



# Transit Ridership Modeling at the Bus Stop Level: Comparison of Approaches Focusing on Count and Spatially Dependent Data

Samuel de França Marques<sup>1</sup>  · Cira Souza Pitombo<sup>1</sup> 

Received: 23 July 2021 / Accepted: 22 August 2022

© The Author(s) 2022

## Abstract

Boarding and alighting modeling at the bus stop level is an important tool for operational planning of public transport systems, in addition to contributing to transit-oriented development. The interest variables, in this case, present two particularities that strongly influence the performance of proposed estimates: they demonstrate spatial dependence and are count data. Moreover, in most cases, these data are not easy to collect. Thus, the present study proposes a comparison of approaches for transit ridership modeling at the bus stop level, applying linear, Poisson, Geographically Weighted and Geographically Weighted Poisson (GWPR) regressions, as well as Universal Kriging (UK), to the boarding and alighting data along a bus line in the city of São Paulo, Brazil. The results from goodness-of-fit measures confirmed the assumption that adding asymmetry and spatial autocorrelation, isolated and together, to the transportation demand modeling, contributes to a gradual improvement in the estimates, highlighting the GWPR and UK spatial estimation techniques. Moreover, the spatially varying relationships between the variables of interest (boardings and alightings) and their predictors (land use and transport system features around the bus stops), shown in the present study, may support land use policies toward transit-oriented development. In addition, by using an approach with little information, the good results achieved proved that satisfactory boarding and alighting modeling can be done in regions where there is a lack of travel demand data, as in the case of emerging countries.

**Keywords** Boarding per bus stop · Alighting per bus stop · Linear regression · Poisson regression · Geographically Weighted Poisson Regression · Universal Kriging

---

✉ Samuel de França Marques  
samuelmarques@usp.br

Cira Souza Pitombo  
cirapitombo@gmail.com

<sup>1</sup> Department of Transportation Engineering, São Carlos School of Engineering, University of São Paulo, Trabalhador São-carlense Avenue, 400, São Carlos, São Paulo 13566-590, Brazil

## Introduction and Background

Alignment between urban planning and transport is one of the pillars of sustainable city development. Associations between land use and urban mobility support the development of sustainable public policies, which are essential for encouraging Transit Ridership (TR), an important instrument for social inclusion and accessibility. In this context, transport modeling is one of the tools that by quantifying and explaining the effects of urban practices concerning the displacement of people and goods, provide support to urban policies at the most diverse geographic scales.

Generally conditioned by data availability, urban travel modeling encompasses different approaches, which can be differentiated by the spatial unit of analysis used. Regarding Public Transport, studies can be found at the system level (Cervero & Dai, 2014; Hensher & Golob, 2008; Hensher et al., 2014; Joonho et al., 2019; Taylor et al., 2009), on Traffic Analysis Zones (TAZs), neighborhoods or districts (Chiou et al., 2015; Kalaanidhi & Gunasekaran, 2013; Ma et al., 2018; Siddiqui et al., 2015; Tu et al., 2018), bus lines (Kyte et al., 1985; Peng et al., 1997), train stations, metro stations and bus stops (Gan et al., 2019; Pulugurtha & Agurla, 2012; Sun et al., 2016; Zhu et al., 2019), and individual or household (Ewing et al., 2014; Siddiqui et al., 2015) ranging from the most aggregated to the most disaggregated level. In a simplified way, the adopted spatial unit of analysis strongly influences the intervening factors, or explanatory variables, which can be considered in the study.

The urbanized area or system approach allows, for example, the inclusion of covariates such as population, jobs, age and color distribution, regional, meteorological and topographic characteristics, Gross Domestic Product (GDP), income, fleet, fare, capacity, number of Public Transport (PT) stations, modal split, PT network mileage, frequency, characteristics of the road system, etc. Models that analyze only an urbanized area, segmented into Traffic Analysis Zones, neighborhoods or districts, are able to refine the socioeconomic, land use and transportation system covariates, compared to previous approaches. In this case, however, fare variations cannot be analyzed, for situations where it is unique in the city, as well as fleet, climate and other factors.

Research carried out on bus lines, in turn, maintains the aggregated characteristics of the Traffic Analysis Zones, however, considering that they are usually based on time series, the effect of the variation in the fare can be analyzed once more. In addition, covariates related to the type of line are also liable to be included in the models. The more disaggregated approaches (individual and household), on the other hand, in addition to further refining the socioeconomic characteristics of previous treatments, add to the set of factors assessed in the Traffic Analysis Zones, trip characteristics, such as time, distance and cost, and user perception.

Finally, between the most disaggregated level and bus lines, some studies address train, metro stations and bus stops as spatial aggregation units. These models, which consist of one of the most recent approaches of Transit Rider-ship, can efficiently quantify the benefits of transit-oriented development, that is, from urban policies applied in neighborhoods, which converge with urban planners' needs. Traditional Traffic Analysis Zones modeling, in contrast, assumes

an average value of the explanatory variables in each spatial unit, which prevents capturing variations at the local level and can lead to ecological fallacy. On the other hand, considering the bus stop as a unit of analysis, boarding and alighting volume estimates can be obtained using models, quickly and economically, supporting the planning of the PT network (Cervero, 2006). This modeling is carried out based on socioeconomic variables, land use and the transport system around the stops.

The travel data, however, which consist of the variable of interest in these models, show two characteristics of fundamental importance for the performance of the estimates, which are: they refer to counts, that is, they can assume only non-negative integer values and have asymmetry (they are heteroscedastic); and present spatial autocorrelation, which means that travel demand values close to each other in space tend to demonstrate similar behavior. Thus, travel demand models have been improved over the years so as to account for these unique characteristics in the modeling process. Concerning the spatial units of interest for sustainable urban planning (bus stops and stations), studies can be found regarding the modeling of Transit Ridership at the bus stop or station level based on classical linear regression (Cervero, 2006; Gutiérrez et al., 2011; Ryan & Frank, 2009). This traditional model, also known as Ordinary Least Squares (OLS), is appropriate for continuous variables and its residuals cannot be dependent on each other, in which case the OLS assumptions are violated (Yan & Su, 2009) and the statistical inference is compromised, that is, the estimator is no longer the one with the least variance. Solutions such as variable transformations and decay functions were adopted by some authors to avoid such problems, although the real nature of the data has not been considered.

In the 1980s, an expansion of the linear model to other probability distributions introduced Poisson and Negative Binomial regressions that, unlike the normal distribution, model count data. These models, which have also been used to address Transit Ridership at the bus stop and station level (Choi et al., 2012; Chu, 2004; Pulugurtha & Agurla, 2012), can demonstrate a better performance than the traditional OLS. Despite this, these approaches still overlook the spatial autocorrelation found in the response variable.

Attempts to solve this limitation culminated in the emergence of spatial regressions, which can consider autocorrelation based on inclusion, as a covariate, of the spatially lagged dependent variable (Spatial Lags Model - SLM), or through model residuals (Spatial Error Model - SEM), and in both cases, the spatial interaction is captured through a spatial weight matrix, usually based on the distance between the points of the database (Fotheringham et al., 2003). These techniques have also been used in ridership models at the station level (Gan et al., 2019), although, according to Fotheringham et al. (2003), these models do not reflect the spatial heterogeneity of the database on a local level because the autocorrelation is expressed in terms of only one parameter. Geographically Weighted Regression (GWR), which generates a different model for each geographic coordinate, would be more appropriate, in this case, to address the autocorrelation and spatial heterogeneity of the estimated parameters (Brunsdon et al., 1996). In GWR applications to Transit Ridership (Blainey & Mulley, 2013; Blainey & Preston, 2010; Cardozo et al., 2012), the results always demonstrate a better performance than the global models.

Despite being able to deal satisfactorily with the database's spatial dependence, GWR has limitations that, similar to the OLS model, also assumes normality of the variable of interest, which, in the case of Transit Ridership, is not observed. Thus, geographically weighted models for count data have recently been developed, called Geographically Weighted Poisson Regression (GWPR) and Geographically Weighted Negative Binomial Regression (GWNBR). Although these models can be easily found in traffic accident modeling (Bao et al., 2018; Gomes et al., 2017, 2019; Liu et al., 2017; Obelheiro et al., 2020; Xu et al., 2017; Xu & Huang, 2015), using it for ridership forecasting is still rare, and it is restricted to the application of GWPR in the scope of metro stations (Liu et al., 2018) and GWNBR for train ridership (Zhu et al., 2019), which again points to a better performance of local models compared to their global version, Poisson regression and Negative Binomial regression, respectively.

Another multivariate spatial model that, similar to GWR, also addresses spatial dependence and is capable of generating a continuous surface of estimated values, refers to the Geostatistics interpolator known as Universal Kriging (UK). The greatest benefit of this technique is to be able to use the maximum available information on the response and explanatory variables when forecasting the values of interest in non-sampled sites, which makes it highly recommended for dealing with the lack of data, a situation often found in travel demand variables along bus lines. In the context of Transit Ridership, few studies have been found to date: Zhang and Wang (2014) applied UK to estimate the number of Boardings in metro stations. On the other hand, Marques and Pitombo (2021a) tested the suitability of UK to model Boardings at the bus stop level, using different groups of predictors. Although the results were satisfactory, the authors compared UK results only with Linear Regression, and did not account for the potential spatial heterogeneity of the predictors. Models for count data were overlooked as well. The main differences between previous transit ridership studies and the present article are outlined in Table 1.

Based on the studies cited above, the following research gaps can be highlighted: (1) Application of spatial models in the context of bus stops: the approaches found so far are restricted to addressing the asymmetry shown by bus stop travel data, overlooking the spatial autocorrelation potentially found in the models, as well as both characteristics simultaneously. (2) Ridership modeling at the bus stop level: although the approaches by train and metro stations also represent a contribution to sustainable urban planning, bus stops are densely distributed within cities (as opposed to rail stations), allowing the incorporation of characteristics from a higher number of neighborhoods into the modeling. Furthermore, it cannot be said that such data fall into the group of scarce variables, since the information on station boarding and alighting is obtained relatively easily. Bus transit, on the other hand, is a much more popular system than rail transit, which is found only in large cities. (3) In most of the studies whose spatial unit of analysis is bus stops (Dill et al., 2013; Kerkman et al., 2015; Ryan & Frank, 2009), the authors apply only the traditional linear model. Although Chu (2004) applied both the OLS and Poisson regressions, only the results of the count data model are shown. Thus, no comparison is made between the two types of models, which prevents the visualization of the gains provided by using the most appropriate regression. Even in other studies, which address

**Table 1** Methodological differences between the present study and the previous ones

References	Model(s) used	Geographic aggregation unit	Differences from the present study
Dill et al. (2013); Kerkman et al. (2015); Ryan and Frank (2009)	Ordinary Least Squares regression	Bus stop	Count data features, spatial dependence and spatially varying relationships are not addressed.
Chu (2004); Pulugurtha and Agurta (2012)	Poisson or Negative Binomial regressions	Bus stop	Spatially varying relationships are not addressed.
Blainey and Mulley (2013); Blainey and Preston (2010); Cardozo et al. (2012); Choi et al. (2012)	GWR and OLS regression	Train or metro station	Count data features are overlooked. Spatially varying relationships at station level may be different from the bus stop case. Most cities do not have rail transit.
Liu et al. (2018); Zhu et al. (2019)	GWPR and Poisson regression; or GWNBR and Negative Binomial regression	Train or metro station	Spatially varying relationships at station level may be different from the bus stop case. Most cities do not have rail transit. There is no proof that the models can perform better than the OLS regression and GWR, which are simpler models.
Gan et al. (2019)	SLM, SEM, GWR and OLS regression	Metro station	Count data features are overlooked. The spatial aggregation unit refers to rail transit.
Zhang and Wang (2014)	Universal Kriging	Metro station	Count data features and spatially varying relationships are overlooked. The spatial aggregation unit refers to rail transit.
Chiou et al. (2015); Ma et al. (2018); Tu et al. (2018)	GWR and OLS regression or Tobit regression	Traffic analysis zone	The bus stop level is capable of quantifying the effects of transit-oriented development. At the TAZ level, intrazonal variation is overlooked.

the transit demand at the station level and in which more than one type of model is applied (Blainey & Mulley, 2013; Blainey & Preston, 2010; Cardozo et al., 2012; Choi et al., 2012; Gan et al., 2019; Liu et al., 2018; Zhu et al., 2019), the regressions address only one of the characteristics previously mentioned, sometimes asymmetry, sometimes spatial autocorrelation, or the authors do not compare it with the traditional linear model. Thus, improvements can be observed provided by including one or the other particularity in ridership modeling, but never both.

Therefore, the present article aims to model the bus stop boarding and alighting volume from GWR for count data and multivariate spatial interpolators. In addition, we aimed to compare different models from classical linear regression to GWPR and UK, using Poisson global regressions, and traditional GWR as well. This proposal intends to allow the visualization of the gradual gains achieved by addressing asymmetry and spatial autocorrelation separated and, later, together. This analysis will be carried out based on a real case study, based on line 6045-10 in the city of São Paulo, Brazil.

This paper has four sections. “[Materials and Method](#)” section describes the proposed method and the database used, dividing it into the description of the dependent variables, independent variables and modeling procedure. The results and discussions are detailed in “[Results and Discussion](#)” section, which is organized as follows: first, the results referring to Boarding are presented and then those of Alighting. Afterward, goodness-of-fit results from all models of Boardings and Alightings are compared. Still in “[Results and Discussion](#)” section, a subsection is presented to compare the results and characteristics of the present study with previous ones. Finally, “[Conclusions, Main Constraints and Final Recommendations](#)” section outlines the main conclusions reached and suggests themes for future research.

## Materials and Method

The database to be used in the present study is based on the results of a boarding and alighting survey carried out on 8 bus lines in the city of São Paulo, São Paulo State, Brazil. For each direction of the lines (inbound and outbound, resulting in 16 cases), a spreadsheet was made available by *São Paulo Transporte SA (SPTrans)*, containing the number of boardings and alightings per bus stop, encoded by an identifier, in 6 different time bands, covering 24 h of a Tuesday in 2017. Having identified the bus stops and their respective geographic coordinates, also provided by *SPTrans*, it was possible to proceed with the spatialization of this database.

## Dependent Variables

The 16 unidirectional lines underwent an exploratory spatial dependence analysis by calculating the Moran index (Moran, 1948) for the number of boardings and alightings per bus stop in the Morning Peak Hours (MPH, from 5 a.m. to 8.59 am), Between Peak Hours (BPH, from 9 a.m. to 3.59 p.m.), Afternoon Peak Hours (APH,

from 4 p.m. to 7.59 p.m.), Evening Peak Hours (EPH, from 8 pm to 11.59 p.m.) and the total number of Boarding and Alighting passengers from 5 a.m. to 11.59 p.m. The Moran index was calculated in the R environment (Paradis et al., 2004; R Core Team, 2020), using weight matrices based on the inverse of the Euclidean distance between the bus stops of the database.

As we are focusing on spatially dependent data, the line to be chosen should be the one whose boarding and alighting volume demonstrates a strong and significant spatial dependence, that is, higher numbers of the Moran index, (when compared to the other lines and time bands) associated with pseudo *p*-values smaller than 0.05. In this context, within the 8 lines considered by the Boarding and Alighting counts survey, the 6045-10-1 line (inbound trip of the 6045-10 line) with 47 bus stops stood out in relation to Boardings in the total number of trips from 5 a.m. to 11.59 p.m. The Alighting volume in that same period showed high and significant spatial autocorrelation in the outbound trip, line 6045-10-2 with 49 bus stops. Thus, the number of Boardings on line 6045-10-1 and Alightings on line 6045-10-2 were established as dependent variables, both referring to the set of trips made from 5 a.m. to 11.59 p.m. Figure 1 shows both directions of line 6045-10 and respective bus stops in the city of São Paulo.

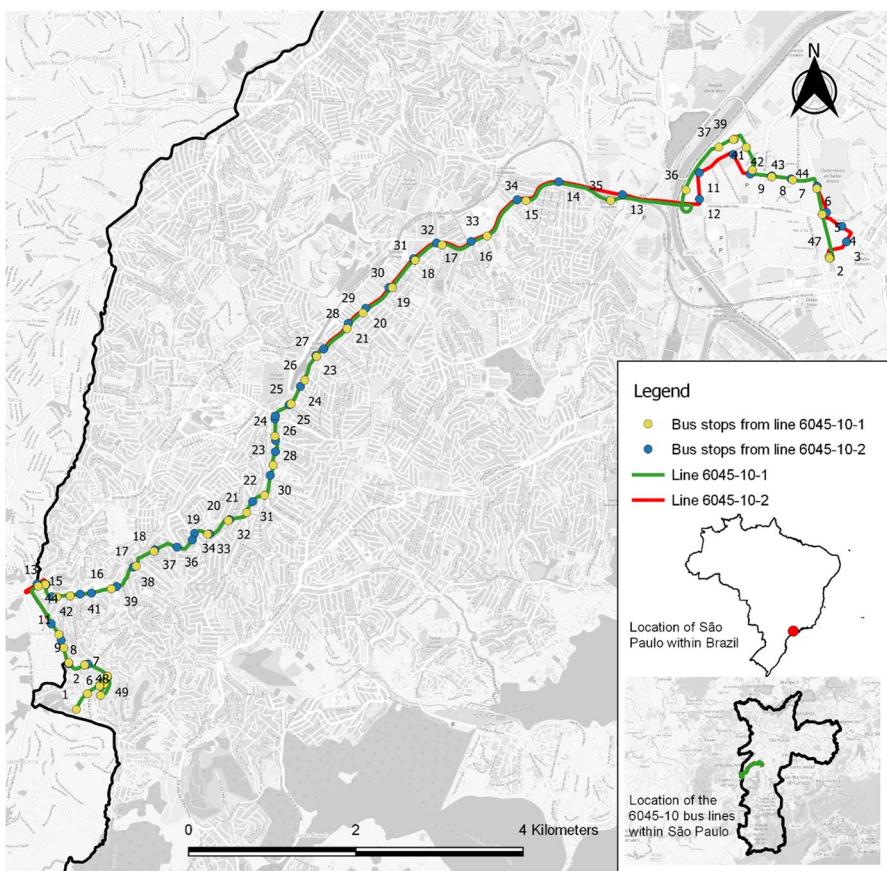
From the bus stop numbering, it can be seen that the inbound trip, line 6045-10-1, starts in the southwest region of the map and ends in the northeast portion. The outbound trip, in turn, line 6045-10-2, originates in the northeast and ends its itinerary in the southwest corner.

## Independent Variables

As mentioned in “[Introduction and Background](#)” section, transit ridership modeling at the bus stop level basically covers three groups of explanatory variables: socioeconomic, land use and the transport system variables. Table 2 summarizes the boarding and alighting models at the bus stop level found in the literature.

As can be observed, the independent variables that model the boarding and alighting volume can also be classified as variables related to Transit Ridership supply or demand. Supply variables include those related to the transport system, while socioeconomic and land use predictors fall into the category of independent variables related to potential demand. Based on this, in the case of the present study, potential predictors were collected both related to bus stops and referring to their area of influence, comprising a 400 m radius buffer centered on the bus stops (Zhao et al., 2003). 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 Geographic Information System (GIS) environment. Table 3 consolidates the potential predictors raised, as well as the database on the basis of which they were calculated.

The potential predictor collection was carried out in a GIS environment. The population variable was calculated based on the areal interpolation of the shapefile of the 2017 Origin/Destination (O/D) Survey (Metrô, 2019), given in Traffic Analysis Zones. The area, in hectares, of the 16 predominant land use categories was



**Fig. 1** Map showing lines 6045-10-1 and 6045-10-2 with their 47 and 49 bus stops, respectively

obtained through the shapefile available on the GeoSampa website, which details the land use in São Paulo, in blocks, in 2016. Among the land use categories available, the following can be found: horizontal and vertical residential, commerce and services, industry and warehouses, public facilities, schools etc. All 16 land use types are cited in Table 3. These data were also used together to calculate the entropy index (Song et al., 2013) around the bus stops, which reflects the mix of land uses found in the region. The other independent variables of potential demand, which include socioeconomic information surrounding the bus stops, were collected from the average of the households sampled by the O/D survey that were covered by the buffer, and, in areas that did not contain any households, the results of the areal interpolation of the aggregated data by Traffic Analysis Zone were used.

To avoid multicollinearity and parameter redundancy, as well as to identify the variables with the greatest potential to explain Boardings and Alightings, Pearson's linear correlation coefficient ( $R$ ) among all the variables in the database was calculated. When a pair of potential predictors had a value of  $R$  equal to or greater than

**Table 2** 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	Income, No-vehicle households, Female (%), Hispanic (%), White (%), Age, No. of inhabitants, No. of jobs, Pedestrian factor
Ryan and Frank (2009)	Boarding+alighting (logarithm)	OLS (log-linear)	Level of service (no. of routes/average waiting time)	Household income, No-vehicle households, Asian population, Residential area	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
Pulugurtha and Agurla (2012)	Boarding	Negative binomial with log-link	On-network characteristics		
Dill et al. (2013)	Boarding+alighting (logarithm)	OLS (log-linear)	Transit service variables, Transportation infrastructure variables		
Kerkman et al. (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 et al. (2015)

**Table 3** Potential predictors for Boarding and Alighting modeling

Predictor(s)	Originated from	Variable type	Source
Distance, in meters, to the nearest bus terminal	Bus stop	Supply	GeoSampa shapefile
Distance, in meters, to the nearest train or metro station (station distance)			
Distance, in meters, to the nearest bus terminal, train or metro station (intra/intermodal dist)			
Number of bus lines passing by the bus stop, except the 6045-10 line			2017 GTFS data provided by SPTrans
Average frequency, in trips per hour, of the bus lines, except the 6045-10 line (frequency)			
Population, in inhabitants (population)	Catchment area	Demand	2017 Origin and Destination survey shapefile, given in Traffic Analysis Zones (Metrô, 2019)
Area, in hectares, of the following land uses: no information; low standard horizontal residential; medium/high standard horizontal residential; low standard vertical residential; medium/high standard vertical residential; commerce and services (com serv area); industry and warehouses; residential, commerce and services; residential, industry and warehouses; commerce, services, industry and warehouses; public facilities; schools; empty land; and without predominance			GeoSampa shapefile, given in blocks
Entropy Index			-
Average household income, in BRL (income)*			Household data from the 2017 Origin and Destination survey (Metrô, 2019)
Average car ownership			
Female (%)			
Population with no complete higher education (%)			
Workers and students (%)			
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			
Number of roads	Supply		Open Street Map
Road length, in meters			
Number of intersections			
Number of intersections per meter of road			
Number of Points of Interest			

\* BRL 1.00 is equivalent to USD 0.18 (Feb. 2021)

0.60, the variable with the lowest correlation with Boardings and Alightings was

discarded. As R values up to 0.60 indicate only a moderate correlation (Profillidis & Botzoris, 2019), this threshold was deemed acceptable in order to combat the omitted variable bias. It is worth mentioning that the variables listed in Table 3 were collected for the bus stops and areas of influence of both lines separately as the inbound and outbound trips are not exactly coincident.

## Modeling

After completing the Boarding and Alighting database with its predictors, we proceeded to the modeling stage. At this stage, for each type of model, we sought to find the combination of explanatory variables that optimized the estimates by minimizing the sum of squares of the differences between the real and estimated values, known as Squared Error (SE, Eq. 1) (Hollander & Liu, 2008). Thus, for each type of regression, all the possibilities resulting from the combinations between the covariates selected in “Independent Variables” subsection were considered. The modeling step was performed in the R environment (R Core Team, 2020), an open and free programming interface, and in the GWR4.09 free software.

$$SE = \sum_{i=1}^n [y_i - y_i^*]^2 \quad (1)$$

Where  $y_i$  and  $y_i^*$  are the real and estimated values of the dependent variable in geographical position  $i$ ; and  $n$  is the number of bus stops. Initially, the traditional linear model was calibrated, whose structure is shown in Eq. 2 (Yan & Su, 2009).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon \quad (2)$$

Where the response variable  $y$  comprises the linear combination of explanatory variables  $x_k$  added to a random error  $\epsilon$ . The  $\beta$  parameters to be estimated are numbers that reflect the contribution of each covariate to explaining the variance of  $y$ . From the Ordinary Least Squares estimator, which, in the case of linear regression, coincides with the Maximum Likelihood estimator, the  $\beta$  coefficients can be obtained according to Eq. 3 (Yan & Su, 2009).

$$\beta = (X^t X)^{-1} X^t Y \quad (3)$$

Where  $X$  and  $Y$  are, respectively, the explanatory variable matrix and the vector of observations of the dependent variable. In R, the traditional linear regression was generated and optimized using the “olsrr” package (Hebbali, 2020). Then, the non-normal count data were analyzed using the Poisson regression, represented by Eq. 4 (Myers et al., 2010).

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (4)$$

Where  $\mu$  is the expected value of the response variable. The Poisson regression, unlike the linear one, admits that the variance of the information to be modeled is not constant, but that this variance varies as a function of  $\mu$  (Hilbe, 2014),

converging with the nature of the count data. Afterward, the isolated treatment of autocorrelation and spatial heterogeneity was addressed by the traditional GWR model (Eq. 5) (Brunsdon et al., 1996; Fotheringham et al., 2003).

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \epsilon_i \quad (5)$$

Where  $(u_i, v_i)$  represent the coordinates of the  $i$ -th point in space and  $\beta_k(u_i, v_i)$  refers to the realization of the continuous function  $\beta_k(u, v)$  at point  $i$  (Fotheringham et al., 2003). In the case of GWR, the spatial interaction between the point at which the model will be estimated and other points in the database is given by a weight that varies depending on the distance between these points and a maximum radius (bandwidth -  $b$ ) outside of which it is assumed zero spatial dependence. Equation 6 (Brunsdon et al., 1996) shows how  $\beta$  parameters are calculated in traditional GWR.

$$\beta_i = (X^t W_i X)^{-1} X^t W_i Y \quad (6)$$

Where  $W_i$  refers to the weight assigned to the remaining points in the database at the time of the calibration of the geographically weighted model in point  $i$ . Finally, the local spatial model that also considers the non-normal count data is structured in Eq. 7 (da Silva & Rodrigues, 2014; Nakaya et al., 2005).

$$\ln(\mu_i) = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} \quad (7)$$

As in the global model, two probability distributions for the response variable are allowed: Poisson and Negative Binomial. Within the scope of GWR, GWPR and GWNBR, the model can be optimized by selecting the weighting function (kernel) and respective bandwidth that minimize the Akaike Information Criterion (AIC) (Sakamoto et al., 1986) of the regression or a Cross-Validation (CV) metric. Based on this, in a simplified preliminary analysis, the Gaussian and bi-square kernels were analyzed, both with adaptive distance. The second was the one that showed the lowest AIC values and, consequently, comprised all the geographically weighted models. In turn, the adaptive bandwidth was chosen over the fixed one because it allows both points located in a region with a high density of bus stops and those located in areas with a lack of bus stops to receive the same amount of data when the model is calibrated. In this case,  $b$  corresponds to the distance between each bus stop where the model will be estimated and the most distant neighbor to be considered in the calibration, that is, in areas with a high density of points,  $b$  will be small, whereas regions with a lack of bus stops will receive a greater bandwidth. Thus, for each of the possible Boarding and Alighting models, two different bandwidths were obtained: the first minimizing the CV criterion, which is based on the Squared Error; and the second, minimizing the AIC. Afterward, the model was generated from these two optimal bandwidths and the bi-square kernel, structured in Eq. 8 (Fotheringham et al., 2003).

$$W_{ij} = \begin{cases} \left[ 1 - \left( \frac{d_{ij}}{b} \right)^2 \right]^2 & \text{if } d_{ij} < b \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Where  $W_{ij}$  refers to the weight assigned to point  $j$  at the time of calibration of the model in  $i$ ;  $d_{ij}$  is the distance between points  $i$  and  $j$ ; and  $b$  is the optimal bandwidth. Finally, we selected the model whose combination of covariates and bandwidth resulted in the smallest SE. This procedure was carried out according to codes available in the “sp” packages (Bivand et al., 2013; Pebesma & Bivand, 2005) and “GWModel” (Gollini et al., 2015; Lu et al., 2014) of R.

The last model to be applied to the database refers to Universal Kriging (UK). As this technique is not commonly used to address spatially discrete variables, the following subsection brings a more detailed discussion about it.

## Universal Kriging

Universal Kriging is one of the spatial interpolators from Geostatistics, a tool that deals with spatial autocorrelation using a probabilistic approach of regionalized variables (Matheron, 1971). Inspired by the work of Krige (1951) Geostatistics was first created to model spatially continuous variables, that is, variables that can assume a value at each geographic coordinate within the field in which they occur. As it is impossible to collect the real value of these variables throughout the whole spatial field, geostatistical interpolators seek to use the maximum information from collected samples to generate a continuous surface of estimated values covering both sampled and non-sampled points. Based on a probabilistic approach, geostatistical interpolators are unbiased and with minimum variance, providing uncertainty measures as well (variance of estimate), features not present in deterministic interpolators. Because of the convenience of Geostatistics to estimate in non-sampled locations, studies addressing spatially discrete variables started to apply geostatistical interpolators to overcome the lack of data caused by obstacles in the field collection (cost, access, topography). In this context, applications can be found in epidemiology, aquiculture, agriculture, forest sciences (Carvalho et al., 2015; Goovaerts, 2009; Kerry et al., 2016; Stelzenmüller et al., 2005), and, more recently, in the transportation engineering area, including accidents/road safety and travel demand modeling (Gomes et al., 2018; Klatko et al., 2017; Majumdar et al., 2004; Marques & Pitombo, 2021a, b; Pinto et al., 2020; Selby & Kockelman, 2013; Wang & Kockelman, 2009; Yang et al., 2018).

The bibliographic review by Marques and Pitombo (2020) highlighted the significant contributions from Geostatistics to various studies involving travel demand variables, which are usually spatially discrete. Research addressing the modal choice in the context of households/individuals (Chica-Olmo et al., 2018; Pitombo et al., 2015), trip generation in Traffic Analysis Zones (Lindner et al., 2016), traffic volume in road segments (Selby & Kockelman, 2013; Yang et al., 2018) and boardings and alightings at stations or bus stops (Marques & Pitombo, 2021a, b; Zhang & Wang, 2014) can be found. Most methods (field surveys, automatic counters, sensors etc.)

that support the exhaustive collection of this information require high financial resources, which may not be available for emerging countries like Brazil.

Unlike some geostatistical models that depend only on the variable of interest, Universal Kriging allows the inclusion of external explanatory variables. According to Fotheringham et al. (2003), it fits into the group of spatial regressions, however, unlike the SLM and SEM models, the spatial interaction between bus stops in the database, in the case of kriging, occurs in terms of the semivariogram function (Eq. 9) (Matheron, 1971; Cressie, 1993; Goovaerts, 1997).

$$\gamma(h) = \frac{1}{2N} \sum_{i=1}^N [Z(x_i + h) - Z(x_i)]^2 \quad (9)$$

In this case,  $Z(x_i)$  expresses the residual between the real and predicted values at point  $i$ ; and  $N$  is equivalent to the number of pairs located at a distance  $h$ . If the residuals show spatial autocorrelation, their values will be similar to each other at close bus stops in space and less similar as the distance between the bus stops increases. Thus, the semivariogram function graph presents an increasing form, from the origin or in its neighborhood, until reaching a sill, which refers to the maximum possible difference between the residuals and occurs at a distance beyond which there is no more spatial dependence between the database points.

The UK structure is similar to that of linear regression (Eq. 2), that is, the estimates are calculated both through the linear combination parameters of explanatory variables, known as large-scale variation, and the theoretical semivariogram model, which reflects the short-range variation (spatial dependence) and is part of the kriging error term (Cressie, 1993). Regarding the theoretical semivariogram, the adjustment of three models typically used was tested: exponential (*exp*), Gaussian (*gau*) and spherical (*sph*) (Olea, 2006). Using the restricted maximum likelihood estimator, Universal Kriging estimates are given by Eq. 10 (Cressie, 1993; Selby & Kockelman, 2013; Zhang & Wang, 2014).

$$y^*(x_0) = [X_o][\beta] + [V_{s_0}^T][V_s^{-1}][\varepsilon] \quad (10)$$

Where  $X_o$  is the matrix of explanatory variable observations of point  $x_0$ ;  $\beta$  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 ( $\gamma$ ) functions are related according to Eq. 11, where  $c_o$  and  $c_1$  stand out, respectively, for the nugget effect and partial sill parameters from the theoretical semivariogram.

$$V(h) = c_0 + c_1 - \gamma(h) \quad (11)$$

UK estimates were calculated in R using the “georob” package (Papritz, 2020a, b).

Although the explanatory variables used in the modeling stage showed a good correlation with Boardings and Alightings, not all of them had statistically

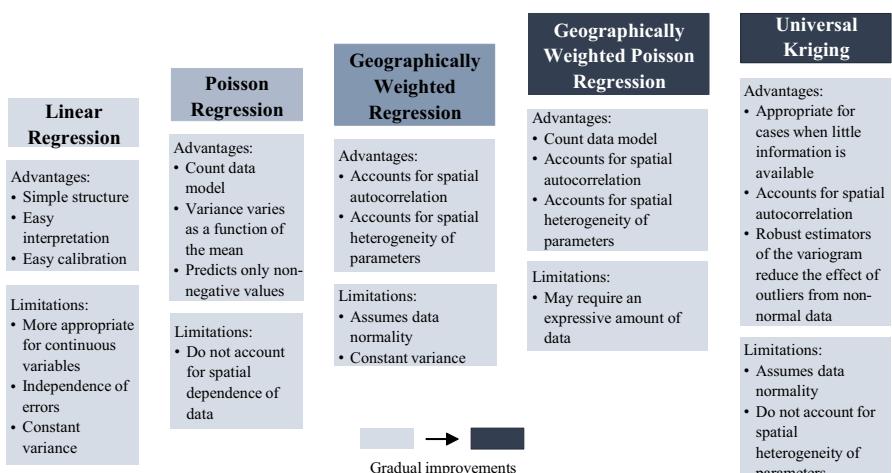
significant parameters in all the models in which they participated. Thus, in the case of global models (linear and Poisson regressions), it was established that, in addition to presenting the lowest Squared Error among the models analyzed in each category, the model with the best performance should also contain only variables whose parameters were statistically significant for a level of at least 90% confidence interval ( $p < 0.10$ ).

Figure 2 illustrates the modeling structure adopted in this article, from a simpler to a more complex approach. The figure summarizes the formulations previously described, illustrating the disadvantages and advantages in each stage of the sequence of models tested here.

The comparison between the best models in each category was performed using various goodness-of-fit measures, namely: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) (Hollander & Liu, 2008) and percentage of error, which must be close to 0 to reflect a good performance of the technique. To verify the best fit of local models over the global ones, the Akaike weight (Fotheringham et al., 2003) was calculated for the following pairs of models: (1) GWR and Linear Regression; and (2) GWPR and Poisson regression. Based on the AIC, which helps to choose the most parsimonious model from a set of competing models, the Akaike weight ( $w$ ) for model  $i$  is given by Eq. 12.

$$w_i = \frac{\exp(-AIC_i/2)}{\sum_j \exp(-AIC_j/2)} \quad (12)$$

As the Akaike weights of models being compared sum to 1, this measure represents the likelihood that each model is the best. So, the greater the weight, the greater the probability of the respective model being the best (Fotheringham et al., 2003). The results and discussion about these points are described in “Results and Discussion” section.



**Fig. 2** Comparison of models focusing on count and spatially dependent data

## Results and Discussion

Table 4 consolidates the descriptive measures of the data used in the present study. Figure 3, in turn, shows the spatial variation of the variables of interest, and of population, income, land use and stations around the 6045-10 bus line. From the linear correlation analysis, five predictors went to the Boarding modeling stage: population; commerce and service area (com serv area); distance to the nearest train or metro station (station distance); distance to the nearest bus terminal, train or metro station (intra/intermodal dist), replacing the previous variable; and average family income (income). For Alightings, the following predictors were selected: population; average frequency (frequency); distance to the nearest train or metro station (station distance); and average family income (income). For the sake of brevity, only the explanatory variables that were maintained in the final models of each of the categories described in the previous section are shown, as well as the dependent variables.

Despite the effort to collect the other variables, many pairs of potential predictors showed a statistically significant ( $p < 0.10$ ) Pearson coefficient correlation greater than 0.60. Bearing in mind that, in the presence of multicollinearity, the addition of more covariates does not significantly improve the performance of the model but can lead to misunderstandings in the value of the parameters, several covariates of Table 3 were discarded. In addition, adding more information to the modeling can lead to high costs due to data collection, making it difficult to apply the equations. However, even though several predictors were discarded, the set of variables chosen has both data related to potential demand and supply, that is, information regarding land use, socioeconomic features and the transport system around bus stops.

It is observed that both dependent variables demonstrate the positive asymmetry commented in “Materials and Method” section: their median is less than the mean and, in the case of Alightings, this difference is even more substantial. The null number of users boarding and alighting occurs only once in the set of trips made from 5 a.m. to 11.59 p.m.; at the last bus stop for Boardings, and at the first one, for Alightings, as expected.

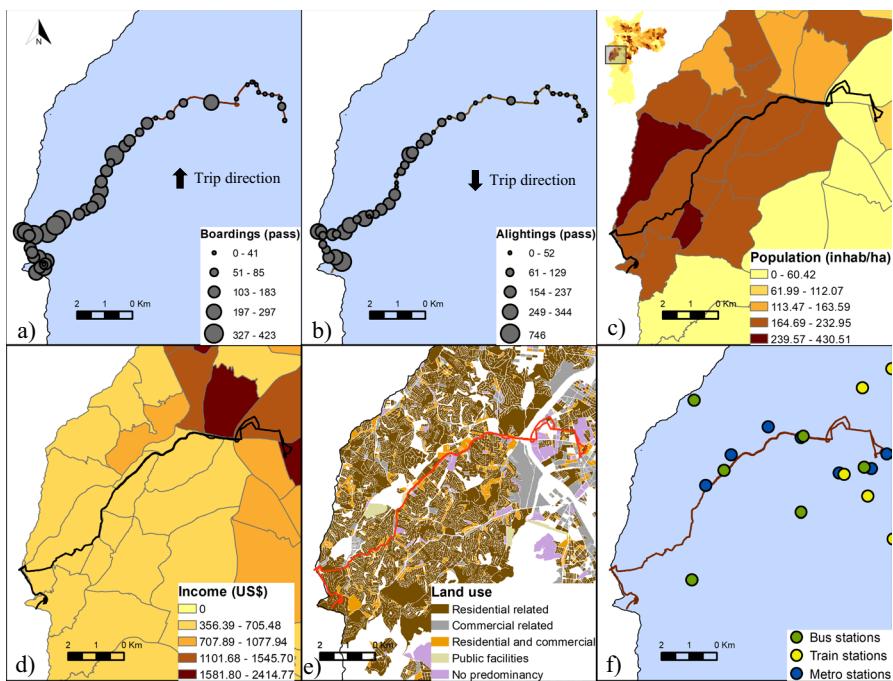
Moran's I results for Boardings and Alightings were 0.34 and 0.26, respectively. Both of them had associated  $p$ -value equal to 0. The spatial autocorrelation of Boardings and Alightings is illustrated by Fig. 3, which reveals that most passengers enter the 6045-10-1 line at its first bus stops, in the southwest region of the map. However, there are other peaks along the route until it reaches its last bus stops, in the northeast portion of the map. The inverse direction (6045-10-2 line) shows the opposite, as the number of passengers alighting is low in its first stops and starts to increase as the line runs along its route.

Despite there being some spatial correlation between the two variables of interest, the authors decided to perform the modeling separately as a way to compensate for the small number of bus stops with Boarding and Alighting data available. Therefore, it would be possible to verify the consistence of the models' results. In addition, as the 6045-10-1 and 6045-10-2 lines share only one bus stop, adding Boardings and Alightings was not an option.

**Table 4** Descriptive measures of the variables used

Direction	Variable\Descriptive	Mean	Std. Dev.	Min.	Max.	25%	50%	75%
Inbound (47)	<i>Boarding</i>	143.87	121.70	0.00	423.00	35.00	112.00	239.00
	Population (inhab.)	3,592.37	2,486.54	250.06	9,062.22	1,197.82	3,134.31	5,795.36
	Commerce and service area (com serv area)	0.52	1.18	0.00	6.22	0.00	0.00	0.61
	Station distance (m)	2,384.82	1,827.63	55.14	5,272.01	664.56	1,535.62	4,455.66
	Intra/intermodal dist (m)	1,437.26	819.36	55.14	2,842.85	664.56	1,376.62	2,192.67
	Average household income (income, BRL*)	4,371.29	2,725.12	1,232.85	14,500.00	2,452.84	3,431.91	5,630.27
Outbound (49)	<i>Alighting</i>	118.92	132.57	0.00	746.00	27.50	76.00	180.50
	Population (inhab.)	3,522.00	2,144.84	490.93	8,510.98	1,587.71	3,211.40	4,993.73
	Average frequency (frequency, trips/h)	4.27	0.79	2.30	5.82	3.77	4.00	5.26
	Station distance (m)	1,943.43	1,605.61	129.33	4,882.16	546.97	1,307.78	3,337.53
	Average household income (income, BRL*)	4,215.28	2,078.05	1,976.34	9,669.07	2,479.38	3,551.73	5,691.72

\* BRL 1.00 is equivalent to USD 0.18 (Feb. 2021)



**Fig. 3** Maps of **a** Boardings along 6045-10-1 line; **b** Alightings along 6045-10-2 line; **c** Population density at the TAZ level; **d** Average household income at the TAZ level; **e** Predominant land use at the block level; and **f** Bus, train and metro stations in the vicinity of the lines of interest

Figure 3c shows that the case study lines are situated in a densely populated area in the southwest region of São Paulo, whose main center also corresponds to the city's geographic center. This area is characterized by households with low-to-medium income (Fig. 3d), however high-income households are present at the end of the inbound trip (Boardings) and the beginning of the outbound trip (Alightings). In the case of the commerce and service area variable, which explained Boardings only, there is a preponderance of null values in its distribution of approximately 60%. In fact, as shown in Fig. 3e, the 6045-10 line runs through a predominantly residential area, with a few blocks of commercial or residential and commercial related use.

The following subtopics detail the results of Boarding and Alighting modeling. As defined in “Materials and Method” section, for each type of model, all possible combinations of covariates were considered, in order to find the set of predictors that generated the smallest Squared Error. Thus, for Boardings, there were 23 possible models in each category; and for Alightings, there were 15. The 8 surplus Boarding models refer to cases in which the station distance variable was replaced by the intra/intermodal dist variable, which showed a better performance in some situations. Since, in the case of geographically weighted models, the bandwidth can be optimized in two different ways, within the scope of GWR and GWPR, the number of possible models was twice that of the other categories.

## Boardings

Table 5 shows the results of the sequence of calibrated models for the Boarding dependent variable. In a preliminary analysis, the Negative Binomial regression, which can model the overdispersion phenomenon, that is, when the variance of the data exceeds its mean, was also considered. However, their results were worse than those of the Poisson regression, both in the case of Boarding and Alighting. Thus, only the Poisson model, among the models for count data, will be shown in the present study. Bearing in mind that, in the generalized global model analysis, the Poisson regression performed better than the Negative Binomial, we did not use GWNBR in the modeling stage, which is why this model does not appear in the results.

The linear regression model for Boardings is

$$\text{Boardings} = -47.43 + 0.02 * \text{Population} + 34.17 * \text{Comservarea} + 0.04 * \text{Stationdistance}.$$

The sign obtained for the predictors' coefficients, positive in the three cases, reveals that the greater the number of inhabitants and the provision of commerce and services around the bus stops, the greater the number of Boardings at them. For example, for each 1 new hectare of commerce and services area, the Boardings volume is likely to increase by 34 passengers, if the other attributes are held constant. In the case of population, an increase of one passenger boarding is expected to occur only if the number of inhabitants increases by  $1/0.02 = 50$ , *ceteris paribus*. Recall that the set of trips embedded in the dependent variable covers a typical full day (from 5 a.m. to 11.59 p.m.), therefore users who may have jobs along the 6045-10-1 line may have used any of its 47 bus stops to return home at the end of the day. In addition, considering that this line departs from a distant region of train and metro stations, slowly

**Table 5** Global and local models for Boardings along the 6045-10-1 line ( $N=47$ )

Model/Predictor	Intercept	Population	Com serv area	Station distance	Intra/inter-modal dist	Income
Linear regression	-47.43522	0.02357	34.17471	0.03721		
Poisson regression	3.55100	0.00021	0.07332		0.00058	-0.00010
GWR ( $N^* = 19$ )	Min -5.95226	-0.02893	-33.92728			
	25% -2.46590	0.01392	-31.05199			
	50% 68.28957	0.02083	7.94510			
	75% 105.03428	0.04605	46.72116			
	Max 278.22104	0.05380	89.67780			
GWPR ( $N^* = 19$ )	Min 1.93027	-0.00013	-0.72354		-0.00236	-0.00018
	25% 2.51549	0.00012	-0.56295		-0.00027	-0.00016
	50% 3.22689	0.00029	0.01206		0.00011	-0.00003
	75% 4.15367	0.00034	0.14317		0.00103	0.00001
	Max 5.61469	0.00036	0.56053		0.00120	0.00016
UK	-112.90000	0.03755		0.05203		

\* Number of neighbors corresponding to the optimal bandwidth. In local models GWR and GWPR, the bus stop where the model is calibrated does not participate in the calibration. GWR: Geographically Weighted Regression; GWPR: Geographically Weighted Poisson Regression; UK: Universal Kriging

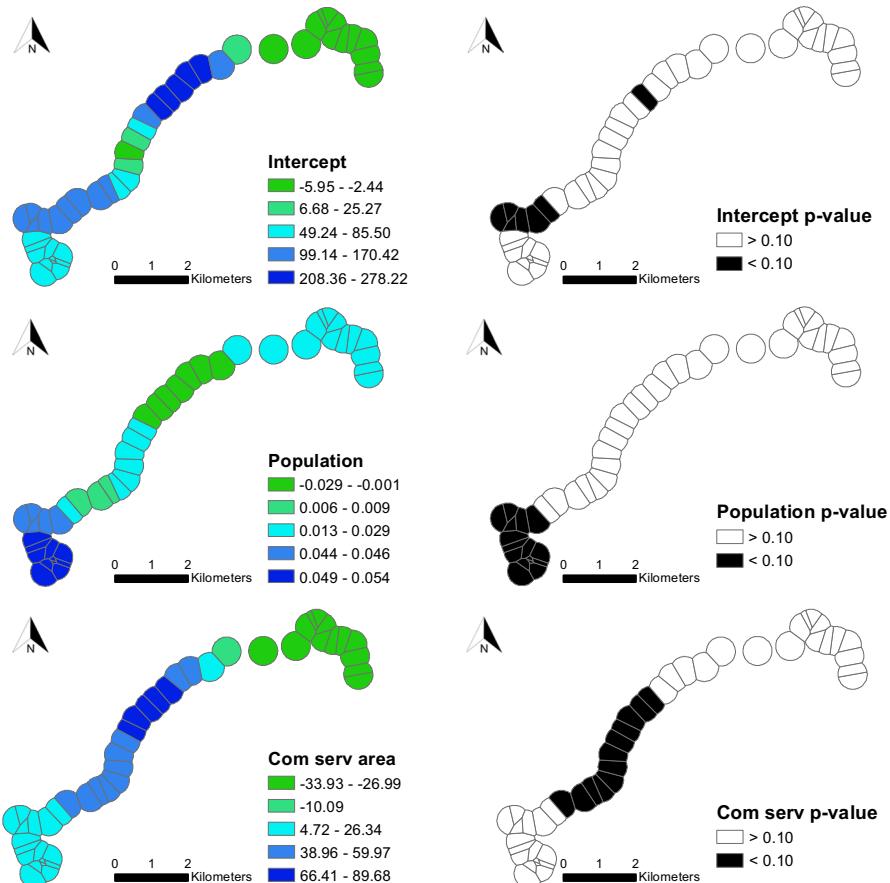
approaching some of them as it travels the route, the sign of the third explanatory variable is also plausible, that is, the largest volumes of Boardings are observed in the most distant areas of the central regions, and it is in these environments where the train and metro stations are usually located.

The resulting model for Poisson regression is

$$\text{Boardings} = \exp(3.55 + 0.00021 * \text{Population} + 0.07332 * \text{Comservarea} + 0.00058 * \text{Intra/interdistance} - 0.00010 * \text{income})$$

Income, as expected, appears with a negative sign, that is, bus stops with areas of influence characterized by a population with lower income, tend to generate a greater number of Boardings than those located near high-income regions. Therefore, if the household income increases by BRL 100.00 (USD 18.00 (Feb. 2021)) and the other attributes are held constant, the associated decrease in the number of Boardings is of  $[\exp(-0.00010 * 100) - 1] = 1.00\%$ .

The maps in Fig. 4 show the spatial variation of the estimated parameters of the GWR and respective *p*-values. It can be seen that the parameters of certain bus stops were negative. The resulting conclusion would be that the larger the population around these bus stops, the lower the number of Boardings. Probably, increasing



**Fig. 4** Estimated Boarding GWR parameters and respective statistical significance

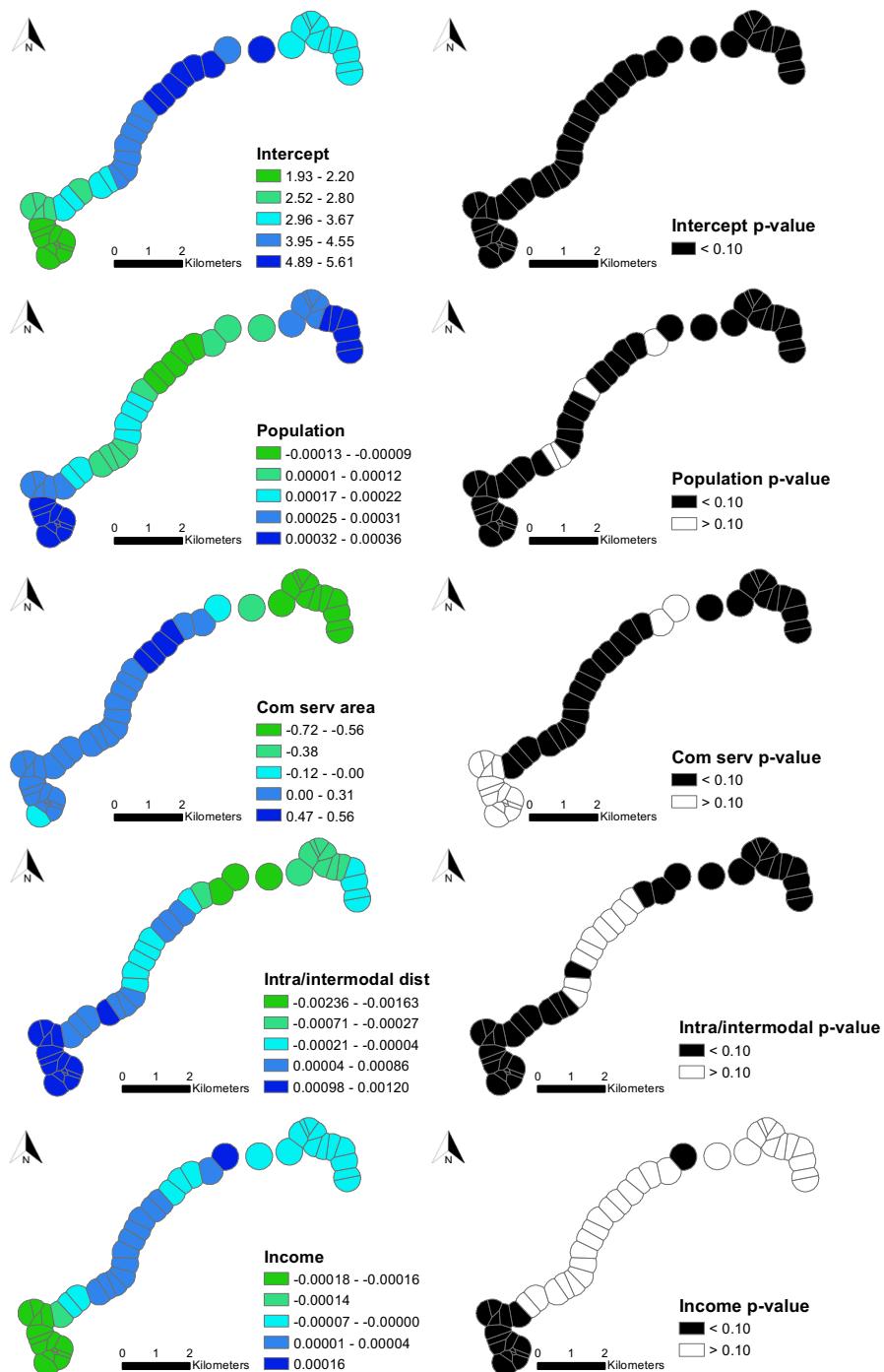
the population in these regions would stimulate the use of other travel modes. The negative sign may also reflect the fact that, for these bus stops, the model lacks other explanatory variables. However, both in the case of population and area of commerce and services, all negative parameters were not statistically significant, which means that these variables probably do not contribute to explaining the variation of Boardings at these bus stops. The local  $R^2$  ranged from 0.569 to 0.972, with 70% of the models calibrated for each bus stop showing a determination coefficient greater than 0.700.

The parameters estimated in the Boarding GWPR and respective  $p$  values are shown in Fig. 5. When comparing the map of population coefficients and Fig. 3c (population density), it can be seen that negative signs of population occur in densely populated regions but with a relatively small Boardings volume. On the other hand, the first and last bus stops of the inbound trip correspond to high and low density areas, respectively, showing proportionately bigger and smaller values of Boardings, which justifies the positive coefficient of the population for these bus stops. In addition, while the Poisson regression indicates an increase of only 7.61% in the number of Boardings if the commerce and services area increases by 1 hectare, there are points with an associated increase ranging from 60 to 70% in GWPR (the last category in Fig. 5), if the other attributes are held constant.

In the case of the intra/intermodal distance variable, the negative coefficients can be explained as follows: some bus stops located very far from bus terminals, metro or train stations may have their Boarding volume negatively impacted as they are unable to serve as elements of intra and intermodal integration. Stronger positive impacts of the proximity to stations can be seen at bus stops situated near the end of the inbound trip, densely supplied by stations (Fig. 3f), and where there is a peak in the Boardings volume. In the case of income, positive signs indicate bus stops with a Boardings volume proportional to the surrounding average income, while negative effects can be seen in areas with high Boardings volume, but low income, and in areas with low Boardings volume, but high income. Statistically significant positive coefficients belong to bus stops surrounded by low-to-medium income areas, and where an intermediate Boardings volume occurs, as shown in Fig. 3.

Figure 5 shows that there is a much larger number of bus stops with statistically significant parameters in GWPR compared to GWR. Bearing in mind that, in the calibration of both models, the data used was the same, this result may suggest that GWPR is more suitable for Boarding modeling at the bus stop level than GWR. Based on the number of bus stops whose parameters were statistically significant ( $p < 0.10$ ), the covariates can be ranked by degree of importance as follows: population, intra/intermodal distance, area of commerce and services, and income. Another important observation is that all explanatory variables have bus stops with statistically significant positive and negative coefficients, which corroborates the spatial heterogeneity of the parameters estimated in the stop-level Transit Ridership modeling.

Afterward, the UK model is presented. Note that the UK with the lowest SE retained only two explanatory variables: population and station distance, both with expected signs. This result emerges from the formulation of this regression itself: comprising a linear combination of predictors and  $\beta$  coefficients to be estimated, UK



**Fig. 5** Estimated Boarding GWPR parameters and respective statistical significance

assumes that the spatial autocorrelation of the database is present in the residuals of the model. Thus, the best fit for the UK occurs in cases where the explanatory variables are able to clearly discriminate this spatial dependence on residuals, and the inclusion of too many predictors may compromise this function.

## Alightings

Table 6 consolidates the models calibrated for Alightings.

The optimal linear regression for Alightings contained, in a similar way to that of Boardings, three explanatory variables, two of which exactly coincide with those of the previous model: population and intermodal distance. While the intensity of the distance effect remains similar to the case of Boardings, the impact of the population variable is greater regarding Alightings. If the population increases by 50 inhabitants, the number of Alightings is expected to increase by 1.5, *ceteris paribus*.

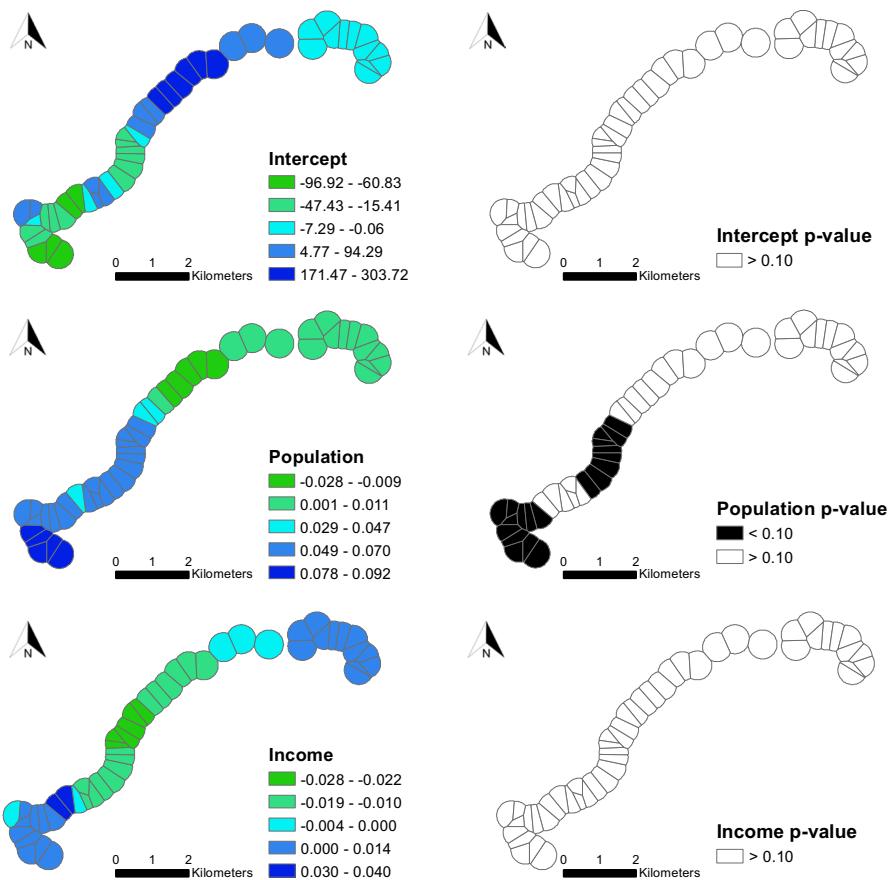
The negative sign of the average frequency variable reveals that regions with a dense coverage of the PT network present a volume of passengers alighting less than areas less supplied by the system. This conclusion shows that most trips on the 6045-10-2 line are attracted to places with less accessibility to PT than in the central regions. This destination may refer to the household of PT users, indicating that, probably, the return line 6045-10-2 serves a considerable portion of work-home trips. Assuming that on line 6045-10-2, return trips from work prevail, it can be stated that the sign of the income coefficient in the Poisson regression is also consistent with that expected.

The spatial variability of the parameters estimated in GWR and GWPR, together with their statistical significance, is shown in Figs. 6 and 7, respectively. The local  $R^2$  for GWR ranges from 0.265 to 0.996, in which 70% of the bus stops have an  $R^2$  value above 0.600.

Following the same pattern of Boardings, GWPR also maintained the same predictors that appeared in the final Poisson regression. As both 6045-10-1 and

**Table 6** Global and local models for Alightings along the 6045-10-2 line ( $N=49$ )

Model\Predictor	Intercept	Population	Frequency	Station distance	Income
Linear regression	228.42797	0.03072	-64.70247	0.03009	
Poisson regression	5.33800	0.00019	-0.37750	0.00020	-0.00007
GWR ( $N=17$ )	Min	-96.91727	-0.02829		-0.02806
	25%	-22.90870	0.01017		-0.01540
	50%	-3.88202	0.04913		-0.00006
	75%	13.60073	0.06111		0.00055
	Max	303.72456	0.09157		0.04025
GWPR ( $N=20$ )	Min	1.17558	-0.00016	-0.49110	-0.00023
	25%	1.55685	0.00027	-0.19205	-0.00023
	50%	2.79466	0.00030	-0.00026	-0.00003
	75%	4.39613	0.00034	0.15553	0.00030
	Max	7.68902	0.00058	0.35559	0.00035
UK		-159.20000	0.04209		0.06845

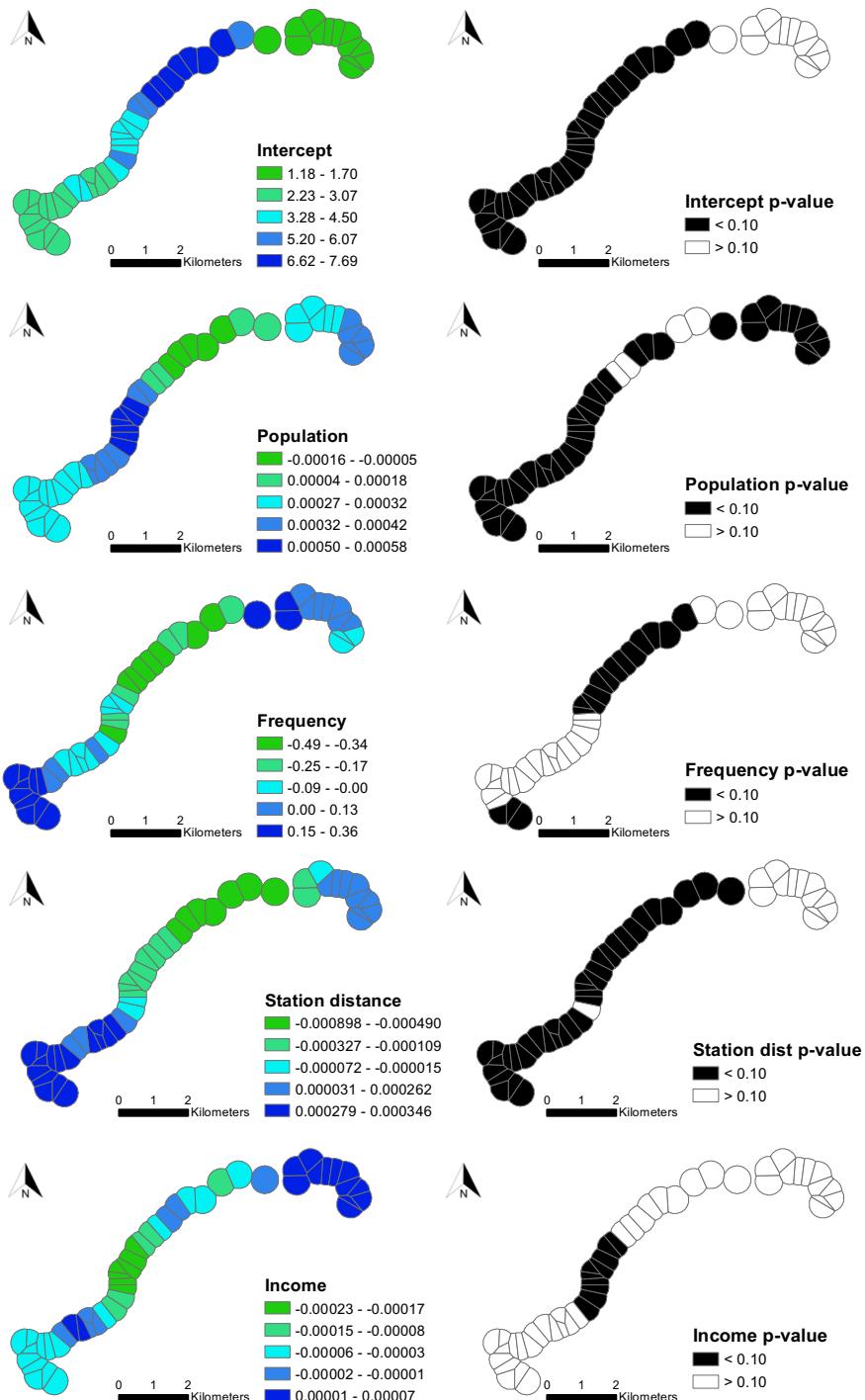


**Fig. 6** Estimated Alighting GWR parameters and respective statistical significance

6045-10-2 lines have itineraries close to each other in space, similar relationships between Boardings and Alightings and their predictors are expected.

The expected effects of the frequency variable on Alightings vary from  $-38.80\%$  to  $+42.70\%$  if the frequency increases by 1 trip/hour and the other attributes are held constant. This impact is only  $-31.44\%$  in the global Poisson model. Bus stops whose average frequency of the other lines that pass through them positively impacts the volume of Alightings possibly serve as intramodal integration nodes.

The  $p$  values found suggest the following classification of the degree of importance of the parameters to explain Alightings: population, intermodal distance, frequency and average household income. An interesting result is that the two most important explanatory variables for Boardings and Alightings were the same in GWPR: population and distance to the nearest station, or to the nearest station or bus terminal. It is important to note that population is part of the group of independent variables of potential demand, and intermodal or intra/intermodal distance comprises the group of supply variables. Therefore, the local modeling that also



**Fig. 7** Estimated Alighting GWPR parameters and respective statistical significance

accounts for the asymmetry of the travel demand variables, contains explanatory variables with statistically significant parameters from both categories of predictors.

The best performing Alighting UK, in turn, presents the same explanatory variables as its Boarding counterpart, but with slight differences in the estimated coefficients. In both cases, the theoretical semivariogram with the best performance was also the same: the exponential model.

### Goodness-of-fit Comparison of All Models

Table 7 summarizes the results of the goodness-of-fit measures applied to the global and local models of Boardings and Alightings.

In general, the techniques can be ranked, from the weakest to the strongest performance, as follows: (1) traditional linear regression; (2) Poisson regression; (3) GWR; (4) GWPR; and (5) UK. Note that, in the case of Boardings, the Poisson regression was better than the linear regression only regarding the MAE. However, the classic linear model, which assumes continuous variables of interest, has the drawback of allowing the prediction of negative values for Boardings and Alightings, which does not occur in the Poisson regression.

Based on this and using the MAE results for Alightings, the advantages, that is, relative reductions in error arising from the incorporation of asymmetry and autocorrelation, in isolation and together, to the process of modeling, can be illustrated as follows: -17.11% in the Poisson regression; -40.86% in GWR; -42.58% in GWPR; and -92.27% in the UK with the best performance. In Boardings, the following sequence is verified: -1.14%, -27.50%, -38.02% and -92.41%, using, as a reference, in both cases, the absolute mean error of the linear regression.

Basically, the global models differ from the local ones in that, in the second type of regression, bus stops with equal values of the explanatory variables are unlikely to have an identical predicted value for Boardings and Alightings, since, in this case, the result also depends on the spatial arrangement of the bus stops. However, while GWR and GWPR are considered local models because they allow, among other conveniences, the discrimination of the spatial heterogeneity of the model parameters, the local character of UK comes from the semivariogram function, which presents advantages in the case of data that are difficult to acquire. Thus, when comparing the two best performing methods, it can be seen that GWPR contributes to the knowledge of the way in which the Transit Ridership in each region would respond locally to changes in land use and in the transport system, guiding the transit-oriented urban development. As shown in Tables 5 and 6, the range of variation of the parameters in the GWR and GWPR corroborates the existence of this spatial heterogeneity. UK, in turn, provides accurate estimates with a small amount of information.

When it refers to goodness-of-fit measures based on the log-likelihood, the AIC of local models was lower in comparison with global models. The Poisson and Linear Regression for Boardings had an AIC of 2,332 and 552, respectively, while the AIC for GWPR and GWR was, respectively, 773 and 520. For Alightings, the results showed the same pattern: AIC of 579 and 542 for Linear Regression and GWR,

**Table 7** Goodness-of-fit measures for global and local models of Boardings and Alightings

Route direction	Model	Predictor(s)\\Measure	SE	MAE	RMSE	MedAPE (%)
Inbound (Boarding)	linear	population + com serv area + station distance	283,382	55,045	77,649	34,796
	Poisson	population + com serv area + intra/intermodal dist + income	297,238	54,417	79,525	37,521
	GWR	population + com serv area	139,195	39,909	54,421	32,450
	GWPR	population + com serv area + intra/intermodal dist + income	110,371	34,115	48,459	28,020
	UK	population + station distance	1,602	4,180	5,839	3,204
	linear	population + frequency + station distance	315,091	62,126	80,190	71,623
Outbound (Alighting)	Poisson	population + frequency + station distance + income	234,029	51,496	69,109	50,383
	GWR	population + income	138,073	36,740	53,083	33,723
	GWPR	population + frequency + station distance + income	129,507	35,670	51,410	34,451
	UK	population + station distance	2,409	4,800	7,011	4,339

SE, MAE, RMSE and MedAPE refer to the squared error, absolute mean error, root mean squared error and median of the absolute error in percentage

respectively; and 2,186 and 992, respectively, for Poisson Regression and GWPR. As the linear and Poisson models come from different probability distributions, the AIC results must not be used to compare all models simultaneously, but they confirm once again the better fit of local models over their global counterparts. The Akaike Information Criteria for the UK Boardings and Alightings was, respectively, 554 and 580. Although these values are higher than those for LR, the AIC from LR does not take into account the semivariogram part of UK, which is nonlinear. Therefore, the comparison between UK and its non-spatial counterpart (LR) should be made by the measures shown in Table 7.

Table 8 displays the Akaike weights from the comparison between local, GWPR and GWR, and global models, PR and LR, respectively. Based on these weights of evidence, GWR and GWPR are certainly better options than their global counterpart.

### Comparison with Previous Studies

Table 9 summarizes the characteristics of the models for Boarding and Alighting at the bus stop level already developed, together with their respective goodness-of-fit measures. The models presented in the present study were also included, for comparison purposes.

Attention is drawn to the fact that most of the models are from the USA, with only one representative in the Netherlands. This is probably due to the difficulty of acquiring reliable data on the movement of passengers along bus lines, that is, Boarding and Alighting per bus stop. The traditional Boarding and Alighting counts survey, which supports the collection of such information, is quite expensive and few municipalities have resources for this purpose. Automatic passenger counters, which could replace Boarding and Alighting survey, have not yet been popularized, especially in less developed countries. An alternative would be to synchronize the smart card data with the GPS of the buses, however, even in this case, some assumptions would have to be made to estimate the Boarding and Alighting bus stops, which could end up affecting the accuracy of the results. Thus, the present research, by providing Boarding and Alighting models per bus stop in a developing country, contributes to knowledge of how the relations between land use and transit ridership on a bus stop level take place in these regions.

It can also be observed that the studies address, as a dependent variable, only the number of Boardings or the sum of Boardings and Alightings. Although it is not wrong to assume that there is some correlation between Boardings and Alightings,

**Table 8** Akaike weights

Model	Boardings Akaike weights	Alightings Akaike weights
LR	0.000	0.000
GWR	1.000	1.000
PR	0.000	0.000
GWPR	1.000	1.000

**Table 9** Features and results of ridership models at the bus stop level

Reference	Country	Dependent variable	Model	Number of bus stops	Number of predictors	Goodness-of-fit measures
Chu (2004)	USA	Boarding	Poisson	2,568	15	Log-likelihood = -18,072
Ryan and Frank (2009)	USA	Boarding + alighting (logarithm)	OLS (log-linear)	3,582	7	Adjusted $R^2$ = 0.328
Pulugurtha and Aguria (2012)	USA	Boarding	Negative binomial	3,582	8	Adjusted $R^2$ = 0.330
Dill et al. (2013)	USA	Boarding + alighting (logarithm)	OLS (log-link*)	2,857	12	Corrected quasi-likelihood = 4,431
Kerkman et al. (2015)	Netherlands	Boarding + alighting (logarithm)	OLS (log-linear)	7,214	29	Adjusted $R^2$ = 0.69
The authors	Brazil	Boarding	OLS	1,400	29	Adjusted $R^2$ = 0.62
			Poisson	350	29	Adjusted $R^2$ = 0.53
			GWR	1,232	18	Adjusted $R^2$ = 0.772
			GWPR	1,284	18	Adjusted $R^2$ = 0.762
			UK	47	3	Adjusted $R^2$ = 0.555
		Alighting	Poisson	47	4	Log-likelihood = -1,161
			GWR	49	3	Adjusted $R^2$ = 0.717
			GWPR	49	4	Pseudo $R^2$ = 0.847
			UK	47	2	AIC = 554,459
			OLS	49	3	Adjusted $R^2$ = 0.602
			Poisson	49	4	Log-likelihood = -1,088
			GWR	17	2	Adjusted $R^2$ = 0.760
			GWPR	20	4	Pseudo $R^2$ = 0.832
			UK	49	2	AIC = 580,242

\* Best model shown in the respective study

the present study shows that the variables that explain Boardings and Alightings can be different and, even those that are repeated in both cases, result in different coefficients. Thus, the effect of such variables on Boardings and Alightings may vary from case to case.

As described in “[Introduction and Background](#)” section, the studies found had not yet provided a spatial approach to Boardings and Alightings. Table 9 also shows that the number of bus stops used in previous studies is considerably greater than that of the present case study, which reveals the availability of variables of interest for almost all or the whole bus network in such cities. This coverage, however, is difficult in regions that have a lack of technology or resources for this purpose.

Regarding the number of predictors, on the other hand, the present study had an extensive set of possible explanatory variables. However, the multicollinearity analysis reduced this group to only four predictors, both in the case of Boardings and Alightings, which did not prevent us from achieving good results. In fact, as the available database has a small number of points (47 and 49), the inclusion of more predictor data into the modeling would cause the parameters from these predictors to have statistical significance issues ( $p$ -value  $> 0.10$ ), especially in the case of GWR and GWPR, as they use only part of the database for calibration. Because the main focus of the modeling was to predict well Boardings and Alightings, we decided to test all possible combinations of predictors (considering only those without or with low correlation between them) that could achieve the best performance in goodness-of-fit measures. Bearing in mind that each model has its own characteristics, the set of predictors was different for the five models compared. When it refers to the spatial models (GWR, GWPR and UK), for example, the group of predictors selected would be the one that highlights the spatial dependence remaining in the residuals of the model, which is an issue that can be found when a small number of specific predictors is used (in the present case study, the resulting set of predictors was not able to control the spatial dependence of Boardings/Alightings in the non-spatial models). Thus, following this method enabled us to address a problem faced by municipalities with a lack of data on travel demand and its intervening factors. However, even when more predictor data is included in the model, testing for spatial dependence on residuals of the non-spatial models must not be overlooked, and if autocorrelation is present, spatial/local models are preferred.

We also recognize that a fairer comparison between the five approaches would be possible only if all models had the same set of explanatory variables. However, the decision to improve the goodness-of-fit measures for each type of model, as a method to achieve the best boarding and alighting estimates, could not retain the restriction of the same predictors for all models. This analysis can be tested in future studies.

Finally, the UK results are surprising: using only two explanatory variables, the Boarding and Alighting UK generated estimates with the median absolute error of 3.20% and 4.34%, respectively (Table 7). The goodness-of-fit measures obtained in the present study indicate that, even though there is not a considerable number of predictors, it is possible to develop models with satisfactory prediction performance. Although it is recognized that several of the potential predictors shown in Table 3 influence the passenger demand, the excess of information embedded in the

model makes it difficult to use it to forecast the number of Boarding and Alighting in hypothetical and/or future scenarios or in other cities/regions, since, for this, all predictors would also need to be estimated for the same condition. In addition, transit ridership models with many explanatory variables are only possible when the number of bus stops considered is also large, otherwise problems arise in the statistical significance of the estimated parameters. Thus, the present study also contributes to Boarding and Alighting modeling in cases in which only a small number of bus stops have data on the variables of interest and the amount of data on land use and transport is scarce.

## Conclusions, Main Constraints and Final Recommendations

The aim of the present study was to assess the gains provided by addressing asymmetry and spatial autocorrelation of stop-level transit ridership in its modeling. Global and local models for continuous and discrete data were applied to the Boarding and Alighting variables along a bus line in the city of São Paulo, Brazil. The results showed that, in fact, there is a gradual improvement in estimates as the two peculiarities of transit ridership are accounted for by the modeling.

In this context, the following topics summarize the research contributions of the present study:

- The solidification and methodological advancement of Boarding and Alighting at the bus stop level, through a comparison of models that consider specific aspects of such variables: asymmetry and spatial autocorrelation.
- The methodological procedure accounts for the lack of data usually faced by developing countries. Even though only a few predictors are used, the proposed models were able to provide good ridership estimates.
- Spatial dependence plays an important role to improve goodness-of-fit measures of stop-level ridership modeling.
- The predictors' effects on Boarding and Alighting can significantly vary from one bus stop to another.

The proposed models (GWPR and UK, for instance) have potential applications to urban and bus network planning. Based on the results achieved, the following recommendations are highlighted:

- The decision on whether to use a local model (GWPR, for instance) or UK for ridership prediction may be a matter of availability of data or policy. Coefficients from local models can be used to guide urban planning towards increasing transit patronage. However, if the main objective is only to achieve accurate ridership predictions, UK may be preferred.
- Results suggest that population and station distance (proxy for accessibility) are important predictors for Boarding and Alighting and, as such, they should not be overlooked in a transit ridership modeling by either GWPR or UK.

- The proposed models can support the analysis of ridership change in future or hypothetical scenarios, based on variations in the predictor information. In addition, they can provide Boarding and Alighting estimates for bus stops that lack these data.
- When ridership estimates are required for an exhaustive number of bus stops, the predictor data can be interpolated by means of kriging (or any other method). Therefore, a continuous surface of estimated ridership values, covering all the bus stops, can be obtained from the spatial models.
- Boarding and Alighting estimates for all bus stops of a route will provide municipalities with sufficient information to carry out the bus fleet sizing, as well as the bus frequency.

The main constraints of the present study can be outlined as follows:

- Given the small sample available for performing the modeling, the results can hardly be generalized. However, the proposed method had the former intention of stimulating the use of spatial and local models in the bus stop context, making it possible for forthcoming studies with bigger Boarding and Alighting datasets to use them and contribute to strengthening the results achieved.
- The dependent variable covers only passengers entering or leaving each specific line. However, the desired scenario would be to have the sum of passengers who enter or leave all bus lines that pass through the sampled bus stops so we could use the models to predict the total ridership in any bus stop.
- Only one of the eight lines was used as a case study. However, the proposed method can be easily applied to the remaining lines as well, separately.

In order to stimulate the consolidation of the appropriate transit ridership modeling at the bus stop level, some topics may be recommended for future work, such as:

- Calculating the goodness-of-fit measures based on a validation sample apart from the calibration sample used in the present analysis. This procedure would enable us to verify if the techniques of better performance in the calibration would also stand out in the validation.
- To address the cases with more than one line, including the analysis of overlapping between lines.
- To test semiparametric geographically weighted models, which admit both predictors of fixed and spatially varying parameters.
- Bearing in mind that UK was the only geostatistical model used, future research could also benefit from the comparison between UK and another multivariate interpolator from Geostatistics, such as Cokriging.
- To address the boarding and alighting data from multiple time bands in a disaggregated way, using geographically weighted models for panel data and spatio-temporal Geostatistics. In this case, the temporal autocorrelation of travel demand could be accounted for by the modeling, together with the already addressed factors: asymmetry and spatial autocorrelation.

**Acknowledgements** The authors would like to thank SPTrans for the Boarding/Alighting survey data used in this study.

**Funding** This research was supported by the São Paulo Research Foundation (FAPESP – grant #2019/12054-4, Brazil) and the National Council of Technological and Scientific Development (CNPq – grant #304345/2019-9, Brazil).

## Declarations

**Conflict of Interest** None.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

Bao, J., Liu, P., Qin, X., & Zhou, H. (2018). Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data. *Accident Analysis & Prevention*, 120, 281–294. <https://doi.org/10.1016/j.aap.2018.08.014>

Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R*, 2nd ed. Springer. Available at: <https://asdar-book.org/>

Blainey, S., & Mulley, C. (2013). Using geographically weighted regression to forecast rail demand in the Sydney region. *Australasian Transport Research Forum*. Brisbane, Australia, 2–4 October 2013. Available at: [https://australiasiantransportresearchforum.org.au/wp-content/uploads/2022/03/2013\\_Blainey\\_Mulley.pdf](https://australiasiantransportresearchforum.org.au/wp-content/uploads/2022/03/2013_Blainey_Mulley.pdf). Accessed August 2022.

Blainey, S., & Preston, J. (2010). A geographically weighted regression based analysis of rail commuting around Cardiff, South Wales. *12th World Conference on Transport Research*. Lisbon, Portugal, 11–15 July 2010. Available at: [https://www.researchgate.net/profile/John-Preston-10/publication/229020242\\_A\\_geographically\\_weighted\\_regression\\_based\\_analysis\\_of\\_rail\\_commuting\\_around\\_Cardiff\\_South\\_Wales/links/00463525a825e632ab000000/A-geographically-weighted-regression-based-analysis-of-rail-commuting-around-Cardiff-South-Wales.pdf](https://www.researchgate.net/profile/John-Preston-10/publication/229020242_A_geographically_weighted_regression_based_analysis_of_rail_commuting_around_Cardiff_South_Wales/links/00463525a825e632ab000000/A-geographically-weighted-regression-based-analysis-of-rail-commuting-around-Cardiff-South-Wales.pdf). Accessed August 2022.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4), 281–298. <https://doi.org/10.1111/j.1538-4632.1996.tb00936.x>

Cardozo, O. D., García-Palomares, J. C., & Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, 34(Supplement C), 548–558. <https://doi.org/10.1016/j.apgeog.2012.01.005>

Carvalho, S. D. P. C., e, Rodriguez, L. C. E., Silva, L. D., de Carvalho, L. M. T., Calegario, N., de Lima, M. P., Silva, C. A., de Mendonça, A. R., & Nicoletti, M. F. (2015). Predição do volume de árvores integrando Lidar and Geoestatística. *Scientia Forestalis/Forest Sciences*, 43(107), 627–637.

Cervero, R. (2006). Alternative approaches to modeling the travel-demand impacts of smart growth. *Journal of the American Planning Association*, 72(3), 285–295. <https://doi.org/10.1080/01944360608976751>

Cervero, R., & Dai, D. (2014). BRT TOD: Leveraging transit oriented development with bus rapid transit investments. *Transport Policy*, 36, 127–138. <https://doi.org/10.1016/j.tranpol.2014.08.001>

Chica-Olmo, J., Rodríguez-López, C., & Chillón, P. (2018). Effect of distance from home to school and spatial dependence between homes on mode of commuting to school. *Journal of Transport Geography*, 72, 1–12. <https://doi.org/10.1016/j.jtrangeo.2018.07.013>

Chiou, Y. C., Jou, R. C., & Yang, C. H. (2015). Factors affecting public transportation usage rate: Geographically weighted regression. *Transportation Research Part A: Policy and Practice*, 78, 161–177. <https://doi.org/10.1016/j.tra.2015.05.016>

Choi, J., Lee, Y. J., Kim, T., & Sohn, K. (2012). An analysis of Metro ridership at the station-to-station level in Seoul. *Transportation*, 39(3), 705–722. <https://doi.org/10.1007/s11116-011-9368-3>

Chu, X. (2004). *Ridership models at the stop level*. National Center for Transit Research, University of South Florida. <https://doi.org/10.5038/CUTR-NCTR-RR-2002-10>

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

da Silva, A. R., & Rodrigues, T. C. V. (2014). Geographically Weighted Negative Binomial Regression—incorporating overdispersion. *Statistics and Computing*, 24(5), 769–783. <https://doi.org/10.1007/s11222-013-9401-9>

de Marques, S., & Pitombo, C. S. (2020). Intersecting geostatistics with transport demand modeling: A bibliographic survey. *Revista Brasileira de Cartografia*, 72, 1028–1050. <https://doi.org/10.14393/rbcv72nespecial50anos-56467>

de Marques, S. F., and, & Pitombo, C. S. (2021a). Applying multivariate geostatistics for transit ridership modeling at the bus stop level. *Boletim de Ciências Geodésicas*, 27(2). <https://doi.org/10.1590/1982-2170-2020-0069>

de Marques, S., & Pitombo, C. S. (2021b). Ridership estimation along bus transit lines based on kriging: Comparative analysis between network and Euclidean distances. *Journal of Geovisualization and Spatial Analysis*, 5(1), 7. <https://doi.org/10.1007/s41651-021-00075-w>

Dill, J., Schlossberg, M., Ma, L., & Meyer, C. (2013). Predicting transit ridership at stop level: Role of service and urban form. *92nd Annual Meeting of the Transportation Research Board*. Washington, USA, 13–17 January 2013. Available at: [https://nacto.org/wp-content/uploads/2016/04/1-3\\_Dill-Schlossberg-Ma-and-Meyer-Predicting-Transit-Ridership-At-The-Stop-Level\\_2013.pdf](https://nacto.org/wp-content/uploads/2016/04/1-3_Dill-Schlossberg-Ma-and-Meyer-Predicting-Transit-Ridership-At-The-Stop-Level_2013.pdf). Accessed August 2022

Ewing, R., Tian, G., Goates, J. P., Zhang, M., Greenwald, M. J., Joyce, A., Kircher, J., & Greene, W. (2014). Varying influences of the built environment on household travel in 15 diverse regions of the United States. *Urban Studies*, 52(13), 2330–2348. <https://doi.org/10.1177/0042098014560991>

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

Gan, Z., Feng, T., Yang, M., Timmermans, H., & Luo, J. (2019). Analysis of metro station ridership considering spatial heterogeneity. *Chinese Geographical Science*, 29(6), 1065–1077. <https://doi.org/10.1007/s11769-019-1065-8>

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2015). GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models. *Journal of Statistical Software*, 63(17), 1–50.

Gomes, M. J. T. L., Cunto, F., & da Silva, A. R. (2017). Geographically weighted negative binomial regression applied to zonal level safety performance models. *Accident Analysis & Prevention*, 106, 254–261. <https://doi.org/10.1016/j.aap.2017.06.011>

Gomes, M. M., Pirdavani, A., Brijs, T., & Pitombo, C. S. (2019). Assessing the impacts of enriched information on crash prediction performance. *Accident Analysis and Prevention*, 122, 162–171. <https://doi.org/10.1016/j.aap.2018.10.004>

Gomes, M. M., Pitombo, C. S., Pirdavani, A., & Brijs, T. (2018). Geostatistical approach to estimate car occupant fatalities in traffic accidents. *Revista Brasileira de Cartografia*, 70(4), 1231–1256.

Goovaerts, P. (1997). *Geostatistics for natural resources and evaluation*. Oxford University Press.

Goovaerts, P. (2009). Medical geography: A promising field of application for geostatistics. *Mathematical Geosciences*, 41, 243–264. <https://doi.org/10.1007/s11004-008-9211-3>

Gutiérrez, J., Cardozo, O. D., & 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), 1081–1092. <https://doi.org/10.1016/j.jtrangeo.2011.05.004>

Hebbali, A. (2020). olrr: *Tools for Building OLS Regression Models*. R package version 0.5.3. Available at: <https://CRAN.R-project.org/package=olrr>

Hensher, D. A., & Golob, T. F. (2008). Bus rapid transit systems: a comparative assessment. *Transportation*, 35(4), 501–518. <https://doi.org/10.1007/s11116-008-9163-y>

Hensher, D. A., Li, Z., & Mulley, C. (2014). Drivers of bus rapid transit systems – Influences on patronage and service frequency. *Research in Transportation Economics*, 48, 159–165. <https://doi.org/10.1016/j.retrec.2014.09.038>

Hilbe, J. M. (2014). *Modeling Count Data*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139236065>

Hollander, Y., & Liu, R. (2008). The principles of calibrating traffic microsimulation models. *Transportation*, 35(3), 347–362. <https://doi.org/10.1007/s11116-007-9156-2>

Joonho, K., Daejin, K., & Ali, E. (2019). Determinants of Bus Rapid Transit Ridership: System-Level Analysis. *Journal of Urban Planning and Development*, 145(2), 4019004. [https://doi.org/10.1061/\(ASCE\)UP.1943-5444.0000506](https://doi.org/10.1061/(ASCE)UP.1943-5444.0000506)

Kalaanidhi, S., & Gunasekaran, K. (2013). Estimation of bus transport ridership accounting accessibility. *Procedia - Social and Behavioral Sciences*, 104, 885–893. <https://doi.org/10.1016/j.sbspro.2013.11.183>

Kerkman, K., Martens, K., & Meurs, H. (2015). Factors influencing stop-level transit ridership in Arnhem–Nijmegen City Region, Netherlands. *Transportation Research Record*, 2537(1), 23–32. <https://doi.org/10.3141/2537-03>

Kerry, R., Goovaerts, P., Giménez, D., Oudemans, P., & Muñiz, E. (2016). Investigating geostatistical methods to model within-field yield variability of cranberries for potential management zones. *Precision Agriculture*, 17, 247–273. <https://doi.org/10.1007/s11119-015-9408-7>

Klatko, T. J., Usman, S. T., Matthew, V., & Samuel, L. (2017). Addressing the local-road VMT estimation problem using spatial interpolation techniques. *Journal of Transportation Engineering Part A: Systems*, 143(8), 4017038. <https://doi.org/10.1061/JTEPBS.0000064>

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), 119–139.

Kyte, M., Stoner, J., & Cryer, J. (1985). Development and application of time-series transit ridership models for Portland, Oregon. *Transportation Research Record*, 1036, 9–18. Available at: <http://onlin epububs.trb.org/Onlinelibrary/trb/1985/1036/1036-002.pdf>. Accessed August 2022.

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

Liu, J., Khattak, A. J., & Wali, B. (2017). Do safety performance functions used for predicting crash frequency vary across space? Applying geographically weighted regressions to account for spatial heterogeneity. *Accident Analysis & Prevention*, 109, 132–142. <https://doi.org/10.1016/j.aap.2017.10.012>

Liu, Y., Ji, Y., Shi, Z., & Gao, L. (2018). The influence of the built environment on school children's metro ridership: An exploration using geographically weighted poisson regression models. *Sustainability*, 10(12), 4684. <https://doi.org/10.3390/su10124684>

Lu, B., Harris, P., Charlton, M., & Brunsdon, C. (2014). The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-spatial Information Science*, 17(2), 85–101. <https://doi.org/10.1080/10095020.2014.917453>

Ma, X., Zhang, J., Ding, C., & Wang, Y. (2018). A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. *Computers Environment and Urban Systems*, 70, 113–124. <https://doi.org/10.1016/j.compenvurbsys.2018.03.001>

Majumdar, A., Noland, R. B., & Ochieng, W. Y. (2004). A spatial and temporal analysis of safety-belt usage and safety-belt laws. *Accident Analysis & Prevention*, 36(4), 551–560. [https://doi.org/10.1016/S0001-4575\(03\)00061-7](https://doi.org/10.1016/S0001-4575(03)00061-7)

Matheron, G. (1971). *The theory of regionalized variables and its applications*. Les Cahiers du Centre de Morphologie Mathématique in Fontainebleau.

Metrô (2019). Pesquisa de Origem e Destino de 2017 (Banco de dados). Companhia do Metropolitano De São Paulo, Secretaria Estadual dos Transportes Metropolitanos. Available at: <https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino>. Accessed August 2022.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society Series B (Methodological)*, 10(2), 243–251.

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2010). *Generalized linear models: with applications in engineering and the sciences* (2nd ed.). Wiley. <https://doi.org/10.1002/9780470556986>

Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine*, 24(17), 2695–2717. <https://doi.org/10.1002/sim.2129>

Obelheiro, M. R., da Silva, A. R., Nodari, C. T., Cybis, H. B. B., & Lindau, L. A. (2020). A new zone system to analyze the spatial relationships between the built environment and traffic safety. *Journal of Transport Geography*, 84, 102699. <https://doi.org/10.1016/j.jtrangeo.2020.102699>

Olea, R. A. (2006). A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research and Risk Assessment*, 20(5), 307–318. <https://doi.org/10.1007/s00477-005-0026-1>

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2), 289–290. <https://doi.org/10.1093/bioinformatics/btg412>

Papritz, A. (2020a). *georob: Robust Geostatistical Analysis of Spatial Data*. R package version 0.3–13. Available at: <https://CRAN.R-project.org/package=georob>

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](https://cran.r-project.org/web/packages/georob/vignettes/georob_vignette.pdf)

Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2). Available at: <https://cran.r-project.org/doc/Rnews/>

Peng, Z. R., Dueker, K. J., Strathman, J., & Hopper, J. (1997). A simultaneous route-level transit patronage model: demand, supply, and inter-route relationship. *Transportation*, 24(2), 159–181. <https://doi.org/10.1023/A:1017951902308>

Pinto, J. A., Kumar, P., Alonso, M. F., Andreão, W. L., Pedruzzi, R., Espinosa, S. I., & de Almeida Albuquerque, T. T. (2020). Kriging method application and traffic behavior profiles from local radar network database: A proposal to support traffic solutions and air pollution control strategies. *Sustainable Cities and Society*, 56, 102062. <https://doi.org/10.1016/j.scs.2020.102062>

Pitombo, C. S., Salgueiro, A. R., da Costa, A. S. G., & Isler, C. A. (2015). A two-step method for mode choice estimation with socioeconomic and spatial information. *Spatial Statistics*, 11, 45–64. <https://doi.org/10.1016/j.spasta.2014.12.002>

Profillidis, V. A., & Botzoris, G. N. (2019). Statistical methods for transport demand modeling. B. Romer (Ed). *Modeling of Transport Demand* (p. 163–224). Elsevier. <https://doi.org/10.1016/B978-0-12-811513-8.00005-4>

Pulugurtha, S. S., & Agurla, M. (2012). Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. *Journal of Public Transportation*, 15(1), 33–52.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>

Ryan, S., & Frank, L. (2009). Pedestrian Environments and Transit Ridership. *Journal of Public Transportation*, 12(1), 39–57. <https://doi.org/10.5038/2375-0901.12.1.3>

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*. D. Reidel Publishing Company.

Selby, B., & Kockelman, K. M. (2013). Spatial prediction of traffic levels in unmeasured locations: Applications of universal kriging and geographically weighted regression. *Journal of Transport Geography*, 29, 24–32. <https://doi.org/10.1016/j.jtrangeo.2012.12.009>

Siddiqui, S., Amirhossein, J., & Hossain, F. (2015). *Increasing Transit Ridership in Small Urban Areas: A case study of Streamline in Bozeman, MT*. <https://doi.org/10.13140/RG.2.1.3488.5847>

Song, Y., Merlin, L., & Rodriguez, D. (2013). Comparing measures of urban land use mix. *Computers Environment and Urban Systems*, 42, 1–13. <https://doi.org/10.1016/j.compenvurbsys.2013.08.001>

Stelzenmüller, V., Ehrlich, S., & Zauke, G. P. (2005). Impact of additional small-scale survey data on the geostatistical analyses of demersal fish species in the North Sea. *Scientia Marina*, 69(4), 587–602. <https://doi.org/10.3989/scimar.2005.69n4587>

Sun, L. S., Wang, S. W., Yao, L. Y., Rong, J., & Ma, J. M. (2016). Estimation of transit ridership based on spatial analysis and precise land use data. *Transportation Letters*, 8(3), 140–147. <https://doi.org/10.1179/1942787515Y.0000000017>

Taylor, B. D., Miller, D., Iseki, H., & Fink, C. (2009). Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. *Transportation Research Part A: Policy and Practice*, 43(1), 60–77. <https://doi.org/10.1016/j.tra.2008.06.007>

Tu, W., Cao, R., Yue, Y., Zhou, B., Li, Q., & Li, Q. (2018). Spatial variations in urban public ridership derived from GPS trajectories and smart card data. *Journal of Transport Geography*, 69, 45–57. <https://doi.org/10.1016/j.jtrangeo.2018.04.013>

Wang, X., & Kockelman, K. (2009). Forecasting network data. *Transportation Research Record: Journal of the Transportation Research Board*, 2105, 100–108. <https://doi.org/10.3141/2105-13>

Xu, C., Li, H., Zhao, J., Chen, J., & Wang, W. (2017). Investigating the relationship between jobs-housing balance and traffic safety. *Accident Analysis & Prevention*, 107, 126–136. <https://doi.org/10.1016/j.aap.2017.08.013>

Xu, P., & Huang, H. (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis & Prevention*, 75, 16–25. <https://doi.org/10.1016/j.aap.2014.10.020>

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

Yang, H., Yang, J., Han, L. D., Liu, X., Pu, L., Chin, S. M., & Hwang. (2018). A Kriging based spatiotemporal approach for traffic volume data imputation. *PLoS One*, 13(4), e0195957. <https://doi.org/10.1371/journal.pone.0195957>

Zhang, D., & Wang, X. C. (2014). Transit ridership estimation with network Kriging: A case study of Second Avenue Metro, NYC. *Journal of Transport Geography*, 41, 107–115. <https://doi.org/10.1016/j.jtrangeo.2014.08.021>

Zhao, F., Chow, L. F., Li, M. T., Ubaka, I., & Gan, A. (2003). Forecasting transit walk accessibility: Regression model alternative to Buffer Method. *Transportation Research Record*, 1835(1), 34–41. <https://doi.org/10.3141/1835-05>

Zhu, Y., Chen, F., Wang, Z., & Deng, J. (2019). Spatio-temporal analysis of rail station ridership determinants in the built environment. *Transportation*, 46(6), 2269–2289. <https://doi.org/10.1007/s11116-018-9928-x>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.