Utilizing a random forest classifier for a methodological-iterative discrete choice model specification and estimation

Yuval Shiftan
Department of Civil and Environmental Engineering,
Technion – Israel Institute of Technology, Haifa 32000, Israel
Email: yuvalshiftan@campus.technion.ac.il

Shlomo Bekhor (corresponding author)
Department of Civil and Environmental Engineering,
Technion – Israel Institute of Technology, Haifa 32000, Israel
Email: sbekhor@technion.ac.il

ABSTRACT

This paper proposes a novel approach to discrete choice model estimation, which we term a methodological-iterative (MI) approach. This approach utilizes the results of an ensemble learner, in this case the random forest classifier, to incorporate an algorithm of recursive feature elimination to utility function specification and model estimation. The method is applied to a case study of car allocation among household members.

The MI method reduces the model specification efforts required considerably, can reveal significant explanatory variables that may be overlooked, and may prove helpful when the modeler lacks the behavioral insight of the dataset or the phenomenon being investigated or when the dataset is very large, as it is estimated by an algorithm rather than by the conventional trial and error approach. The method is a hybrid tool of data-driven machine learning and a theory-driven discrete choice model to obtain an improved model and forecast. As such, it harnesses some of the advantages of a powerful data-driven machine learning classifier while obtaining the interpretability of discrete choice models.

The MI method is implemented in this paper to investigate car allocation choice behavior. The MI car allocation choice model produced a forecast that is similar to the data-driven machine learning classifiers prediction wise, but with full interpretability.

Keywords: car allocation; car deficient household; discrete choice, random forest; feature importance, methodological-iterative.

INTRODUCTION

Choice behavior has been studied extensively in many disciplines, while the prevailing methodology has been, for many years, the discrete choice models (especially the multinomial and nested logit models) that are based on the random utility theory. These models are aimed at predicting the choice of a decision-maker among a set of mutually exclusive and collectively exhaustive alternatives. They provide forecast to inform policy and marketing decisions while understanding the behavioral process that led to the specific choice. Although these models proved to be extremely productive, they require an a-priori knowledge of the behavioral processes to specify the model, and specifically, the utility functions. The specification of these utility functions is a difficult, time consuming, prone to error task, which usually requires many iterations of specifications-estimations-evaluations by a trial and error method.

The study of choice behavior and forecasting has also drawn on the interdisciplinary methods of machine learning that have also proved to be extremely productive. These methods usually focus on prediction rather than interpretability, which relies on patterns and inference as opposed to explicit instructions (or explicit specification). Data-driven machine learning algorithms have been very successful in many domains, including transportation research. A motivation of this paper is to harness advantages from the data-driven machine learning discipline toward the challenging task of utility specification within the random utility theory.

We suggest in this paper a novel methodological-iterative (MI) algorithm to the discrete choice model estimation that utilizes the results of the random forest (i.e., feature importance). We apply it to the question of car allocation as a case study. The MI method allows a more efficient model estimation by reducing the time spent on the full model building process and can offer a starting point to model-specification, or a "safety net" that can help find statistically significant explanatory variables. Such variables may be overlooked in the process of model estimation through trial and error of many variables. Moreover, the MI approach does not only assist in finding overlooked variables, but it can find different sets of relationships between the explanatory variables, that the modeler missed or did not consider, as there can be an enormous alternative for such relationships. Detecting these relationships can be a challenging task, especially when the modeler does not have sufficient insight into the behavior process at hand.

The MI approach's goal is to be a tool that helps the modeler in achieving the best possible outcome - not necessarily to find the best outcome automatically. Nonetheless, in this paper, we purposely show the results of the MI method without further enhancing the model to evaluate its performance.

This paper also contributes to the understanding of car allocation choice behavior among household members, as it uses car allocation as a case study. We wish to fill some of the gaps in the literature by investigating the car allocation choice behavior among a wide range of household types and household members. In this aspect, we investigate the factors that influence these household-level decisions, such as the mobility needs of different family members, the family structure within the household, and the influence that each household member has on household decisions. These dynamics between household members affect car allocation, specifically in car deficient households, which are defined here as households with at least one car and more licensed drivers than cars. To this end, we analyze a comprehensive GPS-assisted household survey conducted in the Tel-Aviv metropolitan area, harnessing the advantages of a rich and detailed revealed preference database.

This paper is organized as follows. The next section summarizes the literature on car allocation models and the relevant machine learning and feature selection methods. The subsequent section presents the MI method and the data used for the analysis. The MI method is applied to a case study of car allocation, and estimation results are presented. Then, we analyze the variables found significant in the MI method. We then compare the predictive strength of the random forest and the MI method, and their strengths and weaknesses. The final part of the paper summarizes the main findings and outlines directions for further research.

LITERATURE REVIEW

Car allocation models

The importance of interactions among household members has been recognized in the literature, as these interactions affect the individuals' behavior, (as indicated, for example, by Bhat & Pendyala (2005), Timmermans & Zhang (2009), Ho & Mulley (2015), Munro (2018)). The focus of interaction between household members has been related primarily to the choice of activities (Gliebe & Koppelman, 2005), (Bradley & Vovsha, 2005) and (Maat & Timmermans, 2009). The specific question of how the relations among family members affect mode choice, and particularly the decision to whom to allocate the available household vehicles, has not received the same amount of attention (Ho & Mulley, 2015).

Car allocation has been researched within the random utility model (RUM) framework. Habib (2014) applied a nested logit model to examine commuters' mode choices and car ownership level choices of two-worker households in the context of reverse commuting in the City of Toronto. The author considered single-car homes and finds that female commuters, reverse commuters, and longer-distance commuters are more likely to get the car. In terms of personal attributes, only age and gender entered the car allocation choice model. Arman et al. (2015) developed a model for the joint activity and vehicle allocation of households, considering one-vehicle households. They applied a

Paired Combinatorial Logit model to estimate the share of each household member in the non-mandatory activities, and a multinomial logit to assess vehicle allocation. They found that employment and the presence of a work trip in the daily schedule increase car allocation probability, while homemakers had lower car allocation probability.

Roorda (2006) incorporated within-household interactions into a mode choice model, where the allocations decisions were made at the household level to maximize household utility. Maat & Timmermans (2009) applied binary and multinomial logit models to investigate the influence of the residential and work environment on car use in dual-earner households. They found that in one-car households, all significant parameters concern men, suggesting that the men's characteristics and circumstances are dominant when determining who uses the car for commuting. Moreover, the likelihood of men taking the car was lower if there are young children in the household, indicating that their partners had more responsibility for activities concerning the children.

Kalter & Geurs (2016) employed a multilevel binary logit model to examine the impact of interactions between household members on car use at the tour, individual, and household level, based on two waves of panel data in the Netherlands. They found that these interactions were indeed significant as they account for more than one-third of the total variation in mode choice between household accounts and individual accounts. Weiss & Habib (2020) developed a two-adult household mode choice model (multinomial logit, nested logit, and a mixed nested logit) aimed to capture the trade-offs concerning vehicle allocation and joint travel opportunities. The authors found, among other conclusions, that car deficient households were more likely to travel jointly.

Machine learning methods

Besides the widely used approach of random utility theory, data-driven machine learning methods have been used for a growing number of transportation prediction and classification problems. They are generally classified into four categories: travel choice behavior, traffic incident prediction, traffic time/flow prediction, and pattern recognition (Cheng et al., 2019). For example, Paredes et al., (2017) found that random forest and support vector machines outperformed discrete choice models for the prediction of car ownership. Lee, et al. (2018) found that the artificial neural networks (ANN) are also superior to Multinomial Logit (MNL) regarding mode choice. Other authors established the same result regarding the question of travel mode choice, where random forest performed significantly better than other classifiers, including multinomial logit (Sekhar et al., 2016), (Hagenauer & Helbich, 2017), (Cheng et al., 2019).

The use of these data-driven machine learning techniques holds great promises. Only a limited number of studies focus on the explicit representation of household car allocation decisions in car deficient households using machine learning algorithms. A notable exception is the work of

Anggraini et al. (2008). The authors considered car allocation choice behavior in car-deficient households with two heads and one car, explicitly in the context of an activity-scheduling process, focusing on work activities. The authors applied a decision tree induction method and found that the probability of taking the car to work was considerably higher for men (especially when women have no work activity) and that socio-economic variables had only a limited impact on car allocation. They also found that the probability of the male getting the car decreases as his work duration increases. The same authors also studied the case of non-work tours and found similar results (Anggraini et al., 2012).

Another notable example is the use of cluster regression to study car use in car deficient households from a gender perspective (Scheiner & Holz-Rau, 2012). The authors found that the gendered roles a person takes in the household may impact car allocation decisions, and they did not find support for the claim that car access is a function of intra-household economic power. Notably, having small children in the family decreased men's, and increased women's car allocation probability. Elder partners tended to use the car less, and education level was negatively correlated to car use.

Harnessing the advantages of machine learning algorithms to discrete choice models has recently yielded some novel approaches, specifically in the domain of artificial neural networks (ANN). Sifringer et al. (2018) proposed a "learning multinomial logit," in which the systematic part of the utility specification is divided into an interpretable part and a non-linear learning representation part arising from neural networks. The authors show that the learning model outperforms traditional multinomial logit models. Extending this work, Han et al. (2020) proposed a neural network-embedded multinomial logit. The method employed a neural network to model tastes as flexible functions of individuals characteristics that are then integrated into a parametric MNL to compute the choice probability

Feature selection

Another related research field in the domains of statistics and machine learning is feature selection, which is the process of selecting a subset of relevant features for use in model estimation. Feature selection is used for several purposes, such as reducing overfitting and improving accuracy (Chandrashekar & Sahin, 2014). Feature selection methods can be broadly distinguished into four categories: wrappers, filters, embedded, and hybrid methods (Jović et al., 2015). Hybrid methods were proposed to combine the best properties of wrapper and filter methods (Das, 2001). Filters select features based on a performance measure regardless of the employed model, and they are usually fast to implement and less prone to overfitting. Wrappers consider subsets of variables by the model's performance, which allows, unlike filter methods, to detect the possible interactions between variables, but at the cost of higher computation time and a higher risk of overfitting (Liu & Motoda, 2012).

Feature selection has been investigated and utilized in transportation research. To name only a few examples, Yang (2013) predicted traffic congestion using a filter method of the p-test score and then a wrapper-like method to select the optimal number of features. Lee & Wei (2010) predicted freeway accident duration time using a wrapper method of genetic algorithms. Ou et al. (2017) used a multilevel filter approach, including permutation importance, for short-term traffic flow forecasting. Barsky et al. (2018) achieved real-time congestion forecasting using a random forest wrapper method. Finally, Yang & Ma (2019) developed a feature selection algorithm based on compressive sensing to predict mode choice.

Recently, Hillel et al., (2019) proposed an assisted procedure to the utility function specification using an ensemble learning model, the gradient boosting decision trees (GBDT). The authors investigated the structure of the GBDT model, estimate the feature importance, and enhance the manually estimated MNL. More specifically, the authors investigate the important features, and by analyzing their KDE plots gain insight into the transformation that is most appropriate, and manually include it in the model. This paper also uses the feature importance of an ensemble learner but utilizes it to a fully systematic procedure. The MI method proposed here not only uses the ensemble learner to gain insight on the utility specification but utilizes it to estimate the choice model completely. Therefore, specification time is reduced since it does not require a manually estimated model at its base or manual investigation and modifications and can consider a wider array of possible interactions between the explanatory variables.

While recursive feature elimination is a popular tool in the fields of statistics and machine learning (Jović et al., 2015), to the best of our knowledge, a recursive, ensemble learning-based feature elimination procedure has not been applied for a full random utility model estimation. We speculate that the reason might be that there is an added complexity in discrete choice models, where there are numerous utility functions. As such, the present case study includes a relatively large number of alternatives and explanatory variables, as will be shown next.

METHODOLOGY

MI approach

This section presents a novel methodological-iterative (MI) approach for discrete choice model specification and estimation. The idea is based on feature selection, originally motivated by the problem of large datasets that contain many irrelevant and redundant features (Guyon, 2003). The

MI approach combines feature selection and discrete choice models. We propose a starting point to model-specification or a "safety net" that can help find statistically significant explanatory variables, or relationships among these variables, that may have been overlooked in the process of model building. Such a procedure is highly relevant to both researchers and practitioners: the specification of the utility functions of the model is a difficult, time-consuming task, and misspecification may lead to biased parameter estimates, lower predictability, and wrong interpretations (Bentz & Merunka, 2000), (Torres et al., 2011), (Van Der Pol et al., 2014).

The MI method aims to find statistically significant explanatory variables and obtain a model with high goodness-of-fit and prediction power. The MI method can be seen as a two-level backward elimination procedure, in which the upper level checks all the important features found in the random forest classifier, and the lower level checks all the coefficients for a checking feature (i.e. combine/eliminate insignificantly different alternative specific coefficients for a checking feature). We describe the procedure in detailed in the following paragraphs.

The starting point of the approach is the application of a random forest classifier (Breiman, 2001) to predict car allocation, which is an ensemble of decision trees (Breiman et al., 1984). A decision tree is a set of "if-then" statements that partition the space of explanatory variables into a set of mutually exclusive regions to predict the class of an observation. The decision tree implemented in this paper is estimated by a greedy algorithm that selects a binary split condition to maximize the weighted reduction in Gini impurity at each node. The Gini impurity formulation is expressed as follows:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \text{ , where } p_i = \left|C_{i,D}\right| / \left|D\right|$$

Where D is a data partition of training observations, p_i is the probability that observation in D belongs to the class C_i.

This decision tree considers a binary split for each attribute. The weighted sum of the impurity of each resulting partition is:

$$Gini_A = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

The reduction in impurity resulting from a binary split on attribute A is then given by:

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

Random forest (RF) is built of many decision trees in bagging (or bootstrap aggregation) procedure: given a training set, it repeatedly selects a random sample with replacement of the training set and a random subset of the features and fits decision trees to the samples. The RF classifies through a "majority vote" over all the trees to obtain a result. RF is a very efficient, supervised machine learning classifier. It is unbiased, handles outliers well, can handle both categorical and numerical values, does not require scaling of the data, and can handle non-linear parameters efficiently (J. Han et al., 2011).

An RF can be used to estimate the "feature importance" rank, meaning the importance of each variable to the classification problem. Feature importance is calculated here using the mean decrease in impurity, i.e., the total reduction in node impurity (weighted by the proportion of samples reaching that node) averaged over all trees (Louppe, 2014).

The importance of each feature to the decision tree is expressed by:

$$fi_A = \frac{\sum_{j:\text{node j splits on feature A}} ni_j}{\sum_{k \in \text{all nodes}} ni_k}$$

Where fi_A is the importance of feature A.

The normalized feature importance is then obtained by dividing by the sum of all feature importance values:

$$\text{norm}fi_A = \frac{fi_A}{\sum_{j \in \text{all features}} fi_j}$$

Then the feature importance is averaged over all trees in the ensemble:

$$\text{RF}fi_A = \frac{1}{T} \sum_{j \in \text{all trees}} \text{norm}fi_{A,j}$$

Where norm$fi_{A,j}$ is the normalized feature importance for feature A in tree j and T is the total number of trees.

When plotted on a bar graph, the ranking typically shows an "elbow," which can be interpreted as dividing all features into two groups: one group having the variables with more explanatory power (group 1) and the second group with the less explanatory power (group 2). For clarification; we consider a feature as an explanatory variable and a parameter as the coefficient being estimated. For example, in the car allocation problem with 7 alternatives, a possible explanatory variable is the number of household vehicles, which is specified as 6 parameters in 6 different utility functions.

The MI method runs as follows: a random forest is trained on a randomly selected test-set and the feature importance is obtained, which is partitioned into two groups by importance rank. All features in group 1 are specified in all the relevant utility functions. This specification must consider the degrees of freedom in the estimation (e.g., if there are J alternatives, an explanatory variable need be specified in J-1 alternatives). In each iteration, a discrete choice model is estimated, and the last feature in group 1 (by feature importance ranking) is considered. If the parameters corresponding to the current feature in different utility functions are not significantly different, they are combined to one parameter (by pairs, each time to those two who show the greatest resemblance). Then, another choice model is estimated until all parameters of the feature are significantly different from each other. Finally, all parameters that are not significantly different from zero are eliminated, by the same order. That is, the least important variable is considered, and if it is not significantly different than zero it is eliminated, and a choice model is estimated again until all variables for the considered feature are significantly different than one another and zero (see Figure 1 for an illustration).

When an iteration of a feature is complete, the next iteration considers the second worse feature in group 1 and goes on in a backward elimination procedure until all parameters are considered (notice that the precise mean decrease in impurity is not used, but only the order of the features by rank). After all the features in group 1 (those with more predictive power) are considered, the same process is done to group 2 (features with less predictive power) until all the parameters are significantly different from zero. The algorithm terminates when all coefficients are significantly different from zero (see Figure 1 for an illustration).

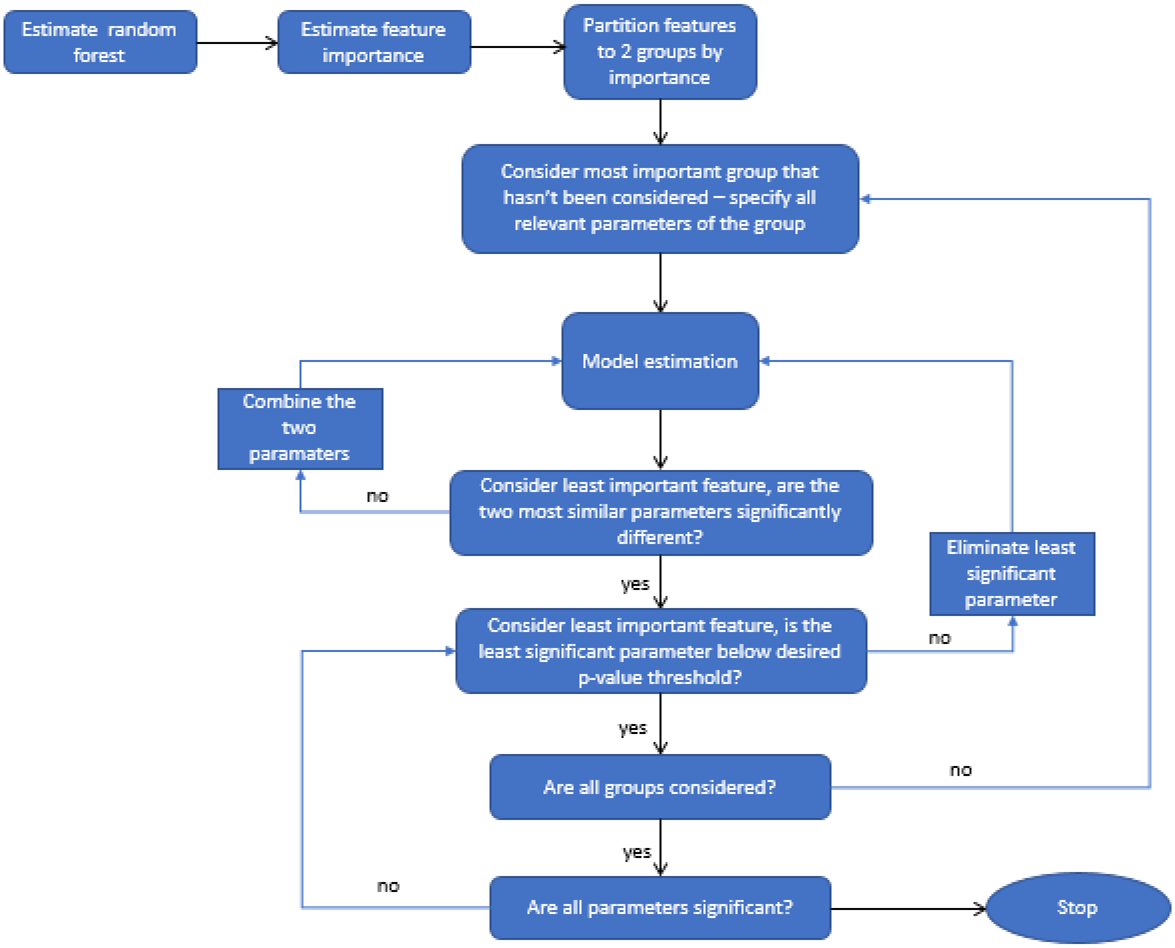
Estimate random forest
Estimate feature importance
Partition features to 2 groups by importance
Consider most important group that hasn't been considered – specify all relevant parameters of the group
Model estimation
Combine the two paramaters
no
Consider least important feature, are the two most similar parameters significantly different?
yes
Eliminate least significant parameter
Consider least important feature, is the least significant parameter below desired p-value threshold?
no
yes
Are all groups considered?
no
yes
no
Are all parameters significant?
Stop

Figure 1. methodological-iterative (MI) model estimation procedure

This paper incorporates the algorithm with the random forest's mean decrease in Gini impurity feature importance due to its many advantages, which we elaborate on, and its ease of implementation, but it is not restricted to them. that MI algorithm can estimate any linear supervised model (i.e., models in which the explanatory variable is specified a-priori), with any ensemble learning model and any feature importance technique. The question of which models and techniques are best suited to different research purposes is beyond the scope of this paper and should be investigated in future research.

The MI procedure incorporates feature selection for the utility specification of the model. It can also provide insight into another challenging task of finding potential non-linear relationships between many explanatory variables and the dependent variable, as the logit framework is restricted to linear coefficients (as suggested by Hillel et al., (2019)). One way to handle this limitation within the discrete choice framework is feature transformation, i.e., a function that transforms features from one representation to another. Because random forest (and other data-driven machine learning models) can handle nonlinear parameters efficiently, if an explanatory variable ranks high by feature importance but is not significant in the MI discrete choice model, this serves as an indication that this specific variable should be furthered researched and evaluated. It might be the case that this particular variable should undergo a certain transformation. Thus, the MI approach further contributes to the practice of model building and to find models with a better overall fit.

Another notable advantage of the MI approach is that it is less prone to overfitting, as it is a hybrid method of feature selection that combines a filter and wrapper methods (Jović et al., 2015). This is because the filter method (i.e. the feature importance) is obtained from a model that is independent of the choice model, with a completely different framework.

Model Specification

This paper applies the MI approach to model the choice behavior of the household members. To do so, we assume that the decision of car allocation is taken among household members, and it is contingent on the activities of its members over a full day. This assumption follows a central premise in activity-based models, that travel decisions are derived from the demand for activities, or simply the daily activity pattern. Individuals who have 24 hours in a day decide how to divide this time between their activities and how they meet travel limitations (Pinjari & Bhat, 2011). For example, two parents who own one vehicle do not decide who will take the vehicle to pick up the child from school at the trip level, but at the daily level. They make this decision regarding their full daily activity schedule coordination with each other.

Since car allocation decision is a decision taken simultaneously by the family members on a daily level, we have constructed the model's decision-maker as the household itself, whose members have to decide how to allocate the household vehicles, and to whom, on a single weekday. There are many different family types concerning household composition. For modeling purposes, the dependent variable is defined as the household member that the car is allocated to among seven possible alternatives:

1. Male adult only (M)
2. Female adult only (F)
3. Anyone of the descendants (D)
4. The male adult and at least one of the descendants (M+D)
5. The female adult and at least one of the descendants (F+D)
6. Both male adult and female adult (M+F)
7. Both male adult and female adult, and at least one of the descendants (M+F+D)

The independent variables that are selected for the analysis comprise of household variables, person variables, and household daily activity variables. The person and daily activity variables were aggregated at the household level.

Because we assume that the car allocation decision is taken daily, we need to define to which household member the car is allocated. Hence, we further assume that if a household member drives the vehicle at least once during a given day, then the car was allocated to this member. For example, if the male adult drives the car in the morning, and the female adult drives the car in the afternoon, then the vehicle is allocated to them both. Only when a household member does not drive the vehicle at all, we assume the car was not assigned to her.

The aggregated variables for each household member are respectively added to the utilities of the alternatives that include this member. For example, the total travel distance of the male adult is only added to the utilities that include the male adult：allocation to the male adult (M), male adult and descendants (M+D), male and female adults (M+F), and male adult, female adult and descendants (M+F+D).

Next, we estimated a random forest. The random forest consists of 150 decision trees which are grown until all leaf nodes are pure, and classification is made by a “majority vote” over all trees. The classification results of the random forest will be compared to the results of the MI method. The MI-MNL was estimated with the feature importance obtained from the random forest

Both models are compared by evaluating their performance in a cross-validation procedure. In this research, we compare all models through the evaluation of confusion matrices (Table 1), which will be a 7*7 matrix as there are 7 possible alternatives (or classes) and three standard scores for classification evaluation：precision, recall, and f1-score (Table 2). For MI-MNL, deterministic results were computed as assigning a class prediction for the highest probability among all alternatives.

Table 1：Confusion matrix

		Predicted class		
		yes	no	Total
Actual class	yes	TP	FN	P
	no	FP	TN	N
	Total	P'	N'	P+N

Table 2: Classifier evaluation measures

Measure	Formula
Precision	TP/(TP+FP)
Recall	TP/P
F1-score	2*Precision*Recall / (Precision + Recall)

DATA

The source of the data is the Tel Aviv household travel habit survey, a comprehensive state of the art survey of the residents in the Tel Aviv metropolitan area that was conducted with a designated mobile phone application. The survey was conducted in two beats, a preliminary survey in 2014, and the main survey in 2017, covering the entire Tel Aviv metropolitan area (Nahmias-Biran et al., 2018). This metropolitan area is the biggest in the country, inhabiting almost half of the county's population – 3.9 million people (2017). Overall, the survey includes information from 13,485 households, which adds up to 39,090 household members and 333,724 trips.

The database of the survey consists of four data sets with rich information: 1) households (e.g., household size, number of vehicles, residential zone), 2) household vehicles (e.g., ownership, fuel cost), 3) household members (e.g., age, relation, gender, employment), and 4) household member's activities and travel diary (e.g., main activity, mode, time, location). A full-day schedule was obtained for every respondent in the survey: the exact time and place of all trips in a single day, the purpose of the tours, the mode, and the traveler's personal information, as well as the information of the person's household.

As indicated in the introduction, we include only car deficient households. Moreover, we consider only households with no more than two adults of different gender, which we will coin type 1 households (i.e., male and female adults, only one male adult, only one female adult). All other household compositions are coined type 2. Table 3 shows the distribution of the sample concerning car deficiency and household type, and Table 4 shows the household composition among type 1 car deficient households.

Table 3. Household sample by car deficiency and household type

Households	Car deficient	Not car deficient	Total
Type 1 households	3,478 (25.8%)	9,852 (73.1%)	13,330
Type 2 households	55 (0.4%)	100 (0.7%)	155
Total	3,533	9,952	13,485

Table 4. Household composition in type 1 car-deficient households

Household type	Count	Percent
Couple with no children	1,047	31%
Parents with children	2,154	62%
A single parent with children	234	6.7%
Roommates (unrelated)	41	1.2%
Roommates (related)	2	0.1%
Total	3,478	100%

We also excluded from the analysis households of roommates because of their small number (42 in the sample). This leaves us with 3,435 households.

The possible options of car allocation in the sample are the male adult in the household (M), the female adult in the household (F), the descendants of these adults (D), any of the combinations between them, other, or no drivers at all. Table 5 presents the car allocation options in the sample.

Table 5. Car allocation distribution among household members in the sample

M	F	D	M+D	F+D	M+F	M+F+D	Other	No Drivers	Total
1,126	699	131	139	99	686	153	64	341	3,435
33%	20%	4%	4%	3%	20%	4%	2%	10%	100%

We excluded households of "no drivers" because we are interested in the car allocation between household members. Similarly, we excluded "other" households because they compose several small groups of different options (e.g., uncle, grandmother, unrelated resident) and add to only 1.9 percent of the sample. Finally, the number of observations for model estimation is 3,030 households (86% of all car deficient households). Selected characteristics are presented in Table 6 and Table 7.

Table 6. The percentage of vehicles and the percentage of adults in 3,030 car-deficient households

vehicles	one adult	two adults	total
1	5%	74%	79%
2 or more	1%	20%	21%
total	6%	94%	100%

Table 7. Selected household characteristics in 3,030 car-deficient households

	Average numbers in the household			Average age		
Car allocated to:	vehicles	descendants	licensed drivers	male	female	descendants
M	1.1	1.2	2.1	50.4	47.1	10.7
F	1.1	1.5	2.2	47.1	45.2	10.2
D	1.4	1.6	2.7	62.3	59.4	27.2
M+D	1.9	2.0	3.3	59.1	54.7	22.4
F+D	1.8	2.2	3.2	54.6	51.7	20.1
M+F	1.3	1.8	2.4	48.4	45.5	10.6
M+F+D	2.3	2.3	3.6	54.7	51.2	19.4

RESULTS

Random forest

As stated in the methodology section, both models are tested on the same test set. Note that the overall number of observations classified in the results is 606, which is the 20% test set of the 3,030 observations of the entire data set. Estimation of the random forest was performed by the machine learning open-source python library "Scikit-learn" (Pedregosa et al., 2011). A comparison and discussion will follow.

Results of the random forest, which is an ensemble of decision trees, are presented in Table 8, showing the confusion matrix, and Table , showing the model evaluation scores.

Table 8. Random forest's confusion matrix

confusion matrix	M	F	D	M+D	F+D	M+F	M+F+D
M	186	15	0	6	0	17	0
F	32	90	1	0	0	21	1
D	2	5	13	1	2	2	0
M+D	8	0	0	13	0	3	6
F+D	1	4	2	2	6	4	3
M+F	48	18	0	0	0	63	3
M+F+D	2	0	0	3	0	13	14

The random forest had a high precision rate for the descendants (D) class and high recall for the male adult class (M). The random forest underperformed on the F+D recall, probably resulting from the low number of instances this label has. The random forest classified relatively well the households that allocated the vehicle to only one family member. This suggests that these households are unique in a way that makes them easier to identify. For example, F had higher scores than M+F, although they have about the same number of instances. Figure 2 shows the result of the feature importance.

Table 9. Random forest's model evaluation

	precision	recall	f1-score	support
M	0.66	0.83	0.74	220
F	0.68	0.62	0.65	145
D	0.81	0.52	0.63	25
M+D	0.52	0.43	0.47	30
F+D	0.75	0.27	0.4	22
M+F	0.51	0.48	0.49	132
M+F+D	0.52	0.44	0.47	32
weighted avg	0.63	0.63	0.62	606

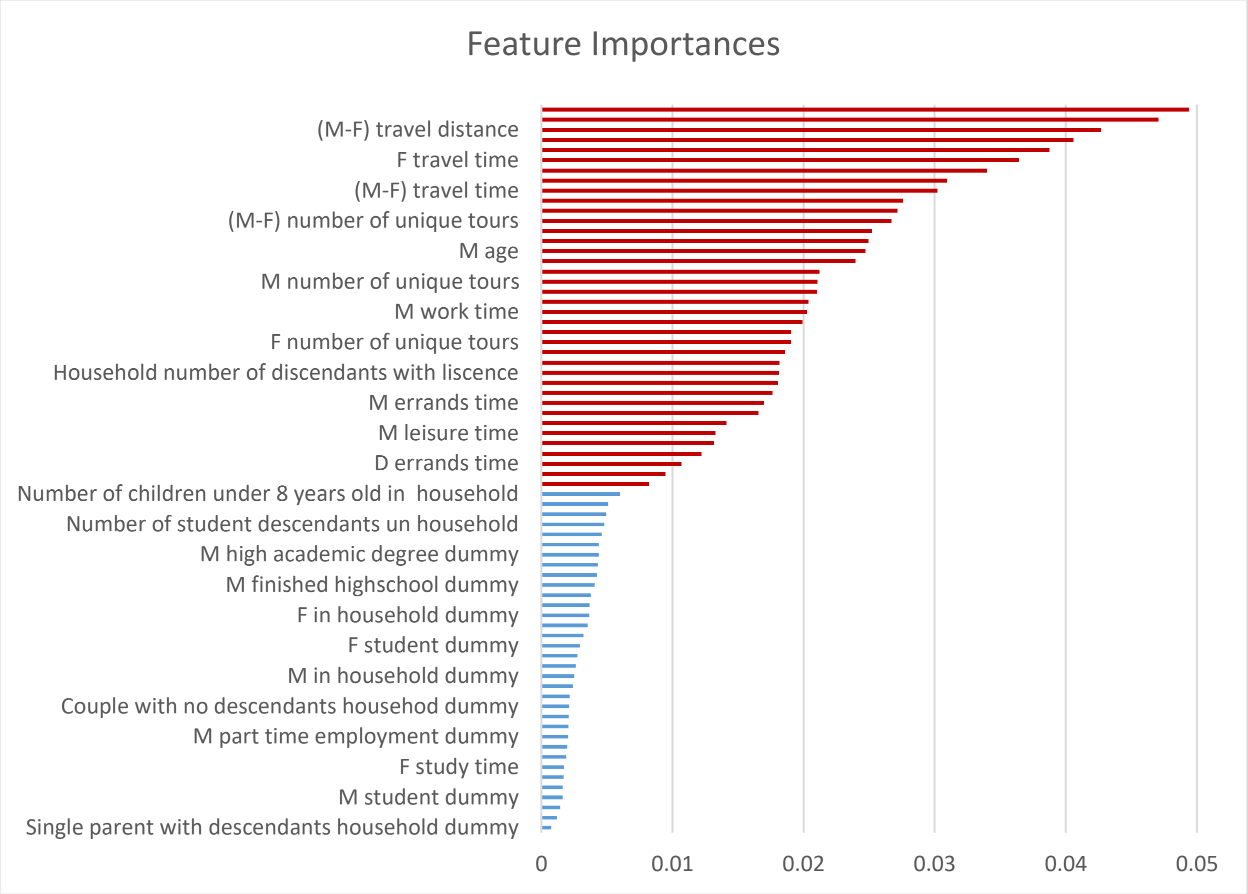

Figure 2. Random forest feature importance (group 1 in red and group 2 in blue)

As described in the methodology section, a feature importance can be generated from the random forest classifier. Figure shows the feature importance on a bar chart and the apparent "elbow" which was used to divide the features into two groups. The features of group 1 (the group with the more essential features to the model) are colored in red, and the features of group 2 are colored in blue. The MI approach first goes through all group 1 features, and only then group over the feature of group 2.

The most important features revealed by the random forest model (by descending rank) are travel distance, number of tours, travel time, number of household vehicles, and household members' age. These variables have been included in the utility specification by the MI algorithm and will be discussed in the next section.

MI Discrete Choice Model

All maximum likelihood estimations throughout the MI approach was performed using the Biogeme software (Bierlaire, 2003). The results for the MI-MNL are presented in Table 8. the first iteration estimated all parameters of the features in group 1. These added to 148 estimated parameters over 38 features, while only 40 of the parameters were significantly different than zero (27%) with a log-likelihood value of -2829. After all group 1 variables were considered, group 2 added another 126 parameters, over the remaining 34 features. Altogether, The systematic MI approach reduced the number of variables from 74 features to 36 features, and from 274 estimated parameters to 58 estimated parameters, all of them significantly different than zero, with a log-likelihood value of -2805 (see Table 10). This result was achieved without any prior knowledge of the phenomena (except for determining the dependent variable, which is required in all supervised machine learning models), as the MI method goes over all parameters in the data set.

Table 8. MI-MNL model estimation results

		MNL	
Variable name	alternative(s)	estimate	p-value
ASC	M	11.5	0
	F	11	0
	D	7.72	0
	M+D	2.68	0.05
	F+D	5.13	0
	M+F	6.37	0
Number of vehicles	M, F	-2.92	0
	D	-2.07	0
	M+D, F+D, M+F	-1.12	0
Number of household bikes	D	-0.232	0.01
Number of descendants	M	0.862	0
	F, M+F	1.07	0
Number of descendants with a driver's license	M, F, M+F	-0.965	0
	D	-0.685	0.01
Number of household children under 8 years old	M	-0.167	0.02
Age of youngest descendant	D	0.0368	0.02
Descendants total number of tours	D, M+D, F+D, M+F+D	0.256	0
Male adult's total number of trips	M, M+D	0.451	0
	M+F, M+F+D	0.749	0
Female adult total number of trips	F, F+D	0.439	0

		MNL	
Variable name	alternative(s)	estimate	p-value
(M-F) total number of trips	M+F, M+F+D	-0.364	0
Descendants total travel distance (km)	D, M+D	0.0051	0.01
Male adults total travel distance (km)	M	0.00771	0
	M+D, M+F, M+F+D	0.00631	0
Female adult total travel distance (km)	F	0.0177	0
	F+D	0.0224	0
	M+F, M+F+D	0.0114	0
Male adult's total travel time (min)	M, M+D	-0.00232	0
Female adult total travel time (min)	F, F+D	-0.00677	0
	M+F	-0.00433	0
(M-F) total travel time (min)	F, M+D	0.0016	0.02
Descendants total work time	D	0.00124	0
Male adults total work time (min)	M, M+D, M+F, M+F+D	0.00175	0
Female adult total work time (min)	F	0.00202	0
(M-F) total work time (min)	F+D, M+F, M+F+D	-0.00101	0
Female adult unemployed dummy	F+D	-1.34	0
Males adult high academic degree dummy	M	0.199	0.07
Females adult high academic degree dummy	F, M+F+D	0.398	0
	M+F	0.679	0
Female adult high school education dummy	F	-0.322	0.07
	F+D	0.768	0.01
	M+F	0.314	0.06
Descendants total errands time	F+D, M+F+D	-0.00226	0.03
Male adults total study time (min)	M	-0.00277	0.01
(M-F) total study time (min)	M, M+D	0.00323	0
Male adult total leisure time (min)	M, M+D, M+F	0.00376	0

		MNL	
Variable name	alternative(s)	estimate	p-value
Female adult total leisure time (min)	F	0.00275	0
(M-F) total leisure time (min)	M, M+D, M+F	-0.0023	0
Male adult's age	M, M+F	-0.0579	0
Female adults age	F, F+D	-0.0604	0
	M+F+D	-0.0457	0.02
(M-F) age	F	-0.0424	0.02
	M+D	0.0543	0.06
At least one of the parents retired and youngest descendants age is above 18 (dummy)	M+D	0.824	0.01
Secular dummy	M, F	-2.33	0
	D	-3.28	0
	M+D, M+F	-1.89	0.01
Jewish orthodox dummy	F+D	2.66	0
Number of estimated parameters:		58	
Number of observations:		3030	
Null log likelihood:		-3923.405	
Final log likelihood:		-2805.24	
Likelihood ratio test to Null		2236.329	

All coefficients are significant at least at a 7% p-value, and most are at 1% or less. Some of the explanatory variables are discussed in the following paragraphs.

As a reminder, the MI method's purpose is to help the model builder as a starting point to model building and finding significant explanatory variables among many features. The results presented in Table 8 are this starting point, i.e., these are the output of the MI approach without further enhancing the model.

Model Variables

This subsection summarizes selected variables that were found significant in the MI model.

Number of household vehicles

The more vehicles in the household, the higher the probability that the car would be allocated to all family members; the male, female, and descendants (M+F+D), which is evident by the fact that all coefficients are negative, while the utility of M+F+D is the reference (hence its parameter is set to zero). Accordingly, a combination of family members has higher chances to use the car than the male or female adult alone. Note that the number of household descendants with a driver's license can be more than one. Hence, the higher the number of vehicles available, there is a higher probability that only the descendants use the car compared to a single adult (either M or F).

Number of household descendants

The more descendants are in the household, the higher probability is that the car would be allocated to the female adult (F), and both parents (M+F). This seems reasonable because the parents are the caregivers of the descendants, and they need the vehicle to take care of them. There is an apparent difference between female and male adults. The more children are in the household, the higher is the probability that the car will be allocated to the mother than to the father (as also found by (Maat & Timmermans, 2009), (Scheiner & Holz-Rau, 2012)). This suggests that the mother might play a greater role in the caregiving of the descendants.

Number of household descendants with a driver's license

As opposed to the number of overall descendants, the more household descendants with a driver's license, the probability the car would be allocated to one of the parents or both of them decreases (M, F, M+F). Again, this is reasonable because the descendants' demand for the vehicle increases as there are more descendants with a driver's license, and they probably share the vehicle with either the male or female adult, thus the parameter for the descendants alone is also negative.

Number of children under the age of eight in the household

The more children under the age of eight are in the household, the lower is the probability that the car will be allocated to the male adult alone (as found in (Maat & Timmermans, 2009), (Scheiner & Holz-Rau, 2012)). This result may indicate the presence of gender roles; that when there are young children, the associated activities and responsibilities concerning the children are shared between the male adult and other members of the household, probably the mother, who needs the car to take care of the young children.

Total number of trips

Note that the model specification includes (M-F) variables, defined as the difference between the males' and females' attribute. The higher the difference between the male number of tours and the female number of tours, the lower the probability that the vehicle would be allocated to both the male and the female together (M+F, M+F+D), and vice-versa; the more trips that the female has than the male, the higher the utility for allocating to M+F and M+F+D. This finding may suggest gender inequality: when the male adult has more trips than the female, there is a lower probability for both adults to share the vehicle, but when it comes to the female, it is the other way around. Besides the difference, as a family member's total number of trips increases, the probability the car would be allocated to him increases.

Total travel distance & total travel time

The higher the total travel distance of the adults, the higher the probability the car would be allocated to them (as found in (Habib, 2014)). There is a tradeoff between travel distance and travel time. For example, while the Male adult's total travel distance increases his probability of using the car, the total travel time has the opposite effect, as the corresponding coefficient is negative. This tradeoff may suggest that while travel distance increases the demand, longer travel time means less available car use for other family members. It should be stressed that all explanatory variables were checked for correlation to ensure there is no high correlation between different features, including travel distance and travel time. For the descendants, the correlation is 0.72, for the female adults 0.67, and the male adults are 0.69.

Total work time

The more time anyone of the family members spends at work, the higher is the probability that the vehicle would be allocated to them. Anggraini et al. (2008) found, on the contrary, that the probability of the male getting the car decreases as his work duration increases, which is also a plausible result and might depend on the specific model formulation.

Unemployment of female adult dummy

If the female adult is unemployed, there is a lower probability that the car would be allocated to her and the descendants (F+D) (which corresponds with (Arman et al., 2015)), probably because unemployed women are considered somewhat secondary to the male adult in the family "hierarchy." This effect was insignificant for male unemployment.

High academic degree dummy

If a female adult has an academic degree higher than a bachelor, the probability of the alternatives that include the female adult is higher (except F+D). this is also the case for male adult only (M), but with less of an effect. It may suggest that having an academic degree promotes status in the household, more so in the case of the female adult. Another explanation can be that education is correlated with income (which was not asked in the survey), and income promotes status in the household.

Adult's age

As the adults are older, the probability the car would be allocated to them decreases. Older individuals may have fewer activities to attend or are not as willing to use the vehicle. This result corresponds with Scheiner & Holz-Rau (2012) that found that elder partners tend to use the car less. Moreover, as the male adult is older than the female adult, the female adult will use the car less, and the male and the descendants (M+D) will be allocated the vehicle more, and vise-versa. This result suggests that the difference in age also affects household status.

Religion dummy

The household that considered themselves secular in the survey showed the least willingness to allocate the vehicle to the descendants, then to only one of the parents, and finally to M+D or M+F, as compared to allocating the vehicle to all family members (M+F+D). On the other hand, households that considered themselves orthodox Jewish showed a higher probability of allocating the vehicle to the adult female and the descendants together (F+D). The result may correspond to the fact that in orthodox Jewish households, the male adults participate less in the workforce than the female adults. Therefore, it might be the case that the male adult has fewer household responsibilities, and therefore the car is less allocated to him.

To compare the prediction forecasts of the MI-MNL model to the decision tree and random forest, we evaluate the MNL with a confusion matrix by cross-validation. The MNL model was again estimated on the same training set (randomly selected 80% of the data), and all non-significant coefficients were removed. Then, the new model estimated the car allocation on the test set. Table 11 presents the MI-MNL confusion matrix Table 12 shows the model evaluation.

Table 9. MI-MNL confusion matrix

confusion matrix	M	F	D	M+D	F+D	M+F	M+F+D
M	170	26	2	2	0	20	0
F	34	88	1	0	2	19	1
D	1	3	18	2	1	0	0
M+D	4	0	2	14	0	7	3
F+D	0	6	3	1	5	2	5
M+F	53	21	0	1	0	52	5
M+F+D	2	0	1	3	0	9	17

Table 10. MI-MNL model evaluation

	precision	recall	f1-score	support
M	0.64	0.77	0.7	220
F	0.61	0.61	0.61	145
D	0.67	0.72	0.69	25
M+D	0.61	0.47	0.53	30
F+D	0.62	0.23	0.33	22
M+F	0.48	0.39	0.43	132
M+F+D	0.55	0.53	0.54	32
weighted avg	0.59	0.6	0.59	606

Model comparison & Discussion

Table 11 Compares the weighted average of the F1-score of MI-MNL and the random forest. The differences are not very large, but prediction wise the random forest obtained the best results, as expected.

Table 11. The weighted average of F1-score for selected models

Model	Weighted average of F1-score
Random forest	0.62
MI-MNL	0.59

Note that many times, as in this case study, we are interested in more than the predictions; we wish to predict the choice among alternatives while understanding the behavioral process that led to the specific choice. A random forest can find feature importance over extensive data sets, in a fast and efficient manner, and handles outliers and nonlinear variables exceptionally well. The primary downside of all the efficient and prediction wise accurate supervised data-driven learning machine models is that they are somewhat of a “black box” that can be interpreted in a limited way, where the reason or the process that led to the outcome remains to a large extent unknown. The MI approach is a hybrid method that is aimed at harnessing machine learning advantages to utility function specification, which remains a difficult task while obtaining the interpretability of discrete choice models. It helps in improving the overall model, especially when there is not sufficient domain knowledge of the investigated phenomenon.

As stated in the research objectives, the MI approach reduces specification time, finds overlooked explanatory variables, and improves the model altogether as a starting point to model building.

Even without doing so (i.e., treating it as the final model and not building on it and improving it) the MI method obtained, without prior knowledge needed, a very good model that makes sense and is consistent with previous research in the field of car allocation. It found explanatory variables that are reasonable and informative and obtained a similar predictive accuracy as the random forest.

The explanatory mechanism discovered by the MI approach shows an expected behavior of supply and demand within the household. When there is a scarce resource, the vehicle, which is essential to basic mobility and participation in different activities, the households need to allocate the vehicle to a family member at the expense of another. This effect is evident by explanatory variables such as the number of household vehicles, total number of trips, total travel distance, and total travel time, for example. The feature importance of the random forest further reinforces this claim, as many of the variables that indicate the need for the vehicle are the most important for the classification model by the Gini gain index.

Moreover, the estimation results suggest a household hierarchy, where overall parents are situated above the descendants, a plausible conclusion as parents are responsible for their descendants' wellbeing. Moreover, different factors may influence household status within this hierarchy, for example, the explanatory variables of employment, education, or the age of the youngest descendant, as discussed in the results. This hierarchy is also apparent by the difference between the adult's unique number of tours compared to those of the descendants.

Within the family hierarchy, there might be some indication of gender inequality and gender roles. For example, the number of descendants has a more positive effect on allocation to the female adult than to the male adult. Furthermore, it seems that female status concerning car allocation is dependent on factors such as employment and education, whereas these variables have less explanatory power in the case of the male adult. The presence of gender roles and gender inequality have also been found in (Srinivasan & Bhat, 2005), (Scheiner & Holz-Rau, 2012) or (Anggraini et al., 2012), for example.

Overall, the MI model results are in line with previous studies on car allocation. The application of the MI method and the rich dataset allowed to discover additional explanatory variables in a systematic procedure that obviates preliminary knowledge of the investigated phenomenon.

SUMMARY AND CONCLUSIONS

Methodologically, this paper proposes a systematic and fast procedure to estimate discrete choice models that utilize some of the advantages of the random forest classifier. Such advantages are efficiently handling large sets of data, outliers, and nonlinear parameters, without an explicit apriory specification, except for the causal direction, which is necessary for all supervised machine learning models. The MI method is intended to help the modeler improve her research by identifying the important variables that she should include in the model specification and more importantly the relationships among these variables. This method is especially helpful when the modeler does not have sufficient knowledge about the investigated phenomena and when the data set is large.

As such, it investigates the complex relationships among explanatory variables. In this case study, the relationships within the household, between adults and adults and their descendants, in allocating available cars among them. The contribution of this research from the perspective of choice-behavior is its broad research subject, using extensive revealed preference data, and covering car deficient household for different travel purposes over a full day. To capture the interrelated decisions of all household members, we have modeled the decision-maker at the household level, rather than at the level of each household member alone. We can consequently implicitly account for the decision process that includes communication and deliberation between household members while taking advantage of an extensive revealed preference survey.

The paper also contributes with a wide range of results that are based on many significant explanatory variables. These variables explain the decision of which household member will drive the vehicle, and what kind of considerations enter the decision process. These include household attributes, household member attributes, and characteristics of the household member's overall trips in a single weekday. The results show that the demand for household vehicles is essential and that there is a trade-off in car-deficient households between the needs of family members. These considerations are possibly affected by family hierarchy and gender roles within the household.

Further Research

This paper implemented the MI method in a semi-automated fashion. Further research will automate the full process, which will allow for improved prediction accuracy. Regarding the methodological-iterative approach, future research should test the MI approach in other case studies and test its applicability to different phenomena and datasets. Furthermore, the MI method should be tested with other ensemble learning models or indexes, and with other feature ranking methods, such as permutation importance. There is also a need to investigate what is the most appropriate method for each case. Finally, future research should improve the MI method altogether, incorporating possible variable transformations systematically, and enhancing the method's flexibility

Car allocation wise, future research should extend the scope of the research to all types of households and develop the study by fully incorporating the car allocation model within an activity-based model. Moreover, in this paper, we assume the direction of the causal effect, where the included explanatory variables explain the car allocation (the dependent variable). This causality may not necessarily be the case, and other techniques, such as unsupervised machine learning, should be implemented to further investigate the relationship between the car allocation decision and other household characteristics. The assumption that the car was allocated to those who drove it should also be examined in a more elaborate procedure, and if there are conditions in which this assumption may be inappropriate.

Author's contribution

Yuval Shiftan: Model Concept, Data Analyses, Model Estimation, Result interpretation, Manuscript Writing.

Shlomo Bekhor: Model Concept, Data Collection, Feedback on Analysis, Result Interpretation, Manuscript Editing.

REFERENCES

Anggraini, R., Arentze, T. A., & Timmermans, H. J. P. (2008). Car allocation between household heads in car deficient households: A decision model. *European Journal of Transport and Infrastructure Research*, *8*(4), 301–319.

Anggraini, R., Arentze, T. A., Timmermans, H. J. P., Anggraini, R., Arentze, T. A., & Car, H. J. P. T. (2012). *Car allocation decisions in car-deficient households : the case of non-work tours*. *8602*. https://doi.org/10.1080/18128602.2010.539415

Arman, M. A., Kalantari, N., & Mohammadian, A. (2015). Joint Modeling of Household Vehicle and Activity Allocation Statistical Analysis and Discrete Choice Modeling Approach. *Transportation Research Record*, *2495*, 121–130. https://doi.org/10.3141/2495-13

Barsky, Y., Gal-Tzur, A., & Bekhor, S. (n.d.). *Methodology for real-time congestion forecasting based on Feature Engineering*.

Bentz, Y., & Merunka, D. (2000). Neural networks and the multinomial logit for brand choice modelling: a hybrid approach. *Journal of Forecasting*, *19*(3), 177–200.

Bhat, C. R., & Pendyala, R. M. (2005). Modeling intra-household interactions and group decision-making. *Transportation*, *32*(5), 443.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models. *Swiss Transport Research Conference*, *CONF*.

Bradley, M., & Vovsha, P. (2005). A model for joint choice of daily activity pattern types of household members. *Transportation*, *32*(5), 545–571. https://doi.org/10.1007/s11116-005-5761-0

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth

Int. *Group*, *37*(15), 237–251.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods q. *Computers and Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society*, *14*(May 2018), 1–10. https://doi.org/10.1016/j.tbs.2018.09.002

Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. *Icml*, *1*, 74–81.

Gliebe, J. P., & Koppelman, F. S. (2005). Modeling household activity-travel interactions as parallel constrained choices. *Transportation*, *32*(5), 449–471. https://doi.org/10.1007/s11116-005-5328-0

Guyon, I. (2003). *An Introduction to Variable and Feature Selection 1 Introduction*. *3*, 1157–1182.

Habib, K. N. (2014). *Household-level commuting mode choices, car allocation and car ownership level choices of two-worker households: the case of the city of Toronto*. 651–672. https://doi.org/10.1007/s11116-014-9518-5

Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, *78*(February), 273–282. https://doi.org/10.1016/j.eswa.2017.01.057

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Han, Y., Zegras, C., Pereira, F. C., & Ben-Akiva, M. (2020). *A Neural-embedded Choice Model: TasteNet-MNL Modeling Taste Heterogeneity with Flexibility and Interpretability*. http://arxiv.org/abs/2002.00922

Hillel, T., Bierlaire, M., Elshafie, M., & Jin, Y. (2019). *Weak teachers: Assisted specification of discrete choice models using ensemble learning*.

Ho, C., & Mulley, C. (2015). Intra-household interactions in transport research : a review. *Transport Reviews*, *0*(0), 1–23. https://doi.org/10.1080/01441647.2014.993745

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205.

Lee, D., Derrible, S., & Pereira, F. C. (2018). Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. *Transportation Research Record*, *2672*(49), 101–112.

Lee, Y., & Wei, C. (2010). A computerized feature selection method using genetic algorithms to forecast freeway accident duration times. *Computer-Aided Civil and Infrastructure Engineering*, *25*(2), 132–148.

Liu, H., & Motoda, H. (2012). *Feature selection for knowledge discovery and data mining* (Vol. 454). Springer Science & Business Media.

Louppe, G. (2014). Understanding random forests: From theory to practice. *ArXiv Preprint ArXiv:1407.7502*.

Maat, K., & Timmermans, H. J. P. (2009). Influence of the residential and work environment on car use in dual-earner households. *Transportation Research Part A*, *43*(7), 654–664. https://doi.org/10.1016/j.tra.2009.06.003

Munro, A. (2018). Intra-Household Experiments: A Survey. *Journal of Economic Surveys*, *32*(1), 134–175.

Nahmias-Biran, B., Han, Y., Bekhor, S., Zhao, F., Zegras, C., & Ben-Akiva, M. (2018). Enriching Activity-Based Models using Smartphone-Based Travel Surveys. *Transportation Research Record*, *2672*(42), 280–291.

Olde Kalter, M. J., & Geurs, K. T. (2016). Exploring the impact of household interactions on car use for home-based tours: A multilevel analysis of mode choice using data from the first two waves of the Netherlands mobility panel. *European Journal of Transport and Infrastructure Research*, *16*(4), 698–712. https://doi.org/10.18757/ejtir.2016.16.4.3166

Ou, J., Xia, J., Wu, Y.-J., & Rao, W. (2017). Short-term traffic flow forecasting for urban roads using data-driven feature selection strategy and Bias-corrected random forests. *Transportation Research Record*, *2645*(1), 157–167.

Paredes, M., Hemberg, E., O'Reilly, U.-M., & Zegras, C. (2017). Machine learning or discrete choice models for car ownership demand estimation and prediction? *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 780–785.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Roorda, M. J. (2006). *Incorporating Within-Household Interactions into Mode Choice Model with Genetic Algorithm for Parameter Estimation*. *May 2014*. https://doi.org/10.3141/1985-19

Scheiner, J., & Holz-Rau, C. (2012). Gendered travel mode choice: a focus on car deficient households. *Journal of Transport Geography*, *24*, 250–261. https://doi.org/10.1016/j.jtrangeo.2012.02.011

Sekhar, C. R., Minal, & Madhu, E. (2016). Mode Choice Analysis Using Random Forrest Decision Trees. *Transportation Research Procedia*, *17*(December 2014), 644–652. https://doi.org/10.1016/j.trpro.2016.11.119

Sifringer, B., Lurkin, V., & Alahi, A. (2018). *Let Me Not Lie: Learning MultiNomial Logit*. 1–32. http://arxiv.org/abs/1812.09747

Srinivasan, S., & Bhat, C. R. (2005). Modeling household interactions in daily in-home and out-of-home maintenance activity participation. *Transportation*, *32*(5), 523–544.

Timmermans, H. J. P., & Zhang, J. (2009). Modeling household activity travel behavior : Examples of state of the art modeling approaches and research agenda. *Transportation Research Part B*, *43*(2), 187–190. https://doi.org/10.1016/j.trb.2008.06.004

Torres, C., Hanley, N., & Riera, A. (2011). How wrong can you be? Implications of incorrect utility function specification for welfare measurement in choice experiments. *Journal of Environmental Economics and Management*, *62*(1), 111–121.

Van Der Pol, M., Currie, G., Kromm, S., & Ryan, M. (2014). Specification of the utility function in discrete choice experiments. *Value in Health*, *17*(2), 297–301.

Weiss, A., & Habib, K. N. (2020). Understanding joint travel and vehicle allocation at the household level through inference of revealed preference data. *Travel Behaviour and Society*, *19*(September 2018), 207–218. https://doi.org/10.1016/j.tbs.2020.01.006

Yang, J., & Ma, J. (2019). Compressive sensing-enhanced feature selection and its application in travel mode choice prediction. *Applied Soft Computing*, *75*, 537–547.

Yang, S. (2013). On feature selection for traffic congestion prediction. *Transportation Research Part C: Emerging Technologies*, *26*, 160–169.