BCG - Bulletin of Geodetic Sciences - On-Line version, ISSN 1982-2170
http://dx.doi.org/10.1590/S1982-21702017000400042

Article

PROPOSAL OF A GEOSTATISTICAL PROCEDURE FOR TRANSPORTATION PLANNING FIELD

Proposta de um procedimento geoestatístico para uso na área de Planejamento de Transportes

Samille Santos Rocha[1]

Anabele Lindner[1]

Cira Souza Pitombo[1]

[1] Universidade de São Paulo, Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, São Carlos, SP, Brasil.

E-mail- samille@usp.br, anabele@usp.br, cirapitombo@usp.br

Abstract:

The main objective of this study is to estimate variables related to transportation planning, in particular transit trip production, by proposing a geostatistical procedure. The procedure combines the semivariogram deconvolution and Kriging with External Drift (KED). The method consists of initially assuming a disaggregated systematic sample from aggregate data. Subsequently, KED was applied to estimate the primary variable, considering the population as a secondary input. This research assesses two types of information related to the city of Salvador (Bahia, Brazil): an origin-destination dataset based on a home-interview survey carried out in 1995 and the 2010 census data. Besides standing out for the application of Geostatistics in the field of transportation planning, this paper introduces the concepts of semivariogram deconvolution applied to aggregated travel data. Thus far these aspects have not been explored in the research area. In this way, this paper mainly presents three contributions: 1) estimating urban travel data in unsampled spatial locations; 2) obtaining the values of the variable of interest deriving out of other variables; and 3) introducing a simple semivariogram deconvolution procedure, considering that disaggregated data are not available to maintain the confidentiality of individual data.

Keywords: travel demand; kriging; semivariogram deconvolution; disaggregated systematic sample; aggregated data.

Resumo:

O principal objetivo deste estudo é estimar variáveis relacionadas ao planejamento de transportes, especialmente valores de produção de viagens por Transportes Coletivo (TCO), propondo um procedimento geoestatístico. O procedimento é uma junção de deconvolução de semivariograma e Krigagem com Deriva Externa (KED). O método consiste em assumir

inicialmente uma amostra sistemática desagregada a partir de dados agregados por unidade de área. Posteriormente, utiliza-se a KED para prever a variável primária, considerando a população como *input* (variável secundária). Esta pesquisa utiliza dois tipos de informações sobre a cidade de Salvador (Bahia, Brasil): dados da pesquisa domiciliar origem-destino, realizada em 1995 e dados do censo de 2010. Além de destacar-se pela aplicação da Geoestatística no campo do planejamento de transportes, este artigo demonstra um método simples de deconvolução de semivariograma para dados agregados de viagens. Estes aspectos foram pouco explorados nesta área de pesquisa. Portanto, este trabalho apresenta três principais contribuições: 1) estimação de dados de viagens urbanas em localizações espaciais não amostradas; 2) obtenção de valores da variável de interesse derivada de outras variáveis; e 3) apresentação de um procedimento simplificado para deconvolução de semivariograma, considerando que dados desagregados, geralmente, não estão disponíveis, para manter a confidencialidade de dados individuais.

Palavras-chave: demanda por viagens; krigagem; deconvolução de semivariograma; amostra sistemática desagregada; dados agregados.

1. Introduction and brief literature review

The transportation planning aims at adjusting the supply with transportation demand, in order to minimize the problems caused by the travel needs. In the field of transportation planning, a few concepts are widely used and some of them are object of study of this paper. According to diverse authors, a trip may be defined as the movement of a single direction leaving from an origin point and heading to a certain destination; trip production (focus of this research) is the amount of trips originated in a specific Traffic Analysis Zone (TAZ) or region; trip attraction is the amount of trips attracted to a particular TAZ or region; trip generation is the sum of trip production and trip attraction per TAZ or region (Papacostas and Prevedouros, 1993; Ortúzar and Willumsen , 2011).

Trip generation model is the first step of the classic four step transport model: trip generation, distribution, travel mode choice and trip assignment. Trip generation models aim to estimate the production and attraction of trips in each TAZ or trips generated by household. This first step is essential to transportation planning, since it is analyzed the main factors that determine the trip generation. The most common models for representing trip generation are Multiple Linear Regression and Cross Classification (Mahmoud and Abdel, 2004; Al-Taei and Taher, 2006; Ortúzar and Willumsen, 2011).

However, these traditional models of trip generation forecasting assume that the observed data have no association with the spatial location (Lopes et al., 2014). It is understood that urban travel issues are associated with individual and household features, as well as the spatial location of each household, destination and the activities distribution in the urban environment (Páez et al., 2013). Therefore, the incorporation of spatially correlated variables and the spatial position into studies of travel demand become important to enhance the quality of the estimations (Bhat and Sener, 2009; Páez and Scott, 2005; Ben-Akiva et al., 2004; Pitombo et al., 2015; Lindner et al., 2016; Gomes et al., 2016; Rocha et al., 2016).

The influence of spatially correlated data and the increasing availability of georreferenced data enabled the incorporation of spatial data into studies on travel modeling. Such types of researches have been identified as emerging lines of study (Páez et al., 2013).

Bhat and Zhao (2002) identified spatial factors that must be incorporated in travel demand models. The authors formulated a multilevel mixed logit to estimate trip generation and activities in the Boston Metropolitan Area. Adjemian et al. (2010) demonstrated that the travel mode choice is associated and depends on the spatial location. Páez et al. (2013) introduced a spatial indicator that was incorporated to discrete choice models for household-based travel estimation. Bhat and Sener (2009) introduced a multivariate logistic distribution copula-based approach to address spatial dependency and endogeneity issues in binary discrete choice models.

Among the spatial statistical techniques, geostatistical methods stand out as they enable the study of features in which the variables are spatially correlated. This makes possible to estimate values of a specific variable when considering a coordinate where there is no prior knowledge of the values. Kriging has been an accepted tool in spatial statistics since it moved from geology and geochemistry into other applications in the 1990s. Although in the 1990s there were some studies in the literature on travel demand forecasting with applications of estimation techniques related to Kriging (Ickstadt, K. et al., 1998), its application is still recent and has not yet been fairly explored in the line of research of Transportation Planning (Yoon et al., 2014; Chen et al., 2015). However, it is possible to find diverse studies in "Transportation Engineering" field (Ciuffo et al., 2011; Mazzella et al., 2011; Zhang and Wang, 2013) and also Emission of vehicular gases (Pearcea et al., 2009; Kassteele and Velders, 2006; Kassteele and Stein, 2006).

Different levels of data aggregation, known in the field of spatial analysis as the change of support, influence the results of geostatistical analysis. The support defines the method, spacing and/or volume of data acquisition (Journel and Huijbregts, 1978). The change of support is related to the adopted scale and the existence of the Modifiable Areal Unit Problem (MAUP) and ecological fallacy (when occurring the misinterpretation of results).

In the field of Transportation Planning, disaggregated data (e.g. household's or individual data) is the best geometric support for forecasting urban travel. That is, this paper has two important steps towards the travel data adequacy to geostatistical modeling: proposing a simple method of disaggregated systematic sampling from two types of aggregated datasets (the 1995 origin-destination dataset and the 2010 census data - IBGE 2010) – the semivariogram deconvolution; and estimating and mapping the *transit trip production* by using Kriging with External Drift (KED).

Hence, this paper has three main objectives: 1) to use geostatistics to estimate urban travel data in locations where the values are unknown or unobserved; 2) to obtain the values of the variable of interest deriving out of a secondary variable (more easily available) by using Kriging with External Drift and; 3) proposing a disaggregated systematic sampling that uses aggregate data (a simple procedure of semivariogram deconvolution).

This article is organized into four sections besides this Introduction. Section 2 presents the materials (techniques, study area and dataset) and the method. Section 3 presents the results, and finally, Section 4 describes the main conclusions.

2. Materials and method

2.1 Sampling in Geostatistics: Systematic Sampling

A sample is a set of values from a spatial event that must have a good representation of reality, in the sense to present certain characteristics, in order to allow the modeling of studied phenomenon. Geostatistical approaches calculate the uncertainty from estimations based on samples. Three types of samples are identified in Geostatistics: Simple Random Sampling (SRS), Stratified Sampling and Systematic Sampling (Webster and Oliver, 2007).

In the SRS, the random components are the geographic coordinates. The selection is arbitrary and all the components must have the same probability of occurrence. The Stratified Sampling works by partitioning the population into groups and randomly collecting the samples in each stratum.

When considering the Systematic Sampling, the intervals are regular and established in advance. In the case of Geostatistics, the Systematic Sample has a particular advantage, since it covers the entire study area and avoids empty areas or grouping of samples. Hence, a Systematic Sampling may produce more accurate estimates (Wang et al., 2012).

This paper introduces a Systematic Sampling aiming at enhancing the geostatistical estimations in the field of Transportation Planning. The procedure consists of considering the geographic coordinates of TAZ's centroids and building a regular grid mesh that turns the dataset regularly distributed in space. The use of the Systematic Sampling is relevant for softening the spatial discontinuity effects of travel variables in terms of the aggregated data, for instance.

2.2 Multivariate Geostatistics

The main point of using Geostatistics is to characterize the spatial (and/or spatial/temporal) dispersion of an event, assessing uncertainty parameters, concerning its spatial variability and obtaining a continuous surface estimation. Geostatistics is better defined as the following steps: 1) variographic analysis, 2) cross validation, and 3) kriging.

2.3 Variographic Analysis

The first concept regarding the use of Geostatistics is the definition of the regionalized variable theory. If a particular variable is distributed in space and suggests various trends in a set of different points, it consists of a regionalized variable. Matheron (1970) first showed that the spatial variation of a regionalized variable could be expressed by: a structural component, having a constant trend that is spatially dependent; a random, but spatially correlated component; and a spatially uncorrelated random noise.

The investigation of the spatial dispersion of each variable makes it possible to predict the structure of a random process, the main direction of spatial continuity (case of anisotropy) or if there is an omnidirectional property. This step of the proposed geostatistical approach aims at investigating if the study variables are indeed regionalized by plotting an experimental semivariogram.

An experimental semivariogram represents the spatial variation of a regionalized phenomenon in quantitative terms. The semivariogram is the mathematical description of the relationship between the variance among each pair of observation (points) and the distance between these observations *(h)*. The experimental semivariogram function is given by Equation 1, where *N(h)* is the set of all pairwise, and $z(x_i)$ and $z(x_j)$ are data values at spatial locations x_i and x_j, respectively (Equation 2).

$$y(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[Z\left(x_i\right) - Z\left(x_j\right) \right]^2 \quad (1)$$

$$N\left(h\right) = \left\{ \left(xi, xj\right) : xi - xj = h \right\} \quad (2)$$

Following the calculation of the experimental semivariogram, a theoretical model must be set. This model is the one that better fits the calculated variance function ($\gamma\ (h)$) (Figure 1). Various semivariogram models are available in different computational software products that provide geostatistical tools. The most popular theoretical models among the literature are the spherical, exponential and Gaussian, as they are usually able to explain the spatial variability of most spatial events.

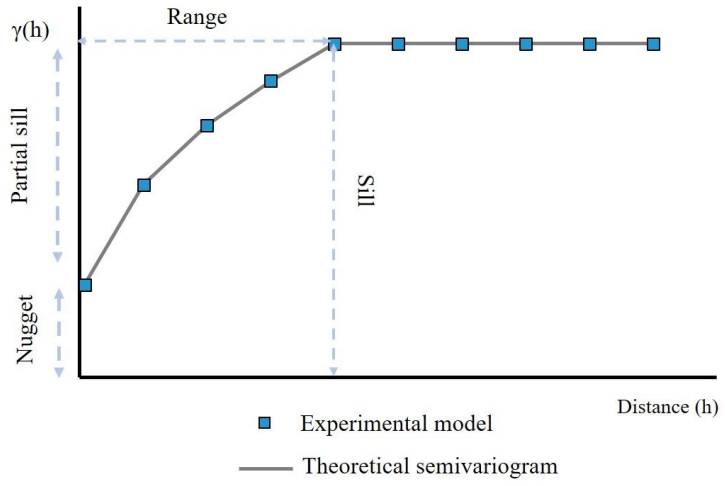

Figure 1: Representation of the semivariogram parameters
Source: Adapted from Isaaks and Srivastava (1989)

The parameters presented in Figure 1 are described as follows:

(1) Range (*r*): distance in which the observations are spatially correlated;

(2) Sill (C_0+C): the maximum variance value (γ) of the model. The range (r) represents the correspondent abscissa value;

(3) Nugget effect (C_0): the first point of the curve, i.e., where the curve reaches the ordinate axis (γ). The nugget effect reflects how short distances are alike or not. A high value to this parameter means that there is a high variance between close observations;

(4) Partial sill (C): the difference between the sill and the nugget effect.

These graphical parameters are used together with the distance matrix to estimate the variables' value at unsampled locations by Kriging with External Drift, for instance.

2.3.1 Cross-Validation

The cross-validation, also known as fictitious test point, is the technique that predicts values in the sampled locations using the corresponding values of the neighboring points and the theoretical semivariogram model. Thus, for each point there will be an original value (sampled) and the estimated value (Dubrule, 1983). In this way, statistical measures such as errors can be calculated. In this study, the validation step was carried out by splitting the sample into two sets. A set consisted of 60% of the sampled data and aimed at calculating a theoretical semivariogram model; 40% of the data was separated to validate the geostatistical estimations.

2.3.2 Kriging with External Drift (KED)

Many studies have a multivariate nature and the interdependency between the attributes must be incorporated into the analysis. In various situations, a variable may be assessed by means of an easier or an inexpensive way when compared to other means. That is, this interdependency between the attributes may be used as tools to predict variable's values from another variable.

KED enables the estimation of a primary variable given a secondary variable (Matheron, 1982). In this case, in addition to the observed points of the primary variable, the values of the secondary variable are previous known for the entire surface to be estimated (Olea, 1999). Considering $Z(x)$ the primary variable and $Y(x)$ the secondary one, both represent the linear dependency (Equation 3) according to Wackernagel (2010). Equation 4 provides the kriging estimator with external drift.

$$E\left[\mathrm{Z}\left(x\right)\right] = a_0 + b_1 Y\left(x\right) \tag{3}$$

$$Y\left(x_0\right) = \sum_{i=1}^{n} \lambda_i Y\left(x_i\right) \tag{4}$$

Where a_0 and b_1 are the coefficients to be implicitly estimated together with the variable *Z(x)*, and *Y(x)* is the value of the secondary variable.

KED is adequate when a main variable (primary variable) presents a relationship of dependence with the auxiliary variable (secondary variable). This study presented the estimation of *transit trip production* obtained from information of *the population* in the Salvador Metropolitan Area (Brazil). Therefore, this paper considers the parameters given by the theoretical semivariogram models of both the primary and the secondary variable.

2.4 Semivariogram Deconvolution

The method of obtaining a point-semivariogram $\gamma(h)$ using a regularized model $\gamma_V(h)$ (area or blocks) is known as the deregularization or deconvolution of a model $\gamma_V(h)$ (Journel and Huijbregts, 1978).

After choosing an initial point of a support model $\gamma^0(h)$, the deconvolution method is iterative and seeks to calculate the support model that minimizes the difference between the theoretical model of a regularized semivariogram and the point-support semivariogram model to be estimated (Goovaerts, 2008). The optimization criterion is the relative difference (D) between both semivariogram models. The $\gamma^1(h)$ model is chosen, if $D^1 < D^0$. This procedure is repeated until the lowest value of *D* is achieved (Goovaerts, 2006).

Techniques as the semivariogram deconvolution may be of interest to the application of geostatistics in the transportation planning field, considering the availability of aggregated data. This study presents a simplified method of semivariogram deconvolution applied to the estimation of aggregated travel data in Traffic Analysis Zones (TAZs).

2.5 Study area and dataset

The Salvador Metropolitan Area (SMA) has an extension of 4,375 km², 13 municipalities and a population of over 3.6 million inhabitants; however, approximately 2.7 million live in the city of Salvador (IBGE, 2010). That is, Salvador represents the city in the SMA where most of the trips are concentrated.

The first dataset consisted of the origin-destination dataset based on a home-interview survey carried out in 1995. This survey indicated 3,691,889 daily trips in the study area, with 22% concentrated in the morning peak (between 6 and 8 am). Approximately 55% of the trips referred to public transportation. In this first dataset, the variable of interest was the *transit trip production* and the explanatory variables referred to the socioeconomic characteristics of each unit area.

The 1995 origin-destination survey used the TAZs as unit areas. These TAZs were classified into four main regions: Consolidated Urban Area, Central Area, Seafront and Suburb. The Central Area and the Suburb are the most populated areas in Salvador (Figure 2), corresponding

to 46% of the total population of the city. Standardized values of the regarded variables were considered, as presented in the next section.

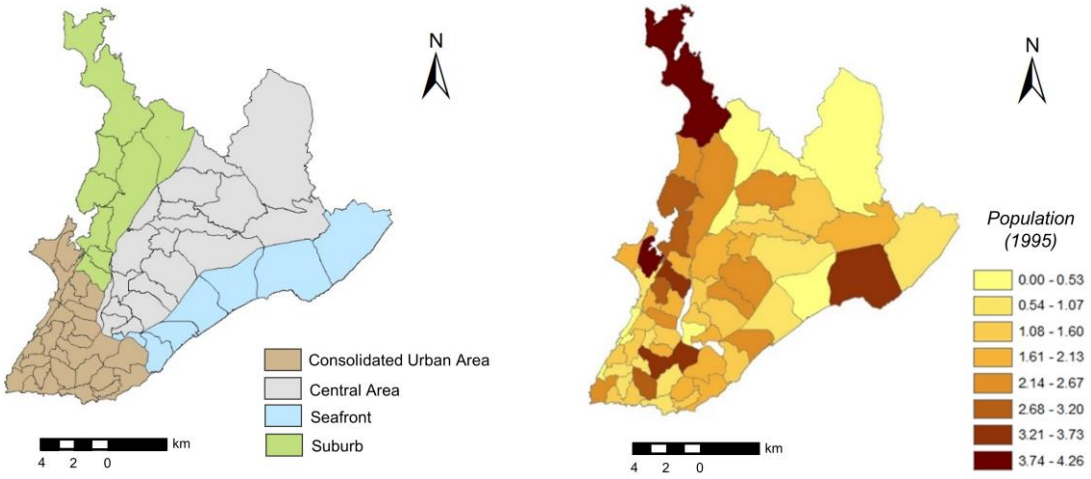

Figure 2: Salvador Metropolitan Area and its population per TAZ (Standardized values)

Source: SETPS (1995)

The second dataset consisted of the 2010 census from the *Instituto Brasileiro de Geografia e Estatística* (IBGE, 2010). The variable of interest was also the *transit trip production*. However, variables of travel demand are more particular and are not usually available in traditional census surveys. In order to forecast *the transit trip production in 2010*, the authors collected information on the socioeconomic features and used a calibrated model obtained from the first dataset (1995).

2.6 Method

The first step of method follows the idea of calibrating a linear model that predicts *transit trip production* by using an origin-destination survey from 1995. The outcome is an equation that uses as independent variables the most significant socioeconomic attributes. This calibrated equation is applied in the second step aiming at predicting the variable *transit trip production* using only socioeconomic data from 2010. Figure 3 demonstrates the proposed method, with five main steps identified.

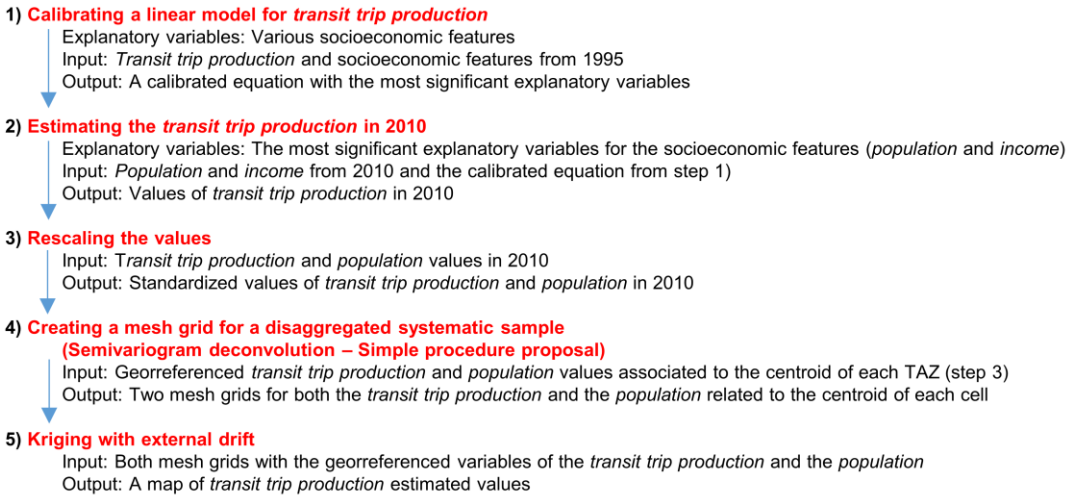

Figure 3: Proposed method

Considering that the dataset consists of different types of variables and different scale of values, the third step suggests that the data should be rescaled. The procedure first considered standardizing the variables as shown in Equation 5. Secondly, as the values could be negative, a normalization was conducted by using the smallest observed value and converting it to zero. Hence, the smallest value was added to all the values and caused the absence of negative values.

$$z = \frac{x - \mu}{\sigma} \tag{5}$$

Where z is the standardized value from x, μ is the mean and σ is the standard deviation.

In the fourth step, the systematic disaggregated sampling is proposed (Semivariogram deconvolution – Simple procedure). This method consists of: (1) defining a grid mesh with uniform spacing of 2,000 meters; (2) defining the centroid of each cell; (3) obtaining, for each centroid, the standardized values for the *transit trip production* and the variable with the greatest association with the former.

Considering that the input data consist of information related to TAZs, it is necessary to measure the occupation of each TAZ in each cell, in order to weigh the value of the *transit trip production* and the secondary variable in each grid:

- From the values of *transit trip production* in 2010 per TAZ, previously calculated in the former steps, a weight was assigned for each TAZ considering its area into each cell.
- Afterwards, a multiplication of each TAZ's weight by the *transit trip production* value of each respective TAZ and summing all the multiplications belonging to a particular cell provide the *transit trip production* value for this considered cell.
- This final value represented the cell's centroid.

The systematic sampling is an attempt to adequate the original data of the origin-destination survey to the application of geostatistics, given that one of the assumptions of such technique is the spatial continuity of the studied event. Aggregated data per unit area are associated to their centroids and, hence, are spatially discrete. Moreover, the geostatistic technique has been developed for regular geographic units. However, in general, travel data are associated to area with different shapes and sizes, that in most cases it is associated to census tracts in order to maintain the confidentiality of individuals.

The fifth step is the multivariate geostatistical procedure. It consisted of using KED as means of mapping the *transit trip production* also based on a secondary variable. The secondary variable is defined in the second step as the variable with the greatest association with the study variable.

It is important to highlight that the procedure intends to predict the *transit trip production* by only aggregated data from 2010 and a calibrated model of 1995. The use of available information, such as census data, makes the proposed methodology interesting in terms of Transportation Planning, especially for developing countries.

The software used in this research were the *IBM - Statistical Package for the Social Sciences (SPSS) version 24* to model a linear equation and *GeoMS 1.0* for the geostatistical processes of the semivariograms and kriging. The software *ArcGIS 9.3* was used to handle with the systematic sampling as well as obtaining graphical representations of the results.

3. Results and discussion

3.1 Calibrating a linear model

The model was initially calibrated with the 1995 origin-destination data and its socioeconomic variables. The Linear Regression Model pointed out two significant variables to predict *transit trip production*: the *population* and the *average income of the head of the family*. The calibrated equation corresponds to Equation 6.

$$Transit\,Trip\,Production = 0.17\,Population - 0.218\,Income \quad (6)$$

The Tables 1 and 2 show the main results of the linear model calculated from the 1995 data. The estimated parameters make sense when explaining the *transit trip production*. The *population* was the most significant variable to estimate the dependent variable.

able 1: Main statistical measures for the calibrated model

Sum of squares	
Regression	$3.75\ 10^9$
Residual	$1.38\ 10^8$
Total	$3.88\ 10^9$
R^2	0.964

Table 2: Calibrated model parameters

Explanatory variables	**Coefficient**	**t**
Constant	0	
Income	-0.218	-3.706
Population	0.177	30.39

3.2 Estimating the transit trip production in 2010

In order to use the calibrated equation (Equation 5) to predict the *transit trip production* in 2010, an adaption regarding the aggregation level of the second dataset was necessary, since the information was at a census tract level. Initially, the census districts from 2010 were made compatible with the TAZs from the origin-destination dataset from 1995.

According to Ortúzar and Willumsen (2011), the traffic zoning must be compatible with other administrative divisions. The areas must be as homogeneous as possible in relation to the land use and the population characteristics.

The compatibility of the aggregation level was conducted in the *ArcGIS 9.3* software. This Geographic Information System made it possible to overlap the census districts of 2010 and the TAZs from the origin-destination survey (1995). Afterwards, the census areas from each TAZ were selected and the *population* and *average income of the head of the family* were outlined. The values of the *average income of the head of the family* were adapted with regard to the updated minimum wage value, in which a 170% increase had occurred from 1995 to 2010.

Having defined the values of the *population* and the *average income of the head of the family* for the 2010 dataset, Equation 5 provided the estimated values of *transit trip production*. The next step addressed the standardization and normalization of the *transit trip production* and its most correlated variable (*population*).

3.3 Creating a mesh grid for a disaggregated systematic sample (Semivariogram deconvolution – A simple procedure)

In order to disaggregate the dataset, a systematic sampling was adopted. Figure 4 presents the cells and their respective centroids.

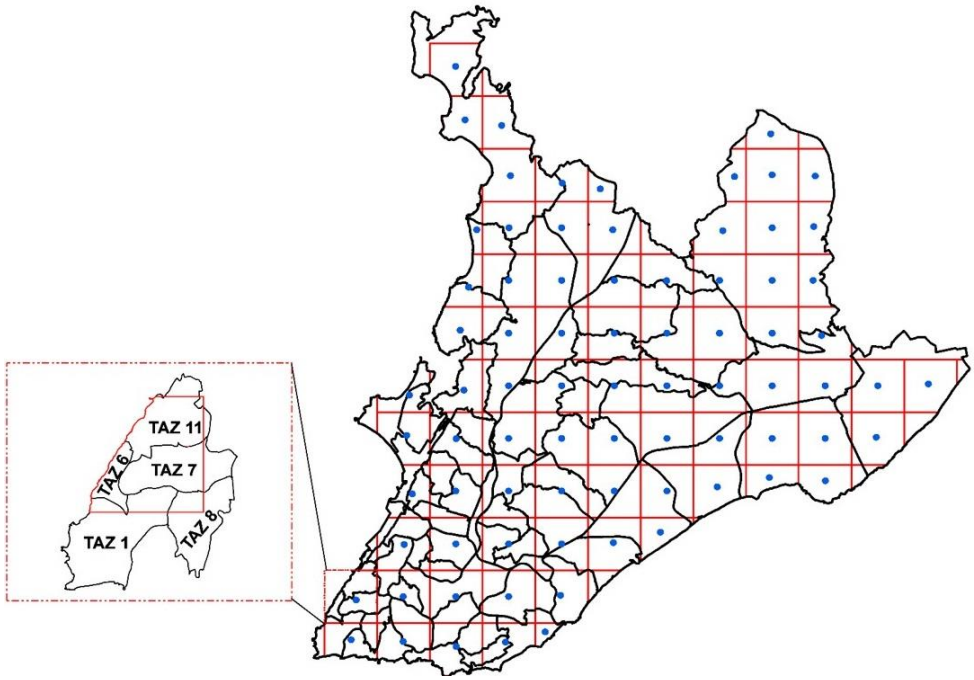

Figure 4: Mesh grid derived from the systematic disaggregated sample
Source: Rocha et al. (2016)

Figure 5 shows the scatter plot of the estimated variables in each cell's centroid (*Transit trip production* and *Population*). It is possible to note that there is a strong correlation between both variables.

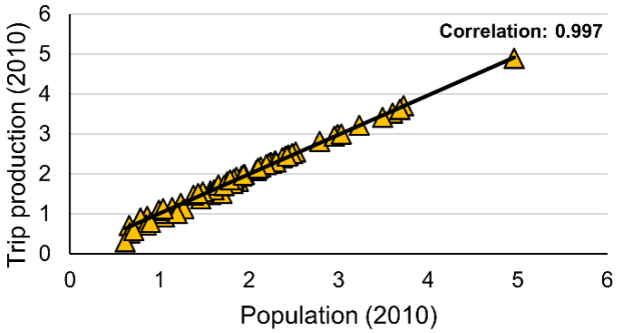

Figure 5: Scatterplot of the estimated values (Transit trip production and population)

3.4 Kriging with External Drift

In this step of the research, it is known that the *population* is the most correlated variable with the *transit trip production*. Thus, the population was interpolated for 2010 and used as auxiliary variable in the kriging procedure.

Aiming at verifying the spatial structure of the *transit trip production*, Figure 6 presents the standardized values and the spatial distribution. It can be noticed that the *transit trip production* tends to increase according to the proximity with the Suburb. This area has the largest population and lowest per capita income.

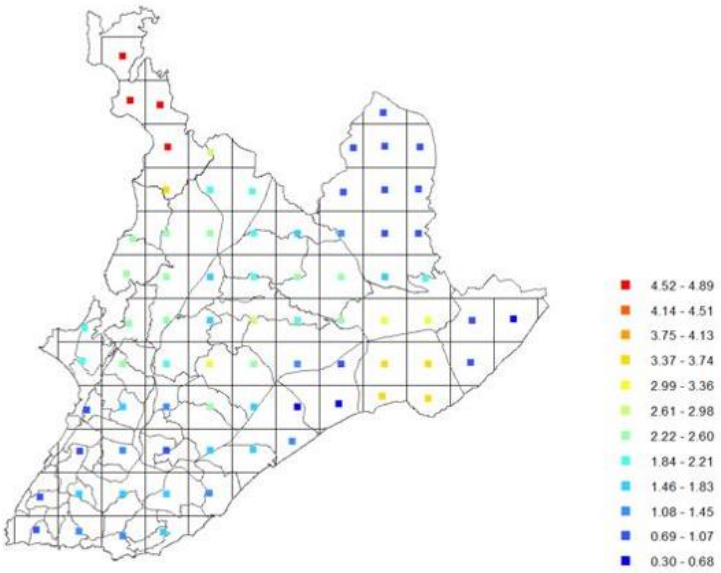

Figure 6: Distribution of standardized values at the Salvador Metropolitan Area: Transit trip production in 2010
Source: Rocha et al., 2016

Some parameters were established to calculate the experimental semivariograms of the *transit trip production* and the *population* variables. Afterwards, the model fitting provided the nugget effect (C_0) and the range (r) parameters. When fitting the theoretical curve, no sill was considered, as the semivariogram presented no stationary through the cut distance. That is, the spatial variability keeps increasing with greater distances between the pair of observations and, thus, they have higher variance values.

Table 3 presents a summary of the modeling parameters of each semivariogram and Figure 7 shows the experimental and theoretical models. The omnidirectional semivariogram was chosen and the spherical theoretical model was the one that best fitted the data.

Table 3: Graphical parameters from the variographic analysis

Variables - Omnidirectional	Nugget Effect (C_0)	Partial Sill (C)	Range (r)	Sill
Population 2010 per TAZ	0.159	2.440	28799.250	2.599
Transit trip production 2010 per TAZ	0.225	2.178	26591.500	2.403

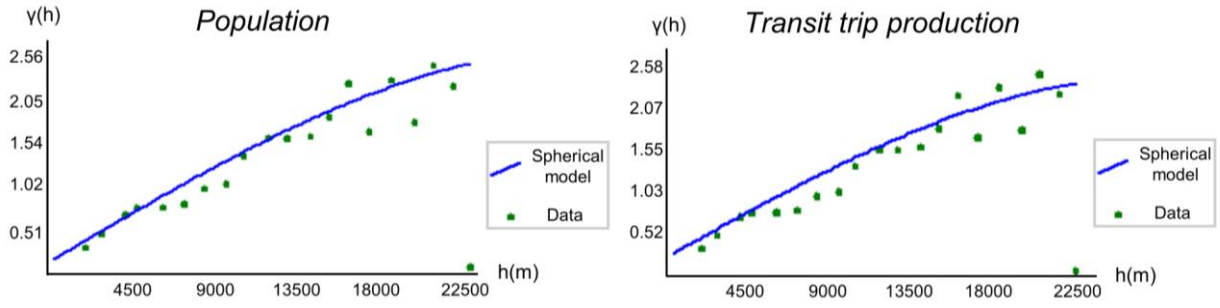

Figure 7: Theoretical semivariogram for the population and the transit trip production in 2010 - omnidirectional case

The results obtained from the cross validation are demonstrated in Table 4. It is noted that the model resulted in low values of errors. Besides, the observed and estimated values are highly correlated, considering the Pearson correlation.

Table 4: Results for the cross-validation

Variable	Pearson correlation	Mean error	Mean squared error
Population (2010)	0.880	0.004	0.279
Transit trip production (2010)	0.998	0.000	0.002

Finally, the travel demand data were estimated by kriging in locations where there was no prior knowledge of the values as well as in sampled locations. The interpolated map of the secondary variable (*population*) was an auxiliary to interpolate the primary variable. Figure 8 presents both maps and the similarity between both is clearly noticed. This is due to the strong association of the *population* and *transit trip production* features.

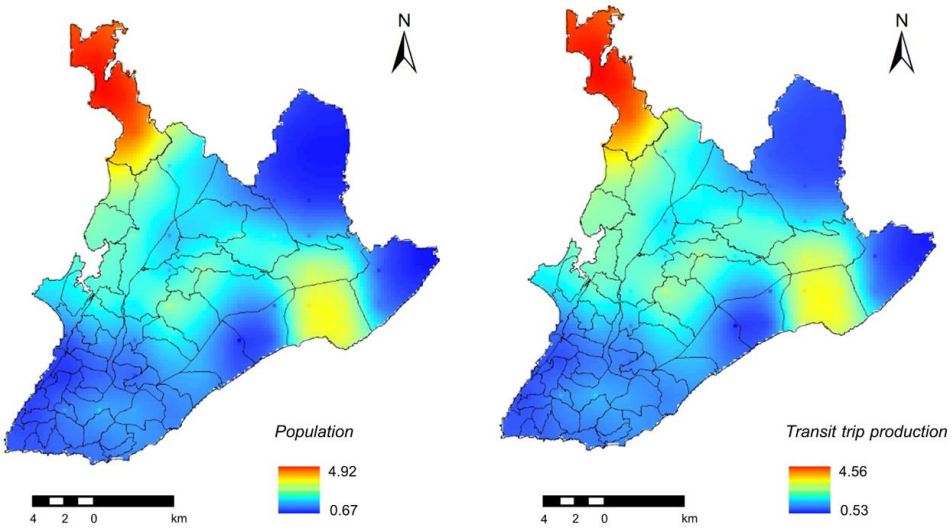

Figure 8: Kriging map for the population and the transit trip production in 2010

Another point to be considered is that there is a great concentration of trip production in the Suburb Area. This tendency decreases insofar as the Consolidated Urban Area is considered.

Figure 9 presents the spatial pattern distribution of observed values of *transit trip production* by areal support through choropleth maps. The representation can lead to an important limitation of choropleth maps, which concerns the biased visual interpretation. For instance, when it is understood that larger areas have more importance than smaller areas, it occurs an ecological fallacy. Such problems can be solved by creating continuous maps of the studied attribute (Goovaerts, 2008).

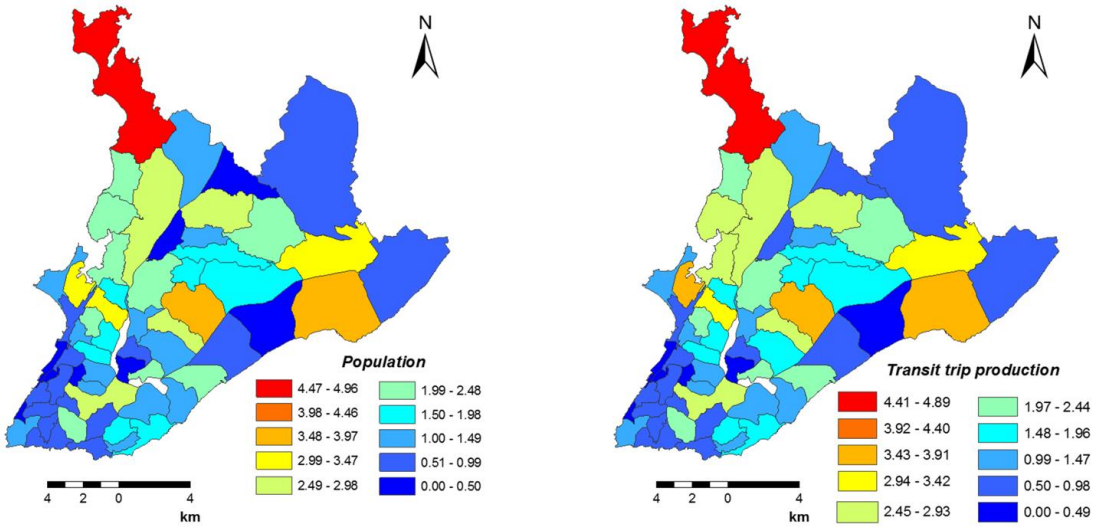

Figure 9: Observed data maps (aggregation of TAZ)

From the maps presented on Figure 9, a similar tendency of trips values between observed data by Traffic Analysis Zones and the estimated surfaces is observed, for example, in the Suburb Area. Ten classes of equal intervals were used in the legend of the maps presented. Despite the mentioned problems regarding the usage of area data and interpretation of choropleth maps, the visual comparison through maps (Figure 8 and Figure 9) is another form of validating, methodologically, the results of geostatistics approach.

4. Conclusions

Methods that consider the spatial correlation of the sampled values have been often used in the area of transportation planning. However, some specific data characteristics, related to this study area, need special attention. Different levels of data aggregation, as well as the presence of irregular areas, are some of the problems identified in the line of health research and transportation planning, for example.

Changes in the level of data aggregation, i.e. change of support, influence the results and cause the Modifiable Areal Unit Problem and ecological fallacy. These two phenomena can occur in spatial analysis due to the modification of the scale or the level of zone aggregation. Moreover, the usual unit areas in travel demand analysis, such as TAZs and census tracts, have irregular shapes. This is an obstacle to the application of Geostatistics in transportation studies, since this technique was developed to be applied in regular supports. Considering this, the present paper proposed to combine a disaggregated systematic sampling method (a simple procedure for semivariogram deconvolution) and a multivariate geostatistical technique (Kriging with External Drift) to estimate *transit trip production*.

Despite the methodological challenges, the results yielded in this research demonstrate a method that has the advantage of estimating values in sampled and non-sampled positions using KED, with the *transit trip production* as primary variable and the *population* as secondary variable. Additionally, the proposed systematic sampling enabled each cell of the created mesh grid (with a size of 2,000 x 2,000 meters each) to be associated with *transit trip production* and *population*

values that were originally at an aggregate level (TAZs and census tract level). This advantage is a significant contribution in the field of Transportation Planning as the data collection may be costly and time-consuming activity.

The kriging procedure presented good results considering the error measures and the correlation coefficient between the estimated and observed values. Besides, the positive results are in line with the study area context: 1) a higher number of *transit trip production* is seen from the Central Area to the Suburb; and 2) locations with lower income and larger population have more tendency to produce transit trips.

This paper is a first step towards the data disaggregation when considering travel demand data and spatial association. The proposed systematic sampling is a basic procedure that still needs to be refined and improved in order to be effectively implemented. The authors suggest, for future studies, the use of simulation techniques. However, the results indicated that the kriging technique is adequate for the intended purposes in this research, especially when considering variables that have similar spatial distribution, such as the case of the *population* and the *transit trip production.*

ACKNOWLEDGEMENTS

This research was supported by the National Counsel of Technological and Scientific Development (CNPq – 462713/ 2014 and 303645/2015-6 - Brazil) and State of Sao Paulo Research Foundation (FAPESP, 25035-1/2013- Brazil).

REFERENCES

Adjemian, M. K. Lin, C. and Williams, J. 2010. Estimating spatial interdependence in automobile type choice with survey data. *Transportation Research Part A*, 44, pp.661-675. doi: 10.1016/j.tra.2010.06.001.

Al-Taei, A. K. and Taher A. M. 2006. Prediction Analysis of Trip Production Using Cross-Classification Technique. *Al-Rafidain Engineering*, 14(4), pp. 51-63.

Ben-Akiva, M. E. Scott, M. R. and Bekhor, S. 2004. Route choice models. Human Behaviour and Traffic Networks. *Springer Berlin Heidelberg*, pp.23-45.

Bhat, C. and Zhao, H. 2002. The spatial analysis of activity stop generation. *Transportation Research Part B 36,* pp.557–575. doi: 10.1016/S0191-2615(01)00019-4.

Bhat, C. R. and Sener, I.N. 2009. A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units. *Journal of Geographical Systems* 11(3), pp.243–272. doi: 10.1007/s10109-009-0077-9.

Chen, X. Xiang, H. Chenfeng, X. Lei, Z. 2015. A Bayesian Stochastic Kriging Metamodel for Simultaneous Optimization of Travel Behavioral Responses and Traffic Management. In.: TRB 94th Annual Meeting Compendium of Papers. Washington. doi:10.1016/j.trpro.2015.06.056.

Ciuffo, B. F. Punzo, V. and Quaglietta, E. 2011. Kriging Meta-Modelling to Verify Traffic Micro-Simulation Calibration Methods. In.: TRB 90th Annual Meeting Compendium of Papers: Washington.

Dubrule, O. 1983. Cross Validation of Kriging in a Unique Neighborhood. *Mathematical Geology,* 15(6), pp. 687-699.

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 Journal of Statistics*, v. 6, pp. 514–527. doi: 10.4236/ojs.2016.63044.

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. *International Journal of Health Geographics*, 5(52). doi: 10.1186/1476-072X-5-52.

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

Ickstadt, K. Wolpert, R. L. and Lu, X. 1998. Modeling travel demand in Portland, Oregon. In: Practical Nonparametric and Semiparametric Bayesian Statistics. New York: Springer, pp. 305-322.

IBGE Brazilian Institute of Geography and Statistics. 2010. Demographic Census 2010. At: http://www.ibge.gov.br/ (last accessed on 20/08/2012).

Isaaks, E. H. and Srivastava, R. M. 1989. *Applied Geostatistics*. New York, Oxford University Press.

Journel, A. G. and Huijbregts, C. H. J. (1978). *Mining Geostatistics*. London: Academic Press.

Kassteele, J. van de. and Stein, A. 2006. A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output. *Environmetrics*, [S.l.], 17(4), pp.309-322. doi: 10.1002/env.771.

Kassteele, J. van de. and Velders, G. J. M. 2006. Uncertainty assessment of local NO2 concentrations derived from error-in-variable external drift kriging and its relationship to the 2010 air quality standard. *Atmospheric Environment*, [S.l.], 40(14), pp.2583-2595. doi: 10.1016/j.atmosenv.2005.12.023.

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

Lopes, S. B. Brondino, N. C. M. and Rodrigues da Silva, A. N. 2014. GIS-Based Analytical Tools for Transport Planning: Spatial Regression Models for Transportation Demand Forecast. *ISPRS International Journal of Geo-information*, 3, pp.565-583. doi: 10.3390/ijgi3020565.

Matheron, G. 1970. La Théorie des Variables Régionalisées et ses Applications. *Les Cahiers de Morphologie Mathématique de Fontainebleau*, Fascicule 5, Thome 1.

Matheron, G. 1982. *Pour une analyse krigeante des données regionalisées.* Report 732.

Mazzella, A. Piras, C. Pinna, F. 2011. Use of Kriging Technique to Study Roundabout Performance. *Transportation Research Record.* [S.l.], 2241. doi: 10.3141/2241-09.

Mahmoud, M. and Abdel, M. 2004. Cross classification trip production model for the city of Alexandria. *Alexandria Engineering Journal*, 43(2), pp. 177-189.

Olea, R. A. 1999. *Geostatistics for Engineers and Earth Scientists*. Massachusetts: Kluwer Academic Publishers, Norwell.

Ortúzar, J. D. and Willumsen, L.G. 2011. *Modelling Transport*. Wiley, 4rd. ed.

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

Páez, A. López, F. A. Ruiz, M. and Morency, C. 2013. Development of an indicator to assess the spatial fit of discrete choice models. *Transportation Research Part B,* 56, pp.217-233. doi:10.1016/j.trb.2013.08.009.

Papacostas C. S. and Provedouros, P. D. 1993. *Transportation Engineering and Planning*. 2rd. ed. Prentice Hall, New Jersey.

Pearcea, J. L. Stephen, L. R. Aguilar-Villalobos, M. Naeher, L. P. 2009. Characterizing the spatiotemporal variability of PM2.5 in Cusco, Peru using kriging with external drift. Atmospheric Environment. [S.l.], 43(12), p.2060–2069. doi: 10.1016/j.atmosenv.2008.10.060.

Pitombo, C. S. Salgueiro, A. R. Costa, A. S. G. and Isler, C. A. 2015. A two-step method for mode choice estimation with socioeconomic and spatial information. *Spatial Statistics,* 11, pp.45-64. doi: 10.1016/j.spasta.2014.12.002.

Rocha, S. S. Pitombo, C. S. and Salgueiro, A. R. 2016. Spatial Interpolation of Transit Urban Trips through an Artificial Systematic Disaggregated Sample (in portuguese). Revista Brasileira de Cartografia, 68 (4), pp.705–715.

Union of Passenger Transport Companies of Salvador – SETPS. 1995. Origin-Destination dataset based on a home-interview survey. Salvador, Brazil.

Wackernagel, H. 2010. *Multivariate Geostatistics: An introduction with applications*. 3rd ed. Springer.

Wang, J. Stein, A. Gao, B. and Ge,Y. 2012. A review of spatial sampling. *Spatial Statistics,* I, pp.1–14. doi:10.1016/j.spasta.2012.08.001.

Webster, R. and Oliver, M. A. 2007. *Geostatistics for environmental scientists*. 2rd ed. Wiley: New York.

Yoon, S. Y, Ravulaparthy, S. K. and 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. *Transportation Research Part A: Policy and Practice,* [S.l.], 68, pp.3-17. doi: org/10.1016/j.tra.2014.01.004.

Zhang, D. and Wang, X. 2013. Traffic volume estimation using network interpolation techniques: an application on transit ridership in NYC Subway System. New York: Final Report.

Received in February 16, 2017.

Accepted in August 28, 2017.