DOI 10.1590/s1982-21702021000100008

ORIGINAL ARTICLE

MODELING SPATIAL EFFECT ON TRAVEL MODE CHOICE USING A SYNTHETIC SPATIALLY CORRELATED DATA SET

Lucas Assirati[1] – ORCID: 0000-0002-0118-2665

Cira Souza Pitombo[1] ORCID: 0000-0001-9864-3175

[1]São Carlos School of Engineering-University of São Paulo (EESC-USP), Department of Transportation Engineering, SãoCarlos-SP, Brazil.
E-mail: assirati@usp.br, cirapitombo@usp.br.

Received in 6th April 2020.
Accepted in 18th February 2021.

Abstract:

Urban dynamics can be characterized more effectively by considering spatial aspects in studies. This paper, using a synthetic spatially correlated data set, aims to model the spatial effect on travel mode choice based on geostatistics precepts. A method was proposed based on three main steps. The first step consists of building synthetic spatially correlated data, using the intrinsic spatial dependence on travel demand data and mathematical principles of bilinear interpolation. The following two steps correspond to the modeling approach. The Exploratory Spatial Data Analysis stage aimed to attest the existence of spatial autocorrelation of the data set using two indicators: Moran and G-SIVAR (Global Spatial Indicator Based on Variogram). The Confirmatory Spatial Data Analysis stage proposed the calibration of two Binomial Logit models. The first model includes only the original database variables (non-spatial model). The second one is analogous to the original but added to spatial covariates obtained by geostatistical concepts (spatial model). A 15% increase in cross-validation hit rates is achieved when spatial variables are included. This paper presents three significant research contributions: (1) The methodological procedure to model spatial effect on travel mode choice; (2) The proposal of spatial covariates based on geostatistical assumptions; and (3) The suggestion of a simple procedure to propose a simulation of a spatially correlated database.

Keywords: Spatial Statistics; MORAN; G-SIVAR; Spatially Dependent Discrete Choice Models.

How to cite this article: ASSIRATI, L.; PITOMBO, C. S. Modeling spatial effect on travel mode choice using a synthetic spatially correlated data set. *Bulletin of Geodetic Sciences*. 27(1): e2021008, 2021.

1. Introduction

Discrete choice models are a traditional tool, widely used in travel demand studies, mainly in travel mode choice problems (Ben-Akiva, 1974; McFadden, 1974; Bhat, 1995; Hess, 2005; Bhat et al., 2008; Ahern and Tapley, 2008; Qin et al., 2017). However, it is worth mentioning that it is challenging to know all the factors that affect individual decisions since there is a heterogeneity underlying travel behavior. Moreover, home and destination locations, the spatial distribution of activities in the urban environment, and land-use variables can influence travel mode choices, as well as personal and alternative attributes.

So, the spatial analysis of travel demand has become a potential line of research, especially given the requirement of including spatial effects on mathematical models (Páez et al. 2013). Hence, many studies (Miyamoto et al., 2004; Zhou, 2012; Páez et al., 2013; Pitombo et al., 2015; Lindner and Pitombo, 2018; Assirati, 2018; Lindner et al., 2021) started to incorporate variables related to geographic location to travel demand forecasting studies, promoting the improvement of estimates. So, this paper mainly proposes to model the spatial effect on travel mode choice, taking into account the spatial analysis of the travel demand research field.

The spatial analysis consists of a set of techniques that are designed to analyze the geographical position of observations of a variable, evaluating the possible relationships between the values of the variable and different locations. Several questions can be answered through spatial analysis. The spatial analysis of a phenomenon comprises a preliminary, exploratory step, in which the degree of spatial dependence, either locally or globally, of a data set is visualized or measured. After confirming the spatial dependence of the data, the confirmatory spatial analysis stage is indicated. Consequently, models started to be employed to include the spatial structure of the data.

Many studies carried out only Exploratory Spatial Data Analysis (ESDA), considering different research fields (Anselin, 1996; Unwin and Unwin; 1998; Messner et al., 1999; Le Gallo and Ertur; 2003). Among the ESDA procedures, the degree of spatial dependence can be measured by spatial indicators (Naizer et al., 2019). Over many decades, different authors proposed spatial indicators to corroborate the existence of spatial autocorrelation (Moran, 1950; Geary, 1954; Getis and Ord, 1992; Anselin, 1995). Global indexes qualify the database, while local indexes characterize the spatial association of observations, identifying spatial association pockets. Most indexes could be applied to quantitative variables (Anselin, 1995; Naizer et al., 2019).

In this paper, the global indicators Moran and G-SIVAR (Global Spatial Indicator Based on Variogram) were used in the ESDA methodological stage. The Moran global indicator is based on correlograms (also known as autocorrelation diagrams, are graphs of the sample's autocorrelations versus distance units). It indicates the similarity of the data and ranges from -1 to 1. Negative values indicate negative spatial autocorrelation, while positive values indicate positive spatial autocorrelation. The value 0 indicates the absence of spatial correlation (Moran, 1950; Anselin, 1995).

The global indicator G-SIVAR (Naizer et al., 2019) is based on assumptions of geostatistics, more specifically, on the semivariogram tool (theoretical and experimental). The G-SIVAR indicator is based on the standardized values of the theoretical variogram (ordinate axis), associated with a hypothesis test for spatial randomness. It indicates the dissimilarity of the data and ranges from 0 to 1, where value 1 shows the complete absence of spatial autocorrelation, and value 0 indicates the entire presence of spatial autocorrelation. The original studies of Moran (1950) and Naizer et al. (2019) should be consulted if the reader wants more details about the mathematical formalism used.

The Confirmatory Spatial Data Analysis (CSDA) includes the set of estimation models, in addition to validation procedures. In recent years, different spatial models have been used in the travel demand field. There are various studies regarding Geographically Weighted Regression application (Chow et al., 2006; Nowrouzian and Srinivasan, 2014; Pitombo et al., 2010) or spatial regression models that consider the spatial autocorrelation including an explanatory spatial variable (Wang, 2001; Lopes et al., 2014). This paper presents a kind of spatial regression model for discrete choice models. The CSDA stage of this study took place applying Binomial Logit models based on the inclusion of spatial covariates.

Decisions between travel modes are a classic example of discrete choice, where one must choose an alternative within a finite set of possibilities (car, public transportation, walking, etc.). The attributes of travel mode options (such as travel time, cost, etc.) and the attributes of individuals (income, age, gender, etc.) are used to calculate the probability of choosing the possible travel mode. It is known, however, that the neighborhood can influence the attributes involved in the discrete choice models. Many techniques have been proposed to deal with discrete choice estimation when spatial dependence is present, and the well-known Spatially Dependent Discrete Choice Models have been applied in recent years (Fleming, 2004; Dugundji and Walker, 2005; Sener et al., 2011; Paez et al., 2013). Then, regionalized or spatial variables started to be incorporated into discrete choice models to promote better estimates. Using an approach based on geostatistics precepts, this research proposes to extract spatial covariates providing increments to parametric modeling. The proposal of an independent spatial variable, based on the geostatistical approach, could be considered a significant methodological contribution.

Thus, this study seeks to attest the spatial association of the data set in the ESDA stage using a synthetic spatially correlated data set. Next, the comparison of two Binomial Logit models is made in the CSDA stage. The first model, with only traditional covariates and the second one incorporating spatial covariates based on geostatistics assumptions.

Additionally, despite the observed advances in population synthesizers for travel data microsimulation, none of the approaches recognize the spatial correlation of data as a relevant input to reproduce travel behavior. Recently, Sequential Gaussian Simulation was presented as a promising simulation tool in Travel Demand Modeling (Lindner and Pitombo, 2019). Then, the specific objective of this paper is to propose a simulation of a spatially correlated database, using the intrinsic spatial dependence on travel demand data and mathematical principles of bilinear interpolation.

This paper presents three notable research contributions: (1) The methodological procedure to model spatial effect on travel mode choice; (2) The proposal of spatial covariates based on geostatistical assumptions; and (3) The suggestion of a simple procedure to propose a simulation of a spatially correlated database.

This article has five sections, in addition to this introduction. Section 2 provides a brief definition of the tools used in the methodological procedure. Section 3 describes the initial database used for data simulation. Section 4 describes the method adopted. Section 5 presents the results and discussions. Finally, Section 6 presents the main conclusions and contributions.

2. Theoretical basis: a brief description of the statistical tools

2.1 Linear and bilinear interpolation

The linear interpolation method is used to extrapolate new values within a discrete range of a previously known dataset. When having a discrete mesh containing some known values, it is useful to infer a value for a location that has no given value. Assuming that to estimate the value A_1 at a point located at the coordinates (x, y_1) and previously knowing values V_{11} and V_{21}, respectively located at the coordinates (x_1, y_1) and (x_2, y_1) as shown in Figure 1a, the value of A_1 from Equation 1 can be inferred:

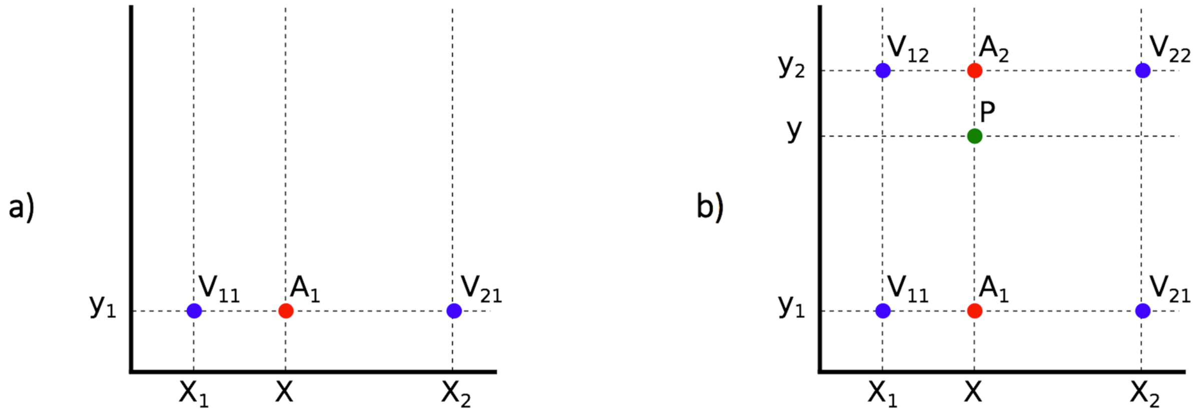


Figure 1: a) An example of linear interpolation. b) An example of bilinear interpolation

$$A_1 = \left(\frac{x_2 - x}{x_2 - x_1}\right) \times V_{11} + \left(\frac{x - x_1}{x_2 - x_1}\right) \times V_{21} \qquad (1)$$

This expression can be interpreted as a weighted average where the weights $\left(\frac{x_2 - x}{x_2 - x_1}\right)$ and $\left(\frac{x - x_1}{x_2 - x_1}\right)$ are inversely related to the distance from endpoints V_{11} and V_{21} to point A_1 whose value is unknown. Thus, point V_{11}, being closer to A_1, has more influence than V_{21} in the interpolation process. The weights should be normalized by the distance $(x_2 - x_1)$ between V_{11} and V_{21} since their sum should be 1.

From the example presented in Figure 1a, one can note that the interpolation for point A_1 involved only V_{11} and V_{21} values and their x-axis coordinates. However, many areas of knowledge deal with data spread in a two-dimensional mesh. Therefore, it is desirable to extend data interpolation to a bidimensional scenario known as bilinear interpolation. Bilinear interpolation can be interpreted as the performance of linear interpolation in one direction, followed by a second application of linear interpolation in the other direction. Although calculations are applied in linear terms at each step, the bilinear interpolation, as a whole, is a quadratic metric at the location whose value is needed to be inferred. In addition to points V_{11}, V_{21} and A_1 of the previous example (Figure 1a), let us now consider points V_{12}, V_{22}, A_2 and P located respectively at the coordinates (x_1,y_2), (x_2,y_2), (x,y_2) and (x,y) presented in Figure 1b.

The bilinear interpolation method must be used to obtain the value of P in the (x, y) coordinate. It implies that for the x-axis, a linear interpolation should be applied to obtain the value of A_1 (Equation 1), and similarly to obtain the value of A_2 (Equation 2):

$$A_2 = \left(\frac{x_2 - x}{x_2 - x_1}\right) \times V_{12} + \left(\frac{x - x_1}{x_2 - x_1}\right) \times V_{22} \qquad (2)$$

Then, for the y-axis, linear interpolation is applied one more time to obtain the value of P finally, according to Equation 3:

$$P = \left(\frac{y_2 - y}{y_2 - y_1}\right) \times A_1 + \left(\frac{y - y_1}{y_2 - y_1}\right) \times A \qquad (3)$$

At first, points A_1 and A_2 were obtained by a weighted average for elements only belonging to the x-axis and values V_{11}, V_{21}, V_{12}, and V_{22}. In a second moment, point P was obtained by a weighted average for elements only belonging to the y-axis and values A_1 and A_2. It is important to note that the final result of bi-linear interpolation does not depend on which axis was used in the first and second calculations. If one initially uses the y-direction for the calculations, followed by the x-direction, an analogous result must be obtained.

2.2 Variograms

Adjusting experimental variograms (obtaining theoretical variograms) is the procedure to measure spatial dependence before and after data simulation. Additionally, the semivariogram was the tool used for the G-SIVAR (Global Spatial Indicator based on the Variogram). G-SIVAR is based on the standardized values of the theoretical variogram (ordinate axis), associated with a hypothesis test.

The variogram is the mathematical description of the relationship between the variance of pairs of observations and the distance separating them (h). In the experimental determination of the variogram, for each value of “h”, all pairs of observations z (x) and z (x + h) are considered, according to Equation 4 (Krige, 1951; Matheron, 1971; Cressie, 1993):

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=l}^{N(h)} [z(x_i) - z(x_i + h)]^2 \qquad (4)$$

where: γ (h) is the variance;

N(h) is the number of measured value pairs; and

z(x) and z(x+h) are pairs of observations separated by the vector “h”.

After the calculation of the experimental variograms, a mathematical model that best represents the variability under study should be established. From the various theoretical models available for variogram adjustments, the most applied in geostatistics are the spherical (Figure 2), Gaussian, and exponential. There are some parameters of the theoretical variogram, illustrated in Figure 2, that must be observed to determine the spatial structure of the variables.

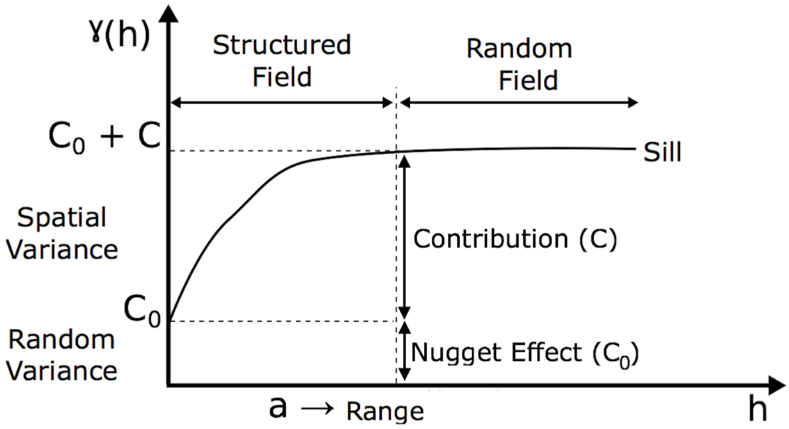


Figure 2: An example of a theoretical variogram (Cressie, 1993).

(1) Range (a): the distance where the samples are spatially correlated; (2) Sill (C_0+C): the maximum value of γ (variance) on the range curve (a). It is obtained by the sum of the nugget effect (C_0) and the contribution (C); (3) Nugget Effect (C_0): is the starting point of the range curve (a) touching the γ-axis. The nugget effect reflects the random variance for short distances. A high value of this parameter indicates that significant variations are found in short distances.

In this paper, taking into account the step of building synthetic spatially correlated data to analyze the spatial structure of the variables (total of six) that compose the initial and simulated data samples, the main parameters of the theoretical variograms of the six variables will be observed.

3. Initial database

Initially, we had a dataset representing a fictitious locality. The set consisted of 60 hypothetical coordinate points that ranged from 0 to 10 in steps of 0.5 for the x-axis and from 0 to 5 in steps of 0.1 for the y-axis. Associated with each of these 60 points were seven variables: Travel distance, Bus travel time, Car travel time, Bus fare, Car travel cost, Income indicator and a dichotomous choice variable that represented the use of bus (0) or car (1) by the individual represented by each of the 60 locations. Figure 3 shows the original points within the fictitious city, with the marks representing the travel mode choices. Table 1 shows the descriptive measures of the variables in the initial database.

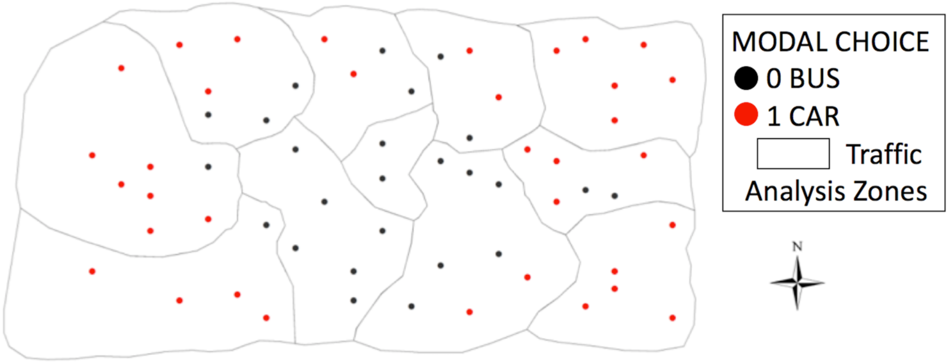


Figure 3: Representation of the fictitious city - 38.33% of the sample is formed by bus users (value 0, black dots) while 61.67% of the sample is formed by car users (value 1, red dots).

Table 1: Descriptive measures of the variables in the initial database.

	Travel distance (Km)	Bus travel time (Min.)	Car travel time (Min.)	Bus fare (Currency Unit)	Car travel cost (Currency Unit)	Income indicator (Scale 1 – 10)
Mean	2.58	5.62	3.87	2.05	2.19	5.23
Std deviation	1.38	2.71	2.07	0.73	1.17	2.50
Max Value	5.50	11.08	8.26	3.00	4.68	10.00
Min Value	0.10	1.17	0.15	1.50	0.09	1.00

4. Methodological procedure

A method was proposed based on three main steps. The first step consists of building synthetic spatially correlated data, using the intrinsic spatial dependence on travel demand data and mathematical principles of bilinear interpolation. The following two steps correspond to the modeling approach. The ESDA stage aimed to attest the existence of spatial autocorrelation of the data set using two indicators: Moran and G-SIVAR (Global Spatial Indicator Based on Variogram). The CSDA stage proposed the calibration of two Binomial Logit models. The first model includes just the original database variables (non-spatial model). The second one is composed by the original variables and some spatial covariates obtained by geostatistical concepts (spatial model). Figure 4 illustrates the methodological procedure. The following subsections describe the methodological stages.

Linear interpolation can be achieved by any method, from manual calculations to computational packages (which we suggest python). The spatial dependency calculations are made using the G-SIVAR R package accessible via https://github.com/pedreirajr/g-sivar.

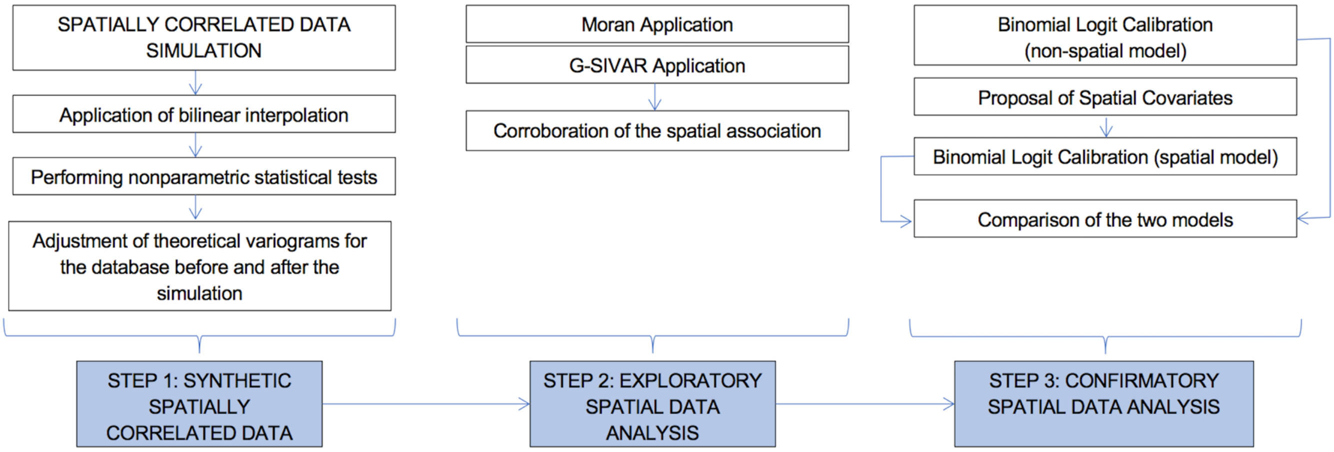

SPATIALLY CORRELATED DATA SIMULATION
Application of bilinear interpolation
Performing nonparametric statistical tests
Adjustment of theoretical variograms for the database before and after the simulation
Moran Application
G-SIVAR Application
Corroboration of the spatial association
Binomial Logit Calibration (non-spatial model)
Proposal of Spatial Covariates
Binomial Logit Calibration (spatial model)
Comparison of the two models
STEP 1: SYNTHETIC SPATIALLY CORRELATED DATA
STEP 2: EXPLORATORY SPATIAL DATA ANALYSIS
STEP 3: CONFIRMATORY SPATIAL DATA ANALYSIS

Figure 4: Methodological procedure scheme.

4.1 Step 1: Synthetic spatially correlated data

The present article proposes a method based on bilinear interpolation to supply absences and simulate a spatially correlated data set from pre-existing values. This simulation is analogous to a disaggregation or even filling gaps left by eventual data loss. The accuracy of the procedure is assessed by nonparametric statistical tests (Kolmogorov-Sminorv and Mann-Whitney) that indicate the similarity of population distributions of the new data generated compared to the original data used as the basis for the simulation. Once desirable accuracy is obtained, experimental variograms of the pre and post-simulation bases are modeled to evaluate whether the spatial structure of the data has been maintained.

4.2 Step 2: Exploratory Spatial Data Analysis (ESDA)

The ESDA of this study consisted of calculating two indicators of global spatial association: Moran and G-SIVAR. To do this, a division of data into performance area ranges was required. Initially, for ESDA, and later, for the proposal of the spatial covariates.

The dimensions of the sample space are ten units on the X-axis and five units on the Y-axis. Thus, the longest possible distance in this space is the main diagonal, with approximately 12 units. This concept is inherited from geostatistics and is called cut-distance. Thus, four neighborhood strips were formed through the equal division of the diagonal into four plots: (I) 0-3 units; (II) 3.1-6 units; (III) 6.1-9 units; (IV) 9.1-12 units. The constitution of the bands is also another concept inherited from geostatistics and is called lag-distance. It refers to the segments, delimited by the cut-distance. All lag-distances are commonly the same size and equidistant.

After calculating both indicators for the given distance ranges, not only the values of the indicators should be evaluated, but also the statistical significance for spatial autocorrelation. Thus, the neighborhoods considered for determining the spatial covariates will be those statistically significant (p-value ≤ 0.05) for the hypothesis tests associated with the calculated global indicators.

4.3 Step 3: Confirmatory Spatial Data Analysis (CSDA)

4.3.1 Proposal of Spatial Covariates

This step was one of the most important of the methodological procedures. It constituted an essential contribution, taking into account the importance of representing spatial dependence through an independent variable. This stage followed the process proposed in previous study (Assirati, 2018).

Based on the geostatistical precepts, the procedure was established to measure the probability that an individual belongs to a particular category (analyzed variable) according to the values of their neighbors (regional character). The approach aims to complement the traditional parametric analysis by including spatial covariates on discrete choice models (Binomial Logit models in this study).

In this article, the spatial covariates are the probabilities of the neighbors (determined from the four neighborhoods) to choose the two categories of travel mode adopted in this study: bus or car.

To calculate the probabilities, the first step consists of establishing the number of neighborhoods. We suggest measuring the maximum distance between the two points most distant from each other and divide this distance into equal portions to determine *n* neighborhoods of interest, which should not overlap.

Once neighborhoods are established, probabilities of belonging to the categories are calculated by individuals, given the categories of their neighbors. For each individual and neighborhood, the number of elements included in a region belonging to each category is counted. The probability of this individual belonging to each of the categories is the ratio between the number of elements in a given category and the total number of elements found in the considered neighborhood. Figure 5 shows an example of the procedure. If one desires to characterize an individual, located at point X (Figure 5), according to the probabilities of belonging to a certain category based on neighborhoods, the procedure should be as follows:

a) Given the maximum dimensions of the data distribution, space is divided into a number *n* of neighborhoods of interest. For the example in Figure 5, the data are arranged in a quarter circumference with a radius of size 4L. Thus, it is possible to delimit *n* = 4 bands of size L each. Each range will, therefore, be a neighborhood for analysis.

b) The number of elements from all categories present in the first neighborhood is counted. For example, in the first strip, there are four red elements of category A and no black elements of category B. Therefore, an individual located in X has the following according to the first neighborhood (*n* = 1):

- P_1 (A) = 1 (4 elements out of a total of 4) of being category A; and

- P_1 (B) = 0 (0 elements out of a total of 4) of being category B.

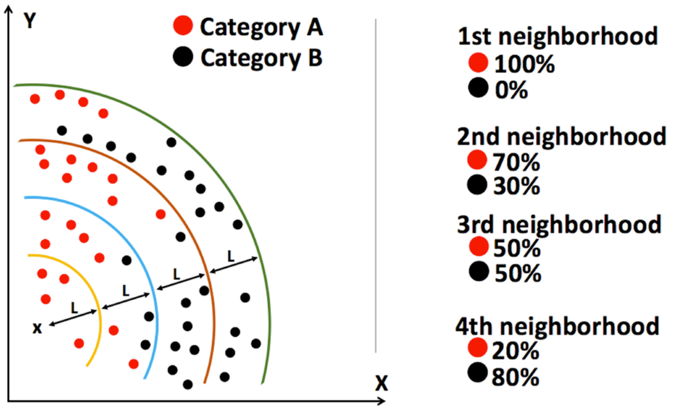


Figure 5: Example of calculating probabilities by category A for point X, according to four neighborhoods of interest.

c) It works similarly for the other neighborhoods (*n* = 2, 3, and 4). Thus, individual X is characterized as:

- P_2 (A) = 0.7 (7 elements out of 10) of being category A;

- P_2 (B) = 0.3 (3 elements out of a total of 10) of being category B;

- P_3 (A) = 0.5 (8 elements out of a total of 16) of being category A;

- P_3 (B) = 0.5 (8 elements out of a total of 16) of being category B;

- P_4 (A) = 0.2 (4 elements out of a total of 20) of being category A; and

- P_4 (B) = 0.8 (16 elements out of a total of 20) of being category B.

d) The previous procedures are repeated for all points in the sample.

Thus, once the spatial dependence is proven in the ESDA stage and the global spatial association indicators are calculated, the spatial covariates are proposed, based on the neighborhoods considered statistically significant in the ESDA stage.

4.3.2 Binomial Logit Calibration (non-spatial model)

This stage corresponds to the calibration of the non-spatial discrete choice model. The determination of utility functions, in this step, is based on the variables that represent the travel mode alternatives (bus travel time, car travel time, bus fare and car travel costs), the trip (travel distance), as well as individual variables (income). The response variable is the binary mode choice (0) bus and (1) car. Equations 5 and 6 represent the utility functions of the non-spatial model.

$$V_0 = \beta_{0\_dist} * dist + \beta_{0\_inc} * income + \beta_{time} * time\_bus + \beta_{cost} * cost\_bus \qquad (5)$$

$$V_1 = CTE + \beta_{1\_dist} * dist + \beta_{1\_inc} * income + \beta_{time} * time\_car + \beta_{cost} * cost\_auto \qquad (6)$$

The variables "income" and "distance" were included in both utility functions V_0 and V_1, although they did not vary between alternatives (only between individuals). Therefore, to guarantee the assumption of distinction between alternatives, it was necessary to have specific parameters $\beta_{0\_dist}.\,\beta_{0\_inc}$ and $\beta_{1\_dist}.\,\beta_{1\_inc}$. for functions V_0 and V_1, respectively. The parameters β_{time} and β_{cost} are generic since the assumption of distinction is guaranteed.

4.3.3 Binomial Logit Calibration (spatial model)

If the ESDA shows that there is a spatial dependence on the concerned variable (travel mode choice), it is expected that the inclusion of spatial information increases the potential for accuracy and quality of the parametric modeling. Then, the calculation of the probabilities of belonging to the categories of the travel mode is applied based on the neighborhood values.

The establishment of probabilities, by category, in each of the neighborhoods provides the informational addition of a regionalized character that was sought to increase the parametric analyses of the discrete choice. Previously, four travel mode attributes were taken into account (bus travel time, car travel time, bus fare, and car travel costs), an individual attribute (income), and a travel attribute (trip distance).

Now, in addition to these, there are eight new attributes of individuals: PI (0); PI (1); PII (0); PII (1); PIII (0); PIII (1); PIV (0); PIV (1), which are, respectively, probability P of an individual choosing to travel by bus (0) from the values of the elements in neighborhood I, probability P of an individual choosing to travel by car (1) from the values of the elements in neighborhood I, and so on, alternating probabilities P of the mode choice until neighborhood IV.

It should be noted that using all neighborhoods is an ideal scenario where they were all identified as statistically significant by the global indicators. The existence of covariates is, therefore, conditioned to the result obtained in the exploratory analysis. Only the spatial covariates related to the regions where at least one indicator pointed to the significance (p-value ≤ 0.05) should be included in the parametric equation. Otherwise, the region (neighborhood) must be dismissed. Locational informational addition now defines new utility functions, represented by Equations 7 and 8.

$$V_0 = \beta_{0\_dist} * dist + \beta_{0\_inc} * income + \beta_{time} * time\_bus + \beta_{cost} * cost\_bus + \beta_{0\_vI} * PI(0) + \beta_{0\_vII} * PII(0) + \beta_{0\_vIII} * PIII(0) + \beta_{0\_vIV} * PIV(0) \quad (7)$$

$$V_1 = CTE + \beta_{1\_dist} * dist + \beta_{1\_inc} * income + \beta_{time} * ttime\_car + \beta_{cost} * cost\_car + \beta_{1\_vI} * PI(1) + \beta_{1\_vII} * PII(1) + \beta_{1\_vIII} * PIII(1) + \beta_{1\_vIV} * PIV(1) \quad (8)$$

4.3.4 Comparison of the two models

For the comparison of the Binomial Logit models (spatial and non-spatial), three metrics will be used: The **adjusted rho-square** values; **the Akaike** criterion value; **the hit rate by cross-validation** and **likelihood values** for validation sample. Such metrics are well known, however, the reader who seeks more details can consult Hosmer and Lemeshow (2000) and Assirati (2018). Equation 9 defines the adjusted rho-square metric. If the models had the same number of parameters, the rho-square metric would suffice. However, we intended to compare models with "original" variables *versus* models that add spatial covariates to "original" variables. The models present a different number of parameters. They, therefore, require evaluation by the adjusted rho-square metric as this allows us to compare models estimated from the same sample of observations, but with a different number of parameters.

$$\rho_*^2 = 1 - \frac{L^* - K}{L_0} \quad (9)$$

Where L_0 is the likelihood value obtained by assuming all parameters beta of the model as zero and L^* is the maximum likelihood value obtained when the parameters β correspond to the estimated values. K is the number of estimated parameters.

The Akaike criterion is defined by Equation 10. The formulation causes the criterion to penalize overfitting (the act of adding too many variables to the equations to obtain better adjustments).

$$A = 2K - 2\ln L^* \quad (10)$$

Lastly, cross-validation consists of segregating a portion of the sample to be used in the parameter estimation process and another part used in the validation process. Then, the modeling is applied along with the calibrated parameters coming from the calibration group in the elements of the validation group. Thus, obtaining estimated values for that group. As the values of the validation group are known, the hit rate can be measured when comparing the real values with the estimated values. Accurate models tend to have high hit rates, and the quality of this value can be measured by calculating the likelihood (Equation 11).

$$L = p^y.(1-p)^{(n-y)} \quad (11)$$

For the total number *n* of elements considered in the test, there is a quantity *y* of correctly evaluated elements. Thus, the ratio p = *ny* between the number of elements is successfully assessed, and the total of elements considered is defined. When the value of *p* tends to unity (100% accuracy), the value of L also tends to unity, and consequently, log L tends to the null value.

5. RESULTS AND DISCUSSION

5.1 Step 1: Synthetic spatially correlated data

After applying the bilinear interpolation technique, we obtained new locations from values for the seven variables involved. This technique allowed the change from 60 starting points to 580 new points, and now the x-axis, set at the original range of 0 to 10, has a step of 0.25 and the y-axis, set at the original limits of 0 to 5, now has a step of 0.05. Figure 6 illustrates the geographical distribution of the binary dependent variable. Table 2 presents the descriptive measures of the simulated data sample variables. The following results were obtained (Table 3) from non-parametric statistical tests. These results corroborate the efficiency of the simulation, attesting the similarity of population distribution between the two samples (initial and synthetic spatially correlated data) for all seven variables under analysis.

Table 2: Descriptive measures of the Synthetic spatially correlated data.

	Travel distance (Km)	Bus travel time (Min.)	Auto travel time (Min.)	Bus fare (Currency Unit)	Car travel cost (Currency Unit)	Income indicator (Scale 1 – 10)
Mean	2.47	5.40	3.70	1.99	2.10	5.43
Std deviation	1.31	2.58	1.97	0.70	1.11	1.47
Max Value	5.52	11.12	8.29	3.00	4.70	10.00
Min Value	0.05	1.08	0.07	1.50	0.04	1.00

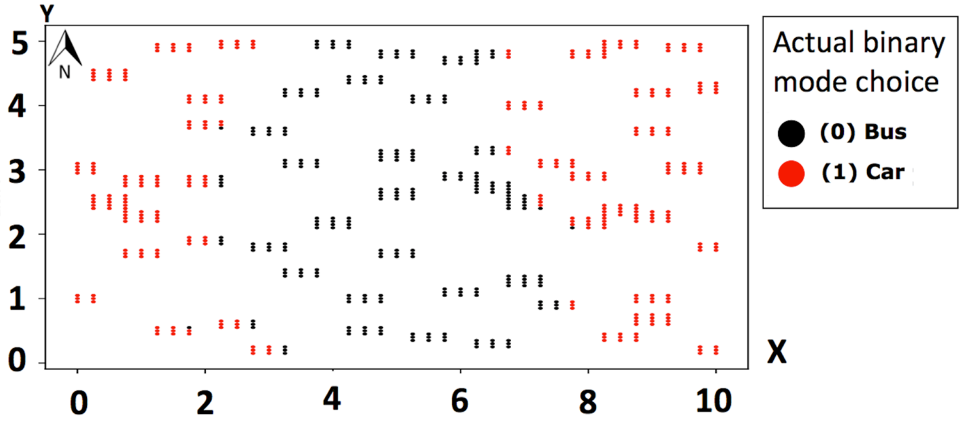


Figure 6: Spatial representation of binary mode choice in the Synthetic spatially correlated data.

Table 3: Results of the nonparametric hypothesis tests (95% of significance level).

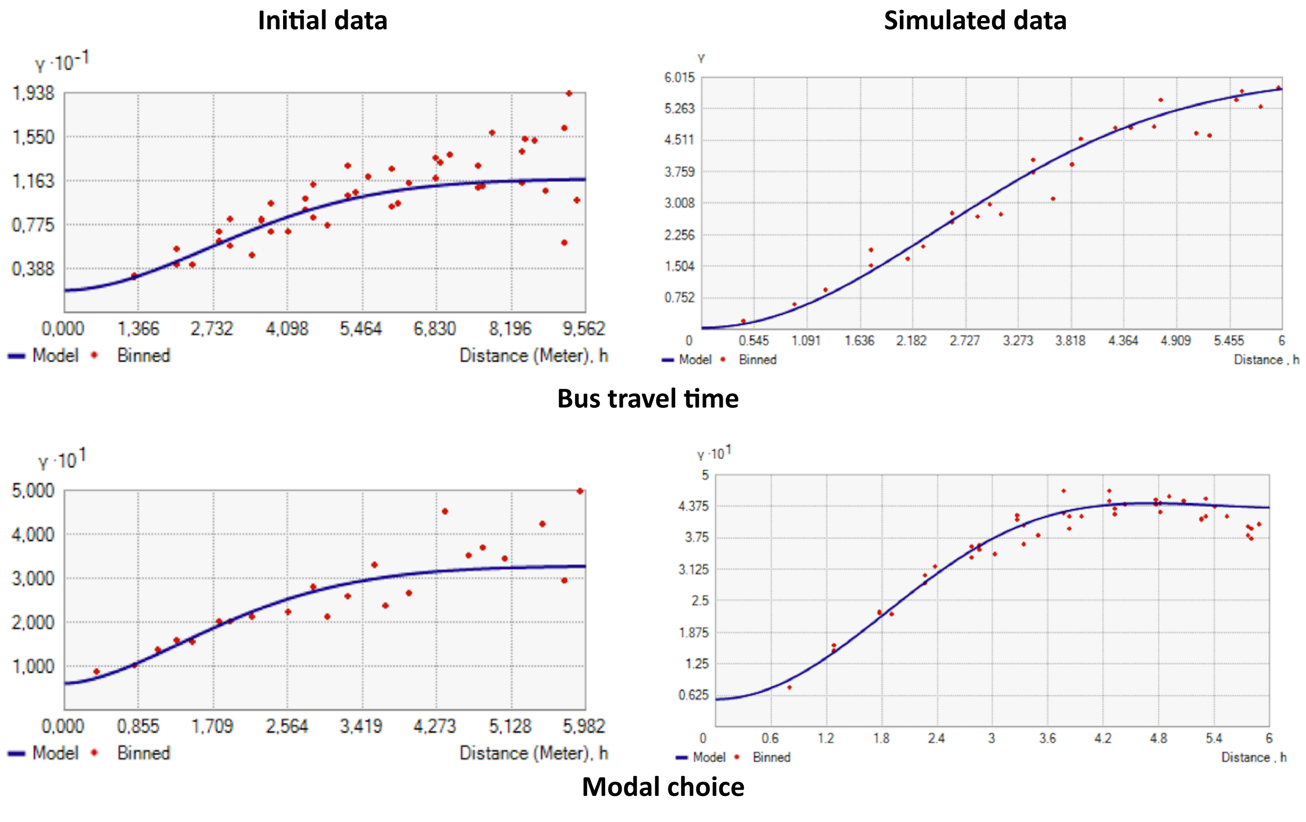
Comparison (Original vs simulated data)	Mann-Whitney		Kolmogorov–Smirnov test	
	Sig.	Null hypothesis	Sig.	Null hypothesis
Travel distance	0.407	Not rejected	0.501	Not rejected
Bus travel time	0.201	Not rejected	0.403	Not rejected
Auto travel time	0.100	Not rejected	0.301	Not rejected
Bus fare	0.08	Not rejected	0.100	Not rejected
Auto travel costs	0.490	Not rejected	0.401	Not rejected
Income	0.290	Not rejected	0.503	Not rejected
Mode choice	0.340	Not rejected	0.508	Not rejected

Null hypothesis: The samples have similar distribution; Alternative hypothesis: The samples have different population distribution.

Finally, the modeling of experimental variograms, a fundamental step for investigating the spatial structure of the variables before and after the simulation of the data, is illustrated in Figure 7. The figure shows examples of adjusted semivariograms for two concerned variables. The parameters of the theoretical variograms (for the seven variables) are shown in Table 4. Similar parameters can be appreciated regarding the theoretical variograms of the initial and simulated data, confirming the similarity of the spatial structure between the two samples.

Table 4: Variogram parameters for the initial and simulated data.

	Main direction (°)		**Tolerance (°)**		**LAG distance (m)**		**C0 [Nugget Effect] (m)**		**C [Contribution] (m)**		**Range (m)**		**Sill [C0 + C] (m)**	
	O	**S**	**O**	**S**	**O**	**S**	**O**	**S**	**O**	**S**	**O**	**S**	**O**	**S**
Travel Distance	262.6°	263 °	17 °	14 °	0.7	0.7	0.55	0	1.92	2.77	7.25	7	2.47	2.77
Income	306.6°	307.1 °	16.5 °	23.2 °	0.7	0.7	3.02	0	2.64	6	6.15	6	5.67	6
Bus travel time	262.7 °	262.6 °	30.6 °	12.9 °	0.8	0.8	1.96	0	7.88	14.09	7.14	10.6	9.84	14.09
Auto travel time	262.4 °	263.5 °	16.7 °	15.1 °	0.8	0.8	1.24	0	4.34	5.59	7.14	5.82	5.58	5.59
Mode Choice	32°	95°	22.8 °	22 °	0.5	0.5	0.06	0.05	0.2	0.325	4.28	6	0.26	0.38



Bus travel time

Modal choice

Figure 7: Variograms for the initial and simulated data.

5.2 Step 2: Exploratory Spatial Data Analysis (ESDA)

Initially, an Exploratory Spatial Data Analysis was performed, using the synthetic spatially correlated data, based on the calculation of the Moran and G-SIVAR indicators, shown in Table 5. It is worth noting here that the neighborhoods chosen were those where the indicators were considered statistically significant by hypothesis tests, associated with at least one of the indicators. The null hypothesis, for the hypothesis tests of both indicators, is spatial randomness. In this case, the neighborhoods of interest are those where the null hypothesis was rejected in at least one of the indicators (Neighborhoods I, II, and III).

Table 5: Result of Moran & G-SIVAR indicators applied to the synthetic spatially correlated data.

Neighborhood	Moran				G-SIVAR			
	Index	Z-Value	p-Value	Null hypothesis	Index	Z-Value	p-Value	Null hypothesis
I	0.269	65.119	0	Reject	0.954	-7.867	0	Reject
II	-0.064	-43.602	0	Reject	1	0	0.6	Not Reject
III	-0.01	-30.338	0	Reject	1	0	0.6	Not Reject
IV	-0.002	0	1	Not Reject	1	0	0.6	Not Reject

It can be seen, from Table 5, that the indicators are almost complementary quantities. The proximity of value 1, for the Moran index, indicates high spatial autocorrelation, while value 1 indicates very low spatial autocorrelation for the G-SIVAR indicator. Note that the values for the statistically significant coefficients are not high (for Moran) or low (for G-SIVAR). This indicates that, even if small, there is some spatial autocorrelation in these locations and it is statistically significant. Thus, the inclusion of covariates extracted from these neighborhoods will provide informational addition to the Binomial Logit model.

5.3 Step 3: Confirmatory Spatial Data Analysis (CSDA)

Subsequently, the probabilities of neighbors who belong to neighborhoods I, II, and III to choose the bus (category 0) or car (category 1) were calculated. It is reinforced that these spatial covariates are individual. Each individual has six probability values from its neighbors: P (1) I; P (1) II; P (1) III; P (0) I; P (0) II; P (0) III.

The next step, related to the calibration of the utility functions of the original variables, was based on Equations 5 and 6. When performing the calibration using the maximum likelihood method, the estimated beta parameters were obtained. However, not all were statistically significant. Therefore, those variables associated with non-significant parameters (in both Equations 5 and 6) were removed from the original utility functions. Consequently, it is necessary to recalibrate the model, dismissing the variables that are not relevant to the study. The results are shown in Table 6a. When segregating the sample in a 7/3 proportion and promoting cross-validation tests with the calibrated coefficients, 74% of hit rates was obtained for the validation sample, with a likelihood value L= 2.26 × 10−44 and log (L) = −100. 49.

Table 6: Binomial Logit Calibration a) non-spatial model) & b) spatial model

	(a)				**(b)**		
Parameter	**Value**	**Error**	**p-Value**	**Parameter**	**Value**	**Error**	**p-Value**
β_{0_dist}	-9.46	1.07	0.03	β_{0_dist}	-19.90	1.80	0.04
β_{1_dist}	9.46	1.37	0.05	β_{0_vI}	13.00	1.41	0.01
β_{cost}	-20.0	0.84	0.00	β_{0_vII}	9.10	1.60	0.02
β_{time}	20.8	1.33	0.00	β_{0_vIII}	8.66	2.71	0.02
CTE	-7,87	1.05	0.00	β_{1dist}	19.90	1.35	0.05
				β_{1_vI}	14.10	5.91	0.03
				β_{1_vII}	7.66	4.67	0.03
				β_{cost}	-47.1	3.20	0.00
				β_{time}	56.1	4.10	0.00
				CTE	1.41	1.66	0.00
Adjusted rho-square	0.738			Adjusted rho-square	0.904		
Akaike criterion	210.455			Akaike criterion	77.323		

Then, the calibration of the spatial model is performed by incorporating the spatial covariates related to neighborhoods I, II and III, per individual. As the hypothesis tests did not show the significance for region IV, the variables related to that region and their respective parameters were taken from original equations. Once more, as in the first modeling, after calibration, there was a non-significant coefficient (coefficient relative to the third neighborhood for V_1) demanding the elimination of its variable from the equation and a new calibration. Therefore, a new calibration is carried out, ignoring the non-significant coefficients and their associated variables. The results are shown in Table 6b. Once more, by segregating the sample in a 7/3 ratio for cross-validation tests with the calibrated coefficients, an 89% hit rate was obtained for the second modeling, with likelihood value L = 1.82 × 10−39 and log (L) = −79. 98.

When comparing the first model, which considered the original significant variables of the database (Table 6a) with the second model that considered the original significant variables and the spatial covariates (Table 6b), one can corroborate the hypothesis that if there is spatial association of variables through ESDA, the spatial aspect can be included in the model as it will bring greater precision to the results.

The informational increase is positive, providing, for this case study, an increase of 15% in the cross-validation hit rates. The values that measure the quality of the model were also positive: adjusted rho-square increased, approaching the unit considerably. The Akaike criterion decreased from 210.455 to 77.323. The importance of such a reduction is emphasized since the Akaike criterion penalizes overfitting, and it is for this reason that lower values are sought for this criterion. The second model is shown to be compatible because, despite having five coefficients more than the first, it presents the Akaike criterion approximately 2.72 times smaller.

6. CONCLUSIONS AND RESEARCH CONTRIBUTIONS

This paper aimed to model the spatial effect on travel mode choice using a synthetic spatially correlated data set. This research presented three major academic contributions mentioned previously. The first one is related to the methodological procedure to model the spatial effect on travel mode choice. The authors presented a sequential spatial method for forecasting mode choice. The procedure is based on an Exploratory Spatial Data Analysis, followed by a Confirmatory Spatial Data Analysis, regarding discrete variables. This procedure could be easily replicable in different study fields regarding correlated spatial data. Additionally, the methodological steps for G-SIVAR calculation are available in an R - code for any interested user (https://github.com/pedreirajr/g-sivar).

The proposal of spatial covariates, based on geostatistical assumptions, is an essential contribution considering the importance of representing spatial effects through an independent variable. In this article, the spatial covariate represents the probability of choosing some alternative based on the choice of different neighbors. The neighborhoods are delimited based on lag distance concepts. Additionally, it is essential to mention that the spatial covariate proposal is dependent on the ESDA stage. By observing the hypothesis tests, associated with the indicators, the neighborhoods to be added to the analysis were determined.

It is also worth mentioning the role of geostatistics in the methodological procedure. The geostatistics tool was fundamental for the composition of the G-SIVAR, to determine the neighborhoods (based on the values of the lags/distances between observations of the variograms), and, indirectly, on the establishment of spatial covariates (in determining the angle and distances from neighbors - based on variograms). Additionally, the adjusted of experimental variograms was applied to attest similarities of spatial structure considering original and simulated samples.

The suggestion of a simple procedure to propose a simulation of a spatially correlated database is the last research contribution, taking into account that, despite the observed advances in population synthesizers for travel data microsimulation, none of the approaches recognizes the spatial correlation of data as an extraordinary input to reproduce travel behavior. Although there is a recent application of Sequential Gaussian Simulation to simulate travel demand data (Lindner and Pitombo, 2019), the procedure proposed in this article requires less conceptual effort.

Finally, the main contribution of this paper is related to a methodological demonstration on how to take into account the spatial characteristics of the data, along with the usual variables. Assirati (2018) applicated this methodology considering a real data and a trip-chaining study case regarding panel data collected by smartphones. This application corroborates that the methodological procedure is feasible to real data, with different urban configuration.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of CNPq (National Council for Scientific and Technological Development, Brazil), Grant #304345/2019-9 and Grant #427525/2018-7).

REFERENCES

Ahern. A. A. e N. Tapley (2008). The use of stated preference techniques to model modal choices on interurban trips in ireland. Transportation Research Part A: Policy and Practice 42(1). 15–27.

Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93-115.

Anselin, L. (1996). Interactive techniques and exploratory spatial data analysis.

Assirati, L. (2018). Análise da influência da vizinhança no comportamento individual relativo a viagens através de dados em painel (Doctoral dissertation, Universidade de São Paulo).

Ben-Akiva, M. E. (1974). Structure of passenger travel demand models (Doctoral dissertation, Massachusetts Institute of Technology).

Bhat, C. R. (1995). A heteroscedastic extreme value model of intercity travel mode choice. Transportation Research Part B: Methodological, 29(6), 471-483.

Bhat. C. R.. N. Eluru. e R. B. Copperman (2008). Flexible model structures for discrete choice analysis. Handbook of Transport Modelling. 5. 75–104.

Chow, L. F., Zhao, F., Liu, X., Li, M. T., & Ubaka, I. (2006). Transit ridership model based on geographically weighted regression. Transportation Research Record, 1972(1), 105-114.

Cressie, N. A. (1993). Statistics for spatial data. John Willy and Sons. Inc., New York.

Dugundji, E. R., & Walker, J. L. (2005). Discrete choice with social and spatial network interdependencies: an empirical example using mixed generalized extreme value models with field and panel effects. Transportation Research Record, 1921(1), 70-78.

Fleming, M. M. (2004). Techniques for estimating spatially dependent discrete choice models. In Advances in spatial econometrics (pp. 145-168). Springer, Berlin, Heidelberg.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. The incorporated statistician, 5(3), 115-146.

Getis. A. e J. K. Ord (1992). The analysis of spatial association by use of distance statistics. Geographical analysis 24(3). 189–206.

Hosmer, D.W. & Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.). Wiley. ISBN 978-0-471-35632-5

Hess. S. (2005). Advanced discrete choice models with applications to transport demand. Ph. D. thesis. University of London.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy, 52(6), 119-139.

Le Gallo, J., & Ertur, C. (2003). Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980–1995. Papers in regional science, 82(2), 175-201.

Lindner, A., & Pitombo, C. S. (2018). A conjoint approach of spatial statistics and a traditional method for travel mode choice issues. Journal of Geovisualization and Spatial Analysis, 2(1), 1.

Lindner, A., & Pitombo, C. S. (2019). Sequential Gaussian Simulation as a promising tool in travel demand modeling. Journal of Geovisualization and Spatial Analysis, 3(2), 15.

Lindner, A., Pitombo, C. S., Assirati, L., Junior, J. U. P., & Salgueiro, A. R. (2021). Estimation of Travel Mode Choice Using Geostatistics: a Brazilian Case Study. Revista Brasileira de Cartografia, 73(1), 182-197.

Lopes, S. B., Brondino, N. C. M., & Rodrigues da Silva, A. N. (2014). GIS-based analytical tools for transport planning: Spatial regression models for transportation demand forecast. ISPRS International Journal of Geo-Information, 3(2), 565-583.

Matheron, G. (1971). The theory of regionalized variables and its applications (Vol. 5). Paris: École National Supérieure des Mines, 211.

McFadden, D. (1974). The measurement of urban travel demand. Journal of public economics, 3(4), 303-328.

Messner, S. F., Anselin, L., Baller, R. D., Hawkins, D. F., Deane, G., & Tolnay, S. E. (1999). The spatial patterning of county homicide rates: An application of exploratory spatial data analysis. Journal of Quantitative criminology, 15(4), 423-450.

Miyamoto. K. V. Vichiensan. N. Shimomura. e A. Páez (2004). Discrete choice model with structuralized spatial effects for location analysis. Transportation research record: journal of the transportation research board (1898). 183–190.

Moran. P. A. (1950). Notes on continuous stochastic phenomena. Biometrika 37(1/2). 17–23.

Naizer, C. C. B. R., Rodrigues, D. S., Pedreira Junior, J. U., & Pitombo, C. S. (2019). G-SIVAR: A Global SpatIal IndIcator based on VAriogram. Boletim de Ciências Geodésicas, 25(4).

Nowrouzian, R., & Srinivasan, S. (2014). A Spatial Quasi-Poisson Model for Car Ownership (No. 14-5382).

Páez. A.. F. A. López. M. Ruiz. e C. Morency (2013). Development of an indicator to assess the spatial fit of discrete choice models. Transportation Research Part B: Methodological 56. 217–233.

Pitombo. C. S.. A. R. Salgueiro. A. S. G. da Costa. e C. A. Isler (2015). A two-step method for mode choice estimation with socioeconomic and spatial information. Spatial Statistics 11. 45–64.

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 (pp. 1-23).

Qin. H.. J. Gao. H. Guan. e H. Chi (2017). Estimating heterogeneity of car travelers on mode shifting behavior based on discrete choice models. Transportation Planning and Technology 40(8). 914–927.

Sener, I. N., Pendyala, R. M., & Bhat, C. R. (2011). Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior. Journal of Transport Geography, 19(2), 294-303.

Unwin, A., & Unwin, D. (1998). Exploratory spatial data analysis with local statistics. Journal of the Royal Statistical Society. Series D (The Statistician), 47(3), 415-421.

Wang, F. (2001). Explaining intraurban variations of commuting by job proximity and workers' characteristics. Environment and Planning B: Planning and Design, 28(2), 169-182.

Zhou. J. (2012). Sustainable commute in a car-dominant city: Factors affecting alternative mode choices among university students. Transportation research part A: policy and practice 46(7). 1013–1029.