Travel Behaviour and Society 14 (2019) 1–10

Contents lists available at ScienceDirect

Travel Behaviour and Society

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

Applying a random forest method approach to model travel mode choice behavior

Long Cheng[a,b], Xuewu Chen[a,*], Jonas De Vos[b], Xinjun Lai[c], Frank Witlox[b,d,e]

[a] Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing 210096, China
[b] Department of Geography, Ghent University, Krijgslaan 281 S8, Ghent 9000, Belgium
[c] School of Electro-Mechanical Engineering, Guangdong University of Technology, No. 100 Waihuan Xi Road, Guangzhou 510006, China
[d] Department of Geography, University of Tartu, Vanemuise 46, 51014 Tartu, Estonia
[e] College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, 29 Yudao Street, Nanjing 210016, China

ARTICLE INFO

Keywords:
Travel mode choice
Prediction performance
Variable importance
Random forest
Nanjing (China)

ABSTRACT

The analysis of travel mode choice is important in transportation planning and policy-making in order to understand and forecast travel demands. Research in the field of machine learning has been exploring the use of random forest as a framework within which many traffic and transport problems can be investigated. The random forest (RF) is a powerful method for constructing an ensemble of random decision trees. It de-correlates the decision trees in the ensemble via randomization that leads to an improvement of forecasting and reduces the variance when averaged over the trees. However, the usefulness of RF for travel mode choice behavior remains largely unexplored. This paper proposes a robust random forest method to analyze travel mode choices for examining the prediction capability and model interpretability. Using the travel diary data from Nanjing, China in 2013, enriched with variables on the built environment, the effects of different model parameters on the prediction performance are investigated. The comparison results show that the random forest method performs significantly better in travel mode choice prediction for higher accuracy and less computation cost. In addition, the proposed method estimates the relative importance of explanatory variables and how they relate to mode choices. This is fundamental for a better understanding and effective modeling of people's travel behavior.

1. Introduction

In order to develop a socially desirable and environmentally sustainable transport system in line with the traveler's demands, transportation planners must improve their understanding of the hierarchy of individual and contextual variables that drive people's travel mode choice. Understanding mode choice is important since it affects how efficiently people can travel and how much urban space is devoted to transportation functions, as well as the range of alternatives available to travelers (De Dios Ortúzar and Willumsen, 1999).

1.1. Determinants of travel mode choice

In fact, a large body of literature shows that travel mode choice is affected by a variety of factors including socio-demographics, built environment and attitudes (Cervero, 2002; Van Acker and Witlox, 2011; Ermagun et al., 2015; De Vos et al., 2016). Behavioral heterogeneity in travel mode choices is observed, varying by age, gender, income, driving license availability, education level, car ownership, and household structure. Li et al. (2012)—exploring travel mode choices in the UK—confirmed that the share of car use decreases at higher ages. With respect to gender, Cheng et al. (2017) found that women rely more on public transit than men. Bhat and Srinivasan (2005) additionally reported that travelers with higher income are more likely to travel by car. Similar findings have been found in the studies where Bhat and Lockwood (2004) observed that people with a high income as well as those who have a driving license drive more frequently. With regard to education level, Plaut (2005) and van den Berg et al. (2011) both revealed that highly educated people conduct more trips (in particular leisure trips) by public transit. Car ownership is an important determinant of car trips (Ding et al., 2017). Finally, people living in larger households are less likely to use non-motorized modes than those living in smaller households (Ryley, 2006).

In addition, a number of key attributes of the built environment have been identified to exert pronounced influences on mode choice, such as building density, land-use mixture, dedicated infrastructure for

* Corresponding author.
E-mail addresses: chenxuewu@seu.edu.cn (X. Chen), jonas.devos@ugent.be (J. De Vos), xinjun.lai@gdut.edu.cn (X. Lai), frank.witlox@ugent.be (F. Witlox).

https://doi.org/10.1016/j.tbs.2018.09.002
Received 23 May 2018; Received in revised form 8 August 2018; Accepted 13 September 2018

pedestrians and cyclists, distance to various facilities, and transportation provisions (Cervero, 2002; Schwanen and Mokhtarian, 2005; Ding et al., 2017). Generally, high density and mixed land use encourage people to walk or cycle, due to relatively short distances and walking/cycling infrastructure, or use the available public transit facilities. It should be noted that dedicated neighborhood design towards walking and cycling advances the use of active modes. Furthermore, the enhanced accessibility tends to have positive effects on walking (Cao et al., 2006). In order to improve accessibility, we need to decrease the travel distance to public facilities as well as increase transport network connectivity.

Recently, the influences exerted by attitudes toward less tangible attributes such as comfort, convenience and travel satisfaction have gained considerable attention (Scheiner and Holz-Rau, 2007; De Vos et al., 2016). Studies indicate that attitudes may be better predictors of mode choice than the traditionally used objective measures. With a sample of Swedish commuters, studies found that attitudes toward flexibility and comfort and a pro-environmental inclination influence the individual's choice of mode (Johansson et al., 2006). Heinen et al. (2011) analyzed the influence of commuters' attitudes toward the benefits of cycling (e.g., convenience, low cost, health benefits) on the mode choice decision for commutes to work. Findings showed that attitudinal factors provide an additional explanation for commuter's cycling choice.

1.2. Modeling approach of travel mode choice

Models have traditionally been estimating travel mode choice using statistical regression framework, e.g. linear regression model, Poisson regression model, multinomial logit model, nested logit model etc. (Cervero, 2002; Bhat and Srinivasan, 2005; Cheng et al., 2016). However, these models have their own model assumptions and require predefined underlying relationships between the dependent and explanatory variables. For example, the multinomial logit model assumes that the choice probabilities of each pair of alternatives are independent of the presence or characteristics of all other alternatives. Violations of these assumptions produce inconsistent parameter estimates and biased predictions. Another critical problem of statistical regression models is that the relative influences of explanatory variables on travel mode choices are not evaluated. Understanding the relative importance of explanatory variables could significantly help travel mode choice prediction and therefore contribute to the improvement of travel demand forecasting. Although the significance test or sensitivity analysis can be conducted for conventional statistical regression models, just one variable is evaluated at one time under the assumption that other variables remain unchanged. As a result, the important interactions among variables might be ignored (Ding et al., 2016).

In contrast to statistical models, methods from the field of machine learning are promising alternatives for modeling travel mode choices. Instead of making strict assumptions, machine learning methods learn to represent complex relationships in a data-driven manner. The usefulness of machine learning methods for predicting travel mode choices has been demonstrated in transportation research, including decision tree (Lindner et al., 2017), neural network (Golshani et al., 2018), and support vector machine (Zhang and Xie, 2008; Semanjski et al., 2017). The common practice of these machine learning methods is to identify the single best performing model and utilize its estimated parameters to predict outputs under different scenarios. However, it is arguable that the development and application of a single model is not necessarily the best approach considering the various sources of error/uncertainty in the analysis of travel mode choices. The input data might contain errors, the sample might be biased, the model itself might be stochastic, and the scenarios used for predicting might not be consistent with the actual evolution of transportation systems (Rasouli and Timmermans, 2012, 2014).

In order to deal with this problem, this study explores the potential of a so-called ensemble method to predict travel mode choices. In machine learning, ensemble methods use multiple learning algorithms, obtaining better predictive performance compared to any of the constituent learning algorithms alone (Ding et al., 2016). Of all ensemble methods, the random forest (RF) method-developed by Breiman (2001) – is popular and shows very good capability in solving prediction and classification problems (Zaklouta and Stanciulescu, 2012; Zhang and Haghani, 2015). Instead of fitting a single "best" tree model, the RF strategically combines multiple simple decision trees to optimize predictive performance. In terms of travel mode choices, the application of the random forest method as a multitude of decision trees means that we allow for differences in travel decision heuristics. Different decision trees in the ensemble may pick up different sources of uncertainty and variability in the data. Thus, from a purely technical viewpoint, the accuracies of model estimation and prediction would be expected to enhance. Drawing on insights and techniques from both statistical and machine learning methods, the random forest method can identify and interpret relevant variables and interactions. The interpretability of this method enables us to better understand model results, and is important to analyze the relationships between mode choice and its contributing factors.

The random forest method has witnessed a wide application to different research fields and achieved great success. In the Appendix section of this paper, we provide a broad summary of recent studies that use the RF method to solve transportation prediction and classification problems. They are generally classified into four categories: travel choice behavior, traffic incident prediction, traffic time/flow prediction, and pattern recognition. Elhenawy et al. (2014), Rasouli and Timmermans (2014), Ermagun et al. (2015) employed RF to predict traveler's behavior, such as driving behavior at the onset of a yellow indication at signalized interactions and travel mode choice. The method has shown to be able to deal with mixed types of data and be effective in predicting multi-category classification problems. Brown (2016) indicated that the predictive accuracy of rail accidents severity improved through the use of RF, and that influential variables could be identified so as to gain a better understanding of the contributors. Rebollo and Balakrishnan (2014), Zhang and Haghani (2015), and Semanjski (2015) applied this method to predict travel time. Their proposed methods can fit complex nonlinear relationships while requiring little data preprocessing. Hou et al. (2015) developed different models to forecast long-term and short-term traffic flows. The experiments suggested that RF has a considerable advantage over other machine learning methods. A number of studies using RF are found in pattern recognition research, including traffic sign recognition, driving posture recognition, vehicle type recognition, trip purpose recognition, travel mode recognition, and drowsy behavior detection (Zaklouta and Stanciulescu, 2012; Zhao et al., 2012; Zhang, 2013; Montini et al., 2014; Shafique and Hato, 2015; Jahangiri and Rakha, 2015; Yang et al., 2016; Kamkar and Safabakhsh, 2016; Wang et al., 2016; Gong et al., 2018).

1.3. Objectives of this research

The review of these studies reveals that RF is a promising data mining approach for its ability to consider different types of variables and determine variable importance with no prior specification of model structures. However, there are limited studies on the application of RF in travel mode choice analysis. To the best of our knowledge, only two studies—conducted by Rasouli and Timmermans (2014) and Ermagun et al. (2015)—used the random forest method to predict travel mode choice. However, they adopted the default RF specification which might not produce the best prediction results. In our analysis, we calibrate the model parameter for achieving better performance.

The contribution of the present study to the literature review is twofold. First, it adds to the existing literature by providing recent developments of random forest method on travel mode choice analysis,