EJTIR

Issue 14(4), 2014
pp. 412-424
ISSN: 1567-7141
tlo.tbm.tudelft.nl/ejtir

Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates

Soora Rasouli[1]
Urban Planning Group, Eindhoven University of Technology, The Netherlands.

Harry J.P. Timmermans[2]
Urban Planning Group, Eindhoven University of Technology, The Netherlands.

The application of activity-based models of travel demand to planning practice has triggered interest in issues that potentially improve the accuracy and/or usefulness of model forecasts. The limited knowledge of uncertainty propagation in complex stochastic model systems has put uncertainty analysis high on the research agenda to differentiate between simulation error and policy effects. Focusing on transport mode choice, this paper draws attention to the use of model ensembles, which has hardly been explored in travel demand forecasting. Prior studies predicting transport mode choice has typically relied on a single equation, relating observed transport mode choices to a set of personal and contextual variables. The estimated single model is then assumed to apply to all individuals. This paper explores the idea of replacing a single equation/representation with an ensemble of model predictions, using the decision tree formalism. Potentially, ensembles better capture the notion that travellers may use different heuristics in their transport mode decisions. The aim of the study is to investigate whether the use of a model ensemble of different decision heuristics will reduce the error/uncertainty in predicting transport mode decisions. Results of the study, conducted in the Rotterdam region, The Netherlands, suggest that the accuracy of predicting transport mode choice is improved, albeit non-monotonically, with increasing ensemble size. Simultaneously, the uncertainty related to these predictions is decreasing. Finally, it is shown that the importance of the selected explanatory variables co-varies with ensemble size. Estimation results tend to become stable in this study with an ensemble size of approximately 20 decision trees

Keywords: *transport mode choice, ensembles of decision trees, uncertainty*

1. Introduction

Over the last decade, activity-based models of travel demand have gradually found increasing application in travel demand forecasting practice (Vovsha, et al., 2005; Henson, et al., 2009; Rasouli and Timmermans, 2014). Contemporary activity-based models of travel demand are based on stochastic model specifications. One of the interpretations of travel demand forecasting models based on random utility theory (e.g., Bowman and Ben-Akiva, 2000; Bhat, et al., 2004; Pendyala, et al., 2005) is that individual travellers hold inherently stochastic preferences. It implies that, under the postulate of utility-maximizing behaviour, travellers may choose different

[1] PO Box 513, 5600 MB Eindhoven, The Netherlands T: +31 040 2473315 E: s.rasouli@tue.nl
[2] PO Box 513, 5600 MB Eindhoven, The Netherlands T: +31 040 2473315 E: h.j.p.timmermans@tue.nl

alternatives on successive choice occasions, even under otherwise equivalent conditions. In the case of rule-based models or computational process models, such as Albatross (Arentze and Timmermans, 2000; 2004; Rasouli and Timmermans, 2013) and Tasha (Roorda, 2005; Roorda and Miller, 2005; Roorda et al., 2005, 2008), stochasticity is reflected in the use of probabilistic decision tables or decision rules. Theoretically, the use of probabilistic decision tables implies the assumption that travellers will activate any of the decision rules, represented by the decision table, with some probability of choice.

The application of these stochastic travel demand forecasting models to predict the effects of policy scenarios involves (Monte Carlo) sampling from the assumed error distributions (in case of econometric models) or from probabilistic decision tables (in case of rule-based models). Consequently, regardless of the theoretical behavioural interpretation of stochasticity of the activity-based model, different sampled error terms or branches of the decision tree, will result in differences in predicted activity-travel patterns, and therefore in possibly different assessments of policy effects. This situation has led to the understanding that Monte Carlo simulation error should be separated from policy effects in applications of micro-simulation and agent-based modelling of travel demand. It generated an increased awareness of the relevance of uncertainty analysis in applied transportation planning practice (Veldhuisen, et al., 2000; Castiglione, et al., 2003; Beger Hugosson, 2005; de Jong et al., 2007, Ziems et al., 2011). Nevertheless, the number of studies on uncertainty analysis in travel demand forecasting is still small (see Rasouli & Timmermans, 2011 for an overview). Existing studies on uncertainty analysis have typically attempted to quantify the amount of uncertainty in particular types of travel demand forecasts (travel behaviour indices, OD-matrices, traffic flows) as a function of input and/or model uncertainty. Uncertainty analysis allows policy makers to assess the confidence levels that are associated with model forecasts due to input and/or model uncertainty. The majority of these studies on uncertain analysis have focused on the four-step model (e.g. Zhao and Koppelman, 2001) or on discrete choice models (e.g. de Jong, et al., 2007; Zhang et al., 2011). Studies, examining uncertainty in forecasts of rule-based models of activity-travel demand have been confined to Albatross (Kwak et al., 2012; Rasouli and Timmermans, 2013) and its Flamish equivalent (Cools et al., 2011; Rasouli, et al., 2012).

The practice of model development and application, however, has not changed: modellers have continued identifying the single best performing model and use its estimated parameters to predict the effects of scenarios in terms of a set of performance indicators. Monte Carlo sampling has been used to estimate the uncertainty surrounding predicted performance indicators and in some rare cases calculate corresponding confidence levels. Although some further approaches dealing with uncertainty in complex model systems can be envisioned to enrich policy recommendations (Rasouli and Timmermans, 2012), it is arguable that the development of a single model/equation is necessarily the best approach considering the uncertainty in forecasting due to inherent variability in travel behaviour of individuals and households, within and between contexts. Current research has examined the effect of the size and distribution of error terms, but not the effect of using a set of different representations of the underlying choice process.

To examine this issue, this paper explores the potential of using *ensembles* of decision trees to predict transport mode choice. Each decision tree combines a subset of the explanatory variables selected to predict the behaviour of interest: in this case transport mode choice. This notion is similar to the concept of random forests in the machine learning research community (Breiman, 2001). Random forest algorithms have found a much wider application in classification studies in many other disciplines, including ecology (e.g., Heung, et al., 2014; Puissant, et al., 2014), pattern recognition (Désir, 2013; Ye, et al., 2013), chemistry (Lee, et al., 2013; Ai, et al., 2014), biomedicine (e.g., Yao, et al., 2013; Taher Azar, et al. 2014), and remote sensing (e.g., Abdel-Rahman, et al., 2014). In travel behaviour research, applications have been primarily limited to the application of

random forests to the detection of transport modes using modern technology such as GPS (e.g., Liu, et al., 2013; Lu, et al., 2013).

However, we are not aware of any study in travel behaviour research, applying this approach to address the issue of behavioural heterogeneity to improve demand model forecasting. Different decision tables may pick up different sources of uncertainty and variability in the data. Thus, from a pure technical point of view, one would expect that error in model forecasting will be reduced by applying model ensembles, as opposed to single model, as for example done in weather forecasting. However, similar to viewing the logit model as a mathematical representation of random utility theory (stochastic preferences and utility-maximizing behaviour), decision tables can be viewed as formalism of decision heuristics, indicating which choices will be made in a particular context by individuals and households with a certain profile. Thus, the application of model ensembles as a multitude of different decision tables theoretically means that we allow for differences in decision heuristics, which is an a largely still underexplored topic in travel behaviour research (Hess and Stathopoulos, 2014). The application of mixed logit and latent class model allow for differences in utility parameters, but the structural specification of the utility function itself is typically identical for each class.

The paper is organised as follows. First, we will we will discuss in detail the problem that we used in this study to investigate the effects of using ensembles of decision trees to predict transport mode choice behaviour. Next, in section 3, we will provide a description of the data that was used for the analysis. This is followed by a discussion of the major findings of the analyses. We will complete the paper by drawing conclusions and discussing possible avenues of future research.

2. The problem

In this study, we consider the problem of transport mode choice. Over the years, a myriad of studies have modeled this choice problem (e.g. Caulfield and Brazil, 2011; Ferdous, et al., 2011; Jiao, et al., 2011; Maley and Weinberger, 2011; Susilo, et al., 2011). The vast majority of these studies has been based on random utility theory and used the conventional multinomial logit model or more advanced discrete choice models to predict the probability that a particular transport mode will be chosen as a function of its attribute levels, relative to the attribute levels of competing transport models, and a set of socio-economic, built environment and context variables.

Fewer studies have used decision trees or decision tables. A decision table consists of an action state (in this case representing which transport mode will be chosen) and a set of condition states, which may represent both attribute levels of competing transport modes and/or categories of socio-demographic variables. A decision table represents the sets of conditions under which a certain transport mode is being selected. Thus,

if $C_1 \in CD_1 \wedge C_2 \in CD_2 \wedge$ .....$\wedge C_j \in CS_{j1} \wedge \ldots. \wedge C_m \in CD_m$ then choose A1
if $C_1 \in CD_1 \wedge C_2 \in CD_2 \wedge$ .....$\wedge C_j \in CS_{j2} \wedge \ldots. \wedge C_m \in CD_m$ then choose A2
etc

where CD_i represents the domain of condition variable C_i, and $CS_{j1} \cup CS_{j2} = CD_j$ and $CS_{j1} \cap CS_{j2} = \varnothing$

In our study, the action states represent different transport modes. The domain of the condition variables includes socio-demographics, mode availability and distance. The requirements $CS_{j1} \cup CS_{j2} = CD_j$ and $CS_{j1} \cap CS_{j2} = \varnothing$ ensure that the domains of any condition variable are exhaustive ($CS_{j1} \cup CS_{j2} = CD_j$) and mutually exclusive ($CS_{j1} \cap CS_{j2} = \varnothing$) . The rules then indicate the set of conditions that need to be met in order to select a particular action (A). Note that different

formalisms, such as IF, THEN, …., ELSE production rules, decision trees, or decision tables may be used to represent these decision heuristics.

Decision tables or decision trees (simply another representation) can be extracted from empirical observations using a variety of algorithms, such as CHAID and C4.5, which differ primarily in terms of the split criteria being used to create the branches of the tree. Examples of the use of decision trees in predicting transport mode choice decisions include Arentze and Timmermans (2000, 2004), Xie, et al. (2003), and Anggraini, et al. (2011).

These prior applications of decision trees (and advanced discrete choice models for that matter) in transportation research have relied on a single tree, which specifies the set of conditions under which a certain transport mode will be chosen. Behaviorally, this implies in the case of deterministic decision trees that all travelers belonging to the same segment according to their socio-demographic profile are assumed to choose the same transport mode. In case of probabilistic decision trees (e.g., Arentze and Timmermans, 2004), all individual with the same profiles exhibit the same probabilistic choice behavior under the same set of conditions. Behavioral heterogeneity is thus incorporated in the sense that sampling from the conditional probabilities may result in different transport mode choices. However, by using probabilistic decision tables or trees which specify the probability that a certain transport mode will be chosen also in this case the conditions that impact the probabilistic choice are the same. Thus, under the same set of condition states, the probability of choosing the different transport modes is the same for all individuals belonging to that segment. There is a single decision tree that splits the condition states into homogeneous segments that are assumed to exhibit identical (probabilistic) choice behavior.

One might argue that different travelers may use different (subsets) of conditions or factors to decide on the choice of transport mode. Technically, this implies that different decision trees with a varying number of factors/conditions should be extracted from the data. One way of achieving that is to use an ensemble of simple decision trees, each producing a response, dependent on a set of conditions. An example of such an approach is the random forest (Breiman, 2001). Ensembles of decision trees are created as follows:

1. Draw *K* bootstrap samples from a subset of the original sample (the training set), and use these for decision tree induction. The remainder of the original sample is used for validation and deriving estimates of the importance of the conditions.

2. For each bootstrap sample, a single decision tree is derived by recursively partitioning that sample using a subset of randomly selected condition variables for each split. The best split on this subset is used to partition the node.

3. The number of randomly selected explanatory variables is held constant during the tree induction process. Trees are not pruned.

4. The final predicted action is the one that was predicted most across the ensemble of decision trees.

Transport mode choice is predicted by combining these decision trees based on the votes over the predictions of the single decision trees. Although we argue that the different trees may show behavioural heterogeneity, it should be mentioned that this approach does not unique link each individual to a particular trees as done in latent class models. In that sense, the use of ensembles has a strong technical component that picks up the variability in the data. Therefore, some scholars may simply see the use of ensembles as a technical improvement/ variation.

Table 1. Frequency distributions of selected variables

Attribute	Categories	Percent	Cumul %	Attribute	Categories	Percent	Cumul %
Number of persons in household	1	40.7	40.7	Income	15000<<22500 Euro/year	13.4	59.5
	2	38	78.7		22500<<30000 Euro/year	12	71.5
	3	7.5	86.2		<30000 Euro/year	12	83.5
	4	9.6	95.8		Unknown	16.5	100
	5	3.2	99	Driver's license	Younger than 18	15.8	15.8
	6	0.7	99.7		Yes	57.4	73.2
	7	0.3	100		No	26.9	99.8
Car availability	Yes	40.5	40.5		Unknown	0.2	100
	No	59.5	100	Number of family member younger than 6	0	92.6	92.6
Gender	Male	45.5	45.5		1	5	97.6
	Female	54.5	100		2	2.4	100
Age	0-17 years old	15.83	15.83	Number of family member 6-11 years old	0	91.6	91.6
	18-49 years old	42.12	57.95		1	5.4	97
	50-74 years old	29.88	87.83		2	2.7	99.7
	75 or older	12.17	100		3	0.3	100
Education	Younger than 12	10.2	10.2	Distance class	No movement	19.3	19.3
	Elementary school	16.4	26.6		0.1-1 km	11.6	30.9
	Low level education	25.6	52.2		1-3.7 km	23.1	54
	Medium and high level education	22.5	74.7		3.7-7.5 km	16.2	70.2
	University	20.5	95.2		7.5-15 km	12.3	82.5
	Other	0.6	95.8		15-30 km	8.3	90.8
	Unknown	4.2	100		30-50 km	3.4	94.2
Income	Younger than 12	10.2	10.2		>50 km	5.8	100
	No income	13.6	23.7				
	<7500 Euro/year	6.2	29.9				
	7500<<15000 Euro/year	16.3	46.1				

A few concepts are critical to interpret the results of this approach. The concept of risk estimate is calculated as the proportion of cases incorrectly classified by the (ensemble of) decision tree(s). The standard error of this risk estimates can also be calculated and in a way says something

about the uncertainty of any misclassifications. The concept of variable importance of a condition variable measures the sum of reduction in uncertainty (misclassification) across all nodes of the decision tree relative to the largest sum found across all condition variables. This means that the variable importance of the condition variable that is most successful in predicting the splits is set at 100.

3. Data

The data used in this study stem form the Dutch National Travel Survey called MON (*Mobiliteit Onderzoek Nederlands* – Mobility Research Netherlands). MON is a continuous travel survey, conducted to obtain travel and activity information of a representative sample of residents in the Netherlands. Although the data collection instrument is primarily based on a trip-diary, details about activities are also collected in addition to the usual individual and household socio-demographics such as age, household composition, education level, income level, vehicle availability, and residential location. Respondents complete a questionnaire about all their trips and activities made within 24 hours of a designated day. This study is based on a subset of the MON 2004 data, pertaining to respondents living in the Rotterdam area. This sub-sample includes 1446 respondents. Half this set was used for training, the other half for testing. The size of the subset of conditions is given by Log (M+1), where M is the number of variables in the training set.

The dependent variable is transport mode choice. The MON data include the following categories for this variable: no transportation, car driver, car passenger, train, bus/tram/metro, motor, bike, walking, and other. In addition, the set of condition variables, listed in Table 1 were selected.

4. Analyses and results

The main goal of the analysis was to examine the effects of an increasing number of decision trees (ensemble size) on the uncertainty in the prediction of transport mode. As transport mode is a categorical variable, uncertainty was measured in terms of the share of off-diagonal elements in the confusion matrix. Note that if the ensemble of decision trees predicts the observed transport mode choice perfectly, the confusion matrix will only have predicted frequencies in its diagonal. Any off-diagonal entries in the confusion matrix depict prediction error.

Figure 1 displays the change in risk estimates as a function of the number of decision trees for the training data. It demonstrates the general tendency of the risk estimate, which represents the proportion of wrong predictions of transport mode, to decrease, albeit not monotonically, as a function of the number of decision trees. In particular, the risk estimate drops substantially from 5 to 20 trees. After that, the drop is less and sometimes it may temporarily increase again. The standard deviation of the risk estimation for the training data, which is portrayed in Figure 2 shows a similar tendency. Note however, that changes in standard errors with increased ensemble size are relatively small.

Figures 3 and 4 depict the corresponding results for the test data set. As shown, the pattern is similar to that observed for the training data set with a higher risk estimate, which is to be expected. These figures also show that sometimes (ensemble size 35 and again at ensemble size 100) uncertainty increases again. It seems to reflect the randomness in the sampling process.

Figure 5 shows the corresponding values for the predictive success of the ensembles as a function of the number of decision trees. As discussed, this graph more or less represents the opposite to the risk estimate as it quantifies the percentage correct predictions. The graph shows that transport mode choices of sample respondents are better predicted with increasing ensemble size.

The highest increment in the percentage correct predictions is obtained when the number of decision trees is increased from five to ten. After that, predictive success tailors off.

Finally, Figure 6 shows the chances in the importance of the variables. Results shows that distance is the most important explaining transport mode choice, followed by age and purpose, income, number of persons, car availability, driver's license, household age and gender. Variable importance is not invariant across ensemble size. For example, Figure 6 shows that purpose takes a second position until the largest number of decision trees where it drops to the third position. Especially, up to an ensemble size of 10, the importance of particular variables may swap. After 20 number of decision trees, variable importance tends to stable more or less.

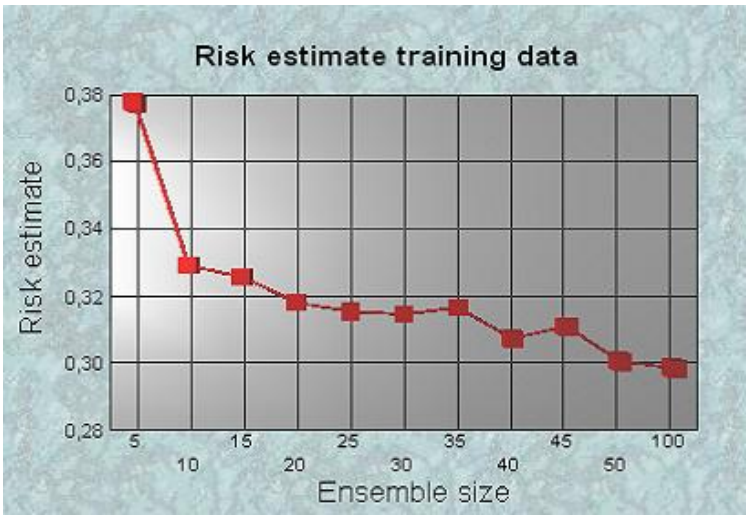

Figure 1: Relationship between ensemble size and risk estimate for training data

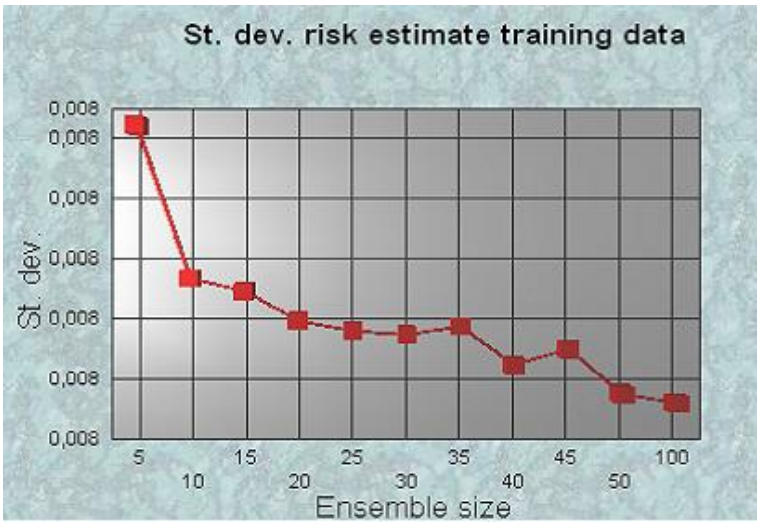

Figure 2: Relationship between ensemble size and standard deviation of risk estimate for training data

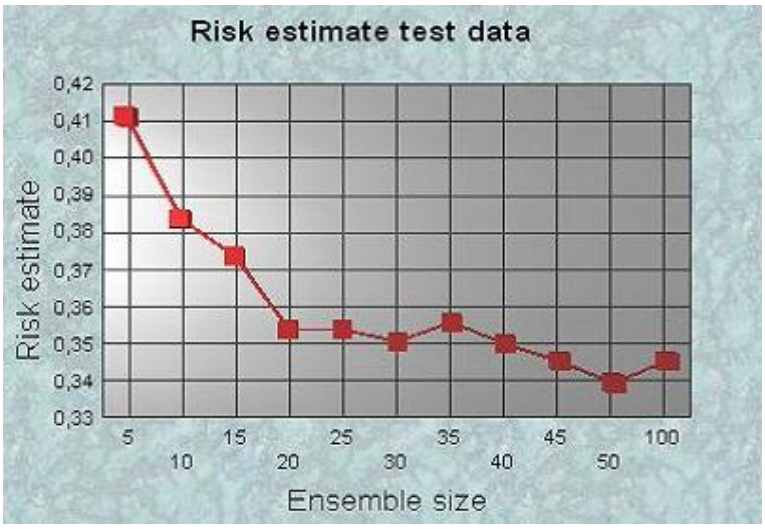

Figure 3: Relationship between ensemble size and risk estimate for test data

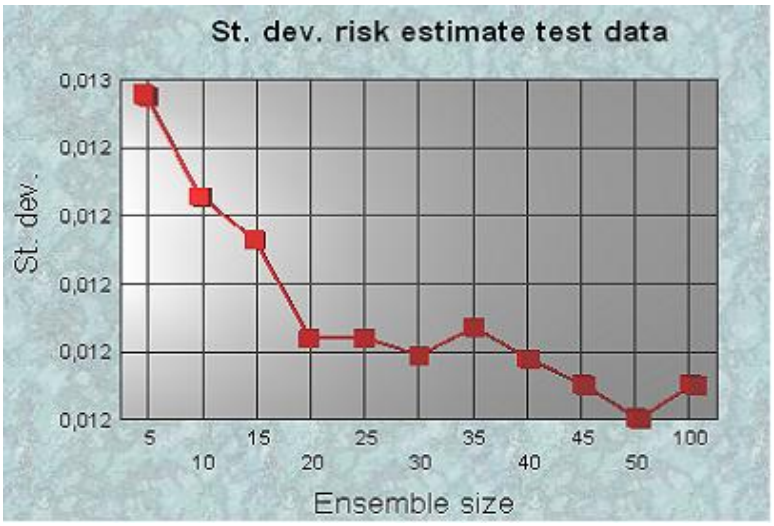

Figure 4: Relationship between ensemble size and standard deviation of risk estimate for test data

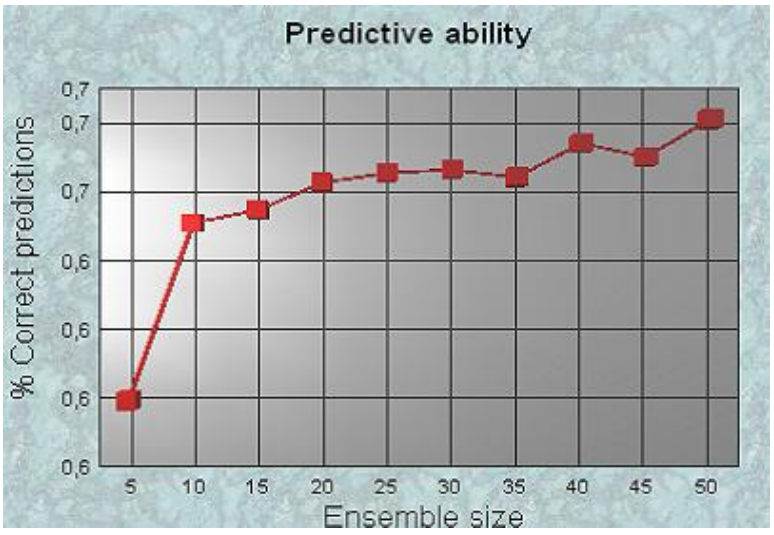

Figure 5: Relationship between ensemble size and predictive ability

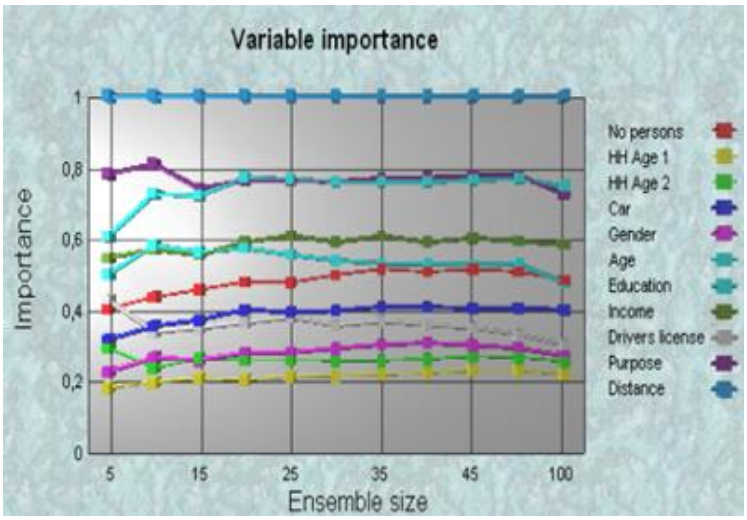

Figure 6: Relationship between ensemble size and variable importance

5. Conclusions and discussion

Analysis and modeling of travel demand in general and transport mode choice in particular have typically relied on a single equation or alternative single representations such as a decision tree. Behaviorally, this implies that researchers have invariably assumed that travelers apply the same utility function or set of decision heuristics when choosing between alternative transportation modes. In reality, however, different people may have different rules. If this hypothesis would be true, the success of predicting transport mode choice would increase if different sets of decision rules with variable condition variables would be used. We argued that this behavioral principle is congruent with the notion of using ensembles of models to predict transport mode choice.

In this paper, we report the findings of such a study, focusing on ensembles of decision trees. The predictive performance of the decision tree formalism is investigated as a function of ensemble size for a subset of the 2004 national travel survey data pertaining to the Rotterdam metropolitan area. Various measures of predictive success and uncertainty in predictions are used to examine the effects of increased ensembles size. In addition, the impact of increased ensembles size on the predicted importance of conditions is investigated.

A few relevant conclusions can be drawn. First, predictive success tends to increase with increasing ensemble size and the corresponding uncertainty in predicted transport modes tends to decrease with increasing ensemble size. Second, this process is not monotonic; in particular predictive success and corresponding uncertainty tend to fluctuate for predictions that involve less than 20 decision trees. More or less stable results are obtained for ensemble sizes higher than 20, although again the statistics are still not necessarily monotonic with such ensemble size. Third, the predictive accuracy of the ensemble of decision trees tends to increase with increasing ensemble size. Fourth, jointly with changes in predictive accuracy, the importance of predictor variables varies as a function of ensemble size, both in absolute and relative terms. Consequently, the importance of a particular variable relative to that of other predictor variable may change with an increasing number of decision trees used to predict transport mode choice.

Although this study suggests that including a richer behavioral spectrum in the predictions to reduce uncertainty in forecasting, this seems a rather technical stance. Behavioral heterogeneity is constructed by bootstrapping procedures and random selection of socio-demographic and contextual variables. In that sense, behaviorally, results depend on the bootstrapping and random selection approach. This approach is akin to the current discrete choice approach that is explicitly formulated against a theoretical framework. Hence, to improve the behavioral

foundations of ensembles, it would be interesting in future research to develop a more systematic approach in which the various decision tables are not derived by a theory-neutral algorithm but rather on an approach that systematically identifies maximum behaviorally distinct segments, using different decision heuristics.

The aim of the present study has been primarily on the relationship between ensemble size and uncertainty. However, the application of random forests involves another layer of many operational decisions that may affect the ultimate results. Examples include the initial categorization of input variables, some of which are small probabilities or represent missing answers. Overall, the impact of these categories will be small, but they may result in differences. Other examples include the selection of training and test samples, the nature of the cross-validation, etc. Future research should also address the sensitivity of the findings to these operational decisions that are associated with the application of random forests and machine learning methods in general.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Program ([FP7/2007-2013] under grant agreement n° 248488.

References

Abdel-Rahman, E., Mutanga, O., Adam, E. and Ismail, R. (2014). Detecting *sirex noctilio* grey-attacked and lightning-struck pine trees using airborne hyperspectral data, random forest and support vector machines classifiers. *Journal of Photogrammetry and Remote Sensing*, 88, 48-59.

Ai, F.F., Bin J., Zhang, Z-M., Huang, J-H., Wang, J-B., Liang, Y-Z., Yu, L. and Yang, Z-Y. (2014). Application of random forests to select premium quality vegetable oils by their fatty acid composition. *Food Chemistry*, 143, 472-478

Anggraini, R., Arentze, T.A. and Timmermans, H.J.P. (2012). Car allocation decisions in car-deficient households: The case of non-work tours. *Transportmetrica*, 8, 209-224.

Arentze, T.A. and Timmermans, H.J.P. (2000). *Albatross: A Learning-Based Transportation Oriented Simulation System*. EIRASS, Eindhoven University of Technology, Eindhoven, The Netherlands.

Arentze, T.A. and Timmermans, H.J.P. (2004). A learning-based transportation oriented simulation system. *Transportation Research Part B,* 38, 613-633.

Beser Hugosson, M. (2005). Quantifying uncertainties in a national forecasting model. *Transportation Research A*, 39, 531-547.

Bhat, C.R., Guo, J.Y., Srinivasan, S., and Sivakumar, A. (2004). A comprehensive micro-simulator for daily activity-travel patterns. *Proceedings of the Conference on Progress in Activity-Based Models*. EIRASS, Maastricht.

Bowman, J. L. and Ben-Akiva, M. (2000). Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research A, 35*, 1-28.

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

Castiglione, J., Freedman, J. and Bradley, M. (2003). Systematic investigation of variability due to random simulation error in an activity-based micro-simulation forecasting model. *Transportation Research Record*, 1831, 76-88.

Caulfield, B. and Brazil, W. (2011). Examining factors that affect mode choice for frequent short trips. *Proceedings of the 90th Annual Meeting of the Transportation Research Board.* Washington DC.

Cools, M.,] Kochan, B., Bellemans, T., Janssens, D. and Wets, G. (2011). Assessment of the effect of microsimulation error on key travel indices: Evidence from the activity-based model Feathers. Proceedings *of the 90th Annual Meeting of the Transportation Research Board,* Washington, DC.

Désir, C., S. Bernard, C. Petitjean, and L. Heutte (2013). One class random forests, *Pattern Recognition*, 46, 3490-3506.

Jong, G. de, Daly, A., Pieters, M., Miller, S., Plasmeijer, R. and Hofman, F. (2007). Uncertainty in traffic forecasts: Literature review and new results for The Netherlands. *Transportation*, 34, 375-395.

Ferdous, N., Pendyala, R.M., Bhat, C.R. and Konduri, K.C. (2011). Modeling the influence of family, social context, and spatial proximity on non-motorized transport mode use *Proceedings **of** the 90th Annual Meeting of the Transportation Research Board,* Washington, DC.

Henson, K.M., Goulias, K.G. and Golledge, R.G. (2009). An assessment of activity-based modeling and simulation for applications in operational studies, disaster preparedness, and homeland security. *Transportation Letters*, 1, 19 – 39.

Hess, S. and Stathopoulos. A. (2014). A mixed random utility - random regret model linking the choice of decision rule to latent character traits. *Journal of Choice Modelling*, In Press

Heung, B., Bulmer, C.E. and Schmidt, M.G. (2014). Predictive soil parent material mapping at a regional scale: A random forest approach. *Geoderma*, 214–215, 141-154.

Jiao, J., Moudon, A.V. and Drewnowski, A. (2011). Grocery shopping: how individuals and built environments influence travel mode choice, *Proceedings of the 90th Annual Meeting of the Transportation Research Board,* Washington DC.

Kwak, M., Arentze, T.A., de Romph, E. and Rasouli, S. (2012). Activity-based dynamic traffic modeling: Influence of population sampling fraction size on simulation error. *Proceedings International Association of Travel Behavior Research Conference*, Toronto, Canada.

Lee, S., Choi, H., Cha, K. and Chung, H. (2013). Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. *Microchemical Journal*, 110, 739-748.

Liu, F., Janssens, D., Wets, G. and Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Systems with Applications,* 40, 3299-3311.

Lu, Y., Zhu, S. and Zhang, L. (2013). Imputing trip purpose based on GPS travel survey data and machine learning methods. *Proceedings of the 92nd Annual Meeting of the Transportation Research Board*. Washington, DC.

Maley, D. and Weinberger, R. (2011). Food shopping in the urban environment: Parking supply destination choice and mode choice. *Proceedings of the 90th Annual Meeting of the Transportation Research Board*, Washington DC.

Pendyala, R.M., Kitamura, R., Kikuchi, A., Yamamoto, T. and Fujii, S. (2005). Florida activity mobility simulator, overview and preliminary validation results. *Transportation Research Record,* 1921, 123-130.

Puissant, A., Rougier, S. and Stumpf, A. (2014). Object-oriented mapping of urban tress using random forest classifiers. *International Journal of Applied Earth Observation and Geoinformation*, 26, 235-245.

Rasouli, S., Cools, M., Kochan, B., Arentze, T.A., Bellemans, T. and Timmermans, H.J.P. (2012). Uncertainty in forecasts of complex rule-based systems of travel demand: Comparative analysis of the Albatross/Feathers model system. *Proceedings International Association of Travel Behavior Research Conference,* Toronto, Canada.

Rasouli, S. and Timmermans, H.J.P. (2011). Uncertainty in travel demand forecasting models: Literature review and research agenda. *Transportation Letters*, 4, 55-73.

Rasouli, S. and Timmermans, H.J.P. (2012). Uncertainty, uncertainty, uncertainty: Revisiting the study of dynamic complex spatial systems. *Environment and Planning A*, 44, 1781-1784.

Rasouli, S. and Timmermans, H.J.P. (2013). Assessment of model uncertainty in destination and travel forecasts of models of complex spatial shopping behaviour. *Journal of Retailing and Consumer Services,* 20 (2), 139-146.

Rasouli, S. and Timmermans, H.J.P. (2013). Probabilistic forecasting of time-dependent OD matrices by a complex activity-based model system: Effects of model uncertainty. *International Journal of Urban Sciences*, 17, 350-361.

Rasouli, S. and Timmermans, H.J.P. (2013). Uncertainty in complex, rule-based systems of travel demand: Results for the Albatross model system. In: Hesse, M., Caruso, G., Gerber Ph. and Viti, F. (eds.), *Proceedings BIVEC/GIBET Transport Research Days,* Luxemburg, pp. 233-241.

Rasouli, S. and Timmermans, H.J.P. (2014). Activity-based models of travel demand: Promises, progress and prospects. *International Journal of Urban Sciences*, 18, 31-60.

Roorda, M.J. (2005). *Activity-Based Modelling of Household Travel*. PhD thesis, University of Toronto.

Roorda, M. J., Doherty, S.T. and Miller, E.J. (2005). Operationalising household activity scheduling models, addressing assumptions and the use of new sources of behavioral data. In Lee- Gosselin, M. and Doherty, S.T. (eds), *Integrated Land-Use and Transportation Models, Behavioural Foundations*, Elsevier, Oxford, pp. 61-85.

Roorda, M.J. and Miller, E.J. (2005). Strategies for resolving activity scheduling conflicts: An empirical analysis. In Timmermans, H.J.P. (ed.), *Progress in Activity-Based Analysis*, Elsevier, Oxford, pp. 203-222.

Roorda, M.J., Miller, E.J. and Habib, K.M.N. (2008). Validation of TASHA: A 24-h activity scheduling microsimulation model. *Transportation Research Part A*, 42, 360-375.

Susilo, Y.O., Hanks, N. and Ullah, M. (2011). Exploration of shoppers travel mode choice in visiting convenience stores in the United Kingdom. *Proceedings of the 90th Annual Meeting of the Transportation Research Board*, Washington DC.

Taher Azar, A., Ismail Elshazly, H., Ella Hassanien, A. and Mohamed Elkorany,A. (2014). A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, 113, 465-473.

Veldhuisen, K.J., Timmermans, H.J.P. and Kapoen, L.L. (2000). Microsimulation model of activity-travel patterns and traffic flows: Specification, validation tests and Monte Carlo error. *Transportation Research Record,* 1706, 126-135.

Vovsha, P., Bradley, M. and Bowman, J.L. (2005). *Activity-based travel forecasting models in the United States, Progress since 1995 ands prospects in the future.* In Timmermans, H.J.P. (ed.) *Progress in Activity-based Analysis*, Elsevier, Oxford, pp. 389-415.

Xie, C., Lu, J. and Parkany, E. (2003). Work travel mode choice modeling using data mining: Decision trees and neural networks. *Proceedings of the 83rd Annual Meeting of the Transportation Research Board,* Washington, DC.

Yao, C., Spurlock, D.M., Armentano, L.E. Page Jr., C.D. VandeHaar, M.J. Bickhart, D.M. and Weigel, K.A. (2013). Random forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. *Journal of Dairy Science*, 96, 6716-6727.

Ye, Y., Wu, Q., Zhexue Huang, J. Ng, M.K. and Li, X. (2013). Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition,* 46, 769-787.

Zhang, T., C. Xie and Waller, S.T. (2011). An integrated equilibrium travel demand model with nested logit structure: Problem formulation and uncertainty analysis. *Proceedings of the 90th Annual Meeting of the Transportation Research Board*, Washington DC.

Zhao, Y. and Kockelman, K.M. (2001). The propagation of uncertainty through travel demand models: An exploratory analysis. *Annals of Regional Science* 36, 909-921.

Ziems, S., Bhargava, S., Plotz, J. and Pendyala, R.M. (2011). Stochastic variability in microsimulation modeling and convergence of corridor-level characteristics. *Proceedings of the 90th Annual Meeting of the Transportation Research Board*, Washington DC.