New perspectives on the performance of machine learning classifiers for mode choice prediction

Tim Hillel *

4th July 2020

Report TRANSP-OR 200704
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.hillelepfl.ch

Abstract

It appears to be a commonly held belief that Machine Learning (ML) classification algorithms should achieve substantially higher predictive performance than manually specified Random Utility Models (RUMs) for choice modelling. This belief is supported by several papers in the mode choice literature, which highlight stand-out performance of non-linear ML classifiers compared with linear models. However, many studies which compare ML classifiers with linear models have a fundamental flaw in how they validate models on out-of-sample data. This paper investigates the implications of this issue by comparing out-of-sample validation using two different sampling methods for panel data: (i) trip-wise sampling, where validation folds are sampled independently from all trips in the dataset and (ii) grouped sampling, where validation folds are sampled grouped by household/person.

This paper includes two linked investigations: (i) a *dataset investigation* which quantifies the proportion of matching trips across training and validation data when using trip-wise sampling for Out-Of-Sample (OOS) validation and (ii) a *modelling investigation* which compares OOS validation results obtained using trip-wise sampling and grouped sampling. These investigations make use of the data and methodologies of three published studies which explore ML classification of mode choice.

The results of the dataset investigation indicate that using trip-wise sampling with travel diary data results in significant data leakage, with up to 96% of the trips in typical trip-wise sampling validation folds having matching trips with the same mode choice in the training data. Furthermore, the modelling investigation demonstrates that this data leakage introduces substantial bias in model performance estimates, particularly for flexible non-linear classifiers. Grouped sampling is found to address the issues associated with trip-wise sampling and provides reliable estimates of true OOS predictive performance.

The use of trip-wise sampling with panel data has led to incorrect conclusions being made in two of the investigated studies, with the original results substantially overstating the performance of ML models compared with linear Logistic Regression (LR) models. Whilst the results from this study indicate that there is a slight predictive performance advantage of non-linear classifiers (in particular Ensemble Learning (EL) models) over linear LR models, this advantage is much more modest than has been suggested by previous investigations.

Acronyms

AB AdaBoost.

ANN Artificial Neural Network.

CEL Cross-Entropy Loss.

DCA Discrete Classification Accuracy.

DCM Discrete Choice Model.

DT Decision Tree.

EL Ensemble Learning.

GBDT Gradient Boosting Decision Trees.

LOS Level of Service.

LPMC London Passenger Mode Choice.

LR Logistic Regression.

LTDS London Travel Demand Survey.

ML Machine Learning.

MNL Multinomial Logit.

NB Naïve Bayes.

NL Nested Logit.

NTS National Travel Survey.

OOS Out-Of-Sample.

RF Random Forest.

RUM Random Utility Model.

SMBO Sequential Model-Based Optimisation.

SVM Support Vector Machine.

1 Introduction

Mode choice modelling has traditionally been tackled with statistical Discrete Choice Models (DCMs) based on the random-utility framework (McFadden 1981). However, there has recently been an increased focus in the literature on applying Machine Learning (ML) classification techniques for mode choice prediction, following the numerous successes of ML approaches in other fields.

It seems intuitive that ML classifiers should afford greater predictive performance than linear Random Utility Models (RUMs); ML techniques are typically much more flexible than RUMs and the majority of ML classifiers can automatically capture complex non-linear relationships between input features and choice probabilities. Indeed, there are several existing studies that find ML classifiers significantly outperform RUMs in terms of their predictive performance. However, a recent systematic review conducted by Hillel

et al. (2020) identifies a number of technical limitations and pitfalls in the methodologies used in these studies. Of particular note, the review highlights that several previous studies make use of *trip-wise* sampling methods for hierarchical *panel data*, where each individual contributes more than one trip to the dataset.

Trip-wise sampling forms validation folds by sampling uniformly across all trips in the dataset. As such, trip-wise sampling with panel data allows for matching return, repeated, and shared trips made by the same individual or members of the same household to occur across training and validation folds. This allows a model to observe trips matching the validation data during model training, and so allows for *data leakage* of the chosen mode from the training data to the validation data. This violates the key principle of validating ML classifiers on unseen Out-Of-Sample (OOS) data and may result in biased model performance estimates, in particular for highly flexible models which are able to *overfit* to the data.

This paper explores the implications of using trip-wise sampling for panel data and introduces grouped sampling, where validation folds are sampled grouped by household/ person, as an alternative. Two approaches are used to investigate the implications of the different sampling methods:

1. **Dataset investigation:** identifies the *mechanism* for data leakage from trip-wise sampling by quantifying the proportion of matching trips in different travel diary datasets for different validation schemes.

2. **Modelling investigation:** assesses the *impacts* of the data leakage from trip-wise sampling by comparing OOS validation results obtained using trip-wise sampling and grouped sampling.

Three previous studies are used for this analysis:

- two studies which compare the performance of multiple ML classifiers using trip-wise sampling with multi-day travel diaries
  - Study 1: Chang et al. (2019), and
  - Study 2: Hagenauer and Helbich (2017)
- and a third study which investigates the use of Gradient Boosting Decision Trees (GBDT) classifiers using grouped sampling with a single-day travel diary
  - Study 3: Hillel, Elshafie, and Jin (2018).

In each case, the dataset in each paper is analysed, and the modelling is repeated using both trip-wise and grouped sampling.

The rest of the paper is laid out as follows. Firstly, Section 2 summarises existing ML and RUM performance comparisons in the literature, including their use of trip-wise/ grouped sampling, and introduces the three papers investigated in this study in more detail. Next, Section 3 presents the dataset investigation to quantify the proportion of matching trips in different travel diary datasets for different validation schemes. Section 4 then presents the modelling investigation to compare OOS validation results for different classifiers using trip-wise sampling and grouped sampling. Finally, Section 5 provides the conclusions for the study, and introduces possible directions for further work.

2 Literature review

2.1 Existing Machine Learning and Random Utility Model performance comparisons

There exist several papers in the literature which investigate mode choice prediction using ML models. This includes applications of Artificial Neural Networks (ANNs) (S. Wang and Zhao 2019; Golshani et al. 2018; Lee, Derrible, and Pereira 2018; Nam et al. 2017; Omrani et al. 2013; Cantarella and de Luca 2005; Hensher and Ton 2000; Subba Rao et al. 1998); Decision Trees (DTs) (Pitombo, Costa, and Salgueiro 2015; Tang, Xiong, and Zhang 2015; Karlaftis 2004); Ensemble Learning (EL) (Chapleau, Gaudette, and Spurr 2019; Cheng et al. 2019; Ding, Cao, and Y. Wang 2018; Hillel, Elshafie, and Jin 2018; Liang et al. 2018; Ermagun, Rashidi, and Lari 2015); and Support Vector Machines (SVMs) (Pirra and Diana 2019); as well as three papers which perform a comparative study of multiple classification approaches (Chang et al. 2019; Hagenauer and Helbich 2017; Zhou, M. Wang, and Li 2019). These papers typically compare the ML approach to a Logistic Regression (LR) model, which is used as a benchmark for a statistical RUM.[1]

Several ML mode choice studies make use of travel diary data, where individuals or households report all of the trips they make during the study period, which can range from a single day to several weeks. This includes two studies which use multi-day travel surveys to compare the relative performance of several ML classification algorithms alongside LR models (Chang et al. 2019; Hagenauer and Helbich 2017). Both of these studies find that ML classifiers, and in particular EL models, substantially outperform RUMs/LR.

Chang et al. (2019) use continuous six-week travel diary data to compare the performance of SVMs, Naïve Bayes (NB), and three EL algorithms (Random Forest (RF), AdaBoost (AB), and GBDT) alongside Nested Logit (NL) and LR models, for predicting choice out of nine different modes (walk, cycle, motorcycle, vehicle driver, vehicle passenger, bus, light-rail transit, rail, other). They find the highest performing classification is obtained with a RF, which achieves substantially higher test accuracy than a LR model (87.41% vs 63.71%).[2] Similarly, Hagenauer and Helbich (2017) use continuous six-day travel diaries to compare six ML classifiers (SVMs, NB, ANN, RF, GBDT, and a bagging ensemble) with a LR model for a four-mode problem (walk, cycle, public transport, car). They also obtain highest performance with the RF classifier, which achieves a mean test accuracy of 91.4%, compared to 56.1% for a LR model.

2.2 Trip-wise sampling

Travel diary data is *panel data*; each individual or household records more than one trip in the final dataset. As such, the dataset may include return, repeated, and shared trips, which are highly likely to be made by the same mode. Trips in travel diary data can therefore not be considered as independent.

A recent systematic review by Hillel et al. (2020) identifies 13 ML studies in the

[1] Note that typically the LR models in these papers are regularised using automated L1/L2 regularisation using the C hyperparameter, whereas traditional RUM are typically regularised using parameter significance tests for manually specified utility functions

[2] RF used in combination with a Denoising Autoencoder to pre-process the data

literature which make use of complete travel diary data (Yang and J. Ma 2019; Cheng et al. 2019; Chang et al. 2019; F. Wang and Ross 2018; Zhu et al. 2017; Hagenauer and Helbich 2017; Semanjski, Lopez, and Gautama 2016; Tang, Xiong, and Zhang 2015; Papaioannou and Martinez 2015; T.-Y. Ma 2015; Rasouli and Timmermans 2014; Shukla et al. 2013; Pulugurta, Arun, and Errampalli 2013; Biagioni et al. 2008). Of these studies, all 13 (including the two comparative studies referenced above) make use of trip-wise sampling to generate the validation datasets. Trip-wise sampling forms validation sets by sampling trips independently from the dataset and so allows trips made by the same individual or household to occur across the training and validation data. This includes return/repeated/shared trips.

Trip-wise sampling with panel data therefore allows for *data leakage* of the chosen mode from the training data to the validation data for matching return/repeated/shared trips, and so violates the key principle of validating ML classifiers on unseen OOS data. This could therefore result in unreliable predictions of predictive performance.

This problem is not unique to mode choice modelling, and a systematic review of cross-validation sampling for papers investigating medical diagnosis using panel data from wearable devices identified a similar prevalence of incorrect sampling methods in the literature (Saeb et al. 2017).

2.3 Grouped sampling

As opposed to trip-wise sampling, grouped sampling forms validation sets by sampling trips by household/person, such that all trips made by an individual or household are sampled together. Grouped sampling prevents trips made by the same individual or members of the same household from appearing in both the training and validation data, including any return, repeated, or shared trips. This therefore prevents the data-leakage occurring from matching trips. Grouped sampling is first used for ML investigations for mode choice modelling by Hillel, Elshafie, and Jin (2018), who train GBDT classifiers on trip records from a single-day travel diary to predict mode choice for a four-mode problem (walk, cycle, public transport, private vehicle).

3 Dataset investigation: matching trips

This investigation quantifies the number of matching trips which occur when using trip-wise sampling with travel-diary data. Three types of matching trips are investigateds: return, repeated, and shared. Formal definitions for each, as well as a general definition for any matching trip are given in Table 1. Note that a single trip can belong to multiple sets of trips. For example, a return trip may both be repeated, and shared by multiple members of a household.

3.1 Datasets

The papers investigated in this study have either made the dataset used in the study publicly available alongside the paper (Hillel, Elshafie, and Jin 2018; Hagenauer and Helbich 2017), or make use of publicly available open data (Chang et al. 2019). An overview of the dataset used in each study is given in the following sections.

Table 1: Definitions of matching trips.

Type	Definition
Return trip	A trip made by an individual from origin x to destination y is made in reverse between origin y and destination x. Note, these do not necessarily need to be part of a return-tour, and could have other trips made between them, or even be made as part of different tours.
Repeated trip	A trip made by an individual from origin x to destination y is made again between the same origin x and destination y (i.e. in the same direction).
Shared trip	A trip is made by two or more members of the same household with matching origin x, destination y, and departure time t. Note this can only occur in household data.
All matching trips	Any sets of trips made by the same individual/members of a household with matching origin and destination, or with reversed origin and destination.

3.1.1 Study 1: Mobidrive

Study 1 uses the 6-week Mobidrive data, which is publicly available online (Chalasani and Axhausen 2004). The dataset contains 52 265 trips made by 361 individuals in 162 households. The data includes both household and individual IDs, and so grouped sampling can be performed using the household ID.[3]

Origin and destination zone IDs are given for each trip. In total, there are 496 different reported zones in the study area. Matching trips are identified for each individual (return and repeated trips) and household (shared trips) using the origin and destination zones. Note, this may identify some false positives, as multiple different destinations could be in the same zone and therefore have the same IDs. However, it is expected that trips between matching zones should also be highly correlated for the same individual/household in terms of mode choice and features describing the trip.

3.1.2 Study 2: Dutch National Travel Survey

Study 2 uses a dataset adapted from a 6-day travel diary from the Dutch National Travel Survey (NTS). The data has been augmented with environmental data describing the land use diversity and proportion of green space of the residential postcode location, the average urban density of the selected route, and the weather conditions at the time of travel. The dataset contains a total 230 608 of trips made by 69 918 individuals. The number of households is not given. The dataset is made available as supplementary material to the paper.

The shared version of the data has no individual and/or household IDs. However, the data contains unique values for the land use diversity and proportion of green space for each postcode area in the study. As such, individual IDs can be imputed by enumerating

[3]The household IDs in the public data are non-unique, and so new household IDs are formed from unique combinations of the household ID, study code, and city code.

the unique combinations of postcode area (using the *diversity* and *green* features in the data) and socio-economic data (combination of age, gender, ethnicity, education, income, cars, bikes, and driving license ownership). If two people in the survey data from the same postcode area have the same socio-economic details they will be grouped as the same person. As such, this approach identifies less unique individuals than in the original data (45 036 vs 69 918). However, these IDs are sufficient for ensuring data from the same individual does not occur in both the train and test data.

Similarly, the dataset does not contain any details of the start or end-location of each trip. As such, matching (combined return and repeated[4]) trips are identified for each individual using exact matches of the straight-line distance value (given accurate to 100m) with the imputed individual IDs. This may result in false positive matches, as two different sets of origins and destinations may have the same distance. As there is no location information in the dataset, there is no way of determining the difference between return and repeated trips, and so results are only obtained for *all* matching trips (see Table 1).

As with many travel datasets, the NTS data is inherently imbalanced in that there are many more trips made by some modes (e.g. car) than others (e.g. public transport). To address this imbalance, the authors of the original paper use a combined over/undersampling scheme to ensure equal numbers of trips are present in the dataset. For modes with less trips than the mean trips per mode (bike, walk, and public transport), the trips are sampled with replacement until the number of trips for that mode is equal to the mean (i.e. they are *oversampled*). For modes with more trips than the mean (car), the trips are randomly removed until the number of trips for that mode is equal to the mean (i.e. they are *undersampled*).

The effect of this sampling scheme are investigated by comparing the number of matching trips for both the original unsampled data and the sampled data.

3.1.3 Study 3: London Passenger Mode Choice

Study 3 uses the London Passenger Mode Choice (LPMC) dataset, which is adapted from a single day travel diary from the London Travel Demand Survey (LTDS). The data has been augmented with imputed Level of Service (LOS) variables describing the choice-set at the time of travel. The dataset contains a total of 81 096 trips made by 31 954 individuals in 17 616 households. The dataset is made available as supplementary material to the paper. The data includes both household and individual IDs, and so grouped sampling can be performed using the household ID.

Matching trips are identified for each individual (return and repeated trips) and household (shared trips) using the reported start and end locations.[5] As the final year of data (2014/15) is used for external validation in the paper, matched trips are only identified using the training and validation data (2012/13–2013/14). This sample contains a 54 766 trips made by 21 399 individuals in 11 725 households.

3.2 Methodology

The dataset investigation has two parts, with methodologies as follows.

[4] as the data is grouped by individual and not household, shared trips cannot be identified

[5] the spatial data is omitted from the publicly available version of the data for privacy reasons

1. Quantify proportion of matching trips in each dataset:
    (a) Identify pairs/sets of return, repeated, shared, and all matching trips in each dataset through targeted searches of the features and household/person IDs.
    (b) Calculate the total number of matching trips of each type and the average set sizes for each dataset.
    (c) Calculate the proportion of each type of matching trip that has matching mode, and the overall proportion of trips in the dataset with matching mode of each type.
2. Quantify the proportions of all matching trips across training and validation data which occur with trip-wise sampling using Monte-Carlo simulation:
    (a) Sample a simulated validation set trip-wise, with the unsampled data forming the simulated training set.
    (b) Calculate the proportion of the validation set with matching trips with the same mode choice in the training.
    (c) Repeat and average over 10 000 repetitions (draws) of this process for each validation set size (30% for holdout validation, 10% for cross-validation).

Two different sample sizes are investigated for the Monte-Carlo simulation. A 30% sample is used to represent the most commonly used sample size for OOS validation for ML mode choice studies (Hillel et al. 2020). A 10% sample is then simulated to represent 10-fold cross validation, which is the most commonly used number of folds used for cross validation in the literature (Hillel et al. 2020).

3.3 Results

The numbers and proportions of matching trips in each dataset are given in Table 2. For the NTS, the values are given for the original unsampled data as well as the sampled data using the over/undersampling scheme proposed in the paper to ensure equal numbers of trips of each mode. As discussed, the analysis for the NTS data uses reconstructed individual IDs and so shared trips within a household are not identified, though are reasonably expected to occur in similar proportions to the other two datasets. Furthermore, as repeated trips in the NTS are determined using the straight line distance only, it is not possible to distinguish between return and repeated trips for this data.

From the table, it is clear to see that a substantial proportion of trips in each dataset belong to sets of matching trips with the same mode choice (shown in the column *P. dataset*). This shows that the individual trips in each dataset can clearly not be considered as independent. Trips in matching sets are likely to have highly correlated or matching features, including distances, LOS variables, departure times, socio-economic variables, etc.

The highest proportion of matching trips with the same mode choice is for the Mobidrive data, in which 91.7% of all trips have one or more matching trips with the same mode choice. This is likely due to the Mobidrive data having a long study period (six weeks) compared to the other datasets. The effect of the period of the trip diary can be

Table 2: Summary statistics for each dataset, for return, repeated, shared, and all matching trips.
N. = number of. . . ; P. same mode = proportion of matching trips that have the same mode choice; P. dataset = proportion of dataset in matching sets with same mode choice.
* the values for the sampled dataset are averaged over 10 000 draws.

Dataset	Type	N. sets	N. trips	Av. set size	N. trips same mode	P. same mode	P. dataset
Mobidrive (n=52 265)	Return	5644	47 416	8.4	45 736	0.965	0.875
	Repeated	5672	43 493	7.7	40 781	0.938	0.780
	Shared	5241	11 384	2.2	5271	0.463	0.101
	All	4415	49 771	11.3	47 922	0.963	0.917
NTS original (n=230 608)	**All**	73 015	168 868	2.3	163 642	0.969	0.710
NTS sampled* (n=230 608)	**All**	47 238	182 157	3.9	179 372	0.985	0.778
LPMC (n=54 766)	Return	15 605	32 471	2.1	30 898	0.952	0.564
	Repeated	1315	2711	2.1	2496	0.921	0.046
	Shared	8541	20 623	2.4	20 051	0.972	0.366
	All	15 814	40 520	2.6	39 357	0.971	0.719

seen in the average set-size of return and repeated trips in the Mobidrive dataset (8.4 and 7.7 respectively) compared to the single-day LPMC data (2.1 and 2.1 respectively). This occurs as trips are likely to be repeated more times during the longer travel diary. As would be expected, the set size for shared trips between the two datasets is similar, as this represents the average party size for shared household trips, which should not be affected by the study period.

Overall, the original (unsampled) NTS survey has a slightly lower proportion of matching trips with he same mode choice than the LPMC data (71.0% vs 71.9%), despite it being a longer period travel diary (six-day vs single-day). This is at least in part due to the fact that shared trips are not accounted for in the NTS data (as there is no way of identifying individual households), which represent a substantial proportion of the matching trips in the LPMC data.

The results also show that the sampling scheme used on the NTS data in Study 2 increases the proportion of matching trips. In particular, the original sample of 9300 public transport trips has been oversampled to create a sample of 57 652 trips, meaning each public transport trip has been repeated an average of 6.2 times.

The proportion of matching trips which have matching mode (P. same mode for *all* types), is above 96% for all datasets. This shows that matching trips are highly likely to share the same mode choice, suggesting that there is high potential for data-leakage if trip-wise sampling is used with these datasets. Shared trips in the Mobidrive data have a lower proportion of matching modes (46.3%). This is because vehicle driver and vehicle passenger are treated as separate mode choices in the Mobidrive dataset, whereas the other two datasets have a single vehicle mode which covers both driving and passenger. As such, multiple household members in the same vehicle will have a different mode choice.

3.3.1 Monte-carlo simulation

Table 3 shows the results of the monte-carlo simulation of trip-wise sampling for holdout validation with a 30% sample and 10-fold cross-validation. This data is represented visually in Fig. 1.

Table 3: Average proportion of matching trips with the same mode choice in each dataset using trip-wise sampling for holdout validation (30% sample) and 10-fold cross-validation (10% sample). Results estimated over 10 000 draws.

Dataset	**Holdout** (p=0.3)	**CV** (p=0.1)
Mobidrive	0.941	0.956
NTS original	0.535	0.653
NTS sampled	0.707	0.757
LPMC	0.579	0.675

The results show that for a 30% holdout validation set sampled trip-wise, over half of the trips in the validation set have matching trips with the same mode choice in the training set for all datasets. This indicates substantial data-leakage from the training data

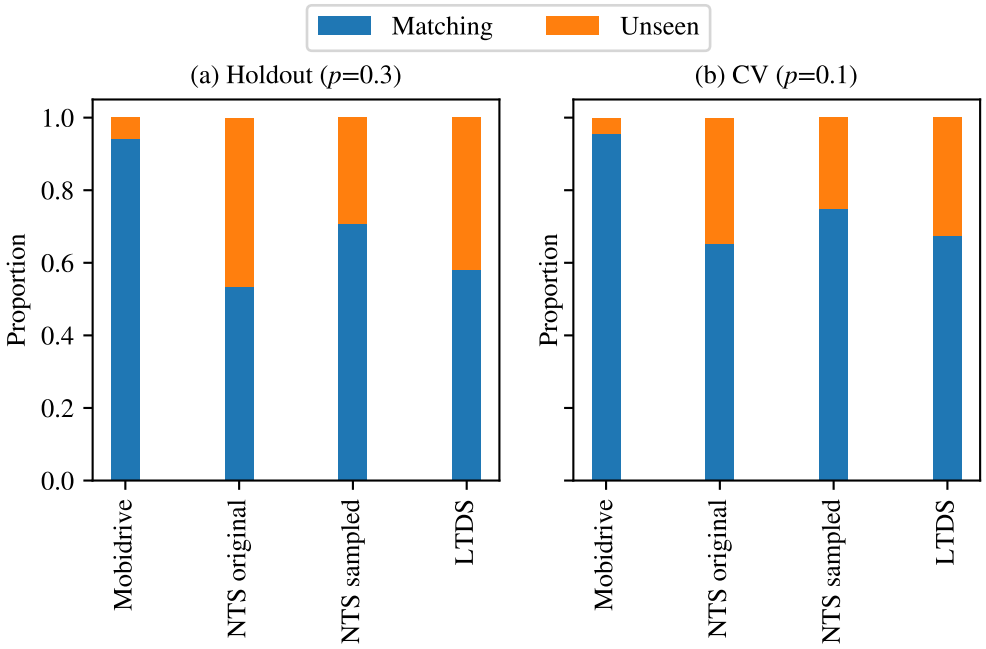

Figure 1: Average proportion of matching trips (with same mode choice) and unseen trips for each dataset over 10 000 draws of Monte-Carlo sampling for (a) holdout validation (30% sample) and (b) 10-fold cross validation (10% sample)

to the validation data. The proportion is even higher for 10-fold cross-validation, where over 65% of the validation data has matching trips with the same mode choice in the training data. The rate of matching trips is highest for the long-period Mobidrive data, where in for 10-fold cross validation, almost all (over 95%) of the the validation data is effectively repeated in the training data.

As expected, the over/undersampling of trips by each mode used on the NTS data in Study 2 results in higher proportions of matching trips between training and validation data.

4 Modelling investigation: Out-Of-Sample validation

4.1 Methodology

In order to investigate the impacts of the data leakage from trip-wise sampling methods, the modelling results from each study are repeated using both trip-wise and grouped sampling.

The two comparative studies (Study 1-2) use a grid search to identify suitable hyperparameters for each classifier, and then estimate performance using k-fold cross-validation, with the validation folds sampled trip-wise. The models are used as discrete classifiers (i.e. each observation is assigned purely to a single class), and model performance is quoted in terms of Discrete Classification Accuracy (DCA) (i.e. the total proportion of correctly classified observations in the validation data).

The experiments in these papers are first repeated on the same data with trip-wise

sampling, using hyperparameters specified in the paper where possible. The experiments are then repeated a second time with the same hyperparameters, this time using grouped sampling for the cross-validation folds. The cross-validation results are then compared for trip-wise and grouped sampling.

Whilst Hillel et al. (2020) identify the use of purely discrete classification as a methodological pitfall, the classifiers in these studies have been optimised (using a grid-search) for discrete classification. Optimal hyperparameters are different for discrete and probabilistic classification, and certain discrete classifiers (e.g. SVM) must be specifically calibrated to give probability-like outputs. As such, only discrete metrics are reported for these studies.

In contrast, Study 3 performs a sequential hyperparameter using k-fold cross-validation with grouped sampling. The optimised models are then validated *externally* on a future year of data. The models are treated as probabilistic classifiers, and are evaluated in terms of their Cross-Entropy Loss (CEL),[6] with the DCA also provided for reference.

Study 3 only evaluates the GBDT model, and does not contain a benchmark LR model. As such, this study uses the modelling methodology from the paper to additionally estimate a LR benchmark. The sequential hyperparameter search is repeated twice for each classifier (GBDT and LR), once using trip-wise sampling and once using grouped sampling, in order to obtain two different model specifications for each classifier. Each optimised model is then validated on the same external sample. Both the cross-validation results and external validation results can therefore be compared, alongside the identified hyperparameters for each classifier.

All of the classifiers for all studies are implemented in Python, using *scikit-learn*, *xgboost*, and *hyperopt*.

4.1.1 Study 1

Five classifiers are implemented from Study 1: (i) LR (referred to as Multinomial Logit (MNL) in the paper), (ii) SVM, (iii) RF, (iv) AB, and (v) GBDT. The NL, fusion models, hybrid models, and feature selection methods are not tested. Reliable results are not found for the NB model, which is among the worst performing classifiers in Study 1, and so this classifier is also omitted.

The authors of the original paper use grid-searches to identify optimal hyperparameters for each classifier. However, the selected hyperparameters for each algorithm are not given in detail in the paper, with only certain hyperparameter values specified. Furthermore, the search spaces for the grid-searches are not specified and so it is not possible to repeat them. The analysis in the original paper makes use of the *python* library *scikit-learn*. As such, hyperparameters from the original paper are used where given, otherwise the default hyperparameter values for each classifier from *scikit-learn* are used. Full details of the hyperparameters used in this study for each classifier are given in Table 12 in Section A.1.

The authors do not mention whether features are normalised prior to model fitting. This should be applied for models which measure distances between points (e.g. SVMs) or apply spatial regularisation (e.g. L2 regularisation in LR). For this study, standard

[6]The CEL is equivalent to the negative log-likelihood - in ML it is typically normalised by diving by the size of the dataset to allow comparison between different dataset sizes.

normalisation (zero-mean, unit-variance) is applied to the input features for the LR and SVM classifiers. EL methods based on DTs are scale invariant, and so scaling is not applied for these classifiers.

Five-fold cross validation is used to evaluate the models in terms of the DCA, following the methodology used in the original paper. As the proposed models are only evaluated as discrete classifiers in the paper, only discrete metrics are reported in this study. The cross-validation is repeated using trip-wise and grouped household-wise sampling.

The paper has a further pitfall in the methodology, in that the modelling makes use of variables which are dependent on the mode choice in the feature vector. This includes the recorded trip duration and recorded trip expenses from the travel survey. The recorded duration and expenditures are both included explicitly as separate features in the dataset. The recorded duration is also implicitly included through the inclusion of both the arrival and departure times (i.e. the duration can be calculated from these values). This is a methodological pitfall, as the duration implicitly provides the model with the achieved speed of the trip (distance divided by duration), which is highly correlated with the mode choice (see Hillel et al. (2020)), and the recorded expenditure can easily be used to differentiate trips made by free modes (e.g. walking) from paid modes (e.g. public transport). Furthermore, these variables could not be known in advance of a trip being made, and so classifying based on these variables represents *mode identification* from the recorded data (as is frequently applied to GPS data where the travel mode is unknown) and not *mode choice prediction*. As such, the experiments for this paper are repeated twice; once using the original data from the paper, and once using a corrected set of features with the reported duration and departure times omitted.

4.1.2 Study 2

Six classifiers are implemented from Study 2: (i) LR (referred to as MNL in the paper), (ii) SVM, (iii) ANN (iv) GBDT (referred to as boosting in the paper), (v) bagging, and (vi) RF. Reliable results are not found for the NB model, which is among the worst performing classifiers in Study 2, and so this classifier is omitted.

The authors of the original paper use grid-searches to optimise the hyperparameters for each classifier. The search spaces and selected hyperparameters for the algorithms are given as supplementary material to the paper. However, only some hyperparameters are optimised, and values for the other hyperparameters are not given. The analysis in the original paper makes use of the *R* package *caret* to interface with the following packages: *nnet*, *klaR*, *ipred*, *e1071*, *randomForest*, and *gbm*. For the *scikit-learn* implementation in this study, hyperparameter values from the original paper are used where specified, otherwise hyperparameter values are chosen to match the default parameters from the respective *R* library. Full details of the hyperparameters used in this study for each classifier are given in Table 13 in Section A.2. An further explanation of the implementation of the LR and ANN classifiers is given in Section A.2.1.

The authors do not mention whether features are normalised prior to model fitting. For this study, standard normalisation (zero-mean, unit-variance) is applied to the input features for the LR, SVM, and ANN classifiers.

Ten-fold cross validation is used to evaluate the models in terms of the DCA, following the methodology used in the paper. This cross-validation is repeated using trip-wise and

grouped person-wise sampling.

As discussed, the methodology in the original paper uses a combined over/undersampling scheme to compensate for the imbalanced data by ensuring equal numbers of trips each mode. The resultant models are only validated on the sampled data using 10-fold cross-validation, and are never tested on a representative holdout sample. As discussed in Hillel et al. (2020), issues with imbalanced data arise from using discrete classification, where the *accuracy paradox* means that classifiers are unlikely to predict low probability classes. These issues should therefore be addressed by using probabilistic classification and proper continuous scoring metrics, such as the CEL. However, as the proposed models are evaluated only as discrete classifiers in the original paper, only discrete metrics are reported in this study.

To investigate the impacts of the sampling scheme, the experiments for this study are repeated twice; once using the over/undersampling methodology from the paper, and once using the original representative sample. The discrete mode-shares are then investigated for the for the two sampling schemes.

4.1.3 Study 3

In contrast to the previous studies which use trip-wise sampling and discrete classification, Study 3 makes use of grouped household-wise sampling and probabilistic classification.

For this study, the GBDT trained on the augmented data is re-implemented (referred to as the *choice-set* model in the original paper). In addition, the proposed modelling methodology is also used to estimate a reference LR model as a comparison. The data are normalised to zero-mean unit-variance for the LR classifier.

In this study, the final models are tested using external validation on a future year of data. Ten-fold cross-validation is instead used to select optimal model hyperparameters during Sequential Model-Based Optimisation (SMBO). As such, for this study, the hyperparameter optimisation is repeated using both grouped household-wise and trip-wise sampling for each classifier (therefore obtaining two different sets of hyperparameters for each classifier). The resulting classifiers are then tested on the external validation sample.

The same search space as the original paper is used for the GBDT model, given in Table 14 in Section A.3. The search space for the LR model is given in Table 15.

4.2 Results

4.2.1 Study 1

The repeated trip-wise results for Study 1 are shown alongside the results from the original paper in Table 4. Whilst there is relatively close agreement between the scores for the EL classifiers, the repeated results for the MNL and SVM classifiers are much higher than the original scores. This suggests features may not have been scaled for these classifiers in the original paper. The remaining differences in results are likely due to differences in the hyperparameters used; as discussed, only certain hyperparameter values are specified in the original paper and default values are used for the remaining hyperparameters (see Table 12). In particular, the default parameters in scikit-learn for the RF classifier allow the DTs in the ensemble to split until all leaf nodes belong purely to one class, whereas the

GBDT and AB classifiers have maximum depths of three and one respectively, restricting the flexibility of the ensemble.

Table 4: Trip-wise sampling modelling results (DCA) for 5-fold cross-validation on full dataset (including durations and expenses) alongside results from original paper for Study 1

Classifier	**Original (trip-wise)**	**Repeated (trip-wise)**
LR	0.533	0.671
SVM	0.366	0.734
RF	0.804	0.879
AB	0.588	0.615
GBDT	0.778	0.754

The results for grouped (household-wise) sampling using both the full and corrected datasets are shown alongside the repeated trip-wise results in Table 5. The group-wise sampling results are lower than the trip-wise sampling scores for all classifiers, showing that the data-leakage from trip-wise sampling identified in the dataset investigation results in biased performance estimates. In particular, the DCA scores for cross validation are substantially lower for the flexible non-linear classifiers (RF, GBDT, and SVM). This shows that these classifiers are able to easily overfit to the leaked data. The linear LR model and AB model with very shallow trees are not able to easily overfit to leaked data, and so perform comparatively worse with trip-wise sampling than their true performance with grouped sampling. As a result of the different effects of the data-leakage on the classifiers, the relative performance of the classifiers with the incorrect trip-wise sampling is not a reliable indicator of the performance on the corrected group-sampling scheme and the rankings of performance in this study are not consistent with the original paper.

Table 5: Grouped (household-wise) sampling modelling results (DCA) for 5-fold cross-validation on full dataset and corrected dataset alongside trip-wise sampling results for Study 1

Classifier	**Trip-wise**	**Grouped**	**Grouped with corrected data**
LR	0.671	0.639	0.582
SVM	0.734	0.534	0.498
RF	0.879	0.663	0.611
AB	0.615	0.601	0.581
GBDT	0.754	0.619	0.515

The modelling results for the corrected data (i.e. with recorded trip durations and costs removed), shown in the last column of Table 5, are lower again for all classifiers, showing that all classifiers were able to fit to the input features which are dependent on the output mode choice.

Overall, the combination, the use of trip-wise sampling and including features dependent on the mode choice results in heavily biased performance estimates for the flexible

non-linear ML models, with a difference in the reported accuracy of the RF model of over 26 percentage points between trip-wise sampling with the full dataset and grouped sampling with the corrected dataset. As such, as shown by Fig. 2, the true differences in performance estimates are much closer for the different models than is suggested by the original paper. The original results show a difference of over 27 percentage points between the DCA of the RF and LR models, whereas this difference is less than 3 percentage points for the grouped sampling results with corrected and appropriately scaled data.

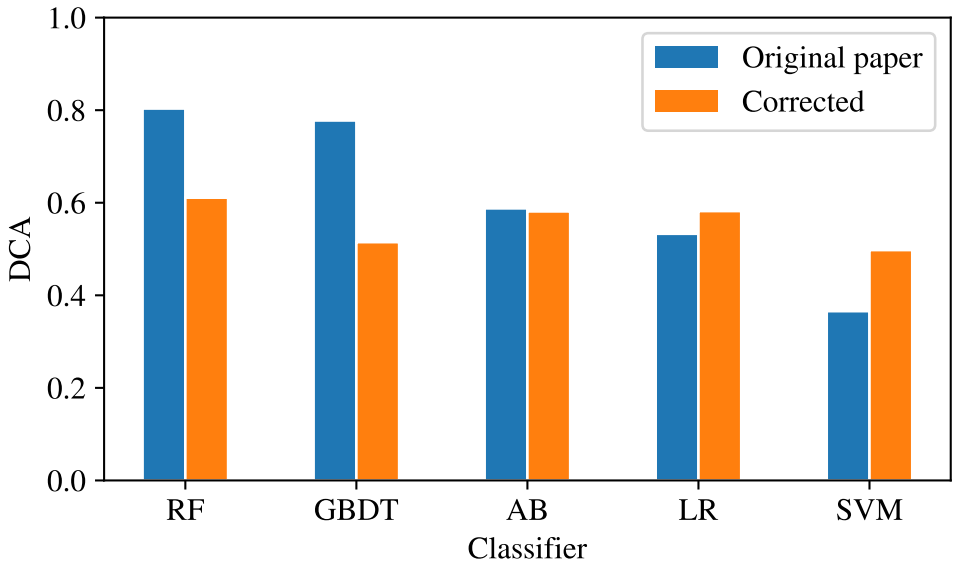

Figure 2: Comparison of results from original paper (using trip-wise sampling) and corrected results (using grouped sampling and corrected dataset) for each classifier in Study 1

4.2.2 Study 2

The trip-wise results for Study 2 are shown for the sampled dataset alongside the original results from the paper in Table 6. There is close agreement between the original results and repeated results for all classifiers. This suggests that the classifiers have been accurately reimplemented in this study, and that feature normalisation was used for the LR, SVM, and ANN classifiers in the original paper.

The results for grouped (person-wise) sampling using the sampled dataset are shown alongside the repeated trip-wise results in Table 7. As with Study 1, the use of trip-wise sampling on the panel results in drastically inflated performance estimates for the non-linear EL and SVM classifiers. The use of grouped sampling results in a reduction of the apparent DCA of the RF model of over 30 percentage points compared to the incorrect trip-wise sampling, whereas it has no discernible impact on the results for the LR model.

The corrected results from this study are shown alongside the results from the original paper in Fig. 3. As with Study 1, there are much smaller differences in the corrected results compared to the original paper, with nearly identical performance of the LR and RF models with grouped sampling, compared to a difference of over 35 percentage points

Table 6: Trip-wise sampling modelling results (DCA) for 5-fold cross-validation on sampled dataset alongside original results from paper for Study 2

Classifier	Original (trip-wise)	Repeated (trip-wise)
LR	0.561	0.578
SVM	0.825	0.799
ANN	0.606	0.613
GBDT	0.801	0.814
BAG	0.906	0.904
RF	0.914	0.914

Table 7: Grouped (person-wise) sampling modelling results (DCA) for 5-fold cross-validation on sampled dataset alongside trip-wise sampling results for Study 2

Classifier	Trip-wise	Grouped
LR	0.578	0.577
SVM	0.799	0.452
ANN	0.613	0.609
GBDT	0.814	0.614
BAG	0.904	0.573
RF	0.914	0.583

in the original paper. As before, the rankings of performance are also not consistent between the sampling schemes, with the GBDT model achieving the highest DCA on the sampled data with the given hyperparameters, instead of the RF model for the trip-wise data.

Impact of sampling scheme for imbalanced data The grouped sampling results for both the sampled and unsampled data are shown in Table 8. The results appear to suggest that the models perform better on the unsampled data, despite there being less data leakage from repeated trips (see Table 3), though this is in fact a result of discrete classification and the *accuracy paradox*; as there is higher class imbalance in the unsampled data, the model can achieve a higher accuracy by simply predicting the majority class.

To illustrate this further, Table 9 shows the predicted mode shares for the GBDT classifier (the highest performing classifier with grouped sampling) for both the sampled and unsampled data. Whilst the accuracy score of the GBDT classifier is higher for the unsampled data, the mode shares show that this is a result of the model substantially over-predicting car trips, and under-predicting the less likely modes (in particular public transport).

As shown by Table 9, discrete classification is unable to generate representative mode shares, even when the data has been sampled to be completely balanced for each mode.[7]

[7]Whilst results for all classifiers are not given in this paper for conciseness, none of the tested classifiers are found to give representative mode-shares with discrete classification for either the sampled or unsampled data, and the GBDT is found to be the highest performing model in terms of predicted mode shares.

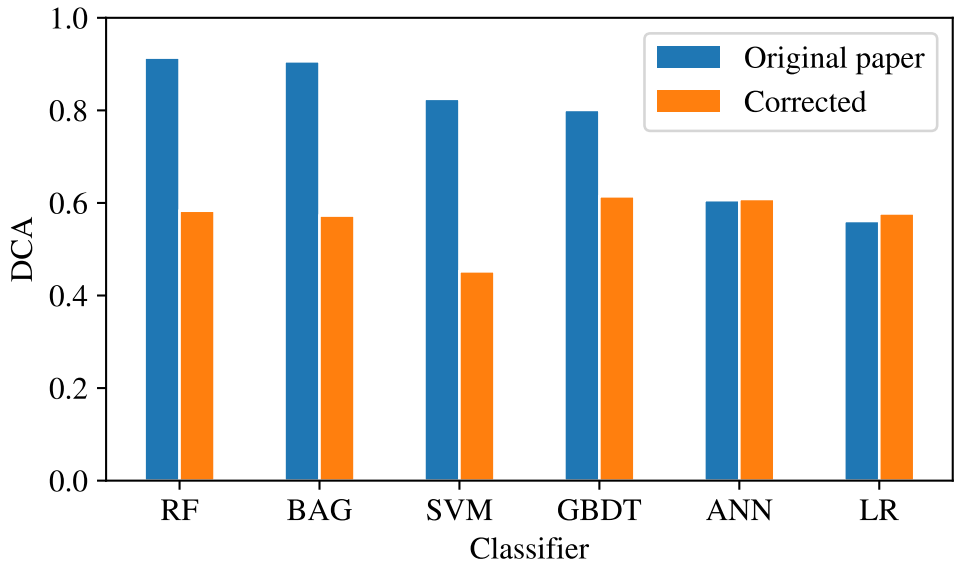

Figure 3: Comparison of results from original paper (using trip-wise sampling) and corrected results (using grouped sampling) for each classifier in Study 2

Table 8: Grouped sampling modelling results (DCA) for 5-fold cross-validation on sampled and unsampled dataset for Study 2

Classifier	**Sampled data**	**Unsampled data**
LR	0.577	0.659
SVM	0.452	0.605
ANN	0.609	0.670
GBDT	0.614	0.688
BAG	0.573	0.678
RF	0.583	0.687

This shows that discrete classification is not appropriate for mode choice prediction, or any application where representivitiy of the output is of importance. Instead, probabilistic classification should be used with strictly proper scoring metrics such as CEL (Hillel, Elshafie, and Jin 2018; Gneiting and Raftery 2007).

Table 9: Predicted and true mode-shares and ratios for the sampled and unsampled datasets for the ANN classifier for Study 2

	Sampled			Unsampled		
	Predicted	**True**	**Ratio**	**Predicted**	**True**	**Ratio**
Walk	52 360	57 652	0.908	28 933	37 571	0.770
Bike	62 137	57 652	1.078	51 043	56 298	0.907
PT	42 158	57 652	0.731	5494	9300	0.591
Car	73 953	57 652	1.283	145 138	127 439	1.139

4.2.3 Study 3

The CEL and DCA for the repeated grouped sampling GBDT model are shown alongside the original results from the paper in Table 10. The cross-validation scores show the predicted OOS performance using 10-fold cross-validation with grouped (household-wise) sampling of the classifiers optimised with a SMBO search also using grouped sampling. The training data (first two years of data) is used for both the model optimisation and the cross-validation performance estimation. The external validation results then show the performance of the model with the hyperparameters identified in the SMBO search with grouped sampling when estimated on all of the training data and used to predict an unseen future year of data (i.e. data collected separately from the training data). There is a close agreement between the original results and the repeated results in this study, suggesting the repeated hyperparameter search has found similar values to the original paper. This is confirmed by the hyperparameter values for each classifier shown in Table 16 in Section A.3; the repeated grouped sampling model has matching maximum depth and a similar number of trees in the ensemble to the original paper. The results in Table 10 indicate that the grouped sampling cross-validation results are a good indicator of true external validation performance, as there are only small differences (approximately 0.015 difference in CEL and one percentage point of DCA) between the cross-validation and external validation results.

Table 10: Grouped sampling modelling results (CEL and DCA) for 10-fold cross-validation and external validation alongside original results from paper for Study 3

Classifier	**Validation**	**Metric**	**Original (grouped)**	**Repeated (grouped)**
GBDT	Cross-validation	CEL	−0.634	−0.635
		DCA	0.759	0.758
	External validation	CEL	−0.651	−0.651
		DCA	0.748	0.748

The grouped and trip-wise sampling results for the GBDT and LR classifiers are shown in Table 11. An explanation of the grouped sampling results is given above. For the trip-wise sampling results, the SMBO search is performed using trip-wise validation. As such, the resultant models have different hyperparameters to the grouped sampling models. The different hyperparameters are shown in Tables 16 and 17 in Section A.3. Note that the parameters identified with trip-wise sampling have lower regularisation than with grouped sampling (i.e. they allow the model to have more flexibility with higher variance and lower bias). For instance, the trip-wise sampling GBDT model has a much higher maximum depth (13 vs 6) and a greater number of trees in the ensemble (2547 vs 1446) and the trip-wise LR model has a higher C value, indicating less regularisation.[8] This allows the trip-wise models to more easily overfit to the data-leakage identified in the dataset investigation.

Table 11: Grouped and trip-wise modelling results (CEL and DCA) for 10-fold cross-validation and external validation for Study 3

Classifier	Validation	Metric	Grouped	Trip-wise
GBDT	Cross-validation	CEL	−0.634	−0.467
		DCA	0.759	0.830
	External validation	CEL	−0.651	−0.730
		DCA	0.748	0.747
LR	Cross-validation	CEL	−0.679	−0.676
		DCA	0.738	0.739
	External validation	CEL	−0.693	−0.693
		DCA	0.736	0.736

As with Studies 1 and 2, comparing the grouped and trip-wise sampling cross-validation results indicates that trip-wise sampling provides heavily biased performance estimates for the highly non-linear GBDT classifier, whilst having very little visible effect on the linear LR model. This is confirmed by the external validation results; whilst the differences between the cross-validation and external validation scores for the GBDT model are small for grouped sampling, they are much larger for trip-wise sampling.

Further to the bias in performance estimates introduced by trip-wise sampling, the external validation results reveal that the GBDT model optimised using trip-wise sampling actually performs worse for true OOS prediction. This shows that using trip-wise sampling during model optimisation will introduce a bias towards high-variance models (which can overfit to the leaked data) that perform worse in practice.

As with the previous studies, the performance differences between EL models and the linear LR model indicated by trip-wise sampling are much larger than the true performance indicated by grouped sampling.

4.2.4 Summary

As shown by the repeated modelling experiments from each study, trip-wise sampling with panel data results in significantly biased performance estimates for non-linear clas-

[8] C is the inverse of the regularisation strength, i.e. the lower the value, the stronger the regularisation.

sifiers, whilst having much lower impact on results of linear models. As such, the true performance differences between non-linear ML classifiers and linear models is much lower when using grouped sampling than suggested by trip-wise sampling results. This is illustrated by Fig. 4, which shows the absolute difference in DCA between the highest performing classifier and the linear LR model for each study, using both trip-wise and grouped sampling.

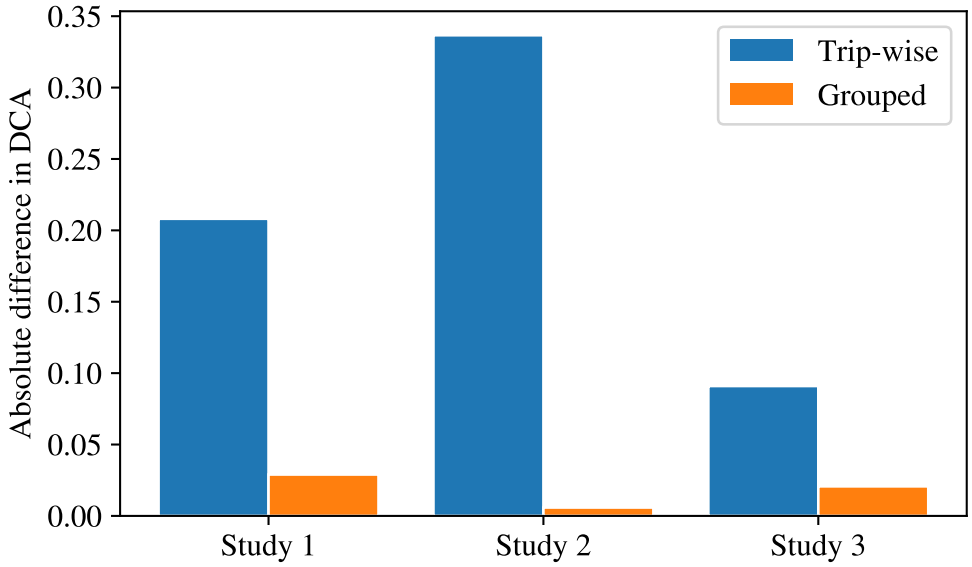

Figure 4: Apparent absolute performance difference (in percentage points of DCA) between highest performing classifier* and linear LR model for each study using trip-wise and grouped sampling.
* Highest performing classifier from original paper for each study: Study 1 - RF, Study 2 - RF, Study 3 - GBDT.

5 Conclusions

This paper conducts an experimental review of sampling methods for ML model validation, using the data and methodology of three published studies which investigate ML approaches for mode choice prediction.

Firstly, the datasets from each study are analysed to identify the proportion of matching trips, including repeated, return, and shared trips. Monte-carlo simulation of different validation schemes is used to identify that for all the datasets, trip-wise sampling results in over half of the validation data having matching trips with the same mode choice in the training data, and for one dataset this figure is well over 90%. This indicates that there is substantial data leakage from the training data to the validation data when using trip-wise sampling with panel data, therefore violating the principle of OOS validation.

The experiments in each study are then repeated using both trip-wise and grouped sampling. For all three studies, the repeated results indicate that flexible non-linear classifiers, including EL models and SVMs with non-linear kernels, are able to easily overfit

to the leaked data from trip-wise sampling. In the case of two of these studies, this has lead to incorrect conclusions from the results, with the authors suggesting there are substantial performance benefits (over 20 and 35 percentage points difference in DCA respectively for Study 1 and 2) of non-linear ML models when compared to RUMs (represented in these studies by the linear LR model). As identified by a recent systematic review of ML techniques for mode choice modelling, this is a wide-spread issue, with all included studies which make use of complete trip-diary data using trip-wise sampling.

As revealed by the grouped sampling and external validation results, whilst there is a slight predictive performance advantage of non-linear classifiers, and in particular EL models, in reality the differences are much more modest (< 3 percentage points of DCA for all studies) than has been suggested by previous investigations.

Two further issues are investigated in this study, (i) the use of input features dependent on output mode choice, and (ii) the use of discrete classification. These issues are found to introduce bias to model performance estimates and/or result in incorrect model outputs (e.g. mode-shares).

It is hoped this study will help researchers and practitioners to evaluate the relative advantages and disadvantages of ML approaches more fairly against traditional RUMs. Whilst the results may indicate that the advantage in predictive performance of ML classifiers may be more limited than initially thought, they do also highlight that there is a consistent marginal advantage of these techniques over linear models across the three repeated studies. From the prevalence of the issues investigated in this paper in the literature, there is a clear need for a deeper understanding of how these techniques should be applied. As such, planned further work includes developing a standardised modelling framework which covers both ML and RUM approaches for choice modelling.

As a result of the modest predictive performance advantages of ML classification techniques identified in this study, it is the author's belief that future studies should focus on how to use ML methodologies to augment and improve the random-utility approach, for example through assisted and/or automated utility specification, rather than on replacing RUMs with ML classifiers.

Acknowledgements

This work was supported by the Transport and Mobility Laboratory at EPFL. The author would like to thank Michel Bierlaire for providing valuable comments on the manuscript.

Appendices

A Hyper-parameter values

A.1 Study 1

Table 12: Hyperparameters used for each classifier for Study 1

Classifier	Parameters
LR	`LogisticRegression()` and `StandardScaler()`
specified:	`C=1`
default:	`penalty='l2', dual=False, tol=0.0001, fit_intercept=True, class_weight=None, solver='lbfgs', max_iter=100, multi_class='multinomial'`
SVM	`SVC()` and `StandardScaler()`
specified:	`kernel='rbf'`
default:	`C=1.0, gamma='scale', shrinking=True, probability=False, tol=0.001, class_weight=None, decision_function_shape='ovr'`
RF	`RandomForestClassifier()`
specified:	`n_estimators=100, criterion='gini'`
default:	`max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, class_weight=None, ccp_alpha=0.0, max_samples=None`
AB	`AdaBoostClassifier()`
specified:	`n_estimators=100, learning_rate=0.1, base_estimator=DecisionTreeeClassifer()`
default:	`algorithm='SAMME.R', max_depth=1`
GBDT	`GradientBoostingClassifier()`
specified:	`n_estimators=100, loss='deviance', learning_rate=0.5`
default:	`subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=3, min_impurity_decrease=0.0, min_impurity_split=None, max_features=None, max_leaf_nodes=None, n_iter_no_change=None, ccp_alpha=0.0`

A.2 Study 2

Table 13: Hyperparameters used for each classifier for Study 2.
* value is modified to match *R* default parameters.

Classifier	Parameters
LR	`MLPClassifier()` and `StandardScaler()`
specified:	`hidden_layer_sizes=()`, `alpha=0,001`, `solver='lbfgs'`
default:	`max_iter=100*`, `tol=0.0001`
SVM	`SVC()` and `StandardScaler()`
specified:	`kernel='rbf'`, `C=1.25`, `gamma=0.4`, `decision_function_shape='ovo'`
default:	`shrinking=True`, `probability=False`, `tol=0.001`, `class_weight=None`
ANN	`MLPClassifier()` and `StandardScaler()`
specified:	`hidden_layer_sizes=(48,)`, `alpha=0.1`, `activation='logistic'`, `solver='lbfgs'`
default:	`max_iter=100*`, `tol=0.0001`
GBDT	`GradientBoostingClassifier()`
specified:	`n_estimators=300`, `learning_rate=0.2`, `max_leaf_nodes=49`, `max_depth=None`, `min_samples_leaf=10`
default:	`loss='deviance'`, `subsample=0.5*`, `criterion='friedman_mse'`, `min_samples_split=2`, `min_weight_fraction_leaf=0.0`, `min_impurity_decrease=0.0`, `min_impurity_split=None`, `max_features=None`, `max_leaf_nodes=None`, `n_iter_no_change=None`, `ccp_alpha=0.0`
BAG	`BaggingClassifier()`
specified:	`n_estimators=350`, `base_estimator=DecisionTreeeClassifer(max_depth=None)`
default:	`max_samples=1.0`, `max_features=1.0`, `bootstrap=True`, `bootstrap_features=False`, `oob_score=False`
RF	`RandomForestClassifier()`
specified:	`n_estimators=450`, `max_features=3`
default:	`criterion='gini'`, `max_depth=None`, `min_samples_split=2`, `min_samples_leaf=1`, `min_weight_fraction_leaf=0.0`, `max_leaf_nodes=None`, `min_impurity_decrease=0.0`, `min_impurity_split=None`, `bootstrap=True`, `oob_score=False`, `class_weight=None`, `ccp_alpha=0.0`, `max_samples=None`

A.2.1 Study 2 - hyperparameter notes

The LR and ANN models are estimated using the *nnet R* package in the original study. This package uses the *BFGS* optimisation algorithm, and so the equivalent solver (*lbfgs*) is used in *scikit-learn* (Venables and Ripley 2002). The *sigmoid* activation function is used for the hidden layer in *nnet* (Venables and Ripley 2002), which is referred to as *logistic* in *scikit-learn*. The hyperparameter for *weight-decay* is equivalent to the *alpha* parameter for L2 regularisation when using the *BFGS* algorithm.

The `interaction.depth` parameter in the *gbm R* library specifies the total number of permitted splits, and is not equivalent to the `max_depth` hyperparameter in *scikit-learn*, which limits the total permitted layers in the tree. As each split in a binary tree adds one leaf node to the tree, and the tree starts with a single node, the *interaction.depth* of 48 in *gbm* is equivalent to a maximum number of leaf nodes of 49 in *scikit-learn*

A.3 Study 3

Table 14: Hyperparameter search space for GBDT model for Study 3

Hyper-parameter	**Distribution**	**Range**
`max_depth`	Uniform	1–14
`gamma`	Log-uniform	0–5
`min_child_weight`	Log-uniform (int)	1–100
`max_delta_step`	Log-uniform (int)	0-10
`subsample`	Uniform	0.5–1
`colsample_bytree`	Uniform	0.5–1
`colsample_bylevel`	Uniform	0.5–1
`reg_alpha`	Log-uniform	0–1
`reg_lambda`	Log-uniform	1–4
`learning_rate`	Fixed	0.01
`n_estimators`	See methodology	1–6000
`extra_stopping_rounds`	Fixed	50

Table 15: Hyperparameter search space for LR model for Study 3

Hyper-parameter	**Distribution**	**Range**
`penalty`	Uniform choice	L1, L2
`C`	Log-uniform	1e−5–1e5
`class_weight`	Uniform choice	None, *balanced*
`solver`	Fixed	*saga*
`multi_class`	Fixed	*multinomial*

Table 16: Optimised hyperparameter values for GBDT for Study 3

Hyper-parameter	**Original (grouped)**	**Grouped**	**Trip-wise**
max_depth	6	6	13
gamma	5.439e−3	2.596e−3	0.03855
min_child_weight	36	26	1
max_delta_step	4	2	9
subsample	0.65	0.5	0.9
colsample_bytree	0.65	0.5	0.55
colsample_bylevel	0.55	0.85	0.8
reg_alpha	4.823e−4	0.3747	0.03022
reg_lambda	2.572	2.4	3.473
n_estimators	1472	1446	2547

Table 17: Optimised hyperparameter values for LR model for Study 3

Hyper-parameter	**Grouped**	**Trip-wise**
penalty	L1	L1
C	0.7715	1.228
class_weight	None	None

References

Biagioni, James P., Piotr M. Szczurek, Peter C. Nelson, and Abolfazl Mohammadian (2008). "Tour-Based Mode Choice Modeling: Using an Ensemble of (Un-) Conditional Data-Mining Classifiers". In:

Cantarella, Giulio Erberto and Stefano de Luca (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.

Chalasani, V. Saikumar and Kay W. Axhausen (May 2004). *Mobidrive: A Six Week Travel Diary*. Working Paper. IVT, ETH Zurich.

Chang, Ximing, Jianjun Wu, Hao Liu, Xiaoyong Yan, Huijun Sun, and Yunchao Qu (Nov. 29, 2019). "Travel Mode Choice: A Data Fusion Model Using Machine Learning Methods and Evidence from Travel Diary Survey Data". In: *Transportmetrica A - Transport Science* 15.2, pp. 1587–1612.

Chapleau, R., P. Gaudette, and T. Spurr (2019). "Application of Machine Learning to Two Large-Sample Household Travel Surveys: A Characterization of Travel Modes". In: *Transportation Research Record: Journal of the Transportation Research Board* 2673.4, pp. 173–183.

Cheng, Long, Xuewu Chen, Jonas De Vos, Xinjun Lai, and Frank Witlox (Jan. 2019). "Applying a Random Forest Method Approach to Model Travel Mode Choice Behavior". In: *Travel Behaviour and Society* 14, pp. 1–10.

Ding, Chuan, Xinyu Cao, and Yunpeng Wang (Dec. 2018). "Synergistic Effects of the Built Environment and Commuting Programs on Commute Mode Choice". In: *Transportation Research Part A - Policy and Practice* 118, pp. 104–118.

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.

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, Ramin Shabanpour, Seyed Mehdi Mahmoudifard, Sybil Derrible, and Abolfazl Mohammadian (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.

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.

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.

Hillel, Tim, Michel Bierlaire, Mohammed ZEB Elshafie, and Ying Jin (2020). "A Systematic Review of Machine Learning Classification Methodologies for Modelling Passenger Mode Choice". In: *Journal of Choice Modelling* (Forthcoming).

Hillel, Tim, Mohammed ZEB Elshafie, and Ying Jin (2018). "Recreating Passenger Mode Choice-Sets for Transport Simulation: A Case Study of London, UK". In: *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction* 171.1, pp. 29–42.

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

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.

Liang, LeiLei, Meng Xu, Susan Grant-Muller, and Lorenzo Mussone (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.

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.

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.

Nam, Daisik, Hyunmyung Kim, Jaewoo Cho, and R. Jayakrishnan (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.

Omrani, Hichem, Omar Charif, Philippe Gerber, Anjali Awasthi, and Philippe Trigano (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 (Jan. 2, 2019). "A Study of Tour-Based Mode Choice Based on a Support Vector Machine Classifier". In: *Transportation Planning and Technology* 42.1, pp. 23–36.

Pitombo, Cira Souza, Aline Schindler Gomes Da Costa, and Ana Rita Salgueiro (2015). "Proposal of a Sequential Method for Spatial Interpolation of Mode Choice". In: *Boletim de Ciências Geodésicas* 21.2, pp. 274–289.

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.

Rasouli, Soora and Harry J.P. Timmermans (2014). "Using Ensembles of Decision Trees to Predict Transport Mode Choice Decisions: Effects on Predictive Success and Un-

certainty Estimates". In: *European Journal of Transport and Infrastructure Research* 14.4, pp. 412–424.

Saeb, Sohrab, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording (May 1, 2017). "The Need to Approximate the Use-Case in Clinical Machine Learning". In: *GigaScience* 6.5, pp. 1–9.

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.

Shukla, Nagesh, Jun Ma, Rohan Wickramasuriya Denagamage, and Nam N. Huynh (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.

Subba Rao, P. V., P. K. Sikdar, K. V. Krishna Rao, and S. L. Dhingra (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.

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.

Venables, W. N. and B. D. Ripley (2002). "Non-Linear and Smooth Regression". In: *Modern Applied Statistics with S*. Ed. by W. N. Venables and B. D. Ripley. Statistics and Computing. New York, NY: Springer, pp. 211–250.

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, Shenhao and Jinhua Zhao (2019). "An Empirical Study of Using Deep Neural Network to Analyze Travel Mode Choice with Interpretable Economic Information". In: Transportation Research Board 98th Annual Meeting.

Yang, Jie and Jun Ma (2019). "Compressive Sensing-Enhanced Feature Selection and Its Application in Travel Mode Choice Prediction". In: *Applied Soft Computing* 75, pp. 537–547.

Zhou, Xiaolu, Mingshu Wang, and Dongying Li (July 2019). "Bike-Sharing or Taxi? Modeling the Choices of Travel Mode in Chicago Using Machine Learning". In: *Journal of Transport Geography* 79, UNSP 102479.

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