Wiley
Journal of Advanced Transportation
Volume 2024, Article ID 7967141, 27 pages
https://doi.org/10.1155/2024/7967141

WILEY

Research Article

Spatial Modeling of Travel Demand Accounting for Multicollinearity and Different Sampling Strategies: A Stop-Level Case Study

Samuel de França Marques[1], **Cira Souza Pitombo**[1], **and J. Jaime Gómez-Hernández**[2]

[1]*Department of Transportation Engineering, São Carlos School of Engineering, University of São Paulo, São Carlos, Brazil*
[2]*Institute of Water and Environmental Engineering, Universitat Politècnica de València, Valencia, Spain*

Correspondence should be addressed to Samuel de França Marques; samuelmarques@usp.br

Received 4 February 2024; Revised 7 April 2024; Accepted 22 June 2024

Academic Editor: Socrates Basbas

Stop-level ridership data serve as a basis for various studies toward increasing bus patronage and promoting sustainable land use planning. To address limitations found in previous studies, this study proposes a novel approach based on Geographically Weighted Principal Component Analysis (GWPCA) and Ordinary Kriging to predict the stop-level boarding or alighting data along bus lines in São Paulo (Brazil), considering four different sampling methods. The main contributions are as follows: by accounting for the spatial heterogeneity of the predictor dataset, the GWPCA can identify the most important factor affecting transit ridership even in bus stops with no information on boarding and alighting; the spatial modeling of stop-level ridership data using GWPCA components as explanatory variables allows visualizing the spatially varying effects from predictors on ridership, supporting the land use planning at a local level; GWPCA coupled with kriging simultaneously addresses the multicollinearity of predictor data, its spatial heterogeneity, and the spatial dependence of the stop-level ridership variable, thus enhancing the goodness-of-fit measures of the transit ridership prediction in unsampled stops; and a balanced sample on predictor data and well-spread in the geographic space might be preferred to accurately estimate missing stop-level ridership data. In addition to solve the lack of stop-level ridership data, supporting a reliable bus system planning, the proposed method indicates what predictors should be addressed by policymakers to stimulate a transit-oriented development. The method can be successfully applied to other travel demand variables facing a lack of data such as traffic volume in road segments and mode choice at the household level.

1. Introduction

Stop-level boarding and alighting data are important pieces of information for decisions regarding land use and bus network planning. Decisions on selecting the best location to place a new bus stop, which bus stop could be removed along a bus line and adjustments in the bus routes often rely on stop-level ridership data [1]. In addition, optimal fleet sizing can be achieved based on the route-segment-level loading information, which is obtained from boarding and alighting data [2]. Stop-level ridership data have also been used to analyze stops' level of service and sizing [3], as an exposure variable for crime research [4, 5], and to support decisions on where amenities, such as shelter, for example, should be installed [6].

However, previous studies reveal that municipalities often face limitations when collecting boarding and alighting data [7–9]. To solve this problem, authors have relied on various modeling approaches, but only a few of them assessed the model performance when predicting the ridership data in a nonsampled point [10–12]. Unlike some travel demand information, data on explanatory variables may be not so difficult to obtain, given the gradual advances in geographic information systems and the relatively easy access to it. However, in addition to the low representativity of missing data evaluations in stop-level research, these studies face another problem: predictor data multicollinearity. The potential presence of high-correlated independent variables has been a matter of concern in most ridership studies at the bus stop level [6–11, 13–18]. Detecting the existence of multicollinearity in predictor datasets has also been carried out in the context of road segments [19–21], traffic analysis zones [22, 23], rail stations [24–26], and pedestrians [27].

Multicollinearity is often disclosed by analyzing the variance inflation factor or the Pearson linear correlation coefficient. Given a specified threshold, one of the variables in a pair of high-correlated variables is eliminated from the model [10, 11, 14–17, 23, 26]. Maintaining pairs of correlated predictors can result in misleading interpretations of the estimated parameters. For example, Kerkman et al. [9] reported the effect from population in their case study to be underestimated probably because of a high correlation between population and residential areas. In turn, Mucci and Erhardt [28] found a potentially overestimated effect from frequency on the ridership, which could be due to the correlation between frequency and other predictors, such as employment. At the same time, excluding a predictor that has proven to affect the variable of interest may not be a wise solution. When dealing with a lack of stop-level ridership data, using all information available to predict boarding and alighting at an unsampled point is fundamental to achieve reliable estimates.

The main goal of this study is to perform the spatial modeling and prediction of a stop-level ridership variable, which has proven to be spatially dependent in previous studies [10, 12, 15, 16, 29], accounting for multicollinearity and the influence of the sampling strategy. As the predictor data may present multicollinearity and spatial variation of effects simultaneously, we propose a conjoint approach based on Geographically Weighted Principal Component Analysis (GWPCA) and Ordinary Kriging to improve the ridership prediction. To the best of our knowledge, this is the first paper applying GWPCA coupled with Ordinary Kriging. In the literature, we have already found the combination between the standard PCA and kriging [30], and GWPCA as a single model or combined with other techniques, such as clustering [31–36]. However, we have not yet found the combined approach between GWPCA and Ordinary Kriging.

This study has five sections. The bibliographic review that supported the main goal of the study is presented in Section 2. Section 3 provides a detailed description of the database used as a case study and the method steps applied. Results are discussed in Section 4. Section 5 summarizes the conclusions, some practical recommendations, limitations, and topics for future research.

2. Research Background

Among the methods for collecting stop-level ridership data, three can be cited: Automatic Passenger Counter (APC), smart cart data, and boarding and alighting count survey. Table 1 summarizes stop-level ridership studies found in the literature which have reported the collection method used. Some limitations described by the authors regarding the collection methods are also presented.

The collection method most available among published studies is the APC. Limitations regarding this technology refer mainly to the APC coverage and accuracy. As the APC is not commonly installed in all buses at the same time, authors have reported working with a sample of trips, with extrapolated data (not accounting for the spatial dependence of ridership data), or with data coming from a short period of days, in which the APCs were assigned to all bus routes. Regarding accuracy, the APC device is more efficient in counting alightings than boardings, as some passengers may bunch when entering the bus [13].

Smart card data, coupled with a Global Positioning System (GPS) in the vehicles, can provide information of interest at a lower cost. However, in this case, the accuracy problem is inverted. If the passengers do not tap the card when leaving the bus, assumptions have to be made to estimate the alighting stop. Moreover, users that do not have the card are not counted in either way.

Together with the smart card method, the boarding and alighting survey had only one representative among the cities used as a case study (São Paulo, Brazil). As the collection is performed manually, the accuracy may not be a problem in the case of the boarding and alighting survey. Conversely, the need for a qualified team of researchers, and the high cost and time required for performing the survey are the main problems faced by this type of collection. In São Paulo, only 8 lines out of more than 1 thousand routes were visited.

From the limitations faced by municipalities in gathering a comprehensive stop-level ridership dataset, the boarding and alighting modeling has been used as a solution to predict missing ridership data. However, only a few studies [11, 12, 28] have carried out a validation analysis, using a validation sample aside from the calibration one. Even when the research performs a missing data evaluation, it tests only one type of sampling approach, ignoring the effect that the selection of the calibration/validation samples may have on the models' prediction power. Table 2 summarizes a bibliographic review on validation analyses over several studies addressing the spatial modeling of a travel demand variable. Geographic units other than the bus stop were also included. Stop-level studies shown in Table 1 that have not performed a validation analysis are omitted in Table 2.

In general, a validation step is found only at the bus stop, road segment, and household levels. In the road segment case, the traffic volume is only obtained directly in segments provided with counting devices (sensors, radar), survey stations, tolls, cameras, and others. The household level is mostly related to mode choice issues. However, as household-based surveys usually cover only a predefined sample, these studies often face a lack of data on the variable of interest. In short, the availability of travel demand data at the bus stop, road segment, and household levels is defined by budget constraints.

Table 1: Data collection methods in stop-level ridership studies.

Reference	Case study location	N. bus stops	Dependent variable	Collection method	Limitations
Cui et al. [8]	Portland, USA	6,261	Daily boarding	APC	Faulty APC recordings caused trips to be removed from the database.
Dill et al. [14]	Portland, USA Lane County, USA Rogue Valley, USA	7,214 1,400 350	Weekday average boarding + alighting (logarithm)	APC	Ridership data were collected by sampling transit trips for each route at different days.
Kerkman et al. [9]	Arnhem-Nijmegen, Netherlands	1,232 1,284	Average daily boarding + alighting (logarithm)	Smart card data	Passengers buying paper tickets were not included in the data as their trips were not registered.
Mucci and Erhardt [28]	San Francisco, USA	6,261	Average of the number of passengers boarding and alighting at each route-stop	APC	A portion of the bus fleet was equipped with APC technology. The counting device was randomly assigned to buses each day, and therefore all routes were included in the survey after several days. However, to cover the entire system, the authors used an extrapolation method that does not account for spatial dependence.
Chu [13]	Jacksonville, USA	2,568	Weekday total boarding	APC	APCs were permuted over all vehicles so that at least a one-way bus trip was counted once. The level of accuracy of the APC technology was reported to be over 95%.
Ngo [37]	Lane County, USA	1,500	Boarding + alighting from 5am to 11pm	APC	
Lanza and Durand [38]	Austin, USA	1,610	Boardings per day (13 h to 18 h)	APC	
Shi et al. [18]	King County, USA	96	Weekday boarding/weekday alighting	APC	
Chakour and Eluru [7]	Montreal, Canada	8,000	Boarding/alighting (AM peak; PM peak; off-peak day; off-peak night)	APC	APCs covered 15% of the bus fleet. The authors reported working with estimates from a representative sample of trips.
Miao et al. [6]	Salt Lake City, USA	5,879 5,854	Average boardings per bus trip at an individual bus stop for weekdays (ln) Average boardings per bus trip at an individual bus stop for weekends (ln)	APC	APC covered 50% of the fleet. The authors took the average of boardings per trip as a way to reduce the sampling bias. Another shortcoming of using an averaged measure as the variable of interest is that it does not provide within-day variation in ridership, such as peak and off-peak passenger flows.
Marques and Pitombo [16]	São Paulo, Brazil	57	Boarding from 20 h to 23 h59 in a typical day (logarithm)	Boarding and alighting count survey	The survey covered only 0.6% of the bus lines.
Marques and Pitombo [17]	São Paulo, Brazil	96	Boarding from 20 h to 23 h59 on a typical day	Boarding and alighting count survey	The survey covered only 0.6% of the bus lines.
Marques and Pitombo [15]	São Paulo, Brazil	47 49	Weekday boarding Weekday alighting	Boarding and alighting count survey	The survey covered only 0.6% of the bus lines.
Marques and Pitombo [10]	São Paulo, Brazil	97	Boarding + alighting from 5 h to 23 h59 on a typical day	Boarding and alighting count survey	The survey covered only 0.6% of the bus lines.

Table 2: Validation analyses in travel demand modeling.

Geographic unit	Reference	Calibration sample (% of data)	Sampling method
Bus stop	Mucci and Erhardt [28]	Models were calibrated based on data from 2009; data from 2016 were used in the validation step	
	Pulugurtha and Agurla [11]	95.71%	Random selection
	Rahman et al. [12]	93.32%	Not reported
	Marques and Pitombo [10]	100.00%, 85.00%, 70.00%, 55.00%, 40%, 25%	Density of points
Road segments	Eom et al. [39]	17.33%	Random systematic sampling
	Wang and Kockelman [40]	80.00%	Not reported
	Selby and Kockelman [41]	80.00%–90.00%	Random sampling
	Sarlas and Axhausen [42]	80.00%	Random sampling (100 replications)
	Kim et al. [43]	10.00%, 20.00%, 30.00%, 50.00%, 70.00%	Random sampling (10 iterations per sample size)
	Klatko et al. [44]	90.00%	Not reported
	Yang et al. [45]	99.00%, 98.60%, 98.20%, 97.20%, 95.60%, 94.00%, 91.70%, 86.10%, 79.30%, 68.90%, 63.90%	Random
	Mathew and Pulugurtha [19]	75.00%	Random (ArcGIS subset features)
	Pulugurtha and Mathew [20]	75.00%	Random (ArcGIS subset features)
	Marques et al. [46]	100.00%, 70.00%	Density of points
	Chi and Zheng [47], Shamo et al. [48], and Song et al. [21]	No validation	
Traffic analysis zone (TAZ)	Ma et al. [22]	100.00%, 80.00%, 70.00%, 60.00%	Sampling method not reported; a goodness-of-fit measure was reported only for the calibration samples
	Chiou et al. [49] and Tu et al. [23]	No validation	
Station	Blainey and Mulley [24], Blainey and Preston [50], Cardozo et al. [25], Liu et al. [26], and Zhu et al. [51]	No validation	
Household	Pitombo et al. [46]	70.00%	Random
	Linder and Pitombo [30]	70.00%	Not reported
	Gomes et al. [52], Chica-Olmo et al. [53]	No validation	

In Table 2, there is a clear predominance of a single sampling method: the random sampling. However, this type of sampling may not be the best representation of the phenomenon under analysis as the spatial distribution of travel demand cannot be considered as purely random. Often, higher passenger flows are concentrated around some points in the spatial field considered [10, 44, 46]. The spatial distribution of bus stops and sampled road segments, for collecting travel demand data, is also concentrated [22, 39, 41, 42, 46], following the main activity centers. However, the selection of a sample for validation using a random method may overlook the spatial distribution of the geographic units under analysis.

Efforts to account for the spatial variation of collected data can be found in Table 2. Eom et al. [39] used a systematic sampling method based on a 10-mile squared grid system to select counting locations in a traffic volume case. Marques et al. [46] and Marques and Pitombo [10] applied a sampling method based on the density of points in the original dataset to selecting traffic counting locations and bus stops, respectively. Both methods were able to reproduce the spatial concentration of data in the original dataset. In addition, the method based on density of points does not require dividing the spatial field into regular areas and is more convenient to point-based data than the systematic sampling.

Moreover, Wang and Kockelman [40] reported that installing a counting device in a road segment can be influenced, among other features, by level of congestion and road design, which are intervening factors of traffic volume [19, 20, 39, 41, 42]. In this case, a more accurate representation of the phenomenon under analysis would be a sampling method accounting for the spatial distribution of both counted points and predictors of the travel demand variable of interest. Another situation emerges when the collected data are used to calculate a travel demand variable in points outside the original spatial field; that is, the initial data are extrapolated. Zhang and Wang [54], for example, modeled the transit ridership data from one metro line in New York and estimated this variable for another line to be implemented. However, as real transit ridership data on the new line were still not available, the authors could not assess the prediction accuracy of the extrapolation carried out by them.

The representativity and prediction power of the sampling conditions discussed above have not been addressed in the transportation engineering area so far. Another issue related to travel demand modeling, but which has been given little attention in the stop-level literature is the spatial heterogeneity of predictors' effects, discussed in Subsection 2.1.

2.1. Spatial Heterogeneity of Predictors' Effects. Spatially varying impacts of predictors on travel demand variables have already been explored in various spatial scales: traffic volume in road segments [19, 41], passenger demand at the TAZ level [22, 23, 49], stations [25, 26, 51], pedestrian [27], and bus stop [10, 15]. Explanatory variables such as road density [19, 22, 23, 26], residential land use [22, 27], commercial land use [22, 26, 27], income [15, 23, 49], employment [22, 23], population [15, 19], trip frequency [10, 15, 49], station distance [15, 22, 26, 27], and land use mix [23, 27] have shown both positive and negative impacts in more than one spatial scale. Although only a few authors provided results on the statistical significance of the estimated parameters [10, 15, 49], there are studies that tested whether or not a great spatial variation of coefficients was detected in the geographically weighted models [10, 22, 23, 26, 27]. These authors consistently reported a great variability in the effects from intervening factors, indicating that spatial heterogeneity is, in fact, an important feature of travel demand predictors.

Regarding the bus stop level, Marques and Pitombo [10] found a statistically significant spatial variation in two (overlapping bus stops and frequency) of the five predictors used by them to model a transit ridership variable. A significant spatial variation was detected even in a predictor showing only negative effects, pointing out that spatial heterogeneity does not necessarily mean the presence of reverse signs.

2.2. Research Gaps. Based on the literature review conducted, the following research gaps can be enumerated: (1) in the scope of our literature review, no study was found addressing the potential effect of the sampling method when predicting a travel demand variable in a missing data point; (2) only a few stop-level studies perform a validation analysis, making it difficult to assess the performance of proposed models when predicting the transit ridership in a nonsampled stop; (3) the spatial variation of predictors' effects on stop-level ridership has been little explored; and (4) no method has been proposed to treat multicollinearity of spatial predictor data without having to exclude highly correlated predictors.

This study tackles all cited research gaps by proposing the application of a Geographically Weighted Principal Component Analysis (GWPCA) on transit ridership predictor data and using its components as predictors to stop-level boardings and alightings. Four different sampling strategies are considered, and the model performance is assessed in both calibration (available data) and validation (missing data) samples. The convenience of GWPCA in the transportation engineering area relies on the fact that it incorporates not only the predictor data multicollinearity but also its spatial heterogeneity into the modeling. By doing so, a novel contribution of GWPCA is to identify the most important predictor to the travel demand variable of interest at each point of the database, even the nonsampled ones.

3. Materials and Method

Figure 1 illustrates the research workflow followed in this study. Except for the literature review, the remainder text details each one of the highlighted blocks.

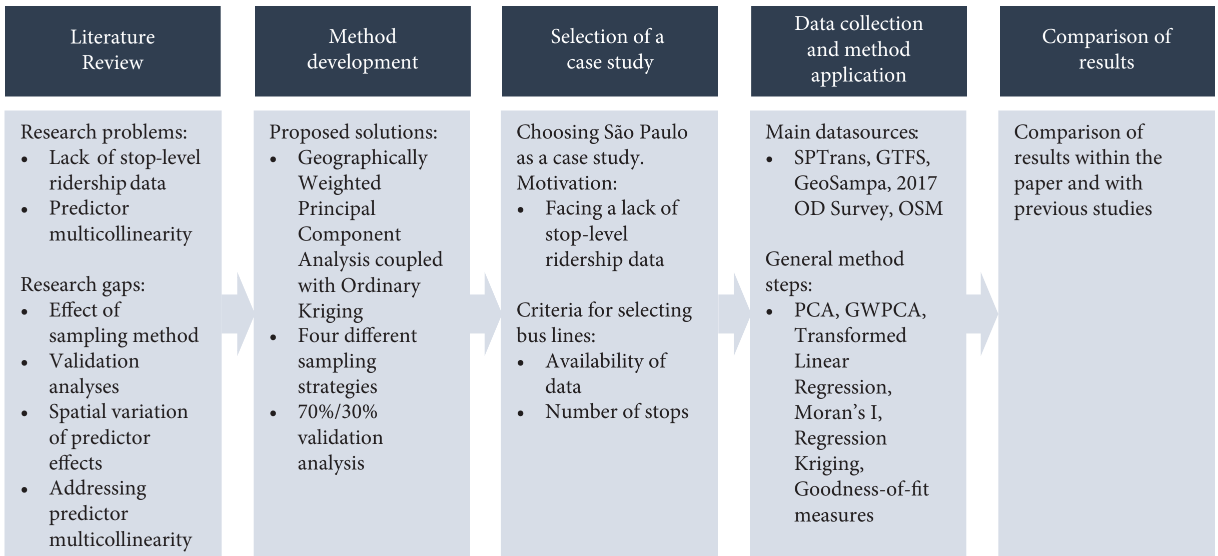

Figure 1: Research workflow.

The case study takes place in São Paulo (Brazil), the most populous city in South America [55] and main economic center of Brazil. Although there is a high representativity of the individual motorized travel mode in São Paulo, bus transit remains as the most used public travel mode in the city [56].

Two datasets compose the analyses carried out in the study: First, a database containing 19,900 bus stops in São Paulo, and second, a database comprising 207 stops of four bus lines in São Paulo for which information on boarding and alighting was available. *SPTrans*, the administrator of the São Paulo bus service, made available the 2017 results of a boarding and alighting count survey along 8 lines of São Paulo, which, separated by direction, comprise 16 unidirectional lines. Among them, four lines were selected for a case study: line 6045-10-2 with 49 bus stops, line 6913-10-1 with 52 bus stops, line 809L-10-2 with 45 bus stops, and line 577T-10-1 with 61 bus stops. Two main criteria guided the line selection: availability of data regarding all independent variables and a reasonable number of bus stops. Figure 2 shows the location map of São Paulo, the lines visited by the survey, and the lines chosen for our case study.

3.1. Dependent Variables. For each bus line, the original variable of interest was the number of boardings or alightings at its bus stops from 5 h to 23h59 in a typical day (Tuesday, 2017-11-07). After verifying that boardings and alightings had a right-skewed distribution, a Box–Cox transformation [57] was applied to their raw data. Thus, for modeling purposes, the dependent variable was the Box–Cox-transformed number of boardings, for some lines, and alightings, for others (more details in Subsection 3.5).

3.2. Independent Variables. Based on a thorough bibliographic review by Marques and Pitombo [10], we collected predictor data from each bus stop in São Paulo using, as a catchment area, the region defined by a radius of 400 m centered in the bus stops [58]. Table 3 summarizes the independent variables collected, their source, and some descriptive measures.

Although the original dataset contained 19,900 bus stops, 571 of them did not have information related to some of the predictors listed in Table 3. Therefore, the method steps described as follows were carried out using the remaining 19,329 bus stops. The predictor data for these 19,329 stops can be accessed through the file provided in the supplementary material section.

3.3. Principal Component Analysis. Before proceeding to the data dimensionality reduction, two tests were applied to confirm the suitability of the predictor dataset (Table 3) to the principal component analysis: the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy [60] and the Bartlett test of sphericity [61]. A good adequacy of the dataset is achieved when the independence hypothesis of Bartlett's test is rejected [62] and the KMO measure reaches a value close to 1 [63]. After confirming that a data dimensionality reduction technique would be useful to the predictor dataset, a traditional PCA [64] was applied to it, and only components with eigenvalue greater than 1 were retained.

3.4. Geographically Weighted Principal Component Analysis: Addressing the Multicollinearity of Spatial Data. The Geographically Weighted Principal Component Analysis (GWPCA) corresponds to a local version of the traditional PCA [65]. In this case, a different PCA is carried out at each point of the database, using weighted neighbor data. An underlying assumption is that the principal component structure follows a spatial pattern, as closer points are more similar than distant ones [66]. Therefore, the loading values vary from one geographic coordinate to another, and it is

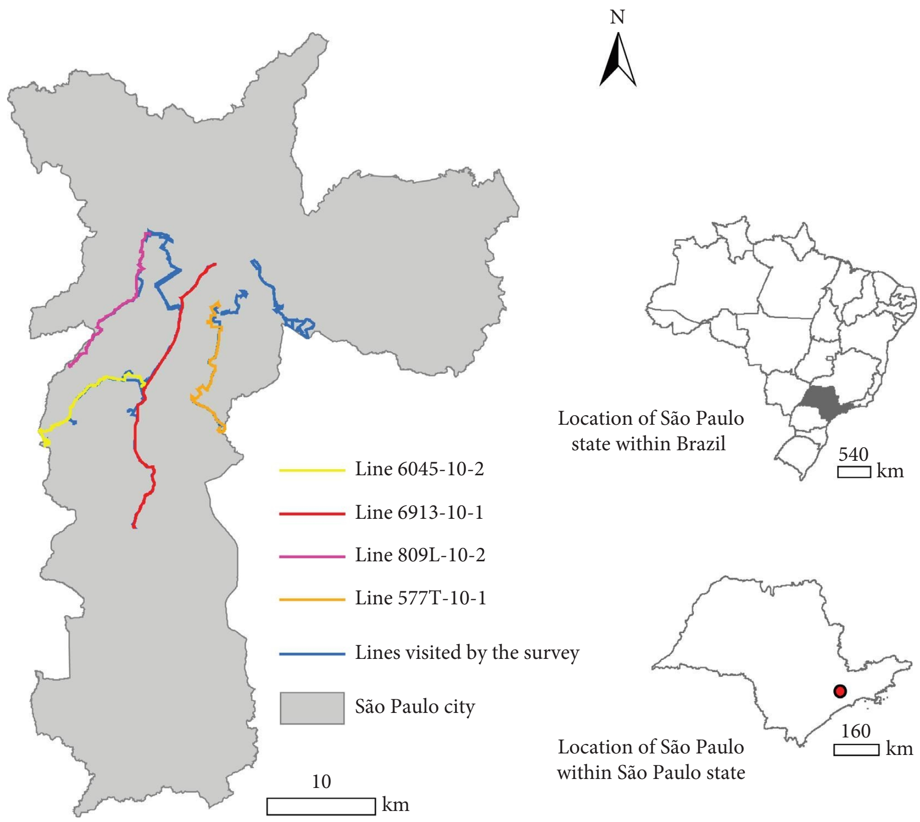

FIGURE 2: Case study databases.

possible to map the predictor having the highest absolute loading value for all PCs, commonly called the winning variable.

In the GWPCA (local PCA), the variance-covariance matrix Σ of a dataset X varies as a function of the location i, with coordinates (u,v), as shown in the following equation [65]:

$$\sum (u, v) = X' W (u, v) X, \tag{1}$$

where $W(u, v)$ is a weight matrix representing the spatial interaction between the database points. In this study, the elements of W are given by the bisquare kernel (2) [65].

$$W_j(i) = \begin{cases} \left(1 - d_{\mathrm{ij}}^2/b^2\right)^2, & \text{if } d_{\mathrm{ij}} \le b, \\ 0, & \text{if } d_{\mathrm{ij}} > b, \end{cases} \quad j = 1, 2, \ldots n, \tag{2}$$

where d_{ij} is the Euclidean distance between the neighbors i and j and b is the bandwidth. The bandwidth can be thought as the region in space within which the points are spatially dependent. In our case study, this bandwidth is the number of nearest neighbors. Using the same number of components retained in the traditional PCA, the bandwidth was optimized by a cross-validation goodness-of-fit measure as described by Harris et al. [65].

Finally, the geographically weighted PCs can be written as (3), in which each location i has its own loading L and variance values V for the defined principal components [65]. Both spatial and nonspatial PCAs were based on correlation matrices.

$$\mathrm{LVL}' \mid (u_i, v_i) = \sum (u_i, v_i). \tag{3}$$

Comparisons between GWPCA and PCA in this study were based on the percentage of variance extracted by the retained PCs and the bandwidth obtained in the GWPCA. If the database spatial pattern yields a large bandwidth (e.g., close to the total number of points minus 1), results from both approaches will be similar. Therefore, using GWPCA may not be justified in this case. Conversely, a smaller bandwidth would indicate the presence of a clear spatial/local structure in the predictor dataset.

3.5. Modeling. In the modeling step, the GWPCs retained from GWPCA and the Box–Cox-transformed boarding and alighting variables went through a linear correlation analysis using the Pearson correlation coefficient. Initially, a correlation analysis was conducted on all possible combinations of dependent variable (boarding or alighting) and geographically weighted principal components (scores) using the complete line databases. Based on an inspection of the highest correlations in the complete line databases, only one interest variable (boarding or alighting) was adopted for each bus line. However, the most correlated GWPC could vary from one sampling method to another as the GWPC most correlated to each specific sample was always selected.

Table 3: Stop-level ridership predictor data.

Variable	Description	Originated from	Source	25%	50%	75%
tot_lines	Number of lines passing through the bus stop	Bus stop	2017 GTFS data	1.00	2.00	5.00
Headway	Average headway (seconds)			746.84	939.18	1,168.42
bus_dist	Distance to nearest bus terminal (m)		GeoSampa	2,122.06	4,341.68	9,768.87
metro_dist	Distance to nearest metro station (m)			1,708.24	4,335.74	9,870.15
train_dist	Distance to nearest train station (m)			2,565.13	5,287.93	9,192.92
center_dist	Distance to the city center (Sé Square) (m)			8,214.43	12,647.06	17,929.54
n_shelters	Number of bus stop shelters		SPTrans	0.00	0.00	1.00
pop	Population (inhabitants)	Catchment area	2017 Origin and Destination Survey [56]	4,444.08	6,275.22	8,076.37
fem	Female percent			0.48	0.53	0.59
educ_level	Percent of people with complete higher education			0.06	0.15	0.31
Youth	Youth percent (up to 17 years)			0.11	0.18	0.25
older_adults	Percent of older adults (60+ years)			0.07	0.14	0.22
perc_noveh	Percent of no-vehicle households			0.30	0.48	0.64
Income	Average household income, in BRL*			2,625.90	3,411.32	4,851.97
Employment	Number of jobs		Morelli et al. [59]	315.97	798.47	2,056.31
low_standard_h_res	Low standard horizontal residential area (km^2)		GeoSampa	0.00	16,988.11	113,126.92
medhigh_standard_h_res	Medium/high standard horizontal residential area (km^2)			8,956.52	61,554.21	142,802.94
low_standard_v_res	Low standard vertical residential area (km^2)			0.00	0.00	30.42
medhigh_standard_v_res	Medium/high standard vertical residential area (km^2)			0.00	7,906.73	60,652.29
com_serv	Area of commerce and services (km^2)			0.00	8,313.25	29,583.92
res_com_serv	Residential, commerce, and service area (km^2)			8,342.62	27,348.98	54,174.91
res_ind_wareh	Residential, industrial, and warehouse area (km^2)			0.00	0.00	16,027.82
comserv_indwareh	Commerce, service, industrial, and warehouse area (km^2)			0.00	0.00	6,440.51
Institutional	Institutional area (km^2)			0.00	0.00	2,662.35
no_predominancy	Area with no predominant land use (km^2)			0.00	0.00	12,084.90
park_area	Area of parks (km^2)			0.00	0.00	10,050.33
Entropy	Entropy index: a measure of land use mix (varies from 0 to 1)			0.45	0.54	0.61
buslanes_length	Length of bus lanes (m)			0.00	0.00	900.31
bikenet_length	Length of cyclepaths (m)			0.00	0.00	941.11
arterial_length	Length of arterial roads (m)			0.00	867.06	1,911.65
Intersections	Number of intersections		OpenStreetMap	24.00	46.00	76.00
sameline_overlap	Number of bus stops having at least one line in common with the reference bus stop		GTFS	2.00	4.00	6.00

*1 BRL equals to 0.19 USD (Feb. 2023).

Preliminary results including other correlated components did not show a significant improvement in the prediction accuracy. Therefore, only one component was used for each case.

Having found the pairs of dependent variable and most correlated GWPC, a Transformed Liner Regression (TLR) was calibrated for each bus line. Afterward, spatial dependence on the transformed regression was disclosed by applying the Moran index [67] on its residuals. To calculate Moran's I, we adopted a weight matrix based on the distance between points along the bus line, which is termed "network distance." Spatial dependence on residuals from the transformed linear regression was addressed by a spatial interpolator called Ordinary Kriging (OK) [68–70], in which the data spatial variance was modeled using network distances and the exponential model [71]. The final estimates of the conjoint approach between GWPCA and OK were obtained through the following equation:

$$Z_{x_0}^* = \alpha + \beta S + \sum_{i=1}^{n} \lambda_i e(x_i), \quad (4)$$

where α and β are parameters from the transformed regression, S represents the scores of the most correlated component, λ are the OK optimum weights, $e(x_i)$ is the TLR residual for the neighbor x_i, and n is the number of sampled neighbor points. Predictions from the nonspatial model include only the first two terms on the right side of (4) Coupling a regression model with the kriging interpolation of residuals has been referred by some authors as "Regression Kriging (RK)" [72, 73]. This is the term we use hereby to refer to the estimates from (4).

3.6. Validation and Cross-Validation. The calibration sample of previous studies varied from a minimum of 10% up to 99% of the total data (Table 2). Percentages between 60% and 90% represent half of the case studies. Based on this, we selected a percentage of 70% for the calibration samples and the remaining 30% for the missing data. Ridership estimates were obtained for both calibration and validation samples.

Estimated values were back-transformed, so they could return to the same scale as the observed values. Then, we compared the performance of the transformed regression with the Regression Kriging approach using three goodness-of-fit measures: Root mean squared error, median of absolute percentage error, and mean absolute error [74]. The modeling and cross-validation/validation steps were repeated for different types of sample collection, which are detailed in Subsection 3.7.

3.7. Sampling Strategies. Four sampling methods were considered in the validation step: simple random sampling, density of points, balanced sampling with geographical spreading, and sample for extrapolation. They are described as follows.

3.7.1. Simple Random Sampling. Considering a simple random sampling, all points in the dataset have the same probability of being chosen [75].

3.7.2. Density of Points. In the sampling strategy based on the density of points, bus stops located in regions with a high density of bus stops have a higher probability of being selected [76]. An assumption underlying this method is that areas with a high concentration of bus stops are also richly served by bus lines. The higher the number of lines, the higher the chance of having information on boarding and/or alighting available.

3.7.3. Balanced Sampling with Geographical Spreading. This method involves two concepts: balanced and well-spread sampling. Knowing the population mean of a covariate that is related to the variable of interest, a balanced sample on this covariate will choose points whose mean is equal to the population mean [75]. Therefore, points are selected in such a way that the variation of the covariate is well-represented by the sample.

However, a balanced sample can result in a poor geographical spreading. To avoid clustering of points and assure a good balancing on both the covariate values and geographic coordinates, the balanced sampling with geographical spreading accounts for these two factors simultaneously. This sampling method was performed using, as a covariate, the principal component most correlated to the Box–Cox transit ridership variable. As it is required to know which component is the most correlated prior to generating the sample, we initially used the component most correlated to the transformed ridership data in the complete line dataset. If the ridership data in the resulting sample had a weak correlation with the component considered (absolute value of Pearson correlation coefficient lower than 0.3 [77]), another sample was generated using the component most correlated to the ridership data in the sample based on the density of points or the simple random sample, which are methods more usual in practice than the extrapolation one.

3.7.4. Extrapolation. The sample for extrapolation seeks to reproduce an extreme scenario in which the ridership data from one line are intended to be used in the ridership prediction for points from neighboring lines. In our case study, the calibration sample in the extrapolation strategy was generated as follows: 15% of points in the beginning and 15% of points in the end of each line were regarded as the validation sample (missing data); the remaining 70% of bus stops, belonging to the more internal segments of the case study lines, were used as a calibration sample. The sequence of method steps is illustrated in Figure 3.

3.8. Computational Tools. Table 4 summarizes the computational tools that supported each method step. Most of the procedures were carried out in the open-source software R, making it easier for the method to be replicated in other databases.

4. Results and Discussion

This section is divided into five subsections: results from the traditional and the geographically weighted PCA are presented in Subsection 4.1; afterward, we discuss the spatial and nonspatial modeling outcomes. Subsection 4.3

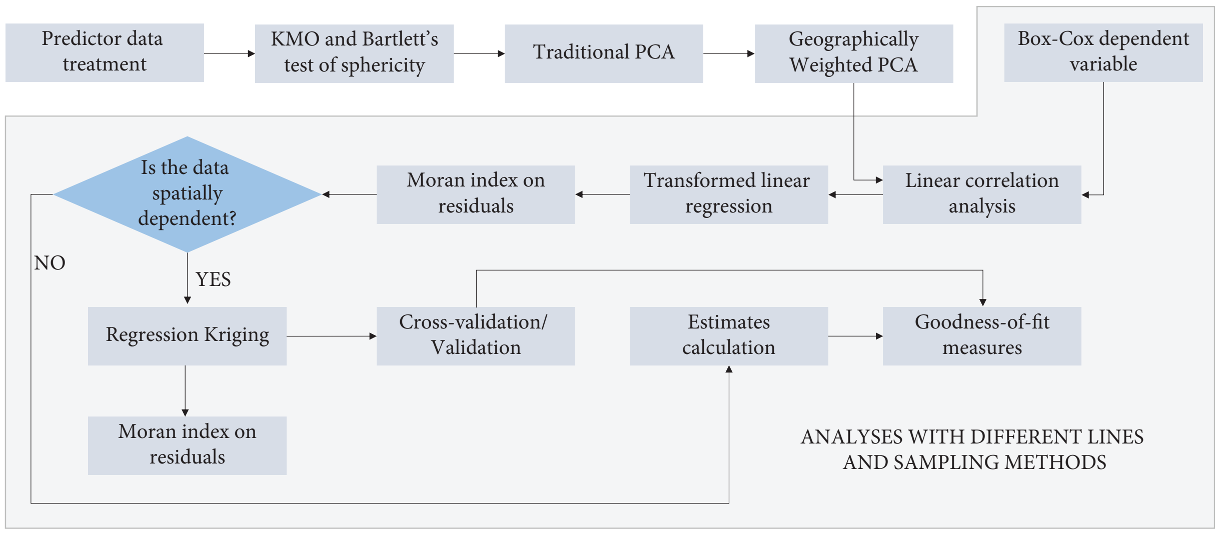

Figure 3: Flowchart of method steps.

illustrates an example of how coefficients from a geographically weighted principal component can be interpreted. Subsection 4.4 provides insights into the best modeling approach and sampling strategy. A comparison between results from this study and previous studies is presented in Section 4.5.

4.1. Global and Local PCA. As shown in Table 5, the KMO measure and Bartlett's test confirmed the adequacy of the predictor dataset to the principal component analysis.

In the PCA, 10 components were retained, which had an eigenvalue greater than 1. These 10 PCs extracted 62.52% of the variance in the original database. Nonrotated loading values are presented in Table 6, highlighting the highest absolute values of the loadings for each component.

PC1 contrasts high accessibility bus stops. Negative loading values for the distances to the center, to the nearest bus terminal, train, and metro stations, reveal that the higher the distance between the bus stops and these elements is, the lower the PC1 score will be. The educational level and income are also important variables composing PC1. In turn, PC2 represents commercial areas, with large amounts of jobs and transport infrastructure, from both motorized and active modes. Therefore, PC1 and PC2 could be named as intra/intermodal proximity and central areas, respectively. In short, PC3, PC4, PC5, PC6, PC7, PC8, PC9, and PC10 are measures of population, industrial land use, age, bus network spatial coverage, bus network temporal coverage, institutional areas (or areas with a low occupation density), lower-income female population, and bus stop facilities, respectively.

Although maximum loadings in Table 6 assume only moderate values, this is not rare in PCA (see, for example, Jolliffe [64]). Loading values depend on factors such as the restriction adopted for maximizing the variance extracted by each component and rotation [64]. Rules for discriminating maximum loadings among moderate loading values, as the one applied in Table 6, can be consulted in Jolliffe [64].

Figure 4 shows the winning variables for the first and second principal components in the GWPCA. The winning variable is the predictor with the highest absolute value in the local PCA. All 19,900 stops have information on the highest loading value and respective winning variable. Loading values for each variable in the ten retained components have been provided as supplementary material (see supplementary material section).

In the global PCA, the first component was mainly represented by eight variables (Table 6). This number increases to 14 in the local PCA. While most variables comprising PC1 are measures of intramodal and intermodal integration, 20% of bus stops had the number of jobs as the winning variable of GWPC1. An interesting result is that these stops are concentrated in the center of São Paulo (orange), which shows the highest employment densities in the city.

Three other variables represented 10% or more of the bus stops in GWPC1: the educational level, low standard horizontal residential area, and entropy index. The education level is highlighted in bus stops from the northwest and southeast regions (green), while low standard horizontal residential areas characterize stops in the south of São Paulo (light blue). Bus stops in the extreme south had a higher importance of the variable entropy, probably because they refer to areas with a high variation in the land use mix index.

In the GWPC2, 54.39% of the bus stops were mainly characterized by a land use category (low standard vertical residential area, or commercial, services, industrial, and warehouse area) or by a variable related to the bus system (number of bus stop shelters or bus lane length). Of these predictors, only the length of bus lanes appears as one of the main features composing the second component of the global PCA. Figure 5 presents the percent of variance extracted by GWPC1 and GWPC1 plus GWPC2.

The first two components were able to account for more than 30% of the variance in the original dataset for some bus stops in the center and extreme south of São Paulo. GWPC1

Table 4: Software associated with each method step.

Method step	Tool	Source
KMO measure and Bartlett's test	IBM SPSS Statistics 22	—
Principal component analysis	R	R Core Team [78]
Geographically weighted PCA	R library "GWmodel"	Gollini et al. [79]; Lu et al. [80]
Box–Cox transformation	R library "EnvStats"	Millard [81]
Moran index	R library "ape"	Paradis et al. [82]
Network distance calculation	GRASS GIS	Bundala et al. [83]
Linear correlation analysis	IBM SPSS Statistics 22	—
Transformed linear modeling	R	R Core Team [78]
Semivariogram modeling	R library "KrigLinCaution"	Ver Hoef [84]
Checking the possibility of occurrence of negative variances in the kriging interpolation	R library "KrigLinCaution"	Ver Hoef [84]
Geostatistical cross-validation	R library "KrigLinCaution"	Ver Hoef [84]
Geostatistical validation	R	The authors
Simple random sampling	R	R Core Team [78]
Density of points	R library "spatialEco"	Evans [76]
Balanced and well-spread sampling	R library "BalancedSampling"	Grafström and Lisic [85]

Table 5: Suitability of the dataset for PCA.

Measure		Value
Kaiser–Meyer–Olkin measure of sampling adequacy		0.776
	Approx. χ^2	257281.195
Bartlett test of sphericity	df	496
	Sig.	0

Table 6: Principal component analysis on stop-level ridership predictor data ($N = 19{,}329$).

Predictor	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
tot_lines	0.09	0.22	−0.02	0.12	−0.21	**0.38**	**−0.43**	0.05	0.06	−0.09
headway	0.03	−0.07	−0.22	−0.13	−0.05	−0.09	**−0.31**	0.23	−0.02	−0.14
pop	−0.02	0.03	**0.43**	**0.32**	**0.32**	0.02	0.06	0.04	0.08	0.06
low_standard_h_res	**−0.24**	0.05	0.20	0.11	0.11	0.04	−0.05	−0.03	−0.03	0.22
medhigh_standard_h_res	0.08	**−0.39**	0.15	0.05	**−0.28**	−0.01	−0.06	−0.05	−0.28	−0.03
low_standard_v_res	−0.07	0.05	0.13	0.11	0.09	0.29	0.23	−0.18	**0.41**	0.02
medhigh_standard_v_res	0.21	−0.17	−0.09	0.16	**0.33**	0.02	0.01	0.20	0.22	−0.10
com_serv	0.12	**0.32**	−0.02	0.00	0.02	−0.18	−0.12	0.00	−0.19	0.09
res_com_serv	0.15	0.14	0.14	**0.27**	−0.02	−0.16	−0.08	0.11	0.02	−0.20
res_ind_wareh	0.07	−0.04	0.22	**−0.27**	−0.25	0.09	0.2	0.21	0.05	**−0.37**
comserv_indwareh	0.09	0.13	0.03	**−0.41**	−0.19	0.07	0.16	0.17	0.25	−0.03
institutional	0.10	0.08	−0.16	−0.06	−0.01	0.18	0.14	**−0.47**	−0.01	0.16
no_predominancy	0.09	0.13	0.01	−0.13	0.05	0.06	−0.07	**−0.32**	0.10	0.08
entropy	0.19	0.10	**0.35**	0.03	−0.18	0.21	0.21	−0.04	0.11	−0.09
employment	0.18	**0.27**	−0.01	0.08	0.17	−0.24	−0.08	0.15	0.03	0.04
fem	0.01	−0.05	0.07	0.07	−0.10	−0.24	−0.29	−0.12	**0.40**	−0.13
educ_level	**0.28**	−0.14	−0.22	0.22	0.10	0.07	0.02	0.05	0.05	−0.03
youth	−0.16	0.11	0.07	**−0.26**	**0.27**	0.23	−0.04	0.21	−0.13	−0.08
older_adults	0.13	−0.14	−0.03	0.18	**−0.43**	−0.29	0.03	−0.23	0.20	0.09
perc_noveh	−0.15	**0.24**	0.20	−0.14	0.03	**−0.34**	−0.20	−0.10	0.11	0.10
income	**0.24**	−0.19	**−0.25**	0.19	0.1	0.18	0.05	0.06	−0.02	0.01
bus_dist	**−0.29**	0.11	−0.16	**0.28**	−0.17	0.02	0.16	0.11	0.03	−0.09
metro_dist	**−0.33**	0.09	−0.16	0.16	−0.17	0.01	0.08	0.10	0.02	−0.07
train_dist	**−0.32**	0.09	−0.13	**0.26**	−0.09	0.02	0.11	0.05	0.03	−0.09
center_dist	**−0.33**	0.05	−0.14	0.09	−0.11	0.07	−0.02	0.06	0.06	−0.04
buslanes_length	0.18	**0.36**	−0.04	0.15	−0.13	0.00	0.14	−0.02	−0.14	−0.13
bikenet_length	0.18	**0.23**	−0.06	0.12	0.01	−0.08	0.22	0.04	−0.13	−0.11
arterial_length	**0.22**	**0.34**	−0.08	0.04	−0.12	0.03	0.09	0.03	−0.18	0.09
park_area	−0.10	0.07	−0.23	−0.09	0.13	0.10	−0.03	**−0.38**	−0.21	−0.17
intersections	−0.01	−0.13	0.31	0.16	−0.19	0.10	−0.03	0.12	**−0.39**	**0.35**
sameline_overlap	−0.01	0.02	0.20	0.14	0.01	0.23	**−0.39**	−0.26	−0.13	**−0.46**
n_shelters	0.07	0.12	−0.09	0.01	−0.17	**0.34**	−0.30	0.20	0.22	**0.46**
Proportion of variance (%)	20.89	7.48	6.57	5.27	4.40	4.05	3.72	3.53	3.45	3.17

Bold values highlight the highest values in each column.

alone could extract a portion of variance higher than 22% for stops in the extreme south of the city. Recall that the database with 32 predictors was collected based on a thorough bibliographic review on factors affecting the stop-level transit ridership. Therefore, overall, the winning variables shown in Figure 4 may represent the most important features influencing the bus patronage at each stop of São Paulo. Although information on boarding and alighting is available only for a few bus stops, decisions regarding the land use and bus network planning toward increasing the number of passengers might benefit from the GWPCA results.

Together, the 10 retained GWPCs managed to extract from 64.94% to 76.36% of the variance in the original database, surpassing the unique value of 62.50% obtained for all bus stops in the traditional PCA. In addition, the bandwidth of GWPCA covered the nearest 5,830 neighbors, which means that only 30%, approximately, of all points were used to calculate the local PCs at each bus stop. This result confirms the existence of a spatial structure in the predictor dataset and suggests the better adequacy of GWPCA over PCA for addressing the multicollinear nature of stop-level ridership predictors.

4.2. Nonspatial and Spatial Modeling. One major concern addressed by this study is whether the sampling strategy affects the spatial prediction of a transit ridership variable at the bus stop level. The modeling step was carried out for four different lines, separately, and considering calibration samples based on four sampling methods. Figures 6, 7, 8, and 9 present the spatial variation of the transit ridership variable

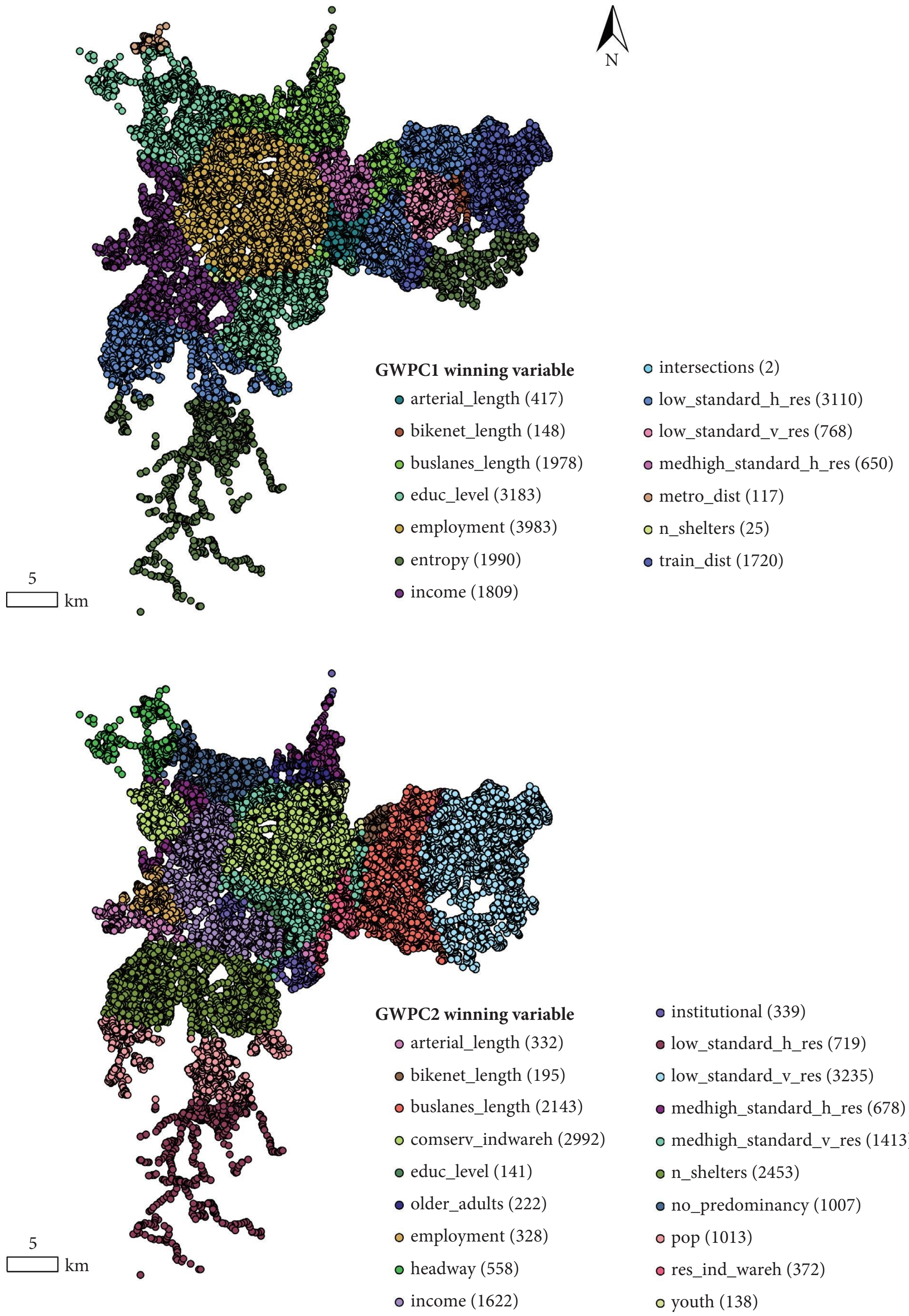

Figure 4: GWPC1 and GWPC2 winning variables ($N = 19{,}900$).

selected for modeling along calibration and validation samples of lines 6045-10-2, 6913-10-1, 809L-10-2, and 577T-10-1, respectively. Calibration samples are shown on the left and validation samples on the right.

Calibration samples of lines 6045-10-2, 6913-10-1, 809L-10-2, and 577T-10-1 had 35, 36, 31, and 43 bus stops, respectively. On the other hand, validation samples covered 14, 16, 14, and 18 stops, respectively. Based on a linear

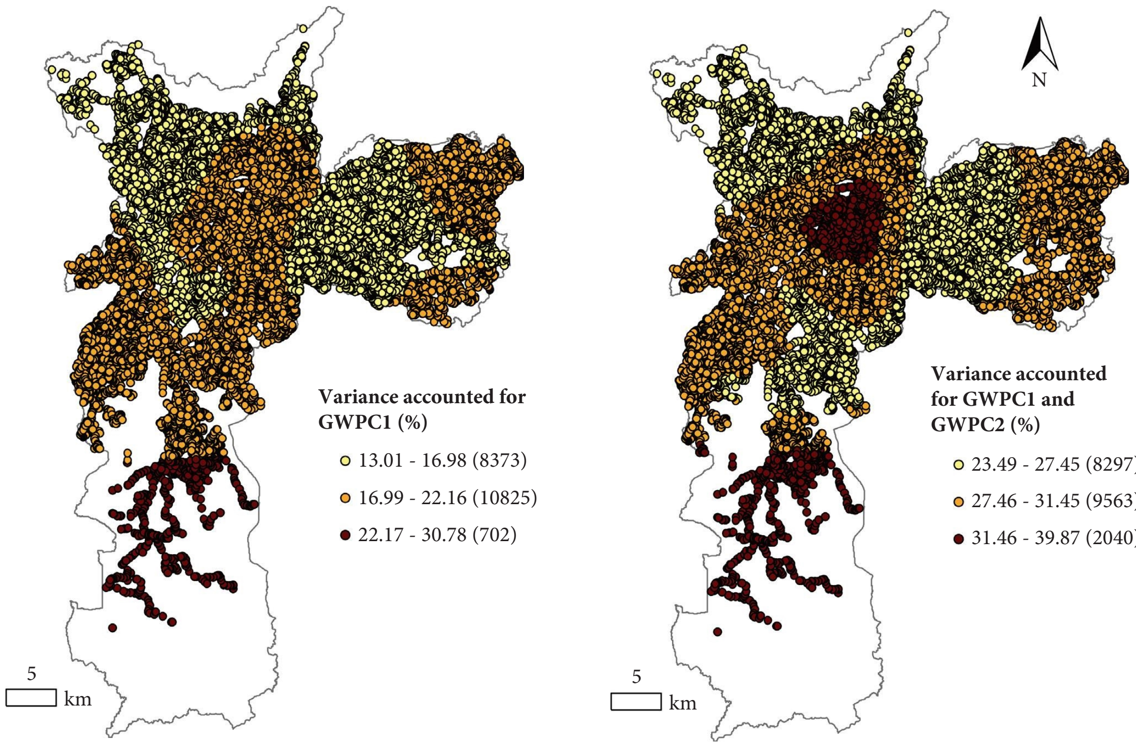

Figure 5: Local percent of variance ($N = 19{,}900$).

correlation analysis, the number of alightings was the variable of interest for lines 6045-10-2 and 809L-10-2, whereas lines 6913-10-1 and 577T-10-1 had the number of boardings as the dependent variable. As both boardings and alightings correspond to data from an entire day (from 05 h to 23 h59), higher passenger flows occur near activity centers and densely populated areas. However, activity centers usually have a higher concentration of the transit system, that is, higher bus stop and bus line densities than residential areas.

It can be observed that each sampling method reproduced, in fact, their main objective, as described in Subsection 3.7. Table 7 summarizes the results from the modeling step.

As the kriging interpolation was applied only on the residuals from the Transformed Linear Regression, the parameters of intercept and the GWPCs are identical for both TLR and RK. These two parameters were statistically significant in all scenarios analyzed (p value <0.05). The GWPCs comprise information on 32 scaled predictors. Therefore, interpretation of their effect on the corresponding ridership variable is not straightforward. Subsection 4.3 discusses what insights can be drawn from a GWPC coefficient using the line 6045-10-2 results as an example.

4.3. Interpreting the Effect of a Geographically Weighted Principal Component on Stop-Level Transit Ridership. To assist the interpretation of the GWPC5 effects on alightings, Figure 10 presents the variable of interest along line 6045-10-2, the scores, and the first and second winning variables of GWPC5. The spatial pattern of GWPC1, GWPC3, GWPC7, and GWPC9 along the remaining case study lines is provided in Figures 11, 12, and 13. Score values for the 19,329 bus stops used in the case study were provided in the supplementary material section.

A negative value for the parameter associated with GWPC5 (Table 7) reveals that stops with lower values of the GWPC5 scores show higher volumes of alightings. The number of alightings along line 6045-10-2 is low at its first bus stops and increases as the bus travels the itinerary (from northeast to southwest) until reaching a maximum value of 746 passengers in the last stop. A clear spatial dependence can be visualized in this variable of interest. This pattern is inverted when it comes to the GWPC5 scores, which show negative values in points with high passenger demand and positive values in stops with a lower number of users. Therefore, a negative parameter for GWPC5 is understandable.

Four predictors appear as the winning variable of GWPC5 along line 6045-10-2, with a high representativity of the no predominant land use feature. The second winning variable (i.e., the variable with the second highest absolute loading value) was more diverse than the first. This time, three predictors prevailed: intersections, low standard vertical residential area, and population. Given the negative parameter obtained for GWPC5, predictors having a negative loading probably exert a positive effect on alightings, while those showing a positive loading are likely to decrease the number of alightings. Intersections and sameline_overlap showed negative loading values in both first and second winning variables. The number of intersections characterizes walkable neighborhoods, while higher

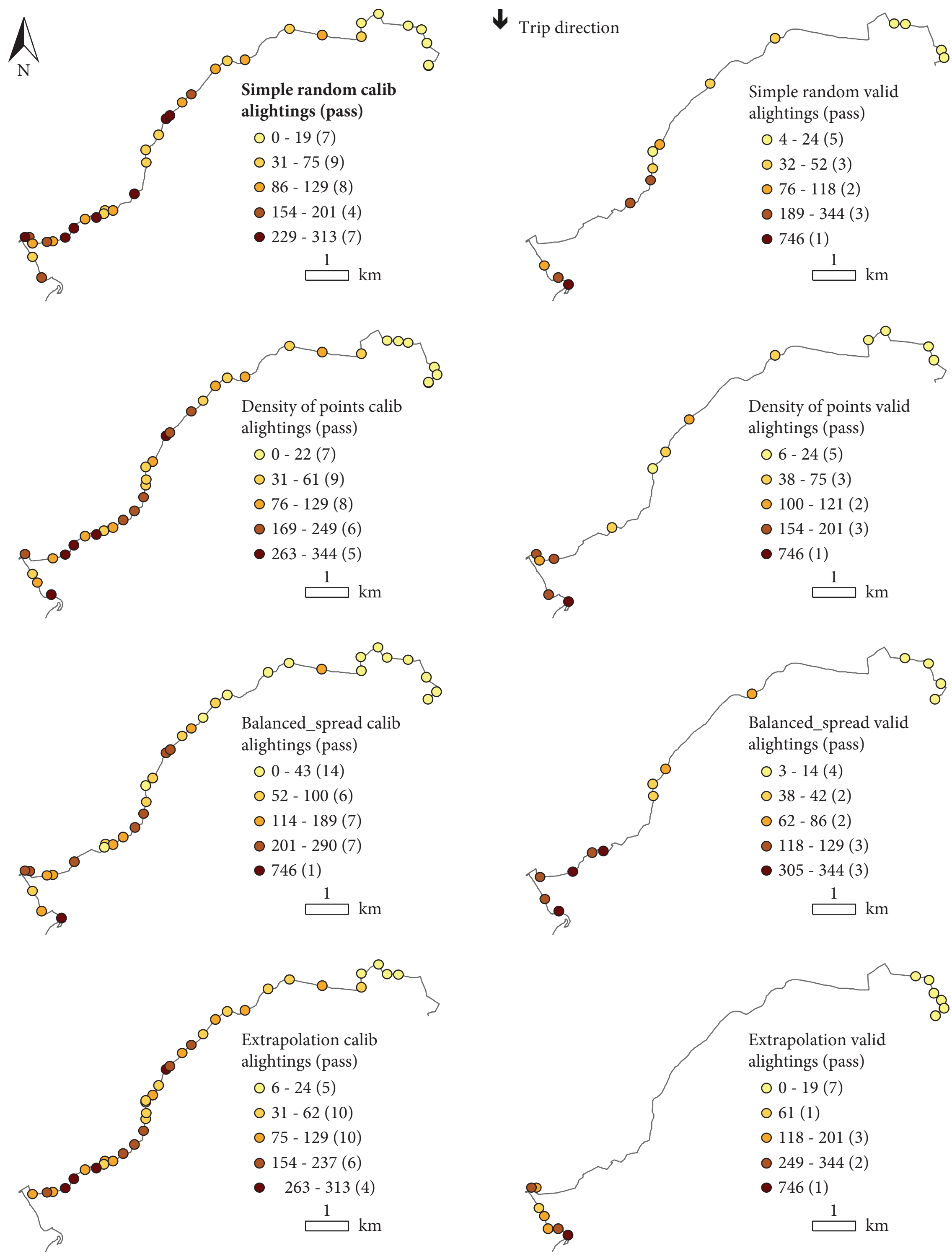

FIGURE 6: Alightings along calibration and validation samples of line 6045-10-2.

concentrations of bus stops indicate a higher coverage of the bus network. Other predictors with negative loadings in the first winning variable are as follows: no predominant land use area and low standard vertical residential area, pointing to the positive contribution of a diverse land use and the low-income population to the transit ridership.

4.4. Performance Evaluation of the Models and Sampling Strategies. The decision of adopting a spatial approach was attested by two methods: Moran's I and goodness-of-fit measures. Recalling the results summarized in Table 7, Moran's I confirmed the presence of a statistically significant spatial dependence on residuals from the transformed linear

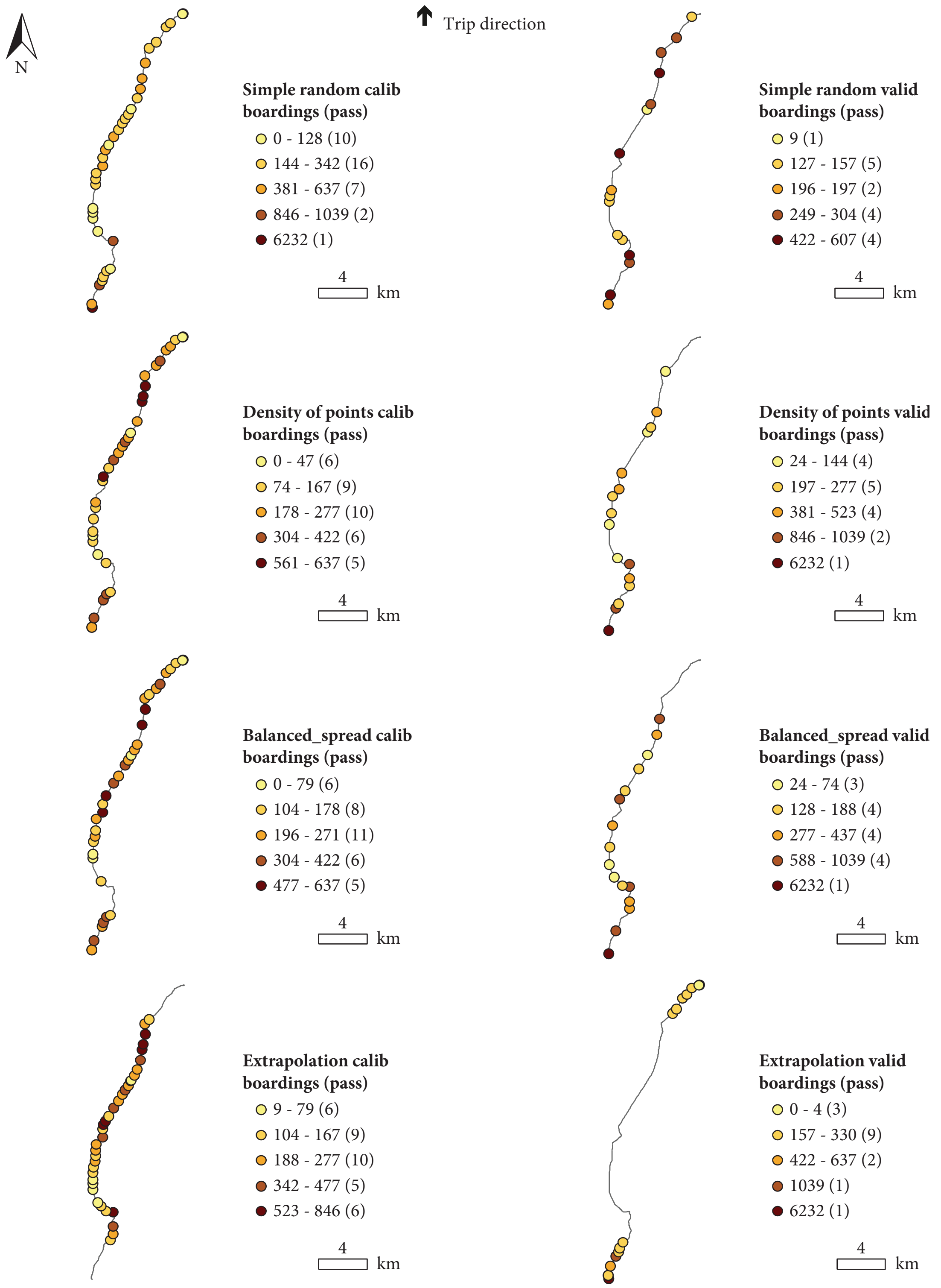

Figure 7: Boardings along calibration and validation samples of line 6913-10-1.

regression in most combinations of bus lines and sampling strategies. However, after the kriging interpolation, the null hypothesis of no autocorrelation was accepted.

Table 8 presents the goodness-of-fit measures results, which are separated by calibration and validation samples, sampling strategy, and bus line. The cases where Regression

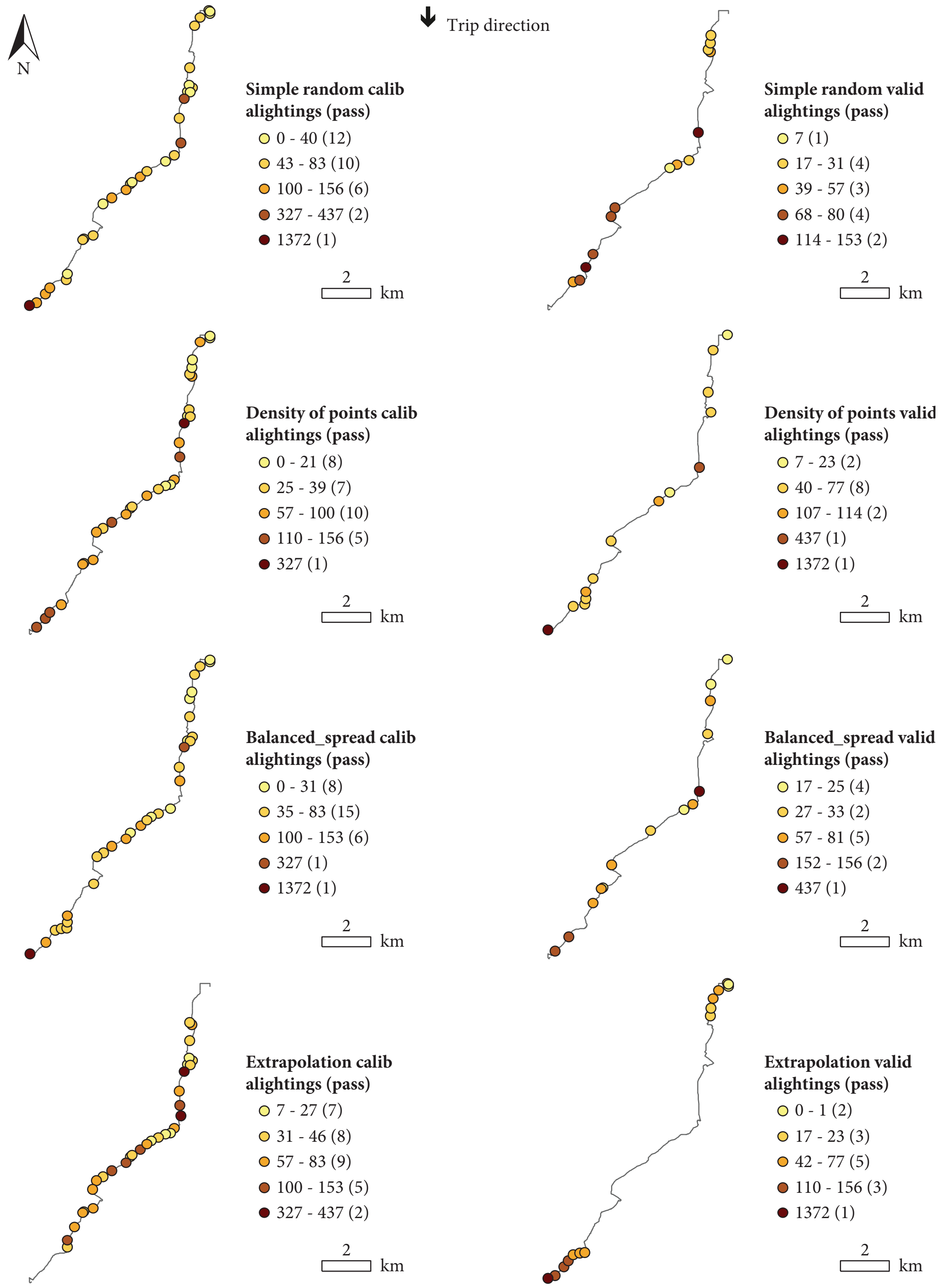

Figure 8: Alightings along calibration and validation samples of line 809L-10-2.

Kriging performed better than the Transformed Linear Regression are highlighted in bold. Blank spaces in the RK columns refer to the cases where no spatial dependence was detected in the TLR model.

Considering the three goodness-of-fit measures (MedAPE, RMSE, and MAE), there are 78 pairs of comparison between TLR and RK, as not all cases involved the application of RK. RK performed better than TLR in more

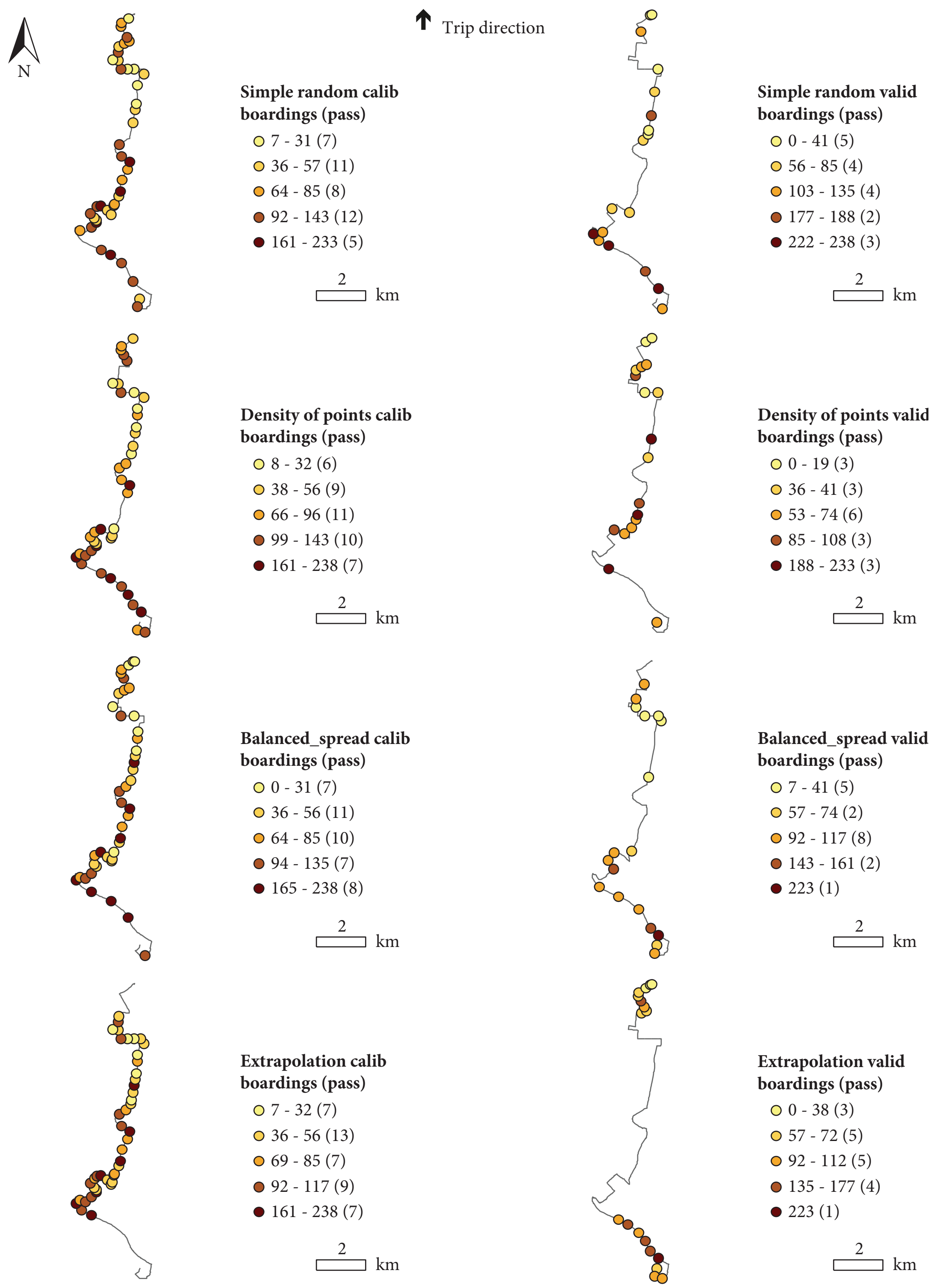

FIGURE 9: Boardings along calibration and validation samples of line 577T-10-1.

than half of these 78 cases. The improvements of RK over TLR, measured as the reduction in the error provided by RK compared to TRL, vary from 0.27% (MAE of the line 577T-10-1 calibration sample in the extrapolation case) to 48.59% (MedAPE of the line 6045-10-2 validation sample in the balanced and well-spread case). Improvements provided by RK reach higher values in the validation samples.

Table 7: Spatial and nonspatial modeling of stop-level transit ridership.

Line (variable of interest)	Sampling method	Model	Intercept	GW PC1	GW PC3	GW PC5	GWPC7	GW PC9	Moran index (p values)	Nugget effect	Partial sill	Range (m)
6045-10-2 (alightings)	Simple random	TLR	13.40			−1.75			0.38 (0.00)			
		RK	13.40			−1.75			−0.17 (0.07)	0.00	27.13	566.87
	Density of points	TLR	10.46			−1.96			0.32 (0.00)			
		RK	10.46			−1.96			−0.03 (0.90)	0.00	16.77	682.34
	Balanced and well-spread	TLR	7.81			−1.05			0.20 (0.00)			
		RK	7.81			−1.05			−0.10 (0.35)	0.00	7.25	743.89
	Extrapolation	TLR	7.33			−0.77			0.22 (0.00)			
		RK	7.33			−0.77			−0.02 (0.80)	0.00	4.41	461.27
6913-10-1 (boardings)	Simple random	TLR	7.95	−0.42					0.25 (0.00)			
		RK	7.95	−0.42					−0.00 (0.26)	0.00	7.59	1873.78
	Density of points	TLR	32.51	−2.56					0.09 (0.20)			
		RK	32.51	−2.56					—	—	—	—
	Balanced and well-spread	TLR	57.00	−4.81					−0.08 (0.54)			
		RK	57.00	−4.81					—	—	—	—
	Extrapolation	TLR	17.13		−0.88				0.22 (0.00)			
		RK	17.13		−0.88				−0.14 (0.06)	0.00	42.42	880.63
809L-10-2 (alightings)	Simple random	TLR	5.36					−1.08	0.15 (0.03)			
		RK	5.36					−1.08	−0.05 (0.85)	0.00	3.09	1017.89
	Density of points	TLR	7.96				1.66		0.11 (0.09)			
		RK	7.96				1.66		−0.03 (0.75)	7.76	2.97	1405.78
	Balanced and well-spread	TLR	4.90					−0.97	0.19 (0.00)			
		RK	4.90					−0.97	−0.01 (0.44)	0.00	2.70	1083.74
	Extrapolation	TLR	3.44					−0.25	−0.01 (0.70)			
		RK	3.44					−0.25	—	—	—	—
577T-10-1 (boardings)	Simple random	TLR	9.65	0.45					0.10 (0.03)			
		RK	9.65	0.45					−0.02 (0.75)	5.61	2.37	2015.27
	Density of points	TLR	9.54		−0.53				0.12 (0.01)			
		RK	9.54		−0.53				−0.01 (0.45)	5.10	3.07	9826.14
	Balanced and well-spread	TLR	11.68		−1.01				0.08 (0.04)			
		RK	11.68		−1.01				−0.01 (0.35)	15.01	7197278.00	$1.28E+10$
	Extrapolation	TLR	6.87		−0.41				0.22 (0.00)			
		RK	6.87		−0.41				−0.01 (0.60)	1.63	1.53	746.14

Note. TLR and RK express transformed linear regression and regression kriging, respectively.

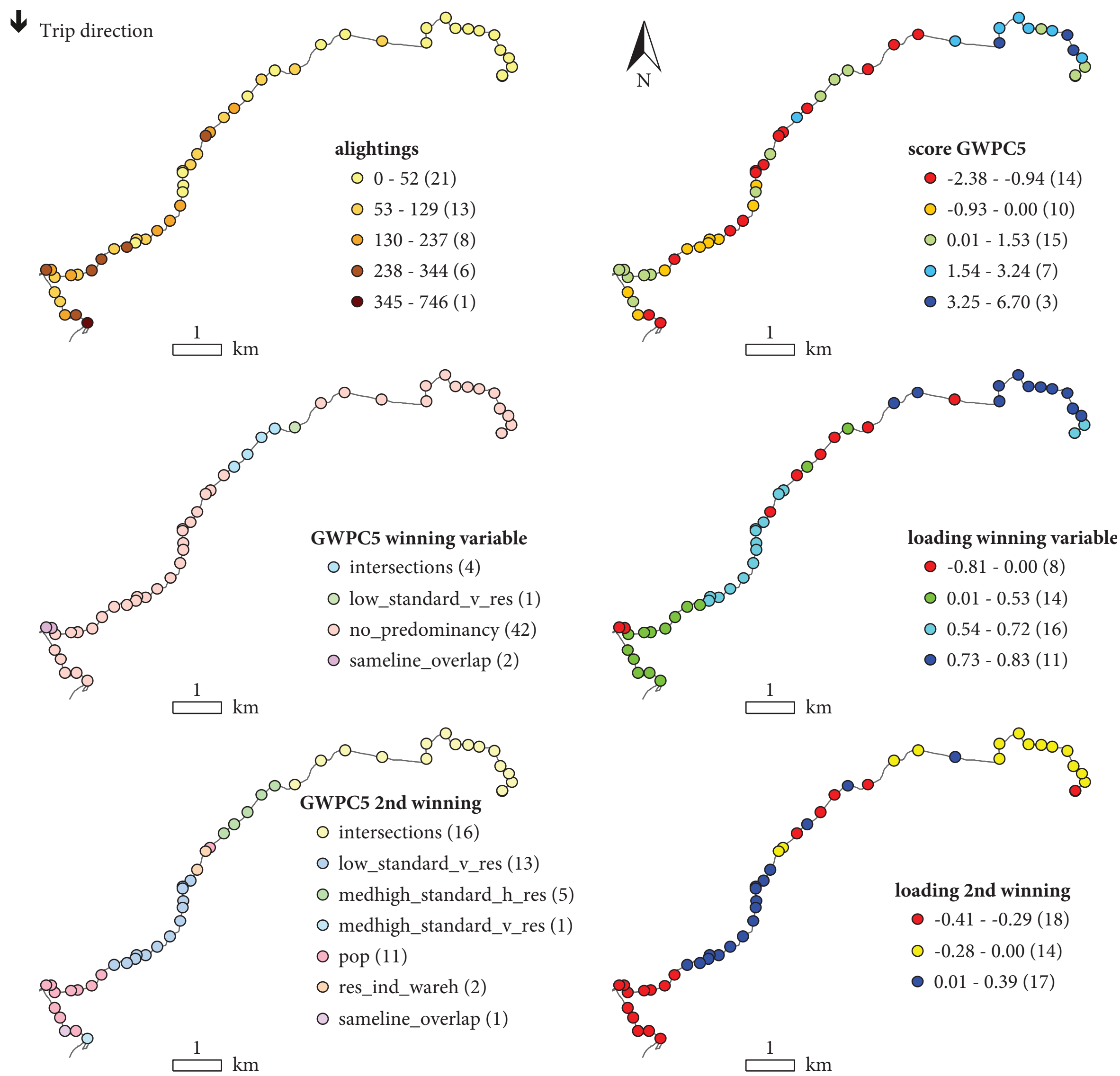

Figure 10: Spatial pattern of alightings and GWPC5 along line 6045-10-2.

The reason why RK did not perform better than TLR in some spatially dependent cases may be the uncertainty in the calculation of empirical and theoretical semivariograms. As no optimization procedure was used to obtain the parameters from these semivariograms, RK results may not be the optimum ones. Optimization techniques applied to kriging with network distances emerge as an interesting topic for future research.

As an effort to identify the sampling strategy having the best performance, we initially searched for the smallest error in each numeric column in Table 8, separated by the type of sample (calibration and validation). This procedure yielded 24 cases for calibration samples and 24 for validation ones. However, some of these cases had a number of elements lower than 4 due to the absence of RK results (blank spaces in Table 8). The RK modeling was not carried out for cases with no autocorrelation detected in the residuals from TLR. Maintaining only the 4-element comparisons, to allow a fair comparison among cases, 36 comparison groups (18 from calibration samples and 18 from validation ones) were listed. Afterward, we identified the sampling strategy corresponding to the smallest error in each group and summed the number of times each sampling method had the smallest error. Simple random, density of points, and extrapolation had the best performance in five calibration cases each, and the balanced and well-spread sampling stood out in three cases. Regarding the validation samples, the balanced and well-spread sampling showed the lowest error values in nine cases, that is, half of the analyzed cases. The simple random and extrapolation methods had the best performance in four cases each, and the density of points in only 1 case. In general, the balanced and well-spread sampling had consistently good results in both calibration and validation samples.

Splitting the comparison groups by bus line, it is more difficult to find a pattern of the best sampling method in both calibration and validation simultaneously. In most cases, some sampling strategies performed better in calibration and

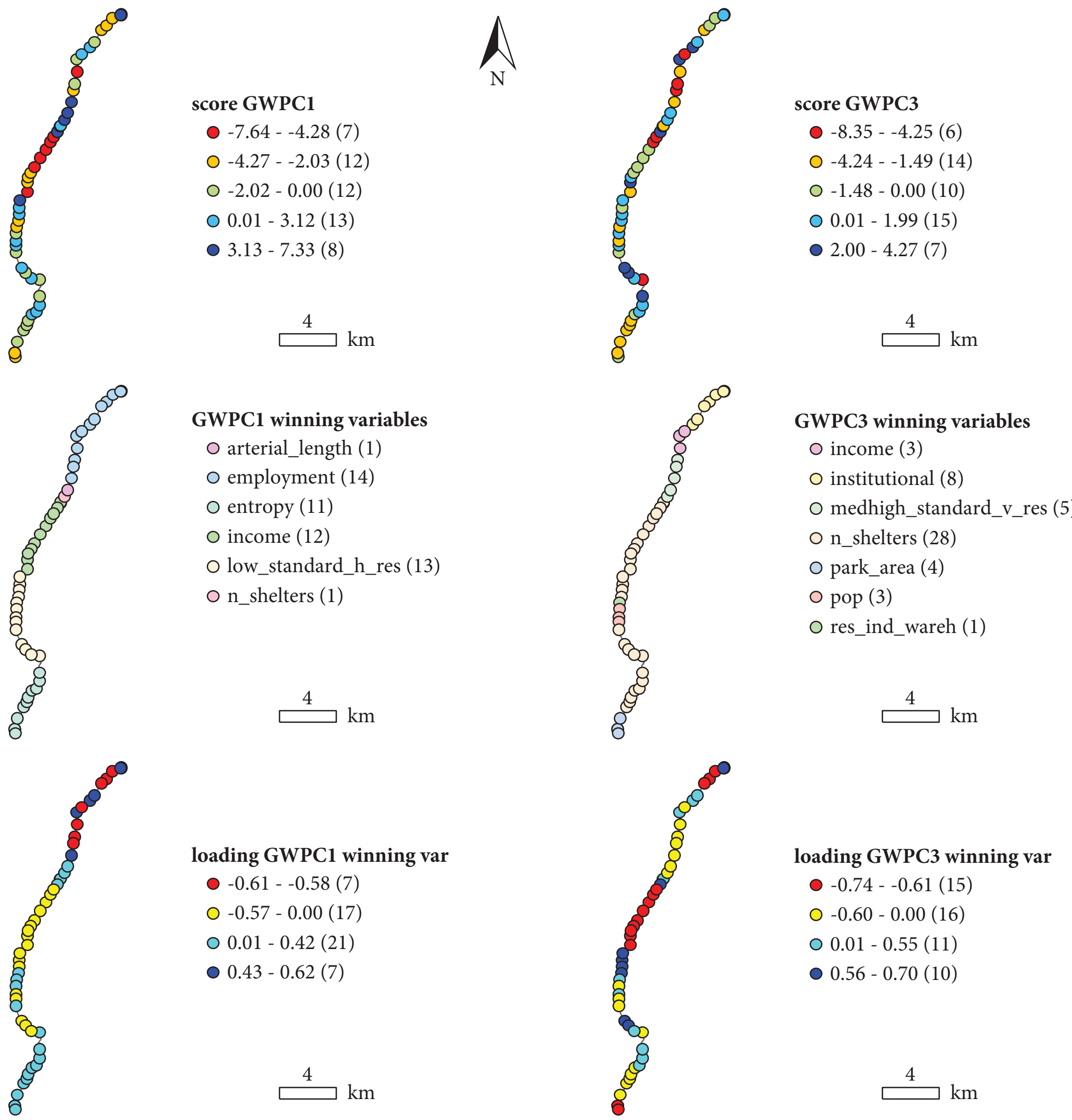

Figure 11: Spatial pattern of GWPC1 and GWPC3 along line 6913-10-1.

others in validation. However, analyzing the standard deviation of goodness-of-fit measures, we found that results from calibration samples of different sampling methods tended to show much less variation than validation samples. This reveals that the sampling strategy had a higher influence in the prediction accuracy of missing data compared to calibration data. In line 6045-10-2, the balanced and well-spread sample had the best validation results. The simple random sampling stood out in the validation results from lines 6913-10-1 and 809L-10-2. In line 577T-10-1, the extrapolation and balanced and well-spread sampling were the best ones in an equal number of times.

Although the sampling based on the density of points is able to reproduce the spatial concentration of data in the complete bus line dataset, one issue may arise from it: missing data points located in regions with a low density of calibration points will have no or a low number of sampled neighbors inside the autocorrelation range to be used in the estimation process (see, for example, Figure 6).

Another problem refers to the spatial variation of transit ridership data: in our case study, all bus stops with available data on both independent and dependent variables were used in the analysis, including points representing bus terminals. Terminals often have a passenger volume much higher than the adjacent neighbors. In the sample based on the density of points, this "outlier" point fell in the validation sample of the first three lines (Figures 6, 7, and 8), making it difficult for both RK and TLR to perform well as a large portion of variation in the dependent variable had not been accounted for when calibrating the models' parameters. This problem is also seen in the extrapolation sample of the first three lines, which are lines with a clear identification of the bus terminal, located at the beginning or the end of the route. However, the extrapolation sample had the best validation results in the last bus line (Figure 9), probably because large amounts of transit ridership are distributed along more than one bus stop on this route.

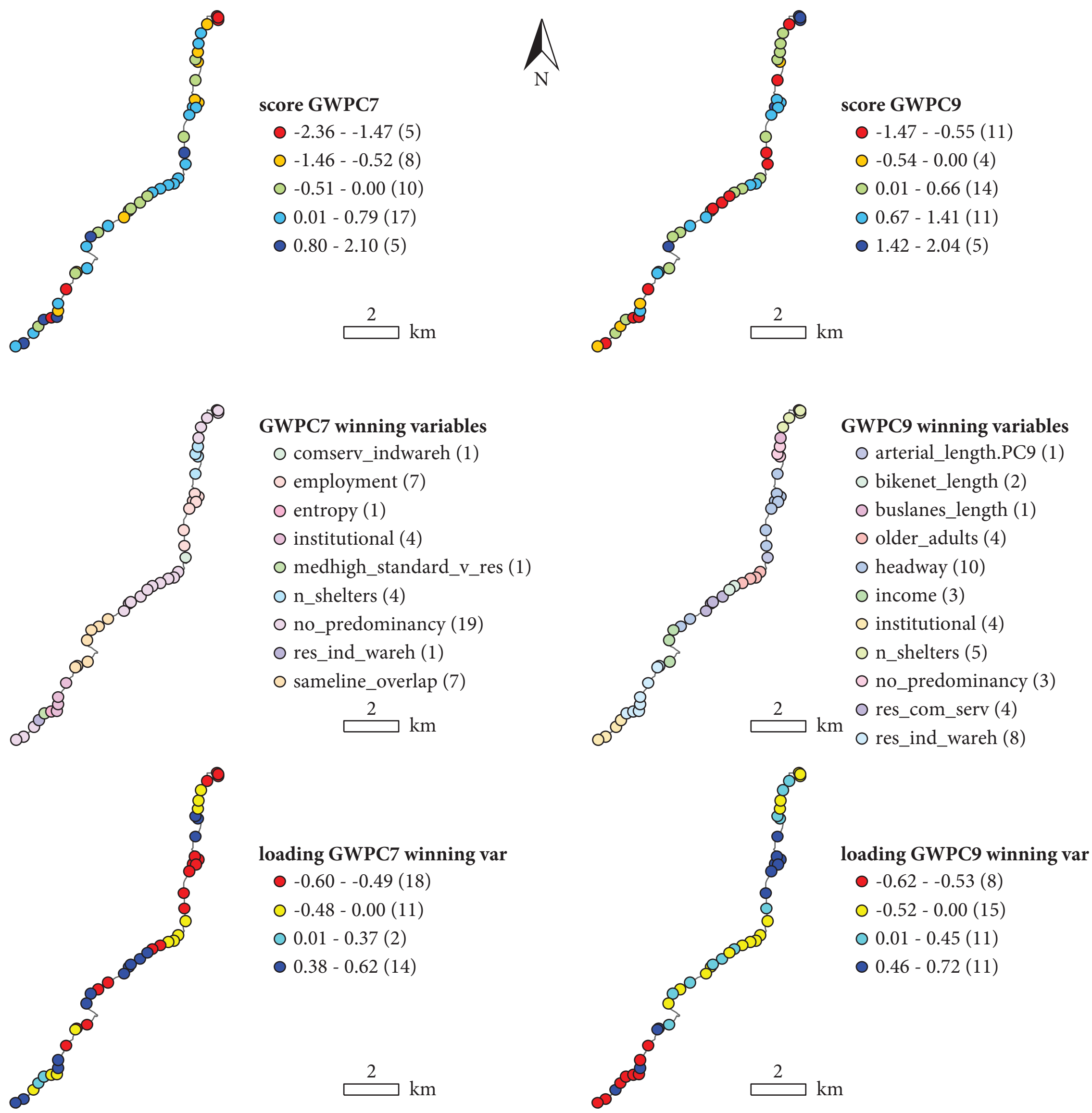

Figure 12: Spatial pattern of GWPC7 and GWPC9 along line 809L-10-2.

4.5. Comparison with Previous Studies. Comparison of this study with previous research can be done based on three main topics: dimensionality reduction, spatially varying effects, and goodness of fit. PC1, PC2, and PC3, from Lindner et al. [30], gather features from PC3, PC1, and PC2 from the present study, respectively. They used the first component (low-income population) as a predictor to model the transit ridership at a TAZ level based on Kriging with External Drift. However, as only sociodemographic features were included in the original dataset, the effect of bus service and transport system variables could not be accommodated. On the other hand, the winning variables from GWPCs in Table 7 (Figures 10, 11, 12, and 13) reveal an important influence of predictors, such as intersections, headway, and number of bus stop shelters on the transit ridership at some stops.

Varying the most important predictor from one point to another, as in GWPCA, is like having spatially varying effects in a geographically weighted regression. Winning variables in Figure 10, such as population, no predominant land use, intersections, medium-high standard horizontal residential area and overlapping, corroborates previous stop-level studies [10, 15], which have shown that effects from these predictors on transit ridership can vary spatially. However, the need to exclude highly correlated predictors resulted in a MedAPE of 33.72% and 34.45% from geographically weighted regressions applied to the alighting variable along line 6045-10-2 [15]. Both values are higher than the one from the current study (33.20%).

In addition, averaged MedAPE results in validation samples of the balanced and well-spread cases were lower than a 30% missing data scenario from Marques and

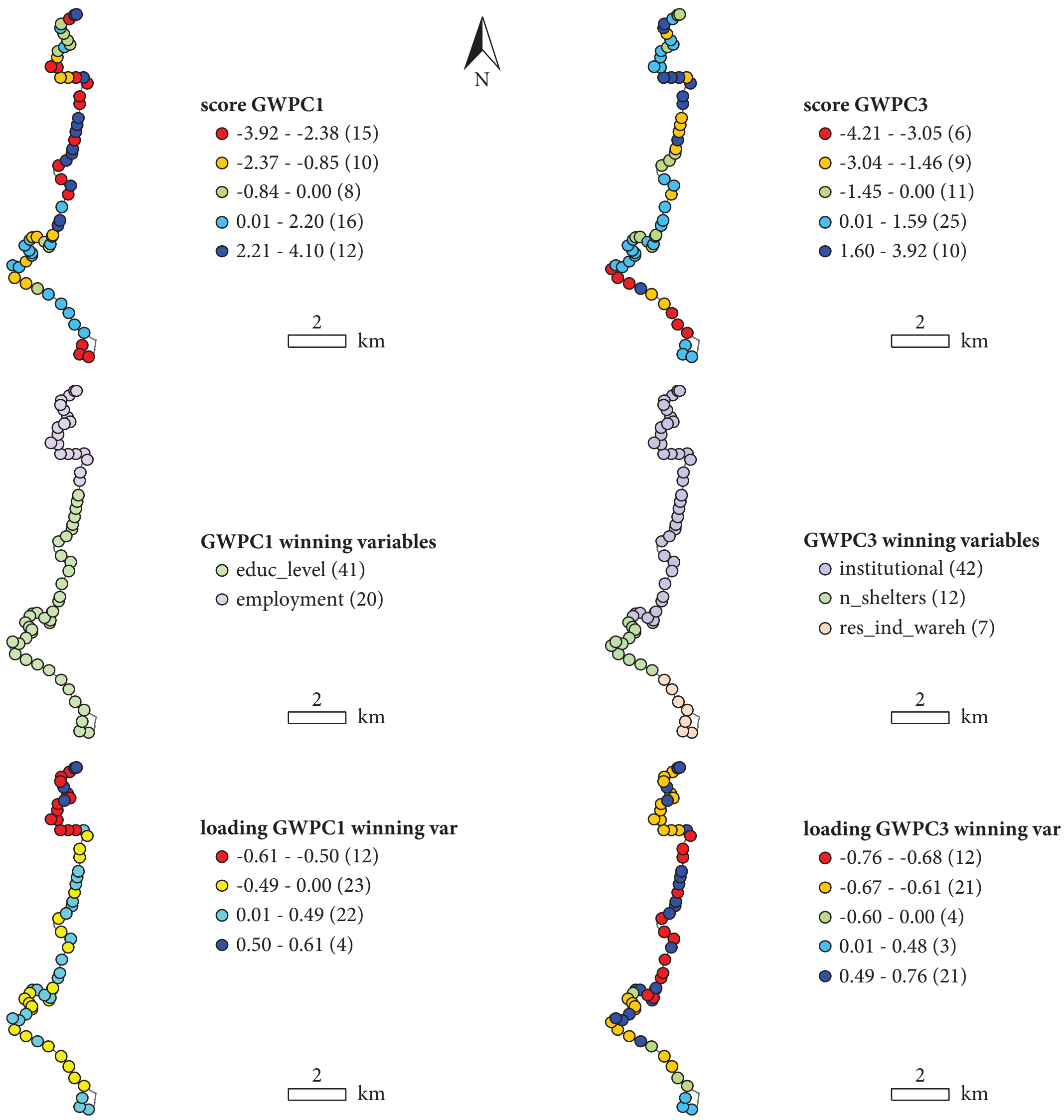

Figure 13: Spatial pattern of GWPC1 and GWPC3 along line 577T-10-1.

Table 8: Performance of four sampling methods in predicting transit ridership at the bus stop level.

Line	Sampling method	Sample	MedAPE (%) TLR	MedAPE (%) RK	RMSE TLR	RMSE RK	MAE TLR	MAE RK
6045-10-2	Simple	Calibration	52.55	**51.07**	82.83	**71.86**	65.91	**54.67**
	Simple	Validation	91.14	**57.89**	172.03	**159.19**	106.07	**76.12**
	Density	Calibration	56.96	57.80	89.14	**81.73**	69.29	**56.58**
	Density	Validation	67.09	**50.98**	165.66	**111.71**	97.20	**61.04**
	balanced_spread	Calibration	63.29	**52.31**	118.76	**102.86**	75.28	**55.94**
	balanced_spread	Validation	64.58	**33.20**	104.03	**77.15**	77.38	**45.94**
	Extrapolation	Calibration	56.63	**45.17**	82.85	**64.53**	65.59	**45.34**
	Extrapolation	Validation	70.15	**61.04**	180.10	**176.23**	107.77	**101.94**
6913-10-1	Simple	Calibration	75.03	**53.85**	1033.96	1216.37	350.30	437.18
	Simple	Validation	56.36	**45.61**	175.46	189.76	153.62	**146.90**
	Density	Calibration	56.33		164.45		135.01	
	Density	Validation	52.28		1524.79		576.33	
	balanced_spread	Calibration	47.80		137.34		117.06	
	balanced_spread	Validation	75.61		1521.23		611.19	
	Extrapolation	Calibration	41.16	**38.74**	173.46	**139.97**	135.57	**112.77**
	Extrapolation	Validation	39.46	51.90	1525.11	1526.19	530.93	542.54

Table 8: Continued.

Line	Sampling method	Sample	MedAPE (%) TLR	MedAPE (%) RK	RMSE TLR	RMSE RK	MAE TLR	MAE RK
809L-10-2	Simple	Calibration	50.83	72.33	243.47	286.44	90.44	121.80
	Simple	Validation	75.13	**61.48**	44.99	82.01	40.55	48.66
	Density	Calibration	48.94	57.72	62.78	63.09	40.72	**38.88**
	Density	Validation	46.02	63.46	368.53	**362.67**	147.13	168.35
	balanced_spread	Calibration	44.79	53.31	240.52	**234.03**	78.05	79.55
	balanced_spread	Validation	54.83	**48.11**	93.73	258.87	51.30	111.64
	Extrapolation	Calibration	47.11		87.42		44.87	
	Extrapolation	Validation	52.65		350.42		124.44	
577T-10-1	Simple	Calibration	41.18	**39.42**	51.13	**49.34**	39.01	**36.37**
	Simple	Validation	50.14	**40.06**	84.96	**75.89**	71.29	**60.38**
	Density	Calibration	34.66	**31.00**	52.72	**49.24**	41.54	**37.18**
	Density	Validation	40.83	54.92	60.18	65.66	45.62	45.88
	balanced_spread	Calibration	38.55	45.73	58.54	**55.97**	44.74	**41.66**
	balanced_spread	Validation	35.79	40.07	47.11	47.25	37.66	39.16
	Extrapolation	Calibration	43.04	53.10	59.37	60.09	43.74	**43.62**
	Extrapolation	Validation	33.55	34.80	48.47	**46.67**	40.76	**39.63**

Note. MedAPE, RMSE, MAE, TLR, and RK are, respectively, median of absolute percentage error, root mean squared error, mean absolute error, transformed linear regression, and regression kriging. The best results are in bold.

Pitombo [10], which used a geographically weighted regression and a sample based on the density of points. These outcomes were also better than the validation MedAPE results from one of the 30% missing data scenarios analyzed by Marques et al. [46], which, again, applied the point-density sampling method, but used Ordinary Kriging for prediction. This indicates a good performance of both RK (GWPCA coupled with OK) and the balanced and well-spread sampling over other modeling approaches and sampling methods, respectively.

5. Conclusions and Final Considerations

This study proposed a two-step method based on Geographically Weighted Principal Component Analysis and kriging interpolation to predict the number of boardings and alightings in uncounted bus stops, considering the effect of the sampling strategy. GWPCA was carried out using all bus stops in São Paulo (Brazil), and the outcomes of it served as an input to a regression modeling accounting for the spatial dependence of the stop-level ridership data.

Outcomes from the spatial PCA can be useful to travel demand modeling in two ways: (1) by highlighting the most important intervening variables even in points with no ridership data, and (2) by acting as a predictor to the travel demand estimation in unsampled points. In our case study, the contribution of spatial interpolation was higher in the missing points than in the calibration ones. In addition, validation results were more sensitive to the sampling strategy compared to the calibration results. When selecting the most appropriate sampling design, the spatial pattern of transit ridership data may play an important role. The balanced sampling with geographic spreading had the best validation results in bus lines with different spatial distributions of stop-level passenger volume. The simple random sampling appears as a possible solution when no knowledge on the most correlated predictor is available. In turn, extrapolation could be recommended for cases where extreme data values are not highly concentrated in the spatial field considered. Although only four lines could be selected to the case study, they were able to reproduce spatial patterns of transit ridership common to various bus lines in the city.

The method proposed is not restricted to stop-level ridership cases. It can successfully support predicting missing data in other geographic units and travel demand variables. An advantage of GWPCA is the fact that, once it is generated, it can be used as a basis for various additional analyses, such as classification, clustering, and creation of indexes. Exploring other contributions of GWPCA is recommended in future research.

Data Availability

The datasets used to support the findings of this study are included in supplementary materials. The datasets used to collect the predictors can be found on the GeoSampa website (https://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx), the 2017 Origin and Destination Survey website (https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino), and the SPTrans website (https://www.sptrans.com.br/desenvolvedores/). The transit ridership dataset analyzed during the current study is not publicly available due to the fact that the data are held by SPTrans but can be requested from it through the Electronic Citizen Information System (https://esic.prefeitura.sp.gov.br/Account/Login.aspx).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The authors would like to thank *SPTrans* for the Boarding/Alighting survey data used in this study. This research was supported by the National Council of Technological and Scientific Development (Grant #403508/2023 and Grant #305973/2023-1, Brazil) and the São Paulo Research Foundation (Grant #2019/12054-4, Brazil).

Supplementary Materials

Four txt. files have been provided as supplementary material. "predictor data.txt" contains the predictor data from 19,329 stops used in the case study. "gwpca_scores.txt" provides the GWPCA scores for the ten components retained. "gwpc1to5_loadings.txt" and "gwpc6to10_loadings.txt" present the loading values for each predictor in the ten GWPCs retained. In this case, loading values were obtained for 19,900 stops in São Paulo. (*Supplementary Materials*)

References

[1] V. R. Vuchic, *Urban Transit: Operations, Planning, and Economics*, John Wiley & Sons, Hoboken, NJ, USA, 2005.

[2] A. Ceder, *Public Transit Planning and Operation: Modeling, Practice and Behavior*, CRC Press, Boca Raton, FL, USA, 2007.

[3] Trb, "Stop, station, and terminal capacity in: transit capacity and quality of service manual," 2003, https://onlinepubs.trb.org/onlinepubs/tcrp/docs/tcrp100/Part7.pdf.

[4] V. Ceccato, O. Cats, and Q. Wang, in *The Geography of Pickpocketing at Bus Stops: An Analysis of Grid Cells BT - Safety and Security in Transit Environments: An Interdisciplinary Approach*, V. Ceccato and A. Newton, Eds., pp. 76–98, Palgrave Macmillan, London, UK, 2015.

[5] R. Zahnow and J. Corcoran, "Crime and bus stops: an examination using transit smart card and crime data," *Environment and Planning B: Urban Analytics and City Science*, vol. 48, no. 4, pp. 706–723, 2019.

[6] Q. Miao, E. W. Welch, and P. S. Sriraj, "Extreme weather, public transport ridership and moderating effect of bus stop shelters," *Journal of Transport Geography*, vol. 74, pp. 125–133, 2019.

[7] V. Chakour and N. Eluru, "Examining the influence of stop level infrastructure and built environment on bus ridership in Montreal," *Journal of Transport Geography*, vol. 51, pp. 205–217, 2016.

[8] B. Cui, J. DeWeese, H. Wu, D. A. King, D. Levinson, and A. El-Geneidy, "All ridership is local: accessibility, competition, and stop-level determinants of daily bus boardings in Portland, Oregon," *Journal of Transport Geography*, vol. 99, 2022.

[9] K. Kerkman, K. Martens, and H. Meurs, "Factors influencing stop-level transit ridership in arnhem–nijmegen city region, Netherlands," *Transportation Research Record*, vol. 2537, no. 1, pp. 23–32, 2015.

[10] S. de F. Marques and C. S. Pitombo, "Local modeling as a solution to the lack of stop-level ridership data," *Journal of Transport Geography*, vol. 112, 2023.

[11] S. S. Pulugurtha and M. Agurla, "Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods," *Journal of Public Transportation*, vol. 15, no. 1, pp. 33–52, 2012.

[12] M. Rahman, S. Yasmin, A. Faghih-Imani, and N. Eluru, "Examining the bus ridership demand: application of spatio-temporal panel models," *Journal of Advanced Transportation*, vol. 2021, Article ID 8844743, 10 pages, 2021.

[13] X. Chu, "Ridership models at the stop level," *National Center for Transit Research*, University of South Florida, FL, USA, 2004.

[14] J. Dill, M. Schlossberg, L. Ma, and C. Meyer, "Predicting transit ridership at the stop level: the role of service and urban form," in *Proceedings of the Transportation Research Board 92nd Annual Meeting*, pp. 1–19, Washington, DC, USA, January 2013.

[15] S. de F. Marques and C. S. Pitombo, "Transit ridership modeling at the bus stop level: comparison of approaches focusing on count and spatially dependent data," *Applied Spatial Analysis and Policy*, vol. 16, no. 1, pp. 277–313, 2023.

[16] S. de F. Marques and C. S. Pitombo, "Applying multivariate Geostatistics for transit ridership modeling at the bus stop level," *Boletim de Ciências Geodésicas*, vol. 27, no. 2, 2021.

[17] S. de F. Marques and C. S. Pitombo, "Spatial modeling of transit ridership along bus lines with overlapping sections," in *Proceedings of the 35o Annual Conference of the Brazilian National Association for Transportation Research and Teaching. Brazilian National Association for Transportation Research and Teaching*, pp. 1568–1580, Virtual, January 2021b.

[18] X. Shi, A. V. Moudon, P. M. Hurvitz, S. J. Mooney, C. Zhou, and B. E. Saelens, "Does improving stop amenities help increase Bus Rapid Transit ridership? Findings based on a quasi-experiment," *Transportation Research Interdisciplinary Perspectives*, vol. 10, 2021.

[19] S. Mathew and S. S. Pulugurtha, "Comparative assessment of geospatial and statistical methods to estimate local road annual average daily traffic," *Journal of Transportation Engineering, Part A: Systems*, vol. 147, no. 7, 2021.

[20] S. S. Pulugurtha and S. Mathew, "Modeling AADT on local functionally classified roads using land use, road density, and nearest nonlocal road data," *Journal of Transport Geography*, vol. 93, 2021.

[21] Y. Song, X. Wang, G. Wright, D. Thatcher, P. Wu, and P. Felix, "Traffic volume prediction with segment-based regression kriging and its implementation in assessing the impact of heavy vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 232–243, 2019.

[22] X. Ma, J. Zhang, C. Ding, and Y. Wang, "A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership," *Computers, Environment and Urban Systems*, vol. 70, pp. 113–124, 2018.

[23] W. Tu, R. Cao, Y. Yue, B. Zhou, Q. Li, and Q. Li, "Spatial variations in urban public ridership derived from GPS trajectories and smart card data," *Journal of Transport Geography*, vol. 69, pp. 45–57, 2018.

[24] S. Blainey and C. Mulley, "Using geographically weighted regression to forecast rail demand in the Sydney region," in *Proceedings of the Australasian Transport Research Forum 2013*, Brisbane, Australia, October 2013.

[25] O. D. Cardozo, J. C. García-Palomares, and J. Gutiérrez, "Application of geographically weighted regression to the direct forecasting of transit ridership at station-level," *Applied Geography*, vol. 34, pp. 548–558, 2012.

[26] Y. Liu, Y. Ji, Z. Shi, and L. Gao, "The influence of the built environment on school children's metro ridership: an exploration using geographically weighted Poisson regression models," *Sustainability*, vol. 10, no. 12, p. 4684, 2018.

[27] S. Kim, S. Park, and K. Jang, "Spatially-varying effects of built environment determinants on walking," *Transportation Research Part A: Policy and Practice*, vol. 123, pp. 188–199, 2019.

[28] R. A. Mucci and G. D. Erhardt, "Evaluating the ability of transit direct ridership models to forecast medium-term ridership changes: evidence from san francisco," *Transportation Research Record*, vol. 2672, no. 46, pp. 21–30, 2018.

[29] S. de F. Marques and C. S. Pitombo, "Ridership estimation along bus transit lines based on kriging: comparative analysis between network and euclidean distances," *Journal of Geovisualization and Spatial Analysis*, vol. 5, no. 1, p. 7, 2021.

[30] A. Lindner, C. S. Pitombo, S. S. Rocha, and J. A. Quintanilha, "Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study," *Geospatial Information Science*, vol. 19, no. 4, pp. 245–254, 2016.

[31] T. Basu and A. Das, "Formulation of deprivation index for identification of regional pattern of deprivation in rural India," *Socio-Economic Planning Sciences*, vol. 74, 2021.

[32] A. J. Comber, P. Harris, and N. Tsutsumida, "Improving land cover classification using input variables derived from a geographically weighted principal components analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 119, pp. 347–360, 2016.

[33] C. D. Lloyd, "Analysing population characteristics using geographically weighted principal components analysis: a case study of Northern Ireland in 2001," *Computers, Environment and Urban Systems*, vol. 34, no. 5, pp. 389–399, 2010.

[34] N. Losada, E. Alén, T. R. Cotos-Yáñez, and T. Domínguez, "Spatial heterogeneity in Spain for senior travel behavior," *Tourism Management*, vol. 70, pp. 444–452, 2019.

[35] C. Wu, N. Peng, X. Ma, S. Li, and J. Rao, "Assessing multiscale visual appearance characteristics of neighbourhoods using geographically weighted principal component analysis in Shenzhen, China," *Computers, Environment and Urban Systems*, vol. 84, 2020.

[36] C. Wu, W. Hu, M. Zhou, S. Li, and Y. Jia, "Data-driven regionalization for analyzing the spatiotemporal characteristics of air quality in China," *Atmospheric Environment*, vol. 203, pp. 172–182, 2019.

[37] N. S. Ngo, "Urban bus ridership, income, and extreme weather events," *Transportation Research Part D: Transport and Environment*, vol. 77, pp. 464–475, 2019.

[38] K. Lanza and C. P. Durand, "Heat-Moderating effects of bus stop shelters and tree shade on public transport ridership," *International Journal of Environmental Research and Public Health*, vol. 18, no. 2, p. 463, 2021.

[39] J. K. Eom, M. S. Park, T.-Y. Heo, and L. F. Huntsinger, "Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1968, pp. 20–29, 2006.

[40] X. Wang and K. Kockelman, "Forecasting network data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2105, no. 1, pp. 100–108, 2009.

[41] B. Selby and K. M. Kockelman, "Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression," *Journal of Transport Geography*, vol. 29, pp. 24–32, 2013.

[42] G. Sarlas and K. W. Axhausen, "Prediction of AADT on a nationwide network based on an accessibility-weighted centrality measure," *Arbeitsberichte Verkehrs- und Raumplanung*, vol. 1094, 2015.

[43] S. Kim, D. Park, T.-Y. Heo, H. Kim, and D. Hong, "Estimating vehicle miles traveled (VMT) in urban areas using regression kriging," *Journal of Advanced Transportation*, vol. 50, no. 5, pp. 769–785, 2016.

[44] T. J. Klatko, S. T. Usman, V. Matthew, L. Samuel, F. J. D, and S. K. C, "Addressing the local-road VMT estimation problem using spatial interpolation techniques," *Journal of Transportation Engineering, Part A: Systems*, vol. 143, no. 8, 2017.

[45] H. Yang, J. Yang, L. D. Han et al., "A Kriging based spatiotemporal approach for traffic volume data imputation," *PLoS One*, vol. 13, no. 4, p. e0195957, 2018.

[46] S. de F. Marques, R. Favero, and C. S. Pitombo, "Should we account for network distances or anisotropy in the spatial estimation of missing traffic data?" *Transport*, vol. 31, no. 1, p. e2822, 2023.

[47] G. Chi and Y. Zheng, "Estimating transport footprint along highways at local levels: a combination of network analysis and kriging methods," *International Journal of Sustainable Transportation*, vol. 7, no. 3, pp. 261–273, 2013.

[48] B. Shamo, E. Asa, and J. Membah, "Linear spatial interpolation and analysis of annual average daily traffic data," *Journal of Computing in Civil Engineering*, vol. 29, no. 1, 2015.

[49] Y. C. Chiou, R. C. Jou, and C. H. Yang, "Factors affecting public transportation usage rate: geographically weighted regression," *Transportation Research Part A: Policy and Practice*, vol. 78, pp. 161–177, 2015.

[50] S. Blainey and J. Preston, "A geographically weighted regression based analysis of rail commuting around Cardiff, South Wales," in *Proceedings of the 12th World Conference on Transport Research*, Lisbon, Portugal, July 2010.

[51] Y. Zhu, F. Chen, Z. Wang, and J. Deng, "Spatio-temporal analysis of rail station ridership determinants in the built environment," *Transportation*, vol. 46, no. 6, pp. 2269–2289, 2019.

[52] V. A. Gomes, C. S. Pitombo, S. S. Rocha, and A. R. Salgueiro, "Kriging geostatistical methods for travel mode choice: a spatial data analysis to travel demand forecasting," *Open Journal of Statistics*, vol. 06, pp. 514–527, 2016.

[53] J. Chica-Olmo, C. Rodríguez-López, and P. Chillón, "Effect of distance from home to school and spatial dependence between homes on mode of commuting to school," *Journal of Transport Geography*, vol. 72, pp. 1–12, 2018.

[54] D. Zhang and X. C. Wang, "Transit ridership estimation with network Kriging: a case study of Second Avenue Subway, NYC," *Journal of Transport Geography*, vol. 41, pp. 107–115, 2014.

[55] Ibge, *São Paulo*, IBGE cities, Geneva, Switzerland, 2021.

[56] Metrô, *Origin and Destination Survey*, Metrô, Mumbai, India, 2019.

[57] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society-Series B: Statistical Methodology*, vol. 26, no. 2, pp. 211–243, 1964.

[58] F. Zhao, L. F. Chow, M. T. Li, I. Ubaka, and A. Gan, "Forecasting transit walk accessibility: regression model alternative to buffer method," *Transportation Research Record*, vol. 1835, no. 1, pp. 34–41, 2003.

[59] A. B. Morelli, A. de C. Fiedler, and A. L. Cunha, "Um banco de dados de empregos formais georreferenciados em cidades brasileiras," 2023, https://arxiv.org/abs/2303.09602.

[60] H. F. Kaiser, "A second generation little jiffy," *Psychometrika*, vol. 35, no. 4, pp. 401–415, 1970.

[61] M. S. Bartlett, "Tests of significance in factor analysis," *British Journal of Statistical Psychology*, vol. 3, no. 2, pp. 77–85, 1950.

[62] C. D. Dziuban and E. C. Shirkey, "When is a correlation matrix appropriate for factor analysis? Some decision rules," *Psychological Bulletin*, vol. 81, no. 6, pp. 358–361, 1974.

[63] H. F. Kaiser and J. Rice, "Little jiffy, mark IV," *Educational and Psychological Measurement*, vol. 34, no. 1, pp. 111–117, 1974.
[64] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, Cham, 2002.
[65] P. Harris, C. Brunsdon, and M. Charlton, "Geographically weighted principal components analysis," *International Journal of Geographical Information Science*, vol. 25, no. 10, pp. 1717–1736, 2011.
[66] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Economic Geography*, vol. 46, pp. 234–240, 1970.
[67] P. A. P. Moran, "The interpretation of statistical maps," *Journal of the Royal Statistical Society-Series B: Statistical Methodology*, vol. 10, no. 2, pp. 243–251, 1948.
[68] N. A. C. Cressie, *Statistics for Spatial Data*, John Wiley & Sons, Inc, Hoboken, NJ, USA, 1993.
[69] G. Matheron, "Principles of geostatistics," *Economic Geology*, vol. 58, no. 8, pp. 1246–1266, 1963.
[70] G. Matheron, *The Theory of Regionalized Variables and its Applications*, Les Cahiers du Centre de Morphologie Mathematique in Fontainebleu, Paris, 1971.
[71] R. A. Olea, "A six-step practical approach to semivariogram modeling," *Stochastic Environmental Research and Risk Assessment*, vol. 20, no. 5, pp. 307–318, 2006.
[72] T. Hengl, G. B. M. Heuvelink, and D. G. Rossiter, "About regression-kriging: from equations to case studies," *Computers & Geosciences*, vol. 33, no. 10, pp. 1301–1315, 2007.
[73] I. O. A. Odeh, A. B. McBratney, and D. J. Chittleborough, "Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging," *Geoderma*, vol. 67, no. 3-4, pp. 215–226, 1995.
[74] Y. Hollander and R. Liu, "The principles of calibrating traffic microsimulation models," *Transportation*, vol. 35, no. 3, pp. 347–362, 2008.
[75] D. J. Brus, *Spatial Sampling with R*, CRC Press, Boca Raton, FL, USA, 2022.
[76] J. S. Evans, "spatialEco. R package version," 2021, https://cran.r-project.org/web/packages/spatialEco/index.html.
[77] V. A. Profillidis and G. N. Botzoris, "Statistical methods for transport demand modeling," in *Modeling of Transport Demand*, B. Romer, Ed., pp. 163–224, Elsevier, Amsterdam, The Netherlands, 2019.
[78] R Core Team, "R: a language and environment for statistical computing," 2021, https:///R:%20A%20language%20and%20environment%20for%20statistical%20computing.
[79] I. Gollini, B. Lu, M. Charlton, C. Brunsdon, and P. Harris, "GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models," *Journal of Statistical Software*, vol. 63, no. 17, pp. 1–50, 2015.
[80] B. Lu, P. Harris, M. Charlton, and C. Brunsdon, "The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models," *Geospatial Information Science*, vol. 17, no. 2, pp. 85–101, 2014.
[81] S. P. Millard, *EnvStats*, Springer eBooks, Cham, 2013.
[82] E. Paradis, J. Claude, and K. Strimmer, "APE: analyses of phylogenetics and evolution in R language," *Bioinformatics*, vol. 20, no. 2, pp. 289-290, 2004.
[83] D. Bundala, W. Bergenheim, and M. Metz, *v.net.allpairs-Computes the Shortest Path between All Pairs of Nodes in the Network*, GRASS GIS, San Diego, CA, USA, 2022.
[84] J. M. Ver Hoef, "Kriging models for linear networks and non-Euclidean distances: Cautions and solutions," *Methods in Ecology and Evolution*, vol. 9, pp. 1600–1613, 2018.
[85] A. Grafström and J. Lisic, "Balanced Sampling: balanced and spatially balanced sampling R package version," 2019, https://cran.r-project.org/web/packages/BalancedSampling/index.html.