Sequential Gaussian Simulation as a Promising Tool in Travel Demand Modeling

Anabele Lindner[1] · Cira Souza Pitombo[1]

Abstract

The aim of this article is to propose an alternative approach to disaggregate data using sequential Gaussian simulation, considering the difficulty in obtaining disaggregated data and the fact that these data are more interesting for transportation planning policies. The study area is the São Paulo Metropolitan Area (Brazil), and the 2007 dataset is associated to the number of transit trips per each traffic analysis zone. The main advantages of the proposed method when compared to traditional simulation methods for travel demand are (1) using less information, (2) including the spatial association of the variables, (3) mapping the simulated value, (4) estimating values in non-sampled locations, and (5) mapping uncertainty parameters, such as conditional variances and confidence interval. The main interest of this research for urban planning policies has been shown with the advantage of mapping critical scenarios for travel demand using a spatially correlated variable. The benefit of providing a map of transit trips associated to a disaggregated unit area, originated within an aggregated dataset, supports decision makers to yield more efficient public transportation systems considering significant cost reduction.

Keywords Transit trips · Public transportation · Simulation · Geostatistics · Spatial statistics · Change of support

Introduction

Transport planning studies are based on forecasting future travel demand. Analysts use travel demand models to evaluate the sensitivities of demand to operational variables, such as costs, charged prices, fleet, and frequency of public transport. Transport planners also assess models to predict whether new facilities should be implemented or if there should be an attempt to better operate the existing ones (Kitamura and Fujii 1998; Ortúzar and Willumsen 2011). The need of using these models is unquestionable as they aim to make the urban mobility plan more efficient.

The classic approach for travel demand is the four-stage model, also known as the trip-based model. Its basic unit addresses origin-destination pairs (commonly) at an aggregate form while it neglects the heterogeneity among different individuals (Zhang and Levinson 2004). This method was outlined as a result of practices in the 1960s (Ortúzar and

Willumsen 2011) given the rapid growth of urban population and motorization.

Evidently, it is reasonable to adapt former approaches to suit present conditions. Planners have made efforts to develop pioneer methods that overcome shortcomings seen in past models, i.e., the fact that trip-based models present unrealistic behavioral characteristics (for further information concerning previous research on the mentioned issue, the authors recommend reading Kitamura (1988), Ben-Akiva and Bowman (1998)). To address this, human behavior was represented at an individual level (Moeckel et al. 2003), especially due to recent developments in computing technology and increased data availability (Buliung and Kanaroglou 2007), which allowed analysts to provide refined outcomes. This framework, recognized as an activity-based model—and first discussed by Recker et al. (1986a, 1986b))—requires disaggregated data that are not often available. Yagi and Mohammadian (2010) point out that activity-based models are more accepted in developed countries rather than in developing countries. On one hand, this is due to the great amount of both dis- and aggregated information demanded in activity-based models. On the other hand, this is due to the need for in-depth econometric knowledge, as well as complex computational processes. Therefore, it can be said that the reasons for developed countries accepting more activity models is due to

✉ Anabele Lindner
anabele@usp.br

[1] Department of Transportation Engineering, São Carlos School of Engineering, University of São Paulo, Av. Trabalhador São-Carlense, 400, São Carlos, São Paulo 13566-590, Brazil

the fact that they have more resources to invest in the criteria mentioned. Despite the cost issues, data availability is also linked to confidentiality matters, that is, even if the individual data exist; in most cases, they are not accessible.

Activity-based models are applied using (1) econometric based applications, (2) mathematical programming frameworks, (3) computational process models, and (4) (micro) simulation approaches (McNally and Rindt 2007). The spotlight is on microsimulation as this article deals with disaggregation in a large study area. Besides the benefits, microsimulation models require a large amount of information to be able to predict satisfactory outcomes. Among many different approaches of population synthesizers needed in microsimulation, one aspect stands out: to date, none of the approaches recognizes the spatial correlation of each variable as an important input to reproduce travel behavior. In addition, despite the observed advances in microsimulation associated with using individual travel behavior data and land use data (Landis and Zhang 1998; Arentze and Timmermans 2000; Waddell 2000; Hunt et al. 2001; Moeckel et al. 2003; Salvini and Miller 2005; Pendyala et al. 2012), the spatial association of data was not addressed. Moreover, it should be mentioned that travel and socioeconomic variables may be spatially correlated (Lindner et al. 2016; Rocha et al. 2017; Lindner and Pitombo 2018). That is, by taking into account spatial association, one could optimize the required input variables in travel demand models.

Considering that it is difficult to obtain disaggregated data and that they are more interesting for transport planning policies, the importance of this article is to contribute by proposing an alternative approach in travel demand field to disaggregate data using sequential Gaussian simulation. The main advantages of the proposed method when compared to traditional methods are (1) using less information, (2) including the spatial association of the variables, (3) mapping the simulated values, (4) estimating values in non-sampled locations, and (5) mapping uncertainty parameters, such as conditional variances and confidence interval.

Activity-based microsimulation models are able to disaggregate multiple variables, consider their correlations, and create future scenarios. The method proposed here shows a different perspective that considers a single variable (along with its spatial information) and does not cover dynamic simulations. The authors do not intend to replace renowned and well-accepted travel demand models, but rather propose geostatistical simulation concepts that are conventionally used in natural sciences to be considered in social science issues.

This research paper is divided into six sections. The "Travel Demand Modeling: Simulation and Spatial Analysis" section presents the literature review on travel demand modeling, specifically concerning microsimulation and spatial analysis. The "Geostatistics: Understanding the Sequential Gaussian Simulation" section covers basic geostatistical concepts required to fully understand the proposed framework. The "Materials and Method" section outlines the proposed method and presents the study area and dataset. The "Results and Discussions" section shows the results. Finally, the conclusions are drawn in the "Conclusions" section.

Travel demand modeling: Simulation and spatial analysis

Ballas et al. (2005) describe the steps for microsimulation as (1) the construction of a disaggregated dataset (when not available), (2) random sampling from the sample created in the first step to generate a synthetic population, (3) what-if simulations to evaluate alternative scenarios, and (4) dynamic modeling to evaluate future scenarios or update a disaggregated dataset.

Microsimulation models deal with the change of support in a downscaling process and are widely applied to replicate travel behavior patterns. Given that microsimulation models require disaggregated data, researchers have been using population synthesizers to create disaggregated data associated to households and individuals (Beckman et al. 1996; Moeckel et al. 2003; Arentze et al. 2007; Guo and Bhat 2007; Müller and Axhausen 2011; Barthelemy and Toint 2013; Farooq et al. 2013). Rahman (2009) classifies the techniques for creating synthetic microdata as: reweighting methods and synthetic reconstruction. Reweighting is carried out by generating data using existing survey microdata rather than artificially creating them (Hermes and Poulsen 2012). On the other hand, synthetic reconstruction is the most familiar approach to transport planners and the most long-standing when compared to reweighting techniques. It attempts to reproduce all known constraints by random sampling using a set of conditional probabilities. Synthetic reconstruction methods comprise data matching and the iterative proportional fitting (IPF). Data matching is performed by pairing datasets from different sources. However, this approach may not be convenient as the identification code of a variable, which is used to match datasets, may not be released due to confidentiality issues. IPF is a straightforward method to allocate individuals to zones. It calculates the maximum likelihood for each combination of zone and individual, which is described in the weight matrix (Lovelace and Dumont 2016).

Microsimulation techniques involve the definition of "agents," i.e., populations of individuals or households, as well as their interrelations, and are therefore the basis for agent-based methods (Balmer et al. 1985). Regardless of specific denominations, traditional models for disaggregation in travel demand mainly involve a large amount of socioeconomic and travel variables and do not take into consideration spatial associations.

However, various transportation planning researchers have suggested using spatial factors in deterministic travel demand models, especially when observing technological advances and how easy it is to obtain georeferenced databases (Bhat and Zhao 2002; Miyamoto et al. 2004; Ben-Akiva et al. 2004; Páez and Scott 2005; Páez 2007; Bhat and Sener 2009; Antipova et al. 2011; Kamruzzaman et al. 2011; Morency et al. 2011; Páez et al. 2013).

It is important to emphasize two main drawbacks of adopting the aforementioned techniques concerning travel demand issues. The first is that they do not take into account the uncertainty involved in the development of future states as they are conducted by deterministic models. The second point is that the nature of the method is related to exploratory analyses (Kamruzzaman et al. 2011; Páez et al. 2013) or to a spatial point estimation—in spatial regression (Ben-Akiva et al. 2004; Páez and Scott 2005), autocorrelated models (Miyamoto et al. 2004), and different logit frameworks based on spatial factors (Bhat and Zhao 2002; Bhat and Sener 2009; Antipova et al. 2011), for instance.

The limitation of point-limited estimates can be tackled by a recent approach applied to the transportation field: geostatistics, which enables performing estimated maps and calculating values in non-sampled points (or areas). Therefore, geostatistical techniques are able to provide confirmatory analysis. Researchers have recently explored and demonstrated the benefits of using geostatistics, which is a well-established framework in natural sciences (Lee et al. 2007; Pearce et al. 2009; Orton et al. 2016). In the transportation planning field, some applications have been developed in traffic modeling cases (Mazzella et al. 2011; Ciuffo et al. 2011; Zou et al. 2012; Tong et al. 2013; Song et al. 2018), traffic accidents (Gundogdu 2014; Molla et al. 2014; Manepalli and Bham 2016), and travel demand (Pitombo et al. 2015a, 2015b; Lindner et al. 2016; Gomes et al. 2016; Rocha et al. 2017; Lindner and Pitombo 2018).

Pitombo et al. (2015a, 2015b), Lindner et al. (2016), Gomes et al. (2016), Rocha et al. (2017), and Lindner and Pitombo (2018) demonstrated (using semivariograms) that travel mode choice variables and transit trip production are spatially correlated data. Kriging enabled the authors to evaluate spatial patterns and to map both disaggregated and aggregated study variables. Aggregated travel demand data have been proven to present modifiable areal unit problems (MAUP), i.e., scale and zonation issues (Lloyd 2014) as the information is associated to areas with different shapes and sizes. On the other hand, disaggregated travel demand data presented a great variability (high values of variance) considering nearby observations. A third consideration to be mentioned is that geostatistics is mostly applied to variables with apparent spatial continuity, commonly seen in natural sciences. Transportation databases normally have spatially discrete variables and, despite this being a counterpoint to the traditional geostatistical method, the kriging estimates, applied to model spatially discrete phenomena, can be found in the literature on

health (Goovaerts and Jacquez 2004; Goovaerts 2005, 2006, 2008, 2009; Kerry et al. 2016). Given all of the mentioned concerns involved in the field of transportation engineering, it may well be argued that travel demand variables require specific adjustments when considered in geostatistical approaches, thus this paper proposes prior variable adjustments to the geostatistical application.

The second limitation seen (not only) in studied spatial models for travel demand (but also in current kriging processes applied to travel demand), i.e., not considering random processes, are addressed in this paper by adopting a geostatistical simulation approach: the Sequential Gaussian Simulation (SGS). That is, the presented approach provides transportation analysts with an alternative tool to disaggregate, map variables of interest, and to evaluate critical scenarios leading to better decision making in transportation policies.

Geostatistics: Understanding the Sequential Gaussian Simulation

Geostatistical techniques are spatial statistics methods that were first developed by Matheron (1963, 1965, 1971). The approach is relevant as it enables the characterization of the spatial dispersion of a phenomenon by analyzing uncertainty measures, determining the spatial variability, and creating a continuous map of estimated (or simulated) values.

Geostatistical methods differ from other spatial techniques, as the former use semivariograms (or covariances) as input to kriging systems. Kriging can identify spatial anisotropy, which may be seen as an advantage when compared to simple interpolation methods. Thus, geostatistics allows researchers to analyze aspects of the direction with greater spatial continuity. Moreover, the geostatistical approach is interesting for this research as it enables mapping the study variable, using less input data, i.e., it may use only the spatial association of one variable, instead of a series of covariates needed in conventional models for travel demand. In addition, the Sequential Gaussian Simulation (SGS) is attractive to transport planners as it creates different scenarios and maps critical spots. In the following paragraphs, the applied geostatistical concepts of variographic analysis, kriging and SGS are introduced.

Matheron (1971) defines geostatistics with the regionalized variable theory. Regionalized variables are those that are (regularly or irregularly) spatially distributed, present spatial structure, and may be considered as the result of a stochastic process. Studying a regionalized variable involves, at least, two geometric aspects: the domain in which the variable is defined and the support to which each observation of a sample is associated (Chilès and Delfiner 1999). The geometric domain is the space where the variation of a regionalized variable is considered relevant. The geometric support, on the other hand,

is described by Matheron (1965) as the size, the geometry, and the spatial orientation associated to the collected sample.

The first step of a variographic analysis is to model the spatial structure of a regionalized variable by calculating an experimental semivariogram, which graphically expresses the spatial structure. Equation 1 presents the semivariogram function, formerly defined by Matheron (1963).

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

where $N(h)$ is the set of all pairwise data values $z(x_i)$ and $z(x_i + h)$ at spatial locations i and $i + h$, respectively. Hence, Eq. 1 may be plotted by setting an ordinate axis with the expectation variance between pairs of observations ($\gamma(h)$) and an abscissa axis with the distance between these pairs, also known as *lag* (h). The following step is to detect a theoretical model—an essential input for kriging—that best fits the experimental semivariogram. Usually, researchers adopt cubic, spherical, and/or exponential models.

Kriging is an estimation method applied in geostatistics. The theory was formulated by Georges Matheron in the 1960s, based on research carried out by Daniel G. Krige in the 1950s. The technique consists of a linear prediction process as the estimated values are linear combinations weighted by sampled data. The main concept embedded in kriging processes lies in the fact that surrounding observations tend to have similar values compared to points that are spread apart. Kriging also differs from other interpolators as it recognizes spatial anisotropy.

The theoretical model is, together with the sampled data, used to set weights λ_i for the kriging system. The purpose of assigning weights is to properly express the influence of the sample data on the estimated values. The kriging system consists of weights that aim at leading to unbiased estimates with minimal variance (Journel 1986). One of the most usual kriging types is simple kriging (SK), which is performed for cases in which the population mean is known. This mean is expected to be uniform in the entire sampled area.

Performing a SGS requires a Gaussian distribution of the data, which is not usual considering practical cases. The process of transforming the distribution of a variable in order to meet this requirement is to firstly sort the data into ascending order to classify the first observation (class) as $k_1 = 1$ and the nth-observation as $k_n = n$. The second step is to calculate the proportion of these classes by dividing each class by the total number of observations (n) − or by $n + 1$, according to Eq. 2 (Journel and Huijbregts 1978). The quantiles of the study variable are calculated using the latter division and the scores of the standard normal distribution according to Eq. 2.

$$y(k_i) = G^{-1}\left(\frac{k_i}{n+1}\right) \qquad (2)$$

where $y(k_i)$ is the score, G^{-1} is the inverse Gaussian function and $k_i/n + 1$ is the quantile for $k_i = 1, n$ (Deutsch and Journel 1998).

Considering that the Gaussian transform leads to zero mean and unit variance, the SK estimator to this case is shown in Eq. 3 (Chilès and Delfiner 1999).

$$z^*(x_0) = \sum_{i=1}^{n} \lambda_i \times z(x_i) \qquad (3)$$

where $z^*(x_0)$ is the estimated value in a non-sampled location x_0; λ, $i = 1,..., n$ are the assigned weights applied to n observations and $z^*(x_i)$ are the values of n observations.

The SGS is a stochastic simulation method that explores a set of scenarios related to a phenomenon. The difference between Kriging and SGS lies in the fact that kriging concerns local statistics—reproducing local means; whereas, SGS involves global statistics—reproducing histograms and variances (Deutsch and Journel 1998). Thus, simulation approaches aim at generating a set of alternative outcomes that replicate spatial patterns, not just by a single disaggregation, as performed by kriging. Estimating a single scenario, calculated by kriging, has the effect of smoothing the results due to the fact that it does not consider an error component. This issue, on the other hand, is addressed in the SGS, whose formulation is presented in Eq. 4.

$$z^{(l)}(x_0) = z^{(*)}(x_0) + R(x_0) \qquad (4)$$

where $N[0,1]$; $z^*(x_0)$ is the kriging formulation—to which the SK is preferred as it reproduces the semivariogram function (Deutsch and Journel 1998), and $R(x_0)$ is the associated error. The SGS involves two stochastic aspects: (1) the simulation of the random term in $R(x_0)$; and (2) the simulation method to define the random path that must (once) visit each point in the grid. These issues are addressed by the Monte Carlo method.

The simulations are known as realizations in the geostatistical field. By comparing different realizations, the simulation methods calculate the associated uncertainty. These realizations are then subject to statistical analyses by evaluating the conditional variances, e-type (average of all realizations) and the confidence interval.

This research paper uses a sequential approach that associates SGS to a proposed data transformation to deal with specific obstacles concerning the implementation of geostatistics to a travel demand dataset. The proposed method allows disaggregated realizations (simulations) to be obtained, so that at the end of the process, it is possible to explore critical situations in the study area.

Furthermore, this paper tackles the change of support (scale) to geostatistical models. Young and Gotway (2007) assert that a new variable, with particular spatial and statistical properties, is generated when changing the

support. In the literature for geostatistics, Cressie (1996) and Kyriakidis (2004) have formerly proposed dealing with the change of scale using point-to-point, point-to-area, area-to-area, and block kriging. These approaches enable calculating aggregated unit areas in terms of disaggregated covariances. However, the case study presented in this paper refers to irregular unit areas and cannot, therefore, consider such kriging techniques. To address this, this paper proposes data transformation in addition to the sequential Gaussian simulation.

Materials and Method

Study Area and Dataset

The study area is located in southeastern Brazil. The São Paulo Metropolitan Area (SPMA) is the most vast and populated Brazilian metropolitan area and comprises 19 municipalities, together with São Paulo, the main metropolis of the state (Fig. 1). The SPMA, with an area of 7947 km^2 and a population of over 20 million inhabitants (Emplasa 2018), can be subdivided into 460 traffic analysis zones (TAZ). TAZ boundaries were established by the São Paulo Metropolitan Company (Metrô 2007) considering: the census track level from 2000, the zoning map from 1997, urban installations, physical barriers, protected areas, the municipality, and district boundaries in the city of São Paulo.

The present study assesses an aggregate travel mode choice dataset associated to TAZs, which was created on the basis of the 2007 origin-destination (O/D) survey performed in the SPMA by the urban subway planning company (Metrô 2007). The O/D survey consisted of selecting 30,000 households using a stratified sample based on the family income.

The disaggregated information of the O/D survey (originally related to households) was extrapolated to give rise to an aggregated dataset related to TAZs. Taking into account the expansion factor used by the Metrô (2007), a TAZ was defined as the smallest unit for which the validity and statistics of the data could be guaranteed. Given this context, this research aims at handling even a smaller unit area by embedding geostatistical concepts.

Figure 2 illustrates the TAZs and the corresponding values for transit trip production, which, in the field of travel demand, refers to the number of trips originating in a TAZ and is represented as the variable of interest in the current study.

Method

Figure 3 denotes the steps followed in this research.

As the aggregate variables of interest in travel demand issues are associated to irregular unit areas, the first step of the proposed method is to create variable adaptations in such ways that (1) the change of support can be considered; and (2) the transformed variable becomes consistent with the original one as the former will represent a different disaggregated unit area to which the sum of the values in the same area must result in the total value of the aggregate variable.

The following steps adopt a conventional procedure concerning using the sequential Gaussian simulation. That is, firstly, it is essential to study the spatial patterns, e.g., the spatial continuity/variability, to ensure the case is a regionalized variable. Secondly, in order to calculate different scenarios using the Gaussian simulation, the Gaussian transform is obtained according to Eq. 2.
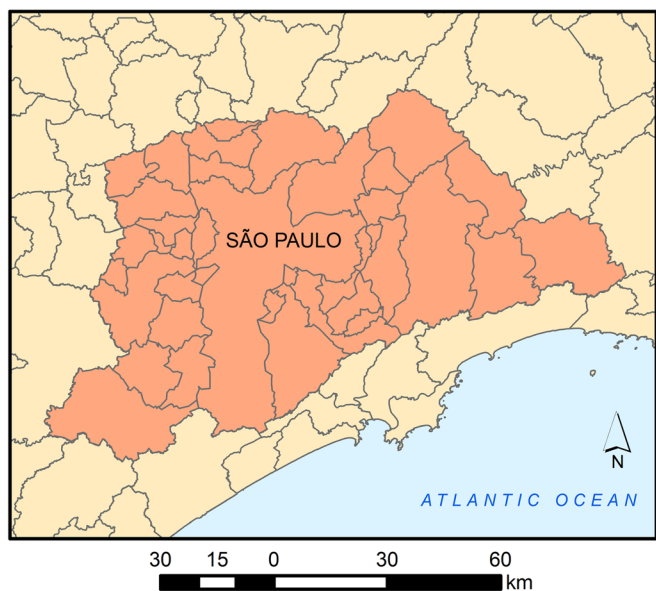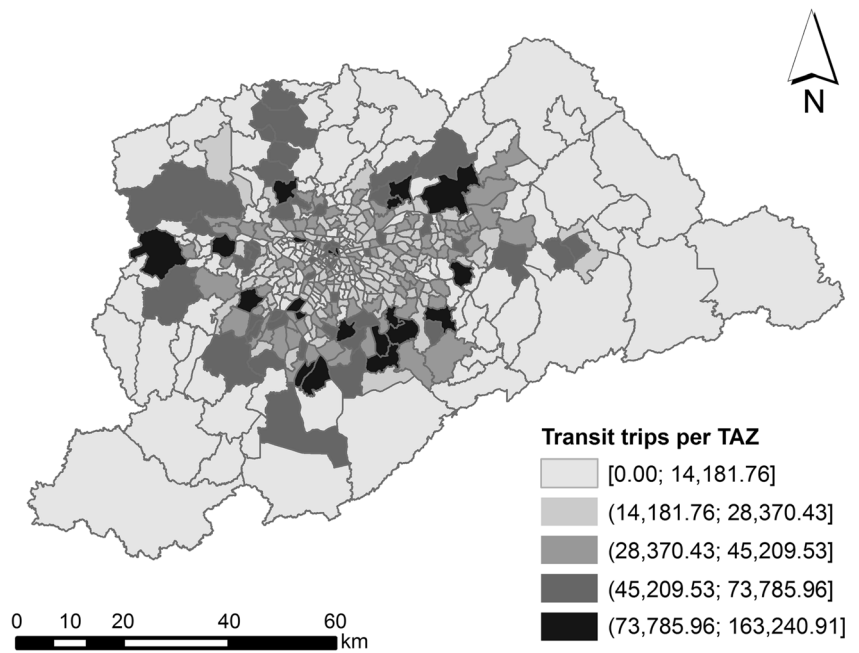

Fig. 1 Localization map and the study area (SPMA and its municipalities)

Fig. 2 Transit trip production per
TAZ

Transit trips per TAZ
- [0.00; 14,181.76]
- (14,181.76; 28,370.43]
- (28,370.43; 45,209.53]
- (45,209.53; 73,785.96]
- (73,785.96; 163,240.91]

Once the Gaussian transform is calculated, the semivariograms may also be calculated, following the geostatistical approach presented in "Geostatistics: Understanding the Sequential Gaussian Simulation" section. The next step is to detect a suitable cell size (within a regular support) to use for the change of support in the simulation process. This research suggests analyzing the most frequent area unit size, considering all records, and assigning weights based on socioeconomic attributes that most affect travel behavior.

Having achieved the former steps, the SGS is then performed—taking into account a sufficient number of realizations that best represent the phenomenon, as well as the previously set criteria (scale and semivariogram parameters). The outcome must be back-transformed into the non-parametric value, and then the results of the minimum/maximum values, e-type and variance can be mapped.

It should be noted that the SGS step deals with the change of support; however, the originated variable would still represent the original support (irregular unit areas) and not the regular scale, which was defined afterwards. In order to address the change of support and the inconsistencies, this paper proposes a heuristic procedure in which

(1) each result should be treated as a density; (2) the sum of the values in the same area must result in the total value of the aggregate variable; and (3) a weighting—related to the corresponding aggregate unit area—shall be assigned. In order to calculate the weightings, each polygon was outlined as an intersection between the unit areas (TAZs) and the cells. Figure 4 outlines an illustration of the procedure by defining three polygons belonging to TAZs 376, 377, and 378.

Finally, the maps for statistical measures of the confidence interval, the minimum/maximum and average values of the simulated maps can be derived. The confidence interval is calculated considering that the population variance is unknown, according to Eq. 5.

$$CI = \overline{X} \pm \Delta_s$$
$$CI = \overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \tag{5}$$

where $\overline{X}$ is the average of all sampled values, $t_{\alpha/2}$ is the critical value considering the significance value α, s is the standard deviation, and n is the number of sampled values.

The computing applications used to calculate geostatistical measures were the R package (maptools, geoR, gstat) and the
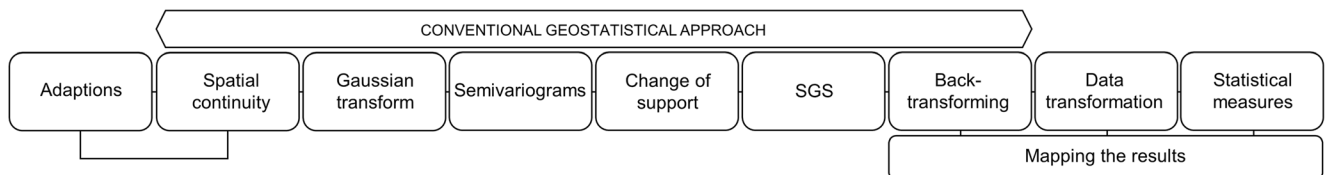

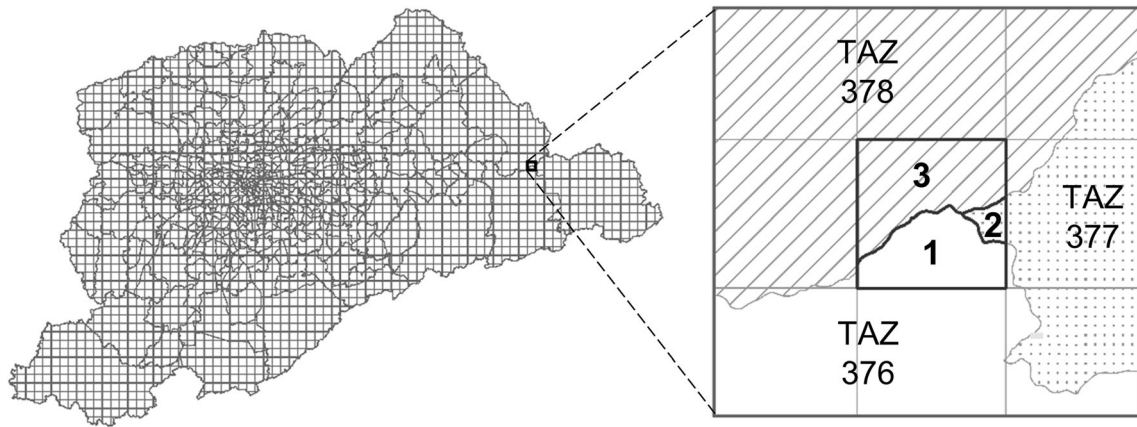**Fig. 3** Illustration of the proposed method

Fig. 4 Approach followed for defining each polygon

SGeMS 3.0. The ArcGIS 10.1 was utilized to exhibit the map of simulated values.

Results and discussions

The results adopting the proposed approach are presented in this section. The scheme shown in Fig. 3 is depicted following the subsections of "Variable Adjustments and Variographic Analysis," "Change of Scale: Choosing a Regular Area Support," "Sequential Gaussian Simulation and Back-Transformation", and finally, "A Heuristic Procedure for Data Transformation."

Variable Adjustments and Variographic Analysis

The proposed method, unlike traditional methods for travel demand, is not straightforward as it requires a few adjustments of the variable throughout the procedure, given the fact that there are changes of support issues and the MAUP.

Firstly, the geostatistical approach estimates values considering that the variable is associated to a point in space. The study variable in the present research is associated to irregular areas and represents the total number of transit trips in those particular areas. However, it is known that the simulation process deals with the change of support and thus the variable must be adapted to fit new conditions, that is, the new support. Adopting a variable of density could reasonably solve this; however, dividing each record into the total area did not lead to a satisfactory outcome, as the variable did not meet the requirements of a regionalized variable. The adopted solution was to use a rate that can be calculated according to Eq. 6.

$$r_n = \frac{V_n}{V_{\text{SPMA}}} \qquad (6)$$

where r is the rate for transit trips in n; V_n is the total number of trips in n; n is the TAZ ranging from 1 to 460 and V_{SPMA} is the total number of transit trips in the SPMA.

The variographic analysis is a geostatistical step that investigates the spatial structure of a variable. The first procedure is to assure that a regionalized variable is involved, i.e., the variable must be spatially distributed with a stochastic spatial structure. Therefore, the variable transit trip rate is spatially represented by the expected variances between pairs of observations in Fig. 5.

Figure 5 corroborates the hypothesis regarding the spatial structure of the variable as the direction of 135° (SE-NW) has greater spatial variability. Accordingly, it can be concluded that the direction of 45° (NE-SW), designated as the main direction, has greater spatial continuity and, therefore, the spatial structure of the variable is anisotropic.

Figure 6 presents the semivariogram for the Gaussian transform in the main and minor directions. The theoretical semivariogram model was selected by visual inspection of the empirical semivariogram model.

The semivariograms presented in Fig. 6 may induce the reader to acknowledge that the data encompasses non-stationarity characteristics. Nonetheless, a further investigation pointed out that the variances tend to remain constant when considering a longer cutoff distance. Furthermore, a variographic analysis using residual input revealed no trend.

The theoretical models (presented in the semivariograms of Fig. 6) are the basis for the kriging (and SGS) processes, according to Eq. 3.

Change of Scale: Choosing a Regular Area Support

Estimation methods often used for transportation planning policies aim to reproduce travel behavior, based on socioeconomic attributes, e.g., it is known that the smaller the aggregation level, the greater the detail level. Thus, information associated with individuals is convenient for traditional travel demand methods. However, by considering spatial methods, the ideal unit area is not necessarily the same. This is due to the fact that each unit area must represent a unique value and, conversely, the traditional travel demand methods may have
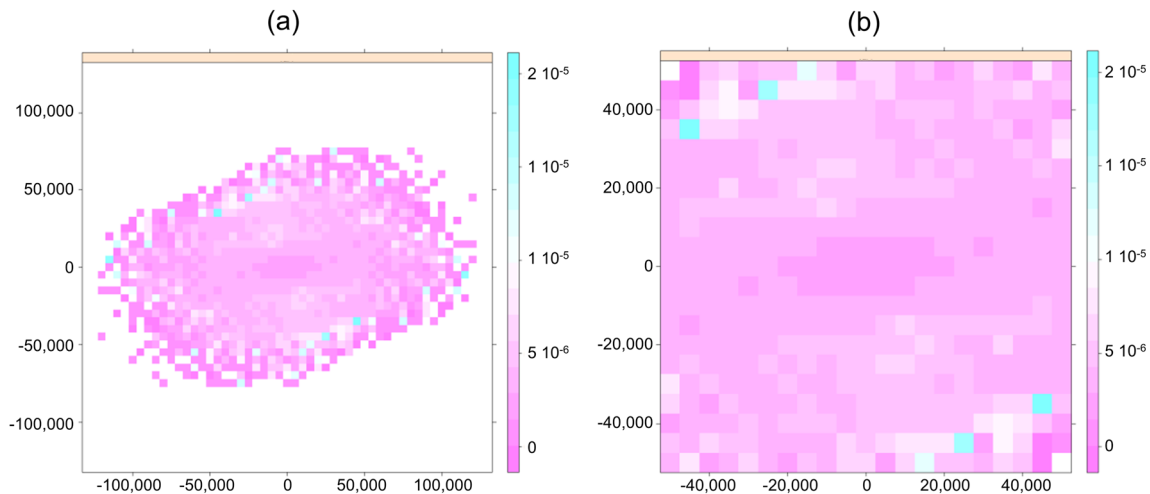
Semivariogram Maps

Fig. 5 Semivariogram maps considering (**a**) the entire study area and (**b**) a cutoff distance of 40 km

different values associated to the same spatial position. In addition, surrounding areas may show similar behaviors, making the aggregation of information interesting. Despite causing loss of information, aggregation may provide advantages to understanding the spatial phenomena or additional analyses that overcome the negative effect. This topic is known in the literature as the change of support (scale). In addition, considering that for different study areas the change of scale is applied, each case study has its own particularities and must be meticulously studied in order to designate the support that best suits the conditions.

In order to consider the change of support for the present study case, it is essential to recognize the correlation between socioeconomic features and trip generation (Ewing et al. 1996). Taking this into account, the method for choosing the ideal support is based on identifying the most frequent area of TAZs and evaluating the influence on the histogram when assigning weightings based on attributes that affect the variable of interest: transit trips. Figure 7 shows (a) the histogram of the TAZ areas; (b) histograms considering weightings for

population density, car ownership, and trip production; and (c) the mean histogram.

The simple histogram shows that TAZs with areas between 225 and 400 ha present more frequency. These areas are emphasized when designating the population density weighting. On the other hand, when assigning weightings that consider the number of cars or the number of trips per TAZ, the frequency of TAZ with areas between 400 and 625 ha becomes greater. Nonetheless, considering the average of the presented histograms, it can be observed that areas between 225 and 400 ha are more frequent. Supposing that the support was a regular square (instead of irregular TAZs with different shapes and sizes), it would be equivalent to assert that the size should be between 1500 and 2000 m. Thus, the support for this research was set with a uniform spacing of 2000 m.

In this study, the spatial structure used for the Sequential Gaussian Simulation is constant, i.e., regardless the spatial support of the output, the semivariogram (used as input) is based on the dataset associated with irregular areas of traffic analysis zones. Thus, parametric techniques, such as Akaike
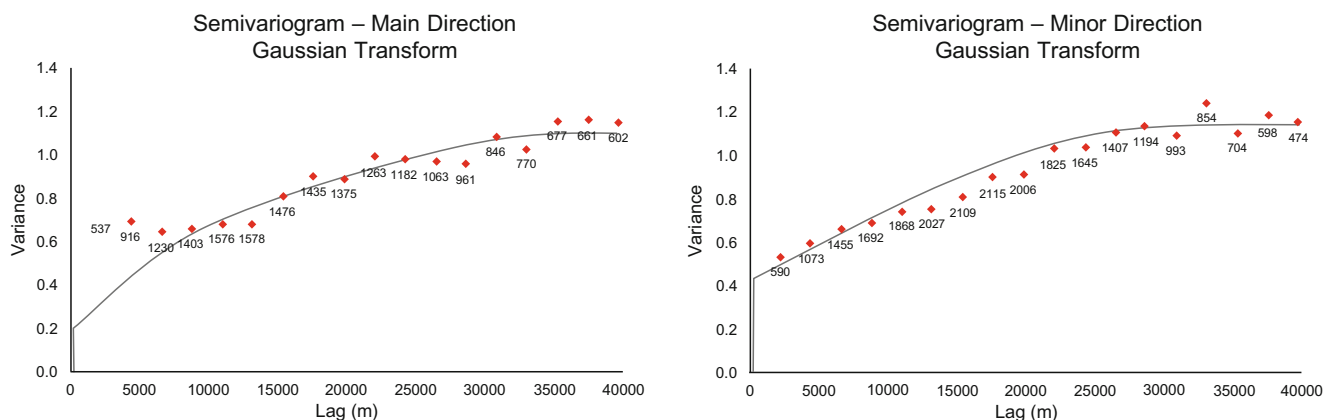

Fig. 6 Experimental and theoretical semivariogram for the Gaussian transform in the main and minor directions
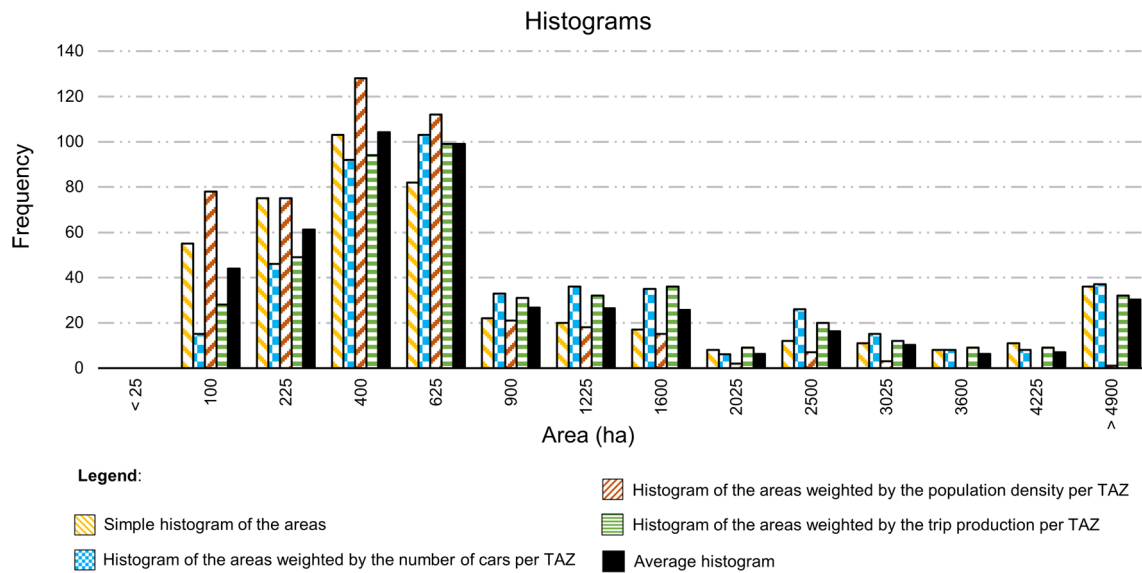
Fig. 7 Histogram for the TAZ areas and weighted histograms

and Bayesian information criteria are not applicable to validate the selection of the adequate scale for the simulation step.

Sequential Gaussian Simulation and Back-Transformation

Sequential Gaussian Simulation (SGS) is a stochastic simulation method in which random numbers are generated to explore a field of possibilities of a phenomenon (Remy et al. 2009). Thus, the simulation aims to generate a set of alternative results (realizations) that reproduce spatial patterns. The theoretical semivariogram presented in Fig. 6 was used for simple Kriging as a procedure intrinsic to the Sequential Gaussian Simulation method. In the same way, the selected scale was used for the Sequential Gaussian Simulation method.

In order to come up with a sufficient number of realizations that best represents the phenomenon, the variance between the realizations was evaluated. Figure 8 shows the variance between combinations of 1000 realizations.
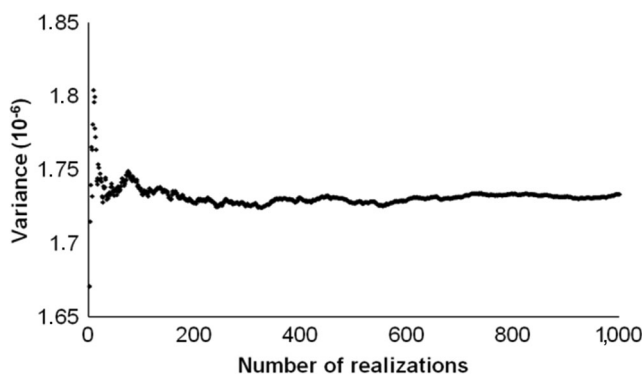

Fig. 8 Average variances from the SGS between combinations of 1000 realizations

It can be observed that the average variances between 2 and approximately 100 realizations fluctuate significantly. Based on a conservative decision, 500 realizations were chosen for this research. As a conservative decision, the number of settled for this research was of 500 realizations. The simulated scenarios were then back-transformed to take into account the non-parametric variable. Figure 9 presents the e-type (average) and variance from the 500 realizations, considering the variable as the transit trip rate per TAZ. In addition, the maximum and minimum values seen in the realizations are presented in Fig. 10.

Despite the advantages of the method allowing researchers to evaluate the uncertainties associated with the method, planners still have to solve problems related to the nature of the transport variable. The spatial aggregation needs to be addressed in such way that each record is associated to the specified regular support.

A heuristic procedure for data transformation

The change of support was addressed by transforming the data following Eq. 7.

$$V_c = r_c V_{\text{SPMA}} \sum_p^P \left[\frac{A_p}{A_c} \left(\sum_z \frac{A_p}{A_z V_z} \right) \right] \tag{7}$$

where V_c is the number of transit trips in cell c; r_c is the estimated (simulated) rate in c; V_{SPMA} is the total number of trips in the SPMA; A_p, A_c, and A_z are the areas in p, c, and in z, respectively; P is the total number of polygons belonging to c; p is a polygon belonging to c; V_z is the number of trips in z and z is the TAZ including the polygon p.

The process of using SGS together with the proposed heuristic data transformation causes the method to no longer be
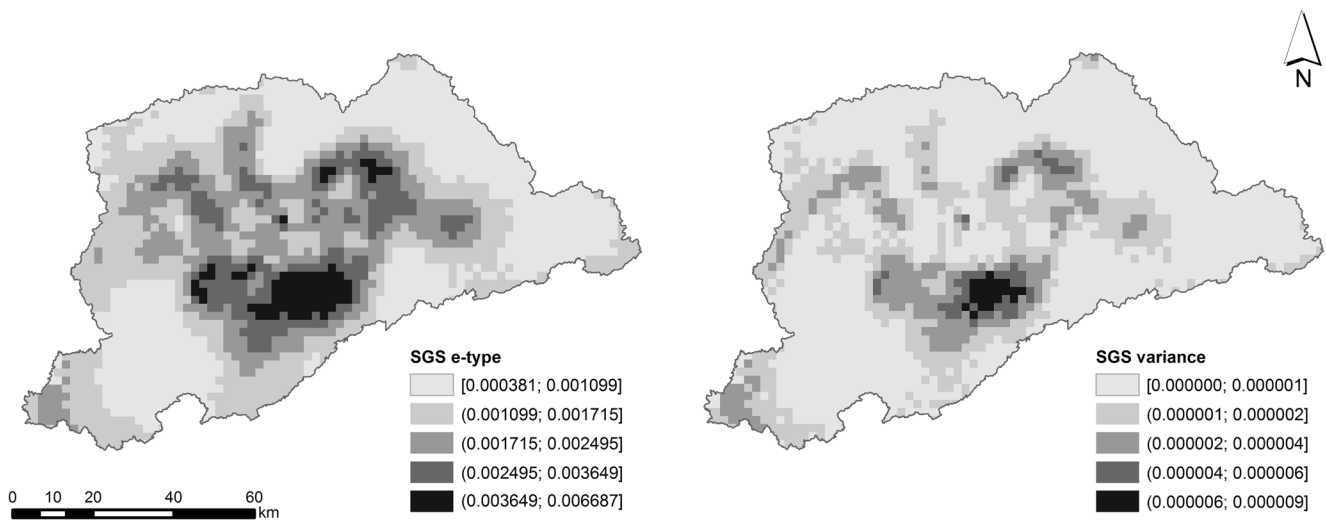
Fig. 9 SGS results—e-type and variance

fully stochastic. Despite this, the method is particularly suitable for social sciences issues, in which the land use and/or human factors can be not only taken into account, but can also be used as constraints in the simulation.

The considered constraints restricted the SGS realizations to assign transit trips at cells belonging to TAZs in which the original number of total trips was not null. In addition, the cells belonging to each TAZ respected the corresponding total number of trips. The maximum and minimum values detected for all 500 simulations in each cell were mapped as presented in Fig. 11.

Figure 12 shows the average of all simulations and the delta (Eq. 5) associated to a 95% confidence interval. Figures 11 and 12 show the critical spots for the number of transit trips. It can be observed that São Paulo city center has the highest values for transit trips. The given

results are consistent with the actual scenario in the SMPA, as mobility rates are greater when individuals consider using public transportation (especially the subway) instead of private motorized transportation, mainly in the center of São Paulo at peak hours. This is due to municipal transportation and circulation plans that have reduced parking spaces and created car restriction policies. In contrast, it can be observed that the outskirts of the SMPA do not present significant values for transit trips. This situation can be explained by the fact that outer TAZs do not have an effective integrated transit system.

The results provided by the proposed sequential method (Sequential Gaussian Simulation and data transformation) may allow analysts to assess areas requiring intervention by overlapping the critical spots for transit trips with the existing public transportation network.
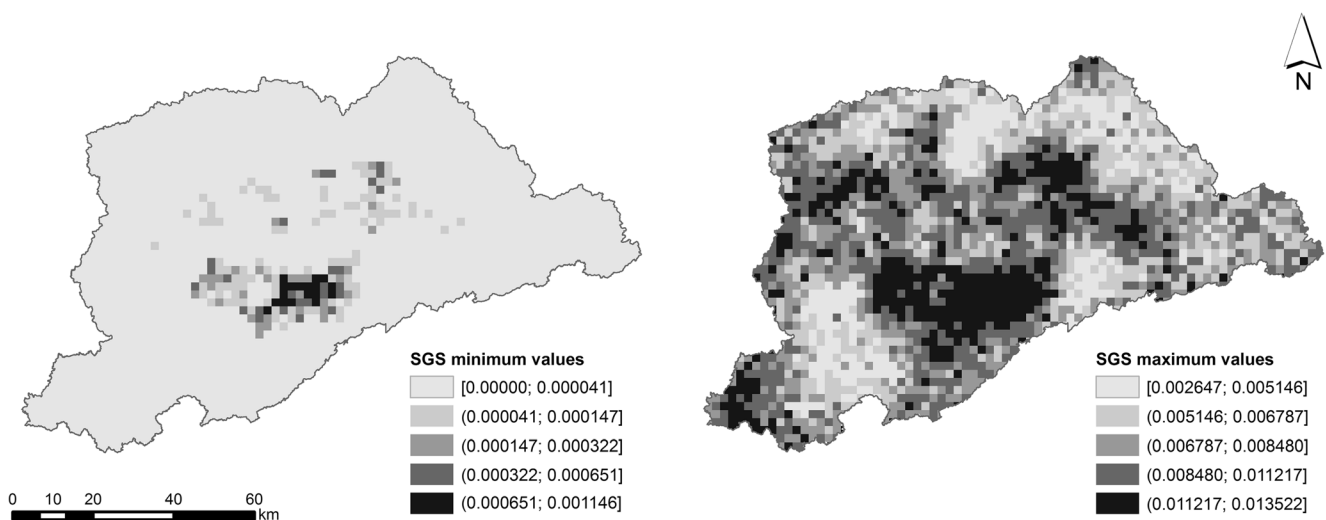

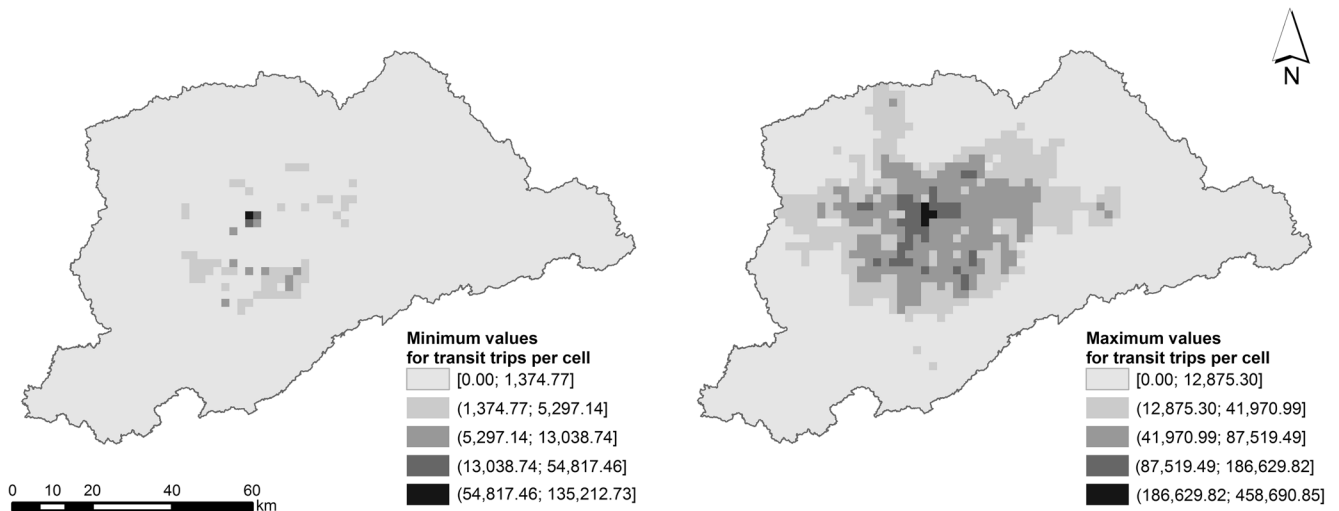**Fig. 10** SGS results—minimum and maximum values

Fig. 11 Data transformation results: minimum and maximum number of transit trips per cell

Conclusions

The main interest of this research for urban planning policies refers to the advantage of mapping critical scenarios for travel demand using a spatially correlated variable. The benefit of providing a map of transit trips associated to a disaggregated unit area helps decision makers to provide a more efficient public transportation system.

Furthermore, having travel variables with a spatial structure is of great interest to the field of travel demand considering that spatially correlated variables may lead to a reduction in the amount of input data in conventional models. The geostatistical analysis conducted in this paper showed that the rate variable of transit trips per TAZ presented characteristics of a regionalized variable, since an anisotropic spatial structure was observed in the semivariogram representation.

The Sequential Gaussian Simulation was applied to the rate variable and—with post-processing; e-type maps, variance, maximum, and minimum were generated. Despite not precisely expressing the number of transit trips related to the current unit area (2000 × 2000-m cells), the method has great potential as it creates continuous maps of different scenarios only using the variable of interest (and its spatial location). Other associated advantages are recognized as estimating values in non-sampled locations and calculating uncertainty parameters. In order to address problems associated to the change of support, the proposed approach for transforming the rate variable was considered, and its results reinforce the idea that the SGS may be applied to social sciences.

Creating disaggregated maps of critical scenarios only using a single variable is the main aim of this research, especially knowing that developing countries do not usually have
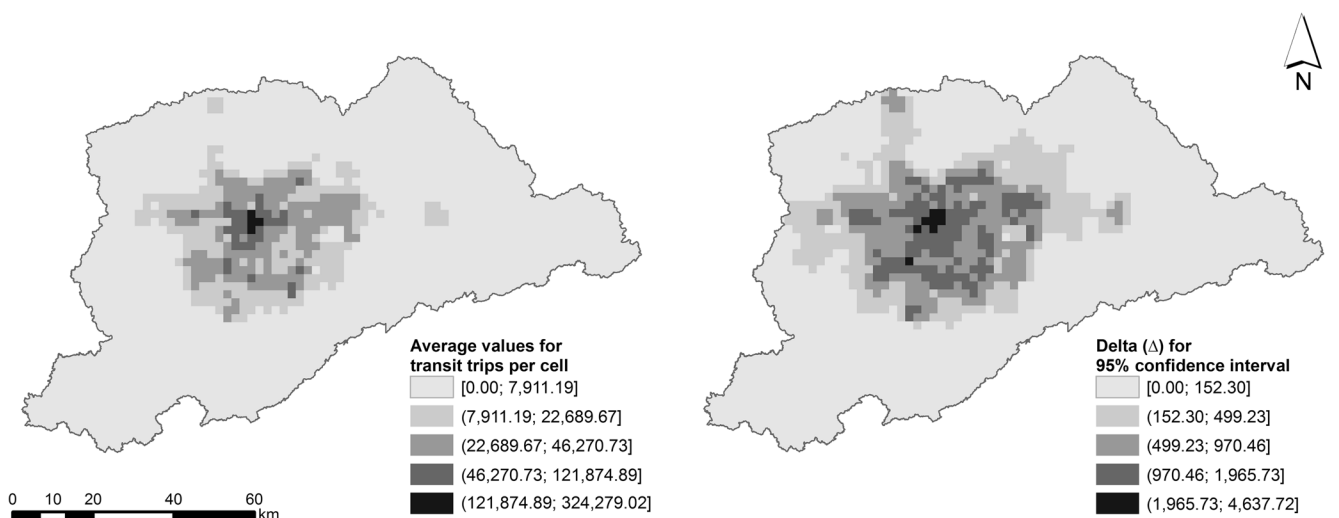

Fig. 12 Data transformation results: average and confidence interval for the number of transit trips per cell

refined information available due to high costs. Despite being very straightforward, the proposed method is innovative and intends to motivate future research aiming to produce practical results for decision making in transportation planning policies, particularly taking into account cost reduction.

Compliance with Ethical Standards

The authors of this paper agree, accept, and comply with all the ethical standards set by the Journal of Geovisualization and Spatial Analysis.

Conflict of Interest   The authors declare that they have no conflict of interest.

Ethical Approval   The authors confirm that this research did not involve any experiment on animals nor on human patients.

Informed Consent   The authors have given the informed consent to publish this article in the Journal of Geovisualization and Spatial Analysis if accepted.

References

Antipova A, Wang F, Wilmot C (2011) Urban land uses, socio-demographic attributes and commuting: a multilevel modeling approach. Appl Geogr 31(3):1010–1018. https://doi.org/10.1016/j.apgeog.2011.02.001

Arentze TA, Timmermans HJP (2000) Albatross: a learning-based transportation-oriented simulation system. Eindhoven University of Technology, Netherlands, Eirass

Arentze T, Timmermans H, Hofman F (2007) Creating synthetic household populations: problems and approach. Transp Res Rec: J Transp Res Board 2014:85–91. https://doi.org/10.3141/2014-11

Ballas D, Clarke G, Dorling D, Eyre H, Thomas B, Rossiter D (2005) SimBritain: a spatial microsimulation approach to population dynamics. Popul Space Place 11(1):13–34. https://doi.org/10.1002/psp.351

Balmer M, Axhausen K, Nagel K (1985) Agent-based demand-modeling framework for large-scale microsimulations. Transp Res Rec: J Transp Res Board 2006:125–134. https://doi.org/10.3141/1985-14

Barthelemy J, Toint PL (2013) Synthetic population generation without a sample. Transp Sci 47(2):266–279. https://doi.org/10.1287/trsc.1120.0408

Beckman RJ, Baggerly KA, McKay MD (1996) Creating synthetic baseline populations. Transp Res A Policy Pract 30(6):415–429. https://doi.org/10.1016/0965-8564(96)00004-3

Ben-Akiva ME, Bowman JL (1998) Activity based travel demand model systems. Equilibrium and advanced transportation modeling. Springer, Boston, MA, pp 27–46. https://doi.org/10.1007/978-1-4615-5757-9_2

Ben-Akiva ME, Ramming MS, Bekhor S (2004) Route choice models. *Human Behaviour and Traffic Networks*. Springer, Berlin Heidelberg, pp 23–45

Bhat CR, Sener IN (2009) A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units. J Geogr Syst 11(3):243–272. https://doi.org/10.1007/s10109-009-0077-9

Bhat C, Zhao H (2002) The spatial analysis of activity stop generation. Transp Res B 36(6):557–575. https://doi.org/10.1016/S0191-2615(01)00019-4

Buliung RN, Kanaroglou PS (2007) Activity–travel behaviour research: conceptual issues, state of the art, and emerging perspectives on behavioural analysis and simulation modelling. Transp Rev 27(2):151–187. https://doi.org/10.1080/01441640600858649

Chilès JP, Delfiner P (1999) Geostatistics modeling spatial uncertainty. John Wiley & Sons, New York, 695p

Ciuffo BF, Punzo V, Quaglietta E (2011) Kriging meta-modelling to verify traffic micro-simulation calibration methods. *TRB 90th Annual Meeting Compendium of Papers*. Washington, D.C.

Cressie NA (1996) Change of support and the modifiable areal unit problem. Geographical Systems 3(2–3):159–180

Deutsch CV, Journel AG (1998) *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, p.18–19; p.119–147

Emplasa - São Paulo Metropolitan Planning Company (2018) São Paulo Metropolitan Area Info. In: https://www.emplasa.sp.gov.br/RMSP (last accessed on 06.05.2018)

Ewing R, DeAnna M, Li SC (1996) Land use impacts on trip generation rates. Transp Res Rec: J Transp Res Board 1518:1–6. https://doi.org/10.3141/1518-01

Farooq B, Bierlaire M, Hurtubia R, Flötteröd G (2013) Simulation based population synthesis. *Transp Res B Methodol* 58:243–263. https://doi.org/10.1016/j.trb.2013.09.012

Gomes VA, Pitombo CS, Rocha SS, Salgueiro AR (2016) Kriging geostatistical methods for travel mode choice: a spatial data analysis to travel demand forecasting. *Open J Stat* 6(03):514. https://doi.org/10.4236/ojs.2016.63044

Goovaerts P (2005) Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *Int J Health Geogr* 4(1):31. https://doi.org/10.1186/1476-072X-4-31

Goovaerts P (2006) Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. *Int J Health Geogr* 5(1):52. https://doi.org/10.1186/1476-072X-5-52

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

Goovaerts P (2009) Medical geography: a promising field of application for geostatistics. Math Geosci, v. 41, n 3, p. 243–264. doi: https://doi.org/10.1007/s11004-008-9211-3

Goovaerts P, Jacquez GM (2004) Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *Int J Health Geogr* 3(1):14. https://doi.org/10.1186/1476-072X-3-14

Gundogdu, IB (2014) Risk governance for traffic accidents by Geostatistical Analyst methods. *International Journal of Research in Engineering and Science*, v.2, Issue 9, p. 35–40

Guo J, Bhat C (2007) Population synthesis for microsimulating travel behavior. Transp Res Rec: J Transp Res Board 2014:92–101. https://doi.org/10.3141/2014-12

Hermes K, Poulsen M (2012) A review of current methods to generate synthetic spatial microdata using reweighting and future directions. Comput Environ Urban Syst 36(4):281–290. https://doi.org/10.1016/j.compenvurbsys.2012.03.005

Hunt JD, Donnelly R, Abraham JE, Batten C, Freedman J, Hicks J, Costinett PJ, Upton WJ (2001) Design of a statewide land use transport interaction model for Oregon, *9th World Conference for Transport Research*, Seoul, South Korea, v 19

Journel AG (1986) Geostatistics: models and tools for the earth sciences. Math Geol 18(1):119–140. https://doi.org/10.1007/BF00897658

Journel AG, Huijbregts CJ (1978) *Mining geostatistics*. Academic press. Reprinted (1991). The Blackburn Press, United States of America 600p

Kamruzzaman M, Hine J, Gunay B, Blair N (2011) Using GIS to visualise and evaluate student travel behaviour. J Transp Geogr 19(1):13–32. https://doi.org/10.1016/j.jtrangeo.2009.09.004

Kerry R, Goovaerts P, Vowles M, Ingram B (2016) Spatial analysis of drug poisoning deaths in the American West, particularly Utah. Int J Drug Policy 33:44–55. https://doi.org/10.1016/j.drugpo.2016.05.004

Kitamura R (1988) An evaluation of activity-based travel analysis. Transportation 15(1–2):9–34

Kitamura R, Fujii S (1998) Two computational process models of activity-travel behavior. Theor Found Travel Choice Modeling, p. 251-279

Kyriakidis PC (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

Landis J, Zhang M (1998) The second generation of the California urban futures model parts 1, 2 and 3. Environ Plann B: Plann Design 25:657–666. https://doi.org/10.1068/b36046t

Lee SY, Carle SF, Fogg GE (2007) Geologic heterogeneity and a comparison of two geostatistical models: Sequential Gaussian and transition probability-based geostatistical simulation. Adv Water Resour, v 30, p. 91914–91932. doi: https://doi.org/10.1016/j.advwatres.2007.03.005

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

Lindner A, Pitombo CS, Rocha SS, Quintanilha JA (2016) Estimation of transit trip production using factorial kriging with external drift: an aggregated data case study. Geo-spatial Inform Sci 19(4):245–254. https://doi.org/10.1080/10095020.2016.1260811

Lloyd CD (2014) Exploring spatial scale in geography, John Wiley & Sons, 255p

Lovelace R, Dumont M (2016) *Spatial microsimulation with R*. CRC Press, Boca Raton, United States 257 p

Manepalli UR, Bham GH (2016) Application of spatial statistics in transportation engineering. Appl Spatial Stat. https://doi.org/10.5772/65051

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

Matheron G (1965) Les variables Régionalisées et leur estimation. Masson et Cie, Paris, 306 p

Matheron G (1971) *The theory of regionalized variables and its applications.* Cahiers du Centre de Morphologie Mathématique, 5. ENSMP, Paris, 212p

Mazzella A, Piras C, Pinna F (2011) Use of kriging technique to study roundabout performance. Transp Res Rec: J Transp Res Board 2241:78–86. https://doi.org/10.3141/2241-09

McNally MG, Rindt CR (2007) The activity-based approach. *Handbook of Transport Modelling*: 2nd edition. Emerald Group Publishing limited. p. 55-73

Metrô - São Paulo Metropolitan Company (2007) Origin-Destination Survey 2007 - São Paulo Metropolitan Area: Summary of information. In: http://www.metro.sp.gov.br/metro/numeros-pesquisa/pesquisa-origem-destino-2007.aspx (last accessed on 19.05.2016)

Miyamoto K, Vichiensan V, Shimomura N, Páez A (2004) Discrete choice model with structuralized spatial effects for location analysis. Transp Res Rec 1898:183–190. https://doi.org/10.3141/1898-22

Moeckel R, Spiekermann K, Wegener M (2003) Creating a synthetic population. *8th International Conference on Computers in Urban Planning and Urban Management*. p. 1-18

Molla, MM; Stone ML; Lee, E (2014) Geostatistical approach to detect traffic accident hot spots and clusters in North Dakota. Upper Great Plains Transportation Institute

Morency C, Paez A, Roorda MJ, Mercado R, Farber S (2011) Distance traveled in three Canadian cities: spatial analysis from the perspective of vulnerable population segments. J Transp Geogr 19(1):39–50. https://doi.org/10.1016/j.jtrangeo.2009.09.013

Müller K, Axhausen KW (2011) Hierarchical IPF: Generating a synthetic population for Switzerland. *51st Congress of the European Regional Science Association*: New Challenges for European Regions and Urban Areas in a Globalised World, 30 August - 3 September 2011, Barcelona, Spain

Orton TG, Pringle MJ, Bishop TFA (2016) A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals. Geoderma 262:174–186. https://doi.org/10.1016/j.geoderma.2015.08.013

Ortúzar JD, Willumsen LG (2011) *Modeling Transport* Wiley, 4th edition

Páez A (2007) Spatial perspectives in urban systems: developments and directions. J Geogr Syst 9(1):1–6. https://doi.org/10.1007/s10109-007-0041-5

Páez A, Scott DM (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

Páez A, López FA, 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

Pearce JL, Rathbun SL, Aguilar-Villalobos M, Naeher LP (2009) Characterizing the spatiotemporal variability of PM2. 5 in Cusco, Peru using kriging with external drift. Atmos Environ 43(12):2060–2069. https://doi.org/10.1016/j.atmosenv.2008.10.060

Pendyala R, Konduri K, Chiu YC, Hickman M, Noh H, Waddell P, Wang L, You D, Gardner B (2012) Integrated land use-transport model system with dynamic time-dependent activity-travel microsimulation. Transp Res Rec: J Transp Res Board 2303:19–27. https://doi.org/10.3141/2303-03

Pitombo CS, Costa ASG, Salgueiro AR (2015a) Proposal of a sequential method for spatial interpolation of mode choice. Bol Ciên Geodésicas 21(2):274–289. https://doi.org/10.1590/S1982-21702015000200016

Pitombo CS, Salgueiro AR, Costa ASG, Isler CA (2015b) 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

Rahman A (2009) Small area estimation through spatial microsimulation models: some methodological issues. *2nd International Microsimulation Association Conference*, Ottawa, Canada, 8-10 June 2009

Recker WW, McNally MG, Root GS (1986a) A model of complex travel behavior: part I - theoretical development. Transp Res Part A: General 20(4):307–318. https://doi.org/10.1016/0191-2607(86)90089-0

Recker WW, McNally MG, Root GS (1986b) A model of complex travel behavior: part II - an operational model. Transp Res Part A: Gen 20(4):319–330. https://doi.org/10.1016/0191-2607(86)90090-7

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

Rocha SS, Lindner A, Pitombo CS (2017) Proposal of a geostatistical procedure for transportation planning field. Bol Ciên Geodésicas 23(4):636–653. https://doi.org/10.1590/s1982-21702017000400042

Salvini PA, Miller EJ (2005) ILUTE: an operation prototype of a comprehensive microsimulation model of urban systems. Netw Spat Econ, v 5, p. 217–234. doi: https://doi.org/10.1007/s11067-005-2630-5

Song Y, Wang X, Wright G, Thatcher D, Wu P, Felix P (2018) Traffic volume prediction with segment-based regression kriging and its implementation in assessing the impact of heavy vehicles. *IEEE Trans Intell Transp Syst*:1–12. https://doi.org/10.1109/TITS.2018.2805817

Tong D, Lin WH, Stein A (2013) Integrating the directional effect of traffic into geostatistical approaches for travel time estimation. Int J Intell Transp Syst Res 11:101–112. https://doi.org/10.1007/s13177-013-0061-0

Waddell P (2000) A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim. Environ Plann B: Plann Des 27(2):247–263. https://doi.org/10.1068/b2627

Yagi S, Mohammadian AK (2010) An activity-based microsimulation model of travel demand in the Jakarta metropolitan area. J Choice Model v. 3, n 1, p. 32–57. doi: https://doi.org/10.1016/S1755-5345(13)70028-9

Young LJ, Gotway CA (2007) Linking spatial data from different sources: the effects of change of support. Stoch Env Res Risk A 21(5):589–600. https://doi.org/10.1007/s00477-007-0136-z

Zhang L; Levinson D (2004) Agent-based approach to travel demand modeling: exploratory analysis. Transp Res Rec: J Transp Res Board, n. 1898, p. 28–36. doi: org/https://doi.org/10.3141/1898-04

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