A systematic review of machine learning methodologies for modelling passenger mode choice

Tim Hillel [*] Michel Bierlaire [*] Ying Jin [†]

October 25, 2019

Report TRANSP-OR 191025
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

[*]École Polytechnique Fédérale de Lausanne (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Transport and Mobility Laboratory, Switzerland, {tim.hillel,michel.bierlaire}@epfl.ch
[†]University of Cambridge, Department of Architecture {Ying.Jin@aha.cam.ac.uk}

Abstract

Machine Learning (ML) approaches are increasingly being investigated as an alternative to Random Utility Models (RUMs) for modelling passenger mode choice. These approaches have the potential to provide valuable insights into choice modelling research questions. However, the research and the methodologies used are fragmented. Whilst systematic reviews on RUMs for mode choice prediction have long existed and the methods have been well scrutinised for mode choice prediction, the same is not true for ML models. To address this need, this paper conducts a systematic review of ML methodologies for modelling passenger mode choice. The review analyses the methodologies employed within each study to (a) establish the state-of-research frameworks for ML mode choice modelling and (b) identify and quantify the prevalence of methodological limitations in previous studies.

A comprehensive search methodology across the three largest online publication databases is used to identify 468 unique records. These are screened for relevance, leaving 60 peer-reviewed articles containing 63 primary studies for data extraction. The studies are reviewed in detail to extract 15 attributes covering five research questions, concerning (i) classification techniques, (ii) datasets, (iii) performance estimation, (iv) hyper-parameter selection, and (v) model selection.

The review identifies ten common methodological limitations. Five are determined to be methodological pitfalls, which are likely to introduce bias in the estimation of model performance. The remaining five are identified as areas for improvement, which may limit the achieved performance of the models considered. A further eight general limitations are identified, which highlight gaps in knowledge for future work.

List of acronyms

AB AdaBoost. 18

ABM Agent Based Model. 29

AI Artificial Intelligence. 9

AMPCA Arithmetic Mean Probability of Correct Assignment. 32

ANN Artificial Neural Network. 5, 6, 9, 17, 18, 31, 35, 36

API Application Programming Interface. 27

BIC Bayesian Information Criterion. 32

BL Bayesian Learner. 18

BN Bayesian Network. 18

CART Classification and Regression Trees. 7

CNL Cross-Nested Logit. 3, 18, 19

DCM Discrete Choice Model. 3, 19

DT Decision Tree. 5, 7–9, 17, 18, 27, 31, 35

EL Ensemble Learning. 5, 7–9, 18, 20, 21, 35, 40

FFNN Feed-Forward Neural Network. 6, 7, 18

FL Fuzzy Logic. 9, 11

GB Gradient Boosting. 18

GBDT Gradient Boosting Decision Trees. 8, 20, 21, 40

GPS Global Positioning System. 13, 21

IVTT In-Vehicle Travel Time. 26

LOS Level of Service. 19, 20, 25, 29, 30

LR Logistic Regression. 5–7, 17, 19–21, 32, 34

ML Machine Learning. 1, 3–6, 9–11, 16, 17, 20, 21, 34–37, 39, 40, 42

MLP Multi-Layer Perceptron. 6, 36

MNL Multinomial Logit. 19

MSE Mean Squared Error. 32, 34

NB Naïve Bayes. 18

NL Nested Logit. 3, 18, 19

OOB Out-Of-Bootstrap. 31

OVTT Out-of-Vehicle Travel Time. 26

PNN Probabilistic Neural Network. 18

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses. 11, 13, 14, 42

PT Public Transport. 26, 27

RBF Radial Basis Function. 8

RBFNN Radial Basis Function Neural Network. 18

RBML Rule-Based Machine Learning. 11, 17, 18

RF Random Forest. 18, 31

ROC Receiver Operating Characteristic. 32

RUM Random Utility Model. 1, 3–7, 9, 17, 19–21, 29, 30, 40, 42

SVM Support Vector Machine. 5, 8, 9, 18, 36

VOC Vehicle Operating Cost. 26

VoT Value of Time. 4, 19, 29

1 Introduction

Solutions used both in industry and academic research for modelling passenger mode choice have traditionally relied almost exclusively on econometric Discrete Choice Models (DCMs) based on the random utility framework (McFadden 1981). However, there have been two key recent drivers which have resulted in researchers exploring alternative approaches. Firstly, the adoption of new transportation-related technologies has driven a step change in the availability of data on passenger movements of several orders of magnitude. Secondly, there have recently been significant breakthroughs in ML research, which have resulted in numerous success stories of ML applications in other similar tasks.

These drivers have resulted in a number of recent research applications of ML techniques to the mode choice problem. The application of ML has the potential to provide valuable new insights into mode choice modelling research questions. However, the existing research is fragmented, and there have been few studies which comprehensively compare ML techniques with each other and with RUMs. Additionally, whilst systematic reviews on RUMs for mode choice prediction have long existed, and the methods have been well scrutinised, the same is not at all true for ML models.

To address these limitations, this paper conducts a systematic review of ML approaches to passenger mode choice modelling. The review focuses on the methodologies employed within each study in order to (a) establish the state-of-research frameworks for ML mode choice modelling and (b) identify and quantify the prevalence of methodological limitations in previous studies.

2 Machine learning for mode choice prediction

The predominant approach used in industry and academic research for modelling passenger mode choice are Random Utility Models (RUMs) (McFadden 1981). These models rely on functions of the input variables, called *utility specifications*, for each option (mode) in the choice set. In a logit model, the utilities (the output values of the utility specifications) are then passed through a logistic function to generate choice probabilities for each option in each considered choice situation. Other model structures, such as the Nested Logit (NL) and Cross-Nested Logit (CNL), allow for these probabilities to be calculated given correlations among the options in the choice-set (Ben-Akiva and Lerman 1985, Chapter 10). The parameter values in the utility specification are estimated using

maximum likelihood estimation, in order to maximise the joint likelihood of the training data given the model.

The utility specifications in the model are defined by the modeller prior to fitting the model. This allows the modeller to incorporate established behavioural theory and expert knowledge into the model. The estimated parameter values can then be used to test hypothesis about the consistency of RUM predictions with expected behaviour. These parameters can also be used to extract key behavioural indicators, such as the elasticities and Value of Time (VoT) (Ben-Akiva and Lerman 1985, Chapter 5).

The nature of all of the relationships between the input variables and the utilities must be defined in the utility specifications. This includes all non-linear transformations of variables and any interactions between them. As the utility functions are specified in advance of estimating the model, this means that the modeller must hypothesise and test these relationships manually.

In ML terminology, a RUM can be considered as a *supervised probabilistic classifier*; the aim of the model is to predict the probability of an individual choosing each mode (i.e. the *classes*), given a set of *features* (variables) describing the choice situation. The modeller has access to a finite *labelled* dataset of choice situations alongside their class labels (the option chosen) to train the model. This task therefore appears to be a natural application for ML classification algorithms, which have shown a great deal of success with similar problems in other research domains, such as image recognition, text classification, and disease detection (Hastie, Friedman, and Tibshirani 2008).

Rather than relying on predefined utility specifications, ML classification algorithms instead attempt to identify the relationship between input features and the class labels directly from the data, without input from the modeller. The majority of ML classifiers have the ability to automatically identify non-linear relationships between the inputs and outputs. The added flexibility in ML classifiers compared to RUMs may allow the model to *generalise* to relationships not previously considered and therefore which would not have been identified using manually defined utility specifications. However, this flexibility presents a much higher propensity for a model to *overfit* to noise in the training data. Additionally, as there is no underlying behavioural model in a ML model, it is not straightforward to check for or ensure for behavioural consistency of the model predictions.

The *generalisation error* measures the ability of a classifier to accurately predict class probabilities for previously unseen data. In ML applications, this is typically estimated by validating model on separate out-of-sample data, unseen by the model during training. The model validation ensures that the model has successfully generalised to valid relationships in the data, without fitting to noise in the data. There is therefore a balance between *underfitting* and *overfitting*, known as the *bias-variance trade-off* (Hastie, Friedman, and Tibshirani 2008). If a model has high *bias* it is not flexible enough to identify valid correlations that are present in the real-world test data (underfitting). If a model has high *variance* it is too flexibile and is replicating noise in the data without generalising to valid correlations between the input features and class labels (overfitting).

The flexibility of an algorithm to fit to the data when training a model is *regularised* using the algorithm's *hyper-parameters*. These are parameters of the algorithm, such as the maximum permissible number of splits in a decision tree, which impact the bias and variance of the fitted model (see section 2.1). Model performance is highly dependent on

chosen hyper-parameter values, and so it is important to select appropriate values for both the task and data (Hoos et al. 2014).

In order for the model validation to be a true estimate of the generalisation error, the test-set must be truly separate from the training data and not seen by the model at any stage prior to final testing. Any exposure of the model to the test-set before final testing (e.g. by selecting hyper-parameters based on test performance, see section 4.4) or shared information between the train and test-sets (e.g. through inappropriate sampling of hierarchical data, see section 4.3) will result in *data-leakage*, and allow the model to fit to the test-set. This will result in the test-error underestimating the generalisation error.

ML classification investigations can be broken into two main processes; *model development* and *model evaluation*. In the model development process, the modeller tries to develop a model with the aim of minimising its generalisation error. This includes hyper-parameter selection, feature processing, and algorithm selection/development. In the model evaluation process, the modeller then estimates the generalisation error of the model, typically by testing on an out-of-sample test set.

If the model development in a study is not appropriate (e.g. if an appropriate hyper-parameter selection is not used) the model will achieve a lower generalisation error than is possible for that algorithm. This means differences in model performance may be due to differences in the model development process, and not to do with the potential performance of the algorithm itself. As such, it is important to consider the model development process when making comparisons between the relative performance of *algorithms* for a given task. Conversely, if the model evaluation process used is inappropriate (e.g. if there is data-leakage from the test-set during model development) than the estimate of generalisation error will be biased. This will result in unreliable evaluation of model performance. As such, it is important to consider the model evaluation process for any evaluation of *model* or *algorithm* performance for a task.

As there is a lot of overlap in the theory and practice in the fields of RUMs and ML, there are a substantial number of equivalent or nearly-equivalent terms between them. As this paper reviews ML methodologies, the ML terminologies have been preferred. For clarity of the associated material, table 1 summarises some of the equivalent and nearly equivalent terms used in this paper.

2.1 Machine learning classification algorithms

In order to provide an understanding of the techniques used, the following sections give an overview of five classes of supervised classification algorithm which have previously been used to investigate mode choice, including introducing their main hyper-parameters: Logistic Regression (LR), Artificial Neural Networks (ANNs), Decision Trees (DTs), Ensemble Learning (EL), and Support Vector Machines (SVMs).

Logistic Regression The Logistic Regression (LR) classifier is very similar to a logit RUM, with linear functions of the input features passed through the softmax (logistic) function to generate class probabilities. This paper distinguishes between the two models on the basis that the LR model does not include an explicit behavioural model or utility specifications. Instead, with LR all features are included uniformly for all modes, with a single weight (equivalent to *parameters* in RUMs) trained for each feature for each mode.

Table 1: Equivalent and nearly-equivalent terms between random utility and ML models.

Random utility	**Machine learning**	**Notes**
Attribute	Feature	Variables of the choice-set.
Covariate	Feature	Socio-economic variables of the individual. No distinction is made between attributes and socio-economic covariates in ML classifiers.
Parameter	Weights	Referred to as coefficients in linear utility functions in RUMs. Weights are used only in parametric ML models (LR and ANNs).
Estimate	Train	Both are often referred to as *fitting* the model.
Logistic function	Softmax	Referred to as the sigmoid function in the binary case.

As the LR classifier does not conform to a behavioural model, it is not subjected to the corresponding constraints on the parameters, and so is typically more flexible than a RUM. As such, the LR classifier has a higher propensity for overfitting. This can be addressed by using *regularisation*, which penalises the model during fitting on the basis of the values of the weights. *L1* regularisation (also known as *lasso* regularisation) penalises the model for the sum of absolute values of the weights. Conversely, *L2* regularisation (also known as *ridge* regularisation) penalises the model for the sum of squares of the weights. The amount of regularisation is controlled using the C hyper-parameter, with a larger value of C indicating more regularisation (higher penalty for the values of the weights).

Artificial Neural Networks Artificial Neural Network (ANN) is a term used to cover a family of classifiers which mimic the network structure of the brain. Whilst there are a huge variety of possible ANN structures for dealing with different input data types (e.g. images, time-series, natural language etc), mode choice applications have typically relied on the Feed-Forward Neural Network (FFNN) (also known as the Multi-Layer Perceptron (MLP)) (Svozil, Kvasnicka, and Pospichal 1997).

A FFNN consists multiple *layers* of *nodes* (neurons), including (i) an input layer, which passes the feature values to the network; (ii) an output layer, which outputs the predicted values from the network; and (iii) any number of hidden layers. For probabilistic classification, the number of nodes in the input and output layers is fixed by the number of features and classes in the data respectively. The hidden layers can each contain any number of nodes. In a fully connected network, every node in one layer is linked to every node in the next layer.

Each node has an activation function, which determines the output of that node from the weighted sum of its inputs. There are many possible activation functions used in practice, including linear, sigmoid, tanh, softplus, softsign, ReLU (rectified linear unit), ELU (exponential linear unit), and SELU (scaled exponential linear unit).

FFNN are highly flexible in terms of relationships they can approximate. It can be shown via the universal approximation theorem (Hornik 1991) that a FFNN containing

one or more hidden layers with a sufficient (finite) number of nodes and *any* non-linear, bounded, continuous activation function can approximate any continuous function.

As with the logit and LR models, the output values of the network are passed through the softmax function to generate classification probabilities. Note that a LR can be thought of as a FFNN with no hidden layer and a linear activation function.

The weights (parameters) for each link in the network are fitted to the input data (equivalent to estimating a RUM). FFNNs are most commonly trained using *mini-batch gradient descent*. This algorithm splits the input data into small batches. The network weights are then updated iteratively on the individual batches. Each time the model sees all of the data once is termed an *epoch*. The number of epochs can then be controlled to limit overfitting. Further regularisation can be applied using *dropout*, where a proportion of the neurons are dropped randomly from the network for each mini-batch of data (Srivastava et al. 2014)

Decision Trees Decision Trees (DTs) (or Classification and Regression Trees (CART)) are classifiers which sort data into groups using a set of sequential splits in a tree-like structure (Breiman 2017). The most commonly used Decision Trees (DTs) are fitted using recursive binary splits, with each split chosen to result in the greatest reduction in the *randomness* of the data at that point (i.e. it is a *greedy* algorithm). Two metrics can be used to measure how shuffled the data are, *Gini impurity* and *entropy*.

To calculate each split, the data at the selected node are sorted according to each feature, and each possible binary split point (less/greater than a certain value) is tested for each feature. The split point which results in the greatest reduction in the impurity or entropy (across all features) of the data is then selected, resulting in two new child nodes. The same algorithm can then be applied recursively to the child nodes. This process is repeated until a stopping condition is met.

The stopping conditions can be set using a combination of different hyper-parameters in order to prevent overfitting. For example, the *maximum depth* specifies the maximum number of sequential splits which can be applied along a branch, the *minimum leaf size* specifies the minimum size *both* nodes of a split must have in order for a split to take place, and the *minimum split size* specifies the minimum number of samples in a node for a split to be considered at that node.

Decision trees can only generate discrete predictions, and so are not suitable for probabilistic mode choice prediction when used independently. However, they can be combined in ensembles to generate probabilistic predictions.

Ensemble Learning Ensemble Learning (EL) algorithms combine several *weak learners* in an ensemble to improve the quality of predictions. Provided the weak learners make errors *independently* (i.e. the learners are uncorrelated), and are more likely to be right than wrong, then combining them in an ensemble reduces their individual uncertainty.

DTs are the predominantly used weak learners for Ensemble Learning (EL). DTs have high variance, making them highly unstable (small changes in the input result in large differences between classifiers). As such, it is relatively easy to train uncorrelated DTs compared to more stable classifiers (e.g. LR). In addition, DTs are algorithmically simple to fit and obtain predictions from. This means that large ensembles of DTs can fit and predict in reasonable time.

By combining the binary splits of a large number of DT in an ensemble, EL algorithms are able to approximate arbitrary complex, non-linear relationships (in particular non-continuous relationships). This makes EL algorithms highly flexible at generalising to relationships in the data. Provided appropriate hyper-parameters are used, EL algorithms can perform well on a wide range of classification tasks. Gradient Boosting Decision Trees (GBDT) in particular have been shown to have best in class performance in several supervised classification tasks (Zhang, Liu, et al. 2017; Brown and Mues 2012; Chapelle and Chang 2011; Caruana and Niculescu-Mizil 2006).

Support Vector Machines The Support Vector Machine (SVM) algorithm makes use of a *kernel* to transform the data into a high-dimension space. The algorithm then finds the optimal linear decision surface (or *hyper-plane*) in the transformed space which divides the data into two classes (Cortes and Vapnik 1995).

There are multiple kernels which can be used to transform the data, including linear (no transformation), polynomial, Radial Basis Function (RBF) (or *Gaussian*), and sigmoid.

For linearly-separable data (within the transformed space), the optimal hyperplane is the one that exactly divides the data without misclassification whilst maximising the possible *margin*. The margin is defined as the perpendicular distance between the hyperplane and the nearest data points (these distances are the *support vectors*). For complex, real-world examples, the input data are not normally linearly-separable, even within the transformed space. As such, there is a balance between the width of the hyperplane and the number of misclassifications of the training data. This is controlled using the regularisation parameter (C). A higher value of C represents a higher importance of the misclassified points (higher variance), whilst a lower value of C will put a higher importance on the width of the hyperplane (higher bias).

Support Vector Machines (SVMs) are inherently binary classifiers. However, they can be used for multiclass classification using either a *one-vs-rest* or *one-vs-one* strategy. The one-vs-rest strategy trains a single binary classifier per class, with each classifier trained to predict whether an instance belongs to a class or not. Each binary classifier is trained on all of the data. The one-vs-one classification strategy trains a binary classifier on each unique pair of classes (i.e. $J(J-1)/2$ binary classifiers for a J-class problem) and predicts which class the instance belongs to. Each classifier is trained on all samples from the data belonging to the corresponding pair of classes (i.e. a walk-vs-cycle classifier would be trained on only the walking and cycling journeys). For both strategies, the confidence scores of the respective classifiers are used to determine the predicted class for unseen data.

SVMs output a continuous score for each prediction. This score can be interpreted as the confidence of the classification. However, these scores do not correspond well to class probabilities (Niculescu-Mizil and Caruana 2005). Methods to calibrate the scores as class probabilities are proposed by Wu, Lin, and Weng (2004) and Platt (1999).

SVM complexity scales at a minimum with $O(n^2)$, where n is the number of instances (rows) in the data (Bottou and Lin 2007). This rises to $O(n^3)$ for high C (regularisation) values. This can cause computational issues and/or considerable fit-times when using SVMs with large datasets.

2.2 Need for a review of machine learning methodologies

As discussed above, ML approaches are increasingly being investigated as an alternative to RUMs for mode choice prediction. However, the research is fragmented, with inconsistent methodologies used in past studies. The implications of different methodological decisions is not yet well understood. As such, there is a need to evaluate the methodologies used in previous studies in order to understand the scope of ML approaches and establish good standard practices.

There exist several review papers in the literature focusing on mode choice modelling, including those by Barff, Mackay, and Olshavsky (1982), Hensher and Johnson (1983), Kruger (1991), Nerhagen (2000), Meixell and Norbis (2008), Ratrout, Gazder, and Al-Madani (2014), Jing et al. (2018), and Minal and Sekhar (2014). However, all but two of these reviews focus exclusively on statistical RUM techniques. Ratrout, Gazder, and Al-Madani (2014) and Minal and Sekhar (2014) explicitly review ML and Artificial Intelligence (AI) approaches within the literature, including ANN approaches to mode choice modelling alongside RUM based studies. The studies conclude that ANN have been successfully used for mode choice modelling, in particular due to their flexibility when dealing with multidimensional non-linear data. Ratrout, Gazder, and Al-Madani (2014) further state that whilst the vast majority of existing studies are based on logit models, it can be expected that the trend of using ML methods will continue in future.

Whilst the studies by Ratrout, Gazder, and Al-Madani (2014) and Minal and Sekhar (2014) evaluate some of the existing ML mode choice research, they have a number of limitations. Primarily, they focus only on ANN (and Fuzzy Logic (FL)) approaches, and as such do not cover any contributions using other ML techniques, including DTs, SVMs, and EL. Secondly, these reviews are intended to be exploratory as opposed to systematic, and do not represent comprehensive coverage of all relevant studies. Additionally, the reviews are intended to be general, and do not focus on specific aspects of the methodologies used in each study. Finally, there have been a substantial number of new studies published since these reviews were carried out. To address these limitations, this paper conducts a systematic review of ML approaches to passenger mode choice modelling.

2.3 Overview of paper

The remainder of the paper is laid out as follows. Section 3 outlines the methodology for the review, including the research questions, review protocol, and study selection. Next, Section 4 presents the results of the review, first giving an overview of the selected studies, before exploring each research question in turn to identify the methodological limitations. The limitations are categorised into

- *technical limitations:* technical issues within the methodologies of specific studies that are likely to have an impact on their results, which are further categorised into
  - *pitfalls:* issues in the model evaluation process which are likely to make the results of an investigation unreliable, and
  - *areas for improvement:* modelling decisions which are not incorrect but could be addressed in order to improve the reliability of the results for comparing the classification algorithms and/or the predictive performance of the models;

and

- *general limitations:* gaps in knowledge or areas across multiple studies that require further investigative work.

Finally, section 5 summarises the findings, identifies potential limitations of the review, and presents the conclusions.

3 Methodology

The procedure for this systematic review is adapted from that given by Kitchenham and Charters (2007). The suggested procedure suggested has 10 stages broken down into three phases:

- *Planning the review*
    1. Identification of the need for a review
    2. Specifying the research questions
    3. Developing a review protocol
- *Conducting the review*
    4. Identification of research
    5. Selection of primary studies
    6. Study quality assessment
    7. Data extraction and monitoring
    8. Data synthesis
- *Reporting the review*
    9. Specifying dissemination mechanisms
    10. Formatting the main report

This review is focused on summarising the methodologies used in each study, and as such, no attempt is made to draw conclusions from the aggregate results or combined findings of the studies. Consequently, no assessment of the quality of each study is made (step 6 in the framework). The review presented in this paper therefore consists of the nine remaining stages presented above.

The focus of this review is the methodologies used in ML approaches to modelling passenger mode choice. In particular, the review serves to investigate the following research questions:

1. Which classification techniques have been used to investigate mode choice?
2. What is the nature of datasets used to investigate mode choice?
3. How is model performance determined?

4. How are optimal model hyper-parameters selected?

5. How is the best model selected?

3.1 Review protocol

This section outlines the protocol for the search strategy, selection criteria, and data extraction strategy.

3.1.1 Search strategy

The search strategy is used to identify relevant papers to the review. In order to ensure full coverage of relevant papers, papers are collated from three databases: the two major online curated publication databases, Web of Science and Scopus; and the Google Scholar search engine. The same search is repeated for each database.

This review focuses on papers with a core focus of mode choice modelling. As such, only papers with the *title* directly relating to mode choice are included. The following initial phrases are tested: *mode choice*, *mode selection*, *travel mode*, *transport mode*, *transportation mode*, and *mode of travel*.

In order to only select papers that discuss ML techniques, only papers with one or more selected phrases relating to ML across all relevant fields are selected. The following initial phrases are tested: *machine learning*, *neural network*, *decision tree*, *ensemble method*, *random forest*, *boosting*, and *support vector*.

Papers from any period up until the search date are included in the search.

The initial search phrases are tested in different combinations across the three databases. The terms *mode of travel* and *mode selection* are omitted from the title search, as they return no relevant papers when used alongside the ML search terms.

Additionally, a number of papers using Fuzzy Logic (FL) (within Rule-Based Machine Learning (RBML)) were found in the initial search results. To reflect this, the phrase *fuzzy logic* is added to the search across all relevant fields.

3.1.2 Selection criteria

The following eligibility criteria are determined for the papers found in the search to be included in the study:

- Studies in peer-reviewed journals or conference proceedings written in English
- Studies which investigate passenger mode choice at disaggregate (individual) level.
- Studies which employ one or more ML technique(s) for predictive modelling.

Paper selection is carried out using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al. 2009). Firstly, duplicates are removed from the search records. Secondly, the record titles and abstracts are screened against the eligibility criteria. Finally, the remaining full-text articles are assessed for eligibility. All stages of the selection criteria are carried out independently by the first author.

Where a paper contains more than one relevant modelling scenario (defined as having separate input datasets and different methodologies), each modelling scenario is treated as a separate study for the analysis.

3.1.3 Data extraction strategy

In order to extract the necessary data from each study without bias, a list of attributes is collected from each study. The attributes, shown in table 2, are intended to be specific, objective, and quantifiable/categorical, in order to limit subjectivity in the data extraction process. Together the attributes provide the evidence for the research questions.

Table 2: Research questions and corresponding attributes of studies for data extraction.

No.	Description
Q1	**Which classification techniques have been used to investigate mode choice?**
Q1a	Classification algorithms used in study
Q1b	Logit model implementation
Q2	**What is the nature of datasets used to investigate mode choice?**
Q2a	Nature of dataset
Q2b	Unit of analysis
Q2c	Dataset availability
Q2d	Modes in choice-set
Q2e	Modelling of mode-alternatives
Q2f	Input features dependent on output choice
Q2g	Hierarchical data
Q3	**How is model performance determined?**
Q3a	Validation method
Q3b	Sampling method
Q3c	Performance metrics used
Q4	**How are optimal model hyper-parameters selected?**
Q4a	Hyper-parameter search method
Q4b	Hyper-parameter validation method
Q4c	Hyper-parameter validation data
Q5	**How is the best model selected?**
Q5a	Model selection technique

Data extraction is carried out independently by the authors. Each paper is reviewed in detail, with each attribute for each study determined and tabulated in a spreadsheet. Separate entries are entered into the spreadsheet for papers containing multiple studies (modelling scenarios).

3.2 Study selection

The following search terms are used to carry out the search strategy outlined in section 3.1.

- **Web of Science**: *TITLE: ("mode choice" OR "travel mode" OR "transport mode" OR "transportation mode") AND TOPIC: ("machine learning" OR "neural network" OR "decision tree" OR "ensemble method" OR "random forest" OR "boosting" OR "support vector" OR "fuzzy logic")*

- **Scopus**: *(TITLE ("mode choice" OR travel mode" OR "transport mode" OR "transportation mode") AND TITLE-ABS-KEY ("machine learning" OR "neural network" OR "decision tree" OR "ensemble method" OR "random forest" OR "boosting" OR "support vector" OR "fuzzy logic"))*

- **Google Scholar**: *(intitle:"mode choice" OR intitle:"travel mode" OR intitle:"transport mode" OR intitle:"transportation mode") AND ("machine learning" OR "neural network" OR "decision tree" OR "ensemble method" OR "random forest" OR "boosting" OR "support vector" OR "fuzzy logic").*
  Due to the restriction on search length in Google Scholar, this search is divided into two separate searches, with the results combined.

The search was carried out on 25/05/2018 on all three databases. Figure 1 shows a PRISMA flowchart of the study selection process.

There were 78 records returned from the Web of Science search, 220 records from Scopus, and 442 records from Google Scholar, for a total of 740 records. Duplicates are then removed, leaving 468 records to be screened. The total number of records after removing duplicates is more than were obtained from any one database, showing that there were results from Web of Science/Scopus which were not returned with the Google Scholar search.

The 468 remaining records are then screened as to whether they meet the eligibility criteria outlined in section 3.1. During screening, 327 papers are excluded for relevance on the basis of their title and abstract. The majority of these records relate to transportation mode detection from Global Positioning System (GPS) data. Of the records which are deemed relevant, a further 45 are excluded as they are not published in peer reviewed publications, (e.g. Thesis/dissertation, unpublished paper, book section), or are not written in English (only having a title and abstract in English).

The full text is obtained for the remaining 96 articles for further review. Of these, a further 36 are excluded on the basis of the selection criteria, as detailed in fig. 1. This leaves 60 selected articles for data-extraction.

Two articles contain multiple modelling scenarios, for a total of 63 separate studies for meta-analysis.

4 Results and discussion

This section presents the results obtained from the systematic review process. Firstly, section 4 provides an overview of the 60 articles used for data extraction, including the publication sources and years. The articles with multiple studies are identified, and each of the 63 studies are given a unique identifier. Sections 4.1 to 4.4 then use evidence from the 63 studies to explore each of the five research questions in turn.

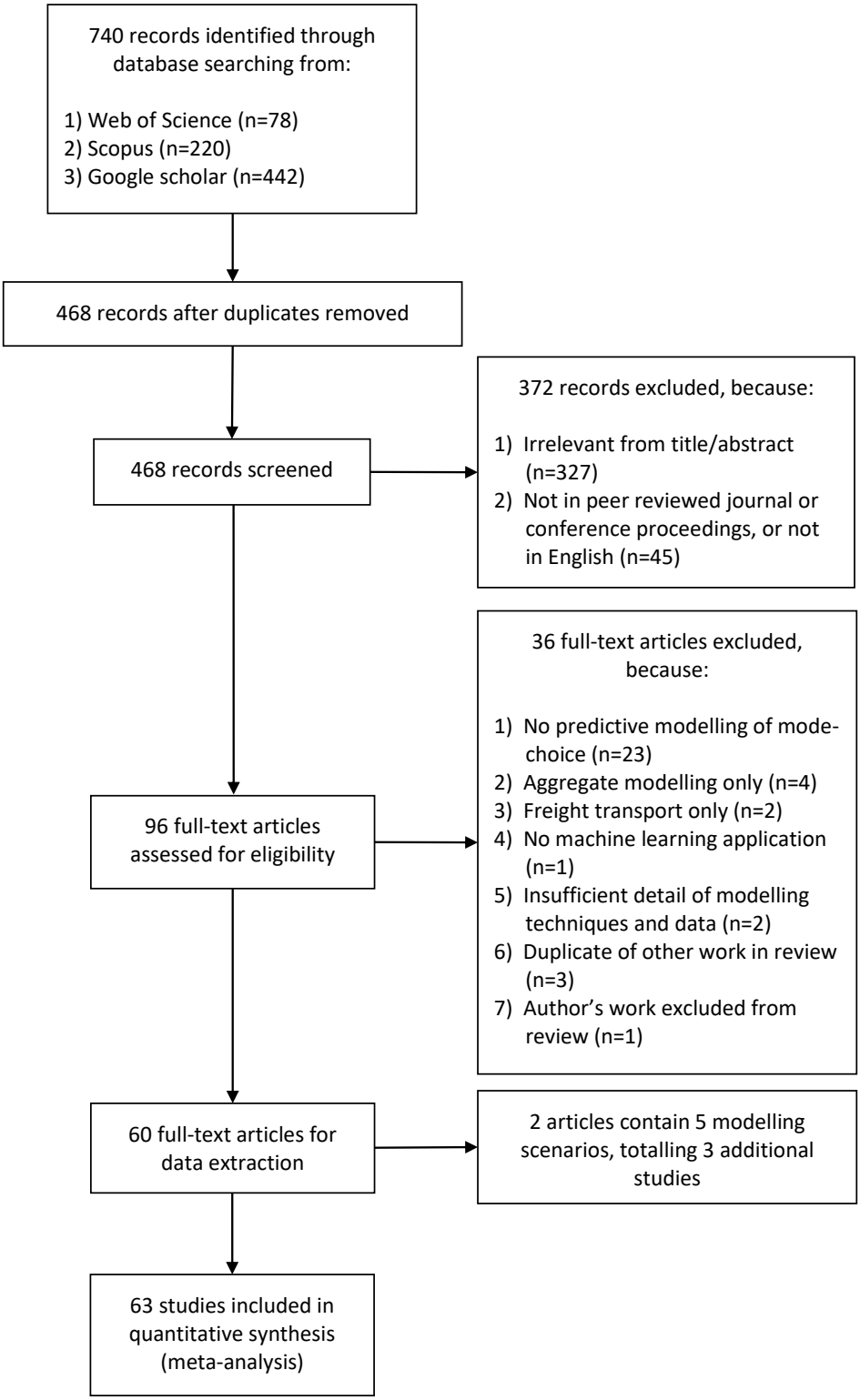

Figure 1: PRISMA flowchart of study selection process.

Articles for data extraction

This section provides an overview of the 60 articles used for data extraction. Table 3 provides a unique identifier for each article, alongside its individual reference.

Table 3: Selected primary articles for review.

No.	Paper	No.	Paper
S1	Andrade, Uchida, and Kagaya (2006)	S31	Moons, Wets, and Aerts (2007)
S2	Assi et al. (2018)	S32	Nam et al. (2017)
S3	Biagioni et al. (2009)	S33	Omrani (2015)
S4	Cantarella and de Luca (2003)	S34	Omrani et al. (2013)
S5	Cantarella and de Luca (2005)	S35	Papaioannou and Martinez (2015)
S6	Chalumuri et al. (2009)	S36	Pirra and Diana (2017)
S7	Cheng et al. (2014)	S37	Pitombo et al. (2015)
S8	Dell'Orco and Ottomanelli (2012)	S38	Pulugurta, Arun, and Errampalli (2013)
S9	Edara, Teodorović, and Baik (2007)	S39	Raju, Sikdar, and Dhingra (1996)
S10	Ermagun, Rashidi, and Lari (2015)	S40	Ramanuj and Gundaliya (2013)
S11	Errampalli, Okushima, and Akiyama (2007)	S41	Rasouli and Timmermans (2014)
S12	Gao et al. (2013)	S42	Seetharaman et al. (2009)
S13	Gazder and Ratrout (2015)	S43	Sekhar, Minal, and Madhu (2016)
S14	Golshani et al. (2018)	S44	Semanjski, Lopez, and Gautama (2016)
S15	Hagenauer and Helbich (2017)	S45	Shafahi and Nazari (2006)
S16	Hensher and Ton (2000)	S46	Shukla et al. (2013)
S17	Hossein Rashidi and Hasegawa (2014)	S47	Subba Rao et al. (1998)
S18	Hussain et al. (2017)	S48	Tang, Yang, and Zhang (2012)
S19	Jia, Cao, and Yang (2015)	S49	Tang, Xiong, and Zhang (2015)
S20	Juremalani (2017)	S50	Van Middelkoop, Borgers, and Timmermans (2003)
S21	Karlaftis (2004)	S51	Wang and Ross (2018)
S22	Kedia, Saw, and Katti (2015)	S52	Wang and Namgung (2007)
S23	Kumar, Sarkar, and Madhu (2013)	S53	Xian-Yu (2011)
S24	Lee, Derrible, and Pereira (2018)	S54	Xie, Lu, and Parkany (2003)
S25	Li et al. (2016)	S55	Yin and Guan (2011)
S26	Liang et al. (2018)	S56	Zenina and Borisov (2011)
S27	Lindner, Pitombo, and Cunha (2017)	S57	Zhang and Xie (2008)
S28	Lu and Kawamura (2010)	S58	Zhao et al. (2010)
S29	Ma (2015)	S59	Zhou and Lu (2011)
S30	Ma, Chow, and Xu (2017)	S60	Zhu et al. (2017)

Two papers [S5; S21] contain multiple modelling scenarios, using separate datasets and a different methodology for each one. A separate identifier is assigned to each modelling scenario in each of these papers, and they are treated as separate studies for the meta-analysis. Table 4 provides the label for the additional studies, alongside a description of each modelling scenario. The two papers have a total of five modelling scenarios. This results in a total of 63 studies for meta-analysis.

Four further papers have multiple modelling phases but are deemed not to be separate studies for the purpose of this review.

S6 includes input datasets for two cities: Visakhapatnam and Nagpur. The datasets are collected as part of the same study and the modelling methodology used for each is identical. As such, both are treated as the same study. The dataset for Visakhapatnam is analysed in section 4.2, with the only difference between the two for the purpose of the review being that the Nagpur city dataset has 27 fewer records (1045 vs 1018).

S16 also contains multiple datasets, one for Sydney and one for Melbourne. However, both datasets are collected using the same methodology, and are used in combination in

Table 4: Primary studies with multiple modelling scenarios in review.

No.	Paper	No.	Scenario
S5	Cantarella and de Luca (2005)	S5.1	VENETO dataset
		S5.2	UNISA dataset
S21	Karlaftis (2004)	S21.1	Interurban mode choice in Australia
		S21.2	Commuter mode choice in Athens, Greece
		S21.3	Commuter mode choice in Las Condes-CBD corridor, Chile

the same model. The combined dataset is used for the analysis in section 4.2.

S13 includes three separate modelling phases, which all model slightly different choice situations. However, each phase uses different subsets of the same dataset, and all use the same methodology. *Phase I*, which models the revealed preference choice between car and plane, is used for the analysis in the review.

Finally, S17 also involves three modelling stages which use different subsets of the same dataset: *Model 1* predicts total number of trips in a day, *Model 2-1* predicts attributes of the first trip in a day made by an individual, and *Model 2-2* predicts attributes of subsequent trips made by using attributes of the previous trip. As *Model 2-2* uses details of the previous trip taken from the dataset (and not as predicted by a model), it is not relevant as a predictive model within the scope of this study. As such only *Model 2-1* is analysed within this review.

Publication source

Table 5 provides details of all journals and conferences/proceedings from which more than one article was selected. The articles come from a wide spread of publications, with a total of 28 different journals and 16 different conferences featured. The majority of the papers (39/60) are published in journals, making up 65 % of the articles, with the remaining 21 papers (35 %) published in conference proceedings.

The top two sources for articles are the Transportation Research Record Journal and the Transportation Research Board Annual Meeting conference, both of which are published by the Transportation Research Board. Together, they make up 22 % (13/60) of the articles.

Publication year

Figure 2 shows the distribution of article publication dates from 1995 to 2018.

There is a clear upwards trend of increasing number of publications regarding ML applications to mode choice per year. Over half of the selected articles (31/60) were published from 2012 onwards. Conversely, only 10 relevant papers were published prior to 2007. Data for 2018 is incomplete as the search was carried out in May 2018, but there are still a considerable number of articles published in this year (5/60).

Table 5: Summary of publication sources contributing more than one paper to review. Multi-conference proceedings are shown in bold, with the individual conferences in italics below.

Publication	**Type**	**No.**
Transportation Research Record	Journal	7
Transportation Research Board Annual Meeting	Conference	6
Transportation Research Procedia:	**Proceedings**	**4**
Euro Working Group on Transportation	*Conference*	*3*
Transportation Planning and Implementation Methodologies for Developing Countries	*Conference*	*1*
Travel Behaviour and Society	Journal	2
East Asia Society for Transportation Studies	Conference	2
International Conference of Chinese Transportation Professionals	Conference	2
Totals (all papers)	**Journal**	39
	Conference	21

4.1 Which classification techniques have been used to investigate mode choice?

The following sections present an overview of the classification techniques used in the 63 studies in the review.

Q1a: Classification algorithms used in study

Based on the responses to Q1a, the classification techniques are grouped into nine categories, as shown in table 6. A brief overview of the classification techniques identified in this paper is given in section 2.1. For each algorithm, an example paper from the systematic review which makes use of that algorithm is provided.

Table 7 shows which classification techniques are used in each study. The majority of studies (40/63) compare ML techniques with statistical RUMs and LR, making logit models the most commonly used classification technique in the studies. The most commonly used ML algorithms are ANNs (30 studies), followed by DTs (16 studies), and RBML (11 studies). The remaining classes of algorithms have been used in 10 or less studies each.

Q1b: Logit model implementation

Whilst the overall focus of this review is the ML methodologies used in the studies, Q1a identifies 40 studies which compare ML approaches with logit models (statistical RUMs and LR). As such, this section gives a brief overview of the logit models used in these studies.

As discussed in section 2.1, a distinction is made between RUMs, which use logistic regression with explicit utility functions for each mode within the random utility framework; and LR classification, where all input features are included uniformly for all classes

Table 6: Classification techniques used in studies in review.

Classification algorithm	Example reference
1. Logit models (Log)	
Logit	Cantarella and de Luca (2005)
Nested Logit (NL)	Hensher and Ton (2000)
Cross-Nested Logit (CNL)	Nam et al. (2017)
2. Artificial Neural Networks (ANNs)	
Feed-Forward Neural Network (FFNN)	Lee, Derrible, and Pereira (2018)
Radial Basis Function Neural Network (RBFNN)	Omrani (2015)
Probabilistic Neural Network (PNN)	Zhou and Lu (2011)
Other neural network structures	Cantarella and de Luca (2003)
3. Decision Trees (DTs)	Karlaftis (2004)
4. Ensemble Learning (EL)	
Random Forests (RFs)	Hossein Rashidi and Hasegawa (2014)
Gradient Boosting (GB)	Wang and Ross (2018)
AdaBoost (AB)	Biagioni et al. (2009)
Bagging	Hagenauer and Helbich (2017)
5. Support Vector Machines (SVMs)	Xian-Yu (2011)
6. Bayesian Learners (BLs)	
Naïve Bayes (NB)	Hagenauer and Helbich (2017)
Bayesian Network (BN)	Ma (2015)
Tree Augmented Naïve Bayes	Tang, Yang, and Zhang (2012)
7. Rule-Based Machine Learning (RBML)	
Fuzzy Inference System	Dell'Orco and Ottomanelli (2012)
Rough Sets Model	Cheng et al. (2014)
Class Association Rules	Lu and Kawamura (2010)
8. Hybrid methods (HM)	
Clustered Logistic Regression	Li et al. (2016)
Boosted logit	Biagioni et al. (2009)
Logit-ANN	Gazder and Ratrout (2015)
9. Miscellaneous (Msc)	
Multivariable Fractional Polynomials	Nam et al. (2017)
Discriminant Analysis	Karlaftis (2004)
Structural Equation Modelling	Papaioannou and Martinez (2015)
Linear regression	Ramanuj and Gundaliya (2013)

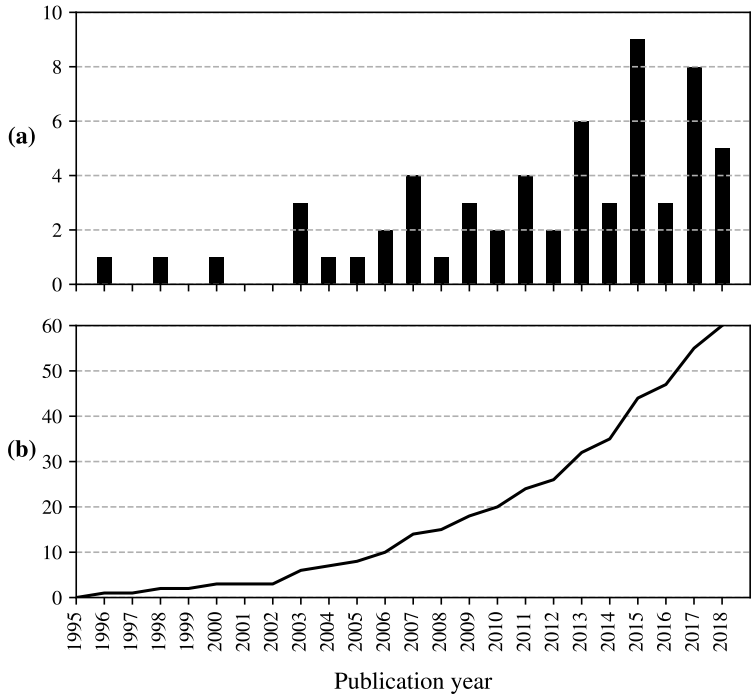

Figure 2: Publication distribution of articles in systematic review (a) per year and (b) cumulative.

in the model, and no utility functions or behavioural assumptions are specified. The distinction is not necessarily clear as to which approach is used in each study, due to the overlapping terms used to describe them. Many studies describe LR classification as Discrete Choice Models (DCMs) or RUMs, and both approaches are frequently referred to as logit models.

For the purpose of this review, a model is deemed to be a RUM only if it uses different utility specifications for the different modes in the model. This can be through including the relevant Level of Service (LOS) variables, e.g. expected journey duration, for each mode (see Q2e); by determining and removing irrelevant features through significance testing and behavioural constraints (e.g. correct sign of parameters); through testing multiple utility specifications to identify one which fits the data best; or a combination of these approaches. Any model where all features are included uniformly across all modes is deemed to be a LR classifier.

Of the 40 studies which use logit models, 19 make use of utility-based RUMs. This includes logit [S1; S3; S4; S5.1; S5.2; S6; S14; S24; S25; S28; S37; S42; S47; S51; S53; S57], NL [S5.1; S10; S16; S32; S53], and CNL models [S5.2; S32].

S25 additionally makes use of a clustered logit structure, where a decision tree is used to segment the population into three clusters on the basis of their socio-economic variables, and separate Multinomial Logit (MNL) models are trained for each cluster on the remaining variables. None of the other studies model input feature interactions - e.g. effect of car ownership on VoT.

Nine studies provide no details of the model structure [S8; S11; S13; S20; S21.2 S34; S49; S56; S59], so that it cannot be determined which approach is used.

Table 7: ML techniques used in each study in review.

No.	Log	ANN	DT	EL	SVM	BL	RBML	HM	Msc	No.	Log	ANN	DT	EL	SVM	BL	RBML	HM	Msc
S1	✓						✓			**S30**	✓					✓			
S2	✓	✓								**S31**	✓		✓		✓				✓
S3	✓		✓	✓	✓	✓				**S32**	✓	✓							
S4	✓	✓								**S33**	✓	✓			✓				
S5.1	✓	✓								**S34**	✓	✓	✓		✓	✓			✓
S5.2	✓	✓								**S35**									✓
S6	✓	✓								**S36**					✓				
S7	✓						✓			**S37**	✓		✓						
S8	✓						✓			**S38**	✓						✓		
S9		✓								**S39**		✓							
S10	✓			✓						**S40**		✓							✓
S11	✓						✓			**S41**				✓					
S12		✓								**S42**	✓						✓		
S13	✓	✓						✓		**S43**			✓	✓					
S14	✓	✓								**S44**					✓				
S15	✓	✓		✓	✓	✓				**S45**							✓		
S16	✓	✓								**S46**		✓	✓						
S17			✓	✓						**S47**	✓	✓							
S18	✓	✓								**S48**		✓				✓			
S19		✓			✓					**S49**	✓		✓						
S20	✓			✓	✓					**S50**			✓						
S21.1			✓							**S51**	✓			✓					
S21.2	✓	✓	✓						✓	**S52**							✓		
S21.3			✓							**S53**	✓	✓							
S22							✓			**S54**	✓	✓	✓						
S23							✓			**S55**		✓							
S24	✓	✓								**S56**	✓		✓						✓
S25	✓							✓		**S57**	✓	✓			✓				
S26	✓			✓						**S58**		✓							
S27	✓	✓	✓							**S59**	✓	✓							
S28							✓			**S60**	✓		✓			✓			
S29	✓					✓				**Sum**	40	30	16	9	10	7	11	2	6

The remaining 12 studies [S2; S7; S15; S18; S26; S27; S30; S29; S31; S33; S54; S60] use LR classification, and include all input variables uniformly for each mode, with no LOS (alternative-specific) variables.

4.1.1 Techniques - Limitations

Three general limitations are identified regarding the ML techniques used to investigate mode choice: (i) the limited number of studies which systematically compare several classification algorithms on the same task; (ii) the relatively low number of investigations into EL algorithms, in particular GBDT; and (iii) the inconsistent representation of RUMs in ML studies.

There are very few studies which systematically compare the performance of a range of classification techniques on the same task. Furthermore, the extensive differences in the datasets and methodologies used in each study make it impossible to make meaningful comparisons of model performances across individual studies. Table 7 shows that the vast majority of studies (52/63) use only one or two types of classification algorithm. Of these, 17 studies use only one type of classification algorithm. In total, only three studies [S3; S15; S34] have attempted to compare more than five types of classification algorithm. All of these studies suffer from various methodological limitations discussed in this review, as shown in table 13. As such, there is a need for a comparative study of classification techniques for mode choice prediction, using a rigorous methodology.

Q1a shows that EL algorithms are used in nine out of the 63 studies. In particular, GBDT models are only used in three studies [S15; S20; S51], despite GBDT models

consistently showing best in class performance in a number of similar tasks (see section 2.1). All of the studies which investigate GBDT show various methodological limitations (see table 13). As such, further investigation into the suitability of GBDT and other EL techniques is required.

Finally, Q1b highlights the inconsistent representation of RUMs in the studies in the review. Twenty-one studies either use uniform LR classification, or do not provide any information on the logit model structure. In the majority of these studies, these models are stated as representing RUMs when compared with other ML classification algorithms. The distinction is important as the utility function allows the modeller to add structural information based on solid behavioural foundations to the model. This structure assists the model in generalising to relationships between the input variables and mode choice, and can help prevent overfitting to training data, thereby improving model performance. Only one of the 40 studies which make use of logit models includes any interaction of input features. This is despite this frequently having a significant impact on modelling results and being common practice in RUM applications.

4.2 What is the nature of datasets used to investigate mode choice?

The following sections discuss the datasets used in the 63 studies in the review, focusing in turn on the nature of the dataset (trip diary/single-trip questionnaire/stated preference survey, etc); the unit of analysis (trip/tour/commute pattern/mobility); the size of the dataset; the dataset availability; the modes in the choice-set; the modelling of mode-alternatives; input features dependent on output choice; and hierarchical data.

Q2a: Nature of dataset

Table 8 shows the description and size of each dataset.

Only three studies [S1; S16; S32] use stated preference data. One study [S42] uses synthetic choice data, where the choice for a hypothetical metro service is synthesised based on a proposed fare structure and the respondent's willingness-to-pay (which is recorded during the interview).

The remaining 59 studies use revealed preference data, which can be largely grouped into two categories: trip diaries, or single-trip questionnaires.

Thirty-one studies make use of trip diary or activity-diary data, over periods ranging from one day to one year. These diaries are collected either from household surveys [S3; S7; S9; S14; S15; S17; S22; S24; S26; S27; S28; S36; S37; S38; S41; S46, S49; S51; S54; S60] or individual surveys [S29; S30; S31; S35; S44; S50]. Five studies which use trip diary data do not specify enough detail to determine if an individual or household survey is used [S12; S48; S53; S57; S59]. One study [S44] uses GPS tracking to log trips automatically, the rest use manually reported trip diaries.

In many studies, a subset of trips is selected from complete trip diaries, e.g. work trips only [S28; S30; S31; S39; S53; S54; S57], education trips only [S22], shopping/social trips only [S14; S60], outbound trips only [S7], trips from home only [S24], first trip of the day only [S17], or random sampling [S12; S48; S59].

Eighteen studies use individual single-trip questionnaires, where an individual is asked about a single trip they have made [S4; S5.1; S5.2; S8; S10; S13; S18; S21.1; S21.2;

Table 8: Nature and size of dataset used in each study in review.

No.	Type	N
S1	Stated preference - individual panel survey (160 people, 6 trips per person)	960
S2	Individual single-trip questionnaire (education commute)	597
S3	Trip diaries from household survey (1-2 day, 116 666 trips, 19 118 tours)	116 666
S4	Individual single-trip questionnaire (mixed purpose urban)	2350
S5.1	Individual single-trip questionnaire (student extra-urban trips)	1116
S5.2	Individual single-trip questionnaire (mixed purpose urban)	2350
S6	Unclear survey	1045
S7	Trip diaries from household survey (5721 outbound trips only, 4831 individuals, 1809 households)	5721
S8	Individual single-trip questionnaire (outbound home-work trip)	361
S9	Trip diaries from household survey (1 year, >100 mile, business trips only)	118 000
S10	Individual single-trip questionnaire (outbound home-school trip)	4700
S11	Unclear individual questionnaire	2868
S12	Trip diaries from unclear survey (650 trips sampled from larger survey, 130 for each mode)	650
S13	Individual single-trip questionnaire (cross-border)	516
S14	Trip diaries from household survey (1-2 day, outbound shopping trips only)	9450
S15	Trip diaries from household survey (6 day, 230 608 trips, 69 918 individuals)	230 608
S16	Stated preference - individual panel survey (3 trips per person)	801
S17	Trip diaries from household survey (1 day, only first trip, 24 807 individuals, 12 568 households)	24 807
S18	Individual single-trip questionnaire (mixed purpose urban)	620
S19	Unclear trip survey (4500 trips sampled from 17 539)	4500
S20	Individual single-trip questionnaire (work commute)	224
S21.1	Individual single-trip questionnaire (mixed-purpose)	210
S21.2	Individual single-trip questionnaire (mixed purpose urban)	7100
S21.3	Individual single-trip questionnaire (work commute)	617
S22	Trip diaries from household survey (education trips only)	409
S23	Individual single-trip questionnaire (work commute)	606
S24	Trip diaries from household survey (home-based trips, sampled to over-represent transit)	4764
S25	Individual single-trip questionnaire (Holiday travel)	731
S26	Trip diaries from household survey (mode choice analysed at household mobility level)	101 053
S27	Trip diaries from household survey (mode choice analysed at household mobility level)	18 733
S28	Trip diaries from household survey (1-day, morning-peak home-work trips only)	9210
S29	Trip diaries from individual survey (1-day, 11 993 trips made by 7235 people)	11 993
S30	Trip diaries from individual survey (1-day, commute patterns extracted)	5040
S31	Activity diaries from individual survey (commute patterns extracted)	1025
S32	Stated preference - individual panel survey	6768
S33	Commute patterns in household economic survey (9500 individuals, 3670 households)	3670
S34	Commute patterns in household economic survey (9500 individuals, 3670 households)	3673
S35	Trip diaries from individual survey (530 trips, <382 individuals)	530
S36	Trip diaries from household survey (Grouped into tours, 39 167 home-based tours, 24 396 individuals)	39 167
S37	Unclear household survey (1 day, mobility of household head only)	1216
S38	Trip diaries from household survey (Unknown trips, 5822 individuals, 2627 households)	?
S39	Unclear household survey (work trips only, 535 trips sampled randomly from 3500)	535
S40	Unclear household survey	1348
S41	Trip diaries from household survey (1 day, unknown trips, 1446 individuals)	?
S42	Individual single-trip questionnaire (synthetic choice)	229
S43	Unclear survey	5843
S44	Trip diaries from individual GPS survey (4 months, 17 040 trips, 292 individuals)	17 040
S45	Unclear household survey	4147
S46	Trip diaries from household survey (1-day)	100 000
S47	Individual single-trip questionnaire (access to rail on work trip)	4335
S48	Unclear daily travel survey (2000 trips sampled from larger survey, 500 for each mode)	2000
S49	Trip diaries from household survey (2-day, 72 536 trips, 31 000 individuals, 14 000 households)	72 536
S50	Activity diaries from individual survey (1 year, >2-day vacations only, 7121 vacations, 2791 individuals)	7121
S51	Trip diaries from household survey (1 day)	51 910
S52	Individual single-trip questionnaire (fixed O-D, mixed-purpose)	366
S53	Travel diaries from unknown survey (work travel mode choice)	4725
S54	Trip diaries from household survey (2-day, 4746 outbound work trips)	4746
S55	Unclear survey	1007
S56	Individual single-trip questionnaire (mixed-purpose)	498
S57	Trip diaries from unknown survey (outbound work trip only)	5029
S58	Individual single-trip questionnaire	100
S59	Trip diaries from unclear survey (500 trips sampled from larger survey, 125 for each mode)	500
S60	Trip diaries from household survey (1-day, home-based social activity)	5213

S21.3; S25; S47; S52; S56, S58] or a commute they make regularly [S2; S20; S23].

Two studies [S33; S34] make use of a household survey, in which each working member of the household details their work commute.

Eight studies [S6; S11; S19; S39; S40; S43; S45; S55] do not describe the data in enough detail to be able to determine the nature of the dataset.

The size of each dataset is also shown in table 8. Twenty studies use small datasets, with between 100-1000 entries. Twenty-nine studies use medium datasets, with between 1000-10 000 entries. Seven studies use large datasets, with between 10 000-100 000 entries. Five studies use datasets larger than 100 000 entries.

Two studies [S38, S41] do not give the exact size of the dataset. They both use the trip diaries of individuals in a household survey (5822 individuals in S38, 1446 individuals in S41).

Q2b: Unit of analysis

Fifty-seven of the studies use a single independent choice as the unit of analysis. The choice can be for a single one-way trip per respondent, a return trip (by assuming each leg is made by the same mode), trip diary data where sequences of trips are treated as independent, a regular commute, or a stated preference. Fifty-five of these studies model the mode choice only, whilst two studies [S17; S60] jointly consider other trip attributes (see Q2e).

Six studies use a different unit of analysis. Four studies analyse *mobility*. S26 and S27 both analyse household mobility by predicting the predominant mode used by a household across all trips made on the survey day. S37 analyses individual mobility, by predicting the predominant mode used by an individual across all trips they make on the survey day. Finally, S9 analyses the mobility within clusters. Clusters of similar trips are generated using k-means clustering (Hartigan and Wong 1979). The proportions of trips made by each mode within these clusters is then predicted.

Two studies use a tour-based approach. S3 uses the predicted mode choice of the first trip in a tour (the *anchor mode*) as an input feature for subsequent trips. S36 groups trips into home-based tours across eight categories and predicts overall mode choice for each tour (including mixed mode tours).

Note that (as discussed in section 4) S17 implements a tour-based analysis, but the subsequent trips in a tour are predicted on the basis of the attributes of the previous trip (including mode choice) as recorded in the dataset, and not as predicted by a model. As such, only the model which predicts attributes of the first trip of the day is analysed in the review (*Model 2-1* in the paper).

Q2c: Dataset availability

An attempt was made to identify and check the availability of the dataset used in each study. The following section discusses all datasets which were found to be openly available. Note that some studies which make use of open data may not have been identified, due to resource constraints when searching for datasets (see section 5).

Eighteen studies are identified as using open or partially open data. The majority use open household travel survey data. Two studies make use of academic datasets made

public by the authors: S21.1 makes use of the CLOGIT dataset, available with the Ecdat R library (Croissant 2016; Greene 2011); and S32 uses the SwissMetro dataset (Bierlaire, Axhausen, and Abay 2001; Bierlaire 2018). Four studies [S29; S30; S33; S34] make use of the partially open LISER PSELL data, which is available on registration (Luxembourg Institute of Socio-Economic Research 2018). Eleven studies use openly available household travel surveys:

- CMAP Travel Tracker Survey, 2007-2008 (Chicago Metropolitan Agency for Planning 2018b) - 3 studies [S3; S14; S24]
- CATS Household Travel Survey, 1990 (Chicago Metropolitan Agency for Planning 2018a) - [S28]
- San Francisco Bay Area Travel Survey, 2000 (Metropolitan Transportation Commission 2018b) - [S54]
- San Francisco Bay Area Travel Survey, 1990 (Metropolitan Transportation Commission 2018a) - [S57]
- Delaware Valley Household Travel Survey, 2012 (Delaware Valley Regional Planning Commission 2018) - [S51]
- National Household Travel Survey, 2009 (Federal Highway Administration 2018) - [S36]
- American Travel Survey, 1995 (Bureau of Transportation Statistics 2018) - [S9]
- Sydney Household Travel Survey (Transport for NSW 2018) - [S46]
- Victorian Integrated Survey of Travel and Activity, 2007-2008 (Transport for Victoria 2018) - [S17]

Only one study [S15] is identified as making the fully processed data openly available, in the format used for modelling within the paper.

Q2d: Modes in choice-set

Figure 3 shows a frequency plot of the number of modes considered in each study, which ranges from two to nine. The most common number of modes considered is four, which is used in 18/63 studies.

Five papers have a different number of classes modelled in the classification problem from the number of modes considered. Three papers perform only one-vs-one or one-vs-rest modelling. S9 and S31 both consider three modes, but in both studies the modelling is performed one-vs-rest across the three modes, so that each model considers two different classes. Unlike other studies which use one-vs-rest modelling, the individual models are not combined to create a multiclass classifier in either study. Similarly, S49 considers four modes, but the modelling is performed one-vs-one. As with S9 and S31, the individual models are not combined to create a single multiclass classifier.

Two models jointly model other variables alongside mode choice. S60 jointly considers four modes across two different time-periods (peak/off-peak), therefore modelling a total of eight classes. Similarly, S17 jointly models three modes, three trip purposes, three departure periods, and four distance categories, for a total of 108 classes, 102 of which are observed in the data.

A total of eleven studies use only binary classification. This includes the eight studies which model only two modes [S2; S6; S11; S13; S25; S27; S35; S42] and the three studies

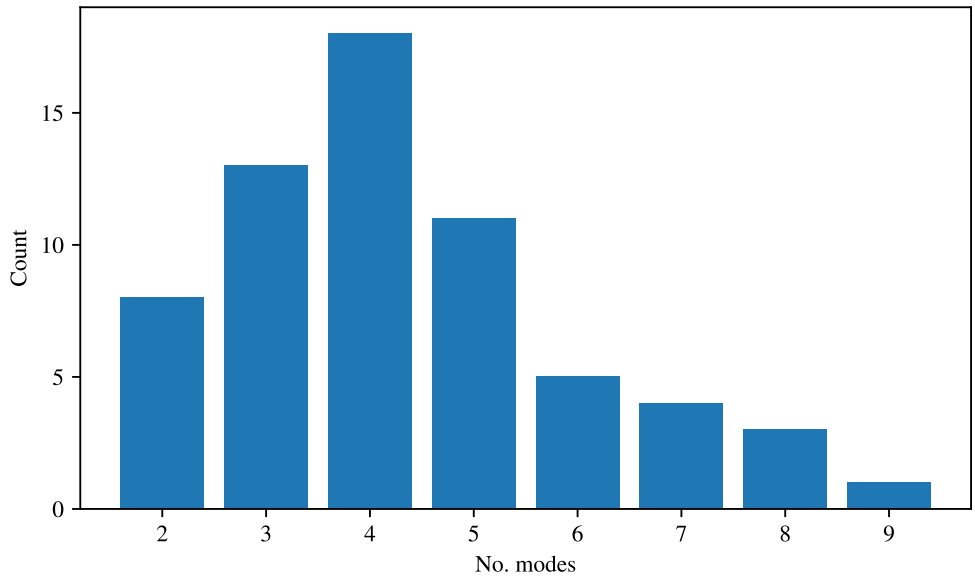

Figure 3: Frequency bar chart of number of modes considered in each study in review.

which use one-vs-rest/one-vs-one modelling without combining individual models to a single classifier [S9; S31; S49].

Figure 4 shows the frequency of each mode/grouping of modes considered in each study. The *car* mode is the most commonly modelled, appearing in 45 studies, followed by *walk* (28 studies) and *public transport* (27 studies). Certain modes either appear individually or grouped. For example, cycling is treated as an independent mode in 21 studies and grouped with walking in nine studies. The grouping of public transport modes cannot be immediately understood from fig. 4, due to different combinations of groupings being possible. For example, for many studies, rail services are not a viable mode of transport, and so *bus* is the only mode considered. Twenty-seven studies consider all public transport modes under one combined *public transport* mode. Of the 25 studies which consider the independent *bus* mode, 14 include *bus* as the only public transport mode. A total of 15 studies consider two or more separate public transport modes.

Q2e: Modelling of mode-alternatives

In order to understand the impact that the transport network has on mode choice, it is necessary for the dataset to include attributes of the mode-alternatives, e.g. the expected duration and cost of travelling by each mode in the choice-set. These are commonly referred to as Level of Service (LOS) attributes in the literature. For revealed preference data, typically only details of the choice made by the passenger are recorded. As such, details of the mode-alternatives need to be synthesised and added to the dataset to be included in the modelling.

Of the 59 studies which use revealed preference data, 27 include no attributes of the mode-alternatives in the choice-set [S2; S7; S12; S13; S15; S17; S18; S19; S20; S22; S25; S26; S27; S29; S36; S37; S40, S41; S44; S46; S48; S50; S52; S53; S54; S56; S59]. A further two studies do not list the input features used in the model with enough clarity

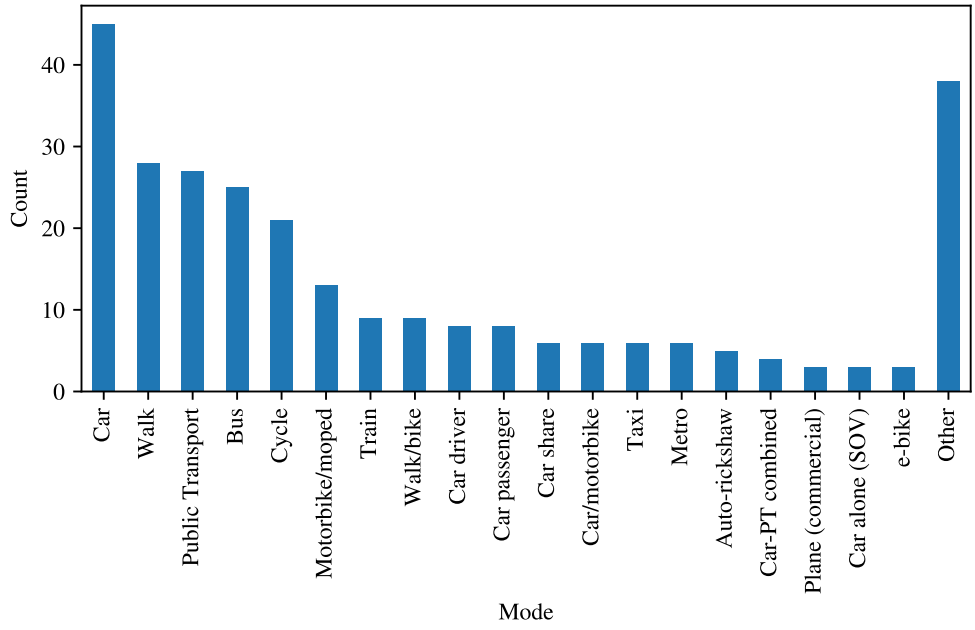

Figure 4: Frequency bar chart of individual modes/grouping of modes in each study in review. The 'Other' category groups all modes/combinations of modes with less than three occurrences across all studies.

to deduce whether any attributes of the mode-alternatives are included [S43; S55].

Table 9 shows the relevant features used in the 30 papers which include attributes of the mode-alternatives. The definition of each term is given below:

- *Duration* - journey time from start-point to end-point (including access, transfers etc.)
- *Cost* - Out of pocket cost (e.g. transport fares, Vehicle Operating Costs)
- *Generalised costs* - Combined duration and cost as a single value of disutility, expressed in the unit of currency
- *Vehicle Operating Cost (VOC)* - the mileage dependent costs of operating a vehicle (e.g. fuel, tires, maintenance, repairs, depreciation)
- *In-Vehicle Travel Time (IVTT)* - the duration spent in vehicle/ on-board public transport services
- *Out-of-Vehicle Travel Time (OVTT)* - the combined access, egress, transfer, and waiting durations for Public Transport (PT)
- *Access* - The walking duration/distance between the start-point and first public transport access stop
- *Egress* - The walking duration/distance between the last public transport stop and the end-point

Fourteen of the 30 studies which model mode-alternatives do not state the methods used to calculate these attributes [S4; S6; S8; S11; S21.2; S21.3; S23; S24; S31; S39; S45; S47; S57; S58]. Fourteen studies use zonal, time-independent (static) transport models to calculate durations and/or costs [S3; S5.1; S5.2; S9; S10; S21.1; S28; S30; S33; S34; S35; S39; S49; S60]. One study [S5.2] additionally makes use of a time-dependent

Table 9: Attributes of mode-alternatives in selected studies in review. Unless stated otherwise, each attribute is a duration. *PT*=Public Transport, *IVTT*=In-Vehicle Travel Time, *OVTT*=Out-of-Vehicle Travel Time, *VOC*=Vehicle Operating Cost

No.	Duration	Cost	Other	No.	Duration	Cost	Other
S3	✓	✓		**S28**			Duration, VOC (Drive), IVTT (train)
S4	✓	✓	Access (Bus)	**S30**	✓	✓	
S5.1	✓	✓	Transfer, access/egress (PT)	**S31**			Duration ratios (Each mode)
S5.2	✓	✓		**S33**			Generalised costs (Each mode)
S6		✓	Access, egress, IVTT (PT)	**S34**			Generalised costs (Each mode)
S8			Generalised costs (Each mode)	**S35**			IVTT, transfer, speed, directness (PT)
S9	✓	✓		**S38**		✓	IVTT and OVTT (Each mode)
S10	✓		Access distance (PT)	**S39**			IVTT and route distance (Each mode)
S11	✓	✓		**S45**			OVTT (Bus)
S14	✓		Access & egress distance (PT)	**S47**	✓	✓	
S21.1	✓	✓	IVTT (PT)	**S49**	✓		
S21.2	✓	✓		**S51**	✓		
S21.3	✓	✓	IVTT (PT)	**S57**	✓	✓	
S23	✓	✓		**S58**	✓	✓	
S24	✓	✓		**S60**	✓	✓	

public transport model to calculate transfer and combined access/egress durations for the PT route at the time of departure. Finally, two studies [S14; S51] make use of an online directions service to generate trip durations.

Q2f: Input features dependent on output choice

In order to be used as a valid predictive model, model input features must be independent of the output choice. Features which are dependent on the choice, e.g. the recorded trip duration (which is dependent on the mode taken) cannot be known until the trip is made, and so cannot be used for prediction.

A significant number of studies (17/63) include input features which are related to the output choice, either directly or indirectly.

Eight studies [S2; S25; S26; S29; S40; S49; S54; S56] include the recorded travel duration of the selected mode as an input feature. Four of these studies also include the trip distance [S26; S29; S40; S49], which would allow the classifier to infer the speed of the mode-selected. A further two studies [S40; S54] additionally include the reported cost of the selected mode.

Two studies [S3; S46] implicitly include the reported duration by including both the reported departure time and arrival time in the feature vector.

Three studies [S21.1; S21.2; S21.3] implicitly include the selected mode in the input feature vector by labelling attributes of the *selected* mode and best *alternative* mode. For example, one node in the DT for S21.2 separates trips between those made by *Auto* and those made by *Metro* on the basis of whether the cost of the selected mode is greater than or equal to 1.6 euro.

Two studies use different definitions of duration in the mode-alternative attributes for the selected mode. [S30] uses the reported duration as the driving duration if the trip is made by car and uses the driving time predicted by a static zonal transport model otherwise. [S51] similarly uses the reported duration for the selected mode, and the duration as predicted by the Google Directions Application Programming Interface (API) for all other modes. In both cases, this may cause leakage of the selected mode into the input

feature vector.

Finally, two studies [S13; S37] include survey questions on reasons for not taking a particular mode in the input feature vector.

As with the modelling of mode-alternatives, two studies [S43; S55] provide insufficient detail of the modelling process to determine whether input features are included which are dependent on the output choice.

Q2g: Hierarchical data

As shown by Q2a, 31 studies make use of trip diary data. Household trip survey data has an inherent hierarchical structure: households are made up of multiple people, each of whom make multiple tours, in which there are multiple legs or trips. Elements within the same groups in the hierarchical structure may show interdependency. This hierarchical structure arises from the specific nature of how trip diary data is collected, and introduces strong correlations which can be observed in the data. Formally, three levels of hierarchy can be considered (each with examples of how the structure could cause interdependency):

- *Household-Person (H-P)* - e.g. multiple members of a household travelling together therefore all travelling by the same mode, one person using the only vehicle in a household meaning that others cannot use that vehicle, all members of a household sharing a tendency to/not to travel by a particular mode, etc.

- *Person-Tour (P-T)* - e.g. individual showing a tendency to/not to travel by a particular mode, individual not being able drive/cycle for all tours due to a vehicle/bike not being available to them on the survey date, individual having a season ticket and therefore being more likely to travel by public transport, etc.

- *Tour-Trip (T-T)* - e.g. return trip being highly likely to be made by the same mode as the outbound trip, vehicle/bike not being available for onwards travel as it was not used for first leg (trip) in tour, vehicle/bike needing to be used for onwards travel as it was used for first leg (trip) in tour and cannot be left behind, etc.

Individual survey trip diaries do not have a household-person grouping, leaving person-tour and tour-trip groupings.

Many of the studies which make use of trip diary data sample the data in a way which removes all/part of the hierarchical structure, e.g. by sampling only outbound trips (removes tour-trip hierarchy), or by sampling only trips made by one member of a household (removes household-person hierarchy). This sampling is presented in table 10.

Additionally, the commute patterns analysed in S33 and S34 are taken from a household survey, with multiple members in each household. As such, these datasets contain a household-person hierarchical structure.

Finally, there may be hierarchical structure in the studies with datasets of unknown nature [S6; S11; S19; S39; S40; S43; S45; S55].

Table 10 shows the levels present in the input dataset (after any sampling/processing) for all studies which make use of hierarchical data.

Whilst S3 uses a tour-based analysis, it still predicts mode choice for individual trips, and so the Tour-Trip hierarchy in the data is still present. In total, there are 35 studies which use hierarchical data, or data which may be hierarchical, after sampling/processing.

Table 10: Details of hierarchies in datasets in relevant studies in review, after sampling/ processing. *H-P*=Household-Person, *P-T*=Person-Tour, *T-T*=Tour-Trip.

No.	H-P	P-T	T-T	Sampling	No.	H-P	P-T	T-T	Sampling
S3	✓	✓	✓	None (complete trip diary, household)	**S36**	✓	✓		Tours from household trip diary
S6	?	?	?	Unclear data	**S37**				Mobility of head of household only
S7	✓	✓		Outbound trips only	**S38**	✓	✓	✓	None (complete trip diary, household)
S9				Mobility of similar clusters	**S39**	?			Work trips only, sampled from larger survey
S11	?	?	?	Unclear data	**S40**	?	?	?	Unclear data
S12	?	?	?	Random sampling from larger survey	**S41**	✓	✓	✓	None (complete trip diary, household)
S14	✓	?		Outbound shopping trips only	**S43**	?	?	?	Unclear data
S15	✓	✓	✓	None (complete trip diary, household)	**S44**		✓	✓	None (complete trip diary, individual)
S17	✓			First trip in day only	**S45**	?	?	?	Unclear data
S19	?	?	?	Unclear data	**S46**	✓	✓	✓	None (complete trip diary, household)
S22	?	?	?	Education trips only	**S48**	?	?	?	Random sampling from trip diaries
S24	✓	✓		Home-based trips only	**S49**	✓	✓	✓	None (complete trip diary, household)
S26				Household mobility only	**S50**		✓		None (complete activity diary, individual)
S27				Household mobility only	**S51**	✓	✓	✓	None (complete trip diary, household)
S28	✓			Morning home-work trips only	**S53**	?			Outbound work trips only
S29		✓	✓	None (complete trip diary, individual)	**S54**	?	?		Outbound work trips only (2-day)
S30				Commute patterns from individual survey	**S55**	?	?	?	Unclear data
S31				Commute patterns from individual survey	**S57**	?			Outbound work trips only
S33	✓			None (Household survey)	**S59**	?	?	?	Random sampling from larger survey
S34	✓			None (Household survey)	**S60**	✓	?		Home-based social trips only
S35		✓	✓	None (complete trip diary, individual)					

This includes 10 studies which use complete, unsampled trip diaries [S3; S15; S29; S35; S38; S41; S44; S46; S49; S51; S60].

4.2.1 Model datasets - limitations

Six limitations are identified in relation to the datasets used to investigate mode choice. Two limitations are technical: (i) studies not including any attributes of the mode-alternatives, (ii) studies using input features dependent on output choice; and four limitations are general: (i) not describing the dataset and modelling process in sufficient detail, (ii) the shortage of studies using large datasets to investigate mode choice, (iii) the lack of relevant, openly available datasets including mode-alternative attributes, (iv) not considering sampling of the data from the population.

Note that using hierarchical data is not an issue in itself, as long as appropriate sampling is used for validation. This is therefore discussed in section 4.3.

Two technical limitations are identified related to datasets. Q2f identifies 27 studies which include no LOS attributes of the mode-alternatives in the choice-set, and a further two studies which do not list the input features used in the model with enough detail to be able to determine whether any attributes of the mode-alternatives are included. In order to model the impact that the transport network has on mode choice, it is necessary for the feature vector to contain attributes of the mode-alternatives, e.g. the expected duration and cost of travelling by each mode in the choice set. As significant correlations between attributes of each mode-alternative and mode choice are likely to exist, not including these variables in the feature vector will result in models with lower predictive performance. Additionally, for statistical RUM models, omitting relevant predictors (features) in the input results in endogenous errors in the parameters of the remaining variables (Train 2009, Chapter 13). This can cause biased, inconsistent estimates of these parameters, which may be important for explaining behaviour (e.g. VoT). Finally, when using the choice model for simulation of future trips under unknown conditions (e.g. in an Agent Based Model (ABM)), the impacts of changes to the transport network on mode choice

cannot be modelled if attributes of the mode-alternatives are not included in the feature vector. These studies therefore do not allow for modelling the impacts of changes to the transport network on the mode choice decisions made by an individual.

Of the studies which do model mode-alternatives, the majority generate LOS variables from static zonal graphs. This means that they do not capture the highly granular spatial and temporal variability of conditions on a transport network.

Q2g identifies 17 studies which include input features which are related to the output choice. These features cannot be known in advance of a trip being made, and so this prevents these models from being used in a predictive context. Additionally, input features which are directly and explicitly dependent on class membership, e.g. travel speed being dependent on travel mode, may be highly correlated with the class membership. As such, this will result in better *apparent performance* of the model than could be achieved using only valid independent variables (i.e. the performance of the model will be overestimated through *data leakage*). As with omitting the mode-alternative LOS variables, including input variables which are dependent on the output in a statistical model (RUM) can introduce endogeneity through reverse causality (Train 2009, Chapter 13). This can also cause biased, inconsistent estimates of model parameters. Again, a further two studies provide insufficient detail of the modelling process to be able to determine whether any input features which are dependent on the output choice are included.

Of the two technical limitations related to datasets, using input features dependent on output choice a *pitfall* that is likely result in incorrect conclusions being drawn from the modelling results. Conversely, not including any attributes of the mode-alternatives is an *area for improvement*, as doing so is likely to improve the performance of the model.

The discussion of the research question also highlights four general limitations. Firstly, the vast majority of studies (49/63) analysed make use of small to medium datasets, with less than 10 000 entries. Only 12 studies make use of datasets with more than 10 000 entries, and as such there is a need for further investigation into problems of this scale.

Secondly, multiple studies do not describe the dataset and modelling process in sufficient detail for the required information for the systematic review to be extracted. This is problematic for repeatability of the mode choice experiments implemented in these studies, particularly when there is such large variation in the methodologies used in each study. In order to ensure repeatability of the results, methodologies should be recorded in detail, and where possible, data and code should be made available.

There is also a need for a relevant, openly available dataset including mode-alternative attributes. There exist several openly available, large datasets for investigating passenger mode choice. Of the 12 studies which use datasets with greater than 10 000 entries, eight make use of openly available datasets [S3; S9; S15; S17; S29; S36; S46; S52]. However, only two of these studies [S3; S9] add mode-alternative information to these datasets, and the processed dataset is not openly available for either study. As mentioned, only the processed dataset for S15 is openly available, and this dataset does not include any mode-alternative attributes.

Finally, no papers checked the representivity of the sample in the dataset with respect to the target population, or discussed how to correct for sampling biases in forecasting. When using a model for forecasting, it is essential to consider the bias in the sample, for example for accurate predictions of market shares.

4.3 How is model performance determined?

The following sections discuss the techniques used to determine model performance in the 63 studies in the review, focusing in turn on the validation method, the sampling method, and the performance metrics used.

Q3a: Validation method

The validation method most commonly used in the studies is holdout validation (non-repeated), which is used in 43 studies. Train-test splits range from 23:77 to 91:9, but the most commonly used splits are 70:30 (10 studies), and 80:20 (nine studies).

Seven studies use repeated holdout validation: S2 runs 3 repetitions of a 75:25 split, S13 runs 10 repetitions of a 75:25 split, S32 runs 10 repetitions of a 70:30 split, S48 runs 50 repetitions of a 70:30 split, S51 runs 100 repetitions of a 75:25 split, and finally S33 and S34 run 100 repetitions of a 60:40 split. Confusingly, S48 only shows the results for both the train and validation data combined, averaged over the 50 runs.

k-fold cross-validation is used in six studies. Four studies use 10-fold cross-validation [S15; S17; S24; S60], and two studies use 5-fold cross-validation [S30; S36]. As well as 10-fold cross-validation, S24 also performs holdout validation (60:40 split).

Three papers use different validation techniques for different models. Whilst they all use in-sample validation for the logit models, S1 uses 80:20 holdout validation for the neuro-fuzzy multinomial logit model; S21.2 uses 60:40 holdout validation for the ANN, DT, and discriminant analysis models; and S26 uses Out-Of-Bootstrap (OOB) error for the Random Forest (RF) model.

Four studies use in-sample validation for all models [S25; S50; S52; S56].

Finally, three studies do not state the validation method used [S3; S20; S35].

Q3b: Sampling method

Of the 35 studies which use hierarchical data, or data which may be hierarchical, none mention the use of grouped (by household or individual) sampling. This includes all 10 studies which make use of complete, unsampled trip diaries. Furthermore, two studies which make use of trip diaries [S3; S35] do not state which validation technique is used at all (see Q3a).

All studies which perform out-of-sample validation appear to use random sampling (either stated explicitly or assumed). Only one study [S16] tests models on data collected separately from, or after, the training data (*external validation*). Each city-specific model (Melbourne and Sydney) is additionally validated on the data from the other city, as well as the holdout test sample.

Q3c: Performance metrics

Table 11 shows the performance metrics used for validation in each study. Performance metrics used only during model estimation/fitting are not recorded, except for studies which use in-sample validation.

Four different discrete classification metrics are shown in the table: accuracy, recall, the confusion matrix, and the predicted mode shares. To calculate these metrics, each trip

is assigned to the mode with highest predicted probability. Precision and specificity are also used by S3 and S31 respectively, but not recorded in the table.

Metrics which evaluate probabilistic classification are grouped together in table 11. Seven different probabilistic metrics are used: percent clearly right/wrong/unclear [S4; S5.1; S5.2], Arithmetic Mean Probability of Correct Assignment (AMPCA) (referred to as fitting factor in S4, S5.1, and S5.2; and average probability of correct assessment in S33), Mean Squared Error (MSE) [S4; S5.1; S5.2; S43], simulated mode shares [S4; S5.1; S5.2; S32], Receiver Operating Characteristic (ROC) curves [S18; S19; S60], log-likelihood [S30; S32], Bayesian Information Criterion (BIC) [S30], and Expected Simulation Recall (ESR) [S33].

Table 11 does not show the metrics used in two studies. S9 performs regression on the total number of trips performed by each mode within a cluster, and so uses regression-based metrics (MSE and average relative variance of regression). S16 uses three metrics: *predicted share less observed share*, *weighted percent correct*, and *weighted success index*. However, no definitions for the performance metrics are provided in the paper, and so it cannot be determined if the metrics are discrete or probabilistic.

In total, 52 of the remaining 61 studies use only discrete classification metrics, whilst nine studies use a combination of probabilistic and discrete classification metrics. Of the studies which use only discrete classification metrics, 29 make use of LR models.

Twelve studies use only one performance metric: accuracy is used as the sole metric in 10 studies [S2; S6; S8; S17; S20; S26; S35; S41; S44; S45], recall per mode in S25, and the confusion matrix in S47.

4.3.1 Model performance estimation - limitations

Four technical limitations are identified in relation the model performance estimation techniques used in the studies: (i) studies using inappropriate validation schemes, (ii) studies using incorrect sampling methods for hierarchical data, (iii) studies not performing external validation, (iv) studies using only discrete metrics.

Q3a identifies 10 studies which make use of inappropriate validation schemes. This includes four studies which use in-sample validation [S25; S50; S52; S56], three studies which use different validation techniques for different models being tested [S1; S21.2; S26], and three which do not state the validation method being used [S3; S20; S35].

In-sample validation uses the same data to fit and validate the model and can be interpreted as using the *train-error* to estimate model performance. As such, it presents only the explanatory power of the model, i.e. the ability of the model to replicate the training data, and not the predictive performance. This is discussed in detail by Shmueli (2010). If a model has high *variance*, it can overfit to noise in the data during model fitting, without generalising to valid correlations between the input and output. This will result in in-sample validation overestimating the predictive performance. Without testing the model on out-of-sample data, there is no way to assess whether overfitting has occurred. Additionally, due to the nature of the *bias-variance tradeoff* (Hastie, Friedman, and Tibshirani 2008, Chapter 2), a classifier will tend to fit partially to noise in the data, even if it does not overfit. As such, the train-error will tend to overestimate predictive performance, even for well specified models which do not overfit.

Formal validation of a model on data separate training data is essential to ensure mod-

Table 11: Summary of performance metrics used for validation in each study in review. **Acc**=Accuracy, **Rec**=Recall, **CM**=Confusion Matrix, **MS**=Mode Shares (Discrete), **Pro**=Probabilistic metric.

No.	Acc	Rec	CM	MS	Pro	No.	Acc	Rec	CM	MS	Pro
S1	✓		✓			**S30**	✓		✓		✓
S2	✓					**S31**	✓	✓		✓	
S3	✓	✓				**S32**	✓			✓	✓
S4	✓	✓			✓	**S33**			✓		✓
S5.1	✓	✓			✓	**S34**	✓	✓	✓		
S5.2	✓	✓			✓	**S35**	✓				
S6	✓					**S36**		✓	✓		
S7	✓			✓		**S37**	✓	✓			
S8	✓					**S38**	✓	✓		✓	
S9	-	-	-	-	-	**S39**	✓	✓			
S10	✓	✓				**S40**	✓	✓	✓		
S11	✓		✓			**S41**	✓				
S12	✓	✓	✓			**S42**	✓		✓		
S13	✓	✓				**S43**	✓				✓
S14	✓	✓				**S44**	✓				
S15	✓	✓	✓			**S45**	✓				
S16	-	-	-	-	-	**S46**	✓			✓	
S17	✓					**S47**			✓		
S18	✓		✓		✓	**S48**	✓	✓	✓		
S19	✓	✓	✓		✓	**S49**	✓		✓		
S20	✓					**S50**	✓	✓		✓	
S21.1		✓	✓			**S51**	✓	✓		✓	
S21.2	✓	✓	✓			**S52**	✓		✓		
S21.3		✓	✓			**S53**	✓		✓		
S22	✓		✓	✓		**S54**	✓	✓	✓	✓	
S23	✓		✓	✓		**S55**	✓	✓	✓		
S24	✓		✓			**S56**	✓	✓			
S25		✓				**S57**	✓	✓	✓		
S26	✓					**S58**		✓		✓	
S27	✓	✓	✓			**S59**	✓	✓		✓	
S28	✓	✓	✓			**S60**	✓	✓	✓		
S29	✓	✓				**Sum**	54	34	29	12	9

els have generalised to the training data without overfitting. As such, in-sample validation is an inappropriate validation scheme. Furthermore, in order to make valid comparisons between performance estimates of different models, the same validation method must be used for all models. Otherwise, any apparent differences in performance may be due to differences in the respective validation schemes.

Q2g identifies 20 studies which make use of hierarchical data, and a further 15 which makes use of data which may be hierarchical. As identified by Q3b, none of these studies sample validation sets or folds grouped by individual or household. As such, trips from the same group (household/person/tour) will occur in both test and training data. These trips inherit correlated features from these groupings, which can allow for data-leakage and overfitting.

There are valid hierarchies in datasets which can be relevant to the modelling scenario. For example, a modeller would be interested if students (socio-economic group) show a tendency towards cycling (correlation), or if trips made at the weekend (temporal grouping) were less likely to be made by public transport (correlation). In both these cases, the hierarchies (groups) are general, and described by information in the feature vector. As such, these correlations are likely to be constant across the training data and future unknown trips, and so are relevant to the modelling scenario.

Conversely, the hierarchies identified by Q2g are not representative of the population (and instead are a feature of our data sampling). As such, these hierarchies are not relevant to the modelling scenario, and will boost the *apparent performance* of the model, whilst in reality causing it to perform worse on true unseen data.

This is particularly problematic for the 10 studies which use complete trip diary data [S3; S15; S29; S35; S38; S41; S44; S46; S49; S51; S60]. Many of these trip diaries are multi-day, compounding the issue. Notably, S15 uses a six-day trip diary (average 3.3 trips per person), and S44 uses sets of GPS trips logged over four months (average 58.4 trips per person). This problem is not unique to mode choice modelling applications. Saeb et al. (2017) conduct a review of sampling methods in studies using ML to make clinical predictions from smartphone or wearable technology data. They review studies which use hierarchical data, where there are multiple *records* for each individual *subject*. They find that of the 62 of the studies included in the meta-analysis, 28 (45 %) use inappropriate *record-wise* sampling, instead of *subject-wise* sampling.

Q3b identifies that only one of the studies reviewed [S16] uses *external validation*, where the model is validated on data collected separately from, or after, the training data. External validation using future data is the only possible method of directly simulating the use case for a mode choice model, of predicting future, unknown trips. External validation can also identify issues with data-leakage, overfitting, and incorrect validation schemes, e.g. the incorrect sampling methods for hierarchical data, as highlighted by Q2g.

Finally, Q3c identifies that the vast majority of studies (51/63) use purely discrete metrics to assess model performance, where each trip is assigned to the mode with the highest predicted probability. This includes 29 studies which assess LR using only discrete classification metrics, despite LR being a statistical technique intended to generate probability distributions. In total, only six studies make use of *proper* continuous scoring metrics, log-likelihood [S30; S32] and MSE [S4; S5.1; S5.2; S43].

There are a number of issues with the use of discrete metrics for choice prediction. Firstly, discretising the classification by assigning each observation to the highest prob-

ability class is likely to result in non-representative mode-shares. Mode choice data is inherently imbalanced, i.e. there are likely more trips made by some modes (e.g. car, walking) than others (e.g. cycling). By assigning each prediction to the class with highest probability, the less frequent classes will be under-represented in the predicted outcomes, and the more frequent classes will be over-represented. For example, consider a biased random coin flip, where heads is 60 % likely to occur, and tails occurs with 40 % probability. The best possible predictive classification model will predict these outcomes at their respective probabilities for each coin flip event. However, assigning the highest probability class for the prediction will result in heads always being predicted (and never tails) as heads is always more likely than tails. This clearly results in non-representative class shares. Non-representative mode-shares are unacceptable for mode choice models, where the mode-shares are a crucial model output.

Similarly, by generating a discrete class for each observation, mode choice is treated as a deterministic instead of a stochastic process. As such, it is assumed that mode choice is constant under the same set of conditions and socio-economic characteristics. In reality, passengers have a degree of intra-heterogeneity, and their choice can be considered as being drawn randomly from a probability distribution (as with the coin-flip example). We define this distribution as the *true model*, which we aim to replicate with the classification model. In order to account for this stochastic heterogeneity in simulation, the predicted mode choice should be drawn from a probability distribution. The metric used to assess model performance should therefore represent how well the predicted probability distributions fit the data.

Additionally, discrete metrics do not assess how right or wrong model predictions are. For example, when using discrete metrics, the contribution to the model's score for a trip where a binary classifier predicts the selected mode at 1 % probability is the same as that for a trip where the classifier predicts the selected mode at 49 % probability. Analysing the probability distributions presents information on where the model performs well or poorly.

Finally, by taking the maximum of the class probabilities, discrete predictions and the associated metrics are discontinuous. This results in discrete metrics having an expected score which is not differentiable or strictly convex. Additionally, accuracy and other discrete metrics are not *strictly proper* scoring rules, and as such do not have unique maximums (Gneiting and Raftery 2007). This makes discrete metrics poor metrics to use during model fitting, particularly where a continuous gradient is required (e.g. gradient descent).

Of the four technical limitations related to model performance estimation, three represent pitfalls (using inappropriate validation schemes, using incorrect sampling for hierarchical data, and using only discrete metrics), and one represents an area for improvement (not performing external validation).

4.4 How are optimal model hyper-parameters selected?

Hyper-parameters are parameters of classification algorithms which are used to regularise the model during model training. The selected hyper-parameters impact the bias and variance of the fitted model. This section discusses the techniques used to optimise model specifications and hyper-parameters for conventional ML algorithms (ANNs, DTs, EL,

SVMs). The 14 studies which do not use at least one these algorithms are therefore omitted from this section of the review [S1; S7; S8; S11; S22; S23; S28; S29; S30; S35; S38; S42; S45; S52].

The following sections review the remaining 49 studies, focusing in turn on the hyper-parameter search method, the hyper-parameter validation method, and the hyper-parameter validation data.

Q4a: Hyper-parameter search method

Of the 49 studies which use at least one conventional ML algorithm, 11 do not mention hyper-parameter values at all within the paper [S12; S20; S21.1; S21.2; S21.3; S25; S26; S41; S43; S46; S56]. A further nine studies either state hyper-parameter values without explanation [S10; S17; S31; S36; S37; S47; S60], or state that they use default values [S27; S59].

This leaves 29 studies which use some form of hyper-parameter optimisation. Twelve studies [S2; S3; S4; S9; S13; S14; S16; S24; S39; S50; S51; S53] perform a manual search, or trial and error, in order to identify model parameters. Of these, S3 searches only for the kernel function in a SVM and uses default values for all other parameters and models, and S39 searches for the number of neurons in a single test layer, again using default values for other parameters.

Nine studies [S6; S18; S33; S34; S40; S48; S49; S55; S57; S58] specify a MLP with a single hidden layer and perform a linear search on the number of neurons in that layer. With the exception of S57, which performs a grid search for the SVM parameters (γ and C), default values are used for all other parameters of all models.

One study [S49] uses a repeated linear search, firstly on the loss-weight ratio of the two classes in each model, and secondly on the number of features used.

Four studies [S5.1; S5.2; S15; S32; S54] make use of a grid search. S32 tests a range of specific ANN structures, with different numbers of hidden layers and nodes, but uses a grid search to select the appropriate dropout ratio and learning rate.

One study [S19], tests two different search strategies in order to find optimal SVM parameters (γ and C): grid search and genetic algorithms. The study finds that whilst the two methods find optimal solutions with similar accuracies, the genetic algorithm finds the solution with the lower penalty parameter (C), and so is preferred.

Finally, one study [S44] states that cross-validation is used to select model parameters but does not state the search method used.

Q4b: Hyper-parameter validation method

Of the 29 studies which use some form of hyper-parameter optimisation, 10 do not state the validation method used to determine optimal values [S2; S3; S9; S13; S24; S39; S40; S48; S55; S58]. S3 states the parameters with best performance are used but does not state how this is determined.

Eight studies [4; 14; 16; 18; 32; 34; 54; 57] use holdout validation. Seven studies [15; 19; 44; 49; 50; 51; 53] use k-fold cross-validation. One study [S33] uses repeated holdout validation. One study [S6] uses in-sample validation.

Finally, two studies [S5.1, S5.2] use a complex multi-criteria assessment, involving relative performance on both the calibration and validation data.

Q4c: Hyper-parameter validation data

Of the 29 studies which use some form of hyper-parameter optimisation, 14 do not state the data used for hyper-parameter validation [S2; S3; S9; S13; S16; S19; S24; S34; S39; S40; S44; S48; S55; S58].

Of the seven studies which use k-fold cross-validation to test hyper-parameter performance, two use only the training data [S53; S57], one uses a random subset of 43 % of the data [S15], two use all of the data [S50; S51], and two do not state the data used (included above). The study which uses repeated validation also uses all of the data [S33].

Of the eight studies which use holdout validation, three use the data reserved for model testing [S4; S14; S32], two use only the train data, dividing it into a new test and train fold [S53; S54], one uses a separate 15 % validation sample which is not used for model testing or training [S18], and two do not state the data used (included above).

Finally, the two studies which use the multi-criteria assessment use both the train and test data.

4.4.1 Model optimisation - limitations

Four limitations are identified in relation the model optimisation techniques used in the studies. Three limitations are technical: (i) studies not performing any type of hyper-parameter optimisation, (ii) studies not using rigorous hyper-parameter search schemes, (iii) studies optimising hyper-parameters on validation data; and one is general: not presenting model hyper-parameters used within the study.

Three technical limitations are identified by the attributes related to hyper-parameter optimisation collected in the systematic review. Of the 49 studies which use one or more conventional ML models to investigate mode choice, Q4a identifies 20 studies which do not perform any type of hyper-parameter optimisation. This includes 11 which do not state hyper-parameter values at all, and nine which use default values or provide values without explanation. Model performance is highly dependent on chosen hyper-parameter values. Additionally, optimal hyper-parameter values are highly task dependent, and will vary for different datasets, metrics, scenarios, etc. Using default hyper-parameter values, or values from previous studies with different modelling scenarios or data, is therefore likely to result in sub-optimal hyper-parameters being used., and the resultant model will perform worse than the optimised model. If the hyper-parameters of each classifier have not been optimised, it is not possible to make valid comparisons between the respective algorithms, as any difference in model performance may be due to better hyper-parameter values selected for one algorithm than another.

Q4a also identifies that no studies use a fully rigorous hyper-parameter search method. Many studies use inconsistent search methods, only searching over one parameter within one model (e.g. number of neurons in a hidden layer), whilst leaving all others with default values. Optimising only the parameters for only a subset of classifiers being compared will tend to improve the performance of those classifiers over those which have not been optimised. Additionally, the search space should cover all dimensions of the

hyper-parameter space, otherwise optimal values are unlikely to be found. Whilst certain hyper-parameters may have little/no effect on model performance, there is no way to determine this unless they are tested.

Finally, search schemes should be used which maximise the probability of finding optimal hyper-parameters in an unbiased manner. Only one study uses an automated sequential search (genetic algorithm in S19) to optimise model hyper-parameters, the rest either using a pre-specified search space (linear search/grid search) or manual search/trial and error.

The primary advantages of manual search are its simplicity and the ability to use the modeller's intuition (from previous trials and similar classification tasks) to influence subsequent guesses. However, manual search presents both high potential for the introduction of bias, and difficulty in reproducing results. Additionally, as the search is manual and cannot be parallelised, it practically limits the modeller to a small number of trials in S.

Grid-search predefines a set of candidate values for each hyper-parameter and use them to define a search space S containing each unique combination of values. Grid-search can be both automated and parallelised, and therefore enables a greater set of candidate values to be searched than with a manual search. However, grid-search is unable to learn from previous evaluations, and so spends a lot of time evaluating candidate values which are unlikely to perform well. Additionally, the same values for each hyper-parameter are repeated for each dimension of the search, limiting the likelihood of evaluating the optimal value for each hyper-parameter. As such, grid-search is highly inefficient for hyper-parameter selection and has been shown to perform poorly in practice at finding optimal hyper-parameter values compared to other search schemes, including random search (Bergstra and Bengio 2012).

Finally, Q4c identifies eight studies which include the validation data in the hyper-parameter search [S4; S5.1; S5.2; S14; S32; S33; S50; S51], as well as 14 which do not state the data used [S2; S3; S9; S13; S16; S19; S24; S34; S39; S40; S44; S48; S55; S58]. Fitting hyper-parameters to the holdout validation data allows the model to select optimal hyper-parameters specifically for that data. In other words, this presents the potential for the model to fit to the validation data using the hyper-parameters (*data leakage*). This will upward bias the performance estimate over that which would be achieved with previously unseen data. This is explored by Varma and Simon (2006), who show that cross-validation provides an upward biased estimate of true performance if it is used for model optimisation.

As discussed, validating a model on previously unseen data is an essential step in predictive modelling. Holdout validation data should not be seen by the model at any time during model development (including hyper-parameter optimisation) until the testing of the finished model.

Of the three technical limitations related to model optimisation, one represents a pitfall (optimising hyper-parameters on validation data), and two represent areas for improvement (not performing hyper-parameter optimisation, and studies not using rigorous hyper-parameter search schemes).

The discussion of Q4c also highlights one general limitation, that studies do not report the model hyper-parameters and hyper-parameter selection schemes with sufficient detail. As with the details of methodologies in Q2, this is problematic for repeatability of the mode choice experiments implemented in these studies. Hyper-parameter values

and selection schemes should be recorded in detail in order to ensure repeatability of the studies.

4.5 How is the best model selected?

This section discusses the model selection techniques (i.e. selecting from different models with optimised hyper-parameters) used in the 63 studies in the review, reviewing the responses to Q5a.

Across all 63 studies, only four [S15; S32; S48; S51] conduct any analysis of the uncertainty or distribution of model performance. S15 uses 10-fold cross validation to estimate the accuracy of seven different classifiers. Firstly, the study uses a Kruskal-Wallis test at a 5 % significance level to test the null hypothesis that the performance estimates of all classifiers tested are not significantly different from one-another. Secondly, a two-sided Wilcoxon rank-sum test is applied pairwise between the classifiers to test whether different pairs of classifiers are significantly different from each other.

Three papers [S32; S48; S51] estimate the standard deviation of the metrics (accuracy in S32 and S48, and accuracy and recall in S51) across each run of k-fold cross-validation/ repeated holdout validation. These estimates of standard deviations are not used to form any formal significance tests in these studies.

Model selection - limitations

One technical limitation is identified in relation the model selection techniques used in the studies: studies not analysing uncertainty in performance estimates.

Q5a identifies that 59 out of the 63 studies do not analyse the expected distribution of the performance estimates. Each evaluation of model performance on a validation sample (whether through holdout validation or repeated cross-validation) is a random variable. If the distributions of the performance estimates are not accounted for, any apparent differences between different classifiers' performance estimates may be due to noise in this variable. Whilst several papers discuss the relative performance of classifiers for the mode choice prediction task, only one [S15] applies any formal test to investigate the statistical significance of differences between the classifiers. Additionally, as a discontinuous scoring metric (accuracy) is used and the number evaluations is low (10 folds of cross-validation) the direct distribution of the metric cannot be analysed, and instead non-parametric pairwise testing is used.

This limitation represents an area for improvement.

5 Conclusions

This paper conducts a systematic review of ML methodologies for modelling passenger mode choice. The review investigates five research questions covering classification techniques, datasets, performance estimation, model optimisation, and model selection.

A comprehensive search methodology across the three largest online publication databases is designed and used to identify 468 unique records. The record titles, abstracts, and pub-

Table 12: Limitations identified within systematic review.

No.	Classification stage	Description	Type
	Technical limitations		
TL1	Datasets	Studies not including any attributes of the mode-alternatives	Area for improvement
TL2	Datasets	Studies using input features which are dependent on output choice	Pitfall
TL3	Performance estimation	Studies using inappropriate validation schemes	Pitfall
TL4	Performance estimation	Studies using incorrect sampling methods for hierarchical data	Pitfall
TL5	Performance estimation	Studies not performing external validation	Area for improvement
TL6	Performance estimation	Studies using only discrete metrics	Pitfall
TL7	Model optimisation	Studies not performing any type of hyper-parameter optimisation	Area for improvement
TL8	Model optimisation	Studies not using rigorous hyper-parameter search schemes	Area for improvement
TL9	Model optimisation	Studies optimising hyper-parameters on test data	Pitfall
TL10	Model selection	Studies not analysing uncertainty in performance estimates	Area for improvement
	General limitations		
GL1	Classification techniques	Limited number of studies which systematically compare several classifiers on the same task	
GL2	Classification techniques	Relatively low number of investigations into EL algorithms, in particular GBDT	
GL3	Classification techniques	Inconsistent representation of RUMs in ML studies	
GL4	Datasets	Not describing the dataset and modelling process in sufficient detail	
GL5	Datasets	Shortage of studies using large datasets to investigate mode-choice	
GL6	Datasets	Lack of relevant, openly available datasets including mode-alternative attributes	
GL7	Datasets	Not considering sampling of the data from the population	
GL8	Model optimisation	Not presenting specific model hyper-parameters	

lication details are screened for relevance, leaving 96 articles. The technical content of the full-text of these articles is assessed according to the eligibility criteria. In total, following the two screening processes, 60 full text peer-reviewed articles containing 63 primary studies are used for data extraction.

The studies are each reviewed in detail to extract 15 attributes covering the five research questions. Through this process, 18 limitations are identified: 10 technical limitations, and eight general limitations. The limitations are summarised in table 12. As shown in the table, each technical limitation belongs to one of the classification stages out of classification techniques, datasets, performance estimation, model optimisation, and model selection.

Of the 10 limitations, five represent *pitfalls* in the methodologies which are likely to impact the results of an investigation, and five are identified as *areas for improvement* which are not incorrect but could be addressed in order to improve the reliability of the results and/or predictive performance of the models.

A full summary of the technical limitations present in each study is given in table 13. All studies have at least three technical limitations within their methodology, and only one study does not have any of the *pitfalls* [S18].

5.1 Recommendations and further work

As this paper shows, there is increasing research focus on ML as an alternative to RUMs for modelling passenger mode choice. This approach has the potential to provide valuable new insights into mode choice modelling research questions when used correctly. However, from the analysis in the systematic review, it is clear that the methodologies used are highly fragmented, and there needs to be further work to establish good standard methodological practice for the use of ML for choice modelling. In particular, almost all of the studies identified in the review show at least one of five methodological pitfalls identified, which will result in biased estimates of model performance. This review has

Table 13: Summary of limitations within each study in systematic review.

Number	Paper	TL1	TL2	TL3	TL4	TL5	TL6	TL7	TL8	TL9	TL10	Sum
S1	Andrade, Uchida, and Kagaya (2006)			✓		✓	✓	NA	NA		✓	4
S2	Assi et al. (2018)	✓	✓			✓	✓		✓	✓	✓	7
S3	Biagioni et al. (2009)		✓	✓	✓	✓	✓		✓	✓	✓	8
S4	Cantarella and de Luca (2003)					✓			✓	✓	✓	4
S5.1	Cantarella and de Luca (2005)					✓			✓	✓	✓	4
S5.2	-					✓			✓	✓	✓	4
S6	Chalumuri et al. (2009)				?	✓	✓		✓		✓	5
S7	Cheng et al. (2014)	✓			✓	✓	✓	NA	NA		✓	5
S8	Dell'Orco and Ottomanelli (2012)					✓	✓	NA	NA		✓	3
S9	Edara, Teodorović, and Baik (2007)					✓	NA		✓	✓	✓	4
S10	Ermagun, Rashidi, and Lari (2015)					✓	✓	✓	✓		✓	5
S11	Errampalli, Okushima, and Akiyama (2007)				?	✓	✓	NA	NA		✓	4
S12	Gao et al. (2013)	✓			?	✓	✓	✓	✓		✓	7
S13	Gazder and Ratrout (2015)	✓	✓			✓	✓		✓	✓	✓	7
S14	Golshani et al. (2018)				✓	✓	✓		✓	✓	✓	6
S15	Hagenauer and Helbich (2017)	✓			✓	✓	✓		✓			5
S16	Hensher and Ton (2000)						?		✓	✓	✓	4
S17	Hossein Rashidi and Hasegawa (2014)	✓			✓	✓	✓	✓	✓		✓	7
S18	Hussain et al. (2017)	✓				✓			✓		✓	4
S19	Jia, Cao, and Yang (2015)	✓			?	✓			✓	✓	✓	6
S20	Juremalani (2017)	✓		✓		✓	✓	✓	✓		✓	7
S21.1	Karlaftis (2004)		✓			✓	✓	✓	✓		✓	6
S21.2	-		✓	✓		✓	✓	✓	✓		✓	7
S21.3	-		✓			✓	✓	✓	✓		✓	6
S22	Kedia, Saw, and Katti (2015)	✓			?	✓	✓	NA	NA		✓	5
S23	Kumar, Sarkar, and Madhu (2013)					✓	✓	NA	NA		✓	3
S24	Lee, Derrible, and Pereira (2018)				✓	✓	✓		✓	✓	✓	6
S25	Li et al. (2016)	✓	✓	✓		✓	✓	✓	✓		✓	8
S26	Liang et al. (2018)	✓	✓	✓		✓	✓	✓	✓		✓	8
S27	Lindner, Pitombo, and Cunha (2017)	✓				✓	✓	✓	✓		✓	6
S28	Lu and Kawamura (2010)				✓	✓	✓	NA	NA		✓	4
S29	Ma (2015)	✓	✓		✓	✓	✓	NA	NA		✓	6
S30	Ma, Chow, and Xu (2017)		✓			✓		NA	NA		✓	3
S31	Moons, Wets, and Aerts (2007)					✓	✓	✓	✓		✓	5
S32	Nam et al. (2017)					✓			✓	✓		3
S33	Omrani (2015)				✓	✓			✓	✓	✓	5
S34	Omrani et al. (2013)				✓	✓	✓		✓	✓	✓	6
S35	Papaioannou and Martinez (2015)			✓	✓	✓	✓	NA	NA		✓	5
S36	Pirra and Diana (2017)	✓			✓	✓	✓	✓	✓		✓	7
S37	Pitombo et al. (2015)	✓	✓			✓	✓	✓	✓		✓	7
S38	Pulugurta, Arun, and Errampalli (2013)				✓	✓	✓	NA	NA		✓	4
S39	Raju, Sikdar, and Dhingra (1996)				?	✓	✓		✓	✓	✓	6
S40	Ramanuj and Gundaliya (2013)	✓	✓		?	✓	✓		✓	✓	✓	8
S41	Rasouli and Timmermans (2014)	✓			✓	✓	✓	✓	✓		✓	7
S42	Seetharaman et al. (2009)					✓	✓	NA	NA		✓	3
S43	Sekhar, Minal, and Madhu (2016)	?	?		?	✓		✓	✓		✓	7
S44	Semanjski, Lopez, and Gautama (2016)	✓				✓	✓		✓	✓	✓	6
S45	Shafahi and Nazari (2006)				?	✓	✓	NA	NA		✓	4
S46	Shukla et al. (2013)	✓	✓		✓	✓	✓	✓	✓		✓	8
S47	Subba Rao et al. (1998)					✓	✓	✓	✓		✓	5
S48	Tang, Xiong, and Zhang (2015)	✓			?	✓	✓		✓	✓		6
S49	Tang, Yang, and Zhang (2012)		✓		✓	✓	✓		✓		✓	6
S50	Van Middelkoop, Borgers, and Timmermans (2003)	✓		✓	✓	✓	✓		✓	✓	✓	8
S51	Wang and Ross (2018)		✓		✓	✓	✓		✓	✓		6
S52	Wang and Namgung (2007)	✓		✓		✓	✓	NA	NA		✓	5
S53	Xian-Yu (2011)	✓			?	✓	✓		✓		✓	6
S54	Xie, Lu, and Parkany (2003)	✓	✓		?	✓	✓		✓		✓	7
S55	Yin and Guan (2011)	?	?		?	✓	✓		✓	✓	✓	8
S56	Zenina and Borisov (2011)	✓	✓	✓		✓	✓	✓	✓		✓	8
S57	Zhang and Xie (2008)				?	✓	✓		✓		✓	5
S58	Zhao et al. (2010)					✓	✓		✓	✓	✓	5
S59	Zhou and Lu (2011)	✓			?	✓	✓	✓	✓		✓	7
S60	Zhu et al. (2017)				✓	✓	✓	✓	✓		✓	6
Sum		29	19	10	34	62	53	20	49	22	62	

not performed any quantitative analysis of the impacts of these pitfalls, or of the relative performance of the classifiers considered.

As identified by the general limitations, there are a limited number of studies which systematically compare several classifiers on the same task. Additionally, there is inconsistent representation of RUMs within papers that compare ML and RUMs.

This leaves three key directions for further work: (i) establish a standardised methodology which can be used for both ML and random utility approaches which addresses the limitations raised in this review, (ii) use the methodology to investigate the impacts of the identified pitfalls on modelling results, and (iii) use the new methodology to systematically compare a range of ML algorithms with RUMs for mode choice modelling.

5.2 Limitations of systematic review

This section analyses the limitations of the review with respect to the recommended PRISMA guidelines (Moher et al. 2009).

Whilst a comprehensive and exhaustive search methodology covering the three largest online databases was used to identify relevant literature, there may have been relevant studies which are not included. Additionally, the review does not consider grey literature or unpublished material. However, in this new, research-led field, the authors are confident that the state-of-the-art techniques are well covered by the sample of papers assembled.

This review focuses purely on the methodologies used and makes no attempt to draw conclusions on the findings reported by each paper. As such, no assessment is made of the quality of each paper, nor the publication bias of the field.

Whilst the procedure for the review is designed to be as objective as possible, the data extraction and discussion is carried out by the first author, under the guidance of the co-authors. This is according to available resources. All results and decisions have been double checked, but there may be remaining errors, which are the responsibility of the authors.

Acknowledgements

This work was supported by the Future Infrastructure and Built Environment Centre for Doctoral Training at the University of Cambridge, funded by the UK Engineering and Physical Sciences Research Council (EP/L016095/1).

Declarations of interest: none

References

Andrade, Katia, Kenetsu Uchida, and Seiichi Kagaya (2006). “Development of Transport Mode Choice Model by Using Adaptive Neuro-Fuzzy Inference System”. In: *Transportation Research Record* 1977, pp. 8–16.

Assi, Khaled J. et al. (2018). “Mode Choice Behavior of High School Goers: Evaluating Logistic Regression and MLP Neural Networks”. In: *Case Studies on Transport Policy* 6.2, pp. 225–230.

Barff, Richard, David Mackay, and Richard W. Olshavsky (Mar. 1, 1982). "A Selective Review of Travel-Mode Choice Models". In: *Journal of Consumer Research* 8.4, pp. 370–380.

Ben-Akiva, Moshe E. and Steven R. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press. 424 pp.

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research* 13 (Feb), pp. 281–305.

Biagioni, James P. et al. (2009). "Tour-Based Mode Choice Modeling: Using an Ensemble of Conditional and Unconditional Data Mining Classifiers". In: *Transportation Research Board 88th Annual Meeting*. Vol. 312. Washington DC, USA: Transportation Research Board, pp. 1–15.

Bierlaire, Michel (2018). *Biogeme Examples - Swissmetro*. URL: `http://biogeme.epfl.ch/examples_swissmetro.html` (visited on 08/26/2018).

Bierlaire, Michel, Kay Axhausen, and Georg Abay (2001). "The Acceptance of Modal Innovation: The Case of Swissmetro". In: *Swiss Transport Research Conference*.

Bottou, Léon and Chih-Jen Lin (2007). "Support Vector Machine Solvers". In: *Large scale kernel machines* 3.1, pp. 301–320.

Breiman, Leo (Oct. 19, 2017). *Classification and Regression Trees*. Routledge.

Brown, Iain and Christophe Mues (Feb. 15, 2012). "An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets". In: *Expert Systems with Applications* 39.3, pp. 3446–3453.

Bureau of Transportation Statistics (2018). *The 1995 American Travel Survey (ATS) - Household Trip Characteristics*. URL: `https://catalog.data.gov/dataset/the-1995-american-travel-survey-ats-household-trips` (visited on 08/26/2018).

Cantarella, Giulio Erberto and Stefano de Luca (2003). "Modeling Transportation Mode Choice through Artificial Neural Networks". In: *Fourth International Symposium on Uncertainty Modeling and Analysis, 2003. ISUMA 2003.* College Park, MD, USA: IEEE, pp. 84–90.

— (2005). "Multilayer Feedforward Networks for Transportation Mode Choice Analysis: An Analysis and a Comparison with Random Utility Models". In: *Transportation Research Part C: Emerging Technologies*. Handling Uncertainty in the Analysis of Traffic and Transportation Systems (Bari, Italy, June 10–13 2002) 13.2, pp. 121–155.

Caruana, Rich and Alexandru Niculescu-Mizil (2006). "An Empirical Comparison of Supervised Learning Algorithms". In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM, pp. 161–168.

Chalumuri, Ravi Sekhar et al. (2009). "Applications of Neural Networks in Mode Choice Modelling for Second Order Metropolitan Cities of India". In: *Proceedings of the Eastern Asia Society for Transportation Studies*. Vol. 7. Surabaya, Indonesia: Eastern Asia Society for Transportation Studies, pp. 1–16.

Chapelle, Olivier and Yi Chang (2011). "Yahoo! Learning to Rank Challenge Overview". In: *Proceedings of the Learning to Rank Challenge*, pp. 1–24.

Cheng, Long et al. (2014). "Modeling Mode Choice Behavior Incorporating Household and Individual Sociodemographics and Travel Attributes Based on Rough Sets Theory". In: *Computational Intelligence and Neuroscience* 2014, pp. 1–9.

Chicago Metropolitan Agency for Planning (2018a). *CATS Household Travel Survey, 1990*. URL: https://datahub.cmap.illinois.gov/dataset/travel-survey-1990 (visited on 08/26/2018).

— (2018b). *Travel Tracker Survey, 2007 - 2008: Public Data*. URL: https://datahub.cmap.illinois.gov/dataset/traveltracker0708 (visited on 08/26/2018).

Cortes, Corinna and Vladimir Vapnik (Sept. 1, 1995). "Support-Vector Networks". In: *Machine Learning* 20.3, pp. 273–297.

Croissant, Yves (Dec. 16, 2016). *Ecdat: Data Sets for Econometrics*. URL: https://CRAN.R-project.org/package=Ecdat (visited on 08/26/2018).

Delaware Valley Regional Planning Commission (2018). *2012 Household Travel Survey*. URL: https://www.dvrpc.org/transportation/Modeling/Data/ (visited on 08/26/2018).

Dell'Orco, Mauro and Michele Ottomanelli (2012). "Simulation of Users Decision in Transport Mode Choice Using Neuro-Fuzzy Approach". In: *International Conference on Computational Science and Its Applications (ICCSA 2012)*. Salvador de Bahia, Brazil: Springer, pp. 44–53.

Edara, Praveen Kumar, Dušan Teodorović, and Hojong Baik (2007). "Using Neural Networks to Model Intercity Mode Choice". In: *Smart Systems Engineering: Computational Intelligence in Architecting Complex Engineering Systems*. Vol. 17. Artificial Neural Networks in Engineering Conference (ANNIE 2007). St Louis, Missouri, USA: ASME Press, pp. 143–148.

Ermagun, Alireza, Taha Hossein Rashidi, and Zahra Ansari Lari (2015). "Mode Choice for School Trips: Long-Term Planning and Impact of Modal Specification on Policy Assessments". In: *Transportation Research Record* 2513, pp. 97–105.

Errampalli, Madhu, Masashi Okushima, and Takamasa Akiyama (2007). "Combined Fuzzy Logic Based Mode Choice and Microscopic Simulation Model for Transport Policy Evaluation". In: *11th World Conference on Transport Research*. Berkley CA, USA: Transportation Research Board.

Federal Highway Administration (2018). *National Household Travel Survey*. URL: https://nhts.ornl.gov/ (visited on 08/26/2018).

Gao, Jian et al. (2013). "Impact of Transit Network Layout on Resident Mode Choice". In: *Mathematical Problems in Engineering* 2013, pp. 1–8.

Gazder, Uneb and Nedal T. Ratrout (2015). "A New Logit-Artificial Neural Network Ensemble for Mode Choice Modeling: A Case Study for Border Transport". In: *Journal of Advanced Transportation* 49.8, pp. 855–866.

Gneiting, Tilmann and Adrian E Raftery (Mar. 2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Golshani, Nima et al. (2018). "Modeling Travel Mode and Timing Decisions: Comparison of Artificial Neural Networks and Copula-Based Joint Model". In: *Travel Behaviour and Society* 10, pp. 21–32.

Greene, William H. (Nov. 21, 2011). *Econometric Analysis*. Pearson Higher Ed. 1230 pp.

Hagenauer, Julian and Marco Helbich (2017). "A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice". In: *Expert Systems with Applications* 78, pp. 273–282.

Hartigan, J. A. and M. A. Wong (1979). “Algorithm AS 136: A K-Means Clustering Algorithm”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1, pp. 100–108.

Hastie, Trevor, Jerome Friedman, and Robert Tibshirani (2008). *The Elements of Statistical Learning*. 2nd ed. Springer Series in Statistics. New York: Springer. 745 pp.

Hensher, David A. and Lester W. Johnson (Aug. 1, 1983). “Alternative Modelling Procedures in Studies of Travel Mode Choice: A Review and Appraisal”. In: *Transportation Planning and Technology* 8.3, pp. 203–216.

Hensher, David A. and Tu T. Ton (2000). “A Comparison of the Predictive Potential of Artificial Neural Networks and Nested Logit Models for Commuter Mode Choice”. In: *Transportation Research Part E: Logistics and Transportation Review* 36.3, pp. 155–172.

Hoos, Holger et al. (2014). “An Efficient Approach for Assessing Hyperparameter Importance”. In: *International Conference on Machine Learning*, pp. 754–762.

Hornik, Kurt (Jan. 1, 1991). “Approximation Capabilities of Multilayer Feedforward Networks”. In: *Neural Networks* 4.2, pp. 251–257.

Hossein Rashidi, Taha and Hironobu Hasegawa (2014). “An Innovative Simultaneous System of Disaggregate Models for Trip Generation, Mode, and Destination Choice”. In: *Transportation Research Board 93rd Annual Meeting*. Washington DC, USA: Transportation Research Board.

Hussain, H. D. et al. (2017). “Analysis of Transportation Mode Choice Using a Comparison of Artificial Neural Network and Multinomial Logit Models”. In: *ARPN Journal of Engineering and Applied Sciences* 12.5, pp. 1483–1493.

Jia, Hongfei, Xiongjiu Cao, and Kaihua Yang (2015). “Residents’ Travel Mode Choice Model”. In: *Traffic Engineering & Control* 56.1, pp. 169–174.

Jing, Peng et al. (Apr. 14, 2018). “Travel Mode and Travel Route Choice Behavior Based on Random Regret Minimization: A Systematic Review”. In: *Sustainability* 10.4, p. 1185.

Juremalani, Jayesh (2017). “Comparison of Different Mode Choice Models for Work Trips Using Data Mining Process”. In: *Indian Journal of Science and Technology* 10.17, pp. 1–3.

Karlaftis, Matthew G. (2004). “Predicting Mode Choice through Multivariate Recursive Partitioning”. In: *Journal of Transportation Engineering* 130.2, pp. 245–250.

Kedia, Ashu Shivkumar, Krishna Bhuneshwar Saw, and Bhimaji Krishnaji Katti (2015). “Fuzzy Logic Approach in Mode Choice Modelling for Education Trips: A Case Study of Indian Metropolitan City”. In: *Transport* 30.3, pp. 286–293.

Kitchenham, Barbara and Stuart Charters (2007). *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. EBSE Technical Report 2007-01. EBSE.

Kruger, J. (1991). “Review of Research on Urban Area Mode Choice Modelling”. In: *13th CAITR Conference, December 12-13, 1991, Cromwell College, The University of Queensland*. Conference of Australian Institutes of Transport Research (CAITR), 13th, 1991, Brisbane, Queensland.

Kumar, Mukesh, Pradip Sarkar, and Errampalli Madhu (2013). “Development of Fuzzy Logic Based Mode Choice Model Considering Various Public Transport Policy Options”. In: *International Journal for Traffic and Transport Engineering* 3.4, pp. 408–425.

Lee, Dongwoo, Sybil Derrible, and Francisco Camara Pereira (2018). "Comparison of Four Types of Artificial Neural Networks and a Multinomial Logit Model for Travel Mode Choice Modeling". In: *Transportation Research Board 97th Annual Meeting*. Washington DC, USA: Transportation Research Board.

Li, Juan et al. (2016). "Cluster-Based Logistic Regression Model for Holiday Travel Mode Choice". In: *Procedia Engineering*. Vol. 137. 6th International Conference on Green Intelligent Transportation System and Safety (GITSS 2015). Beijing, China: Elsevier, pp. 729–737.

Liang, LeiLei et al. (2018). "Travel Mode Choice Analysis Based on Household Mobility Survey Data in Milan: Comparison of the Multinomial Logit Model and Random Forest Approach". In: *Transportation Research Board 97th Annual Meeting*. Washington DC, USA: Transportation Research Board.

Lindner, Anabele, Cira Souza Pitombo, and André Luiz Cunha (2017). "Estimating Motorized Travel Mode Choice Using Classifiers: An Application for High-Dimensional Multicollinear Data". In: *Travel Behaviour and Society* 6, pp. 100–109.

Lu, Yandan and Kazuya Kawamura (2010). "Data-Mining Approach to Work Trip Mode Choice Analysis in Chicago, Illinois, Area". In: *Transportation Research Record* 2156, pp. 73–80.

Luxembourg Institute of Socio-Economic Research (2018). *Socio-Economic Panel of Liewen Zu Lëtzebuerg III (PSELL3)*. URL: `http://dataservice.liser.lu/en_US/dataservice/db=23` (visited on 08/26/2018).

Ma, Tai-Yu (2015). "Bayesian Networks for Multimodal Mode Choice Behavior Modelling: A Case Study for the Cross Border Workers of Luxembourg". In: *Transportation Research Procedia*. Vol. 10. 18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands, pp. 870–880.

Ma, Tai-Yu, Joseph Y. J. Chow, and Jia Xu (2017). "Causal Structure Learning for Travel Mode Choice Using Structural Restrictions and Model Averaging Algorithm". In: *Transportmetrica A: Transport Science* 13.4, pp. 299–325.

McFadden, Daniel (1981). "Econometric Models of Probabilistic Choice". In: *Structural Analysis of Discrete Data with Econometric Applications*. Ed. by Charles F. Manski and Daniel McFadden. MIT Press, pp. 198–272.

Meixell, Mary J. and Mario Norbis (Aug. 15, 2008). "A Review of the Transportation Mode Choice and Carrier Selection Literature". In: *The International Journal of Logistics Management* 19.2, pp. 183–211.

Metropolitan Transportation Commission (2018a). *1990 Bay Area Travel Surveys*. URL: `http://www.surveyarchive.org/` (visited on 08/26/2018).

— (2018b). *San Francisco Bay Area Travel Survey 2000*. URL: `http://www.surveyarchive.org/` (visited on 08/26/2018).

Minal and Ch. Ravi Sekhar (Sept. 2014). "Mode Choice Analysis: The Data, the Models and Future Ahead". In: *International Journal for Traffic and Transport Engineering* 4.3, pp. 269–285.

Moher, David et al. (2009). "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement". In: *PLoS Medicine* 6.7, p. 6.

Moons, Elke, Geert Wets, and Marc Aerts (2007). "Nonlinear Models for Determining Mode Choice". In: *Progress in Artificial Intelligence*. 13th Portuguese Conference on

Aritficial Intelligence (EPIA 2007). Lecture Notes in Computer Science. Guimarães, Portugal: Springer, pp. 183–194.

Nam, Daisik et al. (2017). "A Model Based on Deep Learning for Predicting Travel Mode Choice". In: *Transportation Research Board 96th Annual Meeting*. Washington DC, USA: Transportation Research Board, pp. 8–12.

Nerhagen, L. (2000). "Mode Choice Behaviour, Travel Mode Choice Models and Value of Time Estiamtion. A Literature Review". In: *CTEK working paper*.

Niculescu-Mizil, Alexandru and Rich Caruana (2005). "Predicting Good Probabilities with Supervised Learning". In: *Proceedings of the 22nd International Conference on Machine Learning*. ACM, pp. 625–632.

Omrani, Hichem (2015). "Predicting Travel Mode of Individuals by Machine Learning". In: *Transportation Research Procedia*. Vol. 10. 18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands, pp. 840–849.

Omrani, Hichem et al. (2013). "Prediction of Individual Travel Mode with Evidential Neural Network Model". In: *Transportation Research Record* 2399, pp. 1–8.

Papaioannou, Dimitrios and Luis Miguel Martinez (2015). "The Role of Accessibility and Connectivity in Mode Choice. A Structural Equation Modeling Approach". In: *Transportation Research Procedia*. Vol. 10. 18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands, pp. 831–839.

Pirra, Miriam and Marco Diana (2017). "Tour-Based Mode Choice Study Through Support Vector Machine Classifiers". In: *Transportation Research Board 96th Annual Meeting*. Washington DC, USA: Transportation Research Board.

Pitombo, Cira Souza et al. (2015). "A Two-Step Method for Mode Choice Estimation with Socioeconomic and Spatial Information". In: *Spatial Statistics* 11, pp. 45–64.

Platt, John C. (1999). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Advances in Large Margin Classifiers*. MIT Press, pp. 61–74.

Pulugurta, Sarada, Ashutosh Arun, and Madhu Errampalli (2013). "Use of Artificial Intelligence for Mode Choice Analysis and Comparison with Traditional Multinomial Logit Model". In: *Procedia - Social and Behavioral Sciences*. Vol. 104. 2nd Conference of Transportation Research Group of India (CTRG 2013). Agra, India: Elsevier, pp. 583–592.

Raju, K A, P K Sikdar, and S L Dhingra (1996). "Modelling Mode Choice by Means of an Artificial Neural Network". In: *Environment and Planning B: Planning and Design* 23.6, pp. 677–683.

Ramanuj, P. S. and P. J. Gundaliya (2013). "Disaggregated Modeling of Mode Choice by ANN-a Case Study of Ahmedabad City in Gujarat State". In: *Journal of the Indian Roads Congress* 74.1, pp. 3–12.

Rasouli, Soora and Harry J.P. Timmermans (2014). "Using Ensembles of Decision Trees to Predict Transport Mode Choice Decisions: Effects on Predictive Success and Uncertainty Estimates". In: *European Journal of Transport and Infrastructure Research* 14.4, pp. 412–424.

Ratrout, Nedal T., Uneb Gazder, and Hashim M.N. Al-Madani (Jan. 1, 2014). "A Review of Mode Choice Modelling Techniques for Intra-City and Border Transport". In: *World Review of Intermodal Transportation Research* 5.1, pp. 39–58.

Saeb, Sohrab et al. (May 1, 2017). “The Need to Approximate the Use-Case in Clinical Machine Learning”. In: *GigaScience* 6.5, pp. 1–9.

Seetharaman, Padma et al. (2009). “Comparative Evaluation of Mode Choice Modelling by Logit and Fuzzy Logic”. In: *Proceedings of the Eastern Asia Society for Transportation Studies*. Vol. 7. Surabaya, Indonesia: Eastern Asia Society for Transportation Studies, pp. 1–16.

Sekhar, Ch. Ravi, Minal, and E. Madhu (2016). “Mode Choice Analysis Using Random Forrest Decision Trees”. In: *Transportation Research Procedia*. Vol. 17. 11th Transportation Planning and Implementation Methodologies for Developing Countries, TPMDC 2014, 10-12 December 2014, Mumbai, India, pp. 644–652.

Semanjski, Ivana, Angel Lopez, and Sidharta Gautama (2016). “Forecasting Transport Mode Use with Support Vector Machines Based Approach”. In: *Transactions on Maritime Science* 5.2, pp. 111–120.

Shafahi, Yusof and Sobhaan Nazari (2006). “Disaggregate Mode Choice Analysis for Work Trips Using Genetic-Fuzzy and Neuro-Fuzzy Systems.” In: *Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Soft Computing (ASC 2006)*. Palma de Mallorca. Spain: ACTA Press, pp. 250–255.

Shmueli, Galit (Aug. 2010). “To Explain or to Predict?” In: *Statistical Science* 25.3, pp. 289–310.

Shukla, Nagesh et al. (2013). “Data-Driven Modeling and Analysis of Household Travel Mode Choice”. In: *20th International Congress on Modelling and Simulation (MODSIM 2013)*. Adelaide, Australia: The Modelling and Simulation Society of Australia and New Zealand Inc., pp. 92–98.

Srivastava, Nitish et al. (2014). “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *The Journal of Machine Learning Research* 15.1, pp. 1929–1958.

Subba Rao, P. V. et al. (1998). “Another Insight into Artificial Neural Networks through Behavioural Analysis of Access Mode Choice”. In: *Computers, Environment and Urban Systems* 22.5, pp. 485–496.

Svozil, Daniel, Vladimir Kvasnicka, and Jiri Pospichal (Nov. 1, 1997). “Introduction to Multi-Layer Feed-Forward Neural Networks”. In: *Chemometrics and Intelligent Laboratory Systems* 39.1, pp. 43–62.

Tang, Dounan, Min Yang, and Mei Hui Zhang (2012). “Travel Mode Choice Modeling: A Comparison of Bayesian Networks and Neural Networks”. In: *Applied Mechanics and Materials* 209-211, pp. 717–723.

Tang, Liang, Chenfeng Xiong, and Lei Zhang (2015). “Decision Tree Method for Modeling Travel Mode Switching in a Dynamic Behavioral Process”. In: *Transportation Planning and Technology* 38.8, pp. 833–850.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge university press.

Transport for NSW (2018). *Household Travel Survey 2005/06 – 2016/17*. URL: https://opendata.transport.nsw.gov.au/dataset/household-travel-survey-200506-%E2%80%93-201617 (visited on 08/26/2018).

Transport for Victoria (2018). *VISTA Data and Publications*. URL: https://transport.vic.gov.au/data-and-research/vista/vista-data-and-publications/ (visited on 08/26/2018).

Van Middelkoop, Manon, Aloys Borgers, and Harry Timmermans (2003). "Inducing Heuristic Principles of Tourist Choice of Travel Mode: A Rule-Based Approach". In: *Journal of Travel Research* 42.1, pp. 75–83.

Varma, Sudhir and Richard Simon (2006). "Bias in Error Estimation When Using Cross-Validation for Model Selection". In: *BMC Bioinformatics*, p. 8.

Wang, Fangru and Catherine L. Ross (2018). "Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model". In: *Transportation Research Record* Advanced online publication, pp. 1–11.

Wang, Weijie and Moon Namgung (2007). "Knowledge Discovery from the Data of Long Distance Travel Mode Choices Based on Rough Set Theory". In: *International Journal of Multimedia and Ubiquitous Engineering* 2.3, pp. 81–90.

Wu, Ting-Fan, Chih-Jen Lin, and Ruby C. Weng (2004). "Probability Estimates for Multi-Class Classification by Pairwise Coupling". In: *Journal of Machine Learning Research* 5 (Aug), pp. 975–1005.

Xian-Yu, Jian-Chuan (2011). "Travel Mode Choice Analysis Using Support Vector Machines". In: *11th International Conference of Chinese Transportation Professionals (ICCTP 2011)*. Nanjing, China: American Society of Civil Engineers, pp. 360–371.

Xie, Chi, Jinyang Lu, and Emily Parkany (2003). "Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks". In: *Transportation Research Record* 1854, pp. 50–61.

Yin, Huanhuan and Hongzhi Guan (2011). "Traffic Mode Choice Model Based on BP Neural Network". In: *Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*. Changchun, China: IEEE, pp. 1441–1444.

Zenina, Nadezda and Arkady Borisov (2011). "Transportation Mode Choice Analysis Based on Classification Methods". In: *Scientific Journal of Riga Technical University. Computer Sciences* 45.1, pp. 49–53.

Zhang, Chongsheng, Changchang Liu, et al. (Oct. 1, 2017). "An Up-to-Date Comparison of State-of-the-Art Classification Algorithms". In: *Expert Systems with Applications* 82, pp. 128–150.

Zhang, Yunlong and Yuanchang Xie (2008). "Travel Mode Choice Modeling with Support Vector Machines". In: *Transportation Research Record* 2076, pp. 141–150.

Zhao, Dan et al. (2010). "Travel Mode Choice Modeling Based on Improved Probabilistic Neural Network". In: *Traffic and Transportation Studies 2010 (ICTTS 2010)*. Vol. 383. Kunming, China: ASCE, pp. 685–695.

Zhou, Miaomiao and Jian Lu (2011). "Research on Prediction of Traffic Mode Choice of Urban Residents". In: *11th International Conference of Chinese Transportation Professionals (ICCTP 2011)*. Nanjing, China: American Society of Civil Engineers, pp. 449–460.

Zhu, Zheng et al. (2017). "A Mixed Bayesian Network for Two-Dimensional Decision Modeling of Departure Time and Mode Choice". In: *Transportation* Advanced online publication, pp. 1–24.