ARTICLE IN PRESS

Travel Behaviour and Society xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Travel Behaviour and Society

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

Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model

Nima Golshani*, Ramin Shabanpour, Seyed Mehdi Mahmoudifard, Sybil Derrible, Abolfazl Mohammadian

Department of Civil and Materials Engineering, University of Illinois at Chicago, 842 W. Taylor Street, Chicago, IL 60607-7023, USA

ARTICLE INFO

Keywords:
Neural networks
Hazard duration model
Copula-based model
Mode choice
Departure time

ABSTRACT

The field of travel demand analysis has traditionally been dominated by statistical models. Conversely, Machine Learning (ML) techniques have been rapidly emerging in the past few years, and several studies have demonstrated instances where ML outperformed statistical models, notably in their forecasting potential. In this article, we compare the performance of discrete, continuous, and joint discrete-continuous statistical models with the performance of the neural networks (NN), recognized as a popular ML technique. Specifically, we model two critical trip-related decisions of travel mode and departure time. Overall, we find that in addition to having a much easier and faster implementation process, the NN model offers better predictions for both decision variables. Nonetheless, critiques of NN usually typecast it as a black box due to difficulty of assessing the role of explanatory variables in estimating the target variables. To tackle this issue, we further investigate the contribution of exploratory variables in two steps: (1) estimating the relative importance of each exploratory variable, and (2) conducting sensitivity analysis on the most important variables. The results indicate that beside superior prediction accuracy, the NN is capable of capturing nonlinearities in travel demand, which suggests that it can also be more accurate to capture asymmetrical and non-linear responses for policy analysis purposes.

1. Introduction

Adopting the right Travel Demand Management (TDM) policies has never been more important. As most municipalities aspire for more sustainable, livable, and equitable communities, curbing traffic congestion and air pollution has become a priority in many urban areas. Being able to effectively and accurately model and predict the impact of a policy is therefore essential. Travel demand analysis studies have traditionally modeled many travel components (e.g., travel mode, departure time, trip destination) using statistical methods—most often from the family of random utility maximization (RUM) models. These methods have a sound mathematical background and focus on behavioral interpretation of the estimated parameters, which are obtained by employing predetermined likelihood functions. They are therefore unable to capture higher degrees of nonlinearity in a dataset[1] (Detienne et al., 2003; Karlaftis and Vlahogianni, 2011). Recently, data mining and Machine Learning (ML) techniques have been establishing their place in transportation studies, thanks to their ability to recognize and model highly nonlinear and complicated patterns in datasets without any assumption of their functional forms. Their main shortcoming has lied in the complexity of interpreting their results (Karlaftis and Vlahogianni, 2011).

The literature on the application of statistical models for modeling mode choice behavior is extensive. These models include multinomial logit model (e.g., Koppelman, 1983; Rassam et al., 1971), nested logit model (e.g., Vovsha, 1997; Wen and Koppelman, 2001), and mixed logit model (e.g., Mcfadden and Train, 2000; Hensher and Greene, 2002). Moreover, since mode choice decision is closely intertwined with decision on trip departure time, a number of studies considered them as a joint decision to account for the interrelation or causal effects between them (Nurul Habib, 2012; Nurul Habib et al., 2009; Paleti et al., 2014; Amirgholy et al., 2017; Shabanpour et al., 2017b). Among available methods to jointly model trip attributes, the copula-based modeling approach is able to jointly model discrete and continuous variables without imposing distribution assumptions on the error terms (Bhat and Eluru, 2009; Eluru et al., 2010; Born et al., 2014; Hossein

* Corresponding author.
E-mail addresses: ngolsh2@uic.edu (N. Golshani), rshaba4@uic.edu (R. Shabanpour), smahmo24@uic.edu (S.M. Mahmoudifard), derrible@uic.edu (S. Derrible), kouros@uic.edu (A. Mohammadian).

[1] This problem can be addressed to some extent by using non-parametric approaches in statistical models. However, estimating non-parametric statistical models are much more complex than the parametric models.

http://dx.doi.org/10.1016/j.tbs.2017.09.003
Received 17 October 2016; Received in revised form 19 May 2017; Accepted 14 September 2017

Rashidi and Mohammadian (2016)).

Traditionally departure time has been modeled using discrete choice models (see, for example, Bhat, 1998; Lemp et al., 2010; Shabanpour et al., 2017a; Small, 1982), while some recent studies have argued that because of the continuous nature of time, it is more accurate to consider it as a continuous variable. Furthermore, modeling departure time as a continuous variable frees researchers from constraints of discrete choice models. It also allows them to discretize time during their study based on their research problem, which is not the case the other way around (Habib and Carrasco, 2011). Therefore, this paper considers time as a continuous variable in line with recent studies (see, for example, Gadda et al., 2009; Habib and Carrasco, 2011; Lee and Timmermans, 2007; Nurul Habib et al., 2009; Amirgholy and Gao, 2017) for departure time analysis.

In contrast, only a few studies focused on the application of ML for modeling mode choice behavior. These studies range from support vector machine (Moons et al., 2007) and decision trees (Moons et al., 2007; Shmueli et al., 1996; Xie et al., 2003) to Neural Networks (NN) family models (Sayed and Razavi, 2000; Hensher and Ton, 2000; Mohammadian and Miller, 2002; Cantarella and de Luca, 2005; Andrade et al., 2006). Among these techniques, NNs showed great predictive capabilities in dealing with massive amounts of multi-dimensional data (Karlaftis and Vlahogianni, 2011; Faghri and Hua, 1992; Dougherty, 1995); NNs are constructed based on learning process of the human brain (i.e., consist of a network of neurons that process and transmit information to other neurons).

Furthermore, main differences between NN and statistical models can be summarized in three major aspects. First, their philosophy and application are different, where NN assumes that the dataset is formed by an unknown mechanism and focuses on the efficient implementation of the estimated model (i.e., achieve the best prediction accuracy with the lowest possible development time). Whereas, statistical models assume that the dataset is formed by a stochastic process and focus on estimation and inference of the estimated model (Hand, 2000). The second major difference is their estimation process, in which NN approximates its functional form in the learning process in contrast with statistical models, where researchers usually assume a prior functional form for the models. This characteristic makes the NN models more flexible but may result in additional issues in interpretation of the results where they are usually labeled as a black-box process (Flexer, 1994; Warner and Misra, 1996). The last major difference lies in the ability of dealing with multicollinearity, outliers, noisy data, and missing values. While NN models can easily address these issues, accounting for them in the statistical models needs laborious steps (Gupta and Lam, 1996).

The aforementioned differences have motivated researchers in different fields to compare the modeling process and estimation accuracy of Machine Learning and statistical methods. However, there are only few studies in transportation literature that compared these models. For example, Moons et al. (2007) developed multiple models including support vector machine and regression tree for mode choice analysis and concluded that in overall, they outperform the statistical models when dealing with balanced distributed data. Hensher and Ton (2000) and Mohammadian and Miller (2002) compared NN with multinomial and nested logit models and reported that NN outperforms these statistical models.

To contribute to this ongoing debate, this article investigates and compares the performance of discrete, continuous, and joint discrete-continuous statistical models with the performance of the neural networks as a popular ML technique in the contexts of trip departure time and mode choice behavior. As for the statistical modeling approach, accelerated hazard model is employed to estimate trip departure time and multinomial logit is used to model travel mode choice. The copula-based model is also selected for the joint modeling structure because it is largely seen as one of the best methods to simultaneously model multiple decision variables by the transportation community. Furthermore, using a joint modeling scheme allows for a thorough investigation of the application of ML and statistical models for both discrete and continuous variables.

2. Data preparation

The data used in this study is extracted from the Travel Tracker Survey conducted by the Chicago Metropolitan Agency for Planning (CMAP). The survey was conducted from January 2007 to February 2008 in Chicago, IL, and a total of 10,552 households were asked to fill a complete travel diary for one or two randomly assigned days. The dataset contains detailed information of more than 210,000 trips including their departure time, duration, purpose, and mode, as well as household and individual level sociodemographic characteristics and activity-related variables. After cleaning the dataset and removing invalid records, 9,450 observations of home-based (trips that originated from home) shopping trips were selected for the purpose of this study. The selection of home-based trips for mode choice analysis is due to the strong dependency of mode choice decision of non-home based travels to the home-based trips in their tour.

Moreover, using Google Maps API, some additional variables representing information on travel time between trip origin and destination by various modes, mode availability, and access/egress distances are calculated and added to the dataset. Recently, this approach has attracted considerable interest in transportation-related studies (see, for example, Javanmardi et al., 2015; Shakeel et al., 2016). We also added new variables representing land use and built environment attributes such as road density and population density to the dataset. Table 1 presents the summary statistics of the key variables used (found to be significant) in this study.

3. Model specification

3.1. Statistical models

This study employs a multinomial logit (MNL) formulation to analyze the determinants of travelers' mode choice decision. The utility function for alternative a can be written as:

$$U_{ai} = \beta_a x_{ai} + \varepsilon_{ai} \tag{1}$$

where U_{ia} is the utility of mode a for observation i, x_{ai} are explanatory variables, β_a are estimated parameters, and ε_{ai} is the error term of the

Table 1
Description of variables and summary statistics.

Variable	Definition	Mean	St. dev.
Road_density	Road density of home TAZ (roads lengths/area of TAZ)	0.22	0.19
Walk_TT	Travel time for walk mode (in hours)	2.45	2.89
Bike_TT	Travel time for bike mode (in hours)	0.55	0.64
Auto_TT	Travel time for auto drive mode (in hours)	0.34	0.39
Transit_TT	Travel time for transit mode (in hours)	0.34	0.33
Auto_cost	Travel cost for auto drive mode ($)	1.15	1.37
Transit_cost	Travel cost for transit mode ($)	1.80	1.59
Walk_accessible	1: if walking distance to the destination is less than 0.25 mile, 0: otherwise	0.08	0.27
Transit_egress	Egress distance to destination for transit mode (km)	1.50	3.23
Transit_access	Access distance from origin for transit mode (km)	2.38	4.20
Weekend	1: if the trip is made in weekend, 0: otherwise	0.11	0.31
HH_bikes	Number of bikes in the household	1.37	1.66
HH_size	Household size	2.70	1.36
HH_vehicle	Number of vehicles in the household	1.87	1.03
Part_work	1: if traveler works part time, 0: otherwise	0.14	0.35
Age_20	1: if traveler's age is less than 20, 0: otherwise	0.07	0.26
Age_40_65	1: if traveler's age is between 40 and 65, 0: otherwise	0.51	0.50