has the highest score through the models. Also, to avoid overestimating the positive class, F1-score is defined to ensure that the random forest model performs better than other models.

Figure 3 displays a CTree decision tree with a controlled depth of four levels. The diagram reveals that income significantly impacts the probability of an LD trip occurrence. Additionally, educational level and reference month also play crucial roles in determining whether a person embarks on an LD trip. Those with a higher academic level tend to undertake more LD trips, while people frequently plan such trips in the summer due to favorable weather conditions and vacation time. Furthermore, the data suggests that individuals with lower income and education levels are less likely to take LD trips. Notably, the results also indicate that young people belonging to the same income group tend to undertake more LD trips if they possess a higher educational qualification.

Table 5 Variable coefficient for DT model

Variable	CART	CTREE
Income	0.314	0.236
Educational level	0.308	0.188
Month	0.189	0.097
Age group	0.298	0.068
Population	0.107	0.041
Employment	0.028	0.011
SEX	0.031	0.007

Table 5 shows the variable coefficients for both CART and CTREE models. The variables include Income, Month, Educational level, Population, Age group, Employment, and SEX. The CART model shows that Income has the highest coefficient of 0.314, followed by Month with 0.308.

The CTREE model also shows Income and Month as the two variables with the highest coefficients, but with lower values of 0.236 and 0.188, respectively. Educational level has a coefficient of 0.097, followed by Population with 0.068.

It is worth noting that the coefficients may not have the same interpretation in both models, as CART and CTREE use different algorithms and criteria to build the decision tree. Therefore, the variable importance ranking, and interpretation may differ depending on the model used. In CART, the coefficient represents the change in the predicted response variable, while in CTREE, it represents the difference in the predicted response variable between the two groups formed by splitting on that variable.

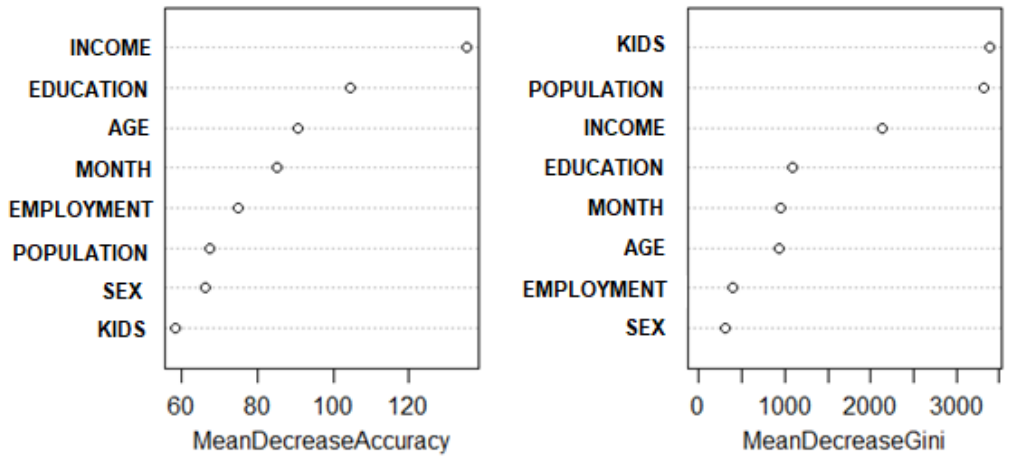


Figure 4 Variable importance using Random Forest model

The interpretation of the variable importance in the random forest model can be made with two crucial graphs represented in Figure 4. This outcome of the random forest model states how vital the variables are in classifying the data. In every tree of the random forest model, the prediction error of OOB data is recorded, and the prediction error of permuting each predictor variable is recorded as well. The average difference of these two errors is normalized by the standard deviation of the differences. This factor is called "*MeanDecreaseAccuracy*" (RColorBrewer 2018).

In summary, this factor is a unitless factor that states how much the accuracy of the model depends on each variable. The variables are presented in descending order. The highest variable plays the essential role variable in the model. "MeanDecreaseGini" is the average of node impurities from splitting on a variable over all trees. In this case, a higher value results in a split with a purer node (RColorBrewer 2018). In other words, a higher value for each variable states how much it contributes to the consistency of the nodes.

An interesting finding from this study is that income and educational levels are the most critical variables in the model. This result aligns with the CTree model, as well as the logit and CART models, indicating that despite the different methods employed, all models demonstrate the importance of these variables in LD trip generation accurately. Although having kids and the Census Metropolitan Area (CMA) of residence may not have a significant impact on the model's accuracy, they do contribute to the purity of the node splits, thereby playing an important role in the analysis.

The performance of the model also confirmed how necessary is the data collection procedure. Also, it confirmed the importance of improving survey methods with a focus on LD trip analysis. Better data collection helps to have more in-depth studies for LD travel and to find how individual characteristics affect their intention to make intercity trips.

6 Conclusion

This paper used data from the Travel Survey for Residents in Canada (TSRC), covering the 2012 to 2017 period, to develop a long-distance (LD) trip generation model. The primary objective of this study is to identify the predictor variables that influence individuals' decision to undertake a long-distance (LD) trip during the study period. During the modeling process, it was discovered that the study suffered from an imbalanced dataset, which required the researchers to employ various balancing techniques and machine learning methods to address this issue. This was necessary to obtain the most accurate model to answer the main research question.

The paper employed several techniques at the data preparation level to address the imbalance distribution issue commonly found in rare event modelling, which is particularly relevant for LD trip generation modelling due to the scarcity of data. Under-sampling was found to yield better performance in the trip generation model.

To identify the most suitable model, CART, CTree, random forest, and generalized linear model were compared. The evaluation metrics used to assess the model performance included overall accuracy, sensitivity, specificity, precision, and F1-score. The study found that decision tree models, random forest, and logit models had the best overall accuracy in that order. However, the imbalanced nature of the data means that overall accuracy might be biased by the majority class, and hence, sensitivity and F1-score are more reliable metrics for evaluating prediction performance.

The results revealed that the random forest model had the best performance in predicting the "having trip" class by individuals, while other models may perform better in predicting the "no trip" class. The choice of the most appropriate model depends on the study's goal and the importance of both positive and negative classes in the expected use of study results.

The study findings highlight the significance of income and educational level in LD trip occurrence. Higher income and academic levels lead to more LD trips, and younger people with a higher level of education make more LD trips even with the same income level. The results suggest the need for improving intercity travel survey methods and other data collection methods to enhance LD trip modelling.

During model estimation, as other studies have stated (Van Nostrand 2013, Llorca 2018), data-related issues are a fundamental challenge in LD travel demand modelling. The TSRC survey dataset comprises mainly categorical variables, except for the "population" variable, which has been ranked based on the population of CMAs in the study and fed them to the model as an ordinal variable. Hence, factors like accessibility to the airport, bus, and train station in terms of travel distance and cost and land-use data might result in better model performance. The TSRC survey was designed for tourism studies. This study proposed a relevant use of these data for travel demand forecasting and insists on improving data collection and survey methods to have opportunities to enhance the mode complexity and provide more insights into the factors having an incidence on LD travel behaviors.

7 Financial Disclosure

The analysis presented in this paper was financed by the Ministère des Transports du Québec.

8 Acknowledgements

The authors wish to acknowledge the support and funding provided by the Ministère des Transports du Québec (MTQ).

9 Author Contribution Statement

The authors confirm contribution to the paper as follows: study conception and design: H. Ali Zadeh, C. Morency, M. Trépanier; data preparation: Hamed Ali Zadeh; analysis and interpretation of results: H. Ali Zadeh. C. Morency, M. Trépanier; draft manuscript preparation: H. Ali Zadeh, C. Morency, M. Trépanier. All authors reviewed the results and approved the final version of the manuscript.

References

Aguiléra, A., & Proulhac, L. 2015. "Socio-occupational and geographical determinants of the frequency of long-distance business travel in France." *Journal of Transport Geography* 28-35.

Atkinson, Anthony B. 1970. "On the measurement of inequality." *Journal of economic theory* 244-263.

Aultman-Hall, L., Harvey, C., Sullivan, J., & LaMondia, J. J. 2018. "The implications of long-distance tour attributes for national travel data collection in the United States." *Transportation, 45(3)*, 875-903.

Aultman-Hall, Lisa, et al. 2018. "The implications of long-distance tour attributes for national travel data collection in the United States." *Transportation* 875-903.

Berliner, R. M., Aultman-Hall, L., & Circella, G. 2018. "Exploring the Self-Reported Long-Distance Travel Frequency of Millennials and Generation X in California." *ransportation Research Record* 208-218.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. 2017. *Classification and regression trees. Routledge.* Routledge.

Chang, Li-Yen, and Wen-Chieh Chen. 2005. "Data mining of tree-based models to analyze freeway accident frequency." *Journal of safety research* 365-375.

Chawla, David A. Cieslak and Nitesh V. 2008. *Learning Decision Tree for Unbalanced Data.* Accessed July 2021. https://www3.nd.edu/~nchawla/papers/ECML08.pdf.

Cieslak, D. A. and N. V. Chawla. 2008. "Learning Decision Trees for Unbalanced Data." *Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg, Springer Berlin Heidelberg.*

Czepkiewicz, M., Heinonen, J., Næss, P., & Stefansdóttir, H. 2020. "Who travels more, and why? A mixed-method study of urban dwellers' leisure travel." *Travel behaviour and society* 67-81.

Enterprise, SAS. 2018. *SAS Institute Inc. Prior Probabilities.* Nov. 29. http://documentation.sas.com/doc/en/emxndg/15.1/p1vqpbjwoo4bv7n1sw77e0z64xxs.htm.

Enzler, H. B. 2017. "Air travel for private purposes. An analysis of airport access, income and environmental concern in Switzerland." *Journal of Transport Geography* 1-8.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2009. *The elements of statistical learning.* New York: Springer.

Guillemette, Y. 2015. *MIEUX COMPRENDRE L'OFFRE ET LA DEMANDE DE DÉPLACEMENTS INTERURBAINS AU QUÉBEC.* Montreal: (Master thesis) Polytechnique de Montreal. https://publications.polymtl.ca/1829/.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. 2017. "Learning from class-imbalanced data: Review of methods and applications." *Expert Systems with Applications* 220-239.

He, H., & Garcia, E. A. 2009. "Learning from imbalanced data." *IEEE* 1263-1284.

Hess, Stephane, et al. 2018. "Analysis of mode choice for intercity travel: Application of a hybrid choice model to two distinct US corridors." *Transportation Research Part A: Policy and Practice* 547-567.

Hothorn, T., Hornik, K., & Zeileis. 2006. "Unbiased recursive partitioning: A conditional inference framework." *Journal of Computational and Graphical statistics* 651-674.

J.J. LaMondia, M. Moore, L. Aultman-Hall. 2015. "Modeling intertrip time intervals between individuals' overnight long-distance trips." *Transportation Research Record* 23-31.

Joyce M. Dargay, Stephen Clark. 2012. "The determinants of long distance travel in Great Britain." *Transportation Research Part A: Policy and Practice* 46 (3): 576-587.

Kabir, Elnaz, Seth Guikema, and Brian Kane. 2018. "Statistical modeling of tree failures during storms." *Reliability Engineering & System Safety* 68-79.

LaMondia, J. J., Aultman-Hall, L., & Greene, E. 2014. "Long-distance work and leisure travel frequencies: Ordered probit analysis across non–distance-based definitions." *Transportation Research Record* 1-12.

Llorca, C., Molloy, J., Ji, J., & Moeckel, R. 2018. "Estimation of a long-distance travel demand model using trip surveys, location-based big data, and trip planning services." *Transportation Research Record* 103-113.

Loh, W. Y., & Shih, Y. S. 1997. "Split selection methods for classification trees." *Statistica sinica* 815-840.

Lu, Jing, et al. 2021. "Modeling hesitancy in airport choice: A comparison of discrete choice and machine learning methods." *Transportation Research Part A: Policy and Practice* 230-250.

Ma, Lukas. 2021. *Modelling rare events using non-parametric machine learning classifiers-Under what circumstances are support vector machines preferable to conventional parametric classifiers?* Göteborgs: Göteborgs universitet , 5-29.

Menardi, G., & Torelli, N. 2014. "Training and assessing classification rules with imbalanced data." *Data mining and knowledge discovery* 92-122.

Miller, E. 2004. "The Trouble with Intercity Travel Demand Models." *Transportation Research Board* 94-101.

Mishra, S., Deeds, N. E., & RamaRao, B. S. 2003. "Application of classification trees in the sensitivity analysis of probabilistic model results." *Reliability Engineering & System Safety* 123-129.

Outwater, M., Bradley, M., Ferdous, N., Trevino, S., & Lin, H. 2015. " Foundational Knowledge to Support a Long-Distance Passenger Travel Demand Modeling Framework: Implementation Report."

Patil, D. D., Wadhai, V. M., & Gokhale, J. A. 2010. "Evaluation of decision tree pruning algorithms for complexity and classification accuracy." *International Journal of Computer Applications* 23-30.

Rachman, A., & Ratnayake, R. C. 2019. "Machine learning approach for risk-based inspection screening assessment." *Reliability Engineering & System Safety* 518-532.

RColorBrewer, S., & Liaw, M. A. 2018. *Package 'randomForest'*. Berkeley: Berkeley, CA, USA: University of California.

Rickard, Julie M. 1988. "Factors influencing long-distance rail passenger trip rates in Great Britain." *Journal of Transport Economics and policy* (1).

Schlosser, L., Hothorn, T., & Zeileis, A. 2019. "The power of unbiased recursive partitioning: a unifying view of CTree, MOB, and GUIDE." *arXiv preprint arXiv*.

Shih, Y. S. 2004. "A note on split selection bias in classification trees." *Computational statistics & data analysis* 457-466.

Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. 2016. "Predicting road accidents: a rare-events modeling approach." *ransportation research procedia* 3399-3405.

Van Nostrand, C., Sivaraman, V., Pinjari, A. R. 2013. "Analysis of Long-Distance Vacation Travel Demand in the United States: A Multiple Discrete–Continuous Choice Framework. Transportation." *Transportation 40*, 151–171.

Vilaça, M., Macedo, E., & Coelho, M. C. 2019. "A rare event modelling approach to assess injury severity risk of vulnerable road users." *Safety* 29.

Yao, Enjian, and Takayuki Morikawa. 2005. "A study of on integrated intercity travel demand model." *Transportation Research Part A: Policy and Practice* 367-381.

Yao, Enjian, and Takayuki Morikawa. 2005. "A study of on integrated intercity travel demand model." *ransportation Research Part A: Policy and Practice 39.4* 367-381.

Zheng, Zijian, Pan Lu, and Denver Tolliver. 2016. "Decision tree approach to accident prediction for highway–rail grade crossings: Empirical analysis." *Transportation Research Record* 115-122.

Zhou, X., Lu, P., Zheng, Z., Tolliver, D., & Keramati, A. 2020. "Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree." *Reliability Engineering & System Safety* 1 (200).