Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 10 (2015) 840 – 849

Transportation Research Procedia

www.elsevier.com/locate/procedia

18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands

Predicting travel mode of individuals by machine learning

Hichem Omrani*

Luxembourg Institute of Socio-Economic Research (LISER), 3, avenue de la Fonte L-4364 Esch-sur-Alzette, Luxembourg

Abstract

Travel mode choice prediction of individuals is important in planning new transportation projects. In this paper, we present four machine learning methods namely artificial neural net-MLP, artificial neural net-RBF, multinomial logistic regression, and support vector machines, for predicting travel mode of individuals in city of Luxembourg. The presented methods use individuals' characteristics, transport mode specifications and data related to places of work and residence. The dataset analyzed comes from a national survey. It contains information on the daily mobility (e.g., from home to work) of individuals who either live or work in Luxembourg. We extracted individual characteristics to relate daily movements (journeys between home and work, in particular) to the characteristics of working individuals. We used the information about public transportation and some geographical location of the residential and work places. We compare the rates of successful prediction obtained by neural networks and several alternative approaches for predicting the travel mode choice using cross-validation. The results show that the artificial neural networks perform better compared to other alternatives. Our analysis can be used to support management decision-making and build predictions under uncertainty related to changes in people's behavior, economic context or environment and transportation infrastructure.

Peer-review under responsibility of Delft University of Technology
Keywords: Data mining; data split; machine learning; validation; neural networks.

1. Introduction

The globalization of the economy and the development of transport and telecommunication technologies has led to an increasing concentration of knowledge-intensive employment and global firms in metropolitan regions, such as Luxembourg City. Between 1985 and 2007, the labor force employed in Luxembourg has more than doubled, from 141700 (among 125600 residents in Luxembourg and 16100 cross-borders workers who commute daily across the borders) to 316500 workers (180250+136250, source: STATEC, Luxembourg). A consequence of this is a dramatic increase in road traffic and congestion in peak hours. In Omrani et al. (2010), we explored the spatial and temporal patterns of commuting to work in Luxembourg. The present study follows up on this by studying also the mode of transport, which is relevant to the sustainable development of the Luxembourg transport infrastructure, and to further understanding of travel behavior of individuals and households.

* Corresponding author. Tel.: +352-585855-313 ; fax: ++352-585855-700.
E-mail address: hichem.omrani@liser.lu

Peer-review under responsibility of Delft University of Technology
doi:10.1016/j.trpro.2015.09.037

We consider the following travel modes: private car, public transport (bus or train) or soft mode (walking or cycling). By modal split we mean the composition (percentages) of commuters who use each of these travel modes. We study the everyday modal split of the workers who reside in Luxembourg by several modeling techniques, and estimate it within categories of social, economic, demographic characteristics, location (places of work and residence) and related variables, such as cost and availability and abundance of public transport. Prediction of the travel mode is a pattern recognition problem (supervised learning), in which several variables (e.g., human characteristics and geographical patterns) explain the choices among the travel modes. We assess models and estimation procedures by the quality of their prediction. Recently, new artificial intelligence models have been applied to predict individual travel mode. They have been introduced as alternatives to complex behavior modeling and pattern recognition.

The standard way of assessing the quality of prediction is by splitting the sample S into a learning and a testing dataset, denoted by L and T ($S = L \cup T, L \cap T = \emptyset$). The model is fitted to L and its performance is evaluated by comparing the fit to the observed values on T. In cross-validation, the sample is split into K subsamples (folds), and a random subset of these subsamples forms L and the remainder forms T. Several splits of S to L and T are drawn at random, and prediction is evaluated on T. We denote this method by L.T(K,R), where K is the number of folds and R the number of replications. The standard approach is $L.T(2, 1)$; $L.T(1, 1)$ corresponds to learning and assessment on the entire dataset, without splitting it. Zhang and Xie (2008) demonstrated by this approach that support vector machine (SVM) outperforms multinomial logit (MNL) in terms of prediction and generalization. He claims that the multilayer feed-forward neural network model, an adaptation of artificial neural network (ANN), is superior to MNL and SVM for fitting (L) but inferior for testing (T). Of course, the concern arises that this conclusion is specific to the dataset or its context. We assess this approach further by applying it to predicting travel mode. The ANN model is known for its high quality of prediction (Yamamoto et al., 2002; Li et al., 2008; Xianyu et al., 2008; Wang and Elhag, 2007; Yang et al., 1993); we highlight its strengths for travel mode over the established models. We assess the performance of the method by cross-validation. Unlike several case studies, in which $L = T = S$ or $L \cap T = \emptyset$ with $R = 1$; we use $R = 100$ replications. We use many replications because the results of one are associated with uncertainty due to sampling (selection of L from S). The results show that ANN predicts individual travel mode better on average than the alternatives.

The paper is organized in five sections. In the next section, we review the recent methodological developments in modeling travel mode, with an emphasis on the relevance and the motivation of ANN. Section 3 describes how travel mode is predicted by ANN using several explanatory variables. Section 4 presents the application, its results, and compares them with several alternatives. The concluding Section 5 discusses a planned implementation of ANN in a user-friendly package with R programming language (Team et al., 2010).

2. Related work

The prediction of the transport mode used by individuals has attracted much attention in recent years. There are many ways of undertaking this, but methods based on artificial neural networks (ANN) are relatively new. Several data mining methods (e.g. SVM) have been relatively successfully applied to solve this kind of problem. But it is not clear from the literature which model is the most appropriate for prediction. In Xie et al. (2003), it has been shown that the ANN is more robust for travel mode prediction than decision tree (DT) and multinomial logit (MNL) models. Moreover, it was demonstrated in (Zhang and Xie, 2008) that SVM and ANN models outperform MNL model. The contribution of this paper is to apply the ANN model to predicting travel mode. We confirm the superior performance of ANN by extensive simulations (cross-validation). Our assessment is based on the average probability of correct assessment (APCA). These efficiency criteria are described in more detail in section 4. To justify the model performance, a lot of papers dealing with the prediction tasks, provide a quantitative comparison to several other methods (using cross-validation technique). In fact, some recent papers present an empirical study for travel mode analysis and they shown that either ANN or SVM model lead to better results than logit or nested logit model (Hensher and Ton, 2000).

This practice of comparison is not very forceful and even not extensively studied to provide a strong justification of model performance. Therefore, there is a need to additional methodological justification (i.e. empirical and not theoretical evidence) to prove the performance in the prediction. A deep study is needed to understand the raisons that one method overcomes another one. Thus the main question is why a given model provides better prediction quality