Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogr
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

Estimation of Travel Mode Choice Using Geostatistics: a Brazilian Case Study

Estimativa da Escolha do Modo de Viagem Utilizando Geoestatística: um Estudo de Caso Brasileiro

Anabele Lindner [1], Cira Souza Pitombo [2], Lucas Assirati [3], Jorge Ubirajara Pedreira Junior [4] e Ana Rita Salgueiro [5]

1 São Carlos School of Engineering - University of São Paulo (EESC-USP), Department of Transportation Engineering, São Carlos-SP, Brazil. anabele@alumni.usp.br
ORCID: https://orcid.org/0000-0002-4487-3650
2 São Carlos School of Engineering - University of São Paulo (EESC-USP), Department of Transportation Engineering, São Carlos-SP, Brazil. cirapitombo@usp.br.
ORCID: https://orcid.org/0000-0001-9864-3175
3 São Carlos School of Engineering - University of São Paulo (EESC-USP), Department of Transportation Engineering, São Carlos-SP, Brazil. assirati@usp.br.
ORCID: https://orcid.org/0000-0002-0118-2665
4 São Carlos School of Engineering - University of São Paulo (EESC-USP), Department of Transportation Engineering, São Carlos-SP, Brazil. jorge.ubirajara@usp.br.
ORCID: https://orcid.org/0000-0002-8243-5395
5 Federal University of Ceará (UFC), Department of Geology, Fortaleza-CE, Brazil. geo.ritasalgueiro@gmail.com.
ORCID: https://orcid.org/0000-0003-3050-4157

Recebido: 04.2020 | Aceito: 09.2020

Abstract: Traditional methods for travel demand estimation are often built on socioeconomic and travel information. The information required to conduct such studies is costly and rarely available in developing countries. Besides, some conventional methods do not consider the spatial relationship of variables and, in general, a large amount of socioeconomic and individual travel data is required. The key aim of this paper is to evaluate the importance of considering spatial information when estimating travel mode choices especially considering the lack of available data. The study area is the São Paulo Metropolitan Area (Brazil) and the dataset refers to an Origin-Destination Survey, conducted in 2007. This research paper analyzes the use of Geostatistics when estimating discrete travel mode choices. The results demonstrated a satisfactory outcome for the geostatistical approach. Finally, although socioeconomic and travel variables have greater explanatory power in predicting travel mode choices, spatial factors contribute to better understand the travel behavior and to provide further information when estimating spatially correlated data.
Keywords: Geostatistics. Indicator Kriging. Travel Mode Choice. Lack of Data.

Resumo: Métodos tradicionais para a estimativa da demanda por transportes são frequentemente delineados considerando informações socioeconômicas e de viagens. As informações para conduzir tais estudos são dispendiosas e raramente disponíveis em países em desenvolvimento. Além disso, métodos convencionais não consideram as relações espaciais das variáveis e, em geral, uma quantidade considerável de dados individuais socioeconômicos e de viagem são requeridos. O objetivo principal deste trabalho é avaliar a importância de considerar informações espaciais no processo de estimativa da escolha do modo de viagem, considerando especialmente a ausência de dados disponíveis. A área de estudo é a Região Metropolitana de São Paulo (Brasil) e o conjunto de dados refere-se à Pesquisa Origem-Destino, conduzida em 2007. Esta pesquisa analisa o uso da Geoestatística na estimativa das escolhas modais. Os resultados provenientes desta abordagem geoestatística se mostraram satisfatórios. Por fim, apesar de as variáveis socioeconômicas e de viagem terem um maior poder explicativo na predição das escolhas modais, fatores espaciais contribuem para um melhor entendimento do comportamento individual relativo a viagens e fornecem informações adicionais no processo de estimativa de dados espacialmente correlacionados.
Palavras-chave: Geoestatística. Krigagem Indicadora. Escolha Modal. Ausência de Dados.

1 INTRODUCTION

Travel demand modeling is an essential tool for transportation planning decisions. It is based upon explanatory factors, such as individual characteristics, trips, urban environment and its facilities (ORTÚZAR; WILLUMSEN, 2011).

This paper focuses on estimating travel mode choice, which is often determined by a behavioral approach. The logit model (and its variations) is one of the most used methods, in which data is obtained from a discrete choice experiment. This model was first explored by Ben-Akiva (1974), McFadden (1974) and Domencich and McFadden (1975).

The traditional application of logit models usually demands a large amount of disaggregated data and is achieved by using independent variables associated to socioeconomic attributes and travel data. However, the spatial distribution of variables is not often taken into account in traditional travel demand modeling.

More recently, renowned authors in the area of travel demand, such as Ben-Akiva, Ramming and Bekhor (2004) and Bhat and Zhao (2002), improved traditional models by considering spatial effects, using kernel estimation and robust logit models. In the field of discrete choice, Bhat and Sener (2009) and Páez et al. (2013) introduced spatial information for logit models.

Some authors have achieved promising results by using spatial information in travel mode choice analysis. Dugundji and Walker (2005) explored empirical estimation of discrete choice models with the effects of social and spatial feedback using the Biogeme toolkit (Bierlaire, 2003). Miyamoto et al. (2004) combined concepts to specify the spatially autocorrelated deterministic and error components within a mixed logit framework. The results endorsed the effectiveness of the proposed method, as well as the advantages as it is an alternative to more time consuming and expensive methods that use a larger number of explanatory variables.

Páez and Scott (2005) indicated some major analytical issues for studying spatial processes in urban analysis, such as spatial association, heterogeneity and the Modifiable Areal Unit Problem (MAUP). They claim that the spatial association can be addressed by spatially autoregressive models and local statistics. In order to deal with heterogeneity, models such as switching regressions, multilevel models and geographically weighted regression can be considered. The MAUP, which refers to aggregate spatial data represented by centroids, can be addressed by strategies based on the Thiessen polygon approach or by the optimization of zoning systems (PÁEZ; SCOTT, 2005).

Therefore, studies found in the literature support the fact that using spatial information complements traditional models, given that travel behavior is also strongly associated with spatial distribution of activities in the urban environment (CERVERO; RADISCH, 1996; KITAMURA; MOKHTARIAN; LAIDET, 1997).

Despite the solid results seen in traditional models combined to spatial approaches, these studies are often limited to either an exploratory analysis – in the cases of kernel estimation and K function (YAMADA; THILL, 2004; DUGUNDJI; WALKER, 2005; XIE; YAN, 2013; KAYGISIZ et al., 2015) – or a point-limited estimate – in spatial regression, autocorrelated models and a mixed logit model based on spatial factors (BHAT; ZHAO, 2002; BEN-AKIVA; RAMMING; BEKHOR, 2004; MIYAMOTO et al., 2004; BHAT; SENER, 2009).

On the other hand, Geostatistics is a potential alternative technique in the field of Transportation Engineering (CIUFFO; PUNZO; QUAGLIETTA, 2011; MAZZELLA; PIRAS; PINNA, 2011; ZOU et al., 2012; PITOMBO; COSTA; SALGUEIRO, 2015a; PITOMBO et al., 2015b; LINDNER et al., 2016; LINDNER; PITOMBO, 2018; LINDNER; PITOMBO, 2019; MARQUES; PITOMBO; 2020; KLAKTO; USMAN; MATTHEW, 2017). Despite the need for further in-depth investigations to such application, Geostatistics may fill in the gaps found in traditional and spatial methods as it allows for spatial interpolation of values and errors. Owing to this, Geostatistics can be employed as a confirmatory analysis.

Some geostatistical approaches have also been successfully applied in travel demand estimation using datasets other than Origin-Destination (O/D) surveys. Rocha et al. (2017) estimated transportation planning variables, in transit trip production, by proposing a geostatistical procedure that combined semivariogram deconvolution and Kriging with External Drift (KED). The method consists of initially assuming a disaggregated systematic sample from aggregate data and subsequently applying KED, considering the

population (census microdata) as a secondary input. Lindner (2019) in turn introduced a heuristic framework to disaggregate data using as input to the procedure information with high availability, such as census microdata.

Geostatistics can also map a phenomenon using observed values and spatial locations of a variable. This feature is an important advantage when compared to traditional models as geostatistical approaches may provide satisfactory results towards transportation planning with less expensive data sources.

Conventional tools for estimating travel demand require socioeconomic and travel data associated to individuals' choices, often collected in household surveys for urban transportation. Nonetheless, in developing countries, e.g. Brazil, such information is rarely available due to the high costs of data collection. Therefore, a method that alternatively uses spatial factors (producing a spatial covariance matrix) for travel mode choice issues can potentially overcome this pitfall.

For spatially correlated data cases, travel mode choice is a behavioral concern and depends on three main factors: (1) individual, (2) travel and (3) location. In this paper, the spatial location is investigated and its importance is evaluated, especially considering the lack of travel and individual information. A comparison to a logit model that uses individual information is provided to assess the spatial approach contribution. Despite this, we highlight that the geostatistical technique does not replace the traditional method since the spatial autocorrelation is only a single factor among several others to estimate travel mode choice. Therefore, our goal is to demonstrate that a single travel demand variable associated to its spatial position can yield satisfactory results for a straightforward estimation in a disaggregated modeling approach. This paper is organized into six sections, including this introduction (Section 1). Section 2 describes some main concepts related to the Geostatistical approach. Section 3 presents the materials (study area and dataset) and the method. Section 4 provides the results, and finally Section 5 draws the main conclusions.

2 GEOSTATISTICS

Geostatistics is a spatial analysis technique that has exploratory and confirmatory features (JOURNEL, 1989). It also enables data visualization to verify existing spatial patterns. Furthermore, a set of estimation models and procedures are employed for its validation. In the approach adopted in the current research, this technique is advantageous as unknown values for non-sampled locations can be inferred.

Geostatistics consists of the following steps: 1) Variographic analysis or spatial covariance, 2) Kriging and 3) Validation (ISAAKS; SRIVASTAVA,1989). The referred sequence may present a reverse order between Kriging and validation in case the cross validation is used.

The variographic analysis is mainly related to the study of spatial structures. It is sequentially performed by detecting Regionalized Variables, calculating empirical semivariograms, adjusting theoretical models and defining the continuity direction (MATHERON, 1963).

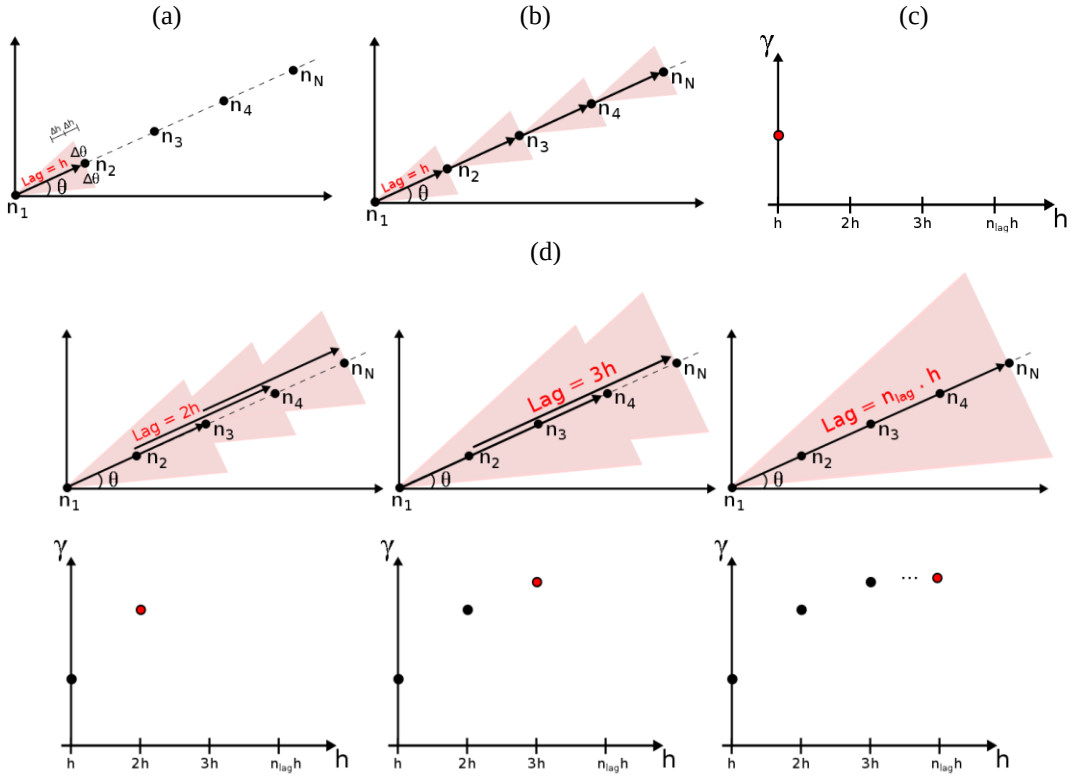
A Regionalized Variable is a spatially distributed variable with spatial structure and a random component (MATHERON, 1971). The spatial structure can be modeled using semivariograms, which suggest the spatial covariance of a Regionalized Variable. The average of the covariances between pairs of observations determines the semivariogram function, according to Eq. (1), as originally defined by Matheron (1963).

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_{i+h})]^2 \quad (1)$$

In which $N(h)$ is the set of all pairwise data values $Z(x_i)$ and $Z(x_{i+h})$ at the spatial locations i and $i+h$, respectively.

From Eq. (1), input parameters are set to construct the graphical semivariogram. These input parameters are: cut distance, lag distance, lag tolerance and angle direction. The calculation of empirical variograms is illustrated in Figure 1, according to the definition defined by Matheron (1963, 1971) and Wackernagel (2003).

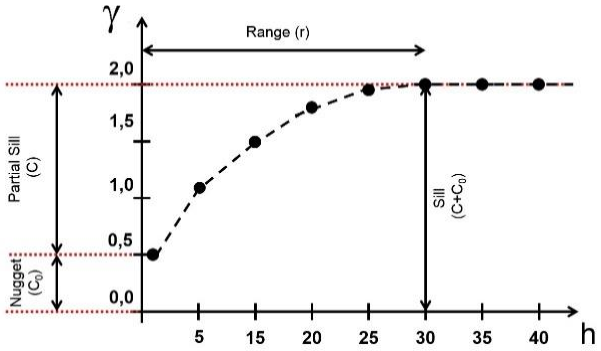
Figure 1 - Basic steps for calculating an empirical semivariogram. Conventions: n_i is the i^{th} observation, h is the predefined lag, Δh is the lag tolerance, θ is the direction of the variographic analysis and $\Delta\theta$ is the angle tolerance. Legend: (a) Starting from n_1. For each observation distanced $h+\Delta h$ and with an angle between $\theta+\Delta\theta$ from n_1, compute the difference between each pair of observations. (b) Calculate the differences between each pair of observations. (c) Having the differences between all pairs of observations (for a lag h), calculate their average (spatial covariance). Hence, this is the first point of the graphical semivariogram, given as half of this average spatial covariance. The semivariogram is presented as half of the covariance in the *y axis* and distance h in the *x axis*. (d) Repeat the process for all equivalent lags ($2h$, $3h$, ..., $n_{lags}h$) to calculate $\gamma/2\ 2h$, $\gamma/2\ 3h$, ..., $\gamma/2\ n_{lags}h$ and, finally, to build the empirical semivariogram.



Source: Adapted from Lindner (2015).

The next steps refer to finding a theoretical model that best fits the empirical semivariogram and to detect the main directions, considering the graphical parameters of a semivariogram, as illustrated in Figure 2.

Figure 2 - Graphical parameters of a semivariogram.



Source: Adapted from Wackernagel (2003).

After setting the theoretical semivariogram model for the major and minor directions of anisotropy, the graphical parameters obtained are used in the interpolation process (Kriging). Kriging consists of a linear prediction since its estimates are weighted linear combinations based on existing data. It assumes that nearby observations tend to be more similar than observations that are spread apart. This assumption is verified by the method's intrinsic weighting (OLIVER; WEBSTER, 2015).

Values of nugget (C_0), partial sill (C) and range (r), presented in Figure 2, are used to define the weights. Eq. (2) exhibits the kriging estimator (WACKERNAGEL, 2003).

$$Z(x_0) = \sum_{i=1}^{n} \lambda_i \times Z(x_i) \quad (2)$$

The weights (λ_i) are generated using the kriging equation system, Eq. (3), which must yield unbiased estimates and minimize the variance. They establish the effect of sampled data on the estimation of new values.

$$\lambda = (G)^{-1} \times M \quad (3)$$

These weights are calculated according to the structural proximity of the data (G) and the disaggregation effect (M) derived from a covariance matrix. This means that the more correlated the samples, the greater the redundancy and the lower the individual weight when building the estimator.

The Indicator Kriging (IK) was applied in this study to estimate the probability of an event. However, this study used a binary variable related to the households' mode choice (car/motorcycle or public transport), rather than a probabilistic approach for the IK.

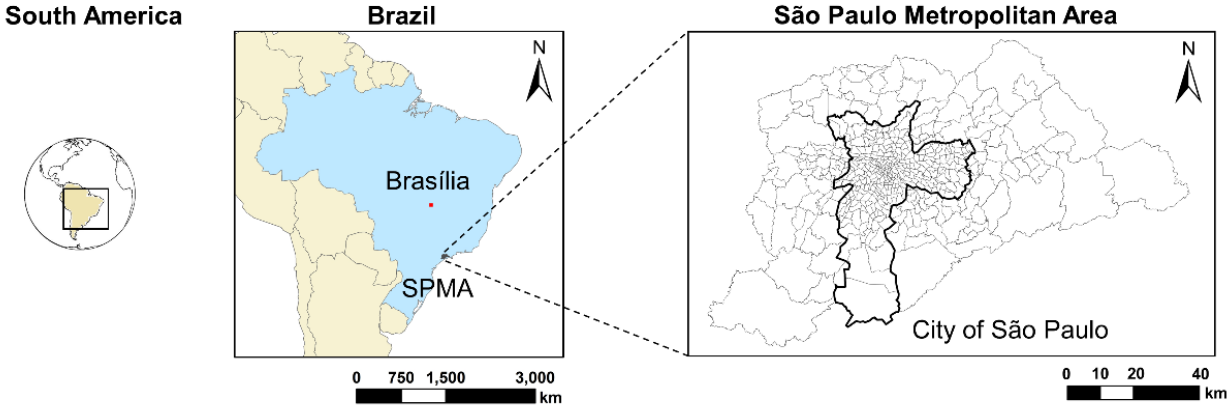
The results of the estimates are represented in a kriged map. This paper used the IK to interpolate the values of the variable of the motorized travel mode choice in the study area in order to validate the sampled households (at each respective spatial position) and to estimate values at non-sampled areas.

3 MATERIALS AND METHOD

3.1 Materials: study area and dataset

This present study case was applied to the São Paulo Metropolitan Area (SPMA), with total area of 7,946.84 km² including 39 municipalities (METRÔ, 2007). Figure 3 illustrates the spatial location of the SPMA and its 460 Traffic Analysis Zones (TAZs). The central part with smaller TAZs refers to dense urban areas and large trip generation hubs. This region concentrates the largest financial center and a great portion of the richest city in Brazil (São Paulo).

Figure 3 - Study area (SPMA, Brazil).



Source: The authors (2021).

According to the São Paulo Metropolitan Company (METRÔ, 2007), the approximately 20 million residents of the SPMA produced roughly 38 million trips per day in 2007. Table 1 shows the frequency of usage of each travel mode in the SPMA and highlights the importance of public and non-motorized transportation, which combined correspond to almost 2/3 of total daily trips in the SPMA.

Table 1 - Distribution of each travel mode considering daily trips.

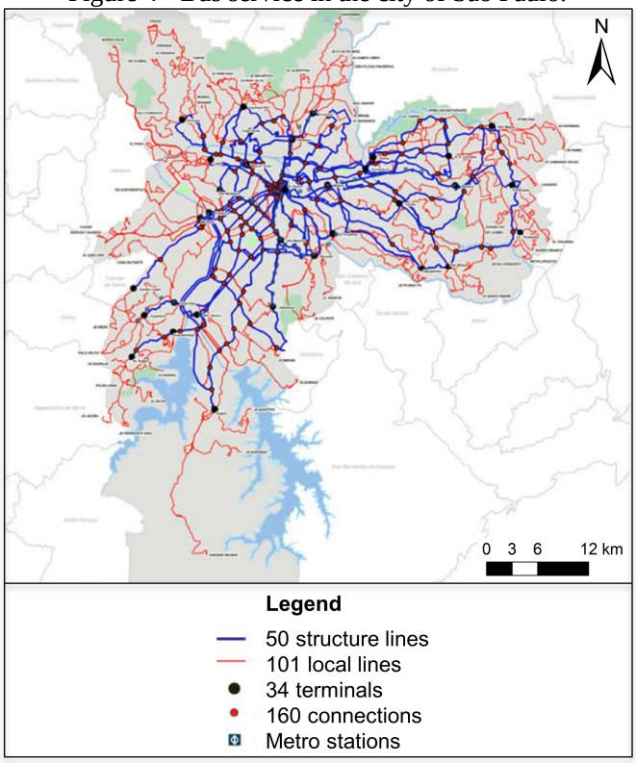
Main travel mode	Travel mode	Trips	%	%
Transit	Bus	9,034,074	23.7%	31.6%
	Metro	2,223,397	5.8%	
	Train	815,177	2.1%	
Private motor vehicle	Motorcycle	721,156	1.9%	21.0%
	Car (driver)	7,276,263	19.1%	
Other motorized	Car (passenger)	3,105,088	8.2%	13.2%
	Chartered transport	513,591	1.3%	
	School transport	1,326,602	3.5%	
	Taxi	90,686	0.2%	
Non-motorized	Bicycle	303,828	0.8%	33.9%
	Walking	12,623,047	33.1%	
Other	Other	61,475	0.2%	0.2%
Total		38,094,385	100.0%	100.0%

Source: Metrô (2007).

The SPMA public transportation system currently comprises buses, trains and a metro system. As for 2007, the rail transport system was run by three companies: CPTM, Metrô and ViaQuatro. ViaQuatro owns a single line, whereas CPTM operates six lines with 257.5 kilometers that cover 19 municipalities in the SPMA (Metrô, 2007). Metrô operates five lines with total length of 68.5 kilometers. The frequency of trains is higher for the Metrô lines as it covers the main part of the city of São Paulo. Twenty nine percent of the users of Metrô lines dwell in the eastern zone and 23% in the southern zone of the SPMA (METRÔ, 2014).

The bus transportation system is managed by the EMTU company in the SPMA and by SPTrans in the city itself (EMTU, 2020). EMTU serves not only the SPMA, but also other municipalities (450 lines, an approximate fleet of 10,000 buses with a ridership of approximately 1.5 million users per day). The EMTU is operated by other 60 companies and the SPTrans is also operated by 16 consortia (total of 1,300 lines and 15,000 vehicles). Figure 4 shows the bus lines and terminals in the city of São Paulo.

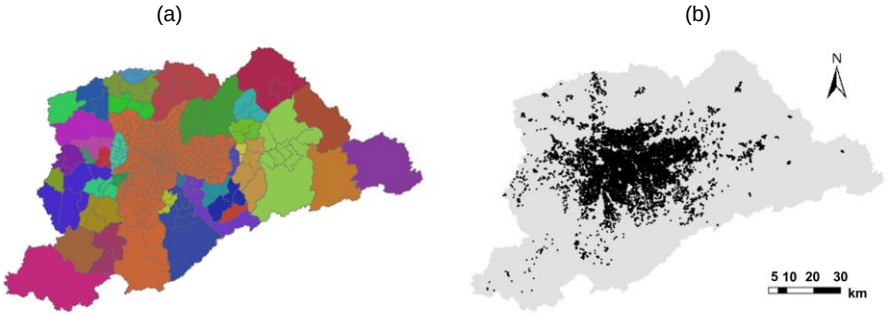
Figure 4 - Bus service in the city of São Paulo.



Source: Adapted from CLN (2017).

The dataset used in this paper refers to an Origin-Destination (O/D) Survey, carried out in the SPMA in 2007 by the São Paulo Metropolitan Company (Metrô). The survey collected data from 30,000 randomly selected households, distributed into 460 TAZs and 39 municipalities, according to Figure 5.

Figure 5 - (a) SPMA and its municipalities (b) and the sampled households.



Source: The authors (2021).

The variables analyzed from the O/D survey are related to daily trips. However, considering that this study requires georeferenced information as an input for the geostatistical analysis, the straightforward use of corresponding positions for heterogeneous trips (i.e., different trips associated to the same household) generate methodological inconsistencies. Taking this into account, the information used was related to each household, whereby its pair of coordinates is mutually exclusive. Table 2 presents the household-related variables used in this study.

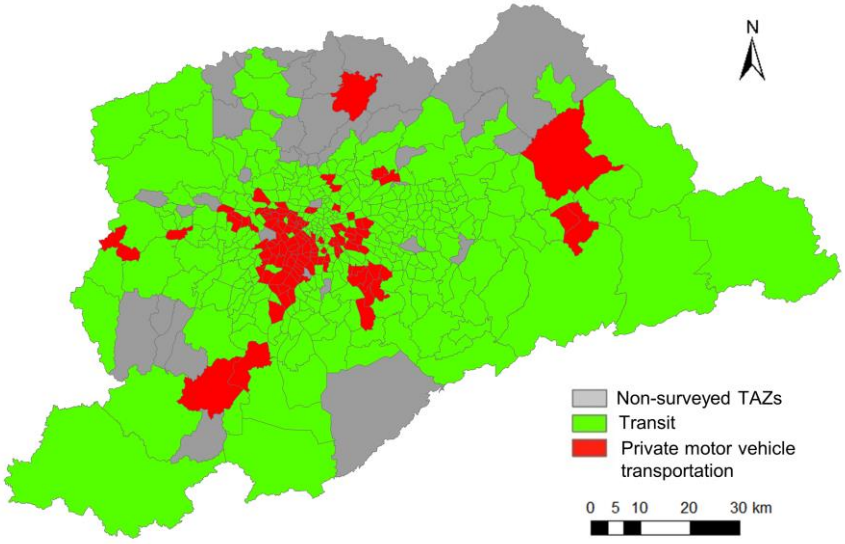
Table 2 - Descriptive statistics of the quantitative variables

Variables per household	average	stand. dev	mode	min.	max.
Number of residents	3.3	1.6	3	1	24
Number of cars	0.9	0.9	1	0	8
Number of motorcycles	0.1	0.3	0	0	9
Number of bicycles	0.5	0.9	0	0	12
Number of trips	6.8	4.6	2	1	68
% of trips using private motor vehicle	27%	32%	0%	0%	100%
% of trips using non-motorized transportation	17%	24%	0%	0%	100%
% of trips using public transportation	33%	32%	0%	0%	100%

Source: Metrô (2007).

Additionally, Figure 6 shows the motorized travel mode choice per TAZ, considering the O/D survey and the choice per household.

Figure 6 - Distribution of the motorized travel mode choice per TAZ.



Source: The authors (2021).

The dependent variable (travel mode choice) was developed based on the level of motorization of each household. The travel mode choice was set as a binary variable and its value was built upon the comparison of

the percentage of trips made using private motor vehicles and the percentage of trips made using public transportation within each household. Higher percentages of trips made using private motor vehicles were represented by the value *0*. Otherwise, the variable was set as *1*.

Therefore, the developed variable indicates which motorized choice is predominant for each household. The variable was created as a way of providing useful information to transport planners, as it shows which locations the households are more likely to use public transportation compared to the private mode. This information is useful, considering that the demand for public transportation can be seen using a kriged map, thus providing useful information for urban planning policies.

The dichotomous dependent variable was analyzed and estimated using two approaches: the traditional (Logistic Regression - LR) and the spatial technique (Geostatistics). This study investigated the usefulness of spatial data to predict motorized travel mode choice. In order to yield the spatial association among observations the geostatistical approach considered the spatial proximity, whilst the traditional approach considered five socioeconomic variables related to each sampled household: 1) number of residents; 2) income criterion (the higher the value, the lower the social class); 3) number of cars; 4) number of motorcycles; 5) number of trips.

The choice set is not exhaustive and excludes all non-motorized modes. This could be an important methodological limitation for logit models. However, this constraint was assumed taking into account that the non-motorized travel mode choices do not have spatial dependence (as no spatial structure was seen for non-motorized choices in this dataset). This fact makes it unfeasible to use non-motorized mode preferences in geostatistical modeling, corroborating the exclusion of this alternative. It is also important to mention that the conclusions are drawn for the sample and not for the entire population since expansion factors were not used.

3.2 Method

This section describes the calibration and validation process for the Logistic Regression and the geostatistical modeling steps. The next subsections describe the data processing and the analysis of results.

3.2.1 LOGISTIC REGRESSION

The traditional modeling consisted of using the motorized travel mode choice as the dependent variable and five socioeconomic variables as explanatory variables. The dataset was split into two subsets: 70% for the calibration set and 30% for the validation set. A logistic regression model was calibrated from the former and evaluated using the latter. The overall goodness of fit was verified adopting the determination coefficients (R^2) proposed by Cox & Snell and Nagelkerke (COX; SNELL, 1989; NAGELKERKE, 1991). The Cox & Snell R^2 compares the log likelihood of the model to the log likelihood of a baseline model, which has a theoretical value lower than 1 even for a perfectly adjusted model. Meanwhile, the Nagelkerke R^2 adjusts the scale of the Cox & Snell R^2 to cover a range from 0 to 1. The estimates were also assessed by the average hit rate and the absolute error. For a further understanding of the econometric concepts associated to the Logistic Regression, refer to Long (1997), Pampel (2000), Hosmer and Lemeshow (2000) and/or Menard (2002).

3.2.2 GEOSTATISTICAL MODELING

The geostatistical approach estimates spatial data using a weighting framework, according to the distances between pairs of observations. Therefore, this case study considered the interest variable and its spatial location for the modeling process. The geostatistical model adopted the same data source from the traditional modeling, however, the information needed by the spatial approach entailed exclusively the travel mode choice values and the location of each observation. Since the dataset also contains households with corresponding locations, a single residence was randomly selected for these cases. This procedure was essential to assure that each location was represented by a unique value of the study variable. Moreover, only 10% of the surveyed households were omitted from this analysis due to their overlayed locations.

The semivariograms were calculated to illustrate the spatial variability. Contrasting scenarios of spatial

isotropy and anisotropy were investigated. Thus, both cases were verified by 1) an omnidirectional semivariogram and 2) orthogonal semivariograms performed for directions multiple of 15° and an angular tolerance of 1°. In case of isotropy, the resulting orthogonal semivariograms would display similar behavior among different directions, suggesting that using an omnidirectional semivariogram may be suitable.

After calculating empirical semivariograms, graphical parameters were analyzed to detect the goodness of fit. Positive aspects suggesting an acceptable semivariogram model are listed as: a) reduced nugget effect; b) adherence to theoretical curves (spherical, exponential and Gaussian); and c) large values for the ratio between the range of pairs of orthogonal directions. These requirements show strong evidence of spatial autocorrelation.

Having the graphical parameters, the weighted estimation and the validation were carried out using Indicator Kriging and the fictitious point method (cross-validation), respectively (SEAMAN, 1983).

Finally, having conducted the cross-validation, the performance of both approaches can be evaluated by calculating statistical measures that compare estimated with observed values (absolute error and hit rate).

4 RESULTS AND DISCUSSION

4.1 Logistic Regression

This approach assumed five travel demand covariates for the motorized travel mode choice estimation. Table 3 shows the parameters resulted from the calibrated logistic model.

Table 3- Parameters resulted from calibrated model for the LR.

Explanatory variable	β	t	sig.
Number of residents	0.128	14.959	0.000
Income criterion	0.231	22.868	0.000
Number of motorcycles	-0.564	-8.353	0.000
Number of cars	-1.251	-24.757	0.000
Number of trips	-0.170	-19.615	0.000
Constant	-16.220	19.070	0.000

Source: The authors (2021).

Assuming that higher predicted values imply higher probability of households choosing transit over private motor vehicles, some conclusions can be drawn from the yielded coefficients (Table 3). It is worth highlighting that all explanatory variables were statistically significant. Furthermore, the variation inflation index for each explanatory variable resulted in values ranging from 1.02 to 1.81, thus corroborating a non-multicollinear scenario (KUTNER et al., 2004).

A positive effect of the income criterion and the number of residents on the choice of public transportation can be noted. These results are expected, since increased income criteria imply lower social class, thus resulting in higher probabilities for households to prefer transit. Additionally, a larger number of residents per household also entail higher preferences for transit. On the other hand, the number of motorcycles and cars are inversely proportional to the preference for public transportation. Therefore, an increase in the number of vehicle ownership implies a strong preference for individual motorized modes. Moreover, large number of trips per household are associated with higher preferences for private transportation.

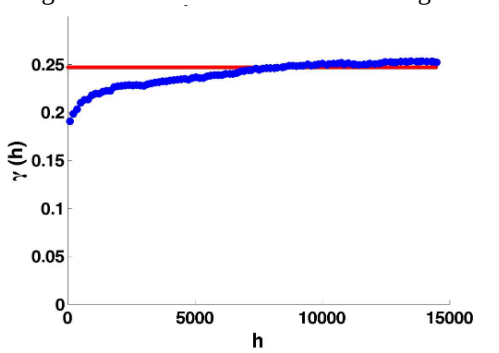
The goodness-of-fit was supported by the determination coefficients proposed by Cox & Snell and Nagelkerke, which resulted in 0.69 and 0.88, respectively. The hit rate of the validation set was of 97% and the absolute error 0.017. These results corroborate the hypothesis that the covariates consistently explain the study variable.

Furthermore, a calibration regression was carried out using a null model – which considered only the constant – resulted in a hit rate of 55%. This validates the interest in including explanatory variables to the model, especially as the hit rate increased by 42%.

4.2 Geostatistical approach

The resulting empirical semivariograms produce evidence of spatial variability. The omnidirectional semivariogram (Figure 7) represents the approximate structure seen in directional semivariograms. The blue points illustrated in Figure 7 correspond to the semivariances for lags of 146 meters. The red line, in turn, indicates the average variance among all pairs of points in the sample.

Figure 7 - Omnidirectional semivariogram.

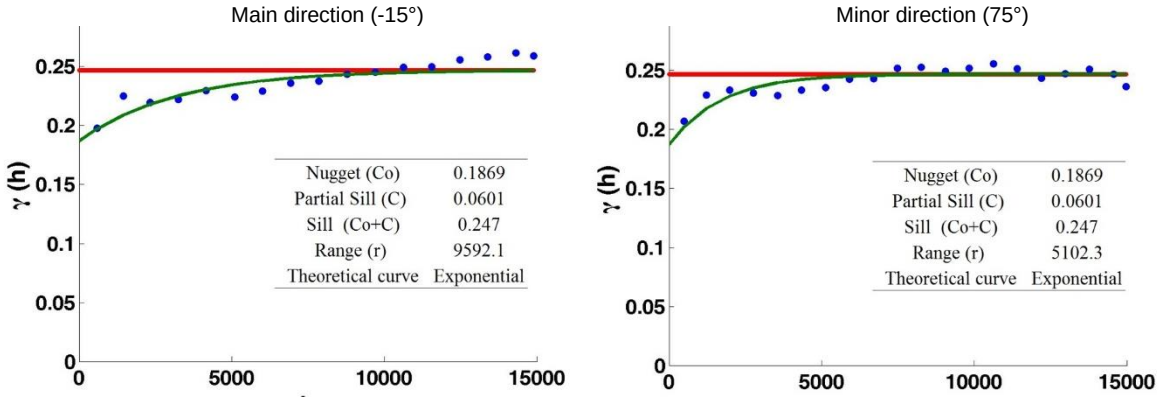


Source: The authors (2021).

A positive association of the variable with the distance between households can be noticed from the semivariogram. This trend is seen until stationarity is reached. The longer the distance between sampled households, the higher the difference between motorized travel mode choices. This occurred until a certain point in which the households were so far apart from each other that the distances no longer affected the variances. The maximum observed variances were very similar regardless the direction. The directional semivariograms (for each 15º axis) differed in terms of their ranges, as the distances between observations were different to reach stationarity. This finding supports existing evidence of geometrical (elliptic) anisotropy, in which the semivariograms present similar sills and distinct ranges. Another observation regarding the semivariogram model is that the nugget effect was approximately 76% of the sill, showing great variability of motorized travel mode choice between nearby households.

Within the spherical, exponential and Gaussian theoretical models, the semivariograms presented aspects of exponential fitting. The fitting process indicates that the directions with higher ratio between ranges were of -15° and 75°. The graphical representation was based on lags of 924 and 787 meters for -15° and 75°, respectively. Figure 8 illustrates the semivariograms for the main and minor directions.

Figure 8 - Theoretical semivariogram models for the main and minor directions.

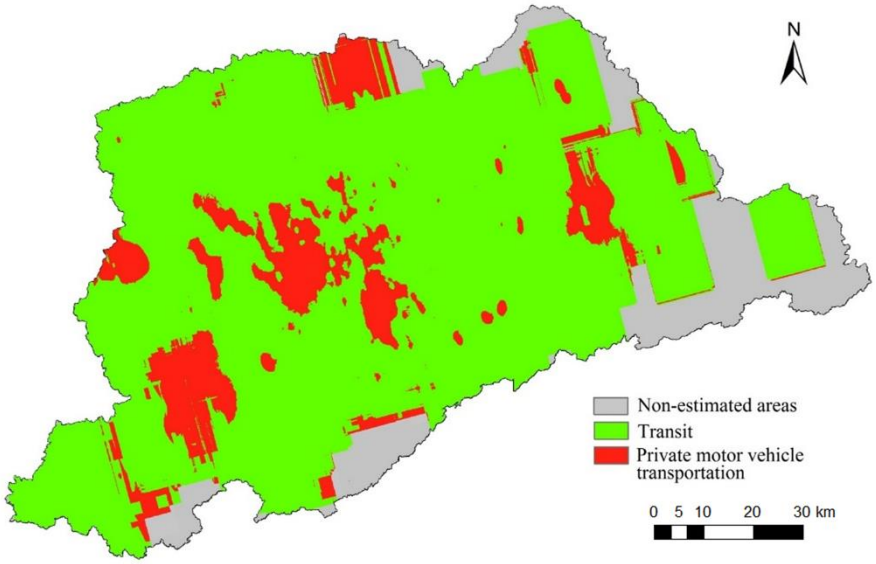


Source: The authors (2021).

The process of mapping the variable (kriging) was based on the spatial structure shown in the graphical parameters (Figure 8) and the binary values of motorized travel mode choice associated to their respective location. With respect to the estimated and the observed values, the statistical measures were calculated, yielding a hit rate of 67% and an absolute error of 0.33. Figure 9 illustrates the kriged map. It is worth

mentioning that the gray areas in the map correspond to regions not estimated by the kriging procedure. Most of them either overlap or are next to TAZs with non-surveyed residences, which resulted in an insufficient number of points for the kriging estimation.

Figure 9 - Kriged map of the motorized travel mode choices in the SPMA.

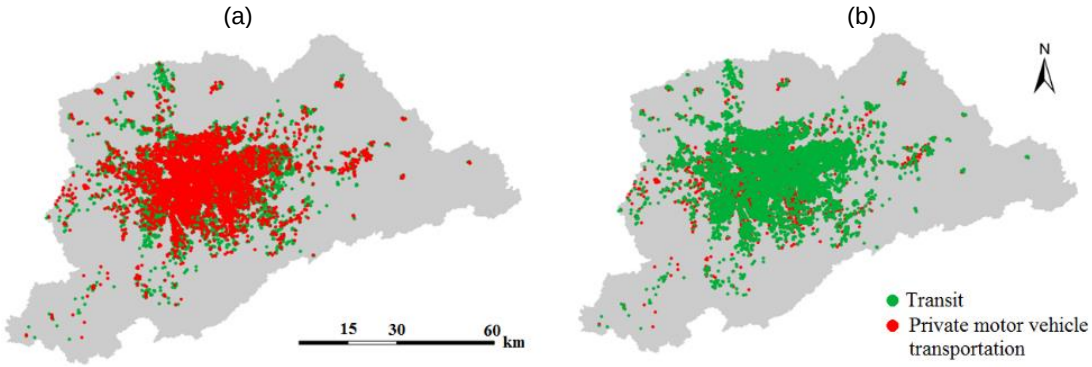


Source: The authors (2021).

The kriged map is appropriate as it enables estimating the variable of interest in some locations that were not previously sampled. Moreover, it is able to detect spatial patterns. The spatial analysis of travel mode choice behavior is a valuable tool when implementing public policies, e.g., when extending public transportation network or/and optimizing routes.

The comparison between both the geostatistical and the traditional approach demonstrates that the mode choice is better explained by socioeconomic variables rather than by geographic coordinates. According to expectations, all the regression coefficients associated with income and vehicle ownership were significant at 1% level. However, the overall objective of this study was to demonstrate the potential of spatial methods, in particular the geostatistical technique, especially considering the lack of available or updated information. In such cases, the spatial position and the respective value of a study variable provide sufficient information to produce a kriged map. This kriged map, in turn, allowed for a fairly hit rate of 67% using considerably reduced amount of data. The outcome can be compared considering the map in Figure 10 showing 18,733 households and their respective motorized travel mode choices.

Figure 10 - Representation of the motorized travel mode choice of households: (a) private motor vehicle – shown in red – superimposed on transit preferences, (b) transit – shown in green – superimposed on private motor vehicle preferences.



Source: The authors (2021).

Figure 10 shows a point-related representation of motorized travel mode choices considering the preference of each household. Figure 10 exhibits two inconclusive representations of the travel mode choice behavior, considering that there is a myriad of observations that overlay each other. On the other hand, Figure 9 presents the kriged map, in which the motorized mode choice is defined for the entire study area without undue overlap.

It is important to highlight that the transit structure in SPMA is composed by a uniform tariff system, thus not being affected by travel distances. Therefore, a trip between the outskirts and the city center may cost the same as a trip between central TAZs. Having that, on average, the distance between suburban TAZs and central TAZs are of approximately 28.2 kilometers. Thus, the spatial tendency shown in Figure 9 could be explained by the uniform travel cost (i.e., regardless the travel distance, which favors suburban residents in long distance-trips) or by other behavioral aspects.

Based on the mapping of motorized choices shown in Figure 9, it can be inferred that there is a spatial pattern, in which there is a greater probability of preferences for individual motorized transportation carried out either by cars or motorcycles in the city center of São Paulo. The preference for transit tends to increase with larger distances from the household to the city center.

5 CONCLUSIONS

Both the Logistic Regression and the geostatistical approach allowed for considerable findings in the estimation of motorized travel mode choices. The Logistic Regression yielded a model that consistently represented the phenomenon and achieved a 97% hit rate. The results show consistency with expected effects for the income, the vehicle ownership (DIELEMAN; DIJST; BURGHOUWT, 2012) and the number of residents per household. The geostatistical technique, in turn, was able to provide valuable contribution to our wider understanding of the spatial behavior of the travel mode choice. The dataset presented large variability for close observations (seen by large values for the nugget effect), showing evidence that individual behavior significantly affects travel demand, as an individual chooses different travel modes compared to others nearby. Therefore, we highlight that the geostatistical technique does not replace the traditional method since the spatial autocorrelation is only a single factor among several others to estimate travel mode choice. In fact, a wide range of human factors may affect this decision. Thus, the use of geostatistical techniques must consider that a large value for the nugget effect is to some extent acceptable when assessing the travel structure of household-related (disaggregated) decisions. Despite this, we suggest an in-depth investigation of the quantitative effect of the support (unit area) to the spatial structure of a travel demand variable (refer to ROCHA, 2019).

However, it is important to mention that the use of disaggregated data overcomes some of the problems and constraints seen in datasets associated to TAZs. In these cases, the geostatistical assumption of spatial homogeneity of the geographical support is violated and the MAUP needs to be addressed (GOOVAERTS, 2008).

Despite the lower hit rate yielded when compared to the logistic regression, the geostatistical approach demands less information to spatially estimate the study variable and is of great interest, especially in cases of scarce, outdated or unavailable data at a disaggregated level. For developing countries, it can be a promising alternative as the data collection for household surveys are considerably costly. In addition, the geostatistical model has the potential of producing even better outcomes, given that covariates can be incorporated within the estimator (Kriging with External Drift). This suggested approach corresponds to the Multiple Logistic Regression, which considers explanatory variables (socioeconomic). Another advantage of applying geostatistical estimation models to travel demand variables is the possibility of estimating mode choice in non-sampled coordinates. Therefore, the results point out to the promising aspects of using spatial information to estimate travel mode choice variables.

Further research should apply the proposed geostatistical approach to stated preference data as means of developing a full picture of different supply scenarios on travel mode choice. This framework enables the investigation of control variables associated with public transport supply. The use of household spatial information could further enhance the quality of such travel demand models, thus providing a more precise representation for decision makers and public transport operators.

Acknowledgments

We would like to thank the Brazilian National Council of Technological and Scientific Development (CNPq 151115/2019-2 and 304345/2019-9) for funding this research.

Authors' Contributions

The first author (Anabele Lindner) conceptualized the research, developed the methodology, curated the data, carried out formal analysis, wrote and revised the manuscript. The second author (Cira Souza Pitombo) supervised the research project and revised the manuscript. The third author (Lucas Assirati) implemented supporting algorithms and conducted data analysis. The fourth author (Jorge Ubirajara Pedreira Junior) wrote and revised the manuscript. The fifth author (Ana Rita Salgueiro) assisted with the variographic analysis.

Conflicts of Interest

The authors declare that there is no conflict of interest.

References

ABEP – **Brazilian Association for Demographic Studies Base LSE 2006/2007.** São Paulo, 2007. Available at: <http://www.abep.org/criterio-brasil>. Last accessed on: 20 feb. 2017.

BEN-AKIVA, M. E. **Structure of Passenger Travel Demand Models**. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, 1974.

BEN-AKIVA, M. E.; RAMMING, M. S.; BEKHOR, S. **Route choice models. Human Behaviour and Traffic Networks**. Springer Berlin Heidelberg, 23-45, 2004. DOI: 10.1007/978-3-662-07809-9_2.

BHAT, C. R., SENER, I. N. A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units. **Journal of Geographical Systems** v.11 n.3, 243–272, 2009. DOI: 10.1007/s10109-009-0077-9.

BHAT, C.; ZHAO, H. The spatial analysis of activity stop generation. **Transportation Research Part B: Methodological**. 36.6, 557–575, 2002. DOI: 10.1016/S0191-2615(01)00019-4.

BIERLAIRE, M. (2003) **BIOGEME: A free package for the estimation of discrete choice models**. Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland.

CERVERO, R.; RADISCH, C. Pedestrian versus automobile oriented neighborhoods. **Transport Policy**. v.3, 127–141, 1996. DOI: 10.1016/0967-070X(96)00016-9.

CIUFFO, B.; PUNZO, V.; QUAGLIETTA, E. **Kriging meta-modeling to verify traffic micro-simulation calibration methods**. 90th TRB Annual Meeting. Transportation Research Board, 2011.

CLN – Central Leste de Notícias. São Paulo, 2017. Available at: <http://itaimpaulista.com.br/portal/uploads/itaimon_1205201560.jpg>. Last accessed on: 20 feb. 2017.

COX, D.R. AND E.J. SNELL. **Analysis of Binary Data**. Second Edition. Chapman & Hall, 1989.

DIELEMAN, F. M.; DIJST, M.; BURGHOUWT, G. Urban form and travel behaviour: micro-level household attributes and residential context. **Urban Studies**, v. 39, n. 3, 507-527, 2002. DOI: 10.1080/00420980220112801.

DOMENCICH, T. A.; MCFADDEN, D. **Urban Travel Demand**. North-Holland Press, Amsterdam, Netherlands, 1975.

DUGUNDJI, E.; WALKER, J. 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: Journal of the Transportation Research Board 1921: 70-78, 2005. DOI: 10.3141/1921-09.

EMTU – Metropolitan Company of Urban Transports. **Transport Network**. São Paulo, 2020. Available at: <http://www.emtu.sp.gov.br/emtu/redes-de-transporte.fss>. Last accessed on: 30 jun. 2020.

GOOVAERTS, P. Kriging and Semivariogram Deconvolution in the Presence of Irregular Geographical Units. **Mathematical geology**, 40(1), 101–128, 2008. DOI: 10.1007/s11004-007-9129-1.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**, 2nd ed. New York: Wiley, 2000.

ISAAKS, E. H.; SRIVASTAVA, R. M. **Applied Geostatistics**. Oxford. University Press, New York, 1989.

JOURNEL, A. **Fundamentals of Geostatistics in five lessons**. Washington, DC: American Geophysical Union, v. 8, 1989.

KAYGISIZ, Ö.; DÜZGÜN, Ş.; YILDIZ, A.; SENBIL, M. Spatio-temporal accident analysis for accident prevention in relation to behavioral factors in driving: The case of South Anatolian Motorway. **Transportation research part F: traffic psychology and behaviour**, v. 33, p. 128-140, 2015. DOI: 10.1016/j.trf.2015.07.002.

KITAMURA, R.; MOKHTARIAN, P. L.; LAIDET, L. A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. **Transportation**. n. 24, 125–158, 1997. DOI: 10.1023/A:1017959825565.

KLATKO, T. J.; USMAN, S. T.; MATTHEW, V.; et al. Addressing the local-road VMT estimation problem using spatial interpolation techniques. **Journal of Transportation Engineering, Part A: Systems**, v. 143, n. 8, p. 4017038, 2017. DOI: 10.1061/JTEPBS.0000064.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J. **Applied Linear Regression Models** (4th ed.). McGraw-Hill Irwin, 2004.

LINDNER, A.; PITOMBO, C. S. A Conjoint Approach of Spatial Statistics and a Traditional Method for Travel Mode Choice Issues. **Journal of Geovisualization and Spatial Analysis**, v. 2, p. 1-13, 2018. DOI: 10.1007/s41651-017-0008-0.

LINDNER, A.; PITOMBO, C. S. Sequential Gaussian Simulation as a promising tool in travel demand modeling. **Journal of Geovisualization and Spatial Analysis**, v. 3, p. 1-15, 2019. DOI: 10.1007/s41651-019-0038-x.

LINDNER, A. **Disaggregated travel demand data analysis using traditional modeling and Geostatistics** (in Portuguese). Master's Dissertation - São Carlos School of Engineering, University of São Paulo. São Carlos, 2015.

LINDNER, A. **Heuristic methods to disaggregate travel demand data using geostatistical simulation** (in Portuguese). PhD Dissertation - São Carlos School of Engineering, University of São Paulo. São Carlos, 2019.

LINDNER, A.; PITOMBO, C. S.; ROCHA, S. S.; QUINTANILHA, J. A. Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study. **Geo-spatial Information Science**, v. 19, n. 4, 245-254, 2016. DOI: 10.1080/10095020.2016.1260811.

LONG, J.S. **Regression models for categorical and limited dependent variables: analysis and interpretation**. Thousand Oaks, CA: Sage, 1997.

MARQUES, S. F.; PITOMBO, C. S. Intersecting Geostatistics and Travel Demand Modeling: A Bibliographic Survey. **Revista Brasileira de Cartografia**, v. 72, p. 1004-1027, 2020. DOI: 10.14393/rbcv72nespecial50anos-56467.

MATHERON, G. Principles of Geostatistics. **Economic Geology** v. 58, n. 8: 1246–1266, 1963. DOI: 10.2113/gsecongeo.58.8.1246.

MATHERON, G. **The theory of regionalized variables and its applications**. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau. École Nationale Supérieure dês Mines de Paris. v. 5, 211p,

1971.

MAZZELLA, A.; PIRAS, C.; PINNA, F. Use of kriging technique to study roundabout performance. **Transportation Research Record: Journal of the Transportation Research Board**, n. 2241, 78-86, 2011. DOI: 10.3141/2241-09.

MCFADDEN, D. The measurement of urban travel demand. **Journal of Public Economics**, v.3, n.4, 1974. DOI: 10.1016/0047-2727(74)90003-6.

MENARD, S. **Applied logistic regression analysis**, v. 106. Sage, 2002.

METRÔ - SÃO PAULO METROPOLITAN COMPANY. **Origin-Destination Survey 2007 - São Paulo Metropolitan Area: Summary of information**. São Paulo, 2007. Available at: <http://www.metro.sp.gov.br/metro/numeros-pesquisa/pesquisa-origem-destino-2007.aspx>. Last accessed on: 14 feb. 2017.

METRÔ - SÃO PAULO METROPOLITAN COMPANY **Caracterização Socioeconômica do Usuário e seus Hábitos de Viagem**. São Paulo, 2014. Available at: <https://transparencia.metrosp.com.br/dataset/pesquisa-de-caracteriza%C3%A7%C3%A3o-socioecon%C3%B4mica-do-usu%C3%A1rio/resource/11cc5c4e-6aa7-4977-ba0a>. Last accessed on: 30 jun. 2020.

MIYAMOTO, K.; VICHIENSAN, V.; SHIMOMURA, N.; PÁEZ, A. Discrete choice model with structuralized spatial effects for location analysis. **Transportation Research Record: Journal of the Transportation Research Board**, n. 1898, 183-190, 2004. DOI: 10.3141/1898-22.

NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. **Biometrika**, 78:691-692, 1991. DOI: 10.1093/biomet/78.3.691.

OLIVER, M. A.; WEBSTER, R. **Basic Steps in Geostatistics: the Variogram and Kriging**. Springer International, SpringerBriefs in Agriculture, Springer, 2015.

ORTÚZAR, J. DE D.; WILLUMSEN, L. G. **Modeling Transport**. London: John Wiley & Sons. 4th ed. 586p, 2011.

PÁEZ, A.; LÓPEZ, F. A.; RUIZ, M.; MORENCY, C. Development of an indicator to assess the spatial fit of discrete choice models. **Transportation Research Part B: Methodological**, v.56, 217-233, 2013. DOI: 10.1016/j.trb.2013.08.009.

PÁEZ, A.; SCOTT, D. M. Spatial statistics for urban analysis: a review of techniques with examples. **GeoJournal**, v. 61, n. 1, 53-67, 2005. DOI: 10.1007/s10708-005-0877-5.

PAMPEL, F. C. **Logistic regression: a primer Sage University papers series on quantitative applications in the social sciences**. Newbury Park, CA: Sage, 2002.

PITOMBO, C. S.; COSTA, A. S. G. D.; SALGUEIRO, A. R. Proposal of a sequential method for spatial interpolation of mode choice. **Boletim de Ciências Geodésicas**, v. 21, n. 2, 274-289, 2015a. DOI: 10.1590/S1982-21702015000200016.

PITOMBO, C. S.; SALGUEIRO, A. R.; Costa, A. S. G.; Isler, C. A. A Two-step method for mode choice estimation with socioeconomic and spatial information. **Spatial Statistics**. v. 11, p. 45-64, 2015b. DOI: 10.1016/j.spasta.2014.12.002.

ROCHA, S. S. **Genetic Algorithms for optimization of geostatistical modeling applied to transport demand** (in Portuguese). PhD Dissertation – São Carlos School of Engineering, University of São Paulo. São Carlos, 2019.

ROCHA, S. S.; LINDNER, A.; PITOMBO, C. S. Proposal of a geostatistical procedure for transportation planning field. **Boletim de Ciências Geodésicas**, v. 23, p. 636-653, 2017. DOI: 10.1590/s1982-21702017000400042.

SEAMAN, R. S. **Objective Analysis accuracies of statistical interpolation and successive correction schemes**. In: Conference on Numerical Weather Prediction, 6 th, Omaha, NE. p. 141-148, 1983.

WACKERNAGEL, H. **Multivariate Geostatistics**. 3rd ed. 388, Berlin, Heidelberg: Springer-Verlag, 2003.

XIE, Z.; YAN, J. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. **Journal of Transport Geography**. v. 31, p. 64-71, 2013. DOI: 10.1016/j.jtrangeo.2013.05.009.

YAMADA, I.; THILL, J. C. Comparison of planar and network K-functions in traffic accident analysis. **Journal of Transport Geography**. v. 12, n. 2, p. 149-158, 2004. DOI: 10.1016/j.jtrangeo.2003.10.006.

ZOU, H. X.; YUE, Y.; LI, Q. Q.; YEH, A. G. O. An Improved Distance Metric for the Interpolation of Link-based Traffic Data Using Kriging: A Case Study of a Large-scale Urban Road Network**. International Journal of Geographical Information Science**. v. 26, n. 4: 667–689, 2012. DOI: 10.1080/13658816.2011.609488.

Main author's biography

Anabele Lindner, Curitiba/PR, Brazil. Civil Engineer graduated from the Federal University of Paraná (UFPR), Curitiba/PR. M.Sc. and Ph.D. in Transportation Engineering from the University of São Paulo (EESC-USP), São Carlos/SP, Brazil. Postdoctoral fellow at the Department of Transportation Engineering, University of São Paulo (EESC-USP), São Carlos/SP, Brazil. Her research interest is on urban transportation planning and infrastructure, spatial analysis and statistics.