Expert Systems With Applications 78 (2017) 273–282

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A comparative study of machine learning classifiers for modeling travel mode choice

Julian Hagenauer[a,*], Marco Helbich[b]

[a] Leibniz Institute of Ecological Urban and Regional Development, Dresden, Germany
[b] Department of Human Geography and Spatial Planning, Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 12 October 2016
Revised 14 January 2017
Accepted 30 January 2017
Available online 13 February 2017

Keywords:
Travel mode choice
Classification
Machine learning
The Netherlands

ABSTRACT

The analysis of travel mode choice is an important task in transportation planning and policy making in order to understand and predict travel demands. While advances in machine learning have led to numerous powerful classifiers, their usefulness for modeling travel mode choice remains largely unexplored. Using extensive Dutch travel diary data from the years 2010 to 2012, enriched with variables on the built and natural environment as well as on weather conditions, this study compares the predictive performance of seven selected machine learning classifiers for travel mode choice analysis and makes recommendations for model selection. In addition, it addresses the importance of different variables and how they relate to different travel modes. The results show that random forest performs significantly better than any other of the investigated classifiers, including the commonly used multinomial logit model. While trip distance is found to be the most important variable, the importance of the other variables varies with classifiers and travel modes. The importance of the meteorological variables is highest for support vector machine, while temperature is particularly important for predicting bicycle and public transport trips. The results suggest that the analysis of variable importance with respect to the different classifiers and travel modes is essential for a better understanding and effective modeling of people's travel behavior.

1. Introduction

The accurate modeling of travel mode choice is important for transportation planning and policy makers to predict travel demand and understand the underlying factors (de Dios Ortúzar & Willumsen, 2011). In fact, a large body of literature shows that travel mode choice is affected by a variety of factors including individual and household characteristics (e.g. Dieleman, Dijst, & Burghouwt, 2002; Schwanen & Mokhtarian, 2005; Böcker, van Amen, & Helbich, 2016) as well as the built environment (e.g. Ewing & Cervero, 2010; Helbich, 2016) and weather conditions (e.g. Böcker, Dijst, & Prillwitz, 2013).

Models of travel mode choice have traditionally been estimated using the discrete choice framework, where travel modes represent mutually exclusive and collectively exhaustive alternatives (Ben-Akiva & Lerman, 1985). The most widely used discrete choice model is the multinomial logit (MNL) model (McFadden, 1973). It is based on the principles of utility maximization and has a mathematic structure which eases parameter estimation (Koppelman & Wen, 1998). For this reasons it has been widely adopted in transportation research (e.g. Ewing, Schroeer, & Greene, 2004; Böcker et al., 2016). A limitation of MNL models is that they assume that the probabilities of each pair of alternatives are independent of the presence or characteristics of all other alternatives (McFadden, 1973). Consequently, the introduction of any alternative has the same proportional impact on the probability of each other alternative. Violation of this assumption yields inconsistent parameter estimates and biased predictions (McFadden, 1973). Other discrete choice models, such as the multinomial probit model (MNP), do not make this independence assumption, but parameter estimation is more difficult than for the MNL model, which hampers their usefulness (Dow & Endersby, 2004).

Methods from the field of machine learning are a promising alternative to statistical approaches for modeling travel mode choice. Instead of making strict assumptions about the data, machine learning models learn to represent complex relationships in a data-driven manner (e.g. Bishop, 2006). The usefulness of machine learning models has already been demonstrated for different areas in transportation research. For example, machine learning models are particularly useful for classifying travel modes

* Corresponding author
E-mail addresses: j.hagenauer@ioer.de (J. Hagenauer), m.helbich@uu.nl (M. Helbich).

http://dx.doi.org/10.1016/j.eswa.2017.01.057
0957-4174/

and inferring trip purposes from global position system and acceleration data (e.g. Shen & Stopher, 2014; Gong, Morikawa, Yamamoto, & Sato, 2014; Shafique & Hato, 2015). Other examples include the prediction of railway passenger demand (e.g. Tsai, Lee, & Wei, 2009) and bimodal modeling of freight transportation (e.g. Tortum, Yayla, & Gökdağ, 2009). However, machine learning is still under represented in research of travel mode choice modeling. Existing studies are limited to a small number of machine learning methods and do not provide comprehensive model comparisons.

Cantarella and De Luca (2003), for example, trained two artificial neural networks (ANNs) with different architectures to model people's travel mode choice behavior. They found that both ANNs clearly outperform a MNL model. Celikoglu (2006) showed that ANNs are effective for calibrating the utility function in travel choice modeling. Zhao, Shao, Li, Dong, and Liu (2010) demonstrated that the accuracy of probabilistic ANNs is similar to basic ANNs for travel mode choice prediction, whereas Omrani, Charif, Gerber, Awasthi, and Trigano (2013) showed that ANNs are more accurate than the other investigated alternatives. A few studies report less promising results for ANNs in comparison to traditional models. Hensher and Ton (2000), for instance, compared the predictive capabilities of ANNs and nested logit models in the context of commuter mode choice and found no performance advantage for ANNs. Similarly, Sayed and Razavi (2000) reported that the classification performance of fuzzy ANNs, MNL, and MNP models is similar.

Classification trees (CTs) have also been applied for travel mode choice analysis. Xie, Lu, and Parkany (2003), for instance, compared CTs and ANNs with MNL models. They conclude that CTs and ANNs perform better than MNL. Moreover, they state that CTs are more efficient and provide better interpretability than ANNs. Rasouli and Timmermans (2014) investigated the relationship between predictive performance and the number of CTs when using ensemble learning. They showed that the accuracy increases non-monotonically with the size of the ensemble. Hierarchical tree-based regression is used by Zhan, Yan, Zhu, Wang et al. (2016) to investigate the travel characteristics of Chinese students and to determine variables that affect students' travel behavior. Tang, Xiong, and Zhang (2015) used CTs to explore travel mode choice for the case where the choice is restricted to two modes in order to investigate people's mode-switching behavior. They confirmed the superior predictive capability of a CT to an MNL model.

Support-vector machines (SVMs) have also been applied in numerous studies. For example, when Zhang and Xie (2008) compared SVM, ANNs, and MNL for modeling travel mode choice, they found that SVM provided the highest accuracy. By contrast, Omrani (2015) showed that ANNs are more accurate than SVMs and MNL models for modeling the travel mode choice behavior of commuters. Xian-Yu (2011) reported that the performance of SVM is superior to ANN and nested logit models.

While the aforementioned studies represent important contributions to the application of machine learning in transportation research, they also have some major limitations. First, these studies deal only with a limited set of machine learning classifiers, even though the number of available classifiers is large (e.g. Fernández-Delgado, Cernadas, Barro, & Amorim, 2014). Advanced classifiers such as random forests or ensemble learners have not been considered in a comparative study, even though it has been shown that these classifiers can produce highly accurate results for many applications (e.g. Fernández-Delgado et al., 2014). Second, model comparisons are not done in a systematically quantitative way using statistical test procedures which take the sampling variability into account (see Hothorn, Leisch, Zeileis, & Hornik, 2005). Third, previous studies do not consider characteristics of the built and natural environment and meteorological conditions, even though these factors substantially influence travel behavior (e.g. Helbich, Böcker, & Dijst, 2014; Liu, Susilo, & Karlström, 2015). Finally, these studies do not thoroughly investigate the importance of variables, particularly with regard to the different models and travel modes, even though such an analysis supports the interpretation and understanding of the results (e.g. Murray & Conner, 2009).

This study addresses the identified shortcomings and adds to the literature as follows: First, it presents a comprehensive comparison of seven machine learning classifiers. Second, the article systematically evaluates the classifiers using strict model validation techniques and test statistics. Third, in addition to individual and household characteristics, it considers characteristics of the built and natural environmental as well as meteorological conditions for model building. Finally, the article investigates the importance of each variable for each classifier and travel mode in detail.

The rest of this article is structured as follows: Section 2 outlines the data and methods used. Section 3 describes the results, followed by a discussion in Section 4. Finally, Section 5 closes the paper with concluding remarks.

2. Materials and methods

2.1. Data

The primary data source for this study is the Dutch national travel survey (NTS) conducted from 2010 to 2012. It is supplied by Onderzoek Verplaatsingen in Nederland (2014) and is based on individual travel diaries. The survey participants were asked to record every trip over the course of six days, which have been randomly selected to cover a whole year in order to account for seasonal effects. To compensate for the lower response rates of non-natives and older participants, both groups were oversampled. In addition to trip-specific data (e.g. travel mode and trip distance), the NTS also provides socio-economic data about the participants (e.g. gender, age, and ethnicity) as well as information on households (e.g. income, number of cars and bicycles). The present study considers only records of participants aged 18 and over to exclude the distinct travel behavior of younger people. Furthermore, records that contain incomplete or erroneous information are also excluded. The resulting sampled data set consists of 69,918 individuals and a total of 230,608 trips. These trips are spatially distributed across all regions of the Netherlands and represent the travel behavior of the Dutch population as a whole. The NTS data can be accessed free of charge from DANS (Data Archiving and Networked Services) through the following link: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:54132/tab/1.

The data is additionally enriched with environmental data (see Fishman, Böcker, & Helbich, 2015). For this purpose, the residential locations of the participants are geocoded by postal codes using the nationwide cadastre database (Basisregistraties Adressen en Gebouwen). The locations are then utilized to derive variables that characterize the built and natural environments and weather conditions. The resulting data set consists of 17 variables, described in Table 1. The proportion of green space and the land use diversity are calculated using the Dutch land use model (Landelikj Grandgebruiksbestand Nederland 2008/2009). The meteorological variables are derived from the daily reports of the nearest weather station of the Royal Dutch Meteorological Institute. There are 36 such stations. Descriptive statistics of each variable are given in Table 2. The data set used for this model competition is provided through the journal's website.

2.2. Classifiers

This article compares seven machine learning methods to classify travel mode choice. These methods have either already been

Table 1
Description of the variables.

Variable	Description
Trip	
distance	Total trip distance in km
weekend	Trip is done at the weekend
mode	Main travel mode (walk, bike, pt, car). pt refers to public transport.
Individual	
age	Age of participant in years
education	Education of participant (lower, middle, higher)
ethnicity	Ethnicity of participant (native, western, other)
license	Participant owns a driver's license (yes, no)
male	Male participant (yes, no)
Household	
bicycles	Number of bicycles per household
cars	Number of cars per household
income	Net annual household income in 1,000€ ($<$20,$\geq$ 20–40,$\geq$ 40)
Build and natural environment	
density	Address density, aggregated over post codes, in 1,000addresses per km^2
diversity	Shannon diversity index of land use classes
green	Proportion of green space per post code area in %
Weather	
precip	Daily precipitation sum in mm
temp	Daily maximum temperature in °C
wind	Daily average wind speed in m/s

successfully used in transportation research or have shown promising results in other fields (e.g. Xu, Li, & Brenning, 2014). The parameters of each classifier are determined by systematically testing values from a manually specified subspace (e.g. Hsu, Chang, Lin et al., 2003). For computational reasons, a random sample (without replacement) of 100,000 trips is used for this purpose. The results can be downloaded from the journal's website.

Because MNL models are frequently used in discrete choice modeling and classification of travel mode choice (Ben-Akiva & Lerman, 1985), they serve in this study as a baseline classifier. The MNL model is estimated using an ANN-based approach (see Ripley, 2007). The ANN used has no hidden layers and is trained by back propagation with a weight decay constant of 0.01.

Naive Bayes (NB) is a simple machine learning method that calculates class probabilities using Bayes theorem while assuming that the features are independent. Predictions are then made for the class with the highest probability. In order to calculate probabilities from continuous features, their probability distributions must be estimated. This is typically done using kernel density estimation (John & Langley, 1995). Even though the independence assumption of NB rarely holds in practice, the classifier has shown to be competitive with more advanced classifiers (e.g. Huang & Ling, 2005). In this study, kernel density estimation with a Laplace correction factor of 0.001 is used.

SVM is a machine learning method for binary classification. It classifies observations by projecting the independent variables into a high-dimensional feature space, where the classes are linearly separable (Cortes & Vapnik, 1995). Since the basic SVM is a binary classifier, a one-against-one-approach is used for multiclass classification. In this approach, $k(k-1)/2$ binary classifiers are trained, with each classifier learning to distinguish a different pair of k classes. For prediction, the class that receives the most votes from all classifiers is chosen. Here, a SVM with a Gaussian kernel is used. The cost of constraint violation is set to 1.25 and the kernel bandwidth is set to 0.4.

Inspired by the biological brain, ANNs consist of a set of artificial neurons and directed connections between them (e.g. Rojas, 2013). Input data is passed through the network where it is summarized and processed by the neurons and weighted by the connections to give a network output. During the training of an ANN, the weights of the connections are adapted to produce a desired network output. The prediction of class membership is

Table 2
Descriptive statistics of the variables of the trip data set.

Variable	Category	%	Min.	Max.	Mean	Std. Dev.
Trip						
distance			0.100	400.000	12.218	23.546
weekend	no	82.066				
mode	walk	20.935				
	bike	24.473				
	pt	2.316				
	car	52.276				
Individual						
age			18.000	98.000	47.661	15.935
education	low	27.370				
	middle	38.293				
	high	34.337				
ethnicity	native Dutch	87.404				
	western	7.707				
	other	4.889				
license	no	10.243				
male	no	54.498				
Household						
bicycles			0.000	10.000	3.357	1.937
cars			0.000	10.000	1.383	0.822
income	$<$20	11.832				
	$\geq$ 20–40	42.123				
	$\geq$ 40	46.044				
Build and natural environment						
density			0.002	11.443	1.569	1.593
diversity			0.000	2.828	1.775	0.493
green			0.000	97.813	54.939	22.172
Weather						
precip			0.000	142.300	2.185	4.675
temp			−9.000	35.900	13.317	7.566
wind			0.400	16.300	4.098	1.915

determined by the neuron with the largest output value. Hornik, Stinchcombe, and White (1989) showed that ANNs can approximate arbitrary continuous functions in Euclidean space to any degree of accuracy. In this study, an ANN with a single hidden layer of 48 neurons is used. The connection weights are trained by back propagation with a weight decay constant of 0.1.

CTs utilize a tree-like data structure for classification. The nodes of the tree represent binary decision rules which recursively split the feature space, while the leaves of the tree represent the classes (Breiman, Friedman, Olshen, & Stone, 1984). Classification trees are easy to interpret and can effectively deal with nonlinear relationships and interactions between variables. However, they are sensitive to noisy data and also have a tendency to overfit (Quinlan, 2014). Tree-based ensemble techniques combine many classification trees in order to form more stable and accurate classifiers than single CTs (Breiman, 1996).

The first tree-based ensemble method selected for comparison is boosting (BOOST). Here, the general idea is to build a sequence of CTs, where each successive tree aims to improve the previously wrong classifications of the preceding trees. Prediction is accomplished by a weighted voting among all CTs. Here, the gradient boosting machine variant (Friedman, 2001) is used. 300 trees are fitted in total. The shrinkage parameter is set to 0.2 and the interaction depth to 48. Additionally, each leaf node must have at least 10 observations.

Bagging (BAG) is a straight-forward application of an ensemble of trees, whereby many CTs are trained in parallel using bootstrap samples of the data. For prediction, class assignment is determined by majority voting among all trees. In this study, 350 classification trees are bagged. Each tree is grown without pruning until the class assignment at each node is unambiguous.

RF is another tree-based ensemble method which is closely related to bagging. While RF also trains many CTs in parallel using bootstrap samples, each split at the nodes of the trees is determined by a random subset of variables (Breiman, 2001) Again, for prediction, a majority vote among all trees determines class membership. In this study, an RF consisting of 450 trees is used and three randomly selected variables are considered for each split at the trees nodes.

All modeling and analyses is done in the R programming environment (R Core Team, 2015) using the 'caret' package (Kuhn, 2008). The 'caret' package provides a common interface for several modeling packages. The relevant modeling packages for this study are 'nnet' (Venables & Ripley, 2002), 'klaR' (Weihs, Ligges, Luebke, & Raabe, 2005), 'ipred' (Peters & Hothorn, 2015), 'e1071' (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2015), 'randomForest' (Liaw & Wiener, 2002), and 'gbm' (Ridgeway, 2015).

2.3. Model comparison

The performance of each classifier is estimated in this study using 10-fold cross-validation (Kohavi et al., 1995). This procedure randomly partitions the data into 10 disjoint subsets. One subset at a time is then used for testing the model, while the remaining sets are used to build the model. Consequently, since the testing and training data sets are independent of each other, bias in performance estimation is reduced (e.g. Kohavi et al., 1995).

The distribution of the dependent variable is imbalanced. For instance, trips by car are done very frequently, while trips by public transport are rare. To account for the class-imbalance, the following procedure is suggested for each training subset of the validation procedure. First, the mean number of trips per travel mode, denoted by n, is calculated. Then, for classes which have less than n cases, observations are sampled with replacement from this class and added to the data set until the class consists of n cases. For classes which have more than n cases, observations are randomly removed from the data set until the class consists of n cases. After this procedure, every class is represented by exactly n observations. Thus, the total size of the data set is not changed by this procedure.

The classification performance is measured using the accuracy and sensitivity statistics. Accuracy measures the overall proportion of correctly classified observations, while sensitivity evaluates the proportion of correctly assigned observations for each class (Japkowicz & Shah, 2011). Hence, sensitivity is particular useful for evaluating classification performance on imbalanced data sets. These statistics are calculated for each model built during the validation procedure and for each repetition.

To evaluate and compare the different classifiers, it is useful to take into account the distribution of the performance statistics (e.g. Hothorn et al., 2005). Following Hothorn et al. (2005), this study evaluates the statistical significance of the classifiers' differences in accuracy as follows. First, the Kruskal–Wallis test with a 5% significance level is used to test the null hypothesis that the performance estimates of all classifiers are not systematically different from each other. Then, the two-sided Wilcoxon rank-sum test is applied to determine the statistical significance of systematic pairwise differences between classifiers. To control the false discovery rate at the 5% level, the p-values are adjusted using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

2.4. Variable importance

The assessment of variable importance is generally an important analysis task, because it allows variable selection and supports meaningful interpretation. However, this remains a complex task due to interactions and correlations among the variables. Seemingly irrelevant variables may become important only in the context of others, while redundancies between variables may lead to an overestimation of importance (e.g. Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). In addition, the assessment of variable importance depends strictly on the model under consideration. If a classifier is incapable of modeling a variable's relationships with a response variable, its importance for the classifier is generally low, while its importance might be high for more powerful classifiers.

Numerous approaches for quantifying variable importance for different models have been proposed (e.g. Olden, Joy, & Death, 2004; Hagenauer & Helbich, 2012; Nathans, Oswald, & Nimon, 2012). In the RF framework, the importance of a variable is commonly evaluated by measuring the change in model performance when randomly permuting the variable in the test data (e.g. Breiman, 2001; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). The more the performance decreases under permutation of a variable, the higher is its importance. This approach can be applied to arbitrary prediction models, given that independent test data for evaluation purposes is available (e.g. Knudby, Brenning, & LeDrew, 2010; Xu et al., 2014; Goetz, Brenning, Petschko, & Leopold, 2015).

In this study, a permutation-based approach to measure the overall importance of each variable is used. This is done by permuting each variable within the test data 10 times for each fold and repetition of the validation procedure and reporting the resulting differences in accuracy. However, in a multiclass classification problem such as travel mode choice, the importance of the variables for the prediction of different classes is also of interest. For example, the ownership of a driver's license might be relevant for predicting car trips, but might be less relevant for the prediction of other travel modes. This study is the first that uses the permutation-based approach for analyzing such importances by reporting the differences in sensitivity for each travel mode under permutation.

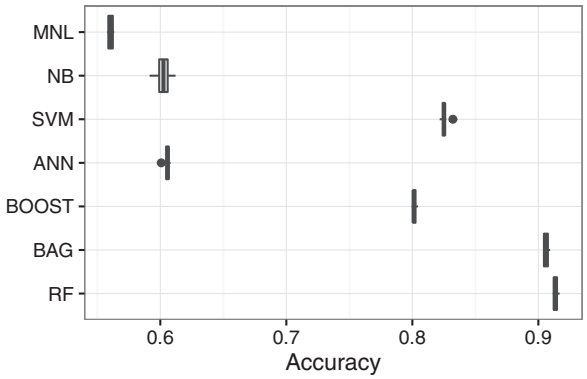

Fig. 1. Accuracy for each classifier.

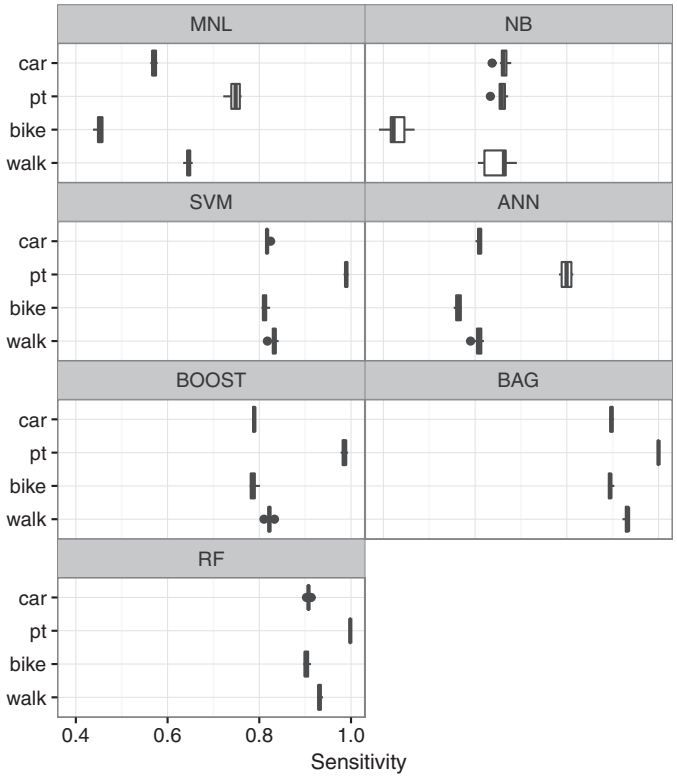

Fig. 2. Sensitivity for each classifier.

Table 3
Results of Wilcoxon rank-sum tests for differences in accuracy. Numbers below the diagonal are *p*-values, the numbers above the estimated differences. The false discovery rate is controlled at 5%.

	MNL	NB	SVM	ANN	BOOST	BAG	RF
MNL		−0.042	−0.265	−0.045	−0.241	−0.346	−0.353
NB	0.000		−0.223	−0.003	−0.199	−0.303	−0.311
SVM	0.000	0.000		0.022	0.024	−0.081	−0.088
ANN	0.000	0.257	0.000		−0.196	−0.301	−0.308
BOOST	0.000	0.000	0.000	0.000		−0.105	−0.122
BAG	0.000	0.000	0.000	0.000	0.000		−0.008
RF	0.000	0.000	0.000	0.000	0.000	0.000	

3. Results

3.1. Classification performance

The accuracy of each classifier is shown in Fig. 1. With respect to median accuracy, RF achieved the best results (0.914), closely followed by BAG (0.906). The third and fourth best classifiers are SVM (0.825) and BOOST (0.801). The accuracy of the other classifiers is substantially lower. The accuracy of ANN (0.606) is only slightly higher than NB (0.602). MNL has the lowest accuracy of all classifiers with 0.561.

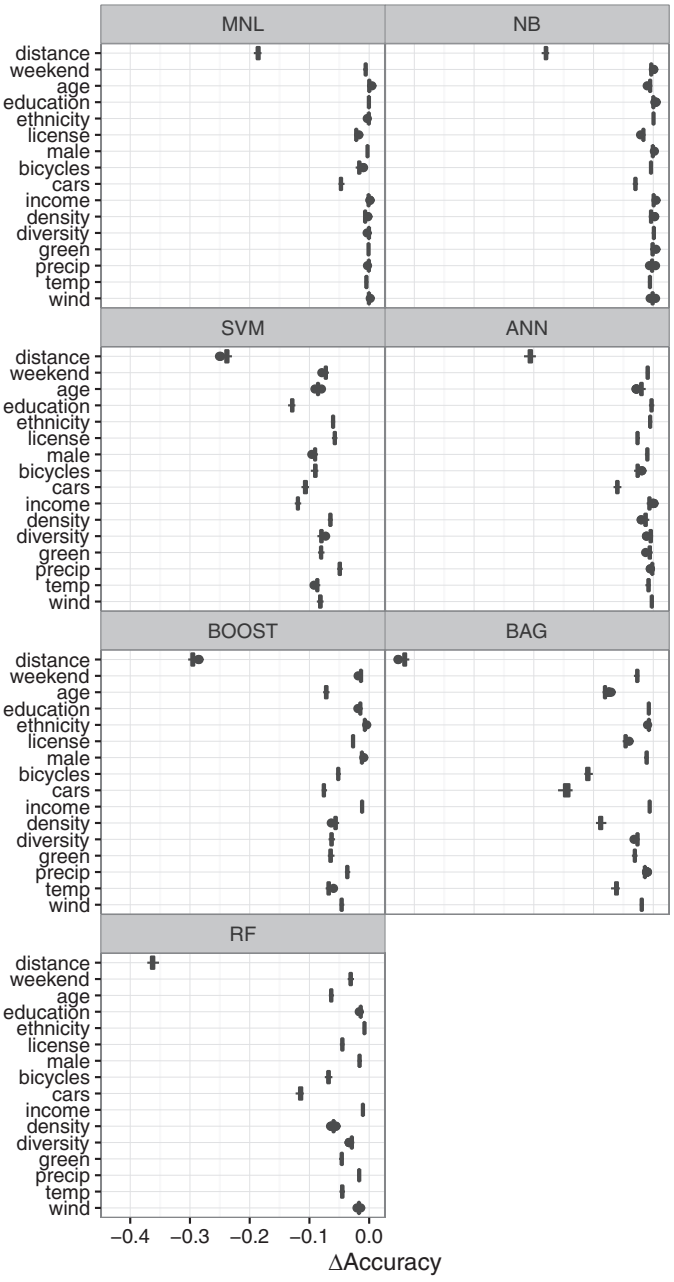

Fig. 3. Overall variable importance.

The null hypothesis of no performance differences between the classifiers was rejected by the Kruskal–Wallis test at 5% significance level. Table 3 shows the results of the two-sided Wilcoxon rank-sum test with adjusted *p*-values. For ANN and NB the null hypothesis that the results are drawn from the same continuous distributions is not rejected. Hence, these classifiers are the only ones whose accuracy is not significantly different.

The sensitivity of the classifiers for each travel mode is shown in Fig. 2. Notably, all classifiers, except NB, predict public transport trips more accurately than other travel modes. NB, however, predicts car trips slightly more accurately than public transport trips. Bike trips are generally less accurately predicted than the

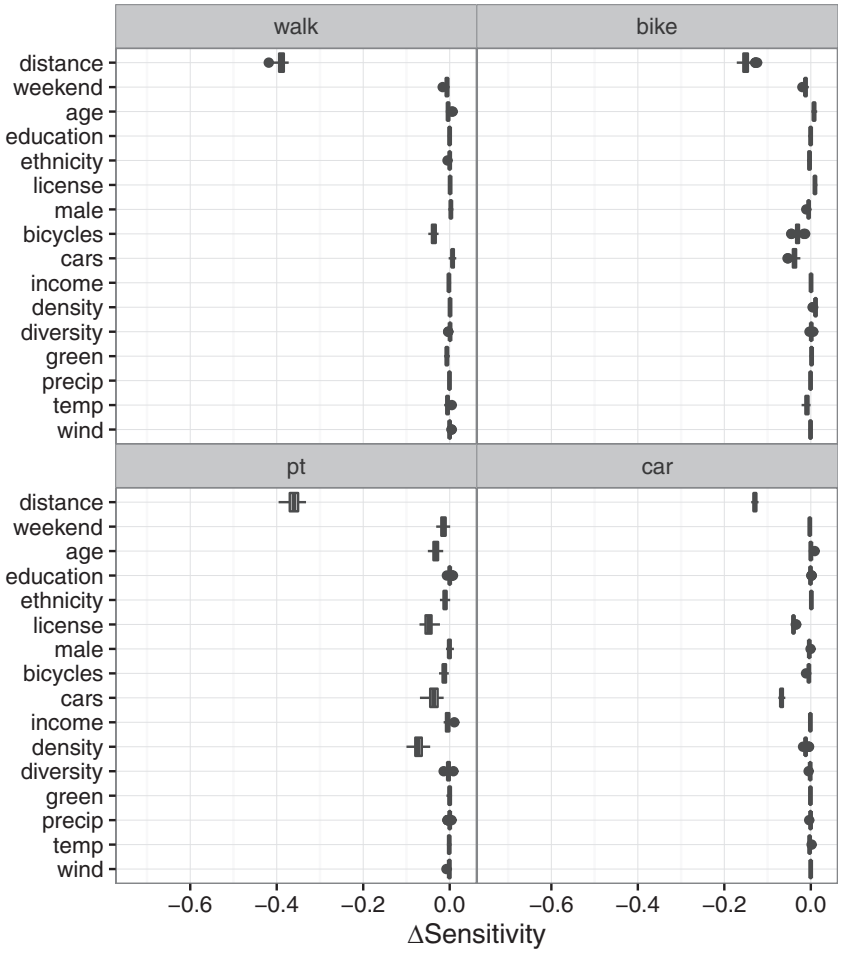

Fig. 4. Variable importance for MNL.

other travel modes, though the absolute sensitivity values of the classifiers differ. For instance, NB predicts bike trips substantially less accurately than the other travel modes, while for RF and SVM the difference in sensitivity between bike and walking trips is only marginal. Moreover, it can be seen that the sensitivity values of SVM, BOOST, BAG, ANN, and RF follow the same patterns. That is, public transport is predicted most accurately, followed by walking trips. The difference in sensitivity between bike and car trips is only marginal. By contrast, the sensitivity values of MNL and NB follow very different patterns.

3.2. Variable importance

Fig. 3 shows boxplots of the importance of each variable for each classifier with respect to accuracy. By far the most important variable for all classifiers is trip distance. For the other variables the ranking of importance is more complex, though address density, age, number of cars and bicycles per household, and the possession of a driving license are of importance for most classifiers. Exceptions are MNL and NB, for which age, address density, and number of bicycles (only NB) are not important. Generally, the number of important variables is smaller for NB, MNL, and ANN than for the other classifiers. For SVM, by contrast, all variables bear substantial importance. In particular, while education and household income are only marginally important for the other classifiers, these variables are the second and third most important variable for SVM. In addition, while in general the meteorological variables are more important for SVM than for the other classifiers, temperature is generally the most important meteorological variable.

Exemplarily, Fig. 4 depicts the importance of the variables with respect to sensitivity for MNL (lowest accuracy), Fig. 5 for BOOST (moderate accuracy), and Fig. 6 for RF (highest accuracy). While trip distance is the most important variable for all travel modes and classifiers, there exist numerous notable differences between classifiers and travel modes. First, the number of important variables varies substantially with travel mode. For example, three variables are substantially important (ΔSensitivity < -0.2) for predicting public transport trips by RF (distance, number of cars, age), but only a single variable (distance) for predicting other trips by RF. Consequently, and second, some variables are more important for certain travel modes than others. For instance, while the address density is important for predicting public transport trips by any classifier, this variable is basically of no importance for predicting car trips. Third, the importance of variables also varies between classifiers. For example, in contrast to MNL, BOOST and RF identify along temperature the proportion of green space as an important variable for predicting public transport and bicycle trips.

4. Discussion

4.1. Classification performance

The tree-based ensemble classifiers performed exceptionally well. This indicates that the flexibility which is obtained by combining multiple CTs is particularly useful for modeling travel mode

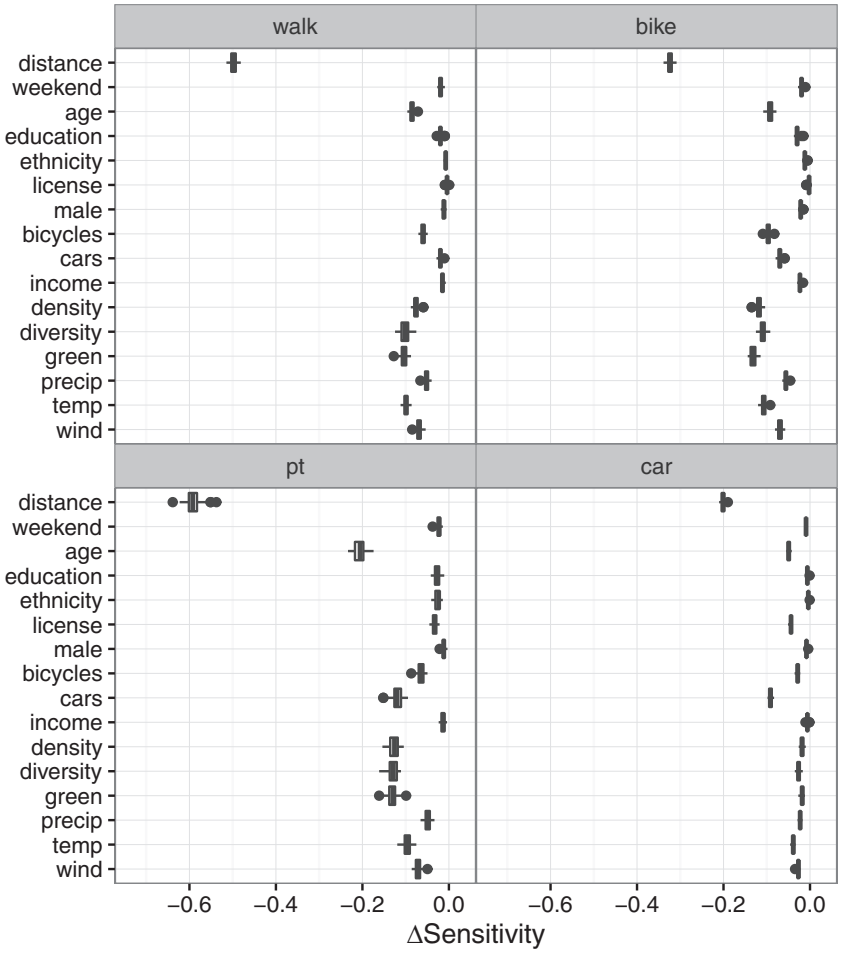

Fig. 5. Variable importance for BOOST.

choice. The performance of RF is significantly better than BAG. This difference can be attributed to the larger diversity among the learned trees of RF, which is a result of the RF's procedure for randomized splitting at nodes. Generally, ensemble classifiers perform better if there is significant diversity among the models (Kuncheva & Whitaker, 2003). However, the performance of BOOST is inferior to both RF and BAG. One explanation can be that boosting methods are primarily designed to minimize model bias and are therefore more prone to overfitting, while RF and BAG conceptually aim to reduce model variance (Ganjisaffar, Caruana, & Lopes, 2011). Furthermore, because boosting methods try to improve previously misclassified data iteratively, outliers can have a critical effect on their performance (Rätsch, Onoda, & Müller, 2001).

SVM uses the one-against-one approach for multiclass classification. This approach generally tends to increase the classifier's variance, because only small subsets of the data are used to learn to distinguish between each pair of classes (Lee, Lin, & Wahba, 2004). In addition, it can lead to inconsistent results in which observations are assigned to multiple classes simultaneously (Bishop, 2006). The similar performance of SVM and BOOST, which is a true multiclass classifier, indicates that these issues do not substantially effect the performance of SVM.

The lowest accuracy was provided by MNL, indicating that its modeling capabilities are generally less effective for modeling travel mode choice and/or that the models' assumptions are substantially violated. These findings are in line with previous studies (e.g. Sayed and Razavi, 2000, Xie et al., 2003, Omrani, 2015). The accuracy of NB, on the other hand, is significantly higher than MNL and close to ANN, indicating that despite the strict independence assumption of NB (e.g. Hand & Yu, 2001) it can be useful for the modeling of travel mode choice.

The results of the sensitivity analysis allow a more detailed investigation of accuracy results. Overall, RF predicts all travel modes with high sensitivity. No classifiers predicts any travel mode more accurately than RF. Thus, RF can be considered the most appropriate classifier for modeling travel mode choice. In addition, since the sensitivity of SVM, BOOST, BAG, ANN, and RF basically follow the same patterns, it can be concluded that neither of these classifiers has distinct properties that make it substantially more useful for predicting certain travel modes.

4.2. Variable importance

The results of the analysis of variable importance with respect to accuracy show that the considered classifiers, except SVM, generally correspond well with regard to the most important variables, though the magnitudes of variable importance between classifiers differ. In particular, even simple classifiers such as MNL and NB are able determine the most important variables. However, because MNL and NB do only consider a rather small set of variables as important for classification and their generally low classification performance, it can be concluded that more advanced and flexible classifiers are required to model the complex interactions and relationships of most variables.

The similar patterns of variable importance of RF and BAG indicates that these classifiers model relationships between variables in a similar manner. However, the magnitude of importance of some variables is different for RF and BAG, even though both

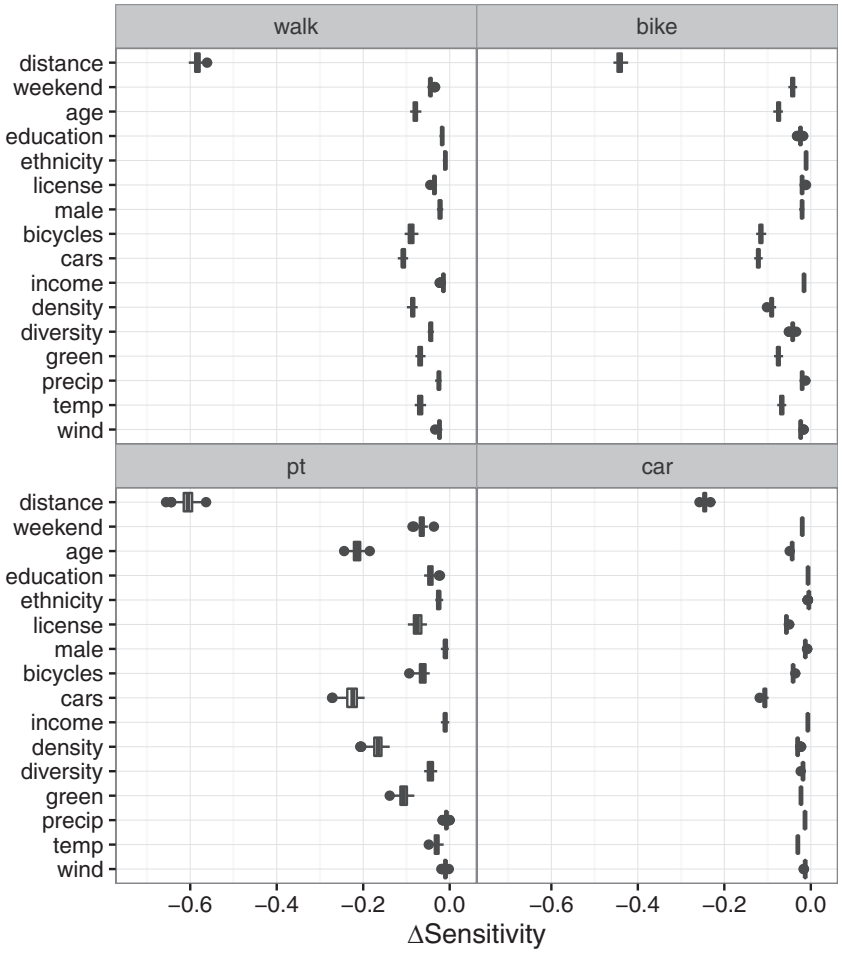

Fig. 6. Variable importance for RF.

classifiers are based on an ensemble of CTs. For instance, address density is more important for BAG than RF, while the proportion of green space and education is more important for RF than BAG. One explanation for these differences can be that RF, in contrast to BAG, creates CTs on random subsets in order to avoid overfitting (Ho, 1998).

The results also confirm the findings of Rasouli and Timmermans (2014), who found that trip distance is the most important variable and, furthermore, that age is more important than income or education when classifying travel mode choice using tree-based ensemble classifiers. While in this article the number of cars is ranked as the second most important variable by all classifiers, except SVM, Rasouli & Timmermans (2014) reported that car availability is not of substantial importance. A reason for this difference could be that this article considered the total number of available cars, while Rasouli & Timmermans (2014) merely considered the general availability of a car, disregarding the total number of cars available.

Previous studies have also identified weather conditions as important variables for making decision about travel modes (e.g. Verplanken, Aarts, & Van Knippenberg, 1997; Garvill, Marell, & Nordlund, 2003; Helbich et al., 2014; Liu et al., 2015), which is confirmed by the results of the present study. However, the results also show that weather variables do not play a dominating role for travel mode choice classification and that temperature is generally more important than precipitation or wind speed. An explanation for the latter can be that temperature also reflects seasonal effects, which are not directly related to weather conditions but nevertheless influence travel mode choice (Clifton, Chen, & Cutter, 2011).

In addition, the results emphasize the overall importance of address density for most classifiers. An explanation for this can be that address density is closely related to variables such as parking costs, distance to public transport stations, and travel time, which have been shown to significantly influence the choice of travel modes (e.g. Frank, Bradley, Kavage, Chapman, & Lawton, 2008; Susilo, Williams, Lindsay, & Dair, 2012).

The analysis of sensitivity allows a more detailed view of the importance of the variables for the different classifiers. For instance, the results show that address density is generally more important for the prediction of public transport trips than for the other travel modes. Considering that address density is a proxy for population density, these results correspond to Limtanakool, Dijst, and Schwanen (2006), who determined that high population density is associated with an increased use of public transport.

Finally, the results also indicate that temperature and proportion of green space are particularly important for predicting bicycle trips by RF. This supports the results of Winters, Brauer, Setton, and Teschke (2010) and Helbich et al. (2014), who showed that the natural and built environment, as well as temperature, substantially affect bicycle behavior. In addition, the results show that these variables are also important for predicting public transport trips. This is in line with the findings of Nankervis (1999), who showed that public transport is a common alternative to cycling, particularly during bad weather conditions.

5. Conclusion

This article presented a systematic comparison of seven different machine learning classifiers for travel mode choice prediction using Dutch travel diary data from the years 2010 to 2012. For this purpose, accuracy and sensitivity analyses have been performed utilizing repeated *k*-fold cross validation.

The results showed that among the investigated classifiers, RF produced the most accurate predictions. The performance of MNL, arguably the most common model for analyzing travel mode choice, is low. In-depth sensitivity analysis revealed that public transport and car trips are predicted with the highest sensitivity by all classifiers, while walking and bicycles trips are predicted with the lowest sensitivity.

Using a permutation-based approach to measure variable importance, the article showed that with regard to accuracy the most important variable is trip distance, followed by the number of cars per household. The importance of the other variables varies with the applied classifiers. Though generally of little importance, the meteorological variables are more important for SVM than for the other classifiers. Furthermore, a detailed analysis of variable importance with regard to sensitivity has shown that variable importance also varies strictly with the travel mode being predicted. Temperature is more important for predicting public transport and bicycle trips than for other travel modes.

The results suggest that it is necessary to analyze alongside the overall classification performance the importance of variables for the different classifiers and travel modes in order to get a better understanding of the relationships within the data and to allow effective modeling of travel mode choice.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.eswa.2017.01.057.

References

Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand*: 9. MIT press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Böcker, L., van Amen, P., & Helbich, M. (2016). Elderly travel frequencies and transport mode choices in greater rotterdam, the netherlands. *Transportation*, 1–22.

Böcker, L., Dijst, M., & Prillwitz, J. (2013). Impact of everyday weather on individual daily travel behaviours in perspective: A literature review. *Transport Reviews, 33*(1), 71–91.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

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

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks.

Cantarella, G. E., & De Luca, S. (2003). Modeling transportation mode choice through artificial neural networks. In *Fourth international symposium on uncertainty modeling and analysis, 2003* (pp. 84–90).

Celikoglu, H. B. (2006). Application of radial basis function and generalized regression neural networks in non-linear utility function specification for travel mode choice modelling. *Mathematical and Computer Modelling, 44*(7), 640–658.

Clifton, K. J., Chen, R. B., & Cutter, A. (2011). Representing weather in travel behaviour models: A case study from sydney, AUS. In *Australasian transport research forum 2011 proceedings* (pp. 28–30).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Dieleman, F. M., Dijst, M., & Burghouwt, G. (2002). Urban form and travel behaviour: Micro-level household attributes and residential context. *Urban Studies, 39*(3), 507–527.

de Dios Ortúzar, J., & Willumsen, L. G. (2011). *Modelling transport*. John Wiley & Sons.

Dow, J. K., & Endersby, J. W. (2004). Multinomial probit and multinomial logit: A comparison of choice models for voting research. *Electoral Studies, 23*(1), 107–122.

Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American Planning Association, 76*(3), 265–294.

Ewing, R., Schroeer, W., & Greene, W. (2004). School location and student travel analysis of factors affecting mode choice. *Transportation Research Record: Journal of the Transportation Research Board*, (1895), 55–63.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research, 15*(1), 3133–3181.

Fishman, E., Böcker, L., & Helbich, M. (2015). Adult active transport in the netherlands: An analysis of its contribution to physical activity requirements. *PloS one, 10*(4), e0121871.

Frank, L., Bradley, M., Kavage, S., Chapman, J., & Lawton, T. K. (2008). Urban form, travel time, and cost relationships with tour complexity and mode choice. *Transportation, 35*(1), 37–54.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Ganjisaffar, Y., Caruana, R., & Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval* (pp. 85–94).

Garvill, J., Marell, A., & Nordlund, A. (2003). Effects of increased awareness on choice of travel mode. *Transportation, 30*(1), 63–79.

Goetz, J., Brenning, A., Petschko, H., & Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & Geosciences, 81*, 1–11.

Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving personal trip data from gps data: A literature review on the existing methodologies. *Procedia-Social and Behavioral Sciences, 138*, 557–565.

Hagenauer, J., & Helbich, M. (2012). Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. *International Journal of Geographical Information Science, 26*(6), 963–982.

Hand, D. J., & Yu, K. (2001). Idiot's bayes—not so stupid after all? *International Statistical Review, 69*(3), 385–398.

Helbich, M. (2016). Children's school commuting in the netherlands: Does it matter how urban form is incorporated in mode choice models? *International Journal of Sustainable Transportation*.

Helbich, M., Böcker, L., & Dijst, M. (2014). Geographic heterogeneity in cycling under various weather conditions: Evidence from greater rotterdam. *Journal of Transport Geography, 38*, 38–47.

Hensher, D. A., & Ton, T. T. (2000). A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. *Transportation Research Part E: Logistics and Transportation Review, 36*(3), 155–172.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832–844.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366.

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics, 14*(3), 675–699.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification. *Technical Report*. Department of Computer Science, National Taiwan University.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on, 17*(3), 299–310.

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 338–345). Morgan Kaufmann Publishers Inc..

Knudby, A., Brenning, A., & LeDrew, E. (2010). New approaches to modelling fish–habitat relationships. *Ecological Modelling, 221*(3), 503–511.

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on artificial intelligence: 14* (pp. 1137–1145).

Koppelman, F. S., & Wen, C.-H. (1998). Alternative nested logit models: Structure, properties and estimation. *Transportation Research Part B: Methodological, 32*(5), 289–298.

Kuhn, M. (2008). Caret package. *Journal of Statistical Software, 28*(5).

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning, 51*(2), 181–207.

Lee, Y., Lin, Y., & Wahba, G. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association, 99*(465), 67–81.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News, 2*(3), 18–22.

Limtanakool, N., Dijst, M., & Schwanen, T. (2006). The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium-and longer-distance trips. *Journal of Transport Geography, 14*(5), 327–341.

Liu, C., Susilo, Y. O., & Karlström, A. (2015). The influence of weather characteristics variability on individuals travel mode choice in different seasons and regions in sweden. *Transport Policy, 41*, 147–158.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics*. New York, NY: Academic Press.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2015). e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien. R package version 1.6–7.
Murray, K., & Conner, M. M. (2009). Methods to quantify variable importance: Implications for the analysis of noisy ecological data. *Ecology, 90*(2), 348–355.
Nankervis, M. (1999). The effect of weather and climate on bicycle commuting. *Transportation Research Part A: Policy and Practice, 33*(6), 417–431.
Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation, 17*(9), 1–19.
Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling, 178*(3), 389–397.
Omrani, H. (2015). Predicting travel mode of individuals by machine learning. *Transportation Research Procedia, 10*, 840–849.
Omrani, H., Charif, O., Gerber, P., Awasthi, A., & Trigano, P. (2013). Prediction of individual travel mode with evidential neural network model. *Transportation Research Record: Journal of the Transportation Research Board*, (2399), 1–8.
Onderzoek Verplaatsingen in Nederland (2014). Onderzoeksbeschrijving ovin 2010–2014. data archiving and networked services (dans). http://www.cbs.nl/nl-NL/menu/themas/verkeer-vervoer/methoden/dataverzameling/korte-onderzoeksbeschrijvingen/ovin-beschrijving-art.htm. Accessed on 27th January 2016.
Peters, A., & Hothorn, T. (2015). ipred: Improved predictors. R package version 0.9-5.
Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. Elsevier.
R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical ComputingVienna, Austria.
Rasouli, S., & Timmermans, H. J. (2014). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *European Journal of Transportation and Infrastructure Research, 14*(4), 412–424.
Rätsch, G., Onoda, T., & Müller, K.-R. (2001). Soft margins for adaboost. *Machine Learning, 42*(3), 287–320.
Ridgeway, G. (2015). gbm: Generalized boosted regression models. R package version 2.1.1.
Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.
Rojas, R. (2013). *Neural networks: A systematic introduction*. Springer Science & Business Media.
Sayed, T., & Razavi, A. (2000). Comparison of neural and conventional approaches to mode choice analysis. *Journal of Computing in Civil Engineering, 14*(1), 23–30.
Schwanen, T., & Mokhtarian, P. L. (2005). What affects commute mode choice: Neighborhood physical structure or preferences toward neighborhoods? *Journal of Transport Geography, 13*(1), 83–99.
Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation, 42*(1), 163–188.
Shen, L., & Stopher, P. R. (2014). Review of gps travel survey and gps data-processing methods. *Transport Reviews, 34*(3), 316–334.
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 1.
Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 1.
Susilo, Y. O., Williams, K., Lindsay, M., & Dair, C. (2012). The influence of individuals environmental attitudes and urban design features on their travel patterns in sustainable neighborhoods in the uk. *Transportation Research Part D: Transport and Environment, 17*(3), 190–200.
Tang, L., Xiong, C., & Zhang, L. (2015). Decision tree method for modeling travel mode switching in a dynamic behavioral process. *Transportation Planning and Technology, 38*(8), 833–850.
Tortum, A., Yayla, N., & Gökdağ, M. (2009). The modeling of mode choices of intercity freight transportation with the artificial neural networks and adaptive neuro-fuzzy inference system. *Expert Systems with Applications, 36*(3), 6199–6217.
Tsai, T.-H., Lee, C.-K., & Wei, C.-H. (2009). Neural network based temporal feature models for short-term railway passenger demand forecasting. *Expert Systems with Applications, 36*(2), 3728–3736.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (4th). New York: Springer. ISBN 0-387-95457-0.
Verplanken, B., Aarts, H., & Van Knippenberg, A. (1997). Habit, information acquisition, and the process of making travel mode choices. *European Journal of Social Psychology, 27*(5), 539–560.
Weihs, C., Ligges, U., Luebke, K., & Raabe, N. (2005). klar analyzing german business cycles. In D. Baier, R. Decker, & L. Schmidt-Thieme (Eds.), *Data analysis and decision support* (pp. 335–343). Berlin: Springer-Verlag.
Winters, M., Brauer, M., Setton, E. M., & Teschke, K. (2010). Built environment influences on healthy transportation choices: Bicycling versus driving. *Journal of Urban Health, 87*(6), 969–993.
Xian-Yu, J. (2011). Travel mode choice analysis using support vector machines. *ICCTP 2011: American Society of Civil Engineers*, 360–371.
Xie, C., Lu, J., & Parkany, E. (2003). Work travel mode choice modeling with data mining: decision trees and neural networks. *Transportation Research Record: Journal of the Transportation Research Board*, (1854), 50–61.
Xu, L., Li, J., & Brenning, A. (2014). A comparative study of different classification techniques for marine oil spill identification using radarsat-1 imagery. *Remote Sensing of Environment, 141*, 14–23.
Zhan, G., Yan, X., Zhu, S., Wang, Y., et al. (2016). Using hierarchical tree-based regression model to examine university student travel frequency and mode choice patterns in china. *Transport Policy, 45*(C), 55–65.
Zhang, Y., & Xie, Y. (2008). Travel mode choice modeling with support vector machines. *Transportation Research Record: Journal of the Transportation Research Board*, (2076), 141–150.
Zhao, D., Shao, C., Li, J., Dong, C., & Liu, Y. (2010). Travel mode choice modeling based on improved probabilistic neural network. In *Traffic and transportation studies 2010* (pp. 685–695). ASCE.