Transportation Research Part C 77 (2017) 16–32

Contents lists available at ScienceDirect

Transportation Research Part C

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

Comparing decision tree algorithms to estimate intercity trip distribution

Cira Souza Pitombo [a,*], Andreza Dornelas de Souza [b], Anabele Lindner [b]

[a] Department of Transportation Engineering, School of Engineering of São Carlos, University of São Paulo, Av. Trabalhador São-carlense, 400, Pq Arnold Schimidt, São Carlos, SP 13566-590, Brazil

[b] Department of Transportation Engineering, School of Engineering of São Carlos, University of São Paulo, Brazil

ARTICLE INFO

Article history:
Received 8 April 2016
Received in revised form 8 December 2016
Accepted 12 January 2017
Available online 25 January 2017

Keywords:
CHAID
CART
Trip distribution modeling
Gravity models
Disaggregated analysis

ABSTRACT

Traditional trip distribution models usually ignore the fact that destination choices are made individually in addition to aggregated factors, such as employment and average travel costs. This paper proposes a disaggregated analysis of destination choices for intercity trips, taking into account aggregated characteristics of the origin city, an impedance measurement and disaggregated variables related to the individual, by applying nonparametric Decision Tree (DT) algorithms. Furthermore, each algorithm's performance is compared with traditional gravity models estimated from a stepwise procedure (1) and a doubly constrained procedure (2). The analysis was based on a dataset from the 2012 Origin-Destination Survey carried out in Bahia, Brazil. The final selected variables to describe the destination choices were *population of the origin city*, *GDP of the origin city* and *travel distances* at an aggregated level, as well as the variables: *age, occupation, level of education, income (monthly), number of cars per household and gender* at a disaggregated one. The comparison of the DT models with gravity models demonstrated that the former models provided better accuracy when predicting the destination choices (trip length distribution, goodness-of-fit measures and qualitative perspective). The main conclusion is that Decision Tree algorithms can be applied to distribution modeling to improve traditional trip distribution approaches by assimilating the effect of disaggregated variables.

1. Introduction

The traditional four-step model aims to estimate travel demand and comprises the steps of Trip Generation, Trip Distribution, Mode Choice and Traffic Assignment. The analysis delineated in this paper is focused on the second stage of this traditional sequential model. Modeling trip distribution is crucial for travel demand forecasting and attempts to estimate the origins and destinations (ODs) of trips within a specific region and, thus to provide a matrix that characterizes the number of trips between each origin and destination. The trip distribution models are calibrated to predict destination choices, based on the trips produced and attracted by each Traffic Analysis Zone (TAZ) or aggregated sociodemographic variables of the TAZs (De Grange et al., 2010; Wilson, 1967).

The interactions between origins and destinations are complex and such models commonly contemplate aggregated variables. Trip distribution is considered an aggregate problem and the models related to it ignore the fact that destination

* Corresponding author.
E-mail addresses: cirapitombo@gmail.com (C.S. Pitombo), andrezadornelas@gmail.com (A.D. de Souza), bele.lindner@gmail.com (A. Lindner).

http://dx.doi.org/10.1016/j.trc.2017.01.009
0968-090X/

choices are made individually and may consider the individual and household features, as well as aggregated factors, such as population, employment and average travel cost.

In this field, the first proposed model was the Fratar Growth one (Evans, 1970; Williams, 1976). However, this model and some successive renowned methods are strongly dependent on the accuracy of the base-year matrix. Moreover, this set of methods is not able to predict the future matrix taking into account different scenarios, such as pricing policies and implementing new travel modes (Ortúzar and Willumsen, 2011).

An alternative to these approaches is the gravity model, which was once widely accepted and still remains as one of the most popular trip distribution models (Cascetta et al., 2007). However, the gravity model is limited to aggregated modeling given the difficulties of including individual travel behavior in the analysis (Mishra et al., 2013).

The disaggregated analysis was introduced to the trip distribution step from the development of the discrete choice techniques, e.g. those formulated from the random utility destination choice model proposed in the 1980s (Fotheringham, 1983; Ben-Akiva and Lerman, 1985).

The traditional methods for discrete choice issues, such as the Multinomial (MNL) and Nested Logistic (NL) regression, are mathematical structures, which require low computational complexity. On one hand, these models are easy to handle and are usually inserted into user-friendly computing platforms. On the other hand, they imply limitations related to an attribute recognized as IIA (Independence of Irrelevant Alternatives).

The IIA attribute involves the constraint whereby random error terms are independent (no correlation) and equally (the same variance) distributed (Koppelman and Wen, 2000). Considering this, the estimation of the travel behavior using traditional approaches may result in biased, and hence, uncertain results. Xie et al. (2003) highlight the effect of IIA in MNL and NL models as the lack of accommodation of unobserved variations and the flexibility reduction of pattern substitution among alternatives.

As well as this main disadvantage regarding the IIA attribute, traditional disaggregated travel behavior modeling is based on simple mathematical assumptions. This means that the entire data is modeled based on a single continuous function. The mentioned disadvantages are overcome by different techniques, e.g. Decision Tree approaches. Classification Trees (CT) are nonparametric techniques that identify patterns and significant explanatory variables, given a dataset. This set of algorithms can be useful to analyze travel demand as they result in convenient estimations through a non-linear function model, or a model with various functions. Furthermore, they do not have any important constraints such as a multicollinear dataset, variable distribution assumptions or IIA.

In this context, travel behavior modeling can be formally described as a task of pattern recognition in which multiple human behavioral attributes represented by explanatory variables can predict a choice among a set of alternatives (Xie et al., 2003).

Some studies considered pattern recognition applied to travel behavior modeling. Shmueli et al. (1996) explored the application of neural networks for a behavioral transportation planning problem while comparing the travel demand patterns of men and women in Israel. Strambi and Van de Bilt (1998) analyzed trip generation rates through a DT algorithm (CHAID). Xie et al. (2003) and Lindner et al. (2017) investigated the performance of decision tree analysis and neural networks for travel mode choice modeling.

Arentze and Timmermans (2007) joined the specific potentials of the rule-based and parametric modeling approaches in the so-called "Parametric Action Decision Tree" (PADT). They replaced the conventional action-assignment rule of the decision tree analysis by a logit model and concluded that the approach can be used to incorporate travel-cost sensitivity on a rule-based model of an activity-travel choice. Pitombo et al. (2011) and Pitombo et al. (2013) analyzed the relations among the socioeconomic, land use, activity participation and travel patterns applying a decision tree approach. The authors performed activity-based approaches using decision tree modeling.

As well as what was previously mentioned, there are some recent studies that considered the application of artificial neural networks (Rasouli and Nikraz, 2013; Mozolin et al., 2005), specifically from the perspective of classifiers' algorithms for the trip distribution modeling step. Recent literature presents the application of decision tree algorithms in activity-based approaches to travel analysis. However, the application of such techniques is still rare in terms of estimating intercity destination choices in a disaggregated dataset (Yang et al., 2014; LaMondia et al., 2009). Fig. 1 summarizes the motivation for applying DT algorithms in destination choice issues.

In order to assess the performance of DT approaches for trip distribution modeling, the main aim of this paper is to propose a disaggregated analysis of destination choices for intercity trips applying nonparametric decision tree algorithms. This article aims to compare the performance of DT algorithms (CHAID - Chi-squared Automatic Interaction Detection and CART - Classification and Regression Trees) when estimating intercity trip distribution. This research also provides estimates of the destination choice through traditional gravity models and compares them with the proposed DT approaches.

The method described in this paper follows four main steps (Fig. 2). Firstly, two DT algorithms were applied to an Origin-Destination dataset related to aggregated and disaggregated features. Subsequently, two gravity models were estimated taking into account the method of least squares in a regression analysis approach and a doubly constrained procedure. Finally, the results from the application of the disaggregated method (DT algorithms) and the gravity models were compared from the perspective of the trip length distribution, goodness-of-fit measures and qualitative perspective.

This article presents five sections, as well as this introductory one. The second section describes the modeling approaches. The third section presents the study area and dataset. The fourth section describes the main results and discussions. The fifth