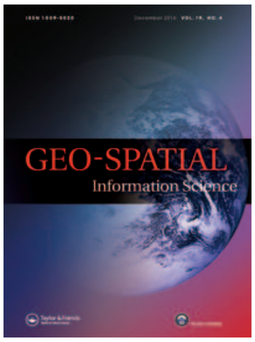


Taylor & Francis
Taylor & Francis Group

Geo-spatial Information Science

ISSN: 1009-5020 (Print) 1993-5153 (Online) Journal homepage: https://www.tandfonline.com/loi/tgsi20

Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study

Anabele Lindner, Cira Souza Pitombo, Samille Santos Rocha & José Alberto Quintanilha

To cite this article: Anabele Lindner, Cira Souza Pitombo, Samille Santos Rocha & José Alberto Quintanilha (2016) Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study, Geo-spatial Information Science, 19:4, 245-254, DOI: 10.1080/10095020.2016.1260811

To link to this article: https://doi.org/10.1080/10095020.2016.1260811

Published online: 07 Dec 2016.

GEO-SPATIAL INFORMATION SCIENCE, 2016
VOL. 19, NO. 4, 245–254
http://dx.doi.org/10.1080/10095020.2016.1260811

Taylor & Francis
Taylor & Francis Group

Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study

Anabele Lindner[a], Cira Souza Pitombo[a], Samille Santos Rocha[a] and José Alberto Quintanilha[b]

[a]Department of Transportation Engineering, São Carlos School of Engineering, University of São Paulo, São Carlos, Brazil; [b]Department of Transportation Engineering, Polytechnic School, University of São Paulo, São Paulo, Brazil

ABSTRACT
Studies in transportation planning routinely use data in which location attributes are an important source of information. Thus, using spatial attributes in urban travel forecasting models seems reasonable. The main objective of this paper is to estimate transit trip production using Factorial Kriging with External Drift (FKED) through an aggregated data case study of Traffic Analysis Zones in São Paulo city, Brazil. The method consists of a sequential application of Principal Components Analysis (PCA) and Kriging with External Drift (KED). The traditional Linear Regression (LR) model was adopted with the aim of validating the proposed method. The results show that PCA summarizes and combines 23 socioeconomic variables using 4 components. The first component is introduced in KED, as secondary information, to estimate transit trip production by public transport in geographic coordinates where there is no prior knowledge of the values. Cross-validation for the FKED model presented high values of the correlation coefficient between estimated and observed values. Moreover, low error values were observed. The accuracy of the LR model was similar to FKED. However, the proposed method is able to map the transit trip production in several geographical coordinates of non-sampled values.

ARTICLE HISTORY
Received 29 April 2016
Accepted 13 August 2016

KEYWORDS
Travel demand forecasting; Geostatistics; Kriging; spatial statistics

1. Introduction

Travel demand forecasting models usually consider explanatory variables, such as Traffic Analysis Zone (TAZ) characteristics, urban environments, transport facilities, travel features, and individual/household factors (Ortúzar and Willumsen 2011) to estimate the trip generation, trip distribution, mode choice, and route choice. These are the four major model components of a travel demand forecasting process known as the sequential Four-Step Model (Ortúzar and Willumsen 2011). The focus of this paper is the trip generation step. Trip generation estimates the number of trips to (attraction) and from (production) in a TAZ. More specifically, this study addresses the trip production. The aggregated trip production model estimates the number of trips originating in a TAZ, whereas the trip attraction model estimates the number of trips to a particular TAZ.

The most common models to estimate trip generation are Multiple Linear Regression (MLR) and Cross-Classification (CC). These two methods can be acceptable to some extent in terms of transportation planning. However, some critical issues are found in each method. On one hand, in the case of MLR, the estimated number of trips is a continuous variable with the assumption of a normal distribution. On the other hand, the CC method estimates travel rates per group of households via the social and economic characteristics of the household. However, the arbitrary choice of independent variables, and consequently the household strata, can be a critical problem (Chang et al. 2014).

Despite the mentioned limitations, MLR and CC are representative methodologies used for this step. They have been widely used for empirical studies and have shown acceptable efficiency for years considering the planning perspective, especially if information regarding the spatial location of the variables is not taken into account. Considering technological progress and the availability of geo-referenced information, spatial analysis of transportation demand forecasting is a potential research subject (Páez et al. 2013).

Significant developments have affected the travel modeling approach and process, eg Geographic Information System (GIS) used in the forecasting process. GIS allows the user to handle and access relevant data, and it employs a fundamental concept in geography, ie nearer objects share more similarities than objects farther apart (Tobler 1979). As a consequence, similar variable values will tend to occur in nearby locations, eg a lower income municipality in a remote region may be neighbors with other low-income municipalities.

CONTACT Anabele Lindner ✉ anabele@usp.br

This spatial clustering implies that many samples of geographic data will no longer satisfy the usual statistical assumption of independence of observations. Thus, the object localization is very important for spatial data analysis (Anselin 1992).

Studies in transportation planning routinely use data in which location attributes are an important source of information. These studies are associated with variables spatially positioned both in an absolute sense (coordinates) and in a relative sense (spatial arrangement, distance), such as densities of residential and socioeconomic activities, proximity between TAZs, and the transportation network.

Regarding the spatial analysis of travel demand modeling, some researchers have realized that travel behavior is correlated to spatial travel features. Bhat and Zhao (2002) highlighted the spatial aspects that need to be recognized when modeling travel demand and proposed a Multi-Level Mixed Logit Model to address these spatial issues. Bhat and Sener (2009) proposed a multivariate logistic distributed copula-based approach to address the spatial dependency and heteroscedasticity issues in binary discrete choice models in travel demand modeling. Recently, Páez et al. (2013) introduced a new indicator of spatial fitness that could be applied to discrete choice models to estimate door-to-door travel choices. Peer et al. (2013) used geographically weighted regression to estimate speed correlations across links and estimated the departure time with choice models.

Concerning spatial statistics, Geostatistics, enables professionals to consider spatial autocorrelation when modeling a problem and to predict the value of a variable in locations where it is unknown or unobserved. Usually, Geostatistics is applied to the cases in which spatial continuity is apparent. Despite this limitation, geostatistical modeling has been used for spatially discrete data for many years (Goovaerts 2009). Generally, travel data are spatially discrete. To deal with this limitation, transportation variables need to be adapted, considering that they are generally discrete variables and have no spatial continuity.

In the literature, however, studies on the application of Geostatistics concerning transportation issues are mainly from traffic engineering. Miura (2010) presented an approach for predicting car travel time by Kriging. This prediction method was shown to be effective for urban districts with links having changeable travel times owing to congestion. Zou et al. (2012) proposed an improved distance metric called Approximate Road Network Distance for solving the problem of the invalid spatial covariance function in Kriging caused by the non-Euclidean distance metric. Following this line of research, recent studies have shown that Geostatistics is able to estimate transportation demand variables and to explain the spatial distribution using maps of Kriging predicted values (Pitombo et al. 2010; Pitombo, Costa, and Salgueiro 2015; Pitombo et al. 2015).

The main aim of this paper is to estimate transit trip production using Factorial Kriging with External Drift (FKED) based on an aggregated data case study. The FKED method consists of a sequential application of Principal Components Analysis (PCA) and Kriging with External Drift (KED), in an aggregate analysis of TAZs in São Paulo city (Brazil). This article is organized into four sections besides this introduction. Section 2 presents the materials (techniques, study area, data-set) and the method. Section 3 presents the results, and finally, Section 4 describes the main conclusions.

2. Materials and method

2.1. Techniques: Geostatistics

Geostatistics was developed as an alternative method to explore events in which the values of a given variable are associated with geographic coordinates. This approach takes general spatial statistics into account because it estimates a continuous surface using a data-set that may be regularly or irregularly spatially distributed. The main point of using Geostatistics is to characterize the spatial (and/or spatial/temporal) dispersion of an event, assessing uncertainty parameters, determining its spatial variability, and obtaining a continuous surface estimation. Geostatistics is better defined as in the following steps: (1) variographic analysis, (2) cross-validation, and (3) Kriging.

The primary tool in geostatistical modeling is the semivariogram, which graphically represents a regionalized variable. The semivariogram function was originally defined by Matheron (1963) and is given by Equation (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.

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

Moreover, the representation of an experimental semivariogram requires further understanding of graphical aspects. Some measures include lag distance and tolerance, cut distance, and direction. Another step of variographic analysis is to model a theoretical semivariogram based on the experimental one. The parameters obtained by this step are: the nugget (C_0), the spatial variation/partial sill (C_1), the sill (C), and the range (r). These parameters are better understood when they are graphically represented in the semivariogram (Figure 1). Furthermore, the major and minor directions can be detected by analyzing semivariograms of all directions.

The next step of geostatistical modeling is the cross validation (fictitious test point), which comprises an analysis of errors, ie it measures the uncertainty of estimation. This test is performed by considering the observed and estimated values in previous sampled geographic coordinates. Cross-validation proceeds by successively removing each validated sample value and estimating a new value using (n - 1) observations. The

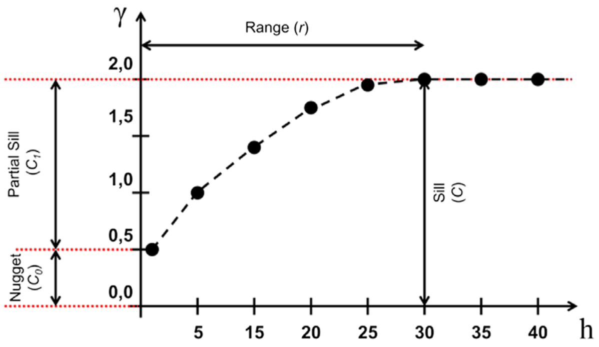


Figure 1. Graphical parameters of a semivariogram.
Source: adapted from Wackernagel (2003).

difference between the estimated and observed value is given by Equation (2):

$$\Delta = Z(x_\alpha) - Z^*(x_{[a]}) \tag{2}$$

where $Z(x_a)$ is a sample value and $Z^*(x_{[a]})$ is an estimated value in location $x_{[a]}$ (Wackernagel 2003).

The geostatistical method follows the Kriging estimation, which is a linear prediction represented by matrix calculus. The aim of Kriging is to predict estimates of one of more variables with a minimum error and variance (optimizing the model) using the parameters defined in the theoretical semivariogram of the major and minor directions, as illustrated in Figure 1. The most common univariate Kriging methods are Simple Kriging, Universal Kriging, and Ordinary Kriging. This paper uses the multivariate method of FKED, which is derived from Ordinary Kriging concepts to emphasize the benefit of including explanatory variables in Geostatistics.

KED is a multivariate geostatistical method that combines the use of multiple variables to co-estimate a correlated variable. FKED follows the same approach using auxiliary variables as the factors/components extracted from a Factorial Analysis or a PCA. This research paper uses *PCA component* 1 as the secondary variable to estimate the correlated primary variable (*transit trip production*) through the FKED approach. Therefore, both multivariate analysis methods, ie PCA and KED are sequentially described here.

FKED can be considered as a link between the classical multivariate analysis and the conceptual multivariate geostatistical method. PCA builds a number of components or regionalized factors, which reflect the main features of the multivariate information of an event (Goovaerts 1992), and which can estimate each weighting and the dependent variable through semivariogram co-localization information. Specific tendencies concerning the occurrence of a phenomenon can be detected using the Factorial Kriging technique (Batista et al. 2001).

2.1.1. PCA

PCA is implemented to assess interrelations among a large number of variables and to understand the variables in terms of their common dimensions, defined as components (Hair et al. 2010). The main idea of the PCA is to reduce the dimension of an associated dataset into non-correlated factors (components), preserving its variance (Jolliffe 2002). The advantages derived from reducing the data are that relevant information about significant variables is retained, and there is an improvement in complex data structures (Sanguansat 2012).

The mathematical formulation associated with PCA is based on a variance–covariance/correlation matrix (matrix ***S***). The variance–covariance matrix is used for data in the same scale of measurement, while the correlation matrix considers data measured at different scales. The variance–covariance denotes the data dispersion. Given a matrix ***S*** ($m \times n$), m is the number of observations, n is the number of variables, and the principal components are yielded by the definition of the matrix ***S*** eigenvectors ($\boldsymbol{v}$).

The eigenvectors ($\boldsymbol{v}$) are calculated as a function of eigenvalues (λ) of the matrix ***S***. ***I*** is the identity matrix, and the eigenvalues (λ) of the matrix ***S*** are scalars that satisfy the characteristic equation (Equation (3)).

$$|\boldsymbol{S} - \lambda \boldsymbol{I}| = 0 \tag{3}$$

Each eigenvalue is associated with an eigenvector, which can be obtained from Equation (4).

$$(\boldsymbol{S} - \lambda \boldsymbol{I})v = 0 \tag{4}$$

In the general case, the eigenvalue matrix is diagonal, where the number of eigenvalues is equivalent to a square matrix ($n \times n$). A new set of variables can be derived by multiplying the eigenvectors and the vectors of the original values. Hence, a square matrix ***A*** is composed using eigenvectors as columns of the matrix. The new set of variables is a linear combination of the original variables, derived from Equation (5).

$$\boldsymbol{W} = \boldsymbol{XA} \tag{5}$$

where *matrix* ***A*** comprises the eigenvectors and ***X*** is the original data vector. The principal components are selected by verifying the fraction of the variance which is explained by a specific component. The higher its proportion, the more relevant the component to the analysis is.

This paper uses *component* 1 as an input of secondary variable to estimate values from a KED method. Another point to be considered is that in this paper, the estimation by the principal components can be given by Equation (3) for a standardized case (Equation (6)):

$$\hat{Y} = \sum_{i=1}^{n} \left(v_i \times \frac{X_i - \bar{X}_i}{S_i} \right) \tag{6}$$

where $\hat{Y}$ is the estimated value for the dependent variable and v_i is the eigenvector associated with the standardized value of each explanatory variable $\left(\frac{X_i - \bar{X}_i}{S_i}\right)$.

2.1.2. KED

Considering the integration of two correlated variables ($Z(x)$ and $Y(x)$) that express the same attribute, Equation (7) defines the basic concept of the KED estimation as a linear function.

$$E^{*}\left[Z(x_0\right] = a_0 + b_1\ Y(x_0) \tag{7}$$

where $Y(x_0)$ is an external drift function to estimate the primary variable $Z(x_0)$ based on the estimated values x_0. KED is given by two basic constraints (Equations (8) and (9)).

$$Y(x_0) = \sum\nolimits_{i=1}^{n} w_i Y(x_i) \tag{8}$$

$$\sum\nolimits_{i=1}^{n} w_i = 1 \tag{9}$$

where x_i is the observed value and w_i is the weight of each value. The estimation variance, as well as the correspondent weights, are yielded by means of the following matrix in Equation (10) (Wackernagel 2003).

$$\begin{bmatrix} C & 1 & Y \\ 1^{\mathrm{T}} & 0 & 0 \\ Y^{\mathrm{T}} & 0 & 0 \end{bmatrix} \begin{bmatrix} w \\ -\mu \\ -\mu \end{bmatrix} = \begin{bmatrix} C_0 \\ 1 \\ Y_0 \end{bmatrix} \tag{10}$$

where C is the covariance function and μ is the Lagrange multiplier that minimizes the estimation variance and both constraints (Equations (8) and (9)).

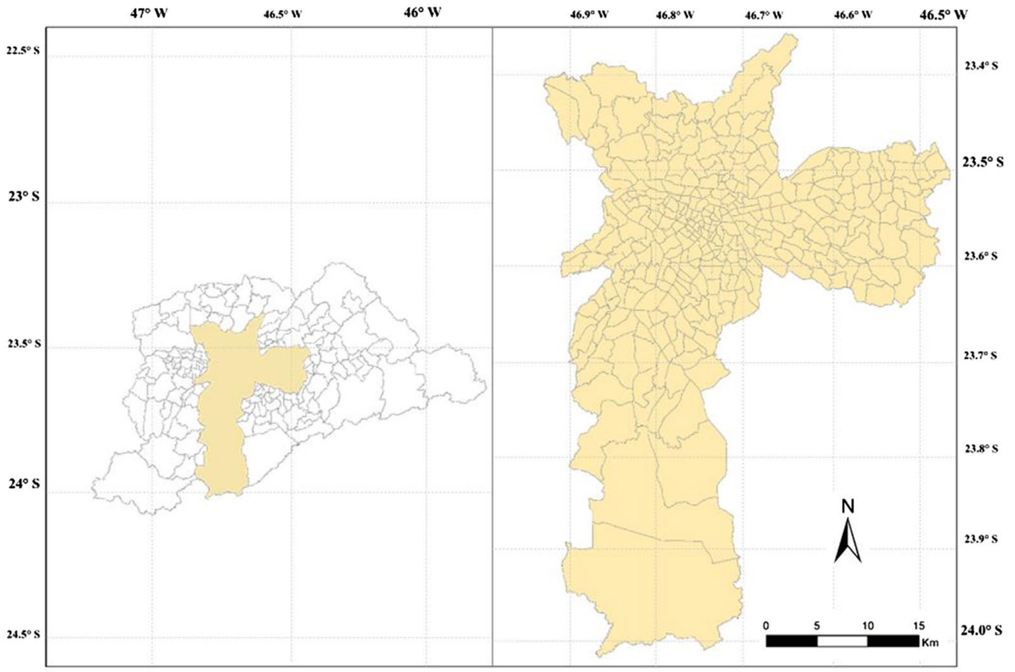
2.2. Study area and data-set

São Paulo is the most populated city in Brazil. Its metropolitan area has a population of over 20 million residents; however, approximately 11.5 million live in the city of São Paulo (IBGE 2010). This research assesses an origin–destination data-set based on a home interview survey carried out in 2007. The original sample includes information from 30,000 households in the São Paulo Metropolitan Area (SPMA). The study area corresponds exclusively to the city of São Paulo divided into 320 TAZs, as shown in Figure 2.

The database consists of 23 socioeconomic variables (Table 1) associated with TAZ units in addition to variables related to trip production. For the purpose of applying FKED, the method considered the values of socioeconomic variables divided by the area of each TAZ.

2.3. Method

Figure 3 demonstrates the proposed method, where three main steps can be identified. The first step was followed to detect the components of the entire data-set and to define its nomenclature. As the second step of the method, *PCA component* 1 was used as input data of the secondary variable to estimate the FKED. Hence, the primary variable (recognized as *transit trip production*) was seen as the most correlated variable with the



(a) São Paulo Metropolitan Area

(b) São Paulo

Figure 2. Representation of the study area: (a) São Paulo Metropolitan Area, (b) São Paulo, Brazil.

Table 1. Set of original variables.

Variable's relation	Description
Income	Household income below $ 350
	Household income from $ 350 to 700
	Household income from $ 700 to 1400
	Household income from $ 1400 to 2626
	Household income above $ 2626
	Average family income ($)
Employment	Total employment
	Employment in the service sector
	Industries
	Commerce
Population	Individuals aged under 10
	Individuals aged 11–17
	Individuals aged 18–39
	Individuals aged 40–59
	Individuals aged over 60
	Number of men
	Number of women
	Population
Vehicle ownership	Private cars
	Households without cars
	Households with one car
	Households with two or more cars
Education	School enrollment

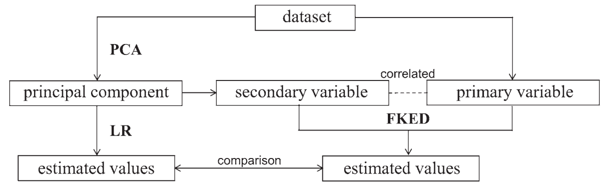


Figure 3. Proposed method.

secondary variable. Besides this, *component* 1 explains the variability of the entire data-set better.

The third step was related to variable estimation through traditional Linear Regression (LR) using the eigenvector of *PCA component* 1. LR models are justified to measure the increase in performance that the proposed method (FKED) can bring to *transit trip production* forecasting. Besides that, LR models are well recognized and can be easily used in travel demand forecasting, especially for trip generation. The comparison between both approaches (traditional non-spatial and spatial) was made considering various goodness-of-fit measures.

The *relative error* or *percent error* is a goodness-of-fit measure, used for a single pair of observed-estimated measures. The relative error (RE) is calculated as follows:

$$\mathrm{RE} = \frac{x_i - y_i}{y_i} \tag{11}$$

where x_i is the estimated measure and y_i is the observed measure.

The histogram of relative error reflects an aggregated way to represent this measure. It is a visual method for observing the existence of high frequencies of erroneous observations around null values (positively skewed distribution), which suggests that the model presents good predictive capability. For further analysis, other measures are assessed, such as those shown in Equations (12)–(18).

$$\mathrm{RMSE} = \sqrt{\frac{\sum (x_i - y_i)^2}{N}} \tag{12}$$

$$\mathrm{MSE} = \frac{1}{N} \sum (x_i - y_i)^2 \tag{13}$$

$$\mathrm{MAE} = \frac{1}{N} \sum (x_i - y_i) \tag{14}$$

$$\mathrm{PCC} = \frac{1}{N-1} \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{\mathrm{SD}_x \mathrm{SD}_y} \tag{15}$$

$$\mathrm{SD}_x = \sqrt{\frac{\sum (x_i - x)^2}{N}} \tag{16}$$

$$\mathrm{SD}_y = \sqrt{\frac{\sum (y_i - y)^2}{N}} \tag{17}$$

$$\mathrm{SE} = \frac{SD}{\sqrt{N}} \tag{18}$$

where RMSE is the root mean square error; x_i is the estimated measure; y_i is the observed measure; N is the number of measures, MSE is the mean squared error; MAE is the mean absolute error; PCC is the correlation coefficient; $\bar{x}$ and $\bar{y}$ are the averages; SD_x and SD_y are the standard deviations; SE is the standard error.

The computing applications used in this research were the IBM – Statistical Package for the Social Sciences (SPSS) Version 22 and the software GeoMS 1.0 for the geostatistical calculation and definition processes of experimental and theoretical semivariograms and Kriging. The software ArcGIS 10.1 was used in order to obtain graphical representations of the results.

3. Results and discussions

This section presents the main results obtained from the following steps:

- The estimation of the FKED method in Section 3.1;
- An LR estimation according to the traditional non-spatial approach in Section 3.2;
- A comparison between the former and the latter validations in Section 3.3.

3.1. FKED

Considering the latent root criterion to extract the components, *PCA component* 1 explains approximately 48% of the data variability in Table 2, which shows the components, their variance percentages, cumulative variance percentages and respective designations or nomenclatures considering the eigenvector matrix.

The designation or nomenclature of each component was determined by analyzing the eigenvector matrix, as presented in Table 3. The variables with larger component scores depict the component to a greater extent. To determine the importance of each variable to each component, the score for the cutoff point was set as greater than or equal to 0.80 (as shown in bold in Table 3).

PCA component 1 presents high values for eigenvectors related to original variables such as the *number of households with low income per TAZ*, *population*, *population of younger individuals*, and *number of households without one car*. This variable group is associated with low-income aspects. However, *PCA component* 2 consists of original variables that represent the high-income population (high values of scores to variables as *household income above $2626*; *number of private cars per TAZ*; *households with two or more cars*). The original values of *PCA component* 3 are associated with employment features, such as *total employment*, *employment in the service sector* and *commerce*. Finally, the original variable *school enrollment* has a high score in *PCA component* 4, suggesting the nomenclature adopted in this paper. As previously mentioned, *PCA component* 1 was selected taking into account the explained variance and the correlation with transit trip production. Using population features as explanatory variables to estimate trip production is well known in the literature on travel demand (Chang et al. 2014; Schmöcker et al. 2005).

Table 2. Explained variance and description of the components.

Principal component	Explained variance (%)	Accumulated variance (%)	Description
1	47.8	47.8	Low-income population
2	22.8	70.6	High-income population
3	12.9	83.5	Employment
4	4.2	87.7	School enrollment

Table 3. Component scores for each component.

	Component			
Variable (TAZ density per area)	1	2	3	4
Household income below $ 350	**0.87**	−0.15	−0.09	0.01
Household income from $ 350 to 700	**0.94**	−0.13	−0.07	−0.01
Household income from $ 700 to 1400	**0.91**	0.16	−0.03	0.03
Household income from $1400 to 2626	0.34	0.78	0.15	0.13
Household income above $ 2626	−0.14	**0.93**	0.13	0.02
Total employment	0.00	0.19	**0.93**	0.21
Employment in the service sector	0.00	0.19	**0.93**	0.21
Industries	−0.11	−0.06	0.65	−0.35
Commerce	0.01	0.01	**0.92**	−0.03
Average family income ($)	−0.25	0.59	0.36	0.36
Individuals aged under 10	**0.93**	0.04	−0.12	−0.04
Individuals aged 11–17	**0.91**	0.11	−0.14	−0.06
Individuals aged 18–39	**0.90**	0.36	0.07	0.09
Individuals aged 40–59	0.66	0.71	0.07	0.07
Individuals aged over 60	0.41	**0.82**	0.12	0.08
Population	**0.88**	0.47	0.02	0.05
School enrollment	0.11	0.10	0.21	**0.88**
Private cars	0.29	**0.94**	0.06	0.02
Number of men	**0.91**	0.40	0.02	0.05
Number of women	**0.84**	0.53	0.03	0.05
Households without cars	**0.86**	0.12	0.21	0.18
Households with one car	0.59	0.74	0.11	0.07
Households with two or more cars	0.01	**0.94**	0.01	−0.04

After selecting the secondary variable (*PCA component* 1), the spatial distribution of the variables concerned was analyzed. Figure 4 presents the distribution of the population density of each TAZ and the spatial distribution of *PCA component* 1, which expresses the low-income population. It can be observed that the central, central-north, east, and central-west areas have the highest population density (Figure 4(a)). The same distribution is noticed in the map of *PCA component* 1 (Figure 4(b)).

In this research, the semivariograms for the primary and secondary variables were obtained for an omnidirectional case and the main direction of 90°, respectively. According to each direction, the semivariograms use pairs of observations in the respective directions. Axis *y* represents the variance, whereas axis *x* represents the distance (meters). The step to calculate the experimental semivariograms and obtain their parameters is a preliminary basis for defining the spatial characteristics of a variable. A good spatial structural semivariogram implies that the data can be represented through a theoretical model.

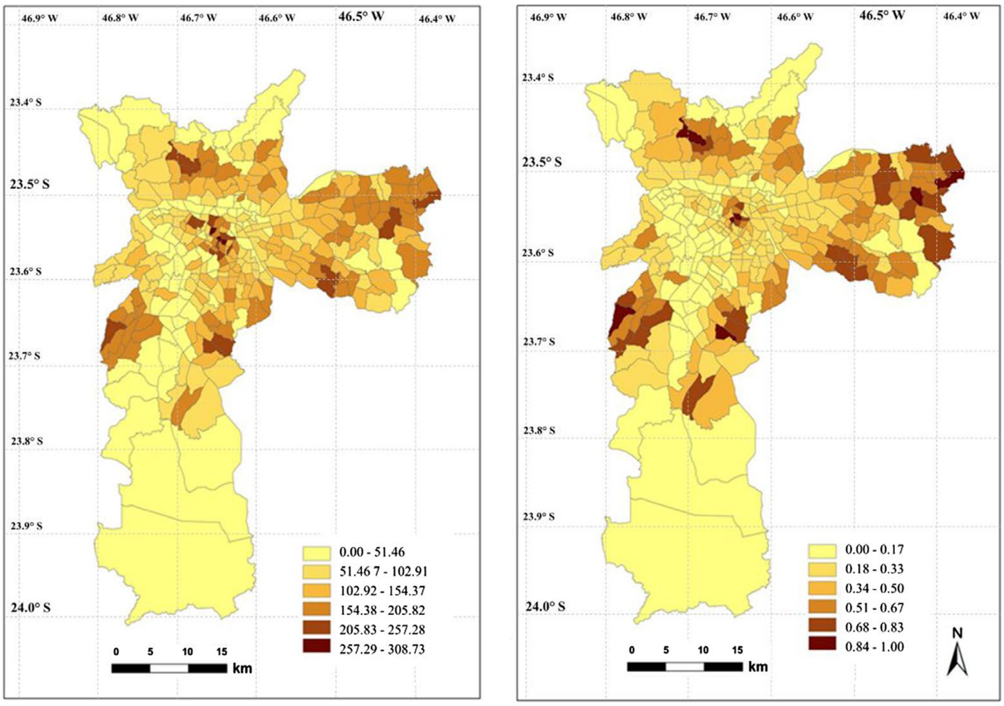
Figure 5 presents the theoretical omnidirectional semivariogram for the variable *transit trip production* (in Figure 5(a)) and the theoretical semivariogram for *PCA component* 1 in the 90° axis (in Figure 5(6)). The theoretical semivariogram model was selected based on visual inspection of the empirical semivariogram. The points of the semivariogram represent the average of variance (γ) in each paired observation with a lag distance of *h*, while the line determined by the sill refers to the average variance of the points in the semivariogram. Table 4 shows the graphical parameters of the theoretical semivariograms.

The parameters presented in Table 4 with the data-set provide the input for weighting and calibrating a geostatistical model through FKED. In other words, the theoretical semivariograms of the primary and secondary variables are used to map the *transit trip production* estimations, which is the primary variable in the FKED and the one most correlated with the secondary variable (*PCA component* 1 called *low-income population*).

Figure 6 presents the map of the FKED estimation of the primary variable, featuring a spatial distribution pattern similar to the maps of the population density and *PCA component* 1 (Figure 4). The Kriging map provides enough evidence to conclude that there is a larger transit trip production trend in areas with a low-income population and higher population density. These areas are located mainly in the center, west and eastern parts of the city.

3.2. Linear Regression: traditional non-spatial approach

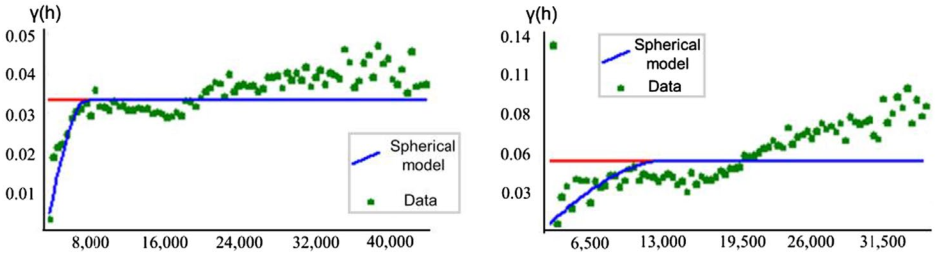
In this paper, PCA was used with LR to predict *transit trip production*. This variable was estimated based on *PCA component* 1 as the explanatory variable. This



(a) Population Density

(b) *PCA component* 1

Figure 4. Distribution of population density (a) and spatial distribution of *component* 1 (b) in São Paulo.



(a) Transit Trip Production - Omnidirectional

(b) *PCA component* 1 - 90°

Figure 5. Theoretical semivariogram models for *transit trip production* (a) and for *PCA component 1* (b).

Table 4. Parameters of the semivariogram models.

Variables	Theoretical model	C_0	C_1	Sill (*C*)	Range (*r*)
Component 1	Spherical	0.003	0.046	0.049	9697.4
Transit trip production	Spherical	0.000	0.035	0.035	4249.9

methodological step was used to compare the proposed spatial method to a usual approach in travel demand forecasting (LR). The parameters of the linear model are described in Table 5.

As expected, *PCA component* 1 (*low-income population*) and *transit trip production* are directly related and the R^2 value could be considered significant for travel forecasting. However, there are some drawbacks of this traditional approach: (1) For future estimations, the structure of this model is awkward for day-to-day use, because its explanatory variable is derived from PCA. A model which directly uses the original variables would be easier to implement and understand. (2) This is a non-spatial procedure and it is not possible to map estimated values of urban trips.

3.3. Validation step

In order to evaluate the accuracy of both models, a validation step was carried out, the results of which are shown in Table 6. For the FKED model, the values of the primary variable were estimated in known values of geographical coordinates through a cross-validation procedure. Afterward, statistical measures were calculated by observed and estimated values of the primary variable. For the LR case, the same goodness-of-fit measures were calculated by observed and estimated values. It can be observed that both procedures have similar error values and Pearson correlation and they could be considered reasonable for trip generation issues.

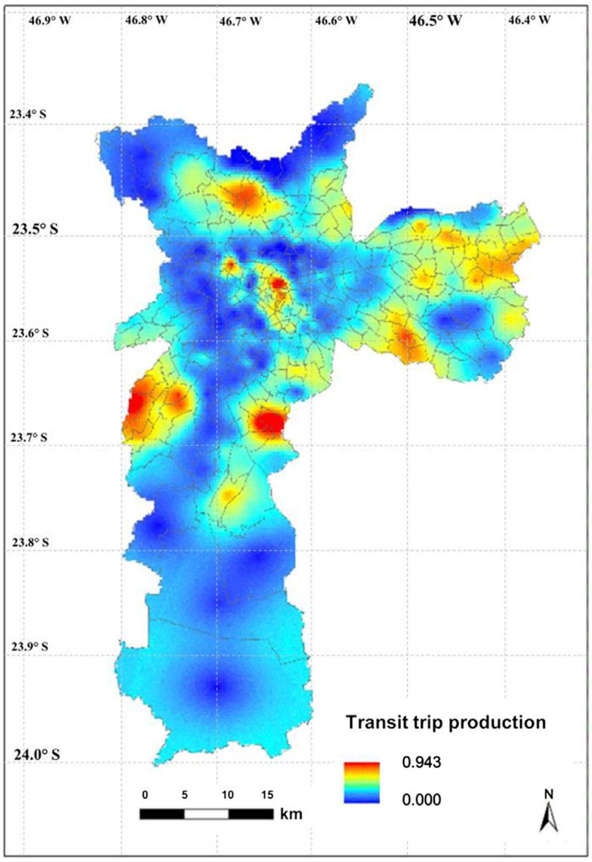


Figure 6. FKED estimation of *transit trip production*.

Table 5. Parameters of the linear model.

Model		Independent variables			
R^2	0.66		Coefficients	*t*	*Sig.*
Sig.	0.000	Constant	0.03	3.035	0.003
F	632.4	*PCA component* 1	0.69	24.970	0.000

Table 6. Validation results: statistical measures for FKED and LR approaches.

Method	MSE	RMSE	SD	SE	MAE	PCC
FKED	0.013	0.113	0.170	0.009	0.080	0.807
LR	0.012	0.109	0.153	0.009	0.075	0.814

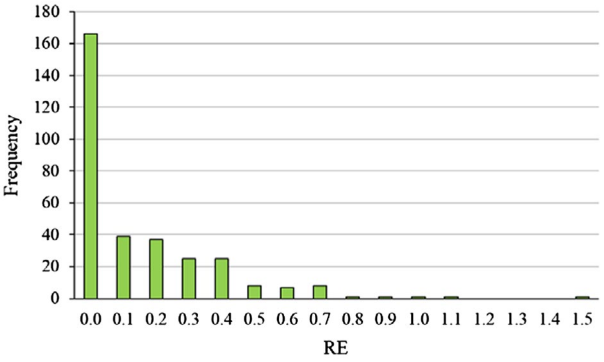


Figure 7. Relative error histogram of the FKED estimation.

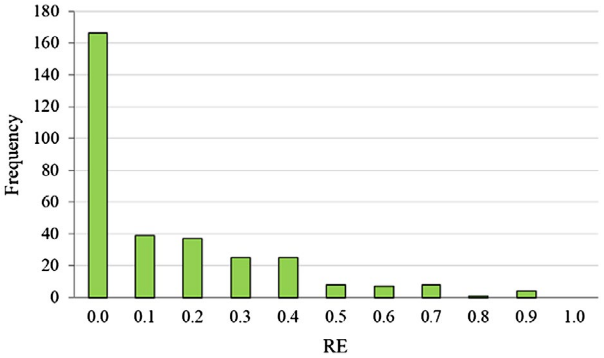


Figure 8. Relative error histogram of the LR estimation.

In addition to the statistical measures presented in Table 6, the relative errors for all observations in the known geographical coordinates were calculated (320 centroids of TAZs). The relative errors of the FKED case are presented in a histogram in Figure 7. It can be observed that there is a high frequency of observations around zero (positively skewed distribution), suggesting that the model has a good predictive capability.

Figure 8 illustrates the histogram of the relative errors when estimating the *transit trip production* through an LR approach using *PCA component* 1 as an explanatory variable. The results showed that both techniques can be used for trip production estimation. Performance measures, such as correlation analysis (r^2 = 0.81) and error analysis, also indicated satisfactory results.

Furthermore, in a subsequent step, if a classical covariate selection (stepwise) was used to estimate the dependent variable through LR – in spite of using the *PCA component* 1, the explanatory variable would be represented as *employment in the service sector*. The correlation would outperform the former approaches (achieving a determinant correlation of 0.9), but it would not result in a high frequency of null errors. According to Figure 9, which shows the relative error histogram, the errors tend to be higher than those seen in the former approaches.

Except for Figures 9, 7 and 8 illustrate the relative error distribution is very similar for both cases, using only *PCA component* 1. The relative error between the predicted and observed values in 320 TAZs presented a higher frequency around zero. Besides this, both approaches presented similar values for the goodness-of-fit measures. However, traditional non-spatial methods do not have the ability to estimate values of the variable in different geographic coordinates, as well as the previously known coordinates (320 TAZs). Hence, this is the main advantage of the proposed multivariate spatial methodology.

4. Main conclusions and methodological limitations

This paper proposed the application of FKED to estimate *transit trip production*. The method was formed by a sequential application of PCA and KED, on an aggregate analysis of TAZs in the city of São Paulo, Brazil.

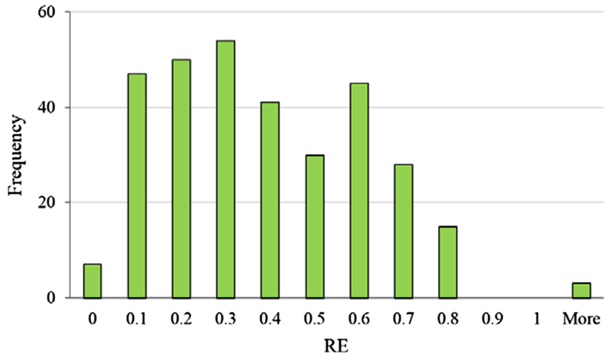


Figure 9. Relative error histogram of the LR model using a stepwise covariate selection.

The proposed methodology and the results of this paper showed that the combined use of PCA and KED can be promising for studies on travel demand forecasting, specifically in the trip production step. The adopted procedure enables the estimation of trip production at various points, as well as the sampled centroids of TAZs. It is important to highlight that the proposed method is not only adequate for future estimations based on explanatory variables, but also provides a continuous map of estimated values of urban trips. These could be considered as the main contributions of this study. The sequential use of PCA and KED (FKED) is also interesting considering that the technique can accomplish the interpolation of trip-related variables regarding secondary information (components), and combines original variables that influence transportation modeling. Table 7 summarizes the drawbacks and benefits of each of the three methods used in this study.

Finally, it is important to mention that, for this study, an initial assumption was taken into account that all geographical units were considered to have the same size and shape. This enabled us to use geographic centroids in semivariogram estimation and Kriging. Another implicit assumption was that the variable values were uniformly distributed within each unit.

Table 7. Advantages and disadvantages of the FKED and LR approaches.

Approach	Advantages	Disadvantages
FKED	Spatial multivariate analysis Future projections Maps estimated values	Secondary variable in FKED derived from PCA Modifiable areal unit problem
LR	Multivariate analysis Future projections	Explanatory variable in LR derived from PCA Does not estimate using spatial associations Does not produce a map of estimated values Modifiable areal unit problem
LR (Stepwise)	Multivariate analysis Future projections Uses original explanatory variables	Does not estimate using spatial associations Does not produce a map of estimated values Modifiable areal unit problem

When performing point Kriging associated to areal data, a practical assumption was made that all habitants of the administrative area live in the same location and the measure refers to this specific location. This assumption is reasonable whenever the aggregation units are small with respect to the spacing of the interpolation grid. However, it is not the case of the research presented in this paper. Therefore, for further analysis and studies, the authors strongly recommend the method proposed by Goovaerts (2006) whereby the size and shape of administrative units, as well as the covariate densities, are incorporated into the filtering of noisy urban trip rates and the creation of isopleth urban trip maps. Furthermore, validation using an independent sample is recommended for future research, such as another region or another year.

Acknowledgments

The authors would like to thank the São Paulo Metropolitan Company for providing the Origin–Destination Survey data from 2007.

Funding

This research was sponsored by the National Counsel of Technological and Scientific Development (CNPq, Brazil), the Coordination for the Improvement of Higher Education Personnel (CAPES, Brazil), and the State of São Paulo Research Foundation (FAPESP-2013/25035-1, Brazil).

Notes on contributors

Anabele Lindner is a civil engineer graduated at Federal University of Paraná and holds Master of Science degree in Transportation Engineering at the University of São Paulo, Brazil. She is currently a PhD student at the University of São Paulo and is involved in projects related to transportation planning, travel demand, Geostatistics, and multivariate and spatial data analysis.

Cira Souza Pitombo is an associate professor at the Transportation Engineering Department in the University of São Paulo, São Carlos. She holds a PhD degree in Transportation Engineering at University of São Paulo. She carried out a postdoctoral project at the University of Lisbon and the University of Leeds. Her areas of expertise include travel demand, modeling traffic accidents, multivariate and spatial data analysis.

Samille Santos Rocha is a geographer graduated at Federal University of Bahia. She is currently a PhD student in Transportation Engineering at the University of São Paulo. Her main experience is in the field of travel demand forecasting using spatial data analysis.

José Alberto Quintanilha is an assistant professor at the Transportation Engineering Department of Polytechnic School of the University of São Paulo. He is a bachelor in Statistics at the São Paulo University. He holds a Master of Science degree in Remote Sensing at the National Institute for Space Research and a PhD degree in Transportation

Engineering at the Polytechnic School of the University of São Paulo. His areas of expertise include GIS, remote sensing, urban and transportation planning.

References

Anselin, L. 1992. "Space and Applied Econometrics: Introduction." *Regional Science and Urban Economics* 22 (3): 307–316. doi:10.1016/0166-0462(92)90031-U.

Batista, A. C., A. J. Sousa, M. J. Batista, and L. Viegas. 2001. "Factorial Kriging with External Drift: A Case Study on the Penedono Region, Portugal." *Applied Geochemistry* 16 (7–8): 921–929. doi:10.1016/S0883-2927(00)00069-X.

Bhat, C., and H. Zhao. 2002. "The Spatial Analysis of Activity Stop Generation." *Transportation Research Part B: Methodological* 36 (6): 557–575. doi:10.1016/S0191-2615(01)00019-4.

Bhat, C. R., and I. N. Sener. 2009. "A Copula-based Closed-form Binary Logit Choice Model for Accommodating Spatial Correlation Across Observational Units." *Journal of Geographical Systems* 11 (3): 243–272. doi: 10.1007/s10109-009-0077-9.

Chang, J. S., D. Jung, J. Kim, and T. Kang. 2014. "Comparative Analysis of Trip Generation Models: Results Using Home-based Work Trips in the Seoul Metropolitan Area." *Transportation Letters* 6 (2): 78–88. doi:10.1179/1942787514Y.0000000011.

Goovaerts, P. 1992. "Factorial Kriging Analysis: A Useful Tool for Exploring the Structure of Multivariate Spatial Information." *Journal of Soil Science* 43 (4): 597–619. doi:10.1111/j.1365-2389.1992.tb00163.x.

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 (1): 52–52. doi:10.1186/1476-072X-5-52.

Goovaerts, P. 2009. "Medical Geography: A Promising Field of Application for Geostatistics." *Mathematical Geosciences* 41 (3): 243–264. doi:10.1007/s11004-008-9211-3.

Hair, J. F., W. C. Black, B. J. Babin, and R. E. Anderson. 2010. *Multivariate Data Analysis: A Global Perspective*. 7th ed. 785. Upper Saddle River, NJ: Pearson Education.

IBGE (Instituto Brasileiro de Geografia e Estatística) 2010. *Demographic Census 2010*. Accessed November 29, 2015. http://www.ibge.gov.br/home/estatistica/populacao/censo2010/sinopse/sinopse_tab_rm_zip.sht

Jolliffe, I. 2002. *Principal Component Analysis*. 2nd ed. 518, Berlin: Springer

Matheron, G. 1963. "Principles of Geostatistics." *Economic Geology* 58 (8): 1246–1266.

Miura, H. 2010. "A Study of Travel Time Prediction Using Universal Kriging." *TOP* 18 (1): 257–270. doi:10.1007/s11750-009-0103-6.

Ortúzar, J. D., L. G. Willumsen. 2011. *Modelling Transport*. 4th ed. 586. Chichester: Wiley.

Páez, A., F. A. López, M. Ruiz, and C. Morency. 2013. "Development of An Indicator to Assess the sPatial Fit of Discrete Choice Models." *Transportation Research Part B Methodological* 56: 217–233. doi:10.1016/j.trb.2013.08.009.

Peer, S., J. Knockaert, P. Koster, Y. Y. Tseng, and E. T. Verhoef. 2013. "Door-to-door Travel Times in RP Departure Time Choice Models: An Approximation Method Using GPS Data." *Transportation Research Part B Methodological* 58: 134–150. doi:10.1016/j.trb.2013.10.006.

Pitombo, C. S., A. S. G. Costa, and A. R. Salgueiro. 2015. "Proposal of A Sequential Method for Spatial Interpolation of Mode Choice." *Boletim de Ciências Geodésicas* 21 (2): 274–289. doi:10.1590/S1982-21702015000200016.

Pitombo, C. S., A. R. Salgueiro, A. S. G. da Costa, and C. A. Isler. 2015. "A Two-step Method for Mode Choice Estimation with Socioeconomic and Spatial Information." *Spatial Statistics* 11: 45–64. doi:10.1016/j.spasta.2014.12.002.

Pitombo, C. S., A. J. de Sousa, J. A. Quintanilha, and M. Birkin. 2010. "Comparing Different Spatial Data Analysis to Forecast Trip Generation." Proceedings of the 12th World Conference on Transport Research Society (WCTR), Lisboa.

Sanguansat, P. 2012. *Principal Component Analysis – Multidisciplinary Applications*, 212, Rijeka, Croatia: Intech.

Schmöcker, J. D., A. Q. Mohammed, R. B. Noland, and M. Bell. 2005. "Estimating Trip Generation of Elderly and Disabled People: Analysis of London Data." *Transportation Research Record: Journal of the Transportation Research Board* 1924: 9–18.

Tobler, W. 1979. "Cellular Geography." In *Philosophy in Geography*, edited by S. Gale and G. Olsson, 379–386, Dordrecht, The Netherlands: Reidel Publishing Company.

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

Zou, H. X., Y. Yue, Q. Q. Li, and A. G. O. Yeh. 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." *International Journal of Geographical Information Science* 26 (4): 667–689. doi:10.1080/13658816.2011.609488.