DOI 10.1590/1982-2170-2020-0069

ARTICLE

APPLYING MULTIVARIATE GEOSTATISTICS FOR TRANSIT RIDERSHIP MODELING AT THE BUS STOP LEVEL

Geoestatística Multivariada Aplicada à Modelagem da Demanda por Transporte Público no Âmbito de Pontos de Parada

Samuel de França Marques[1] - ORCID: 0000-0001-5602-3277

Cira Souza Pitombo[2] - ORCID: 0000-0001-9864-3175

[1] Universidade de São Paulo, Escola de Engenharia de São Carlos, Departamento de Engenharia de Transportes, São Carlos - SP, Brasil.
E-mail: samuelmarques@usp.br

[2] Universidade de São Paulo, Escola de Engenharia de São Carlos, Departamento de Engenharia de Transportes, São Carlos - SP, Brasil.
E-mail: cirapitombo@gmail.com

Received on: 24-Feb-2020

Accepted on: 03-Mar-2021

Abstract:

Travel demand models have been developed and refined over the years to consider a characteristic normally found in travel data: spatial autocorrelation. Another important feature of travel demand data is its multivariate nature. However, regarding the public transportation demand, there is a lack of multivariate spatial models that consider the scarce nature of travel data, which generally are expensive to collect, and also need an appropriate level of detail. Thus, the main aim of this study was to estimate the Boarding variable along a bus line from the city of São Paulo - Brazil, by means of a multivariate geostatistical modeling at the bus stop level. As specific objectives, a comparative analysis conducted by applying Universal Kriging, Ordinary Kriging and Ordinary Least Squares Regression for the same travel demand variable was proposed. From goodness-of-fit measures, the results indicated that Geostatistics is a competitive tool comparing to classical modeling, emphasizing the multivariate interpolator Universal Kriging. Therefore, three main contributions can be highlighted: (1) the methodological advance of using a multivariate geostatistical approach, at the bus stop level, on public transportation demand modeling; (2) the benefits provided by the models regarding the land use and bus network planning; and (3) resource savings of field surveys for collecting travel data.

Keywords: Transit Ridership; Boarding per Bus Stop; Universal Kriging; Ordinary Kriging; Linear Regression; Spatial Statistics

1. Introduction and Background

Increasing concern about the environment and a discussion about sustainability have strongly influenced public policies around the world. In Brazil, law 12,587/2012, known as the Urban Mobility Law, points out that non-motorized and public transportation modes should be

prioritized over motorized and individual ones, respectively. This determination recognizes Public Transportation (PT) as a promoter of sustainable development and social inclusion. However, in order to allow the supply and demand balance of this service, support of appropriate planning is needed to guarantee the properly work of the transportation system.

Among the most traditional models that provide support to travel demand predictions are those that use classical linear regression (George and Kattor 2013; Pendyala, Shankar and McCullough 2000; Varagouli, Simos and Xeidakis 2005). This technique, however, overlooks an important characteristic normally found in travel demand variables: spatial autocorrelation, i.e., the fact that trip data located near each other in space present similar values. Since the traditional linear model assumes independence between sample data (Yan and Su 2009), the outcomes of using it cannot be totally reliable when it refers to travel demand variables as such variables are, generally, spatially dependent.

Thus, linear regression adaptations, seeking to include spatial autocorrelation, as well as new improved techniques, were developed in order to overcome classical model constraints regarding treating Regionalized Variables (RV). Attempts to include spatial dependence of travel demand observations have been made by Gutiérrez et al. (2011) and Pulugurtha and Agurla (2012) from decay functions. This approach represents an advance in the RV modeling, as it basically consists of assigning weights to predictor data according to the distance between the database points and their influence areas (also known as service or catchment areas). Nevertheless, as such models include space only as an attribute, and in a deterministic way, these approaches cannot yet be considered as completely spatial (Fotheringham et al. 2003).

This limitation is overcome by the spatial regression models, which have already been used for travel demand forecasting (Gan et al. 2019; Lopes, Brondino and Rodrigues da Silva 2014; Sarlas and Axhausen 2016; Wang 2001). These models can consider the spatial autocorrelation by means of an explanatory variable, obtained from a spatially lagged dependent variable, or by the residual term of the model, and both of them include a spatial weight matrix normally based on the distance between the points of the database (Fotheringham et al. 2003).

Moreover, when dealing with scarce data, spatial regression models include a new interpolation approach (Krige 1951; Matheron 1963; 1971) that treats Regionalized Variables as random and no longer deterministic functions, allowing the application of statistical inference on the estimates provided by these new techniques. In its application, this science field, known as Geostatistics, presents the advantage of not requiring, necessarily, information about ancillary variables, and the fact that its interpolators generate unbiased and minimum variance estimates. In addition, Geostatistics can use the maximum amount of information available about the variable of interest to estimate its value in non-sampled points, also eliminating the negative effect of using clustered samples (Matheron 1971).

Unlike the traditional spatial regression models, in which spatial interaction is usually captured by a weight matrix based on the distance between points, Geostatistics uses the semivariogram function. This tool, which comes from a probabilistic approach of Regionalized Variables, enables us to model the spatial dependence of the data, and the results of this modeling provide a complete understanding of the spatial structure of the variable of interest, both in visual and numerical ways.

Geostatistics covers different types of estimators. In this paper, we mention three of them: Simple Kriging (SK), Ordinary Kriging (OK) and Universal Kriging (UK). The search for the interpolator that demonstrates the best performance, in goodness-of-fit measures, has led to several studies in

which Simple Kriging results are compared to those of Ordinary Kriging (Daya and Bejari 2015; Taharin and Roslee 2017; Viswanathan et al. 2015), in which Ordinary Kriging is compared to Universal Kriging (Hiemstra et al. 2010; Kiš 2016; Liu et al. 2015; Mubarak et al. 2015; Nalder and Wein 1998; Wang and Zhu 2016), and in which the three techniques are simultaneously compared (Asa et al. 2012; Seo et al. 2015). In short, since UK includes explanatory variables in its formulation, it normally outperforms the other interpolators, especially when there is some large-scale trend present in the interest variable structure. Afterwards, OK, which assumes that the interest variable mean is unknown and varies locally, demonstrates the best results compared to SK, whose mean is global, constant and known.

In spite of several comparative studies already developed, the conclusion reported in these studies is not consensual. In the aforementioned articles, the interpolators´ performance varied substantially according to the type of data under analysis. Regarding the travel demand, not many studies were observed that compare the performance of geostatistical interpolators. In the case study proposed by Shamo, Asa and Membah (2015), the interest variable (Annual Average Daily Traffic) refers only to rural highway segments, which does not offer, *a priori*, a contribution to the urban public transportation planning. Besides this, the authors themselves reinforced the idea that the best kriging technique and semivariogram can only be obtained from the structure present in the available information about the interest variable.

Regarding urban bus transportation planning, which is highly important to the supply and demand balance of the PT system, passenger flow along the bus lines is a valuable information and, often, hard to acquire. Marques and Pitombo (2021), Marques and Pitombo (2019) and Marques (2019) proved that Geostatistics, more specifically Ordinary Kriging, demonstrates an excellent potential in estimating the three variables, collected from a Boarding and Alighting counts survey, that express the passenger demand along a bus route. They are: Boardings and Alightings (number of passengers entering and leaving the bus line at each bus stop, respectively) and Loading (passenger volume inside the bus at each line segment contained between two consecutive bus stops). Since this survey demands high resources, the results found by those authors suggest that it is possible to perform the Boarding and Alighting counts only in some bus line segments and, by kriging, estimate, with relative accuracy, the demand variable for non-sampled bus stops and segments. This study, however, did not make any comparison between OK and other geostatistical interpolators to verify which one of them could best fit the passenger volume estimate along a public transport line.

It is worthwhile mentioning that the spatial modeling of public transportation passengers at the bus stop level and train, metro or bus station is the most detailed treatment that can be applied to PT network planning. Due to this, this approach is the most recent among the techniques that seek to program supply and understand transportation and land use relationships. In the scientific literature, several studies of this kind can be found, most at the station level (Blainey and Mulley 2013; Blainey and Preston 2010; Cardozo et al. 2012; Chakour and Eluru 2013; 2016; Chiou, Jou and Yang 2015; Choi et al. 2012; Chow et al. 2006; Gutiérrez et al. 2011; Sun et al. 2016) and a few at the bus stop level (Chu 2004; Dill et al. 2013; Kerkman, Martens and Meurs 2015; Pulugurtha and Agurla 2012; Ryan and Frank 2009). However, due to the difficulty in acquiring the variables to be modeled (Boardings and Alightings), in the case of the bus stop level, to the best of the authors´ knowledge, these studies have still not provided a spatial approach of ridership until the present moment. Even in the station level cases, the studies retrieved basically focus on applying Geographically Weighted Regression and generalized linear models to ridership data. Only the station level study of Zhang and Wang (2014), which applies Universal Kriging to the Boarding

variable, was found so far, meaning that approaches based on multivariate Geostatistics at the bus stop level were not yet observed.

Thus, the aim of this study is to estimate a public transportation demand variable, along a bus line, by means of a multivariate geostatistical modeling at the bus stop level. As specific objectives, a comparative analysis conducted by applying Universal Kriging, Ordinary Kriging and Ordinary Least Squares Regression for the same variable under analysis is proposed.

Finally, the following main research gaps associated to this study can be enumerated: (1) Multivariate modeling of public transportation demand at the bus stop level by means of a geostatistical approach; (2) Lack of spatial approaches of transit ridership at the bus stop level; (3) The need for assessing the improvement, in goodness-of-fit measures, caused by the inclusion of explanatory variables to the geostatistical modeling; and (4) Passenger volume modeling at the bus stop level as they are the most appropriate elements for performing this analysis.

This article contains 5 sections, including this introduction. The next section summarizes the few studies that perform ridership modeling at the bus stop level. Section 3 introduces the materials used in the case study and the method applied to them. Then, the results, as well as discussions about them, are presented in Section 4. Lastly, Section 5 draws the conclusions and also proposes suggestions for future research.

2. Ridership models at the bus stop level

While the traditional transportation planning (Ortúzar and Willumsen 2011) is done by means of Traffic Analysis Zones and continues as the most popular method for mobility diagnosis and solution proposal, Cervero (2006) argues that ridership modeling at the local level can provide demand estimates quickly and economically. Moreover, in spite of a regional approach, which uses averaged values of data for each Traffic Analysis Zone, boarding and alighting modeling per bus stop, train, metro or bus station can capture the effect of transit-oriented development on public transport demand, i.e., the influence of built environment variables on transit usage.

From smart card data, boarding and alighting per train or metro station are readily available. On the other hand, bus ridership at the stop level is not easy to collect. Concerning this, cities often depend on expensive surveys, such as boarding and alighting surveys, or automatic counters, which are not widely popularized yet. It may be possible to obtain boarding and alighting per bus stop from smart card data and GPS information, but some assumptions have to be made that affect the accuracy of the results, especially in the case of Alighting. Therefore, boarding and alighting surveys remain the only way to collect ridership at the bus stop level accurately. Table 1 shows studies that perform ridership modeling at the bus stop level.

Table 1: Ridership models at the bus stop level

Reference	Dependent variable	Model	Independent variables	
			Supply	Demand
Chu (2004)	Boarding	Poisson	Transit level of service within 1 to 2-5 min of walking	Income, No-vehicle households, Female (%), Hispanic (%), White (%), Age, No. of inhabitants, No. of jobs, Pedestrian factor

Reference	Dependent variable	Model	Independent variables	
			Supply	Demand
Ryan and Frank (2009)	Boarding + alighting (logarithm)	OLS (log-linear)	Level of service (no. of routes/average waiting time)	Income, No-vehicle households, Female (%), Hispanic (%), White (%), Youth (%), Walkability index
Pulugurtha and Agurla (2012)	Boarding	Negative binomial with log-link	On-network characteristics	Household income, No-vehicle households, Asian population, Residential area
Dill et al. (2013)	Boarding + alighting (logarithm)	OLS (log-linear)	Transit service variables, Transportation infrastructure variables	Households below poverty (%), No-vehicle households (%), White (%), Youth (%), elderly (%), Education level, Job accessibility, Employment (no.), Population (no.), Land use area (single-family, multifamily, commercial), Area parks, Pedestrian destinations, Land use mix index, Distance to city center
Kerkman, Martens and Meurs (2015)	Boarding + alighting (logarithm)	OLS (log-linear)	Stop frequency (logarithm), Directions, Frequency per direction, Direct connections, Competitive bus stops, Bus terminus, Transfer stop, Bus station, Dynamic information, Benches, Supply-demand index	Potential travelers (logarithm), Income, Elderly (%), Distance to urban center (km), Land use: residential, Land use: agriculture, Land use: sociocultural facilities, Supply-demand index

Source: adapted from Kerkman, Martens and Meurs (2015)

From Table 1, it can be seen that the models used are limited to ordinary least squares regressions with logarithmic transformation to correct the asymmetry of the interest variable. Models for count data were also applied, but none of them present a spatial approach of bus ridership. Pulugurtha and Agurla (2012) tried to include spatial dependence of boarding through a weighting function, but only in a deterministic way.

Moreover, explanatory variables used in the boarding and alighting modeling can be divided into two groups: demand and supply variables. Demand independent variables intend to capture the effect of sociodemographic and land use features around bus stops on ridership. On the other hand, infrastructure and public transport service characteristics are addressed by the supply independent variables. In order to minimize the amount of information needed for the spatial modeling, the present study proposed a simple method for selecting the best predictors, as described in Section 3.

3. Materials and Method

The dataset used in this case study refers to the Boarding per bus stop data (number of passengers entering the bus line at each bus stop) over line 856R-10 from the city of São Paulo – Brazil. The results, from a Boarding and Alighting count survey performed along this line on a typical day

(Tuesday) in 2017, as well as the geographic coordinates of its 57 bus stops, were provided by *São Paulo Transporte S.A. (SPTrans)*. Boarding and Alighting per bus stop were available for six times bands: 1st (04h to 04h59), 2nd (05h to 08h59), 3rd (09h to 15h59), 4th (16h to 19h59), 5th (20h to 23h59) and 6th (00h to 03h59). This information was then spatialized in the ArcGIS 10.2 software using the SIRGAS 2000 UTM 23S projection system.

In order to compose the group of explanatory variables to be included in Universal Kriging and Ordinary Least Squares Regression, both features from bus stops themselves and from their influence area were collected. From a catchment area of radius 400m centered in the bus stops (Zhao et al. 2003), the following variables were calculated: population (inhabitants) and population density (inhabitants per hectare), based on the 2017 Origin and Destination Survey (Metrô 2019) shapefile, which is given in Traffic Analysis Zones; and averaged values of household income and car ownership, female (%), population with no complete higher education (%), households with no private vehicles (%), percent of people aged up to 14, up to 17, aged between 18 and 22, 18 and 29, 18 and 39 and above 60 years old. These data were obtained from the sampled households of the 2017 O/D Survey that were within the catchment area; area, in hectares, of the 16 predominant land use classes according to the shapefile of predominant land use in 2016 (GeoSampa), which is disaggregated at the block level; and number of roads and intersections, length (meters) and road density (meters per hectare) inside each catchment area, based on the São Paulo road system (Open Street Map) shapefile. The number of points of interest (POI), also given by OSM shapefile, inside each influence area, was also considered. Overlapping catchment areas were prevented by using Thiessen polygons, similar to the method adopted by Zhang and Wang (2014) and Sun et al. (2016), in a GIS environment.

Besides the road system variables collected from Open Street Map, other indicators were adopted as a proxy of accessibility as well. Together with the Boarding/Alighting count survey results, SPTrans also made the General Transit Feed Specification (GTFS) data, from the São Paulo PT network, available. Knowing the code of the 57 bus stops covered by line 856R-10, the following was calculated from GTFS data: the number of bus lines that passed by each of these stops, and the average frequency of those lines; Euclidean and network distance between each bus stop and the nearest bus terminal, nearest metro station and nearest train station. Two intermodal proximity measures considering the shortest Euclidean and network distance between each bus stop and the nearest metro or train station were also included. While Euclidean distance is based on a straight line, network distance is calculated along the road system. These distance measures were obtained from the 57 bus stop shapefiles along with the São Paulo bus terminals, metro stations and train stations shapefiles, and Open Street Map road system. Versions of the populational, road system and accessibility variables, transformed by the natural logarithm, were also considered, and, in the cases where the raw data contained zeros, it was added to 1 before applying the transformation (Bartlett 1947). In order to include only the attributes encompassed by the bus stops´ influence area, the attributes of the original shapefiles went through an aerial interpolation. As stated in Table 1, the data collected for the modeling procedure covers both supply and demand independent variables.

Afterwards, dependent and independent variables were selected using a joint analysis of linear correlation and spatial autocorrelation. In order to choose the variable of interest, the Moran index (Moran 1948) was calculated for the Boarding and Alighting data in the six time bands mentioned above. After that, the degree of association between the cases with the highest and statistically significant values of Moran's index and all explanatory variables was tested by the Pearson linear correlation coefficient (R). In order to eliminate multicollinearity, at this stage, the

R value between two potential predictors was limited to 0.60. Therefore, when a pair of independent variables had a high correlation with the variable of interest, but R with each other above 0.60, the variable with the least correlation with the dependent variable was discarded. This threshold was considered acceptable to avoid the omitted variable bias as well, since a pair of highly correlated variables does not always represent a cause-effect relationship. Other criteria for choosing dependent and independent variables were: expected correlation signal and presence of independent variables from both supply and demand groups. Thus, the number of Boardings, transformed by the natural logarithm in the 5th time band, also known as Night Peak (NP, from 20h to 23h59), was chosen as the dependent variable. As potential predictors, the following variables were kept: population, number of POIs, number of road intersections, road length, number of other bus lines, mean household income and average frequency of other bus lines in the same time band as Boardings, all transformed by natural logarithm; also, population with no complete higher education (%), residential, commercial and services area (ha), and network distance, in meters, between each bus stop and the nearest metro station, were considered.

The modeling step started by initially calibrating a linear regression model. To select the best predictors among those considered, a stepwise method was applied, in which only three independent variables remained. Regarding the modeling area, in general, there is a trade-off between the prediction power of the technique and the number of explanatory variables used in the model, whose data source might be hard to access. The desirable scenario is to have a minimum number of explanatory variables (that are preferably easy to acquire) associated to a satisfactory performance of the model. Based on this, the following procedure was adopted: initially, a simple linear regression model was calibrated with each one of the three explanatory variables, separately; then, three linear regressions were estimated using two predictor combinations; afterwards, a third model considering the three variables as predictors was generated. This approach was repeated in the geostatistical modeling by means of UK as this estimator also includes explanatory variables in its formulation. The purpose of this analysis was to verify whether the models with the least explanatory variable are also competitive in terms of minimizing errors between real and estimated values, and how much the spatial approach improves bus ridership estimates compared to traditional linear regression.

All linear regression models were calibrated using the Ordinary Least Squares method (Yan and Su 2009). Considering only the cases in which all predictors were statistically significant in linear regression ($p < 0.10$), the geostatistical modeling steps were performed. They are: (1) Empirical semivariogram calculation and model fitting; (2) Cross validation; and (3) Estimation by OK and UK.

The semivariogram $\gamma(h)$, or variogram $2\gamma(h)$, is the main graphical tool of Geostatistics as it visualizes the spatial structure of the variable under analysis. The calculation of the empirical, or experimental, semivariogram is given by Equation (1) (Cressie 1993; Matheron 1971).

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

$Z(x)$: value of the Regionalized Variable *Z* in the sampled geographical position x;
N: number of pairs situated at distance h.

Equation 1 refers to Ordinary Kriging, in which the semivariogram is calculated straight based on the RV information. Concerning UK, this calculation is applied to the residual term, in which a spatial structure is assumed. Then, a theoretical model is adjusted to the empirical semivariogram values. The process of fitting a well-defined function to the empirical semivariogram points consists of obtaining three main parameters, the nugget effect, partial sill and range, from a pre-established method (Cressie 1993). In the present case study, geostatistical modeling was performed by means of the three main theoretical semivariogram models: Exponential (Exp), Gaussian (Gau) and Spherical (Sph) (Olea 2006), in order to verify if one of them demonstrates a much better adjustment compared to the others.

The process of kriging a Regionalized Variable basically consists of obtaining the optimum weights for the linear combination of weights and neighboring values that results in a continuous surface of estimated points, which also covers the non-sampled locations. The kriging estimator is given by Equation (2) (Cressie 1993; Matheron 1971).

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

$Z^*(x_0)$: estimated value of Regionalized Variable at the geographic position x_0;
λ_i: optimum weight assigned by kriging to the neighbor i value.

Although both OK and UK are linear combinations, the first one assumes a constant and local, but unknown mean (μ) of the dependent variable observations (Equation (3)), while the latter relaxes this assumption by considering the presence of a large-scale trend over the response variable structure (Equation (4)).

$$Z = \mu + \varepsilon \tag{3}$$

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \tag{4}$$

in which ε is the error term of the model, x_k represents the explanatory variables, and β_{k+1} expresses the linear function parameters to be calibrated. Thus, Universal Kriging assumes that the Regionalized Variable values are affected not only by their neighbors (small range variation), but also that there is a systematic component in their structure, caused by the influence of the built environment around the treatment elements, which are, in this case, the bus stops. Besides this, UK allows this large-scale variation to be modeled through the inclusion of explanatory variables to the kriging estimator. Thus, instead of considering the errors completely as white noise, it is assumed that the RV spatial structure is present in the residual term oscillation, where the semivariogram function is calculated (Cressie 1993).

Ordinary Kriging weights λ_i are obtained from a matrix operation, represented in Equation (5). The resulting nonlinear equations system takes into account three constraints: the (1) non bias, (2) minimum variance, and (3) weight sum equal to 1, in order to guarantee the best linear unbiased estimator (Cressie 1993; Goovaerts 1997; Matheron 1971).

$$\begin{bmatrix} \gamma(h_{1-1}) & \gamma(h_{1-2}) & \ldots & \gamma(h_{1-n}) & 1 \\ \gamma(h_{2-1}) & \gamma(h_{2-2}) & \ldots & \gamma(h_{2-n}) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(h_{n-1}) & \gamma(h_{n-2}) & \cdots & \gamma(h_{n-n}) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(h_{0-1}) \\ \gamma(h_{0-2}) \\ \vdots \\ \gamma(h_{0-n}) \\ 1 \end{bmatrix} \tag{5}$$

The matrix on the left corresponds to the theoretical semivariance between sample points $[K]$; vector $[\lambda]$ in the middle contains the kriging weights; and the vector on the right expresses the theoretical semivariance between the sample points and the point to be estimated $[M]$. Therefore, OK weights are calculated according to Equation (6) for each point to be estimated.

$$[\lambda] = [K]^{-1}[M] \tag{6}$$

On the other hand, Universal Kriging formulation deals with parameters in linear function, which is similar to classical regression, and residual semivariogram. Therefore, its calibration process is complex and must be performed in an iterative way. First, the linear model is calibrated and, after the residual term is calculated, the nugget effect, partial sill and range are obtained. Other values for these parameters, nearby the original ones, are tested until there is some convergence to an optimum error between the observed and estimated value criteria (Cressie 1993; Selby and Kockelman 2013; Zhang and Wang 2014). In short, UK estimates are given by Equation (7).

$$Z^*(x_0) = [X_o][\beta] + \left[V_{s_0}^T\right][V_s^{-1}][\varepsilon] \tag{7}$$

Where X_0 is the matrix of explanatory variable observations of point x_0, β is the vector of linear parameter estimates, V_{s_0} represents the vector of estimated covariances between sample points and point x_0, while V_s expresses the matrix of estimated covariances between sample points. It is worth remembering that covariance (V) and semivariogram (γ) functions are related according to Equation (8).

$$V(h) = c_0 + c_1 - \gamma(h) \tag{8}$$

Where c_0 and c_1 stand out, respectively, for the nugget effect and partial sill parameters from the theoretical semivariogram.

Concerning geostatistical estimates, cross validation is performed by the leave-one-out method (Cressie 1993). This technique consists of removing the database points one by one and calculating their value from the remaining points and theoretical semivariogram parameters (and also the linear function, when it refers to UK). Therefore, from the observed value at the points and respective estimated value, several goodness-of-fit measures can be established to assess the performance of the applied spatial statistics tool. Regarding the linear regression, the estimate considered in this study was the number of Boarding predicted by the model equation. Thus, some of the goodness-of-fit measures suggested by Hollander and Liu (2008) were calculated, which are: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Pearson linear correlation coefficient between the observed and predicted values (R).

The cited goodness-of-fit measures were applied to the results of each estimate and, hence, it was possible to assess and compare the accuracy of results found from such techniques, and to select those that demonstrated the best performance. In the UK cases, results from the semivariogram that provided the smallest errors were selected to compare them with the respective linear regression estimates. The computational resources that gave support to the method stages were: ArcGIS 10.1, QGIS 3.0.3 and GRASS GIS 7.4.0 (Bundala, Bergenheim and Metz 2014) to collect the potential predictors; GeoDa (Anselin 2004; Anselin, Syabri and Kho 2005) for Moran's index

calculation; IBM SPSS 24.0 (IBM 2016) for correlation analysis; and R (R Core Team 2020; Ribeiro Jr and Diggle 2016; Papritz 2020a; Papritz 2020b) for linear regression, Ordinary Kriging and Universal Kriging.

4. Results and Discussion

Figure 1 shows thematic maps for the dependent variable, Boardings, i.e., the number of passengers entering each bus stop on a typical day (Tuesday) in 2017, in the aggregated set of bus trips made from 20:00 to 23:59; and for the three explanatory variables selected by the stepwise method. They are: (1) natural logarithm of population (*lnpop*); (2) residential, commercial and services area, in hectares (*res_com_serv area*); and (3) network distance, in meters, between each bus stop and the nearest metro station (*metrodist_net*). As *lnpop* and *res_com_serv area* belong to the demand variable group, and *metrodist_net* to the supply one, this result was deemed satisfactory.

As expected, bus stops located at regions with more inhabitants tend to have a higher number of Boardings. This pattern can also be noted in the case of residential, commercial and service area, meaning that the higher the land use mixture, the higher Boardings will be. Pearson's correlation coefficient between *ln_boarding* and *ln_pop* and between *ln_boarding* and *res_com_serv area* was, respectively, 0.68 and 0.45. On the other hand, despite some bus stops located near metro stations are showing less passenger flow, there are many points nearer metro stations that do present a high number of Boardings. This relationship resulted in a R value of -0.26 between *ln_boarding* and *metrostation_net*. Thus, it can be stated that most 856R-10 line users, in the period from 20:00 to 23:59, come from metro lines, probably returning from work to home.

Figure 1 also reveals that the number of Boardings per bus stop in line 856R-10 shows, in general, five volume peaks: the first one is next to the beginning of the route, the second and third are halfway, and the last two are near the end of the line. Such peaks interlay with lower passenger flow points, starting at the first bus stops of the line, which present a reduced number of Boardings. This pattern resulted in a Moran's index of about 0.26, which increased to 0.48 with the logarithmic transformation. In both cases, the index value was statistically significant (pseudo p-value < 0.05), proving the presence of spatial dependence in Boardings per bus stop data.

Descriptive statistics of dependent and independent variables are presented in Table 2. Travel demand variables, in general, are given as count data and show asymmetry very often. Thus, their relationship with explanatory variables may not be linear. In this case, logarithmic transformations contribute to linearizing the model equation, addressing the real nature of the data and, hence, improving results.

As shown in Table 2, mean and median measures for *ln_boardings* and *ln_pop* are similar, given their normality. Standard deviation for all variables, as well as minimum and maximum values, reveal the presence of a wide range of values, meaning the inclusion of more diversified data in the modeling, thus making it possible to use the models to estimate ridership for various conditions. Moreover, it is important to mention that Boardings and *res_com_serv_area* were zero for three and five bus stops, respectively. In the case of Boardings, some points at the end of the route did not have any passengers entering the bus line in the period from 20:00 to 23:59, probably because at this time most users are returning home from work and, hence, at the end of the line, most passengers are leaving the vehicle rather than entering it.

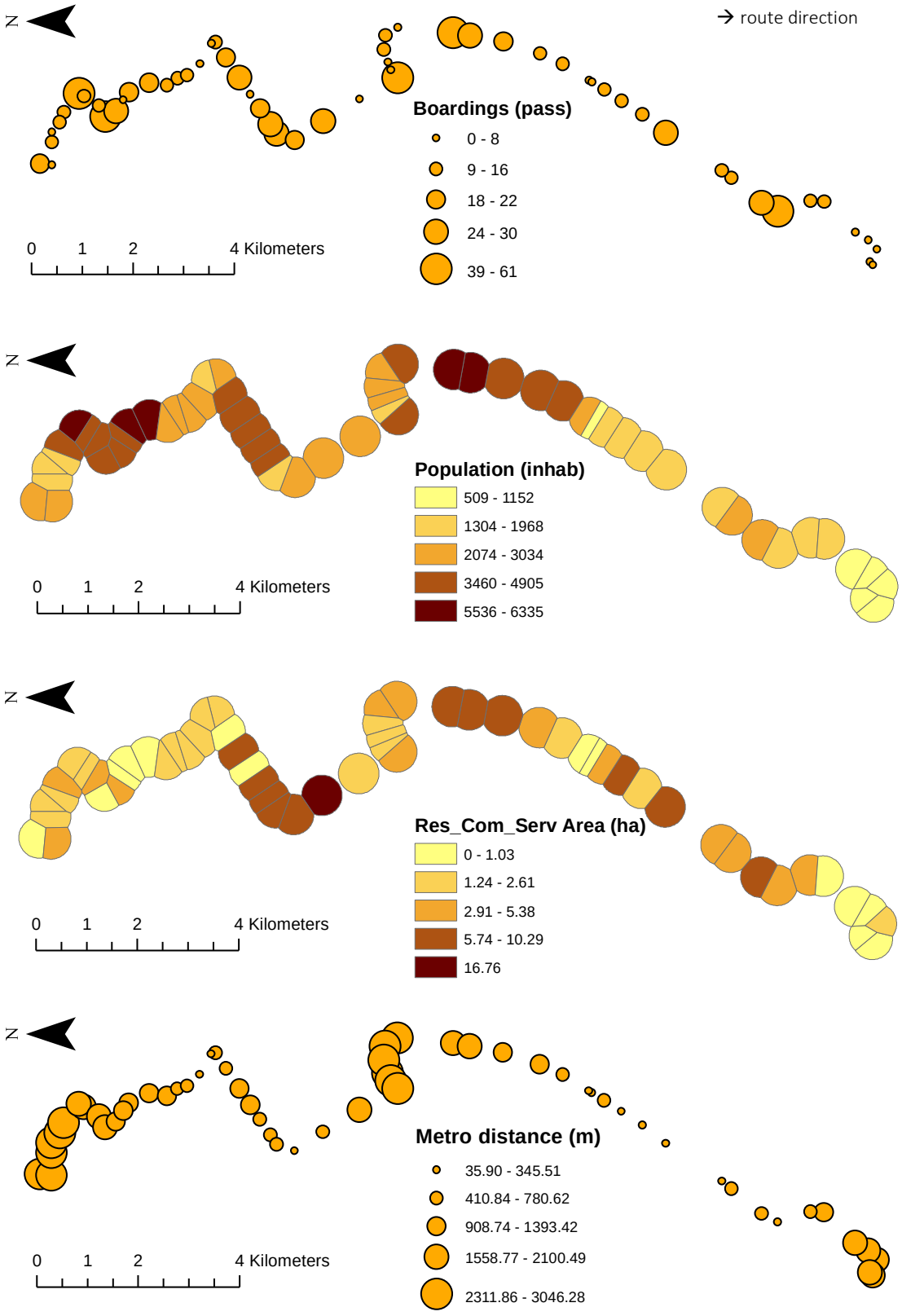

Figure 1: Patterns of (from top to bottom) Boardings; Population; Residential, commercial and services area; and distance to the nearest metro station along the bus line 856R-10

Table 2: Descriptive statistics

	ln_boarding	ln_pop	res_com_serv area (ha)	metrostation_net (m)
N	57	57	57	57
Mean	2.51	7.81	3.23	1319.60
Std. Deviation	0.95	0.61	3.16	909.20
Minimum	0.00	6.23	0.00	35.90
25%	2.20	7.43	1.14	490.64
50%	2.64	7.76	2.21	1136.44
75%	3.11	8.34	4.64	2089.06
Maximum	4.13	8.75	16.76	3046.28

4.1 Univariate step: Ordinary Kriging

Results of Ordinary Kriging are displayed in Table 3. In spite of the low percentages of nugget effect relative to the sill (nugget effect plus partial sill), goodness-of-fit measures are not quite satisfactory. Comparing the three theoretical models, the exponential one provided the best estimates. Experimental semivariogram for *ln_boarding* and the fitted exponential model are shown in Figure 2.

Table 3: Ordinary Kriging results

Measure\Model	Gaussian	Exponential	Spherical
Nugget effect	37.26%	25.24%	35.09%
Partial sill	0.933	1.155	0.813
Range (m)	10000	10000	15000
MAE	9.138	8.308	8.413
RMSE	13.551	12.684	12.870
MAPE	117.25%	96.27%	100.13%
R	0.057	0.296*	0.256*

Note: * statistically significant at the 0.05 level (one-tailed). MAE, RMSE, MAPE and R are, respectively, Mean Absolute Error, Root Mean Square Error, Mean Absolute Percentage Error and Pearson Linear Correlation Coefficient between predicted and observed values.

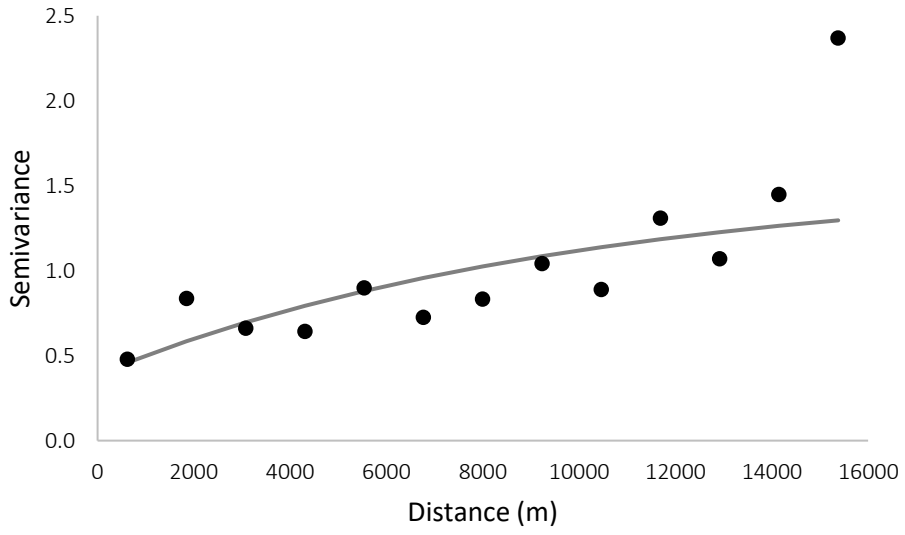

Figure 2: Semivariogram of Boardings with logarithmic transformation

The exponential model had a good fit to the experimental *ln_boardings* semivariogram. On the other hand, the experimental semivariogram seems to increase without bound as the lag distance increases, which could indicate the presence of a large-scale trend in the interest variable that is not being modeled (Oliver and Webster 2015). This might be the reason why Ordinary Kriging estimates are almost twice the observed values, given a Mean Absolute Percentage Error of 96%. It is worth remembering that, although geostatistical and traditional modeling were performed based on the values of Boardings with logarithmic transformation, goodness-of-fit measures were calculated using the estimates with inverse transformation, so they could be directly compared to the real values.

4.2 Multivariate step: Universal Kriging and linear regression

According to the method, 25 different estimates were obtained. They are: Ordinary Kriging with exponential (1), Gaussian (2) and spherical (3) semivariograms, which have already been showed in subsection 4.1; simple linear regression with *ln_pop* (4), *res_com_serv area* (5), and *metrodist_net* (6) as the predictor; multiple linear regression with *ln_pop* and *res_com_serv area* (7); with *ln_pop* and *metrodist_net* (8); and with *res_com_serv area* and *metrodist_net* (9); then with *ln_pop*, *res_com_serv area* and *metrodist_net* (10); UK with *ln_pop* and the three semivariograms (11-13); UK with *res_com_serv area* and the three semivariograms (14-16); UK with *ln_pop* and *res_com_serv area*, and the three semivariograms (17-19); UK with *ln_pop* and *metrodist_net*, and the three semivariograms (20-22); and finally UK with *ln_pop*, *res_com_serv area* and *metrodist_net* as predictors, and the three semivariograms (23-25). The *metrodistance_net* variable was not statistically significant in the simple linear regression (6) neither when coupled with the *res_com_serv area* (9). Thus, these combinations were not repeated in the geostatistical modeling and will not be presented here, for brevity.

Table 4 shows the resulting parameters from Universal Kriging and linear regression. As for Ordinary Kriging, the best semivariogram model, i.e., the theoretical semivariogram that yielded the best goodness-of-fit measures, in all predictor combination cases, was the exponential one. Therefore, for the sake of brevity, Universal Kriging results shown in Table 4 correspond only to those from the exponential model.

Table 4: Results from spatial interpolators and classical linear regression

Model\Parameters	Intercept	ln_pop	res_com_serv area (ha)	metrodist_net (m)	Nugget effect	Partial sill	Range (m)
Universal Kriging	-6.0460***	1.1040***			47.54%	0.3090	1229.0990
Linear regression	-5.8260***	1.0670***					
Universal Kriging	2.1490***		0.1025*		46.69%	0.4350	1365.1720
Linear regression	2.0762***		0.1352***				
Universal Kriging	-5.7216***	1.0207***	0.0864**		68.48%	0.1510	2238.5980
Linear regression	-5.3115***	0.9615***	0.0965**				
Universal Kriging	-5.5912***	1.1012***		-0.0003*	54.75%	0.2380	1288.6310
Linear regression	-5.4770***	1.0700***		-0.0003**			
Universal Kriging	-5.3772***	1.0234***	0.0715*	-0.0002(.)	69.26%	0.1420	2058.5110
Linear regression	-5.1560***	0.9829***	0.0789**	-0.0002*			

Note: ***, **, * and (.) are statistically significant at the 0.001, 0.01, 0.05 and 0.1 level, respectively.

As expected, from the linear correlation analysis, population and residential, commercial and service area have a positive effect on ridership. Although the signal of *metrodist_net* is negative, it means that the closer a bus stop is from a metro station, the higher the number of Boardings at it will be. Moreover, it should be noted that all parameter estimates show little variation across the models (except for the intercept in the second model), which suggests that some factors, such as multicollinearity, that could cause misunderstanding in the coefficient's values, are not present.

Based on statistical significance, one can assume that the order of importance of predictors used might be: *ln_pop*, *res_com_serv area* and *metrodist_net*, which was also the sequence of predictors entering in the stepwise selection method. The percentage of the nugget effect in relation to the sill increased compared to the univariate case. In spite of that, in two of the five models, this parameter remains below 50%. According to Cambardella et al. (1994), variables with nugget-to-sill ratio of 25% up to 75% can still be considered as spatially dependent, in a moderate way. Conversely, range was significantly reduced, showing values from 1.2km to 2.2km, approximately.

Table 5 presents the goodness-of-fit measures applied to models shown in Table 4. Ordinary Kriging results, based on exponential semivariogram, are also displayed.

Table 5: Goodness-of-fit measures

Case	Predictors	Model	MAE	RMSE	MAPE	R
0	-	Ordinary Kriging	8.308	12.684	96.27%	0.296*
1.1	Ln_pop	Universal Kriging	**5.211**	**8.117**	**42.03%**	**0.800****
		Linear regression	7.820	11.028	72.51%	0.537**
1.2	Res_com_serv area	Universal Kriging	5.758	9.500	50.43%	0.703**
		Linear regression	8.686	13.830	81.89%	0.309**
2.1	Ln_pop and res_com_serv area	Universal Kriging	6.071	9.434	48.10%	0.683**
		Linear regression	7.424	**10.694**	62.36%	**0.586****
2.2	Ln_pop and metrodist_net	Universal Kriging	5.437	8.460	43.89%	0.772**
		Linear regression	7.981	11.341	68.58%	0.502**
3	Ln_pop, res_com_serv area and metrodist_net	Universal Kriging	5.926	9.355	46.39%	0.690**
		Linear regression	**7.409**	10.782	**60.44%**	0.571**

Note: ** and * are statistically significant at the 0.01 and 0.05 level, respectively (one-tailed).

Based on the goodness-of-fit measures, Universal Kriging models can be ranked, from the best to the worst, as follows: 1.1, 2.2, 3, 1.2 and 2.1. The best models for linear regression, in turn, were 3 and 2.1, followed by 1.1, 2.2 and 1.2. Comparing all eleven models simultaneously, UK estimates outperformed all other models, meaning that even the UK cases with only one or two predictors showed better results than linear regression with three predictors. Ordinary Kriging, which is a univariate technique, presented a MAE and RMSE lower than those of linear regression with *res_com_serv area* as the predictor.

Although models with more predictors may better explain the variance of interest variable, estimates can show no or little improvement when a new explanatory variable is added to the model, even a statistically significant one. The best results, from both Universal Kriging and linear regression, are highlighted in bold in Table 5. In the case of Universal Kriging, the model with only *ln_pop* as the predictor yielded the best estimates, while for linear regression, the best results are

those from models 2.1, which use *ln_pop* and *res_com_serv area*, and 3, which uses all three predictors.

The reason for that could be the fact that when multiple predictors are added to the linear combination part of UK, spatial structure of residuals starts to get blurred. As shown in Table 4, the nugget effect of cases 2.1 and 3 are the highest ones, corresponding to 70% of sill, approximately. However, even in these cases, estimates can still be improved through geostatistical modeling as Universal Kriging do not overlook the remaining spatial dependence on residuals.

Table 5 also proves that kriging estimates can, in fact, be improved by including explanatory variables in geostatistical modeling. Comparing Ordinary Kriging results with those of the UK last ranked case (model 2.1), there is a reduction in MAE, RMSE and MAPE of about 27%, 26% and 50%, respectively, while R increased 131%. Considering the best model of UK (1.1), these numbers increase to 37%, 36%, 56% and 170%, respectively. Moreover, ridership estimates can also be significantly improved by geostatistical modeling compared to linear regression: the most subtle improvements were for model 2.1, which showed reductions of 18%, 12% and 23% in MAE, RMSE and MAPE, respectively, and an increase of 17% in R. On the other hand, MAE and RMSE reduced 34% and 31%, respectively, in model 1.2, and R increase reached 128%. The best MAPE improvement corresponded, in turn, to model 1.1, with a reduction of 42%. These results indicate that not only geostatistical modeling can provide the best ridership estimates, but also that improvements will depend on what predictors are being used.

Finally, linear regression models from Table 1 exhibited the following adjusted coefficients of determination (adjusted R^2): 0.328 and 0.330 (Ryan and Frank 2009), 0.69, 0.62 and 0.53 (Dill et al. 2013), and 0.772 and 0.762 (Kerkman, Martens and Meurs 2015). Meanwhile, adjusted R^2 for linear regression models in Table 5 was: (1.1) 0.453, (1.2) 0.188, (2.1) 0.545, (2.2) 0.518, and (3) 0.572. It should be noted that despite using much less information, some linear regression results obtained in the present study, which were outperformed by UK, are similar or slightly better than the first two, which suggests that the three predictors used were correctly specified, as they can explain a significant part of the ridership variance, show little variation when a new predictor is added to the model, and are statistically significant.

In order to provide a disaggregated analysis of errors and allow a comparison between models, Figure 3 shows maps of error ratios for Ordinary Kriging (a); Linear regression and Universal Kriging, both with all predictors, which was considered the best result of linear regression (b and c, respectively); Linear regression and Universal Kriging, both with the *ln_pop* as the predictor, which is the best result of UK (d and e, respectively).

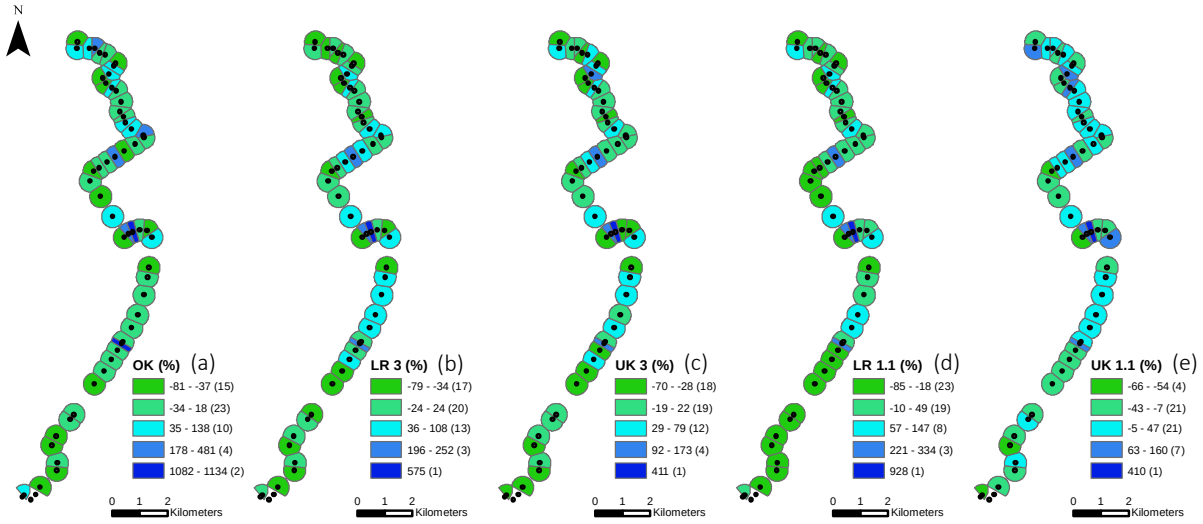

Figure 3: Error ratios of (from left to right) Ordinary Kriging (a), Linear Regression and Universal Kriging with all three predictors (b and c), and Linear Regression and Universal Kriging with *ln_pop* (d and e).

Three bus stops had an observed Boarding value equal to zero. Therefore, the error ratio could not be calculated for these cases, which is the reason why they do not appear in Figure 3. From the minimum and maximum error ratios, as well as the limits for each error group and the amount of bus stops in each group, the following conclusion can be drawn: the best estimates come from UK with *ln_pop*, then UK with all the predictors, followed by linear regression with three predictors, linear regression with *ln_pop*, and lastly Ordinary Kriging.

Despite the fact that Ordinary Kriging showed some very high errors, a detailed analysis of percentiles reveals that OK and linear regression with all predictors had the same amount of bus stops with an error ratio between -30% and 30%, approximately, which corresponds to 37% of the total data. Linear regression with *ln_pop*, UK with three predictors and UK with *ln_pop* showed, respectively, 34%, 45% and 50% of bus stops with an error rate between -30% to 30%, which was considered a satisfactory range of error.

As Ordinary Kriging assumes the interest variable mean is a constant, OK modeling of variables that present a wide range of variation usually yields high errors. Conversely, the same amount of bus stops showed error ratios ranging from -30% and 30% in both OK, which is a univariate technique, and linear regression with all predictors. On the other hand, as it does not include any explanatory variable, OK can only be applied to short-term public transportation planning, in which all built environment and transportation system variables are assumed to remain constant.

Following the bus stop sequence from top to bottom in Figure 3, extreme error ratios occurred at bus stops 32, in all cases, and 42, in Ordinary Kriging estimates. The main reason for that might be the size of catchment areas devoted to these points, which are the smallest ones due to high proximity to neighboring bus stops. This problem could be solved by running an alternative modeling in which all catchment areas would have the same size, overlapping each other, and then include some explanatory variable that could control the occurrence of competitive bus stops, as performed by Kerkman, Martens and Meurs (2015).

5. Conclusions and Final Remarks

Public Transportation plays an important role in the sustainable development of cities and social inclusion. In order to promote the proper functioning of this system, travel demand models have been developed and refined over the years, seeking to consider a characteristic normally found in travel data: spatial autocorrelation. Another important feature of travel demand data is its multivariate nature. However, regarding the bus transit demand, there is a lack of multivariate spatial models that consider the scarce nature of travel data, which are expensive to collect, and also need an appropriate level of detail. Thus, the main aim of this study was to estimate the Boarding variable along a bus line from the city of Sao Paulo - Brazil, by means of a multivariate geostatistical modeling at the bus stop level. As specific objectives, a comparative analysis conducted by applying Universal Kriging, Ordinary Kriging and Ordinary Least Squares Regression to the same travel demand variable was proposed.

In general, results showed that the inclusion of explanatory variables to the kriging estimator contributes, in fact, to increasing the prediction power of the technique. However, the performance of the models with only one predictor did not follow the same pattern in both geostatistical and traditional modeling. This reinforces the opportunity to investigate what would be the best predictors to be used in transportation demand spatial approaches to avoid those that would not bring significant improvements, but whose acquisition would require additional costs. Results also suggested that Ordinary Kriging, which does not require additional information about explanatory variables, can be competitive to linear regression with only one predictor. This comes, probably, from the fact that OK already considers the spatial autocorrelation present in the Boarding variable. However, this interpolator has the disadvantage of not being able, from only the available data about the interest variable, to predict its values for other scenarios, including future ones. This capacity is observed only in Universal Kriging and Linear Regression. In addition, estimates from all geostatistical cases revealed a better adjustment of exponential semivariograms to Boarding data.

Although the results from Universal Kriging may suggest that the lower the number of predictors, the better the estimates will be, we do not encourage ignoring additional information when it is available and contributes, in fact, to explaining interest variables. However, when detailed data is not provided, which is the case of various cities, in development countries, especially the small and medium-sized ones, spatial models with little information available could also yield good estimates. In general, results showed that traditional modeling can always be improved by geostatistical multivariate interpolators, not only in cases where there is only one predictor, but also when a large amount of information is used. Best results from UK showed 50% of bus stops with error between -30% and 30%. In turn, regarding the best results from linear regression, only 37% of bus stops had errors within this range.

Therefore, three main contributions are highlighted: the methodological advance of using a detailed geostatistical approach, the bus stop level, on bus ridership modeling; the benefits provided by the models regarding the land use and bus network planning; and resource savings of field surveys for collecting travel data. In order to compare the achieved results with another spatial method that, similar to the geostatistical interpolators, also creates a surface of estimated values, Geographically Weighted Regression is recommended for the same dataset used in the present study. Nevertheless, it is opportune to compare the OK and UK results to those of generalized linear models (Poisson and Negative Binomial regressions), which consider the

positive asymmetry of count data, and those of geographically weighted models with count distributions for the response variable.

ACKNOWLEDGEMENT

The authors would like to thank the São Paulo Research Foundation (FAPESP, Brazil - Process 2019/12054-4), the National Council for Scientific and Technological Development (CNPq, Brazil - Process 304345 / 2019-9), and SPTrans, for the Boarding/Alighting survey data used in this study.

AUTHOR´S CONTRIBUTION

The first author (Samuel de França Marques) was responsible for Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing - initial draft and Writing - review and editing; the second author (Cira Souza Pitombo) was responsible for Supervision and Writing - revision and editing.

REFERENCES

Anselin, L. 2004. Exploring spatial data with GeoDaTM: a workbook. *Urbana*, 51(61801). Available at: <http://www.csiss.org/clearinghouse/GeoDa/geodaworkbook.pdf> [Accessed November 2020].

Anselin, L. Syabri, I. and Kho Y. 2005. GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis*, 38(1), pp5–22. doi: https://doi.org/10.1111/j.0016-7363.2005.00671.x

Asa, E. Saafi, M. Membah, J. and Billa, A. 2012. Comparison of Linear and Nonlinear Kriging Methods for Characterization and Interpolation of Soil Data. *Journal of Computing in Civil Engineering*, 26(1), pp11–18. doi: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000118

Bartlett, M. S. 1947. The Use of Transformations. *Biometrics*, 3(1), pp 39–52.

Blainey, S. and Mulley, C. 2013. Using geographically weighted regression to forecast rail demand in the Sydney region. In: *Australasian Transport Research Forum 2013*. Brisbane, Australia, 2-4 October 2013.

Blainey, S. and Preston, J. 2010. A geographically weighted regression based analysis of rail commuting around Cardiff, South Wales. In: *12th World Conference on Transport Research*. Lisbon, Portugal, 11-15 July 2010.

Bundala, D. Bergenheim, W. and Metz, M. 2014. *v.net.allpairs - Computes the shortest path between all pairs of nodes in the network*. GRASS GIS code. Available at: <https://trac.osgeo.org/grass/browser/grass/branches/releasebranch_7_2/vector/v.net.allpairs> [Accessed November 2020].

Cambardella, C. A. Moorman, T. B. Novak, J. M. Parkin, T. B. Karlen, D. L. Turco, R. F. and Konopka, A. E. 1994. Field-scale variability of soil properties in central Iowa soils. *Soil science society of America journal*, 58(5), pp1501-1511. doi: https://doi.org/10.2136/sssaj1994.03615995005800050033x

Cardozo, O. D. García-Palomares, J. C. and Gutiérrez, J. 2012. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, 34(Supplement C), pp548–558. doi: https://doi.org/10.1016/j.apgeog.2012.01.005

Cervero, R. 2006. Alternative Approaches to Modeling the Travel-Demand Impacts of Smart Growth. *Journal of the American Planning Association*, 72(3), pp285–295. doi: https://doi.org/10.1080/01944360608976751

Chakour, V. and Eluru, N. 2013. Examining the Influence of Urban form and Land Use on Bus Ridership in Montreal. *Procedia - Social and Behavioral Sciences*, 104(Supplement C), pp875–884. doi: https://doi.org/10.1016/j.sbspro.2013.11.182

Chakour, V. and Eluru, N. 2016. Examining the influence of stop level infrastructure and built environment on bus ridership in Montreal. *Journal of Transport Geography*, 51(Supplement C), pp205–217. doi: https://doi.org/10.1016/j.jtrangeo.2016.01.007

Chiou, Y. C. Jou, R. C. and Yang, C. H. 2015. Factors affecting public transportation usage rate: Geographically weighted regression. *Transportation Research Part A: Policy and Practice*, 78, pp161-177. doi: https://doi.org/10.1016/j.tra.2015.05.016

Choi, J. Lee, Y. J. Kim, T. and Sohn, K. 2012. An analysis of Metro ridership at the station-to-station level in Seoul. *Transportation*, 39(3), pp705–722. doi: https://doi.org/10.1007/s11116-011-9368-3

Chow, L.-F. Zhao, F. Liu, X. Li, M.-T. and Ubaka, I. 2006. Transit Ridership Model Based on Geographically Weighted Regression. *Transportation Research Record*, 1972, pp105–114. doi: https://doi.org/10.3141/1972-15

Chu, X. 2004. *Ridership models at the stop level*. National Center for Transit Research: University of South Florida.

Cressie, N. A. C. 1993. *Statistics for spatial data*. John Wiley & Sons, Inc.

Daya, A. A. and Bejari, H. 2015. A comparative study between simple kriging and ordinary kriging for estimating and modeling the Cu concentration in Chehlkureh deposit, SE Iran. *Arabian Journal of Geosciences*, 8(8), pp6003–6020. doi: https://doi.org/10.1007/s12517-014-1618-1

Dill, J. Schlossberg, M. Ma, L. and Meyer, C. 2013. Predicting Transit Ridership at Stop Level: Role of Service and Urban Form. In: *92nd Annual Meeting of the Transportation Research Board*, Washington, United States of America, 13-17 January 2013.

Fotheringham, A. S. Brunsdon, C. and Charlton, M. 2003. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Gan, Z. Feng, T. Yang, M. Timmermans, H. and Luo, J. 2019. Analysis of Metro Station Ridership Considering Spatial Heterogeneity. *Chinese Geographical Science*, 29(6), pp1065–1077. doi: https://doi.org/10.1007/s11769-019-1065-8

George, P. and Kattor, G. J. 2013. Forecasting Trip Attraction Based On Commercial Land Use Charateristics. *International Journal of Research in Engineering and Technology*, 2(9), pp471–479.

GeoSampa. *São Paulo predominant land use in 2016*. [online] Available at: <http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx> [Accessed February 2020].

Goovaerts, P. 1997. *Geostatistics for Natural Resources and Evaluation*. Oxford University Press.

Gutiérrez, J. Cardozo, O. D. and García-Palomares, J. C. 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography*, 19(6), pp1081–1092. doi: https://doi.org/10.1016/j.jtrangeo.2011.05.004

Hiemstra, P. H. Pebesma, E. J. Heuvelink, G. B. M. and Twenhöfel, C. J. W. 2010. Using rainfall radar data to improve interpolated maps of dose rate in the Netherlands. *Science of The Total Environment*, 409(1), pp123–133. doi: https://doi.org/10.1016/J.SCITOTENV.2010.08.051

Hollander, Y. and Liu, R. 2008. The principles of calibrating traffic microsimulation models. *Transportation*, 35(3), pp347–362. doi: https://doi.org/10.1007/s11116-007-9156-2

IBM 2016. *IBM SPSS Statistics 24 Core System User's Guide*. International Business Machines. [online] Available at: <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/IBM_SPSS_Statistics_Core_System_User_Guide.pdf> [Accessed November 2019].

Kerkman, K. Martens, K. and Meurs, H. 2015. Factors Influencing Stop-Level Transit Ridership in Arnhem–Nijmegen City Region, Netherlands. *Transportation Research Record*, 2537(1), pp23-32. doi: https://doi.org/10.3141/2537-03

Kiš, I. M. 2016. Comparison of ordinary and universal kriging interpolation techniques on a depth variable (a case of linear spatial trend), case study of the šandrovac field. *Mining-geological-petroleum engineering bulletin*, 31(2), pp41–58. doi: https://doi.org/10.17794/rgn.2016.2.4

Krige, D. G. 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6), pp119–139.

Liu, W. Du, P. and Wang, D. 2015. Ensemble learning for spatial interpolation of soil potassium content based on environmental information. *PLoS ONE*, 10(4), pp1-11. doi: https://doi.org/10.1371/journal.pone.0124383

Lopes, B. S. Brondino, C. N. and Rodrigues da Silva, N. A. 2014. GIS-Based Analytical Tools for Transport Planning: Spatial Regression Models for Transportation Demand Forecast. *ISPRS International Journal of Geo-Information*, 3(2), pp565-583. doi: https://doi.org/10.3390/ijgi3020565

Marques, S. F. 2019. *Estimativa do volume de passageiros ao longo de uma linha de transporte público por ônibus a partir da Geoestatística*. MSc. University of São Paulo. doi: https://doi.org/10.11606/D.18.2019.tde-26042019-110232.

Marques, S. F. and Pitombo, C. S. 2021. Ridership Estimation Along Bus Transit Lines Based on Kriging: Comparative Analysis Between Network and Euclidean Distances. *Journal of Geovisualization and Spatial Analysis*, 5, 7. doi: https://doi.org/10.1007/s41651-021-00075-w

Marques, S. F. and Pitombo, C. S. 2019. Estimativa do volume de passageiros ao longo de uma linha de transporte público por ônibus a partir da Geoestatística. *Transportes*, 27(3), pp15–35. doi: https://doi.org/10.14295/transportes.v27i3.2007

Matheron, G. 1963. Principles of geostatistics. *Economic Geology*, 58(8), pp1246–1266.

Matheron, G. 1971. *The Theory of Regionalized Variables and Its Applications*. Paris: Les Cahiers du Centre de Morphologie Mathematique in Fontainebleu.

Metrô 2019. *2017 Origin and Destination Survey*. Companhia do Metropolitano De São Paulo, Secretaria Estadual dos Transportes Metropolitanos. [online] Available at: <http://www.metro.sp.gov.br/pesquisa-od/> [Accessed November 2019].

Moran, P. A. P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society*. Series B (Methodological), 10(2), pp243–251.

Mubarak, N. Hussain, I. Faisal, M. Hussain, T. Shad, M. Y. AbdEl-Salam, N. M. and Shabbir, J. 2015. Spatial Distribution of Sulfate Concentration in Groundwater of South-Punjab, Pakistan. *Water Quality, Exposure and Health*, 7(4), pp503–513. doi: https://doi.org/10.1007/s12403-015-0165-7

Nalder, I. A. and Wein, R. W. 1998. Spatial interpolation of climatic Normals: test of a new method in the Canadian boreal forest. *Agricultural and Forest Meteorology*, 92(4), pp211–225. doi: https://doi.org/10.1016/S0168-1923(98)00102-6

Olea, R. A. 2006. A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research and Risk Assessment*, 20(5), pp307–318. doi: https://doi.org/10.1007/s00477-005-0026-1

Oliver, M. A. and Webster, R. 2015. *Basic steps in geostatistics: the variogram and kriging*. Springer.

Ortúzar, J. D. and Willumsen, L. G. 2011. *Modelling Transport*. John Wiley & Sons.

Papritz, A. 2020a. *georob: Robust Geostatistical Analysis of Spatial Data*. R package version 0.3-13. [online] Available at: <https://CRAN.R-project.org/package=georob> [Accessed November 2020].

Papritz, A. 2020b. *Tutorial and Manual for Geostatistical Analyses with the R package georob*. Available at: <https://cran.r-project.org/web/packages/georob/vignettes/georob_vignette.pdf> [Accessed November 2020].

Pendyala, R. M. Shankar, V. N. and McCullough, R. G. 2000. Freight Travel Demand Modeling: Synthesis of Approaches and Development of a Framework. *Transportation Research Record*, 1725(1), pp9–16. doi: https://doi.org/10.3141/1725-02

Pulugurtha, S. S. and Agurla, M. 2012. Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. *Journal of Public Transportation*, 15(1), pp33–52. Available at: <https://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1095&context=jpt> [Accessed in November 2020].

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. [online] Available at: <https://www.R-project.org/> [Accessed in November 2020].

Ryan, S. and Frank, L. 2009. Pedestrian Environments and Transit Ridership. *Journal of Public Transportation*, 12(1), pp39–57. doi: https://doi.org/10.5038/2375-0901.12.1.3

Ribeiro Jr., P. J. and Diggle, P. J. 2016. *geoR: Analysis of Geostatistical Data*. R package version 1.7-5.2. [online] Available at: <https://CRAN.R-project.org/package=geoR> [Accessed in November 2020].

Sarlas, G. and Axhausen, K. W. 2016. Exploring spatial methods for prediction of traffic volumes. In: *16th Swiss Transport Research Conference (STRC 2016)*. Monte Verità, Switzerland, 18-20 May 2016. doi: https://doi.org/10.3929/ethz-b-000116988

Seo, Y. Kim, S. and Singh, V. P. 2015. Estimating Spatial Precipitation Using Regression Kriging and Artificial Neural Network Residual Kriging (RKNNRK) Hybrid Approach. *Water Resources Management*, 29(7), pp2189–2204. doi: https://doi.org/10.1007/s11269-015-0935-9

Shamo, B. Asa, E. and Membah, J. 2015. Linear Spatial Interpolation and Analysis of Annual Average Daily Traffic Data. *Journal of Computing in Civil Engineering*, 29(1), pp4014022. doi: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000281

Sun, L.-S. Wang, S.-W. Yao, L.-Y. Rong, J. and Ma, J.-M. 2016. Estimation of transit ridership based on spatial analysis and precise land use data. *Transportation Letters*, 8(3), pp140-147. doi: https://doi.org/10.1179/1942787515Y.0000000017

Taharin, M. R. and Roslee, R. 2017. Comparison of Cohesion (c'), and Angle of Internal Friction (Φ') Distribution in Highland Area of Kundasang by using Ordinary Kriging and Simple Kriging. *Geological Behavior*, 1(1), pp16–18. doi: https://doi.org/10.26480/gbr.01.2017.16.18

Varagouli, E. G. Simos, T. E. and Xeidakis, G. S. 2005. Fitting a multiple regression line to travel demand forecasting: The case of the prefecture of Xanthi, Northern Greece. *Mathematical and Computer Modelling*, 42(7), pp817–836. doi: https://doi.org/10.1016/j.mcm.2005.09.010

Viswanathan, R. Jagan, J. Samui, P. and Porchelvan, P. 2015. Spatial Variability of Rock Depth Using Simple Kriging, Ordinary Kriging, RVM and MPMR. *Geotechnical and Geological Engineering*, 33(1), pp69–78. doi: https://doi.org/10.1007/s10706-014-9823-y

Wang, F. 2001. Explaining Intraurban Variations of Commuting by Job Proximity and Workers' Characteristics. *Environment and Planning B: Planning and Design*, 28(2), pp169–182. doi: https://doi.org/10.1068/b2710

Wang, C. and Zhu, H. 2016. Combination of Kriging methods and multi-fractal analysis for estimating spatial distribution of geotechnical parameters. *Bulletin of Engineering Geology and the Environment*, 75(1), pp413–423. doi: https://doi.org/10.1007/s10064-015-0742-9

Yan, X. and Su, X. G. 2009. *Linear regression analysis: theory and computing*. World Scientific.

Zhang, D. and Wang, X. C. 2014. Transit ridership estimation with network Kriging: A case study of Second Avenue Subway, NYC. *Journal of Transport Geography*, 41, pp107–115. doi: https://doi.org/10.1016/j.jtrangeo.2014.08.021

Zhao, F. Chow, L. F. Li, M. T. Ubaka, I. and Gan, A. 2003. Forecasting transit walk accessibility: Regression model alternative to buffer method. *Transportation Research Record*, 1835, pp34–41. doi: https://doi.org/10.3141/1835-05