Transportation Research Interdisciplinary Perspectives 34 (2025) 101735

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives

Bypassing the lack of available travel disaggregated data: A geo-spatial framework to simulate mode choice

Anabele Lindner [a,1], Cira Souza Pitombo [a,2], Samuel de França Marques [a,3,*]

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

ARTICLE INFO

Keywords:
Spatial analysis
Geostatistics
Simulation
Semivariogram deconvolution
Change of support
Travel mode choice

ABSTRACT

Disaggregated data for travel demand are essential resources toward appropriate urban planning, especially regarding public transportation. However, especially in developing countries, access to such information is limited. The current paper addresses this issue by introducing an approach, comprising semivariogram deconvolution, Sequential Gaussian Simulation (SGS), and validation, using regular spatial scales. As input to the procedure, we propose to use information with high availability, such as census microdata. The hallmark of the devised method lies in inferring travel spatial variability of more disaggregated unit areas using synthetic data. The method proposes to calculate synthetic data using the socioeconomic census microdata and a calibrated regression model with travel demand data associated with Traffic Analysis Zones (TAZs) in São Paulo city. The resulting maps and statistical metrics corroborate the original data values associated with TAZs. This paper presents relevant contributions as the method enables: bypassing the lack of available travel disaggregated data; creating different scenarios to reproduce travel spatial behavior; and assessing the associated uncertainty.

1. Introduction

Forecasting travel demand is crucial for urban planning policies. Traditional models for travel demand are usually based on Origin/Destination (O/D) Surveys, whose data are collected by randomly sampling households within the study area. However, the process of collecting such data is cumbersome, time-consuming and requires large financial investments by responsible municipal bodies.

In addition to this scenario, traditional models do not consider spatial factors as important variables to estimate travel demand. Furthermore, simulations used in conventional data disaggregation processes and/or synthetic data acquisition often overlook the spatial autocorrelation of travel demand variables. However, different studies have recognized and advocated the link between travel behavior and the spatial allocation of urban activities (Cervero and Radisch, 1996; Kitamura et al., 1997).

Long-established travel models applied to urban planning policies set out to replicate travel behavior using socioeconomic factors, for example. The lower the level of aggregation, the higher the amount of detail associated with the data. Therefore, individual information is a convenient resource for traditional urban planning methods. However, due to the confidentiality and high financial investment associated with the collection, such information is not regularly available. In the travel demand line of research, the process of obtaining disaggregated data is well consolidated in microsimulation approaches. Although microsimulation is a well-established approach to travel demand issues, Lindner and Pitombo (2019) highlight the potential of using spatial autocorrelation of travel-related variables as key components to input in the microsimulation and generating synthetic data. To the best of the authors' knowledge, this topic has not been explored in any other academic literature. Therefore, this underscores the need to address the main shortcoming identified by Lindner and Pitombo (2019), *i.e.*, obtaining the spatial structure of downscaled information, which is discussed in this paper.

Based on the context that travel demand variables are spatially correlated and that such a feature should be considered for data modeling and disaggregation, this paper proposes using geostatistical procedures. Various authors have already applied geostatistics in the

[*] Corresponding author at: Trabalhador São-carlense Avenue, 400, São Carlos, State of São Paulo, Brazil.
E-mail addresses: bele.lindner@gmail.com (A. Lindner), cirapitombo@gmail.com (C.S. Pitombo), samuelmarques@usp.br (S.F. Marques).

[1] ORCID: https://orcid.org/0000-0002-4487-3650.

[2] ORCID: https://orcid.org/0000-0001-9864-3175.

[3] ORCID: https://orcid.org/0000-0001-5602-3277.

https://doi.org/10.1016/j.trip.2025.101735
Received 14 September 2022; Received in revised form 27 September 2024; Accepted 9 November 2025
Available online 15 November 2025

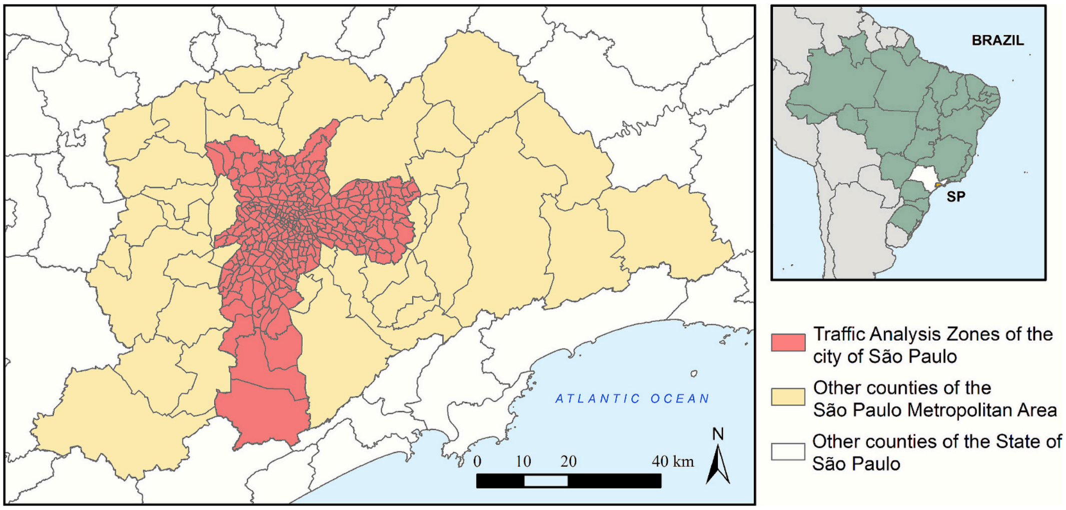
BRAZIL
SP
ATLANTIC OCEAN
0 10 20 40 km
N
Traffic Analysis Zones of the city of São Paulo
Other counties of the São Paulo Metropolitan Area
Other counties of the State of São Paulo

Fig. 1. São Paulo (city) location map and its TAZs.

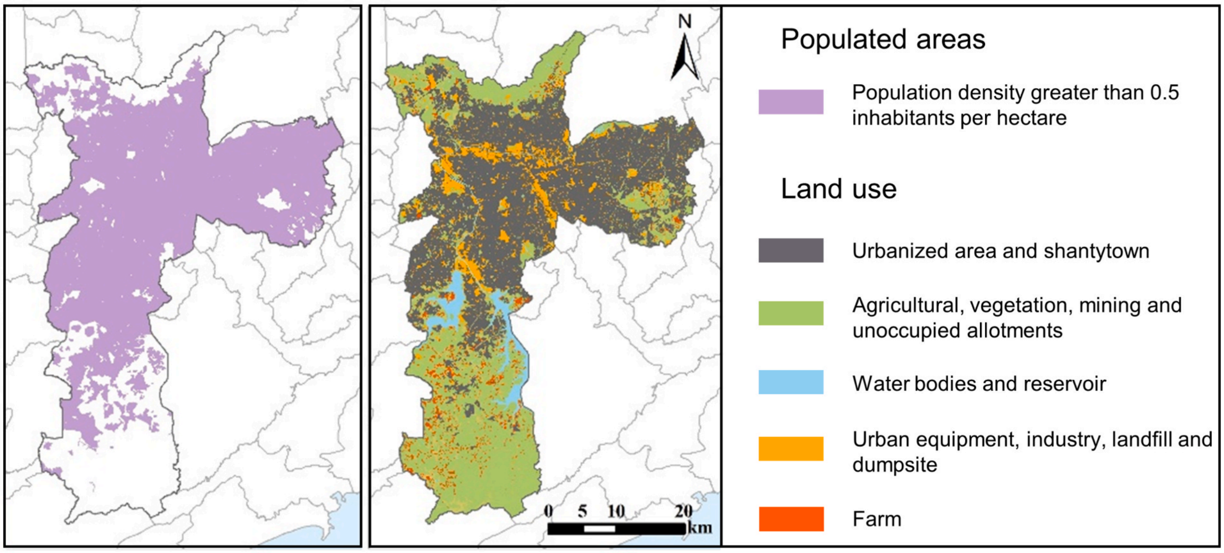
N
0 5 10 20 km
Populated areas
Population density greater than 0.5 inhabitants per hectare
Land use
Urbanized area and shantytown
Agricultural, vegetation, mining and unoccupied allotments
Water bodies and reservoir
Urban equipment, industry, landfill and dumpsite
Farm

Fig. 2. Populated areas and land-use map (adapted from GeoSampa website - https://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx, 2010).

field of transportation (Yoon et al., 2014; Chen et al., 2015; Miura, 2010; Pitombo et al., 2015; Lindner et al., 2016; Rocha et al., 2017; Lindner and Pitombo, 2018; Marques and Pitombo, 2021a, 2023; Marques et al., 2024). This technique enables modeling a variable at spatial positions whereby its values are unknown. Owing to the potential of using Geostatistics for travel demand, this paper sets out to explore geostatistical simulation to devise a heuristic framework for data disaggregation.

Sequential Gaussian Simulation (SGS) is the most popular geostatistical simulation technique. This method facilitates the calculation of equiprobable models that reproduce the spatial correlation and the probability distribution of a continuous variable (Verly, 1993). As several simulations are generated, the associated uncertainty, such as confidence intervals and conditional variances, can be calculated.

Simple downscaling processes using Geostatistics consist of analyzing the spatial structure of the variable (variographic analysis) and defining a semivariogram model as input in a kriging system, which allows for estimating values of a variable at non-sampled positions, based on the distance to their surrounding observations. The original input dataset of this case study is associated with Traffic Analysis Zones (TAZs), which have different shapes and sizes, leading to the Modifiable Areal Unit Problem (MAUP). However, the travel variable, originally related to TAZs, will become associated with smaller regular unit areas. Thus, due to the MAUP and its association with larger and irregular areal units, the initial semivariogram model is incompatible with the output information from the disaggregation process. Goovaerts (2008) proposes solving the MAUP by using a deconvoluted semivariogram, according to the concepts provided by Journel and Huijbregts (1978). However, the classic procedure for semivariogram deconvolution requires disaggregated data to calculate a regularized semivariogram.

The main aim of this paper is to bypass the lack of available travel disaggregated data through a heuristic approach comprising semivariogram deconvolution, Sequential Gaussian Simulation (SGS), and validation, using regular spatial scales. Furthermore, this paper contributes by (1) generating more disaggregated data through information associated with irregular areas to overcome the unavailability of individual/household data; (2) proposing an alternative procedure for semivariogram deconvolution by employing data with higher availability (e.g. census microdata); (3) obtaining different scenarios, with simulated data of the study variable, thus yielding the distribution of possible values of this variable and a map with confidence intervals.

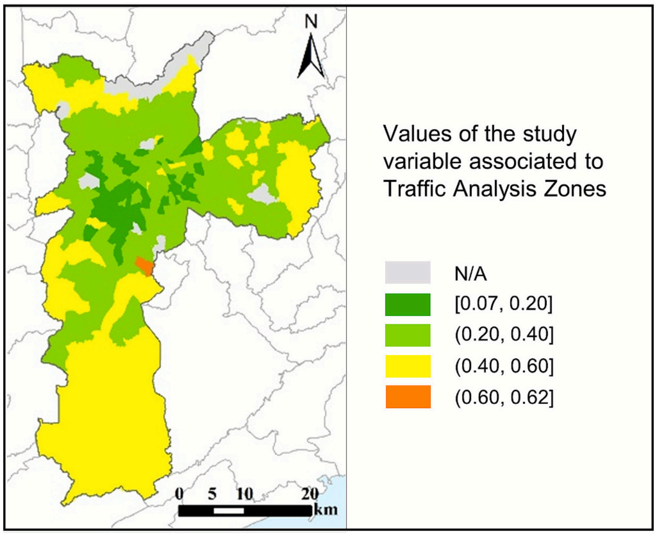
N
Values of the study variable associated to Traffic Analysis Zones
N/A
[0.07, 0.20]
(0.20, 0.40]
(0.40, 0.60]
(0.60, 0.62]
0 5 10 20
km

Fig. 3. Transit trip rate per TAZ (adapted from Metrô, 2007).

2. The path from traditional spatial analysis to geostatistical simulation in travel demand modeling

Socioeconomic factors, cost and service level are widely recognized as explanatory attributes in traditional logistic regressions for travel demand models. It can be further noted that conventional travel mode choice models disregard information related to the spatial position of variables. However, owing to the technological advances and the high availability of geographical data, spatial modeling is seen as an important area of interest in travel demand, especially with the insertion of spatial patterns in mathematical models (Páez and Scott, 2005; Páez et al., 2013).

Various authors have reached valuable results by adding spatial attributes to travel behavior analysis (Yamada and Thill, 2004; Dugundji and Walker, 2005, Xie and Yan, 2013, Kaygisiz et al., 2015). However, Geostatistics may be more advantageous as it enables exploratory and confirmatory analyses by forecasting values of spatially correlated variables at sampled and non-sampled locations using the distance between the observations of the dataset and the theoretical semivariogram function. In addition to incorporating the spatial variability of variables, geostatistical methods also consider aspects of spatial patterns, such as the main direction of continuity (known as spatial anisotropy) (Matheron, 1963). Besides, the estimation (or simulation) of geostatistical approaches is not achieved by simple spatial interpolation, but rather by a kriging process, using theoretical models that most fit the empirical semivariogram.

Geostatistical frameworks demonstrated in travel demand issues have not been sufficiently explored (Yoon et al., 2014; Chen et al., 2015; Marques and Pitombo, 2020). However, current research has shown that the technique may be promising to provide spatial estimates of travel demand variables using Ordinary, Universal, Indicator Kriging and Kriging with External Drift (Miura, 2010; Pitombo et al., 2015; Lindner et al., 2016; Gomes et al., 2016; Rocha et al., 2017; Lindner and Pitombo, 2018; Lindner et al., 2021; Marques and Pitombo, 2021b, 2023; Marques et al., 2024). Simple Kriging, in turn, is suitable for cases where the population mean is known, which applies to the current study. Considering the subject of travel demand modeling, the application of geostatistical procedures still requires in-depth studies, especially regarding the effect of the nature of the variables, which encounter obstacles as they are linked to human behavior.

The optimal geographical scale (support) preferred by specialists in traditional travel demand models may differ from the scale selected for spatial models. Transportation decision-makers usually adopt variables associated with individuals or households, if available, rather than areal data. However, individual information may not be suitable for spatial models, as it is point-related and surveys do not precisely capture individual geographical coordinates. In addition, individuals residing in the same household are likely to behave differently, causing biased georeferenced information. To efficiently work with socioeconomic data, a certain level of aggregation is needed. This issue leads to a change of support.

Geographical information systems applied to social sciences commonly address the disaggregation of demographic data through spatial interpolation (Flowerdew and Green, 1993; Goodchild et al., 1993). However, available methods have the drawback of requiring a disaggregated or regularized semivariogram (Kyriakidis, 2004), which, in turn, demands disaggregated data. Rocha et al. (2017) proposed an initial attempt to develop an alternative semivariogram deconvolution aiming at improving travel demand modeling, considering the assumption of continuity in geostatistical approaches and the MAUP. In spite of the procedure limitations, the authors presented an initial proposal to solve an important issue in travel demand analysis: the unavailability of disaggregated data.

The present paper gives some impetus to the concept of using semivariogram deconvolution for data disaggregation, while addressing the previously identified challenges. The concepts for geostatistical modeling are conducted using the Sequential Gaussian Simulation (SGS). SGS is a stochastic simulation technique that aims to establish a group of distinctive scenarios that reproduce spatial features. Stochastic simulations generate a range of realizations (formal designation for simulations) that may express the associated uncertainty in the spatial simulation or deconvolution method (Goovaerts, 1997; Remy et al., 2009). In addition, SGS enables the change of support (Goovaerts, 2001) and avoids smoothing effects that occur in kriging techniques (Deutsch and Journel, 1998).

An application of the SGS to explore different scenarios of transit production in the São Paulo Metropolitan Area (SPMA), Brazil (Lindner and Pitombo, 2019), demonstrates that the method may also be appropriate for travel mode choice variables. The authors have explored the following benefits of the stochastic simulation for the transportation field: gathering less information as input, incorporating the spatial association, predicting values at non-sampled positions, and mapping the simulated variable and the associated uncertainty using conditional variances and confidence intervals. However, the authors mention that the lack of availability of disaggregated spatial structure (semivariogram) may be seen as a drawback when applying the geostatistical simulation.

3. Case-study context, dataset and method

The case study area consists of the city of São Paulo, located in the east of São Paulo state (SP), Brazil, according to Fig. 1.

São Paulo is the most populous city in Brazil, comprising 320 TAZs with 9 million passenger displacements per business day (SPTrans, 2018). On average, the demographic density is 7,400 inhabitants per square kilometer. Fig. 2 (left side) shows the areas in which the population density is greater than 50 inhabitants per square kilometer – population densities lower than this threshold were assumed as non-populated regions. The land use map, also presented in Fig. 2 (right side), corroborates this assumption. It can be noted that both the northern and the southern areas (in greater part) encompass land use characteristics that do not demand an intense urban transportation network.

The dataset sources include the Brazilian Institute of Geography and Statistics (IBGE, 2010) and the São Paulo Metropolitan Company (Metrô, 2007), which provided the socioeconomic microdata and O/D database, respectively. The microdata pertain to 172,627 non-georeferenced households located at 310 sets of census tracts in the city of São Paulo. The 2007 O/D Survey, in turn, holds 196,699 travel

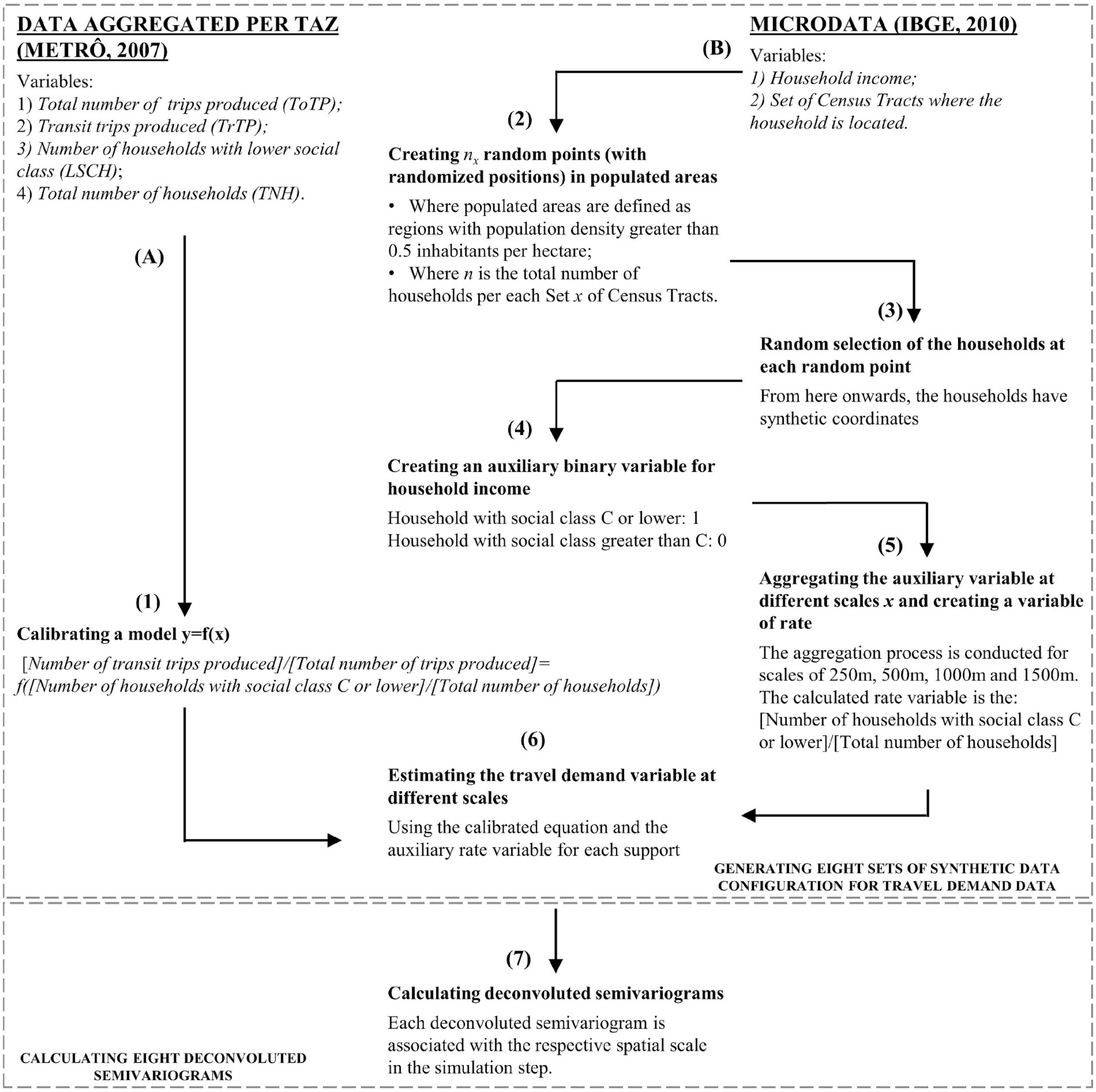
DATA AGGREGATED PER TAZ (METRÔ, 2007)
Variables:
1) Total number of trips produced (ToTP);
2) Transit trips produced (TrTP);
3) Number of households with lower social class (LSCH);
4) Total number of households (TNH).
(A)
MICRODATA (IBGE, 2010)
(B)
Variables:
1) Household income;
2) Set of Census Tracts where the household is located.
(2)
Creating n_x random points (with randomized positions) in populated areas
• Where populated areas are defined as regions with population density greater than 0.5 inhabitants per hectare;
• Where n is the total number of households per each Set x of Census Tracts.
(3)
Random selection of the households at each random point
From here onwards, the households have synthetic coordinates
(4)
Creating an auxiliary binary variable for household income
Household with social class C or lower: 1
Household with social class greater than C: 0
(5)
Aggregating the auxiliary variable at different scales x and creating a variable of rate
The aggregation process is conducted for scales of 250m, 500m, 1000m and 1500m. The calculated rate variable is the: [Number of households with social class C or lower]/[Total number of households]
(1)
Calibrating a model y=f(x)
[Number of transit trips produced]/[Total number of trips produced]= f([Number of households with social class C or lower]/[Total number of households])
(6)
Estimating the travel demand variable at different scales
Using the calibrated equation and the auxiliary rate variable for each support
GENERATING EIGHT SETS OF SYNTHETIC DATA CONFIGURATION FOR TRAVEL DEMAND DATA
(7)
Calculating deconvoluted semivariograms
Each deconvoluted semivariogram is associated with the respective spatial scale in the simulation step.
CALCULATING EIGHT DECONVOLUTED SEMIVARIOGRAMS

Fig. 4. Flowchart for the alternative semivariogram deconvolution.

records associated with 30 thousand surveyed households in the SPMA. In the city of São Paulo, the O/D Survey covered 15,759 households.

The information collected from 15,759 households was aggregated within the TAZs of the study area, resulting in a dataset of 320 records. The O/D dataset was also aggregated into different spatial scales to provide a validation tool (Comparative Method) for the proposed method. It should be noted that the number of observations sampled for census surveys overcomes travel datasets from O/D surveys. Hence, microdata represent a larger sample of the population and may provide alternative perspectives for travel analysis, covering a wider spatial sample, especially when considering geostatistical models.

The OD Survey ensures the representability of the study area by adopting a proportional stratified random sampling based on four levels of household energy consumption, covering various levels of income. This sampling method allowed for calculating the total of produced trips at each TAZ with error margins of less than 5 % (Metrô, 2008). In the Census case, 5 % of the households were interviewed, accounting for the population distribution inside each tract (IBGE, 2013). As a higher number of households were visited in the Demographic Census, the error margins were even lower than in the OD case when estimating total values based on sampled households.

The transit trip rate (*i.e.*, rate of trips by bus, metro and train, considering the main travel mode choice per household) is assessed as a study variable in this paper. Fig. 3 presents the values of the transit trip rate associated with the 320 TAZs in São Paulo, using the 2007 O/D Survey as a source.

It can be observed that the TAZs in southern São Paulo, shown in Fig. 2 as areas with low population density, are represented in Fig. 3 by higher rates of transit preference. Consequently, this may lead to misinterpretations of the associated travel behavior. Furthermore, the TAZs have homogeneous behavior when compared with one another, demonstrating the smooth effect of aggregated data, which results in loss of information. The present paper proposes a heuristic framework to disaggregate data with an alternative semivariogram deconvolution method and SGS, using individual socioeconomic data and travel data associated with TAZs.

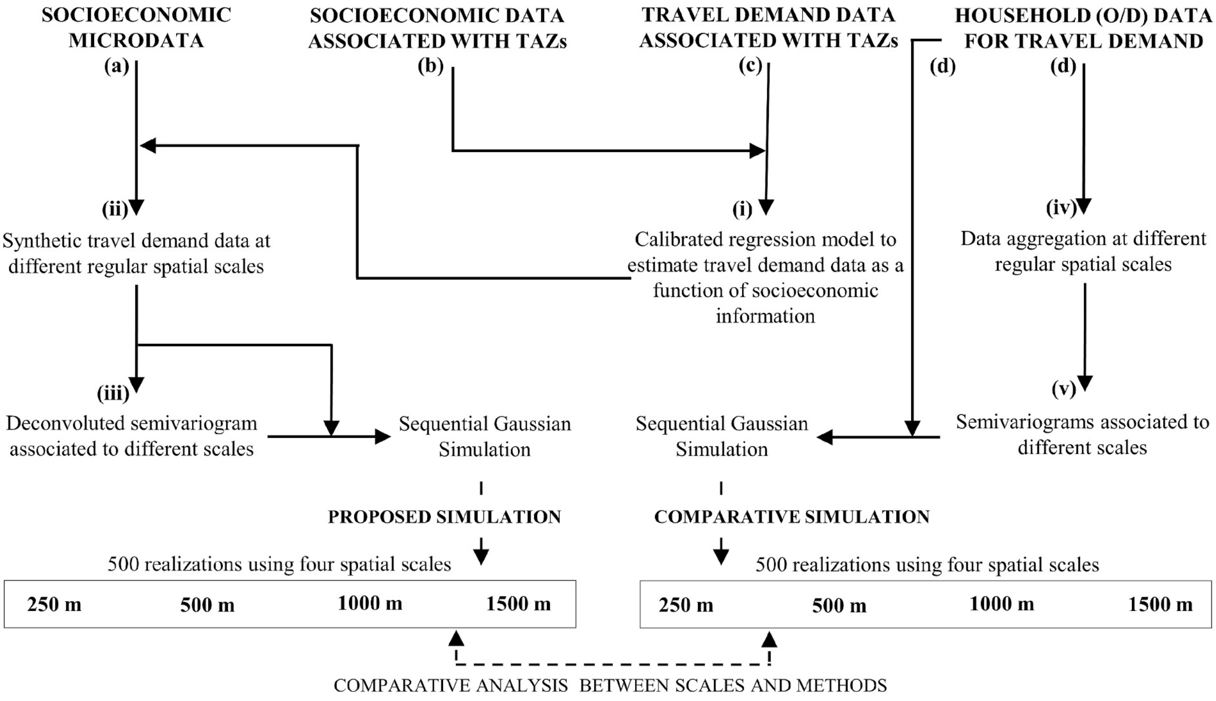
SOCIOECONOMIC MICRODATA
(a)
SOCIOECONOMIC DATA ASSOCIATED WITH TAZs
(b)
TRAVEL DEMAND DATA ASSOCIATED WITH TAZs
(c)
HOUSEHOLD (O/D) DATA FOR TRAVEL DEMAND
(d)
(d)
(ii)
Synthetic travel demand data at different regular spatial scales
(i)
Calibrated regression model to estimate travel demand data as a function of socioeconomic information
(iv)
Data aggregation at different regular spatial scales
(iii)
Deconvoluted semivariogram associated to different scales
Sequential Gaussian Simulation
Sequential Gaussian Simulation
(v)
Semivariograms associated to different scales
PROPOSED SIMULATION
COMPARATIVE SIMULATION
500 realizations using four spatial scales
250 m
500 m
1000 m
1500 m
500 realizations using four spatial scales
250 m
500 m
1000 m
1500 m
COMPARATIVE ANALYSIS BETWEEN SCALES AND METHODS

Fig. 5. Flowchart of the method.

3.1. Alternative semivariogram deconvolution

Fig. 4 summarizes the proposed method for disaggregating data, and the semivariogram deconvolution process. The steps are designated by Arabic numerals (1) to (7), whilst the datasets (and variables) are depicted by the letters (A) and (B).

One main explanatory variable was set as essential to calculate the study variable using a calibrated model: low-income social class, as it links both data sources (Census and O/D Survey). A linear regression model was calibrated considering the transit trip rate per household (study variable) as a function of the low-income household rate, using the O/D data at the TAZ level.

The social class is provided in its original format in the travel dataset (Social Class A, B, C, D and E – representing higher to lower household conditions, respectively). The present study assumes that social classes C, D and E cover low-income households, according to ABEP (2010). Whereas, when considering the census microdata, the social class may be inferred using the provided total income per household. Total incomes lower than BRL 1,541 represent low-income households, according to the criteria set by ABEP (2010).

Two data sources comprised a case study in São Paulo: the 2007 O/D Survey and the 2010 Demographic Census. Different units of analysis were used: census households with synthetic coordinates, O/D households with real coordinates, O/D TAZs, and grid squares. In short, the O/D TAZ data (transit trip rate and low-income household rate) were used to calibrate a linear regression model. Afterward, this equation was used to calculate the transit trip rate with the low-income household rate from the census household data. For both O/D and census data, the transit trip rate was obtained for different aggregation units. However, in the census data case, the transit trip rate was estimated by the regression equation.

The scale is recommended to be empirically set based on the spatial behavior of the study variable, according to minimum distances between the centroids of the source unit areas (TAZs in the present case study) and according to variographic experiments. The authors recommend the process for disaggregating data (1–6) to be repeated until sufficient configurations are explored to reproduce the phenomenon. For the present research, eight sets of synthetic data configurations were set up. At the end of the deconvolution process, the respective semivariogram of each configuration is calculated and the feasibility of using the average experimental semivariogram is assessed, subject to the variability of one another.

3.2. Method

Fig. 5 introduces the flowchart, which depicts the steps followed in the current paper. Letters (a) to (d) present the information assessed from both datasets. The steps are described by the Roman numerals (i) to (v), followed by the SGS. Steps (i) to (iii) are embedded in the semivariogram deconvolution, previously outlined in Fig. 4. Hence, the hallmark of the research approach lies in steps (i) to (iii) and the successive SGS (Proposed Method). This paper recommends steps iv, v and the subsequent geostatistical simulation (Comparative Method) to validate the Proposed Method.

In short, the following methods were applied:

1) Linear regression – to calibrate a model of the transit trip rate as a function of the low-income household rate.
2) Calculation and modeling of the empirical semivariogram – to analyze if the spatial structure of the estimated transit trip rate (based on census data) was similar to the real transit rate (OD Survey) at different aggregation grids. The semivariogram calculation for lower levels of aggregation using census data corresponds to an alternative deconvolution method that does not depend on disaggregated data regarding the interest variable.
3) Sequential Gaussian Simulation – to obtain different scenarios for the spatial distribution of estimated and real transit trip rates at different aggregation grids; calculation of the associated uncertainty (variance and confidence intervals).

The following conditions are required for the variable considered for the SGS: 1) normal distribution with mean 0 and variance 1; and 2) multigaussian assumption, which defines that each linear combination of the variable is distributed by a normal distribution. The kriging system is embedded in the calculation of each realization at the SGS, according to Equation (1).

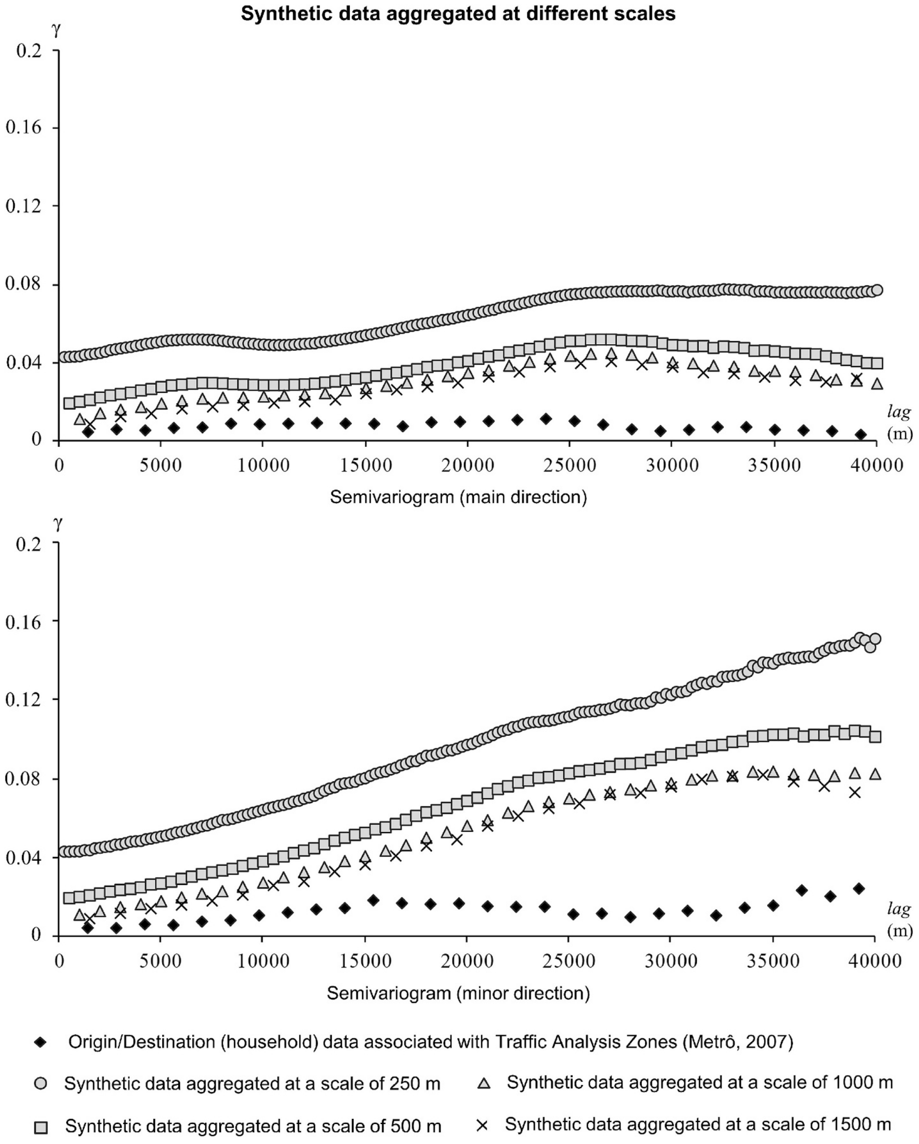

Fig. 6. Deconvoluted semivariograms for the transit trip rate using synthetic data.

$$z^{(l)}(x_0) = z^*(x_0) + \sigma\varepsilon \quad (1)$$

The index $z^{(l)}(x_0)$ is the simulated variable at x_0, $z^*(x_0)$ is the estimated variable at x_0 using kriging; ε is a random component between 0 and 1. Considering different kriging techniques, Deutsch and Journel (1998) suggest that researchers adopt Simple Kriging (SK), as it ensures the reproduction of the semivariogram. In terms of the results, the average value of the realizations at each location (also known as e-type) approximates the estimated value achieved by kriging methods. The result of the variances between the realizations tends to be similar to the kriging variance (Chilès and Delfiner, 1999).

The SGS can be carried out using aggregated travel data associated with the centroid of each TAZ and the theoretical semivariograms (also associated with the TAZ centroids). However, in such a case, despite the MAUP, only aggregated information for the population distribution and the spatial structure would be considered as input. The variographic deconvolution, in turn, enables the incorporation of spatial structure associated with more disaggregated data, using theoretical semivariograms for each analyzed regular support.

Finally, the results are compared using four criteria. The first criterion is the visual inspection of the spatial results processed by the SGS and represented by the average of 500 realizations (e-type), the confidence interval, median and variance. The second criterion refers to analyzing univariate statistical measures. The third criterion tests whether the distributions of both average simulations are similar, using non-parametric hypothesis testing. The last criterion aims to compare the performance of both methods (Proposed *versus* Comparative).

4. Results and discussions

4.1. Data processing, variographic analysis and deconvolution

The inference of travel demand information (based on socioeconomic microdata) was derived from the following equation: $T = 1.056^*H$, where T is the rate between produced trips by transit and total produced trips, and H is the rate between low-income households and the total number of households inside each regular area. The regression model presented a determination coefficient of 0.8 and a statistically significant coefficient for the independent variable H (sig = 0.00). While the regression model was calibrated using the O/D Survey data, the previous equation was used to calculate the study variable T using the census synthetic data.

After estimating T for different aggregation units, a variographic analysis of this variable was conducted, as shown in Fig. 6. The

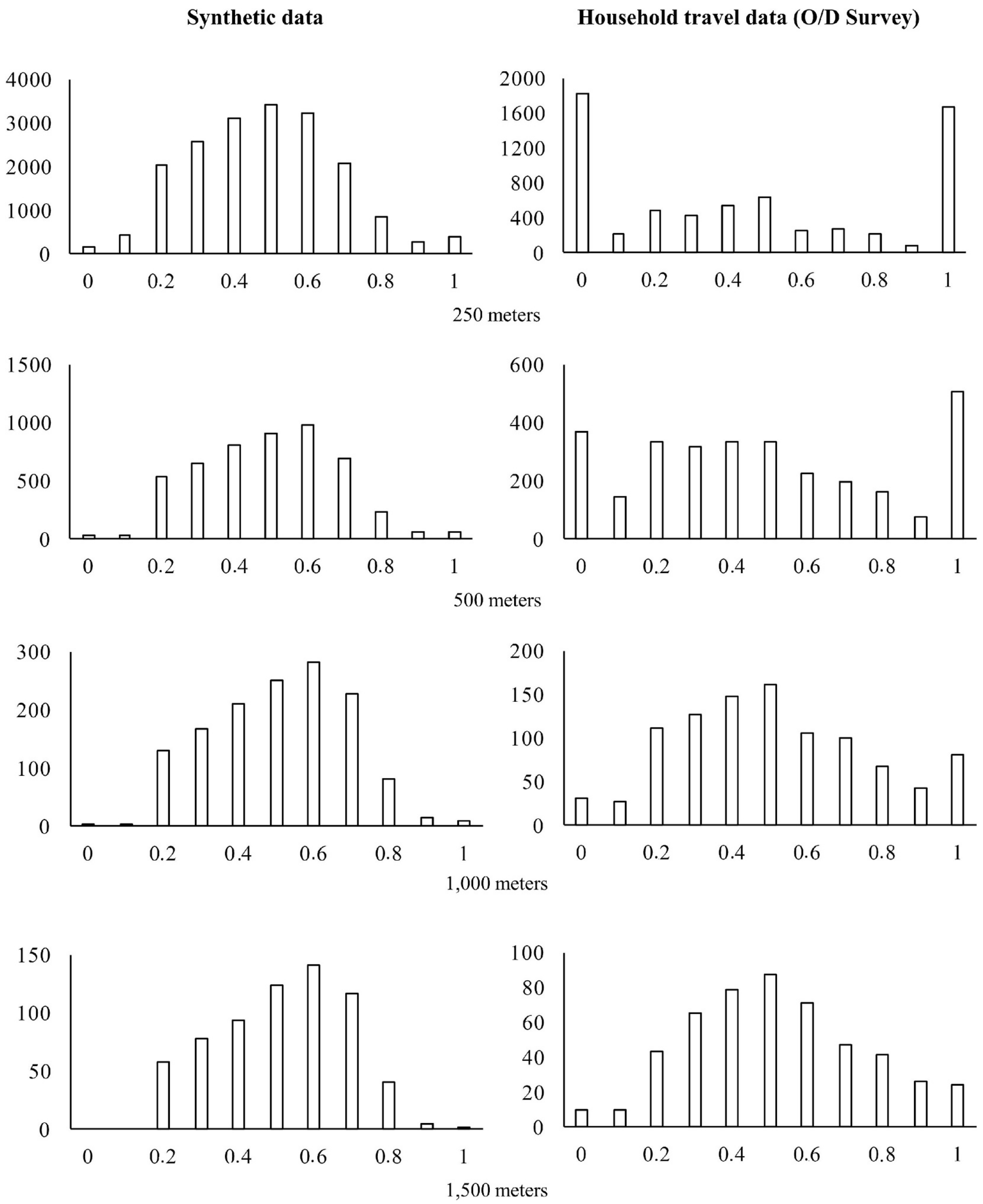

Fig. 7. Histograms for the O/D and the synthetic data at different scales.

omnidirectional semivariograms for the eight sets of synthetic data presented a similar structure, corroborating the assumption of using the average semivariograms for the main and minor directions of data continuity (N-S and E-W, respectively). As the Euclidean distance between TAZs is 1,400 m, it can be concluded that using coarse spatial supports does not provide disaggregated outcomes. However, this study includes a coarse-scale of 1,500 m to compare the results.

The variographic analysis of the deconvoluted semivariograms leads to the conclusion that the more disaggregated the scale, the less the tendency for spatial stationarity. Thus, researchers may consider investigating particular spatial behavior for different scales while exploring geostatistical approaches for travel data. For the present case study, observations 250 m apart (or closer than 250 m) indicate a trend for linear theoretical semivariograms, showing evidence of the stop criteria for downscaling as it violates the intrinsic hypothesis of second-order stationarity, postulated in the formal geostatistical theory.

The disaggregated data to be used in the Comparative Simulation, *i. e.*, household travel data, is then aggregated within unit areas of 250, 500, 1000 and 1500 m. The aggregation of household-related data may lead to distributions with high levels of null values, as a unique observation value may represent its entire respective cell, especially considering that the sample size of the O/D Survey is smaller than the census sample. Fig. 7 shows the histograms for the household data (O/D Survey) and the synthetic data at different scales, using the transit trip rate variable. The household aggregated data distribution tends to normality at coarser scales. In contrast, when considering synthetic data, the more disaggregated, the more similar to a normal distribution.

Fig. 8 exhibits the semivariograms for the household-related data, whose variances between pairs of observations are higher than those associated with TAZs, albeit both are based on the O/D Survey. Comparing Fig. 8 to Fig. 6, we can see that synthetic data, despite encompassing higher variability between pairs of observations than O/D

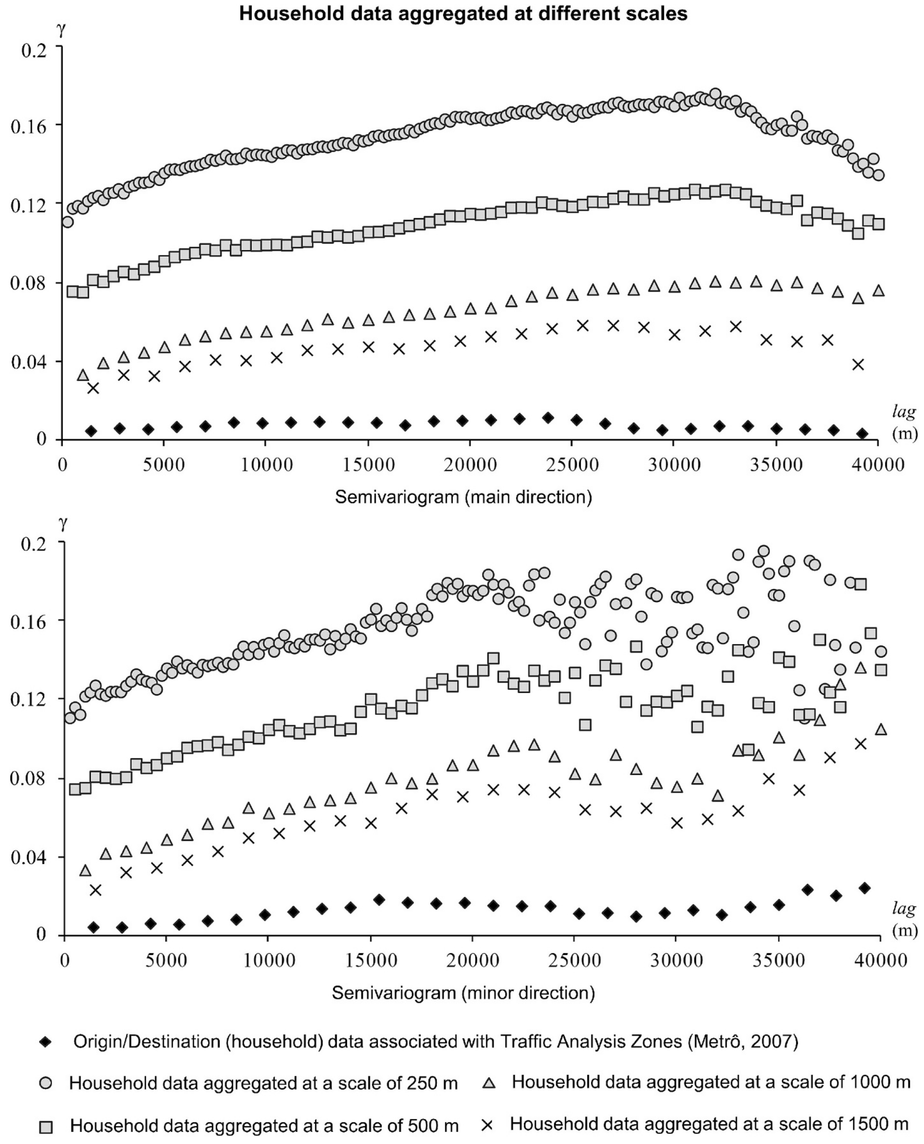

Fig. 8. Semivariograms for the transit trip rate from the O/D data.

Table 1
Descriptive statistics for the utilized datasets.

Support (m)	Data	# Records*	Statistical measures						
			Average	Std. Deviation	Min.	First Quartile	Median	Third Quartile	Max.
250	Household	6,619	0.44	0.39	0	0	0.35	1	1
	Synthetic	18,563	0.43	0.20	0	0.28	0.43	0.57	1
500	Household	2,999	0.46	0.33	0	0.18	0.41	0.70	1
	Synthetic	4,990	0.44	0.19	0	0.30	0.45	0.58	1
1000	Household	1,002	0.47	0.26	0	0.26	0.44	0.64	1
	Synthetic	1,378	0.46	0.18	0	0.32	0.47	0.60	1
1500	Household	503	0.47	0.24	0	0.29	0.46	0.62	1
	Synthetic	659	0.47	0.17	0.10	0.33	0.48	0.60	1
irregular (MAUP)	Household per TAZ	308	0.30	0.10	0.07	0.23	0.30	0.37	0.62

* Non-zero cells.

data aggregated at TAZs, do not hold the same level of variance as the household-related information. Despite the greater level of information that causes higher variability at refined scales in household-related data, it can be recalled that such data entail a limited sample number, as they depend on costly O/D surveys, causing several cells to have no information in spatial models. These mentioned trade-offs must be taken into account when selecting adequate data and/or scale.

Table 1 summarizes the descriptive statistics of the groups of information utilized in this study.

The processed household datasets, derived from the O/D Survey (Metrô, 2007), contain fewer records compared to the synthetic data. In addition, the range of values from the irregular support (associated with

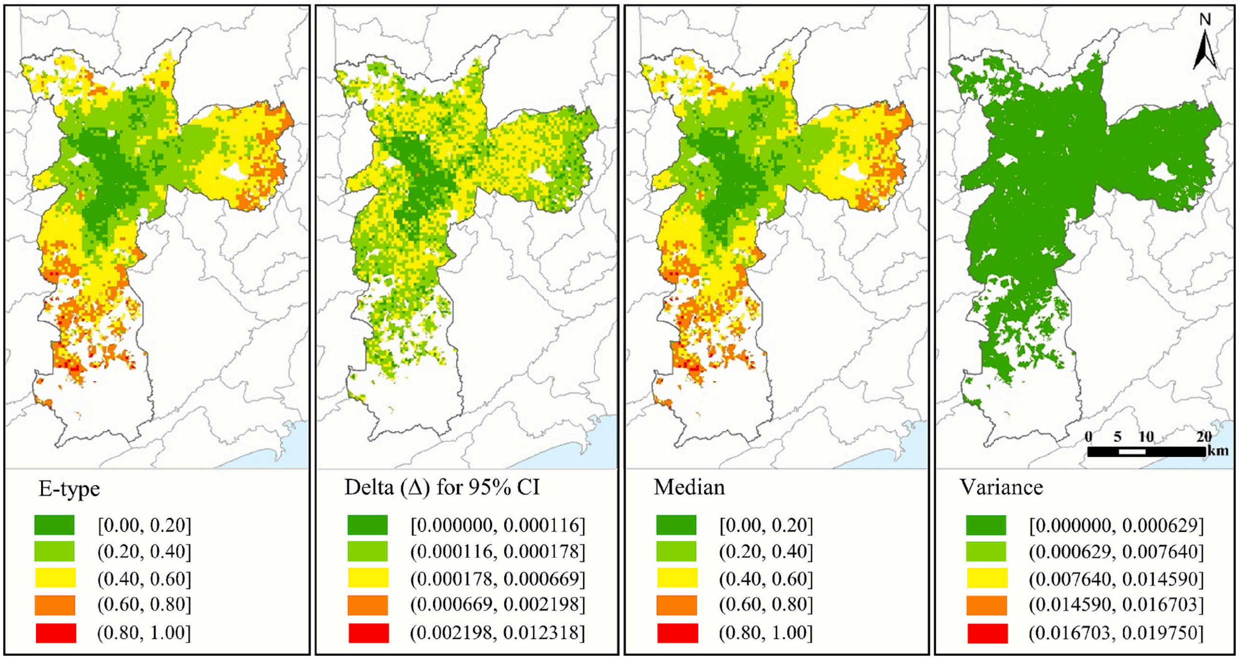

Fig. 9. Statistical maps for the Proposed Simulation using a 500-meter scale.

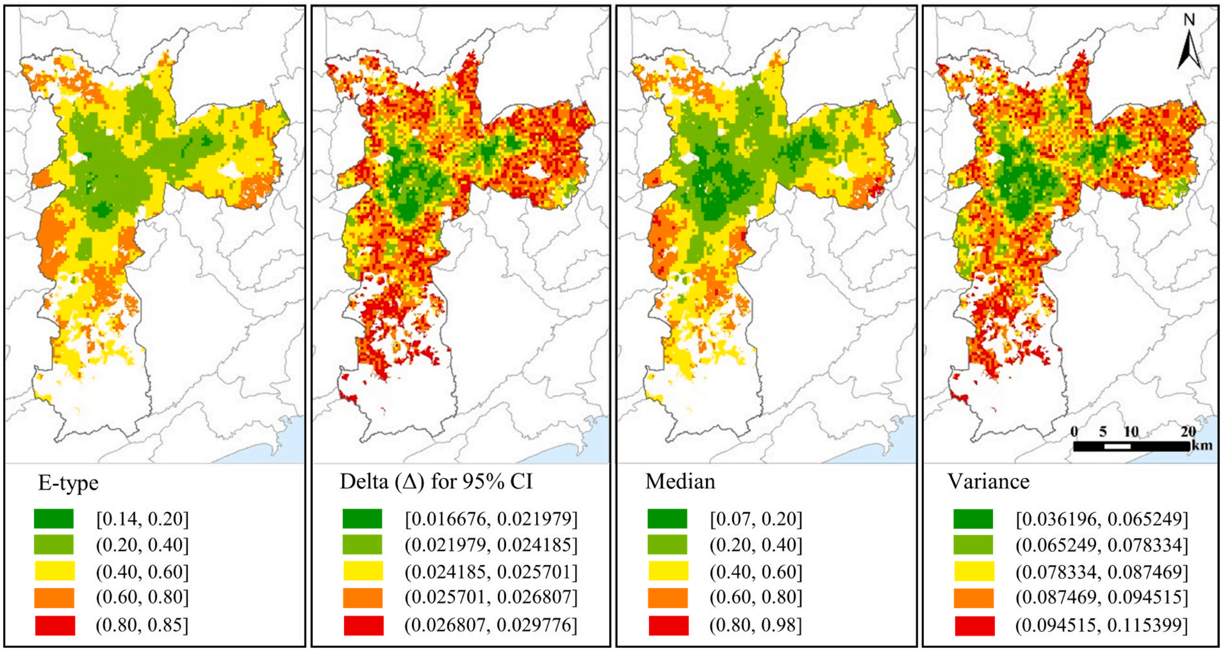

Fig. 10. Statistical maps for the Comparative Simulation using a 500-meter scale.

TAZs) does not reach the attainable range for transit rates (from 0 to 1), instead it provides smooth values for each unit area. Hence, the Proposed Method intends to use (as input to the Simulation) the Synthetic Data, as it not only covers a wider sample area but is also able to generate samples with a higher level of detail, considering the range of transit rate from 0 to 1.

4.2. Geostatistical simulation and comparison analysis

The Proposed Simulation consists of using the spatial structure of each scale (through the deconvoluted semivariograms) and the synthetic data. Fig. 9 presents the average, confidence interval, median and variance of 500 realizations (simulated scenarios), considering a 500-meter scale.

There is a uniform variance trend for the proposed simulation in populated areas. Although a few non-populated cells (not represented in the map) may cover higher variances, such values were not found to be higher than 2 %, which may be seen as a major drawback for the current case study. A comparison with the other supports indicates, however, that the more downscaled the map, the less uniform the spatial variance distribution. The method was also applied to the Comparative Simulation framework.

Fig. 10 gathers statistical maps for the Comparative Simulation at 500 m. The Proposed Simulation resulted in maps with lower variances

Table 2
Descriptive statistics for the simulation approaches.

Scale (m)	Simulation Approach	# Records*	Statistical Measure						
			Average	Std. Deviation	Min.	First Quartile	Median	Third Quartile	Max.
250	Proposed	18,944	0.43	0.20	0	0.28	0.43	0.57	1
	Comparative	18,944	0.49	0.39	0	0.09	0.49	0.96	1
500	Proposed	5,076	0.45	0.19	0	0.30	0.45	0.58	1
	Comparative	5,076	0.49	0.32	0	0.22	0.47	0.76	1
1000	Proposed	1,398	0.46	0.18	0	0.32	0.47	0.60	1
	Comparative	1,398	0.50	0.25	0	0.31	0.48	0.67	1
1500	Proposed	663	0.47	0.17	0.10	0.33	0.49	0.60	1
	Comparative	663	0.47	0.17	0.10	0.33	0.49	0.60	1

[*] Non-zero cells.

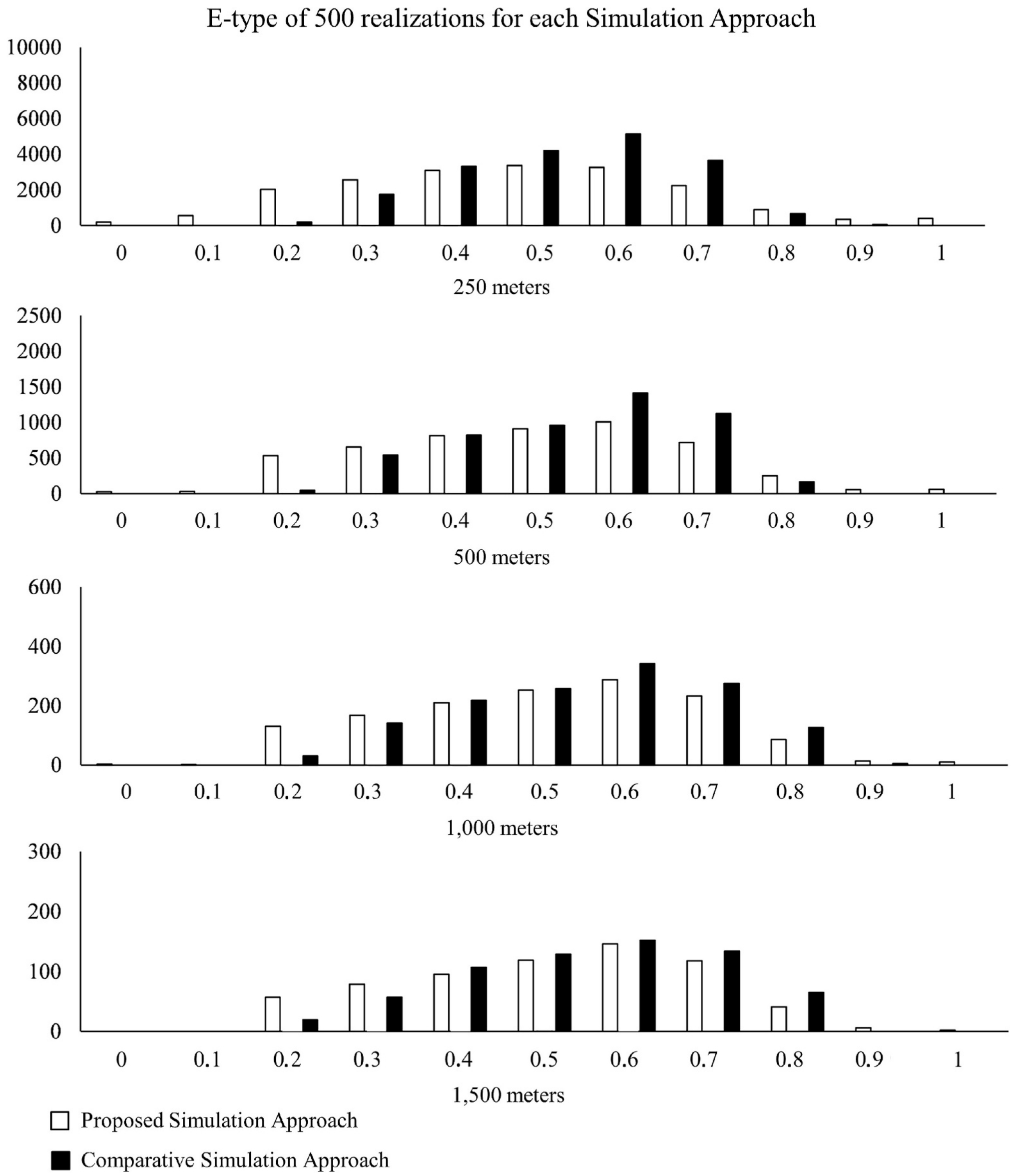

Fig. 11. E-type histograms for the Proposed and the Comparative Simulations.

and confidence intervals in contrast to the Comparative Simulation. This outcome was expected as the semivariograms based on the O/D Survey showed higher variances than those from synthetic data.

Table 2 outlines the descriptive statistics from the simulation approaches at each scale.

In general, it can be observed that that the univariate statistical measures of both approaches tend to become more similar at more aggregated scales. Fig. 11 sets out the distribution of average values (e-type) for both approaches.

Contrary to the expectations of similar distributions from Fig. 11, non-parametric tests (Mann-Whitney and Kolmogorov-Smirnov) have led to rejecting the null hypothesis, suggesting that both average

Table 3
Comparative analysis between both simulation approaches.

Scale (m)	MD	MAD	MSE	RMSE	MPE	MAPE	r
250	0.054	0.139	0.031	0.177	0.107	0.302	0.553
500	0.046	0.113	0.021	0.144	0.098	0.244	0.684
1,000	0.040	0.104	0.018	0.134	0.255	0.077	0.806
1,500	0.032	0.100	0.017	0.129	0.059	0.209	0.710

MD – Mean Deviation; MAD – Mean Absolute Deviation; MSE – Mean Square Error; RMSE – Root Mean Square Error; MPE – Mean Percentage Error; MAPE – Mean Absolute Percentage Error; r – Correlation.

distributions are not equal. Finally, Table 3 shows the comparative analysis between the approaches for each scale.

Table 3 shows no association between the disaggregation level and the correlation, despite all the measures resulting in satisfactory values. The highest correlation was 0.806 for the comparison between approaches at 1000 m-scale. Regardless of the lack of pattern in terms of correlation, coarser scales are more accurate, considering the RMSE. As the MD and MPE can detect eventual trends for over- or underestimations, the results showed that there is no association between the bias and the respective scale. Hence, the selection of the scale depends on the purpose of the data disaggregation.

4.3. Key findings

The key findings of the paper are outlined as follows:

- The spatial structure of travel data might be inferred using socioeconomic microdata and a calibrated regression model, instead of traditional O/D (household) information.
- By using widely available socioeconomic data, the proposed method allows for inferring the spatial pattern of travel demand data at different levels of aggregation.
- The alternative deconvolution procedure validates the spatial structure of disaggregated data, following a regular scale rather than the irregular unit areas of the input dataset.
- Geostatistical simulation tools provided means of creating different scenarios, with respective maps of confidence intervals and variances.

5. Conclusions

The declining public investment in accurate origin–destination (O/D) surveys in Brazil has sparked significant discussions about developing alternative methods for generating travel data using fewer resources. The current paper was formulated based on this context of unavailable travel disaggregated data, as well as on the lack of traditional transportation models that account for the spatial dependence of variables.

The proposed method can be useful for municipalities that have no information on travel demand data. If one is interested in the spatial pattern of transit trip production, only the socioeconomic aspect is required to calculate estimates based on the equation given in Subsection 4.1. Equations for estimating other variables of interest can be transferred from cities with similar characteristics which have an OD Survey. Other requirements would be the following: knowledge of dealing with spatial data in a geographic information system such as randomly assigning geographic coordinates to households sampled in a census; and knowledge of basic statistics (linear regression) and geostatistics. In turn, the modeling steps can be conducted at no cost by using opensource software or a free interface (R Core Team, 2021; Pebesma, 2004; Ribeiro Jr. and Diggle, 2016; Remy et al., 2009; Deutsch and Journel, 1998).

Despite resulting in a uniform variance map (for one of the most disaggregated scales), the proposed framework sheds new light on spatial simulation methods to travel demand analysis. Applications of the proposed method to other variables of interest are highly recommended, mainly in the social sciences, which commonly deal with human behavior. Other recommended topics include making a comparison between the semivariogram deconvolution method proposed in this paper and the one proposed by Journel and Huijbregts (1978); incorporating temporal effects; testing other types of regression modeling; and applying SGS to stop-level ridership data to support the operation planning considering on-peak and off-peak hours.

CRediT authorship contribution statement

Anabele Lindner: Investigation, Writing – original draft, Visualization, Methodology, Data curation, Writing – review & editing, Conceptualization, Formal analysis. **Cira Souza Pitombo:** Supervision, Conceptualization, Writing – review & editing. **Samuel de França Marques:** Writing – review & editing.

Funding

This research was supported by the National Council of Technological and Scientific Development (CNPq – grants #403508/2023-1, #151115/2019-2 and # 305973/2023-1, Brazil) and the São Paulo Research Foundation (FAPESP – grant #2019/12054-4, Brazil).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets are open source and are available at: http://www.metro.sp.gov.br/metro/numeros-pesquisa/pesquisa-origem-destino-2007.aspx https://www.ibge.gov.br

References

ABEP - Brazilian Association of Research Companies (2010). Critério Brasil 2010. http://www.abep.org/criterio-brasil (last accessed on 03.03.2020).

Cervero, R., Radisch, C., 1996. Pedestrian versus automobile oriented neighborhoods. Transp. Policy 3, 127–141. https://doi.org/10.1016/0967-070X(96)00016-9.

Chen, X.M., He, X., Xiong, C., Zhang, L., 2015. A Bayesian stochastic kriging metamodel for simultaneous optimization of travel behavioral responses and traffic management. In: Transportation Research Board. 94th Annual Meeting Compendium of Papers. Washington. https://doi.org/10.1016/j.trpro.2015.06.056.

Chilès, J.P., Delfiner, P., 1999. Geostatistics Modeling Spatial Uncertainty. John Wiley & Sons, New York, p. 695p.

Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, p. 370.

Dugundji, E., Walker, J., 2005. Discrete choice with social and spatial network interdependencies: an empirical example using mixed generalized extreme value models with field and panel effects. Transp. Res. Rec. 1921, 70–78. https://doi.org/10.3141/1921-09.

Flowerdew, R., Green, M., 1993. Developments in areal interpolation methods and GIS. In: Geographic Information Systems, Spatial Modelling and Policy Evaluation. Springer, Berlin, Heidelberg, pp. 73–84. https://doi.org/10.1007/978-3-642-77500-0_5.

Gomes, V.A., Pitombo, C.S., Rocha, S.S., Salgueiro, A.R., 2016. Kriging geostatistical methods for travel mode choice: a spatial data analysis to travel demand forecasting. Open J. Stat. 6 (3), 514. https://doi.org/10.4236/ojs.2016.63044.

Goodchild, M.F., Anselin, L., Deichmann, U., 1993. A framework for the areal interpolation of socioeconomic data. Environ Plan A 25 (3), 383–397. https://doi.org/10.1068/a250383.

Goovaerts, P., 1997. Geostatistics for natural resources evaluation. In: Applied Geostatistics. Oxford University Press on Demand, New York, p. 496. https://doi.org/10.1080/00401706.2000.10485733.

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103 (1), 3–26. https://doi.org/10.1016/S0016-7061(01)00067-2.

Goovaerts, P., 2008. Kriging and semivariogram deconvolution in the presence of irregular geographical units. Math. Geosci. 40 (1), 101–128. https://doi.org/10.1007/s11004-007-9129-1.

IBGE - Brazilian Institute of Geography and Statistics (2013) *Metodologia do Censo Demográfico 2010*. Série Relatórios Metodológicos, 41. https://ftp.ibge.gov.br/

Censos/Censo_Demografico_2010/metodologia/metodologia_censo_dem_2010.pdf (last accessed on 25.03.2024).

IBGE - Brazilian Institute of Geography and Statistics (2010) *Census and Microdata*. https://www.ibge.gov.br (last accessed on 03.03.2020).

Journel, A.G., Huijbregts, C.J., 1978. Mining geostatistics. Reprinted (1991). Academic PressThe Blackburn Press, United States of America, p. 600.

Kaygisiz, Ö., Düzgün, Ş., Yildiz, A., Senbil, M., 2015. Spatio-temporal accident analysis for accident prevention in relation to behavioral factors in driving: the case of South Anatolian Motorway. Transport. Res. F: Traffic Psychol. Behav. 33, 128–140. https://doi.org/10.1016/j.trf.2015.07.002.

Kitamura, R., Mokhtarian, P.L., Laidet, L., 1997. A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. Transportation 24, 125–158. https://doi.org/10.1023/A:1017959825565.

Kyriakidis, P.C., 2004. A geostatistical framework for area-to-point spatial interpolation. Geogr. Anal. 36 (3), 259–289. https://doi.org/10.1111/j.1538-4632.2004.tb01135.x.

Lindner, A., Pitombo, C.S., Assirati, L., Pedreira Junior, J.U., Salgueiro, A.R., 2021. Estimation of travel mode choice using Geostatistics: a Brazilian case study. Rev. Bras. Cartogr. 73, 182–197. https://doi.org/10.14393/rbcv73n1-54210.

Lindner, A., Pitombo, C.S., 2018. A conjoint approach of spatial statistics and a traditional method for travel mode choice issues. J. Geovisual. Spatial Anal. 2 (1). https://doi.org/10.1007/s41651-017-0008-0.

Lindner, A., Pitombo, C.S., Rocha, S.S., Quintanilha, J.A., 2016. Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study. Geo-spatial Inf. Sci. 19 (4), 245–254. https://doi.org/10.1080/10095020.2016.1260811.

Lindner, A., Pitombo, C.S., 2019. Sequential Gaussian simulation as a promising tool in travel demand modeling. J. Geovisual. Spat. Anal. 3 (2), 15. https://doi.org/10.1007/s41651-019-0038.

Marques, S.D.F., Pitombo, C.S., 2020. Intersecting geostatistics with transport demand modeling: a bibliographic survey. Rev. Bras. Cartogr. 72, 1028–1050. https://doi.org/10.14393/rbcv72nespecial50anos-56467.

Marques, S.D.F., Pitombo, C.S., 2021a. Ridership estimation along bus transit lines based on kriging: comparative analysis between network and euclidean distances. J. Geovisual. Spat. Anal. 5 (1), 7. https://doi.org/10.1007/s41651-021-00075-w.

Marques, S.D.F., Pitombo, C.S., 2021b. Applying multivariate Geostatistics for transit ridership modeling at the bus stop level. Boletim De Ciências Geodésicas 27 (2). https://doi.org/10.1590/1982-2170-2020-0069.

Marques, S.D.F., Pitombo, C.S., 2023. Transit ridership modeling at the bus stop level: comparison of approaches focusing on count and spatially dependent data. Appl. Spat. Anal. Policy 16, 277–313. https://doi.org/10.1007/s12061-022-09482-y.

Marques, S.D.F., Pitombo, C.S., Gómez-Hernández, J.J., 2024. Spatial modeling of travel demand accounting for multicollinearity and different sampling strategies: a stop-level case study. J. Adv. Transp. 2024 (7967141), 2024.

Matheron, G., 1963. Principles of geostatistics. Econ. Geol. 58, 1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246.

Metrô - São Paulo Metropolitan Company (2008). Síntese das Informações Pesquisa Domiciliar. Diretoria de Planejamento e Expansão dos Transportes Metropolitanos – DM. https://transparencia.metrosp.com.br/sites/default/files/S%C3%8DNTESE_OD2007_abr09.pdf (last accessed on 25.03.2024).

Metrô - São Paulo Metropolitan Company (2007). Origin-Destination Survey 2007 - São Paulo Metropolitan Area: Summary of information. http://www.metro.sp.gov.br/metro/numeros-pesquisa/pesquisa-origem-destino-2007.aspx (last accessed on 05.10.2019).

Miura, H., 2010. A study of travel time prediction using universal kriging. Sociedad De Estadística e Investigación Operativa. 257–270. https://doi.org/10.1007/s11750-009-0103-6.

Páez, A., López, F.A., Ruiz, M., Morency, C., 2013. Development of an indicator to assess the spatial fit of discrete choice models. Transp. Res. B Methodol. 56, 217–233. https://doi.org/10.1016/j.trb.2013.08.009.

Páez, A., Scott, D.M., 2005. Spatial statistics for urban analysis: a review of techniques with examples. GeoJournal 61 (1), 53–67. https://doi.org/10.1007/s10708-005-0877-5.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012.

Pitombo, C.S., Salgueiro, A.R., Costa, A.S.G., Isler, C.A., 2015. A Two-step method for mode choice estimation with socioeconomic and spatial information. Spatial Stat. 11, 45–64. https://doi.org/10.1016/j.spasta.2014.12.002.

R Core Team, 2021. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.r-project.org/> (accessed 06/03/2023.

Remy, N., Boucher, A., 2009. Wu J (2009). A User's Guide, Cambridge University Press, Cambridge, Applied Geostatistics with SGeMS.

Ribeiro Jr. PJ, Diggle PJ (2016). *geoR: Analysis of Geostatistical Data*. R package version 1.7-5.2. Available at: <https://CRAN.R-project.org/package=geoR> (accessed 06/03/2023).

Rocha, S.S., Lindner, A., Pitombo, C.S., 2017. Proposal of a geostatistical procedure for transportation planning field. Boletim De Ciências Geodésicas. 23 (4), 636–653. https://doi.org/10.1590/s1982-21702017000400042.

SPTrans (2018). Daily report on the number of carried passengers. http://www.prefeitura.sp.gov.br/ (last accessed on 05.10.2019).

Verly, G., 1993. Sequential Gaussian simulation: a Monte Carlo method for generating models of porosity and permeability. In: Generation, Accumulation and production of Europe's Hydrocarbons III. Springer, Berlin, Heidelberg, pp. 345–356. https://doi.org/10.1007/978-3-642-77859-9_28.

Xie, Z., Yan, J., 2013. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. J. Transp. Geogr. 31, 64–71. https://doi.org/10.1016/j.jtrangeo.2013.05.009.

Yamada, I., Thill, J.C., 2004. Comparison of planar and network K-functions in traffic accident analysis. J. Transp. Geogr. 12 (2), 149–158. https://doi.org/10.1016/j.jtrangeo.2003.10.006.

Yoon, S.Y., Ravulaparthy, S.K., Goulias, K.G., 2014. Dynamic diurnal social taxonomy of urban environments using data from a geocoded time use activity-travel diary and point-based business establishment inventory. Transp. Res. A Policy Pract. 68, 3–17. https://doi.org/10.1016/j.tra.2014.01.004.