Spatial Statistics 11 (2015) 45–64

Contents lists available at ScienceDirect

Spatial Statistics

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

A two-step method for mode choice estimation with socioeconomic and spatial information

Cira Souza Pitombo [a,*,1], Ana Rita Salgueiro [b,2], Aline Schindler Gomes da Costa [c,3], Cassiano Augusto Isler [a,1]

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

[b] *Department of Geology, Federal University of Ceará, Ceará, Brazil*

[c] *Master of Environmental and Urban Engineering, Federal University of Bahia, Bahia, Brazil*

ARTICLE INFO

Article history:
Received 25 September 2014
Accepted 17 December 2014
Available online 29 December 2014

Keywords:
Mode choice
Decision tree
Spatial estimation
Ordinary Kriging

ABSTRACT

Individuals choose the travel mode considering their own characteristics, those of the journey and the transport systems. Despite the current wide availability of georeferenced information and the forthcoming of Spatial Travel Demand Analysis as a research field, only a few studies have integrated the mode choice modeling methods and the geographical information. In this context, the goal of this paper is to apply a two-step method to estimate the mode choice based on the geographical position and socioeconomic attributes. From a database of household surveys in the city of São Carlos (Brazil) the first step of the method is to select the attributes which most influence the mode choice with a Decision Tree (*DT*). After comparing the performance of the *DT* with a Multinomial Logit Model, an Ordinary Kriging is applied to predict the mode choice under the spatial locations. The *DT* has shown to be effective in estimating the mode choice and selecting the variables to be kriged, which allowed predicting the mode choice of travelers based on geographical location. The proposed method may be

* Corresponding author.
E-mail addresses: cira@sc.usp.br (C.S. Pitombo), geo.ritasalgueiro@gmail.com (A.R. Salgueiro), ali_sgc@hotmail.com (A.S.G. da Costa), cassiano.isler@usp.br (C.A. Isler).

1 Address: Avenida Trabalhador Sãncarlense, 400, São Carlos - SP, 13566-590, Brazil.

2 Address: Campus do Pici - Bloco 912, 60440-900, Fortaleza - CE, Brazil.

3 Address: Escola Politécnica - Rua Professor Aristides Novis, 2 - Federação, Salvador - BA, 40210-630, Brazil.

http://dx.doi.org/10.1016/j.spasta.2014.12.002

an alternative to the traditional approaches in both non-spatial and spatial modeling, by enabling mode choice estimations given geographic coordinates of households.

1. Introduction

Individuals choose their travel mode considering multiple factors which can be classified by (1) the characteristics of the trip maker (car ownership, income, residential density etc.), (2) the characteristics of the journey (trip purpose, time of the day etc.), and (3) the characteristics of the transport facility (travel time, monetary costs, comfort, convenience etc.). Thus, travel behavior involves household and personal characteristics, travel variables, and spatially correlated factors as the location of households and individuals' destinations, and the activities distribution in the urban environment (Ortúzar and Willumsen, 2011).

Several studies corroborate that travel behavior – especially in the case of mode choice – is strongly related to the spatial distribution of cities (defined by their urban density, e.g., compact *versus* spread cities), the distribution of their economic activities and the presence of Traffic Analysis Zones (TAZ) with mixed activities (Cervero and Radisch, 1996; Kitamura et al., 1997).

Due to the advances in technology, the wide availability of georeferenced information turned Spatial Travel Demand Analysis into a forthcoming research field (Páez et al., 2013) by enabling the recognition of travel patterns taking into account spatial variables and, thus, incorporating the spatial effects in mode choice modeling.

In different applications of this approach, some researchers have found that travel behavior and mode choice are correlated to the spatial features of a trip. Bhat and Zhao (2002) have highlighted the spatial aspects that need to be recognized when modeling travel demand and have proposed a Multi-Level Mixed Logit formulation to address these spatial issues in the context of activity based analysis in the Boston Metropolitan area.

Furthermore, Adjemian et al. (2010) have shown that vehicle type ownership is spatially dependent at both the regional and household-level. More recently, Páez et al. (2013) have introduced a new indicator of spatial fit that can be applied to discrete choice models to estimate the door-to-door travel choices.

Among the Spatial Statistics techniques, the Geostatistics enables one to consider the spatial autocorrelation when modeling a problem and to predict the value of a variable in locations where it is unknown or unobserved. The goal of this paper is to apply a two-step method to estimate the mode choice of travelers whose geographical coordinates are known (sampled households) and also to predict the choices of non-sampled households.

The method is a sequential application of a Decision Tree (*DT*) Analysis and Ordinary Kriging (*OK*). The *DT* approach is applied to estimate the probabilities of the traveler's mode choice located in previously known positions and also to generate the variables to be kriged in the next step (the probabilities of mode choice in sampled households).

Based on the variables obtained in the first stage, the *OK* is applied to extend the mode choice estimations to eventual travelers in locations outside the range of the geographical coordinates of the sampled households. As this method can be applied only to numerical variables, the probabilities estimated from *DT* are effective for this purpose.

The proposed method is an alternative to the traditional approaches of mode choice modeling as it covers the spatial attributes of trips and consists of two main steps: non-spatial and spatial modeling. Here, the Decision Tree Analysis is considered an alternative to the traditional random utility models that may be applied in cases of lack of travel information or answers from a Stated Preference Survey. However, in order to prove the effectiveness of the *DT* approach, a Multinomial Logit Analysis is applied to the dataset described in this paper and its results are compared to the ones obtained with the former method.

In spatial modeling, the Ordinary Kriging proves to be a technique with a major advantage over other spatial confirmatory techniques – for example, the Geographically Weighted Regression – since it enables estimating the probabilities of mode choice of non-surveyed locations.

The remainder of this paper is organized as follows. In the next two sections we summarize a few applications of data mining and mode choice modeling, and studies of Geostatistics applied to transport planning. Section 4 describes the rationale of the proposed method in which a dataset at georeferenced disaggregated level is applied to the *DT* and the *OK* methods. Finally, in Section 5 we condense the overall conclusions drawn from the proposed method.

2. Mode choice modeling

For decades researchers have been investigating the factors that influence the travel mode choice based on different techniques – such as Logit and Probit models based on the Random Utility Maximization Approach – and datasets collected from Stated Preference (SP) and Revealed Preference (RP) Surveys (Sen et al., 1978; Ahern and Tapley, 2008; Grange et al., 2013).

The most common theoretical framework for discrete mode choice based on the Random Utility Maximization considers that (Domencich and McFadden, 1975; Williams and Senior, 1977; Ortúzar and Willumsen, 2011): the individuals whose mode choice are being modeled belong to an homogeneous population Q, act rationally and search for the decision which maximizes their personal utility given a set of constraints; there is a set of alternatives $A = \{A_1, \ldots, A_j, \ldots, A_N\}$ available to be chosen and a set of X attributes related to them such that an individual q is distinguished by the attributes $x \in X$ which define the choice $A(q) \in A$; and, each alternative $A_j \in A$ has an utility U_{ijq} related to individual q.

Since the planner does not have all the required information to define the individual choice, the utility of the jth choice for the qth individual is given by

$$U_{jq} = V_{jq} + \varepsilon_{jq} \tag{1}$$

where V_{jq} is a measurable portion and ε_{jq} is the random error related to the particular features of each individual.

The deterministic portion of Eq. (1) is generally given by a linear function of the attributes of the mode and the individual related to the jth choice as shown in Eq. (2). On the other hand, its random term can be expressed by a random variable with average zero and a given specific probability distribution.

$$V_{jq} = \sum_k \theta_{kj} \cdot x_{jkq} \tag{2}$$

where θ_{kj} is the coefficient (numeric value) related to the kth attribute of the jth choice and x_{jkq} is the level of the kth attribute for the jth choice of an individual q.

Thus, given that the individuals of the homogeneous population choose the alternative of maximum utility, the choice A_j is chosen if

$$U_{jq} \geq U_{jq} \quad \forall A_i \in A(q) \tag{3}$$

or

$$V_{jq} - V_{iq} \geq \varepsilon_{iq} - \varepsilon_{jq} \quad \forall A_i \in A(q). \tag{4}$$

In probabilistic terms, the choice of the jth given alternative can be calculated by:

$$P_{jq} = Prob\left(V_{jq} - V_{iq} \geq \varepsilon_{iq} - \varepsilon_{jq}\right) = Prob\left(\varepsilon_{iq} \leqslant \varepsilon_{jq} + V_{jq} - V_{iq}\right) \quad \forall A_i \in A(q). \tag{5}$$

The Multinomial Logit (Domencich and McFadden, 1975) assumes that the random portion of $g(x)$ is distributed by a IID Gumbel probability function given by the logarithm of the Weibull distribution function as shown in Eq. (6).

$$g\,(x) = \frac{1}{\sigma} \cdot exp\left[\frac{x-\mu}{\sigma} - exp\left(\frac{x-\mu}{\sigma}\right)\right] \quad \forall -\infty \leqslant x \leqslant \infty \tag{6}$$

where μ and σ are the average and standard deviation of the sampled population, respectively.

Thus, from Eq. (6) and after applying the Maximum Likelihood approach to estimate the parameters θ_{kj} of Eq. (2), the probability of a traveler q choosing the mode i for a trip is given by:

$$P_{iq} = \frac{exp\left(V_{iq}\right)}{\sum_{A_i \in A(q)} exp\left(V_{jq}\right)}. \tag{7}$$

Although most of the traditional mode choice models are based on this principle of random utility maximization, the data mining techniques are also effectively applied to predict travel behavior and mode choice. In this sense, mode choice modeling can be formally described as a task of pattern recognition in which multiple human behavioral attributes represented by explanatory variables allow the prediction of a choice among a set of alternatives (Xie et al., 2007).

Some studies have considered these pattern recognitions applied to transport planning and travel modeling. Shmueli et al. (1996) have explored the application of neural networks to a behavioral transportation planning problem while comparing the travel demand patterns of men and women in Israel. Xie et al. (2007) investigate the performance of Decision Tree (*DT*) Analysis and Neural Networks (NN) – two emerging pattern recognition data mining methods – for mode choice modeling of trips for business purpose.

Arentze and Timmermans (2007) combined the specific potentials of the rule-based and parametric modeling approaches in the so called *Parametric Action Decision Tree* (*PADT*). They have replaced the conventional action-assignment rule of the Decision Tree Analysis by a logit model and have concluded that the approach can be used to incorporate travel-costs sensitivity on a rule-based model of activity–travel choice.

Finally, Pitombo et al. (2011) have analyzed the relations among the socioeconomic, land use, activity participation and travel patterns applying a Decision Tree approach.

3. Geostatistics and Transport Planning

Although Geostatistics is commonly applied to natural or social sciences (Goovaerts, 1997), Transportation Planning might be considered as part of this research field – specifically studies related to travel behavior – since it considers the human and social aspects of individuals and groups, and personal attributes as socioeconomic features and behavioral variables.

However, only a few geostatistical methods have been applied to transportation datasets in the recent years, and most of the related publications refer to traffic engineering (Ciuffo et al., 2011; Mazzella et al., 2012; Zhang and Wang, 2013; Tong et al., 2013). Miura (2010) has applied kriging techniques to predict travel times by car between two arbitrary points and has got results with reasonable accuracy.

Zou et al. (2012) have proposed an improved distance metric called *Approximate Road Network Distance* (*ARND*) to solve the problem of invalid spatial covariance function in kriging methods given by the non-Euclidean distance metrics. Besides, Wang et al. (2012) have proposed to estimate the floating car speed with geostatistics.

Despite these examples of geostatistics and transportation planning combinations, the application of these techniques to demand or travel behavior modeling still remains incipient, with only a few recent papers (Ji and Gao, 2010; Peer et al., 2013; Pitombo et al., 2010). In these conditions, the focus of this paper is to propose an alternative method to evaluate the patterns of travel behavior considering not only the traditional techniques but also the spatial attributes of trips.

Generally, geostatistics considers datasets with intrinsic spatial structure. If, for example, measurements are taken at two different points, the difference between the measured values decreases as the two points are closer to each other (Matheron, 1971). The variables in this approach are called regionalized as they are distributed in the space, and modeled as random functions with a given spatial structure or, in other words, with a spatial correlation (Goovaerts, 1997).

The empirical variogram is the quantitative representation of the variation of a regionalized variable in space, where the variogram function is defined as half of the average square difference between

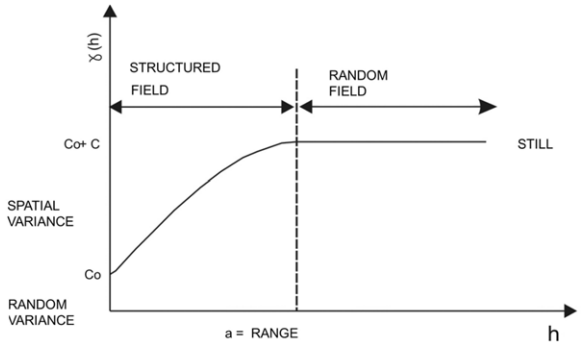

Fig. 1. Elements of a theoretical variogram.

points separated by a distance h (Matheron, 1963) as given by Eq. (8).

$$\gamma(h) = \frac{1}{2 \cdot N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2 \tag{8}$$

where $N(h)$ is the set of all pairwise; $z(x_i)$ and $z(x_i + h)$ are data values at spatial locations i and $i + h$, respectively.

After obtaining the experimental variograms, a mathematical function is fitted to best represent the variability in study. Among the several theoretical models for adjustments of a variogram the most frequently used ones are the Spherical, the Gaussian and the Exponential. In this step, the experimental variogram is replaced by a theoretical variogram function from which is possible to obtain the main parameters for spatial modeling: the nugget effect (C_o), the Range (a) and the Sill ($C + C_o$) as shown in Fig. 1.

From the known variogram of a region, a kriging method is applied to estimate the value of the regionalized variable given a specific position and the information regarding its neighborhood (Wackernagel, 2010). Thus, given the variogram of a regionalized variable its information might be applied to a kriging algorithm.

Ordinary Kriging (OK) is the most widely used kriging method that is related to the acronym B.L.U.E., which stands for "Best Linear Unbiased Estimator". OK is *linear* because its estimations are linear combinations of the available data, it is *unbiased* since it seeks a mean residual equal to zero and it is *best* because it aims to minimize the variance of errors between the observed and estimated data (Isaaks and Srvastava, 1989).

For the prediction of the variable Z at a location $x_0[Z(x_0)]$, the estimator $\hat{Z}(x_0)$ is defined by Goovaerts (1999) as in Eq. (9).

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i \cdot Z(x_i) \tag{9}$$

where the λ_i are the weights from the simultaneous solution of the following equations.

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \cdot \gamma(x_i, x_j) + \mu = \gamma(x_i, x), \quad i = 1, \ldots, n \\ \sum_{j=1}^{n} \lambda_j = 1 \end{cases} \tag{10}$$

where $\gamma(h)$ is the theoretical model for the variogram of the variable Z (fitted to the sample variograms) and μ is a Lagrange multiplier.

Cross Validation is a procedure to compare the various assumptions about either the model (e.g., type of function that best represents the adjusted variogram and its parameters) or the considered

database. In cross validation, each sample value $Z(x_1)$ is removed from the data set and a value $Z^*(x_1)$ located in x_1 is estimated using the remaining $(n - 1)$ samples. The difference between a data value and the estimated value $[Z(x_1) - Z^*(x_1)]$ provides an indicator of how well the data fits into the neighborhood of the surrounding data values (Wackernagel, 2010).

4. Method

The flowchart of Fig. 2 shows the steps of the method proposed in this paper. Firstly, a treatment of the database is required, followed by the application of a Decision Tree Analysis in order to identify the variables to be considered in the following steps, and also to enable to estimate the mode choice of the sampled household and to predict the eventual choices of the non-sampled population.

Although the *DT* is a non-parametric technique, it is an alternative approach to the traditional econometric models with the advantages of being unconstrained regarding the type of input variables (categorical, numeric or dummy) and that the assumptions of normality, linearity and multicollinearity do not require to be satisfied. In order to support this statement, an application of a Logit Multinomial approach to the same dataset of the Decision Tree Analysis is described in the following sections of this paper and the results of both methods are compared.

Finally, the patterns provided by the Decision Tree are analyzed by selecting the variables to be kriged and estimating the probabilities of mode choice, and their results are applied to an Ordinary Kriging method in order to enable predicting the mode choice of travelers on a spatial basis given their geographical position.

4.1. Data base

The study area in which the proposed method of this paper has been applied is the city of São Carlos (São Paulo/Brazil), with 221,936 inhabitants, 96% of which living in an urban area of approximately 105 km^2 (IBGE, 2010).

The sampled data on this population has been provided by the *Household Interview* applied in the occasion of the *Origin–Destination Survey* of 2007/2008 (Rodrigues da Silva, 2008). Moreover, a set of 5% of these interviewed households has been selected randomly from the water supply database from which their respective geographical coordinates have been obtained.

The *Urban Transportation Evaluation Survey* – a qualitative database of the transportation system of São Carlos – has been applied simultaneously with the previous survey to one member of the households and contains 2791 records. The preliminary analysis of the described databases has led to the elimination of records when one of the following situations had been observed: (1) inconsistent or missing data, (2) people who did not travel and (3) households with repeated geographic coordinates.

From this analysis, the final sample has resulted in 1216 records (individuals) defined by the categorical and numeric attributes (variables) shown in Table 1. Such information has been linked to the geographic coordinates (latitude and longitude in meters) of the interviewed members of the sampled households.

4.2. Variable selection and mode choice modeling

4.2.1. Decision tree analysis

Decision Tree (*DT*) Analysis is a set of procedures to evaluate and represent the existing relations among a dependent variable and a set of observed values of independent variables. It consists of a sequential binary partitioning of the dataset considering the values of these variables. Classification and Regression Tree (*CART*), a specific *DT* procedure, are fitted by successively splitting the data to form homogeneous subsets, resulting in a hierarchical tree of decision rules useful for prediction or classification (Breiman et al., 1984).

The *CART* algorithm – a specific *DT* procedure (Classification and Regression Tree) – is a *DT* segmentation modeling technique that has the following properties:

- The hierarchy resulted from its application is called tree and each segment is known as node;
- The root node contains the complete database;

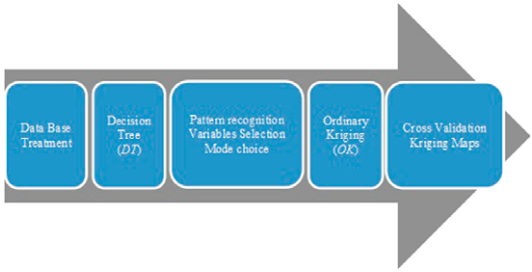

Fig. 2. Flowchart of the proposed method for the spatial mode choice analysis.

Table 1
Main variables considered in the DT application.

Variables	Description
Main problem—non-motorized mode	(1) Risk of running over; (2) Robberies; (3) Poor condition of sidewalks; (4) Lack of trees
Transit capacity	(1) Empty; (2) Suitable; (3) Crowded; (4) Overcrowded
Transit fleet	(1) Very small; (2) Small; (3) Suitable; (4) Upper
Main problem—transit mode	(1) Travel time; (2) Safety; (3) Comfort; (4) Itinerary; (5) Schedules
Main problem—car mode	(1) Traffic jam; (2) Lack of parking; (3) High cost
Driver's license	(1) Yes; (2) No
Gender of household head	(1) Male; (2) Female
Literacy	(1) Complete high school degree or college; (2) Incomplete high school degree or less; (3) Illiterate
Income	(1) 0 to 2 MW*; (2) 2, 1 to 8 MW*; (3) 8, 1 to 20 MW*; (4) Not answered
Main travel mode	(1) Public; (2) Private; (3) Non-motorized
Motorcycle ownership	Number of motorcycles
Car ownership	Number of cars
Age	Household's head age
Amount of trips	Number of trips performed by the household head

* MW = Minimum Wage.

- The root node is divided sequentially, generating child nodes;
- When no further data subdivision is possible, the final subgroups are called terminal nodes or leaves;
- Three main elements should be defined to run the *CART*: a set of questions delimiting data division; a criterion to establish the best division to produce the child nodes; and a termination rule to the subdivisions (*stop-splitting rule*).

The *CART* algorithm has been applied in this paper to investigate the variables that affect the individual mode choice and to predict the probable travel mode chosen. Furthermore, this *DT* analysis has been used to select the relevant numeric variables to estimate the mode choice of travelers considering their geographical position in the Ordinary Kriging step of the method proposed in Fig. 2.

A tree has been generated from the application of the *CART* algorithm to 70% of the 1216 records (851 records) of the database previously described as a calibration step, and 30% of the records have been used to test *DT* model.

The method has been parameterized with a minimum of 25 observations per terminal node (the stop-splitting rule considering the sample size and the desired homogeneity of the groups), the dependent variable has been set as the "main travel mode", which can assume three categories ((1) transit, (2) car/motorcycle and (3) non-motorized). The independent variables have been considered as the socioeconomic features of the sampled household, their travel characteristics and the qualitative measures of the transport system as exhibited in Table 1.

In order to evaluate the performance of the method under the established parameters after this calibration, the 30% remaining records of the dataset have been segregated with the *CART* algorithm as a validation step. Fig. 3 represents the training tree with 70% of the records.

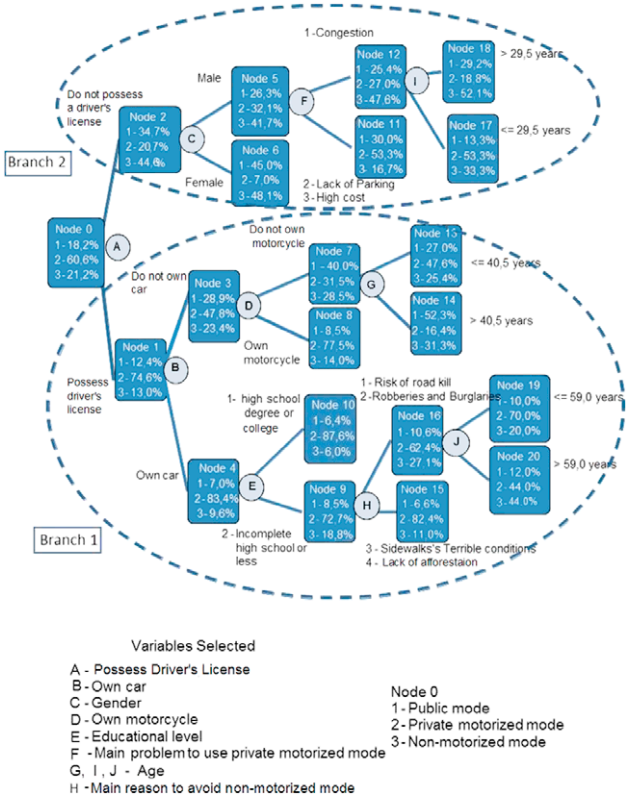

Fig. 3. Results of the Decision Tree application to the dataset. .

Starting with the entire database in the root node, the *CART* algorithm has divided the dataset successively into diverse groups considering the values (levels) of the independent variables (attributes) resulting in 11 terminal nodes by the end of data segregation.

In Fig. 3 we can see that the majority of travelers use private motorized mode (60.6%), followed by the non-motorized mode (21.2%) and then the transit mode (18.2%), and the most important variable (which best explains the data variability considering mode choice) is "Driver's License Possession", which may be split into the following two groups (the two leaves, or nodes, of the second level of the tree):

- Individuals that have driver's license (*Node 1* with 74.6% choosing car/motorcycle, 13.0% choosing non-motorized mode and 12.4% choosing transit);
- Individuals that do not have driver's license (*Node 2* with 44.6% choosing the non-motorized mode, 34.7% choosing transit and 20.7% choosing the car/motorcycle).

A few conclusions can be drawn from this segregation. Firstly, the majority of travelers choose the car/motorcycle mode in cases of driver's license possession and, secondly, the non-motorized mode is preferable when they do not possess a driver license, with the transit mode remaining as the third best choice.

Secondly, the selection of "Driver's License Possession" as the most important variable for the travel mode decision may be justified by the strong correlation of this attribute of travelers with their income and car ownership, which are known to have a high influence on the mode choice.

Further analysis of Fig. 3 has enabled conclusions about the level of each attribute related to the chosen mode of travel considering the 11 terminal nodes of the validation step, as shown in Table 2. For example, a traveler who chooses a car/motorcycle mode has a driver's license, has a car or motorcycle

Table 2
Variables selected by the DT and their relations to the mode choice.

Variables selected by *DT*	Car/motorcycle	Transit	Non motorized
Driver's license	Yes	No	No
Car ownership	At least one car	Zero car	Zero car
Motorcycle ownership	At least one moto	Zero moto	Zero moto
Age	$\Leftarrow 59$	>40	>59
Gender	Male	Female	Female
Literacy	Complete high school degree or college	High school/illiterate	High school/illiterate
Main problem—car mode	Lack of parking/high cost	Lack of parking/high cost	Traffic jam
Main problem—non motorized mode	Risk of running over /robberies	Poor condition of sidewalks/lack of trees	Poor condition of sidewalks/lack of trees

at home, is younger than 59 years old, is male and is worried about the lack of parking and the high cost of car trips.

Finally, the accuracy of the *CART* is calculated comparing the predicted mode choices (the one with highest value of probability among the three alternatives) and the modes actually chosen to perform the trips, both in the calibration and validation steps of the algorithm. These values demonstrate a good accuracy of the *DT* approach with hits of 78% of right estimation for private motorized mode, 83% for transit and 80% for non-motorized mode.

Prior to the geostatistical modeling, a chi-square test has been performed to evaluate the significance of association between the mode choice observed and those estimated by the segregation modeling. In this case, the hypotheses of chi-square test are:

- Null hypothesis: the observed and estimated are not associated.
- Alternative hypothesis: the observed and estimated are associated.

Besides the high accuracy of the *DT* to estimate the mode choice as mentioned before, a high value of chi-square statistic ($\chi^2 = 396$) has been obtained considering a usual significance level of 5% and 4 degrees of freedom, which means that the null hypothesis of non-association between the values estimated by the *DT* and the observed values may be rejected. Therefore, the observed mode choices and the estimations may be considered associated at a statistical significance level.

4.2.2. Multinomial logit

In order to evaluate the results of the *DT* procedure described in the previous section, there has been a first attempt to model the travel behavior with the attributes of Table 1 applying a MNL approach under the following set of utilities for each mode.

$$\begin{aligned} UCAR \;=\; & \beta_{MOTO} \cdot \mathrm{MOTO} + \beta_{CAR} \cdot \mathrm{CAR} + \beta_{AGE} \cdot \mathrm{AGE} + \beta_{LIC} \cdot \mathrm{LIC} + \beta_{GENDER} \cdot \mathrm{GENDER} \\ & + \beta_{STUDY} \cdot \mathrm{STUDY} + \beta_{INSTRU} \cdot \mathrm{INSTRU} + \beta_{MORCONDICA} \cdot \mathrm{MORCONDICA} \\ & + \beta_{MORCONDI\_1} \cdot \mathrm{MORCONDI\_1} + \beta_{MAX\_VIANUM} \cdot MAX\_VIANUM \end{aligned} \tag{11}$$

$$U_{BUS} = ASC_{BUS} \tag{12}$$

$$U_{WALK} = ASC_{WALK} \tag{13}$$

where ASC_{CAR} is the Alternative Specific Constant for CAR mode, ASC_{BUS} is the Alternative Specific Constant for BUS mode, ASC_{WALK} is the Alternative Specific Constant for WALK mode and the remaining coefficients are related to the respective attributes of Table 1.

In this set of equations, the constant term is equal to zero for the mode CAR and estimated for the remaining available modes, which enables comparing the perceived utility of each mode when the level of the other attributes are set to zero. On the other hand, the coefficients related to the socioeconomic variables have been estimated only for this first travel mode since the attributes related only to the personal features of the travelers. Thus, it makes sense to consider these variables only in one utility function since they remain at the same level across all the modes (i.e., they are not dependent on the mode, but only on the traveler's characteristics).

Table 3
Estimated coefficients and significances of the first MNL model.

Coefficient	Value	Standard error	t-test	p-value	
ASC_{BUS}	−3.54	1.28	−2.77	0.01	
ASC_{CAR}	0.00	Fixed	–	–	
ASC_{WALK}	−3.46	1.28	−2.71	0.01	
β_{CAR}	0.674	0.304	2.22	0.03	
β_{MOTO}	0.560	0.268	2.09	0.04	
β_{MAX_VIANUM}	0.411	0.137	3.00	0.00	
β_{LIC}	−1.43	0.347	−4.11	0.00	
$\beta_{MORCONDICA}$	0.150	0.151	0.99	0.32	*
$\beta_{MORCONDI\_1}$	−0.0218	0.0469	−0.46	0.64	*
β_{STUDY}	0.0195	0.100	0.20	0.85	*
β_{INSTRU}	−0.980	0.233	−4.21	0.00	
β_{AGE}	−0.00154	0.00881	−0.17	0.86	*
β_{GENDER}	−0.888	0.302	−2.94	0.00	

Table 4
Estimated coefficients and significances of the second MNL model.

Coefficient	Value	Standard error	t-test	p-value
ASC_{BUS}	−4.23	0.765	−5.53	0.00
ASC_{CAR}	0.00	Fixed	–	–
ASC_{WALK}	−4.15	0.765	−5.42	0.00
β_{CAR}	0.648	0.302	2.15	0.03
β_{MOTO}	0.577	0.267	2.16	0.03
$\beta_{MAXVIANUM}$	0.401	0.135	2.96	0.00
β_{LIC}	−1.45	0.344	−4.22	0.00
β_{INSTRU}	−0.991	0.216	−4.59	0.00
β_{GENDER}	−0.908	0.301	−3.02	0.00

Loglikelihood Maximization has been applied using the software Biogeme version 2.1 (Bierlaire, 2003) to estimate the parameters of the utility functions for the same 70% portion of the described database (just as with the *DT* calibration procedure) in the city of São Carlos. The results of this calibration are shown in Table 3 including the statistical significance (t-ratio and significance) of each attribute with a level of confidence of 95%.

The values of the statistics ρ^2 and Adjusted-ρ^2 for this estimation are 0.361 and 0.345 respectively. Ortúzar and Willumsen (2011) suggest that values of 0.400 for these statistics indicate good fit of data to a specific model, so the use of the proposed utilities look like a reasonable approximation to the collected data.

From these results it can be seen that the variables $\beta_{MORCONDICA}$, $\beta_{MORCONDI\_1}$, $\beta_{MORESTUDA}$ and $\beta_{MORIDADE}$ are not significant at a level of confidence of 95% and, thus, do not affect the value of the utility of the modes significantly. As a second attempt to calibrate the significant parameters, these socioeconomic variables have been removed from Eq. (11) which represents the utility of trips by CAR and the other equations have been fixed as described previously.

$$UCAR = \beta_{MOTO} \cdot \text{MOTO} + \beta_{CAR} \cdot \text{CAR} + \beta_{LIC} \cdot \text{LIC} + \beta_{GENDER} \cdot \text{GENDER} + \beta_{INSTRU} \cdot \text{INSTRU} + \beta_{MAX\_VIANUM} \cdot \text{MAX\_VIANUM}. \quad (14)$$

From Biogeme (Bierlaire, 2003), the results of Table 4 have been obtained from the calibration of these equations with the same 70% of the database. The statistics ρ^2 and Adjusted-ρ^2 considering the changed utility expressions for the CAR mode are 0.360 and 0.349 respectively which are slightly lower than the previous ones but are still close to the recommendation of Ortúzar and Willumsen (2011).

In this case, it can be seen that all the attributes significantly affect the value of the utility of the mode CAR and, thus, the probability of a traveler choosing this or the other modes. In order to compare the performance of the MNL approach with the *DT* segregation, the 30% remaining records of the dataset have been used to estimate the utilities of each mode – and their probabilities of being chosen – considering the level of the attributes of the sample population.

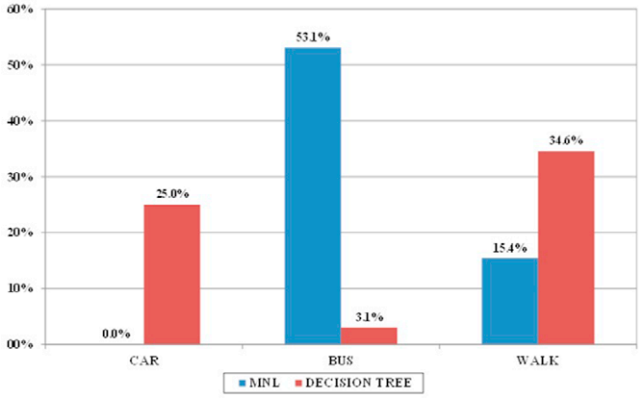

Fig. 4. Comparison of hits between the DT and MNL estimations.

4.2.3. Comparing the chosen variables of DT and LOGIT

The percentage of hits between the actually chosen modes and the Decision Tree and the Multinomial Logit approach are shown in Fig. 4. These values have been calculated by comparing the individuals' choices and those estimated from the probabilistic mode choice models.

From these results it can be concluded that the performance of the *DT* is as good as the MNL since the percentage of hits compared to the real chosen modes are of 68% and 70% respectively. Thus, the former method may be effectively used to select the variable to be kriged and to predict the mode choice according to the travelers' location, with the advantages of the straightforward application of the *DT* techniques.

4.3. Spatial patterns analysis

In order to match the categorical variable "Main Travel Mode" of the described database to the requirements of the geostatistical technique, this variable should be converted to a numeric continuous one for spatial interpolation of the mode choice in the unobserved locations. In this paper, it has been done by considering the mode choice as a regionalized variable given by the probability of a traveler choosing the car/motorcycle, the transit and the non-motorized travel modes for a trip. Here, it is important to highlight that the sum of these three probabilities must be always equal to 1.

Given that good estimations in the spatial interpolation steps of the Ordinary Kriging depend on the spatial structure of the variable to be kriged, three spatial patterns of the mode choice given by exploratory maps of probabilities have been created from the results of the *DT*. Fig. 5(a)–(c) shows the spatial distribution of the probabilities that the sampled households have to choose the private motorized mode, the transit mode and non-motorized travel mode respectively.

From the exploratory spatial analysis of Fig. 5(a)–(c) it can be verified that the variable to be kriged does not have any clear spatial pattern. So, in order to identify any spatial pattern, the studied city has been divided into six regions considering the income and the location of the interviewed households.

Therefore, a two-step clustering method has been applied to the dataset in the city of São Carlos with the categorical variable "Minimum Wage" (Table 1) and the geographic coordinates of households (latitude and longitude in meters). Fig. 6 shows the six segregated regions which are defined by the percentage of households with the lowest range of income (0 to 2 Minimum Wages).

As an example, from the probabilities estimated by the *DT* application (since their results are robust when compared to the MNL modeling) the segregated data of Fig. 6 have been linked to geographic coordinates of 110 sampled households in Region 2 and 518 households in Region 5.

A detailed analysis of the probability maps of each of the six regions separately revealed that Region 2 has the more diffusive spatial pattern for the probabilities of choosing each of the available travel mode, ranging from the center to the periphery of the area as shown in Fig. 7(a)–(c).

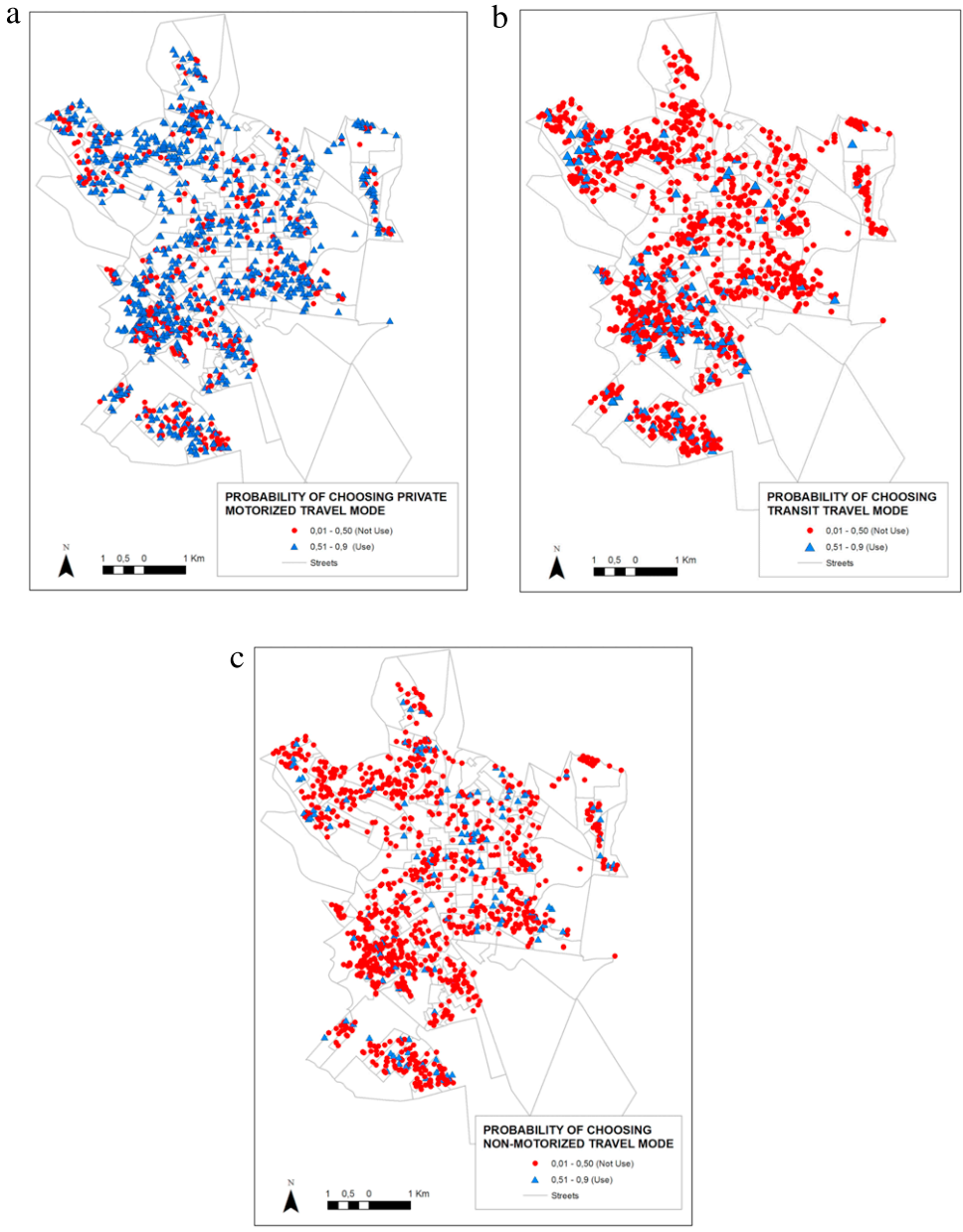

Fig. 5. Spatial distribution of mode choice in the city of São Carlos.

This region consists of both low income (center of the region) and high income (peripheral neighborhoods) households and, in addition, it has a new campus of the already existing university, which significantly affects the mode choice of trips from the locations within the area. Therefore, from the calculated probabilities of the *DT*, the Ordinary Kriging has been applied to the set of georeferenced information of Region 2 (110 observations).

In order to increase the number of observations of this geostatistical modeling, the 518 additional observations of Region 5 have been included to the dataset. Since this is a central and higher household

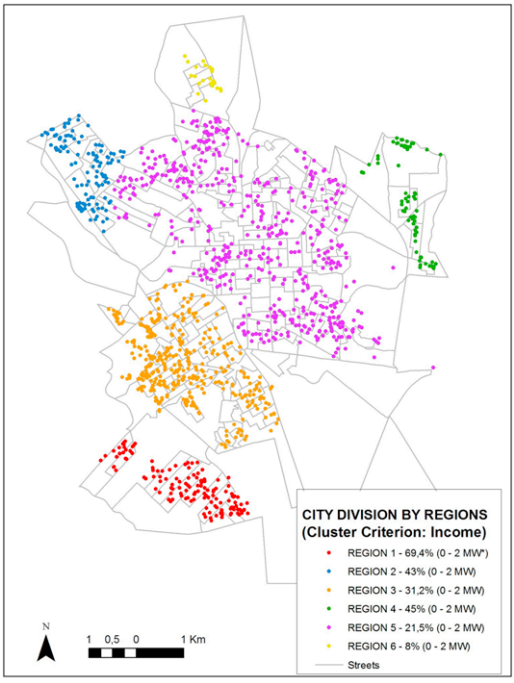

Fig. 6. Clustering results of the mode choice in the city of São Carlos considering the percentage of minimum wage as a metric.

income area – with only 21.5% of the sampled inhabitants earning the lowest range of the minimum wage – the proposed method has been applied to a dataset different from Region 2 to demonstrate the versatility of the method proposed in this paper.

4.4. Ordinary Kriging (OK) application

As mentioned before, in the non-spatial step of the proposed method the probabilities of mode choice considering socioeconomic independent variables have been estimated. The second spatial step described in this section considers the results of the *DT* modeling (probability of mode choice) as variables to be kriged. Three regionalized variables have been considered in the application to the database of the city of São Carlos: (i) the probability of a traveler choosing the car; (ii) the probability of choosing a non-motorized mode; (iii) and the probability of choosing transit mode.

Based on the modeling approach for the *OK*, experimental variograms from the observed points for the three variables have been constructed and the theoretical models have been adjusted with Spherical functions. Table 5 summarizes the parameters of the theoretical variograms of each regionalized variable in Region 2 and Region 5 and Fig. 8(a)–(c) illustrates the theoretical variograms in the main direction for these three regionalized variables in Region 2. Finally, Fig. 9(a)–(c) shows the theoretical variograms of the sampled dataset of Region 5.

4.4.1. Cross validation

In order to assess the accuracy of the models defined by the theoretical variograms, a few statistical measures such as mean of residuals, variance of errors and Root Relative Squared Error (RMSE) have been calculated taking into account the observed and estimated values.

For the geostatistics modeling validation, 30% of the observations from Region 2 and Region 5 have been randomly selected resulting in 30 and 142 georeferenced points in each region, respectively, to be used in cross validation. As mentioned in the literature review section, cross-validation enables the

Fig. 7. Spatial pattern of mode choice in Region 2.

validation of the estimated values and also assesses the goodness of the fitted model and parameters applied to the kriging method.

The statistical measures calculated to evaluate the goodness of fit of the dataset and the models are shown in Table 6. In addition, the percentages of correct estimations of the travel mode against the observed choices in the validation dataset are exhibited in the referred table.

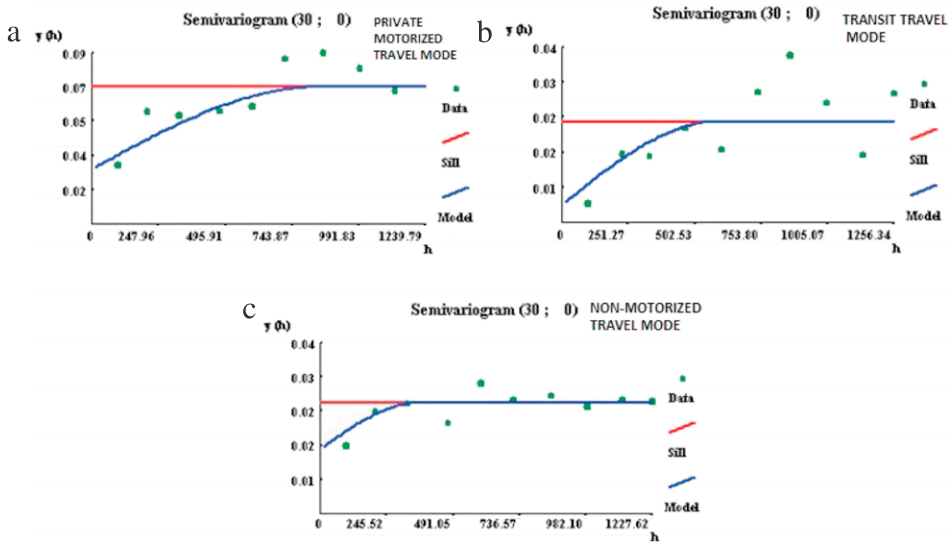

Fig. 8. Theoretical variograms of mode choice for the Region 2.

Table 5
Parameters of the theoretical variograms of regionalized variables in Region 2 and Region 5.

	Direction	Nugget effect (Co)	Range (a) (meters)	Sill (C+Co)	Structure
Regionalized variable—Region 2					
Private motorized mode probabilities	N30E N60W	0.03	833 363	0.072	Spherical
Transit mode probabilities	N30E N60W	0.00	583 188	0.023	Spherical
Non-motorized mode probabilities	N30E N60W	0.02	512 361	0.026	Spherical
Regionalized variable—Region 5					
Private motorized mode probabilities	N30E N60W	0.06	1369 1004	0.081	Spherical
Transit mode probabilities	N25E N65W	0.01	1378 1763	0.015	Spherical
Non-motorized mode probabilities	N25E N65W	0.02	1418 951	0.034	Spherical

From Table 6 it can be concluded that the mean of the residuals (difference between observed and estimated) and their variance are close to zero for all the probabilities of mode choice in both the regions, which represents that the modeled probabilities are closely related to the modes actually chosen by the travelers.

Moreover, the percentage of correct estimations are reasonably high, with values ranging from 88% for the transit mode probabilities in Region 5 to 67% for the motorized mode in the same region, except for this last mode for the Region 2 with a low value of 49%.

4.4.2. Kriging maps

When applying the *OK* method for the interpolation of the three choice probabilities of travel mode as regionalized variables, a grid of cells with 100 m per 100 m has been established based on the distance between the position of the households in the study area of the city of São Carlos.

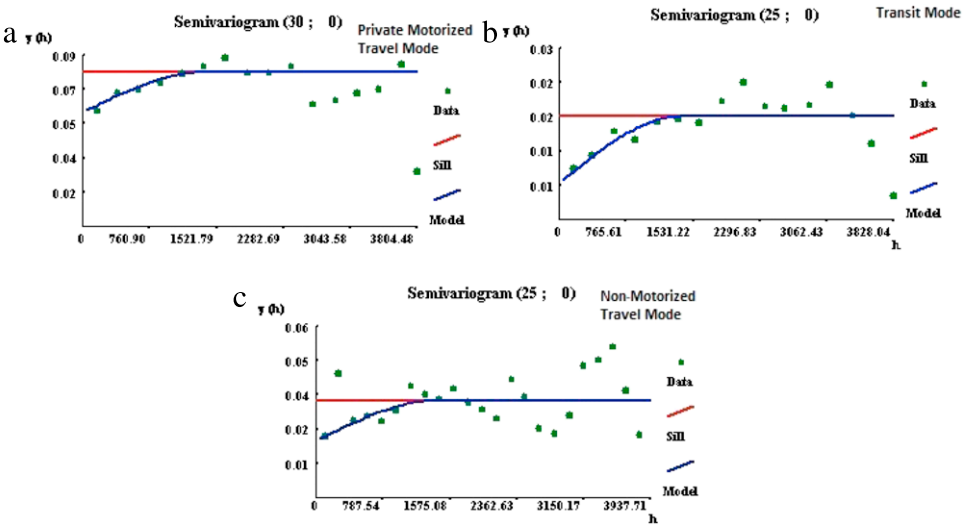

Fig. 9. Theoretical variograms of mode choice for the Region 5.

Table 6
Results of cross validation.

	Root mean squared error	Mean of residuals	Variance of errors	% Correct estimation
Probabilities of regionalized variables—Region 2				
Private motorized mode	42%	−0.003	0.079	49%
Transit mode	33%	−0.004	0.032	73%
Non-motorized mode	10%	0.003	0.027	79%
Probabilities of regionalized variables—Region 5				
Private motorized mode	38%	−0.005	0.069	67%
Transit mode	18%	−0.001	0.007	88%
Non-motorized mode	10%	0.003	0.024	80%

Firstly, some important information that may affect the accuracy of the theoretical variograms and proposed method for mode choice estimation are listed in the following items:

- A data mining technique has been applied to estimate the mode choice with known geographic coordinates and it has been considered that mode choice is a problem of pattern recognition.
- The regionalized variables are not natural (*i.e.*, they derive from a previous database treatment and not from direct measures) and are estimated from a *DT* modeling approach which is a non-parametric model.
- It is important to mention that this two-step method has a propagation of uncertainties since the mode choice models have inherent estimation errors and that the regionalized variables, chosen from their results, are modeled in the kriging step.
- These regionalized variables apparently do not have significant spatial patterns as shown in Fig. 5(a)–(c).

The collected data from the described surveys are consistent but may contain biased information due to the application of the questionnaires and the data transcription.

From this information, the kriging maps have been generated by interpolating the estimated probabilities of each of the three available travel modes as illustrated in Fig. 10(a)–(c).

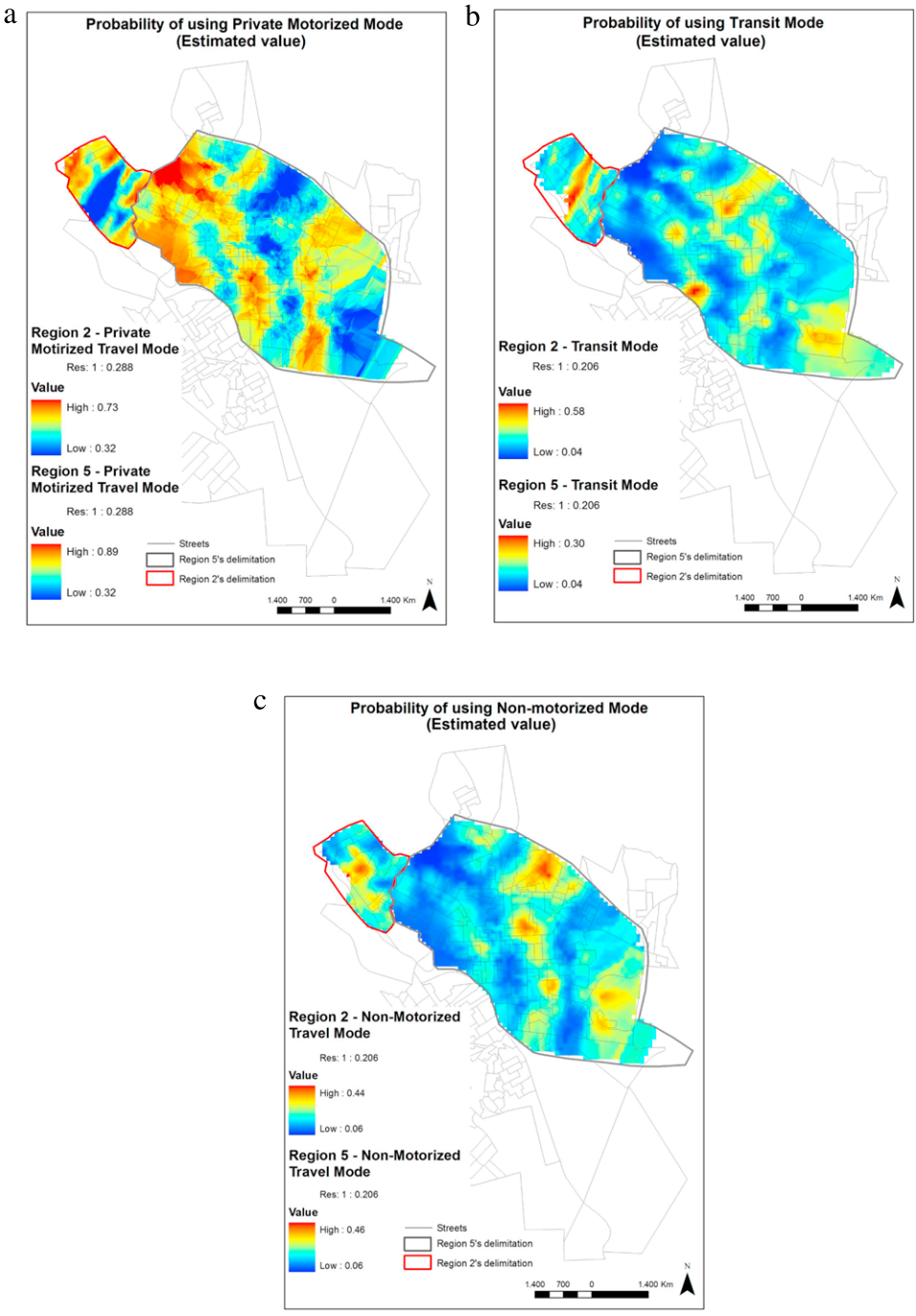

Fig. 10. Kriging maps from the mode choice analysis of Region 2 and Region 5 in the city of São Carlos.

Considering the probabilities of mode choice in Region 2 kriged from the information provided by the *DT* and the household's locations, the probability of choosing the private motorized travel mode

is higher in the periphery of the region. Therefore, the willingness of car usage increases the farthest the traveler is located from its center.

In opposition to the tendencies observed for the motorized mode, the choices for the transit and non-motorized modes increase from the periphery to the center area of Region 5 in a main direction of south to north, opposed to a direction of southeast to northwest in the case of the motorized mode.

The results of spatial interpolation in Region 2 are consistent with the observed conditions of the region where locations with higher probability of car/motorcycle (motorized mode) choice are exactly those of higher income population. On the other hand, the center of this region shows less probability of this mode choice, which corresponds to lower income households' area.

5. Conclusions

In this paper we have estimated the probabilities of mode choice of households in the city of São Carlos considering a set of socioeconomic variables and their geographical position based on observed values from different surveys by jointly applying a Decision Tree analysis and the Ordinary Kriging method. The *DT* technique (*CART*) has shown to be effective since its comparison with the Multinomial Logit approach indicates that the variables chosen by both the methods are the same, with the advantages of the straightforward application of the *DT* method.

The application of the Decision Tree segregation has enabled to predict the individuals' mode choices and their socioeconomic characteristics. An analysis of the results of this procedure has led to interesting conclusions as, for example, the fact that individuals (males) with a higher education degree, who possess car or motorcycle, and driver's license, are more likely to use private motorized travel mode.

From the data partition algorithm (*CART*), the eleven terminal nodes resulted from the *DT* have grouped those individuals susceptible to use a specific travel mode considering their socioeconomic characteristics and those of the transportation system.

The maps obtained through the Ordinary Kriging have shown that there is a trend in the choice of the private motorized travel mode, which increases from the center to the periphery of a region outside downtown. In opposition, a different trend has been found in the central region of São Carlos (São Paulo/Brazil), where there is an increase in motorized mode choice from southeast to northwest of this area.

Regarding modeling, cross validation has provided good results considering the mean and the variance of the residuals from the difference between the observed mode choice (those estimated by the *DT* segregation) and the estimated value from *OK*. A percentage of correct responses over 70% for the non-motorized and transit modes estimations has been achieved for Region 2 (outside downtown). Moreover, in the city center (Region 5) this percentage of correct estimations was over 80% for the non-motorized travel and transit modes.

An important aspect of the analysis is that the regionalized variables are not natural (i.e., not directly measured but estimated with the *DT*, a nonparametric modeling approach), which may have incurred in probable estimation bias from the type of data considered and the sequential application of a non-spatial and a spatial modeling approach.

Nevertheless, the innovation in this study that should be taken into account is the fact that the two-step method presented here is based in unusual techniques of spatial analysis of mode choice: Decision Tree Analysis and Ordinary Kriging. The results of this paper have shown the effectiveness of joining these methods – which has enabled a preliminary assessment of the spatial features of the households regarding their mode choice – given that they are extremely dependent of the quality and robustness of the database.

Acknowledgment

This research has been sponsored by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (*CNPQ*) (307908/2012-7), a Brazilian Federal Higher Education Funding Agency. We also

thank the financial support of the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - 2013/25035-1.

References

Adjemian, M.K., Lin, C., Williams, J., 2010. Estimating spatial interdependence in automobile type choice with survey data. Transp. Res. A: Policy Pract. 44, 661–675. http://dx.doi.org/10.1016/j.tra.2010.06.001.
Ahern, A., Tapley, N., 2008. The use of stated preference techniques to model modal choices on interurban trips in Ireland. Transp. Res. A: Policy Pract. 42, 15–27. http://dx.doi.org/10.1016/j.tra.2007.06.005.
Arentze, T., Timmermans, H., 2007. Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. Transp. Res. B: Methodological 41 (7), 772–783. http://dx.doi.org/10.1016/j.trb.2007.01.001.
Bhat, C., Zhao, H., 2002. The spatial analysis of activity stop generation. Transp. Res. B 36, 557–575. http://dx.doi.org/10.1016/S0191-2615(01)00019-4.
Bierlaire, M., 2003. BIOGEME: A free package for the estimation of discrete choice models. In: Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland.
Breiman, L., Friedman, J.H, Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group, California.
Cervero, R., Radisch, C., 1996. Pedestrian versus automobile oriented neighborhoods. Transport Policy 3, 127–141. http://dx.doi.org/10.1016/0967-070X(96)00016-9.
Ciuffo, B.F., Punzo, V., Quaglietta, E., (2011) Kriging meta-modelling to verify traffic micro-simulation calibration methods. In: TRB 90th Annual Meeting Compendium of Papers.
Domencich, T.A., McFadden, D., 1975. Urban Travel Demand: A Behavioral Analysis. North Holland Publishing, Amsterdam.
Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Applied Geostatistics Series. Oxford University Press, New York, Oxford, p. 483.
Goovaerts, P., 1999. Using elevation to aid geostatistical mapping of rainfall erosivity. Catena 34, 227–242.
Grange, L., González, F., Vargas, I., Muñoz, J.C., 2013. A polarized logit model. Transp. Res. A: Policy Pract. 53, 1–9. http://dx.doi.org/10.1016/j.tra.2013.06.003.
IBGE, 2010. Brazilian Institute of Geography and Statistics. Census Brazilian population in 2010. Available in http://www.ibge.gov.br (in Portuguese) (accessed 20.08.12).
Isaaks, E.H., Srvastava, R.M., 1989. An introduction to Applied Geostatistics. Oxford University Press.
Ji, J., Gao, X., 2010. Analysis of people's satisfaction with public transportation in Beijing. Habitat Internat. 34 (4), 464–470. http://dx.doi.org/10.1016/j.habitatint.2009.12.003.
Kitamura, R., Mokhtarian, P.L., Laidet, L., 1997. A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. Transportation 24, 125–158. http://dx.doi.org/10.1023/A:1017959825565.
Matheron, G., 1971. The Theory of Regionalized Variables and its Applications. Technical Report 5. Paris School of Mines. Cah. Cent. Morphol. Math., Fontainbleau.
Matheron, G., 1963. Principles of geostatistics. Econ. Geol. 58, 1246–1266. http://dx.doi.org/10.2113/gsecongeo.58.8.1246.
Mazzella, A., Piras, C., Pinna, F., 2012. Use of kriging technique to study roundabout performance. Transp. Res. Record: J. Transp. Res. Board 2241/2011, 78–86. http://dx.doi.org/10.3141/2241-09.
Miura, H., 2010. A study of travel time prediction using universal kriging. TOP 18 (1), 257–270. http://dx.doi.org/10.1007/s11750-009-0103-6.
Ortúzar, J.D., Willumsen, L.G., 2011. Modelling Transport, fourth ed. Wiley.
Páez, A., López, F.A., Ruiz, M., Morency, C., 2013. Development of an indicator to assess the spatial fit of discrete choice models. Transp. Res. B 56, 217–233. http://dx.doi.org/10.1016/j.trb.2013.08.009.
Peer, S., Knockaert, J., Koster, P., Tseng, Y.Y., Verhoef, E.T., 2013. Door-to-door travel times in RP departure time choice models: An approximation method using GPS data. Transp. Res. B 58, 134–150. http://dx.doi.org/10.1016/j.trb.2013.10.006.
Pitombo, C.S., Sousa, A.J., Birkin, M., Quintanilha, J.A., 2010. Comparing different spatial data analysis to forecast trip generation. In: World Conference on Transport Research Society, 2010, Lisbon. Proceedings of the 12th WCTR. Lisboa.
Pitombo, C.S., Kawamoto, E., Sousa, A.J., 2011. An exploratory analysis of relationships between socioeconomic, land use, activity participation variables and travel patterns. Transport Policy (Oxford) 18, 347–357. http://dx.doi.org/10.1016/j.tranpol.2010.10.010.
Rodrigues da Silva, A.N., 2008. Preparation of a Travel Database for Assistance of Development Researches in Transportation Planning Area. FAPESP Report, Case No. 04/15843-4. School of Engineering of São Carlos, University of São Paulo, Brazil (in Portuguese).
Sen, A., Sööt, S., Pagitsas, S, E., 1978. The logit modal split model: Some theoretical considerations. Transp. Res. A 12 (5), 321–324. http://dx.doi.org/10.1016/0041-1647(78)90006-0.
Shmueli, D., Salomon, I., Shefer, D., 1996. Neural network analysis of travel behavior: Evaluating tools for prediction. Transp. Res. C: Emerg. Technol. 4 (3), 151–166.
Tong, D., Lin, W.H., Stein, A., 2013. Integrating the direction effect of traffic into geostatistical approaches for travel time estimation on an urban transportation network. Int. J. ITS Res. 11 (3), 101–112.
Wackernagel, H., 2010. Multivariate Geostatistics: An Introduction with Applications, Third ed. Springer.
Wang, Y., Zhuang, D., Liu, H., 2012. Spatial distribution of floating car speed. J. Transp. Syst. Eng. Inform. Technol. 12 (1), 36–41. http://dx.doi.org/10.1016/S1570-6672(11)60182-7.
Williams, H.C.W.L., Senior, M.L., 1977. Model based transport policy assessment. Part II: Removing fundamental inconsistencies from the models. Traffic Eng. Control 18, 464–469.
Xie, C., Jinyang, L., Parkany, E., 2007. Work travel mode choice modeling with data mining: Decision trees and neural networks. Transp. Res. Record: J. Transp. Res. Board 1854, 50–61. http://dx.doi.org/10.3141/1854-06.

Zhang, D., Wang, X., 2013. Traffic Volume Estimation using Network Interpolation Techniques: An Application on Transit Ridership in NYC Subway System. Final Report. http://www.utrc2.org/sites/default/files/pubs/Final-Traffic-Volume-Interpolation_0.pdf.

Zou, H., Yue, Y., Li, Q., Yeh, A.G.O., 2012. An improved distance metric for the interpolation of link-based traffic data using kriging: a case study of a large-scale urban road network. Int. J. Geograph. Inform. Sci. 26 (4), 667–689. http://dx.doi.org/10.1080/13658816.2011.609488.