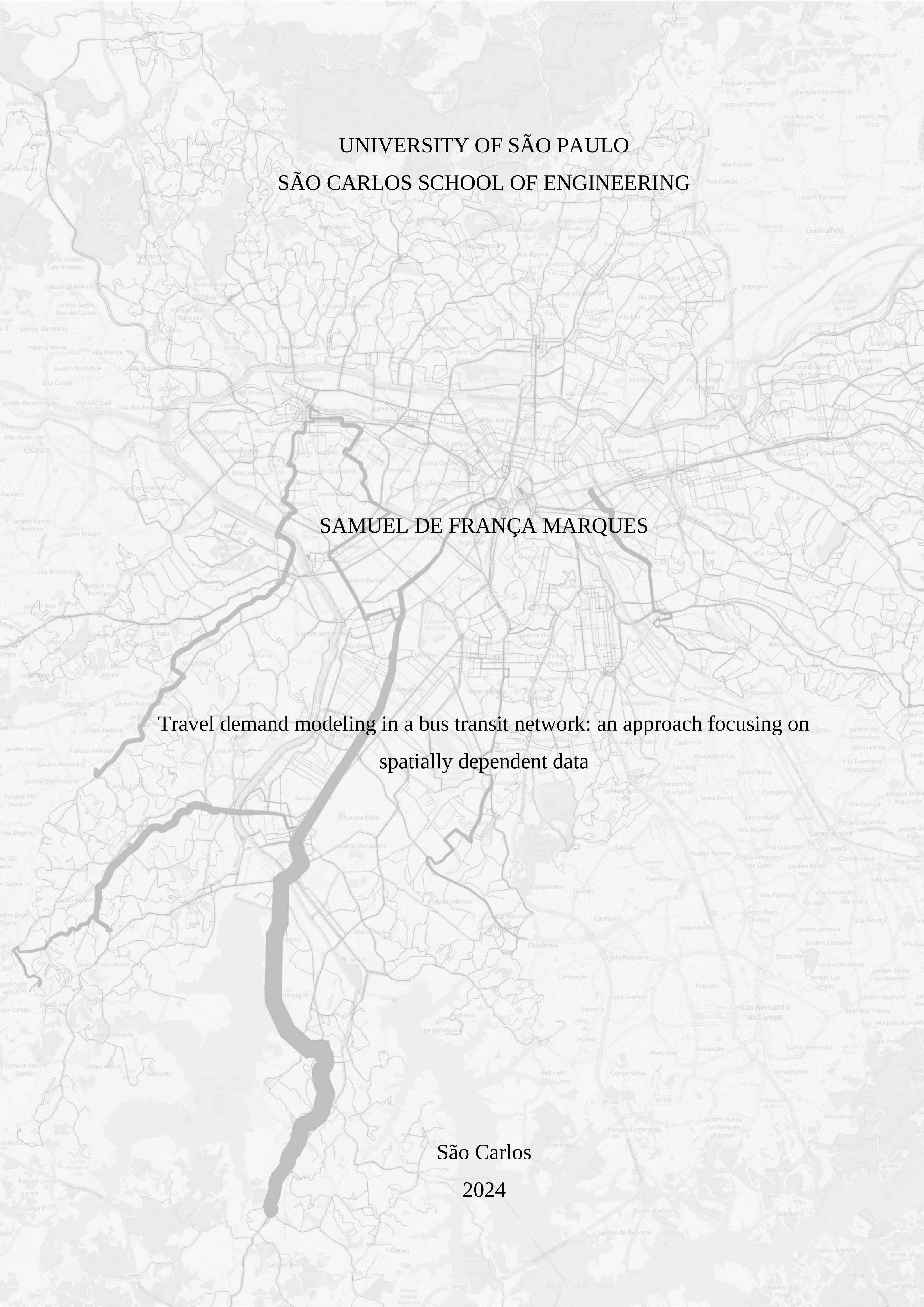
UNIVERSIDADE DE SÃO PAULO
Escola de Engenharia de São Carlos

TRAVEL DEMAND MODELING IN A BUS TRANSIT NETWORK: AN APPROACH FOCUSING ON SPATIALLY DEPENDENT DATA

SAMUEL DE FRANÇA MARQUES

EESC · USP



UNIVERSITY OF SÃO PAULO

SÃO CARLOS SCHOOL OF ENGINEERING

SAMUEL DE FRANÇA MARQUES

Travel demand modeling in a bus transit network: an approach focusing on spatially dependent data

São Carlos

2024

SAMUEL DE FRANÇA MARQUES

Travel demand modeling in a bus transit network: an approach focusing on spatially dependent data

Thesis presented to the São Carlos School of Engineering of the University of São Paulo in partial fulfillment of the requirements for the degree of Doctor of Science in Transportation Engineering.

Subject area: Planning and Operation of Transport Systems

Advisor: Prof. Cira Souza Pitombo.

CORRECTED VERSION

São Carlos
2024

Catalog card prepared by the Library Prof. Sérgio Rodrigues
Fontes of EESC/USP

Marques, Samuel de França.
M357t Travel demand modeling in a bus transit network: an approach focusing on spatially dependent data. / Samuel de França Marques ; advisor Cira Souza Pitombo. São Carlos, 2024.

Thesis(Doctorate - Program in Transportation Engineering and Concentration area of the Planning and Operation of Transport Systems)-- Escola de Engenharia de São Carlos of Universidade de São Paulo, 2024.

1. Transit ridership. 2. Kriging. 3. Geographically Weighted Regression. 4. Count data. 5. Bus stop. 6. Missing data. I. Title.

FOLHA DE JULGAMENTO

Candidato: Bacharel **SAMUEL DE FRANÇA MARQUES.**

Título da tese: "Modelagem da demanda em rede de transporte público por ônibus: uma abordagem com foco em dados espacialmente dependentes".

Data da defesa: 03/07/2023.

Comissão Julgadora	**Resultado**
Profa. Associada **Cira Souza Pitombo** **(Orientadora)** (Escola de Engenharia de São Carlos – EESC/USP)	Aprovado
Prof. Dr. **Jaime Gomez Hernandez** (Universitat Politècnica de València/UPV)	Aprovado
Profa. Dra. **Mariana Abrantes Giannotti** (Escola Politécnica/EP-USP)	Aprovado
Profa. Dra. **Renata Lúcia Magalhães de Oliveira** (Centro Federal de Educação Tecnológica de Minas Gerais/CEFET-MG)	Aprovado
Prof. Titular **Antônio Nelson Rodrigues da Silva** (Escola de Engenharia de São Carlos – EESC/USP)	Aprovado

Coordenador do Programa de Pós-Graduação em Engenharia de Transportes:
Profa. Associada **Ana Paula Camargo Larocca**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Carlos De Marqui Junior**

DEDICATION

To God

ACKNOWLEDGMENTS

The evolution in research that can be seen throughout the papers I wrote overflowed into my personal life. Over the last four years, I have been through many growing pains that contributed to disclosing the real me. The increasing level of commitment required to conduct my research could only be met by a whole me. Therefore, I imagine life knew what it was doing. Fortunately, I survived. The articles reflect my history. Their content embeds this maturity process. Now, I can only thank the beautiful souls that were part of this.

My best friend, God, the architect of it all, always putting me on the right track. I shared with Him all moments of joy and sadness. He made me find real peace. Incidentally, connection with faith is what inspires me to adopt people-focused work when carrying out research.

Doralice and Hadenízio, also known by their superhero names of mom and dad. They are my fist examples of hard workers, always supporting and inspiring me to do my best. Thank you for understanding my moments of absence.

Sara, Débora and Daniela, my perfect sisters, and my precious nephew, André. They are the ones who root for me the most and I am deeply grateful for that.

Cira, who is more than a supervisor, a real partner, someone who I can always count on.

Jorge, Murilo, Godfred, Matheus R., Renata, Sara, Renan, Heber, Matheus F., people that made me believe one more time that everyone has a purpose in a friend’s life.

To my friends from the Department of Transportation Engineering. It was six years (master’s degree plus Ph.D.) of happy encounters and productive conversations.

To Professor Jaime, who kindly received me at the Universitat Politècnica de València.

To the Christian Community Oikos, represented by pastor André Matheus. The message of being a positive influence in society is something that I will never forget.

To all my masters since kindergarten. Their passion for teaching, despite not being properly appreciated sometimes, is what allows people like me to reach the position I am in today.

To the University of São Paulo, the São Carlos School of Engineering and the Department of Transportation Engineering, for all the infrastructure provided for the research development.

This study was supported by three funding agencies: the São Paulo Research Foundation (FAPESP, grant #2019/12054-4, Brazil), the National Council for Scientific and Technological

Development (CNPq, grant #142082/2019-8, Brazil), and in part by the Coordination for the Improvement of Higher Education Personnel (CAPES, Brazil) – Grant Finance Code 001.

To the *São Paulo Transporte S.A.* for providing the Boarding/Alighting survey data used in this study.

This list is not exhaustive. I would like to thank all the people that, even in a small way, had a positive influence on my life. May you feel embraced.

EPIGRAPH

"…sometimes I've believed as many as six impossible things before breakfast."
Alice in Wonderland by Lewis Carroll

"Whatever you do, do it wholeheartedly as though you were working for your real master and not merely for humans"
Colossians 3 23

ABSTRACT

MARQUES, S. de F. **Travel demand modeling in a bus transit network**: an approach focusing on spatially dependent data. 2024. Thesis (Doctor in Science) - São Carlos School of Engineering, University of São Paulo, São Carlos, 2024.

Boarding and alighting per bus stop modeling, along bus lines, plays a fundamental role in Transport Public network planning, in addition to contributing to transit-oriented development. However, in this case, the variables of interest (boarding and alighting) show four characteristics that have implications for the modeling process and/or may affect the estimates' results. They are: (1) spatial dependence; (2) asymmetry; (3) the trips occur along the transport network; and (4) limited sample. As these peculiarities are overlooked by classical modeling and the scientific literature has not addressed them so far simultaneously, the main objective of the present study was to model transit ridership at the bus stop level, including, in the estimating process, the four aforementioned characteristics. As specific objectives, we analyzed the effect of explanatory variables, the effects of using network or Euclidean distances, the amount of missing data, and the sampling type. The text is divided into seven chapters, and four of them are articles, whose contents address one or more specific objectives separately and/or simultaneously. Two classes of spatial models were proposed, geostatistical interpolators and geographically weighted regressions, and a database comprising eight lines in the city of São Paulo (Brazil) was used as a case study. The following conclusions were achieved: models that consider asymmetry and spatial dependence should be prioritized over the ones that overlook these characteristics, as well as the multivariate models over the univariate ones. Public policies toward increasing public transport usage should focus mainly on four groups of explanatory variables: sociodemographic characteristics, bus network coverage, street layout and land use. The spatial models proved to be able to estimate the volume of boardings and alightings in unsampled points accurately, solving the problem of a lack of stop-level transit ridership data. Despite the network distance approach not contributing significantly to improving the models' prediction power, this type of distance may better represent the relationship between the transit ridership and its intervening factors. Prioritizing the use of network distances in the spatial modeling of boardings and alightings is recommended. In addition, a balanced sample on predictor data and well-spread in the geographic space might be preferred to accurately estimate missing stop-level ridership data. Therefore, the present research adds important methodological and practical contributions to the urban planning associated with sustainable transport.

Keywords: Transit ridership. Kriging. Geographically Weighted Regression. Count data. Bus stop. Missing data.

RESUMO

MARQUES, S. de F. **Modelagem da demanda em rede de transporte público por ônibus**: uma abordagem com foco em dados espacialmente dependentes. 2024. Tese (Doutorado) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2024.

A modelagem de embarques e desembarques por ponto de parada, ao longo de linhas de ônibus, exerce uma importância fundamental no planejamento de redes de transporte público, além de contribuir para o desenvolvimento urbano orientado ao transporte sustentável. Contudo, as variáveis de interesse, nesse caso (embarques e desembarques) apresentam quatro características que trazem implicações ao processo de modelagem e/ou podem afetar o resultado das estimativas. São elas: (1) dependência espacial; (2) assimetria; (3) os deslocamentos acontecem ao longo da rede; e (4) amostragem limitada. Tendo em vista que tais peculiaridades são ignoradas pela modelagem clássica e que a literatura científica ainda não as aborda de forma simultânea, o presente trabalho teve o objetivo principal de modelar a demanda por transporte público no âmbito de pontos de parada, incluindo, no processo de estimativa, as quatro características supracitadas. Como objetivos específicos, foram analisados os efeitos de variáveis explicativas, da aplicação de distâncias em rede ou euclidianas, da quantidade de dados faltantes e do tipo de amostragem. O texto está dividido em sete capítulos, dos quais quatro são artigos científicos, cujo conteúdo aborda um ou mais objetivos específicos separada e/ou simultaneamente. Duas classes de modelos espaciais foram propostas, interpoladores geoestatísticos e regressões geograficamente ponderadas, e um banco de dados composto por oito linhas de ônibus da cidade de São Paulo (Brasil) foi utilizado como estudo de caso. As seguintes conclusões foram obtidas: modelos que consideram a assimetria e dependência espacial de embarques e desembarques devem ser priorizados frente aos que ignoram tais características, assim como os modelos multivariados em comparação aos univariados. Políticas públicas para o aumento do uso do transporte público devem focar principalmente em quatro grupos de variáveis explicativas: características sociodemográficas, cobertura da rede de ônibus, desenho viário e uso do solo. Os modelos espaciais provaram ser capazes de estimar o volume de embarques e desembarques em pontos não amostrados com acurácia, solucionando o problema da falta de dados de demanda por transporte público no âmbito de parada. Apesar da abordagem com distâncias em rede não ter contribuído significativamente para a melhoria do poder preditivo dos modelos, esse tipo de distância pode representar melhor a relação entre demanda por transporte público e seus fatores intervenientes. Recomenda-se priorizar o uso de distâncias em rede na modelagem espacial de embarques e desembarques. Além disso, uma amostra baseada em preditores e bem distribuída no espaço geográfico deve ser priorizada para estimar com precisão dados faltantes de demanda por transporte público no âmbito de pontos de parada. Dessa forma, o presente trabalho acrescenta importantes contribuições metodológicas e práticas para um planejamento urbano associado ao transporte sustentável.

Palavras-chave: Demanda por transporte público. Krigagem. Regressão Geograficamente Ponderada. Dados de contagem. Ponto de parada. Dados faltantes.

LIST OF PUBLICATIONS

I. **Marques, S. de F.**, and Pitombo, C. S. (2021a) Applying multivariate Geostatistics for transit ridership modeling at the bus stop level. *Bulletin of Geodetic Sciences*, *27*(2). doi:10.1590/1982-2170-2020-0069

II. **Marques, S. de F.**, and Pitombo, C. S. (2023a) Transit ridership modeling at the bus stop level: comparison of approaches focusing on count and spatially dependent data. *Applied Spatial Analysis and Policy*, *16*(1), 277–313. doi:10.1007/s12061-022-09482-y

III. **Marques, S. de F.**, and Pitombo, C. S. (2023b) Local modeling as a solution to the lack of stop-level ridership data. *Journal of Transport Geography*, *112*, 103682. doi:10.1016/j.jtrangeo.2023.103682

IV. **Marques, S. de F.**, Pitombo, C. S., and Gómez-Hernández, J. J. (2024) Spatial modeling of travel demand accounting for multicollinearity and different sampling strategies: a stop-level case study. *Journal of Advanced Transportation*, *2024*, 7967141. doi:10.1155/2024/7967141

AUTHOR CONTRIBUTIONS

Samuel de França Marques was responsible for Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, and Writing – review and editing.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

1 INTRODUCING THE THESIS

> "*What is the most resilient parasite? Bacteria? A virus? An intestinal worm? An idea. Resilient...highly contagious. Once an idea has taken hold of the brain, it's almost impossible to eradicate. […] It's not just about depth. You need the simplest version of the idea in order for it to grow naturally in your subject's mind. It's a very subtle art.*"
>
> Inception by Christopher Nolan

Travel demand modeling along bus lines is an important tool for Public Transport (PT) network planning, as well as for the promotion of Transit-Oriented Development (TOD). Information on boarding and alighting per bus stop is commonly used to define fleet size and composition, bus line capacity, trip scheduling, route design, decisions on whether to install a new bus stop or remove it and where to apply this, etc. (Ceder, 2007; Vuchic, 2005). In turn, the relationships between these data and the built environment features around bus stops, which are analyzed in the modeling, support decision-making on land use policies and transport system coverage (Cervero, 2006; Cervero and Dai, 2014; Taylor and Fink, 2013).

However, boarding and alighting, which refer to the number of passengers entering the bus line and leaving the bus line, respectively, show some of the peculiarities concerning travel demand variables that ridership modeling has not accounted simultaneously for so far in literature. They are: 1) Spatial autocorrelation: which means that travel demand values at points located near each other in space tend to be more similar than at points distant from each other (Tobler, 1970); 2) Asymmetry: travel demand variables are often collected as a count data and, because of that, they are not normal, as assumed by traditional linear model; 3) Lack of data: information on boarding and alighting is hard to acquire as it often depends on expensive field surveys such as boarding and alighting count survey. Therefore, it is very difficult to collect these data for the entire PT network; and 4) Network distance measure: as the travel activity occurs along a road network, ridership modeling approaches, such as spatial ones, which depend on a distance measure, can also use, as an alternative option, the distance along the network. As the network distance reflects the real path taken by the traveler, using this metric could yield better results than the traditional straight-line distance measure, the Euclidean one (Eom et al., 2006; Wang and Kockelman, 2009).

When these characteristics are overlooked in travel demand modeling, the estimates may not be the best and, in some cases, such as the first two, there could be a misunderstanding in the models' parameters, leading to wrong conclusions about the effect of external variables

on boarding, alighting and loading. In spite of that, ridership models at the bus stop level usually found in the literature do not account for these peculiarities (Chu, 2004; Dill et al., 2013; Kerkman et al., 2015; Pulugurtha and Agurla, 2012; Ryan and Frank, 2009), and even travel demand models at station level overlook one or more of them (Blainey and Mulley, 2013; Blainey and Preston, 2010; Cardozo et al., 2012; Choi et al., 2012; Gan et al., 2019; Liu et al., 2018; Zhu et al., 2019).

On the other hand, the boarding and alighting modeling at the bus stop level can be considered the most recent approach of bus ridership. In general, ridership modeling studies found in the literature can be divided according to the adopted spatial analysis unit: (1) systems/cities; (2) Traffic Analysis Zones (TAZs), neighborhoods and districts; (3) bus lines; (4) train stations, metro stations and bus stops; and (5) households/individuals, ranging from the most aggregated to the most disaggregated level.

In this context, the bus stop level shows some advantages over the other spatial units: (a) it can analyze the effects of TOD on bus ridership at a scale finer than in the case of stations, as bus stops are separated by a distance smaller than that of stations; and (b) as the bus stop is situated in an intermediate level of aggregation, spatial approaches of bus ridership are likely to yield better results than in the case of the other units of analysis. In a more aggregated level, such as the TAZ one, some intrazonal variations cannot be captured by the modeling as the variables are considered uniform across each whole unit and, hence, ecological fallacy may occur. In turn, when the modeling is performed in the context of households or individuals, the high randomness associated to human behavior compromises the estimates' performance. As spatial approaches rely on the assumption of spatial dependence, neighbor values must be similar, which is a situation that is not easy to achieve in the case of households/individuals.

Thus, characteristics of data at the bus stop level not only have some advantages over features from the other units of analysis, but also show the four peculiarities aforementioned regarding the travel demand variables (Spatial autocorrelation; Asymmetry; Lack of data and Network distance measure). The variables considered in this study are boarding and alighting, which represent the transit ridership. As these characteristics have not been addressed in previous studies found so far, mainly at the bus stop level, and given the advantages arisen from this spatial unit of analysis, some research opportunities can be created.

1.1 RESEARCH HYPOTHESES

The main hypotheses of this research are: 1) Spatial approaches of ridership modeling can yield better estimates than non-spatial models; 2) Ridership models considering the asymmetry of travel demand variables are more adequate than those based on the normality assumption; 3) Spatial interpolation modeling of ridership data can overcome the problem of data scarcity, regarding boarding and alighting per bus stop; and 4) The distances along the bus lines can yield better estimates than with the traditional Euclidean distance. Hence, the objectives of the thesis are outlined as follows:

1.2 GENERAL OBJECTIVE

To perform a transit ridership modeling along bus lines based on spatial approaches.

1.3 SPECIFIC OBJECTIVES

a) To investigate what factors affect the stop-level transit ridership.
b) To assess the improvements in the estimates provided by the inclusion of explanatory variables in the geostatistical modeling of bus ridership.
c) To compare the performance of spatial and local models of bus ridership with traditional approaches.
d) To compare the performance of spatial approaches of bus ridership using network distances and Euclidean distances.
e) To evaluate the performance of spatial and local models on predicting stop-level ridership data in unsampled stops.
f) To analyze the effect of the sampling strategy on the prediction accuracy of stop-level ridership models.

1.4 MATERIALS AND GENERAL METHOD OF THE THESIS

The analyses carried out in the thesis used the city of São Paulo (Brazil) as a case study. São Paulo is the most populous city and main economic center in Brazil and South America.

Its Gross Domestic Product is mainly based on the sectors of services, industrial, public administration and agricultural (IBGE, 2021).

The main reason for choosing São Paulo is the availability of data regarding stop-level ridership data and several variables that commonly serve as predictors for boardings and alightings. Data on the variables of interest were provided by *São Paulo Transporte S.A.* (*SPTrans*), the administrator of the São Paulo bus service. In turn, the GeoSampa website (<https://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx> Accessed May 2023) was the source for most independent variables used.

Two classes of predictor data were collected: 1) originating in the bus stop itself, such as bus frequency, number of lines, distance to nearest metro station etc.; and 2) from the bus stop catchment area, considered as the area covering a radius of 400m (Zhao et al., 2003) centered in the bus stop (for example: population, area of various land use categories, income etc.). Predictors from the catchment area were calculated based on areal interpolation, for cases in which the source data had been given in areal units (Traffic Analysis Zones or blocks); or averaged values, in the cases whose original data was point-based (households sampled in an Origin and Destination survey).

The boarding and alighting data corresponds to 2017 and covers 8 of the 1,355 bus lines São Paulo had that year. Together, these 8 lines serve 631 of 20,006 bus stops. *SPTrans* made the boarding and alighting data available for six time bands: 1st (04h to 04h59), 2nd (05h to 08h59), 3rd (09h to 15h59), 4th (16h to 19h59), 5th (20h to23h59) and 6th (00h to 03h59). However, in the boarding and alighting survey that gave rise to the data of interest, only passengers boarding and alighting each line were counted. Therefore, for bus stops serving different bus lines, the transit ridership variable collected is likely to be different as well. Consequently, a better approach would be to work with each bus line separately.

Based on this, the analyses carried out in the following chapters use different bus lines as a case study. Two main criteria supported the selection of bus lines for a case study: 1) number of bus stops (the higher the number of stops, the better); and 2) availability of data on the explanatory variables.

Figure 1.1 characterizes the database used as a case study. It shows maps of the lines visited by the boarding and alighting survey, the transit ridership (boardings plus alightings during an entire day) along these lines, sociodemographic and transport system features.

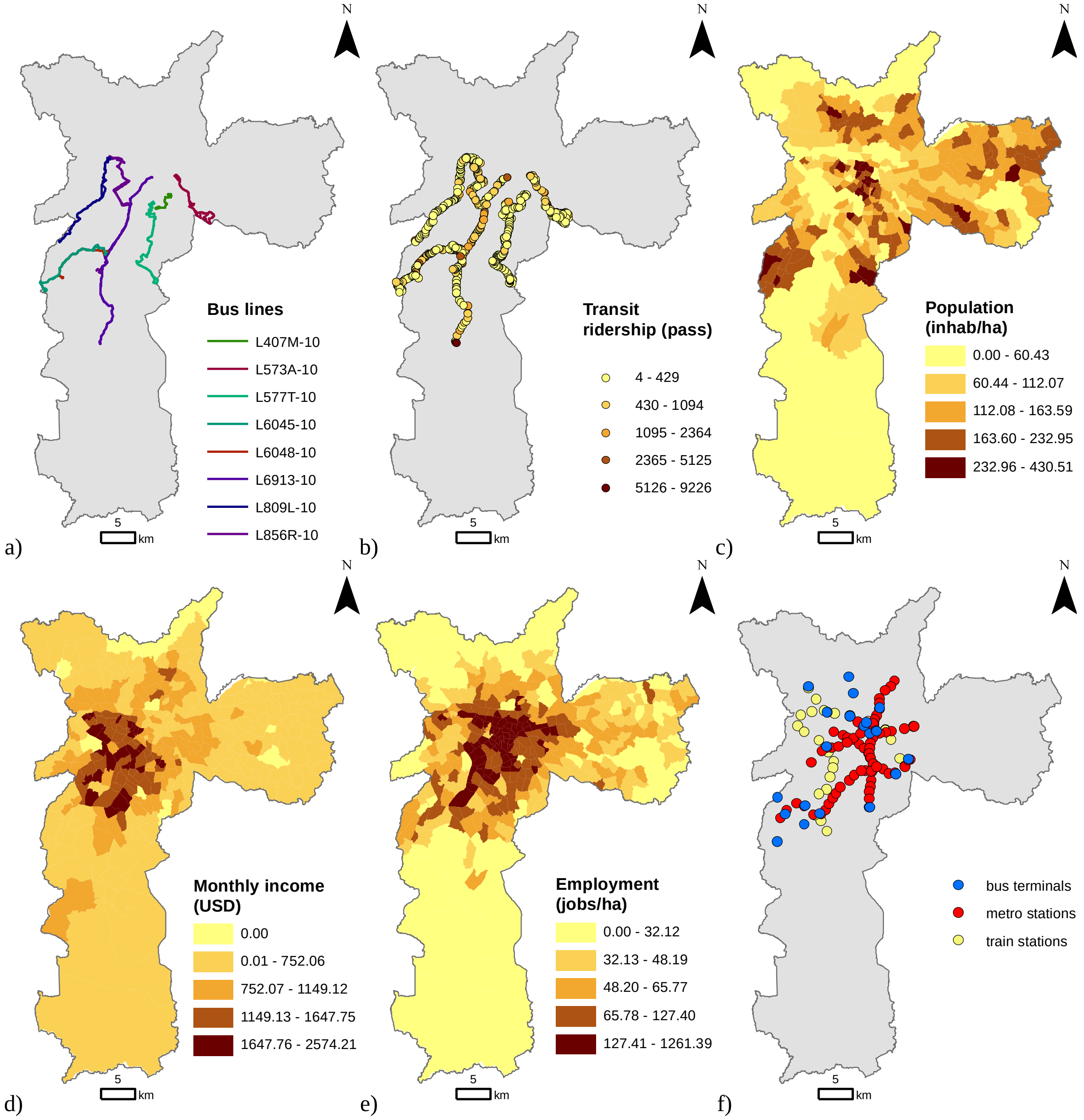


Figure 1.1 - Database characterization. * 1 USD equals 5.21 BRL (Mar. 2023)

Features such as income and number of jobs follow a similar pattern, which is more concentrated around the city center (which coincides with the geographic center of the city). On the other hand, population is more spread over the city. Overlapping these maps helps to understand the origin of the supposed spatial dependence on the ridership data: in a common urbanizing process, most people choose to live near the services and activities they need, which are usually located in the city center (Rodrigue et al., 2016). Therefore, the land use price starts to significantly increase in this area, allowing only wealthier families to continue living there.

In the case of São Paulo, high population densities can be seen outside the main center, in areas that have, in general, low-to-middle levels of income. Spatial dependence is created

when this population needs to move every day from their borough to the city center to work, as most job positions are available in the center. Therefore, a usual pattern shows large flows of passengers boarding from the city periphery, gradually decreasing as the bus lines approach the center. For alightings, it is the opposite. Note that higher numbers of transit ridership are often visualized in regions with high population and employment density.

In addition to the demographic and economic features, the availability of transport infrastructure also reveals the presence of spatial dependence on the bus patronage. This is due to the fact that the placement of a transit station or terminal is often associated to level of activity shown by a given area. In São Paulo, terminals and stations clearly follow the major employment axes.

The travel demand variables referred in the thesis title are boarding and alighting per bus stop along bus lines. Also known as "transit ridership", these variables are of great importance to bus network planning, but municipalities often face obstacles to their collection, as the surveys or devices required to acquire these data are highly expensive. However, as boarding and alighting per bus stop usually show spatial dependence, that is, points close to each other in space are more related than distant ones (Tobler, 1970), the main motivation of the thesis is to overcome the collection limitation by using spatial approaches, such as those of Geostatistics and Geographically Weighted regressions.

Goestatistics covers a set of spatial interpolation techniques whose main objective is to use the information of interest collected in field samples to estimate its value along the entire space in which the variable of interest occurs. Therefore, the result is a continuous surface of interpolated values, covering both sampled and non-sampled points (Cressie, 1993; Matheron, 1963, 1971; Wackernagel, 2003).

On the other hand, Geographically Weighted regressions address not only the spatial dependence of variables of interest, but also the potential spatial heterogeneity of predictor data (explanatory variables). While traditional regressions, such as the linear one, assumes a single predictor parameter for all points in the database, Geographically Weighted regressions calibrate a different model for each point, allowing the explanatory variables' coefficients to vary over space (Brunsdon et al., 1996; Fotheringham et al., 2003).

The general objective of the thesis, namely "To perform a ridership modeling along bus lines based on spatial approaches" is motivated by the main research gap of addressing the spatial dependence in the modeling and estimation of boardings and alightings at the bus stop level. Figure 1.2 shows the general method involving the thesis structure, in which each model or treatment proposal is made to overcome problems found along the path to accomplish the

general objective. Therefore, subsequent methods embrace the solutions previously established and the treatment evolves, becoming more refined, sophisticated and appropriate.

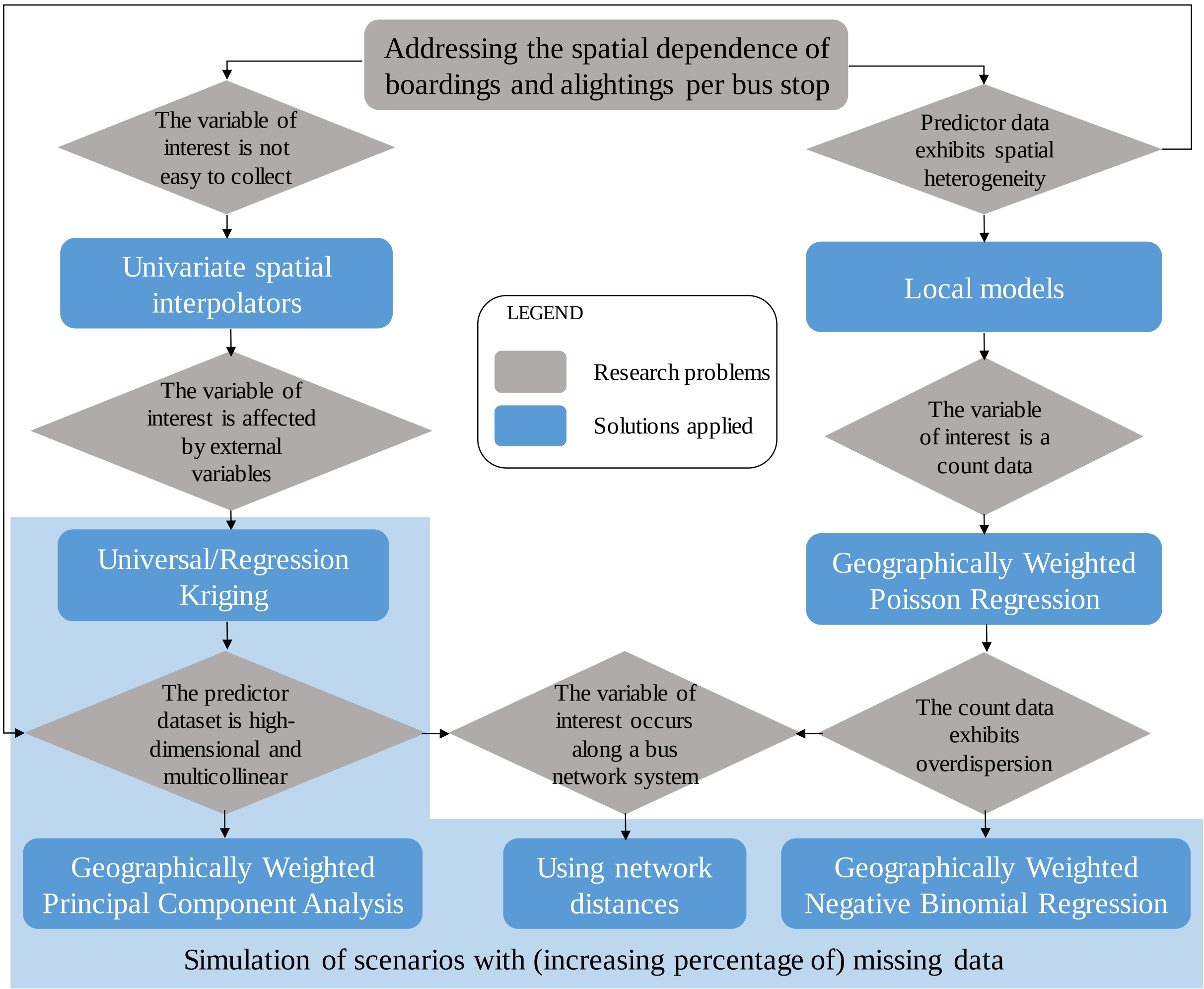

Addressing the spatial dependence of boardings and alightings per bus stop
The variable of interest is not easy to collect
Predictor data exhibits spatial heterogeneity
Univariate spatial interpolators
Local models
LEGEND
Research problems
Solutions applied
The variable of interest is affected by external variables
The variable of interest is a count data
Universal/Regression Kriging
Geographically Weighted Poisson Regression
The predictor dataset is high-dimensional and multicollinear
The variable of interest occurs along a bus network system
The count data exhibits overdispersion
Geographically Weighted Principal Component Analysis
Using network distances
Geographically Weighted Negative Binomial Regression
Simulation of scenarios with (increasing percentage of) missing data

Figure 1.2 - General method of the thesis.

Although the database characterization (Figure 1.1) provides strong evidence for the presence of spatial dependence on the transit ridership data, we do not treat this fact as a rule in the thesis chapters, but as an assumption. Throughout the analyses carried out, spatial dependence on the data of interest is consistently attested by the Moran index (Moran, 1948).

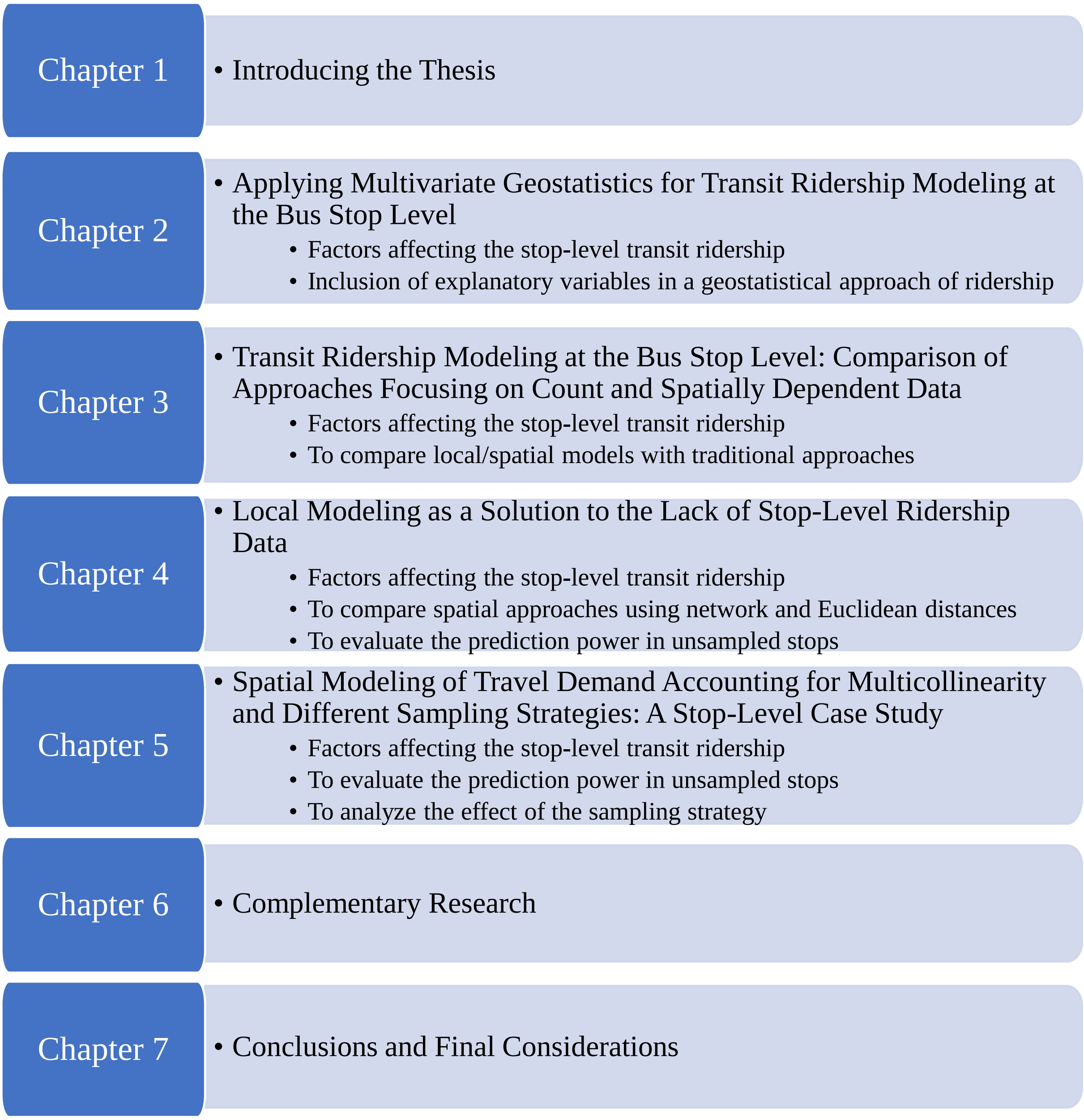
1.5 STRUCTURE OF THE THESIS

The structure of the thesis follows an article collection, in which each chapter corresponds to an article. Thus, the thesis is divided into seven chapters, in which this Introduction is the first one. The next four chapters comprise articles that address some of the specific objectives mentioned above, which, in turn, are associated with research gaps in the

travel demand modeling area. Chapter 6 briefly presents additional research developed during the doctoral period which had an impact on the thesis conclusions. The last chapter, in turn, concludes the thesis, summarizing the main results achieved and suggestions for future research.

The first article (Chapter 2) is entitled “Applying Multivariate Geostatistics for Transit Ridership Modeling at the Bus Stop Level” and it mainly addresses the specific objective (b). The second one (Chapter 3) includes mostly the specific objective (c) and has the title of “Transit Ridership Modeling at the Bus Stop Level: Comparison of Approaches Focusing on Count and Spatially Dependent Data”. The third article (Chapter 4), entitled “Local Modeling as a Solution to the Lack of Stop-Level Ridership Data”, addresses the specific objectives (d) and (e). Finally, the fourth article (Chapter 5), entitled “Spatial Modeling of Travel Demand Accounting for Multicollinearity and Different Sampling Strategies: A Stop-Level Case Study”, addresses the specific objectives (e) and (f). Specific objective (a) is present in all articles.

Figure 1.3 shows the structure of the thesis. The specific objectives associated with each chapter are also shown.



Chapter 1

- Introducing the Thesis

Chapter 2

- Applying Multivariate Geostatistics for Transit Ridership Modeling at the Bus Stop Level
    - Factors affecting the stop-level transit ridership
    - Inclusion of explanatory variables in a geostatistical approach of ridership

Chapter 3

- Transit Ridership Modeling at the Bus Stop Level: Comparison of Approaches Focusing on Count and Spatially Dependent Data
    - Factors affecting the stop-level transit ridership
    - To compare local/spatial models with traditional approaches

Chapter 4

- Local Modeling as a Solution to the Lack of Stop-Level Ridership Data
    - Factors affecting the stop-level transit ridership
    - To compare spatial approaches using network and Euclidean distances
    - To evaluate the prediction power in unsampled stops

Chapter 5

- Spatial Modeling of Travel Demand Accounting for Multicollinearity and Different Sampling Strategies: A Stop-Level Case Study
    - Factors affecting the stop-level transit ridership
    - To evaluate the prediction power in unsampled stops
    - To analyze the effect of the sampling strategy

Chapter 6

- Complementary Research

Chapter 7

- Conclusions and Final Considerations

Figure 1.3 - Structure of the thesis.

The current text is the corrected version of the thesis, which was defended on July 03rd, 2023. The original version of the thesis comprised the published version of the first two articles and the accepted version of the third article. The fourth article (Chapter 5) had been only submitted to a journal. Based on the committee recommendations and the revisions carried out in the fourth article, a corrected version of the thesis was prepared. This corrected version includes the four articles in their published version. The São Carlos School of Engineering stablished that, in a thesis composed by an article collection, the published papers must appear as in their published format. Therefore, the four articles follow the layout of the journal where they were published. No author rights have been violated.

DOI 10.1590/1982-2170-2020-0069

ARTICLE

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

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

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

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

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

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

Received on: 24-Feb-2020

Accepted on: 03-Mar-2021

Abstract:

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

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

1. Introduction and Background

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

How to cite this article: MARQUES, S.; PITOMBO, C. Applying multivariate geostatistics for transit ridership modeling at the bus stop level. Bulletin of Geodetic Sciences. 27(2): e2021019, 2021.

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

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

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

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

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

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

Geostatistics covers different types of estimators. In this paper, we mention three of them: Simple Kriging (SK), Ordinary Kriging (OK) and Universal Kriging (UK). The search for the interpolator that demonstrates the best performance, in goodness-of-fit measures, has led to several studies in which Simple Kriging results are compared to those of Ordinary Kriging (Daya and Bejari 2015; Taharin and Roslee 2017; Viswanathan et al. 2015), in which Ordinary Kriging is compared to Universal Kriging (Hiemstra et al. 2010; Kiš 2016; Liu et al. 2015; Mubarak et al. 2015; Nalder and Wein 1998; Wang and Zhu 2016), and in which the three techniques are simultaneously compared (Asa et al. 2012; Seo et al. 2015). In short, since UK includes explanatory variables in its formulation, it normally outperforms the other interpolators, especially when there is some large-scale trend present in the interest variable structure. Afterwards, OK, which assumes that the interest variable mean is unknown and varies locally, demonstrates the best results compared to SK, whose mean is global, constant and known.

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

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

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

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

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

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

2. Ridership models at the bus stop level

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

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

Table 1: Ridership models at the bus stop level

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

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

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

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

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

3. Materials and Method

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The cited goodness-of-fit measures were applied to the results of each estimate and, hence, it was possible to assess and compare the accuracy of results found from such techniques, and to select those that demonstrated the best performance. In the UK cases, results from the semivariogram that provided the smallest errors were selected to compare them with the respective linear regression estimates. The computational resources that gave support to the method stages were: ArcGIS 10.1, QGIS 3.0.3 and GRASS GIS 7.4.0 (Bundala, Bergenheim and Metz 2014) to collect the potential predictors; GeoDa (Anselin 2004; Anselin, Syabri and Kho 2005) for Moran's index calculation; IBM SPSS 24.0 (IBM 2016) for correlation analysis; and R (R Core Team 2020; Ribeiro Jr and Diggle 2016; Papritz 2020a; Papritz 2020b) for linear regression, Ordinary Kriging and Universal Kriging.

4. Results and Discussion

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

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

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

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

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

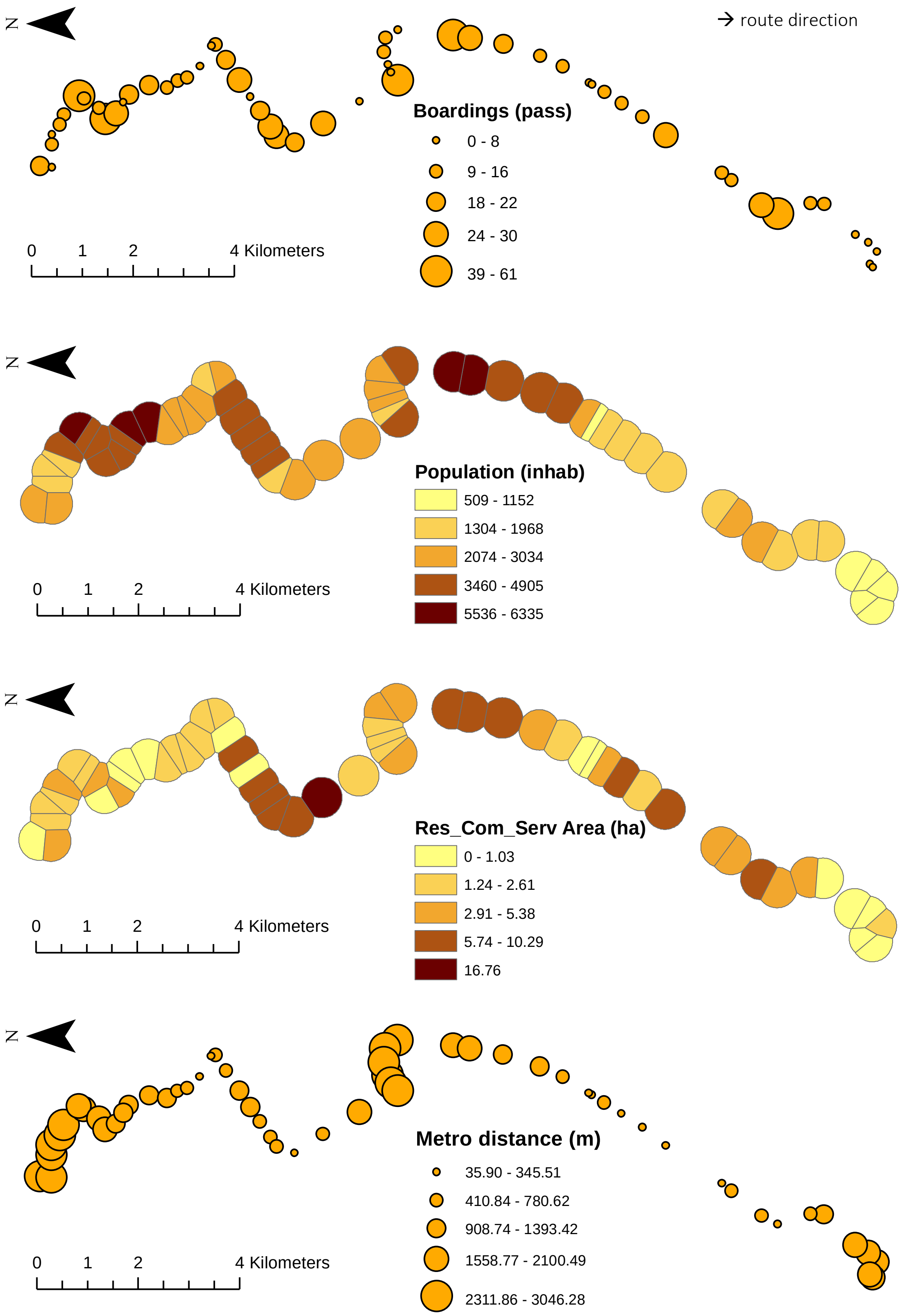


Figure 1: Patterns of (from top to bottom) Boardings; Population; Residential, commercial and services area; and distance to the nearest metro station along the bus line 856R-10

Table 2: Descriptive statistics

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

4.1 Univariate step: Ordinary Kriging

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

Table 3: Ordinary Kriging results

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

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

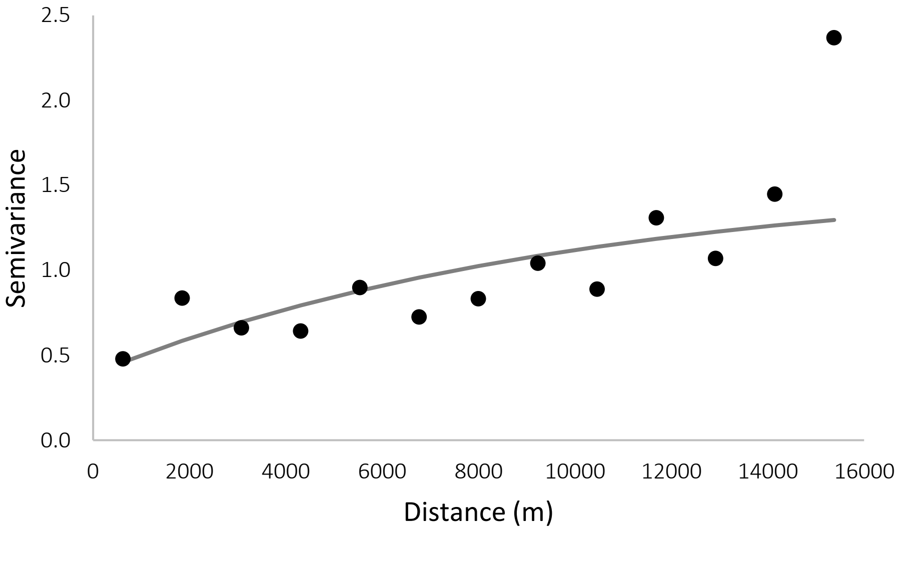


Figure 2: Semivariogram of Boardings with logarithmic transformation

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

4.2 Multivariate step: Universal Kriging and linear regression

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

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

Table 4: Results from spatial interpolators and classical linear regression

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

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

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

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

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

Table 5: Goodness-of-fit measures

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

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

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

Although models with more predictors may better explain the variance of interest variable, estimates can show no or little improvement when a new explanatory variable is added to the model, even a statistically significant one. The best results, from both Universal Kriging and linear regression, are highlighted in bold in Table 5. In the case of Universal Kriging, the model with only *ln_pop* as the predictor yielded the best estimates, while for linear regression, the best results are those from models 2.1, which use *ln_pop* and *res_com_serv area*, and 3, which uses all three predictors.

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

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

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

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

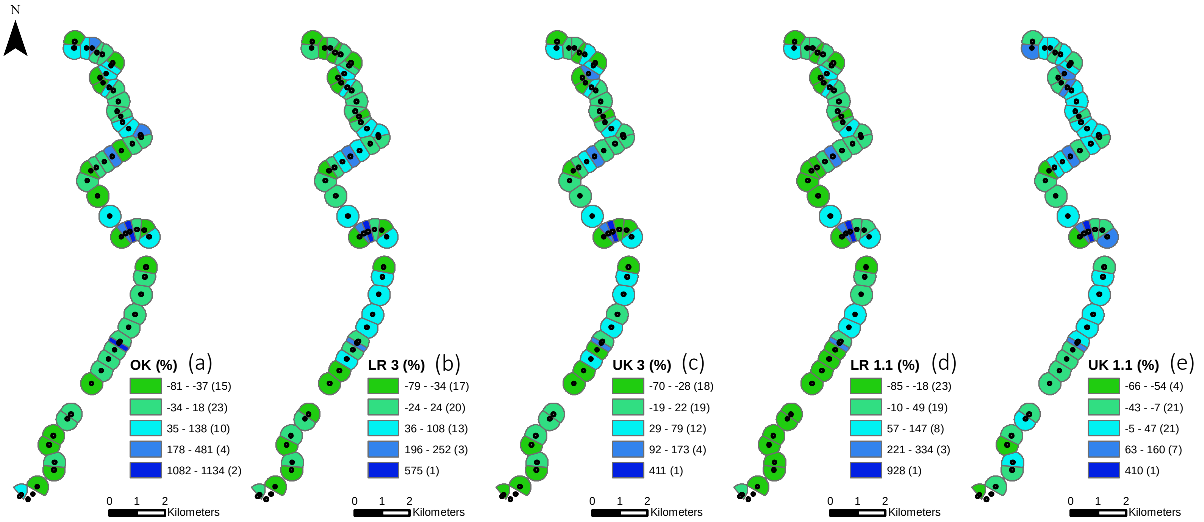


Figure 3: Error ratios of (from left to right) Ordinary Kriging (a), Linear Regression and Universal Kriging with all three predictors (b and c), and Linear Regression and Universal Kriging with *ln_pop* (d and e).

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

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

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

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

5. Conclusions and Final Remarks

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

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

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

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

ACKNOWLEDGEMENT

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

AUTHOR´S CONTRIBUTION

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Applied Spatial Analysis and Policy (2023) 16:277–313
https://doi.org/10.1007/s12061-022-09482-y

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

Samuel de França Marques[1] · Cira Souza Pitombo[1]

Received: 23 July 2021 / Accepted: 22 August 2022 / Published online: 2 September 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

[1] 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 Ridership, 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 Agurla (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

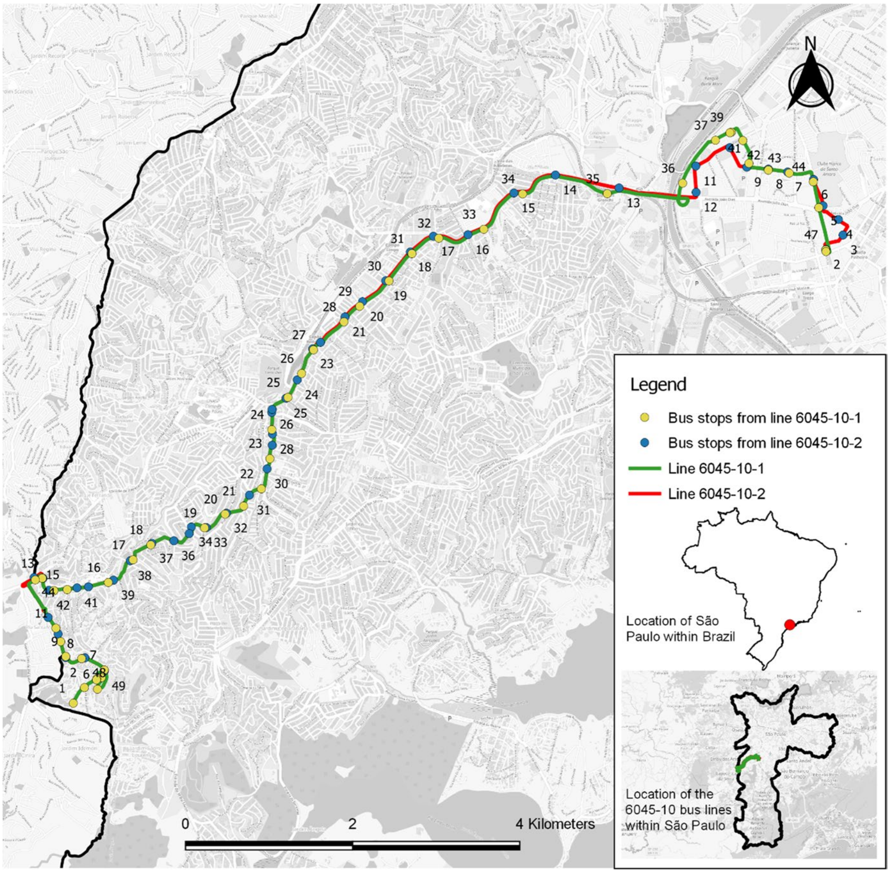

N
Legend
● Bus stops from line 6045-10-1
● Bus stops from line 6045-10-2
Line 6045-10-1
Line 6045-10-2
Location of São Paulo within Brazil
Location of the 6045-10 bus lines within São Paulo
0
2
4 Kilometers

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
Ryan and Frank (2009)	Boarding + alighting (logarithm)	OLS (log-linear)	Level of service (no. of routes/average waiting time)	Income, No-vehicle households, Female (%), Hispanic (%), White (%), Youth (%), Walkability index
Pulugurtha and Agurla (2012)	Boarding	Negative binomial with log-link	On-network characteristics	Household income, No-vehicle households, Asian population, Residential area
Dill et al. (2013)	Boarding + alighting (logarithm)	OLS (log-linear)	Transit service variables, Transportation infrastructure variables	Households below poverty (%), No-vehicle households (%), White (%), Youth (%), Elderly (%), Education level, Job accessibility, Employment (no.), Population (no.), Land use area (single-family, multifamily, commercial), Area parks, Pedestrian destinations, Land use mix index, Distance to city center
Kerkman 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.

$$SE = \sum_{i=1}^{n} \left[y_i - y_i^*\right]^2 \tag{1}$$

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).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \tag{2}$$

Where the response variable y comprises the linear combination of explanatory variables x_k added to a random error ε. The β 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 β coefficients can be obtained according to Eq. 3 (Yan & Su, 2009).

$$\beta = \left(X^t X\right)^{-1} X^t Y \tag{3}$$

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).

$$ln\ (\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{4}$$

Where μ 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 μ (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).

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

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 β parameters are calculated in traditional GWR.

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

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).

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

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).

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

Where *Wij* 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).

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

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 semivariogam 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).

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

Where X_0 is the matrix of explanatory variable observations of point x_0; β is the vector of linear parameter estimates; V_{so} represents the vector of estimated covariances between sample points and point x_0, while V_s expresses the matrix of estimated covariances between sample points. It is worth remembering that covariance (V) and semivariogram (γ) functions are related according to Eq. 11, where c_o and c_1 stand out, respectively, for the nugget effect and partial sill parameters from the theoretical semivariogram.

$$V(h) = c_0 + c_1 - \gamma(h) \tag{11}$$

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.

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

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.

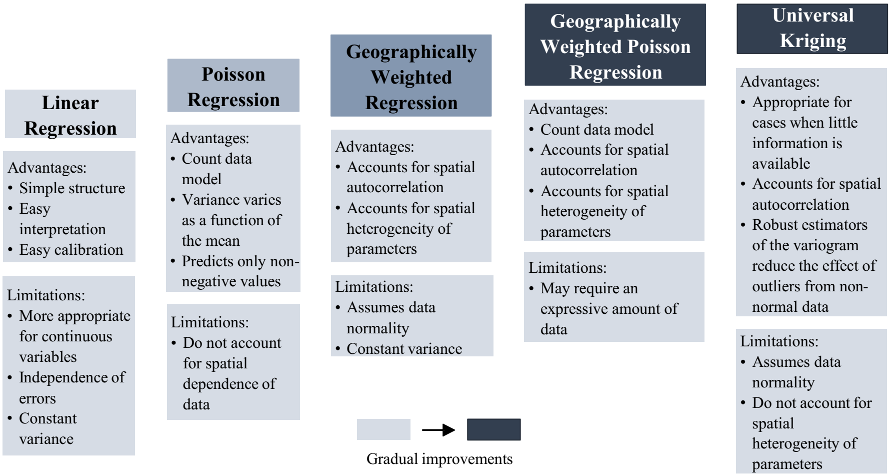

Linear Regression
Advantages:
• Simple structure
• Easy interpretation
• Easy calibration
Limitations:
• More appropriate for continuous variables
• Independence of errors
• Constant variance
Poisson Regression
Advantages:
• Count data model
• Variance varies as a function of the mean
• Predicts only non-negative values
Limitations:
• Do not account for spatial dependence of data
Geographically Weighted Regression
Advantages:
• Accounts for spatial autocorrelation
• Accounts for spatial heterogeneity of parameters
Limitations:
• Assumes data normality
• Constant variance
Geographically Weighted Poisson Regression
Advantages:
• Count data model
• Accounts for spatial autocorrelation
• Accounts for spatial heterogeneity of parameters
Limitations:
• May require an expressive amount of data
Universal Kriging
Advantages:
• Appropriate for cases when little information is available
• Accounts for spatial autocorrelation
• Robust estimators of the variogram reduce the effect of outliers from non-normal data
Limitations:
• Assumes data normality
• Do not account for spatial heterogeneity of parameters
Gradual improvements

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)	*Boarding*	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)	*Alighting*	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)

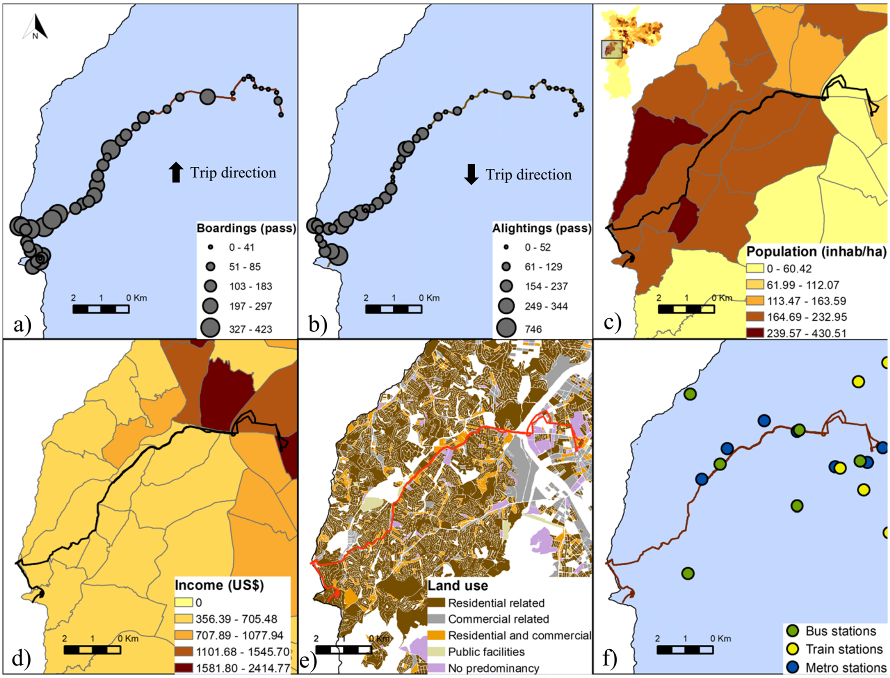


Fig. 3 Maps of **a** Boardings along 6045-10-1 line; **b** Alightings along 6045-10-2 line; **c** Population density at the TAZ level; **d** Average household income at the TAZ level; **e** Predominant land use at the block level; and **f** Bus, train and metro stations in the vicinity of the lines of 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 $Boardings = -47.43 + 0.02 * Population + 34.17 * Comservarea + 0.04 * Stationdistance$. 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 $Boardings = \exp(3.55 + 0.00021 * Population + 0.07332 * Comservarea + 0.00058 * Intra/interdistance - 0.00010 * income)$. 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 $\left[\exp(-0.00010 * 100) - 1\right] = 1.00\%$.

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

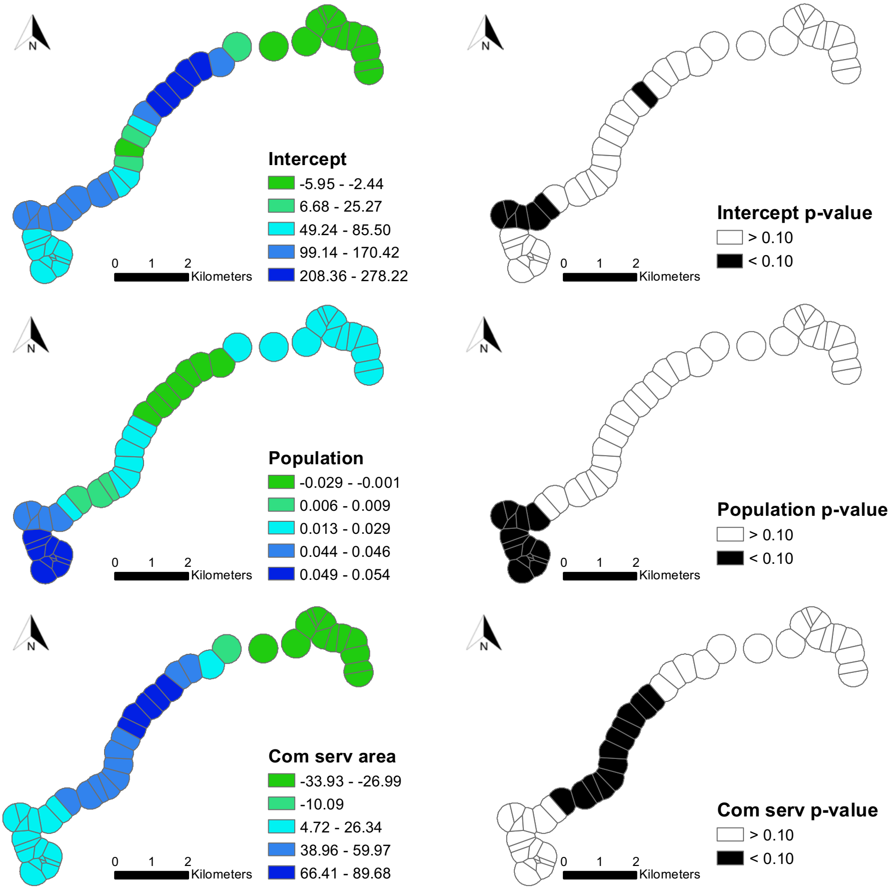


Fig. 4 Estimated Boarding GWR parameters and respective statistical significance

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 β coefficients to be estimated, UK

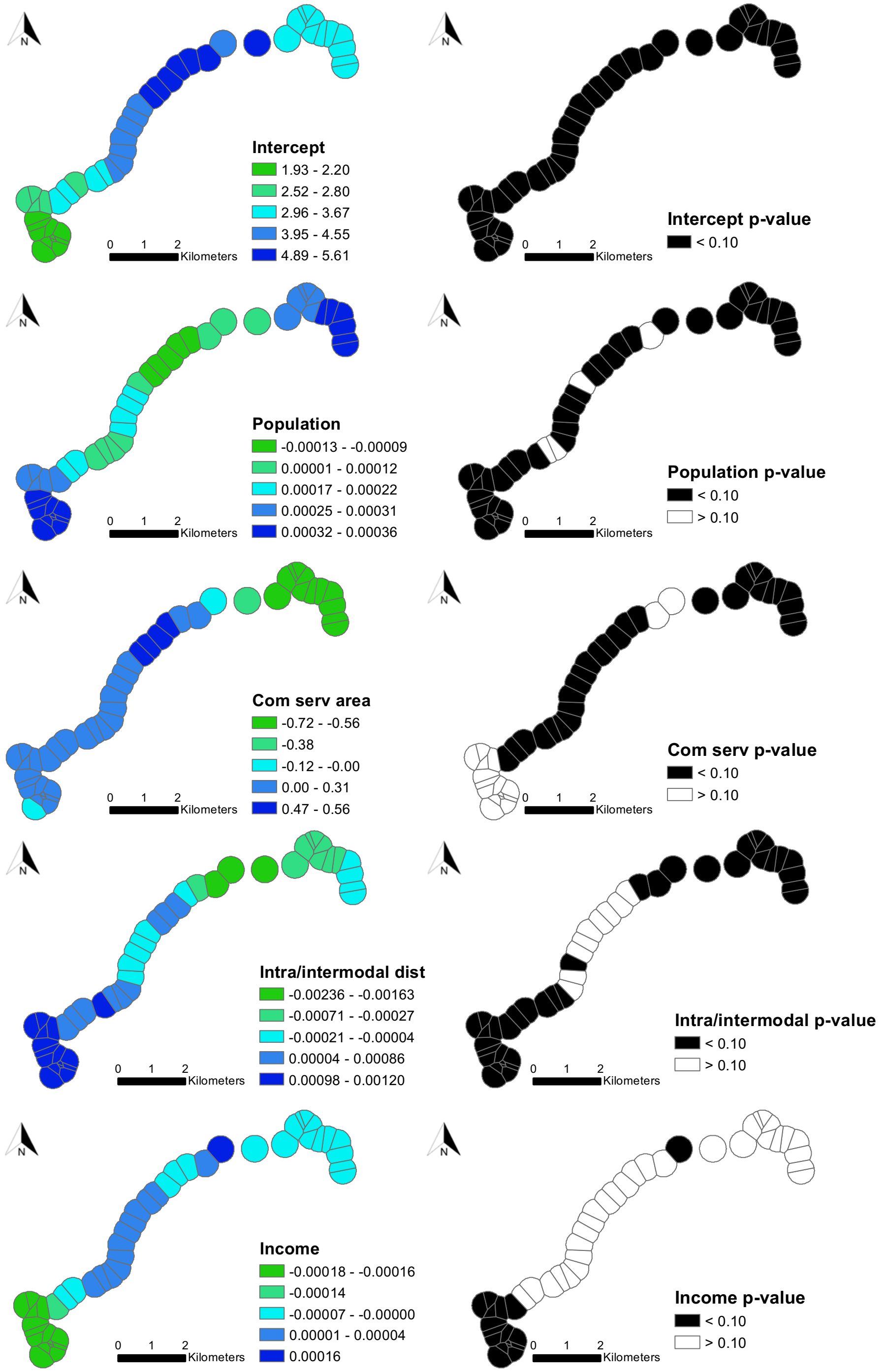


Fig. 5 Estimated Boarding GWPR parameters and respective statistical significance

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 R² for GWR ranges from 0.265 to 0.996, in which 70% of the bus stops have an R² 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	-0.00090	-0.00023
	25%	1.55685	0.00027	-0.19205	-0.00023	-0.00007
	50%	2.79466	0.00030	-0.00026	0.00017	-0.00003
	75%	4.39613	0.00034	0.15553	0.00030	0.00005
	Max	7.68902	0.00058	0.35559	0.00035	0.00007
UK		-159.20000	0.04209		0.06845	

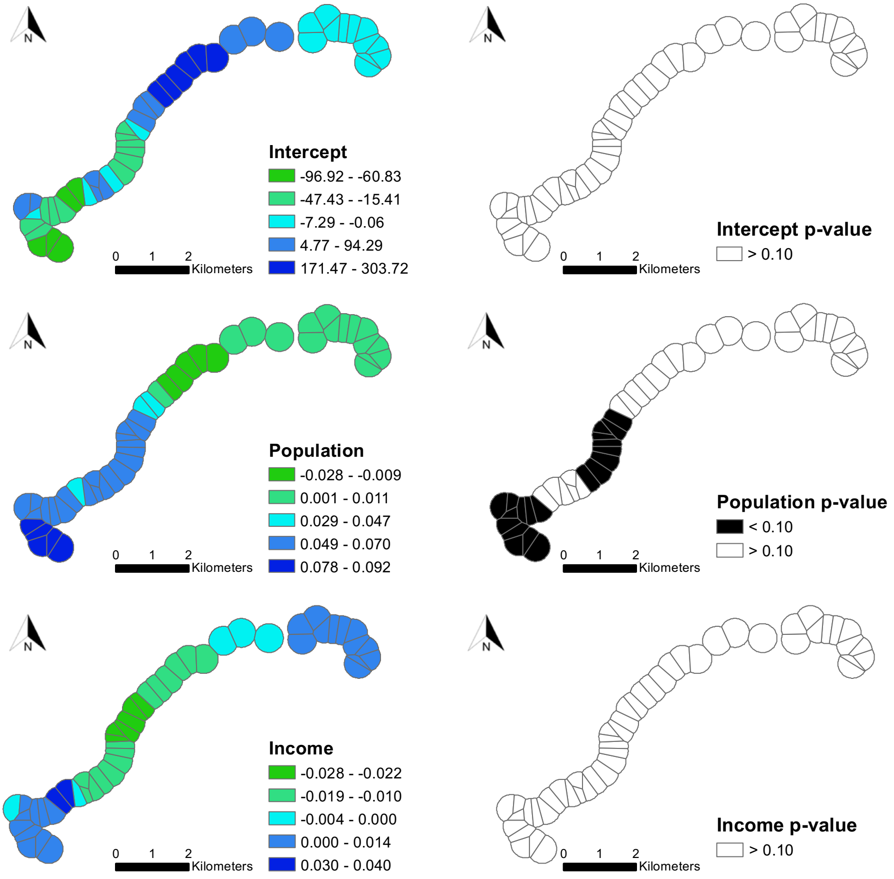


Fig. 6 Estimated Alighting GWR parameters and respective statistical significance

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

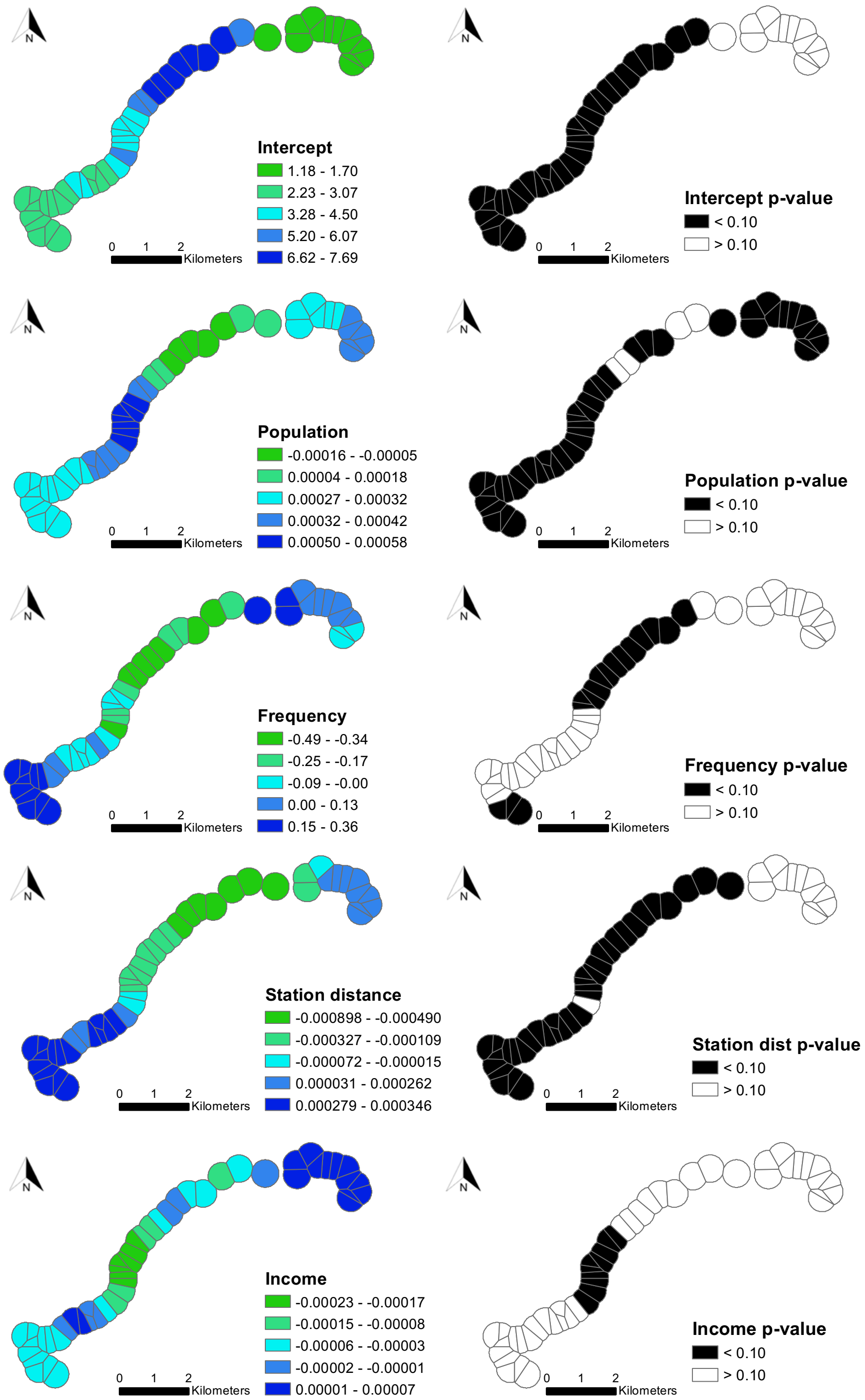


Fig. 7 Estimated Alighting GWPR parameters and respective statistical significance

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
Outbound (*Alighting*)	linear	population + frequency + station distance	315,091	62.126	80.190	71.623
	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 $R^2 = 0.328$
				3,582	8	Adjusted $R^2 = 0.330$
Pulugurtha and Agurla (2012)	USA	Boarding	Negative binomial with log-link*	2,857	12	Corrected quasi-likelihood = 4,431
Dill et al. (2013)	USA	Boarding + alighting (logarithm)	OLS (log-linear)	7,214	29	Adjusted $R^2 = 0.69$
				1,400	29	Adjusted $R^2 = 0.62$
				350	29	Adjusted $R^2 = 0.53$
Kerkman et al. (2015)	Netherlands	Boarding + alighting (logarithm)	OLS (log-linear)	1,232	18	Adjusted $R^2 = 0.772$
				1,284	18	Adjusted $R^2 = 0.762$
The authors	Brazil	Boarding	OLS	47	3	Adjusted $R^2 = 0.555$
			Poisson	47	4	Log-likelihood = -1,161
			GWR	19	2	Adjusted $R^2 = 0.717$
			GWPR	19	4	Pseudo $R^2 = 0.847$
			UK	47	2	AIC = 554.459
		Alighting	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 (poxy 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.

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://australasiantransportresearchforum.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. 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 integrandoLidar 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. 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). olsrr: *Tools for Building OLS Regression Models*. R package version 0.5.3. Available at: https://CRAN.R-project.org/package=olsrr

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

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 methodsto 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://onlinepubs.trb.org/Onlinepubs/trr/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

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

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

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., Ehrich, 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.

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Local modeling as a solution to the lack of stop-level ridership data

Samuel de França Marques [1],[*], Cira Souza Pitombo

Department of Transportation Engineering, São Carlos School of Engineering, University of São Paulo, São Carlos, Brazil

ARTICLE INFO

Keywords:
Transit ridership
Negative binomial regression
Geographically weighted negative binomial regression
Network distances
Missing data
Sensitivity analysis

ABSTRACT

Transit ridership modeling at the bus stop level is an important tool for bus network planning and transit-oriented development. However, many cities, especially in developing countries, face a lack of boarding and alighting data due to the high costs of collection. Solutions based on smartcards often rely on assumptions that negatively affect the data accuracy. Noting that previous studies suggest the existence of spatial heterogeneity and dependence in factors affecting stop-level ridership, the present paper proposes the application of Geographically Weighted Negative Binomial Regression (GWNBR) to modeling the transit ridership along bus lines in São Paulo – SP (Brazil) under missing data conditions. Four important topics are analyzed: 1) whether the spatial variation of predictors' effects is statistically significant; 2) the consistency of parameter estimates; 3) the prediction power sensitivity to missing data scenarios; and 4) the use of network distances replacing the traditional Euclidean ones. Of the five predictors that explained the transit ridership better, overlapping and frequency proved to have coefficients with statistically significant spatial variation. Goodness-of-fit measures indicated that GWNBR is an effective tool to the transit ridership estimation in uncounted bus stops, even when the availability of data is low. GWNBR in missing data scenarios could, in fact, reproduce the spatial pattern of effects shown in the complete database model, for some explanatory variables. Network distances may better represent the spatial relationship between transit ridership and some of its predictors. In addition, GWNBR models were able to address the spatial dependence found in the Negative Binomial Regression.

1. Introduction and background

Passengers boarding and alighting from bus lines at each bus stop of a transit system is valuable information for transit-oriented development and bus transit network planning (Ceder, 2007). In this context, transit ridership (boarding plus alighting) modeling at the bus stop level is an important tool to measure relationships between bus patronage and built environment characteristics around bus stops, giving support to urban policies, and providing estimates of boarding and alighting quickly and economically (Cervero, 2006).

Ridership models at the bus stop level found in the literature can be grouped by their main motivation and the type of model used. Most bus ridership studies have the main intention of identifying the factors affecting the stop-level passenger demand, while assuring good ridership estimates (Chu, 2004; Dill et al., 2013; Kerkman et al., 2015; Rahman et al., 2021). Measuring the impact of built environment and transport system characteristics on bus ridership contributes to developing urban policies towards sustainable development. At the same time, boarding and alighting estimates provide municipalities with enough information for a solid bus network planning, especially in cases of a lack of ridership data. In addition to these main contributions, recent articles have used the statistical modeling as a tool to measure the impact of other travel modes on bus ridership (Rahman et al., 2019) and to analyze how the passenger demand behaves from extreme weather events, considering the infrastructure provided by the bus stop (Lanza and Durand, 2021; Miao et al., 2019; Ngo, 2019). Table 1 consolidates some main characteristics from all studies found so far.

Previous studies have modeled boarding and alighting, separately, but also the total transit ridership, and, depending on the availability of data, different time bands are used. Table 1 reveals that the number of bus stops is a feature of great variability throughout the models, although most cases are concentrated in the United States of America. A reason for this might be the high cost for collecting boarding and alighting data, making their modeling hard to carry out in developing countries. While most case studies use Automatic Passenger Count (APC) data, which provided boarding and alighting information for all or most

[*] Corresponding author.
E-mail address: samuelmarques@usp.br (S.F. Marques).

[1] Permanent adress: Trabalhador São-carlense Avenue, 400, City: São Carlos – State of São Paulo, Zip code: 13566–590, Country: Brazil

https://doi.org/10.1016/j.jtrangeo.2023.103682
Received 14 November 2022; Received in revised form 23 March 2023; Accepted 15 August 2023
Available online 23 August 2023

bus stops, the Brazilian ones (Marques and Pitombo, 2021a,b, 2022) relied on data from a boarding and alighting count survey, whose variables of interest were available only for 0.6% of the bus lines from São Paulo.

A document from São Paulo Mobility and Transport Secretary highlighted its interest in installing APCs on the São Paulo public buses in 2018 (São Paulo, 2018). The document presented a cost per APC unit of BRL 3750.51 and 1308 APCs would be bought, resulting in a total of BRL 5 million, approximately (in 2018, BRL 1.00 varied from USD 0.2388 to USD 0.3129). Despite the affordable cost related to this acquisition, it would cover only 1.5% of the São Paulo bus fleet and was eventually cancelled in 2018. An announcement made in 2022 states that a new Operations Center for *SPTrans* (the administrator of the São Paulo bus system) is under construction and must be finished until the end of 2023 (São Paulo, 2022). According to the announcement, an investment of BRL 63.7 million will provide new onboard technologies such as cameras, APCs, wi-fi, traffic control and service measures. If we consider the installation of APCs in 15% of the São Paulo bus fleet, such as the case of Chakour and Eluru (2016), the cost would be BRL 49 million (with no correction for inflation from 2018 to 2023), which is almost all the resources to implement the new Operations Center and Monitoring System.

Even in developed countries, APC availability often does not cover all bus fleets. In our literature review, 9 studies, carried out in developed countries, were based on APC data (Chakour and Eluru, 2016; Chu, 2004; Cui et al., 2022; Dill et al., 2013; Lanza and Durand, 2021; Miao et al., 2019; Mucci and Erhardt, 2018; Ngo, 2019; Shi et al., 2021). Depending on the coverage of the counters, an estimate of cross-sectional boarding and alighting can be made for all bus stops, by replacing the devices from one vehicle to another in a short-time period. However, in some of the studies cited, the ridership variable is not available for all bus stops (Miao et al., 2019) or a method is used to expand the ridership from sampled buses to the whole system (Erhardt et al., 2017; Mucci and Erhardt, 2018), which does not account for spatial dependence. Chakour and Eluru (2016) reported to work with estimates coming from a representative sample of trips. There are also

Table 1
Ridership studies at the bus stop level.

Reference	Case study location	N. bus stops	Dependent variable	Regression model
Johnson (2003)	Minneapolis - St. Paul, USA	2568	Weekday boarding	Linear Regression
Ryan and Frank (2009)	San Diego, USA	3582	Daily boarding + alighting (logarithm)	Linear Regression (log-linear)
Cui et al. (2022)	Portland, USA	6261	Daily boarding	Linear Regression (log-linear)
Dill et al. (2013)	Portland, USA Lane County, USA Rogue Valley, USA	7214 1400 350	Weekday average boarding + alighting (logarithm)	Linear Regression (log-linear)
Kerkman et al. (2015)	Arnhem–Nijmegen, Netherlands	1232 1284	Average daily boarding + alighting (logarithm)	Linear Regression (log-linear)
Mucci and Erhardt (2018)	San Francisco, USA	6261	Average of the number of passengers boarding and alighting at each route-stop	Linear Regression (log-linear)
Frei and Mahmassani (2013)	Chicago, USA	11,000+	Number of boardings at half hour	Linear Regression (log-log)
Chu (2004)	Jacksonville, USA	2568	Weekday total boarding	Poisson Regression
Pulugurtha and Agurla (2012)	Charlotte, USA	2857	Average daily boarding	Linear, Poisson, Gamma, and Negative Binomial
Ngo (2019)	Lane County, USA	1500	Boarding + alighting from 5 am to 11 pm	Negative Binomial
Lanza and Durand (2021)	Austin, USA	1610	Boardings per day (13 h to 18 h)	Multilevel Negative Binomial
Shi et al. (2021)	King County, USA	96	Weekday boarding / Weekday alighting	Negative Binomial (with conditional tree inference for grouping predictors)
Chakour and Eluru (2016)	Montreal, Canada	8000	Boarding / Alighting (AM Peak; PM Peak; Off Peak Day; Off Peak Night)	Ordered probit model
Rahman et al. (2019)	Orlando, USA	3745	Weekday boarding / Weekday alighting	Joint panel mixed grouped ordered logit model
Miao et al. (2019)	Salt Lake City, USA	5879 5854	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)	Panel regression fixed-effects model
Marques and Pitombo (2021a)	São Paulo, Brazil	57	Boarding from 20 h to 23 h59 in a typical day (logarithm)	Universal Kriging and Linear Regression
Marques and Pitombo (2021b)	São Paulo, Brazil	96	Boarding from 20 h to 23 h59 on a typical day	Linear Regression with Ordinary Kriging of residuals
Rahman et al. (2021)	Orlando, USA	3495	Weekday boarding / Weekday alighting	Spatial error model / Spatial lag model and Linear Regression
Marques and Pitombo (2022)	São Paulo, Brazil	47 49	Weekday boarding Weekday alighting	Universal Kriging, GWPR, GWR, Poisson regression and Linear RegressionUniversal Kriging, GWPR, GWR, Poisson regression and Linear Regression

cases in which faulty APC recordings caused trips to be removed from the database (Cui et al., 2022).

Based on the type of model used, the stop-level ridership studies can be divided into four categories. The first group uses the traditional Linear Regression (LR) (Cui et al., 2022; Dill et al., 2013; Frei and Mahmassani, 2013; Johnson, 2003; Kerkman et al., 2015; Mucci and Erhardt, 2018; Ryan and Frank, 2009). In most LR cases, the positive asymmetry of the dependent variable is dealt with by taking the natural logarithm of the data. The second one consists of studies addressing asymmetry and/or overdispersion of transit ridership by applying count data models: Poisson Regression and/or Negative Binomial Regression or a variation of it (Chu, 2004; Lanza and Durand, 2021; Ngo, 2019; Pulugurtha and Agurla, 2012; Shi et al., 2021). The third group falls into the framework of categorical variable modeling, by using transit ridership data divided into classes (Chakour and Eluru, 2016; Rahman et al., 2019). Finally, the last group covers spatial regression models to address spatial dependence of transit ridership (Marques and Pitombo, 2021a,b, 2022; Rahman et al., 2021). Within the third and fourth groups, some studies have also accounted for temporal autocorrelation of their panel dataset, by using appropriate models for panel data (Miao et al., 2019; Rahman et al., 2019), or adding spatiotemporal lagged variables as predictors (Rahman et al., 2021).

Selecting the most appropriate modeling routine has been conditioned by important characteristics usually present in transit ridership data. Boarding and alighting are count data, they present asymmetry, and often overdispersion as well, if the variance of the dependent variable exceeds its expected mean (Hilbe, 2014). Spatial dependence, which is the fact that points close to each other in space are more related than distant ones (Tobler, 1970), has also been found in ridership data (Marques and Pitombo, 2021a,c, 2022). Under these circumstances, traditional models, such as Linear Regression and Poisson Regression, result in biased standard errors, as their basic assumptions are violated (Hilbe, 2014; Yan and Su, 2009). Hence, the interpretation of resulting coefficients can be misleading. In addition, passenger flow is a variable that occurs along the bus network. As such, proximity measures based on the Euclidean distance between points, aiming to incorporate the spatial dependence of the variable of interest, can be a rough simplification of the analyzed phenomenon (Lu et al., 2014a).

Although boarding and alighting modeling has been increasingly evolving over time, studies found do not account for some of the important features of transit ridership variables aforementioned. Research using spatial approaches overlook the network distance factor and overdispersion. The remaining studies do not address potential spatial dependence of transit ridership, and, hence, the network distance as well. However, research comparing multiple modeling techniques has shown that addressing these features can provide a better fit of transit ridership data: Negative Binomial Regression, which incorporates overdispersion of transit ridership, proved to perform better than Linear Regression and Poisson Regression in the case study by Pulugurtha and Agurla (2012). Spatial regressions have also improved goodness-of-fit measures of bus ridership modeling compared to the LR case (Marques and Pitombo, 2021a,b; Rahman et al., 2021). Incorporating asymmetry and spatial dependence, isolated and together, was fundamental to improve boarding and alighting estimates in the case study by Marques and Pitombo (2022).

A thorough analysis of all cited studies allowed the extraction of the main predictors of transit ridership, which were the ones appearing multiple times with statistical significance in different models. Table 2 shows these explanatory variables, and the sign of their respective effect on boarding, alighting and/or transit ridership.

From the percentage of no-vehicle households to the distance to the nearest station, it can be observed that all predictors showed both positive and negative signs in different case studies. Cases where the same dependent variable had predictors with opposite signs (in the same study) refer to different cities (Dill et al., 2013), different years (Kerkman et al., 2015), or different time band and ridership category

Table 2
Main predictors of transit ridership.

Predictor	Positive sign	Negative sign
No-vehicle households (%)	Frei and Mahmassani (2013)^{1*}, Ryan and Frank (2009)^{3}, Marques and Pitombo (2021b)^{1}, Chu (2004)^{1}	Dill et al. (2013)^{3}
Youth (%)	Johnson (2003)^{1}, Dill et al. (2013)^{3}, Frei and Mahmassani (2013)^{1}, Rahman et al. (2019)$^{1/2}$, Rahman et al. (2021)$^{1/2}$	Ryan and Frank (2009)^{3}, Chu (2004)^{1}
Population	Johnson (2003)^{1}, Dill et al. (2013)^{3}, Kerkman et al. (2015)^{3}, Marques and Pitombo (2021c)^{1}, Cui et al. (2022)^{1}, Marques and Pitombo (2022)$^{1/2}$	Frei and Mahmassani (2013)^{1}, Marques and Pitombo (2021b)^{1}, Marques and Pitombo (2022)$^{1/2}$
Residential area	Johnson (2003)^{1}, Dill et al. (2013)^{3}, Frei and Mahmassani (2013)^{1}, Chakour and Eluru (2016)$^{1/2}$, Mucci and Erhardt (2018)^{3**}, Marques and Pitombo (2021c)^{1}	Johnson (2003)^{1}, Pulugurtha and Agurla (2012)^{1}, Frei and Mahmassani (2013)^{1}, Chakour and Eluru (2016)$^{1/2}$; Rahman et al. (2021)^{2**}
Distance to city center	Dill et al. (2013)^{3}	Dill et al. (2013)^{3}, Chakour and Eluru (2016)$^{1/2}$, Rahman et al. (2019)$^{1/2}$, Rahman et al. (2021)$^{1/2}$
Elderly (%)	Dill et al. (2013)^{3}, Frei and Mahmassani (2013)^{1}	Dill et al. (2013)^{3}, Frei and Mahmassani (2013)^{1}, Kerkman et al. (2015)^{3}, Rahman et al. (2021)$^{1/2}$, Cui et al. (2022)^{1}
Employment	Dill et al. (2013)^{3}, Frei and Mahmassani (2013)^{1}, Chakour and Eluru (2016)^{2}, Mucci and Erhardt (2018)^{3}, Chu (2004)^{1}	Frei and Mahmassani (2013)^{1}, Rahman et al. (2021)^{1}, Chakour and Eluru (2016)^{1}
Bus stops in a buffer (overlapping)	Chakour and Eluru (2016)$^{1/2}$	Dill et al. (2013)^{3}, Kerkman et al. (2015)^{3}, Marques and Pitombo (2021b)^{1}, Mucci and Erhardt (2018)^{3}, Rahman et al. (2019)$^{1/2}$, Rahman et al. (2021)$^{1/2}$, Cui et al. (2022)^{1}, Chu (2004)^{1}
Number of lines through stops or bus route length within buffer	Ryan and Frank (2009)^{3}, Chakour and Eluru (2016)$^{1/2}$, Lanza and Durand (2021)^{1}, Shi et al. (2021)^{2}, Marques and Pitombo (2021b)^{1}, Rahman et al. (2019)$^{1/2}$, Rahman et al. (2021)$^{1/2}$	Chakour and Eluru (2016)$^{1/2}$
Bike lanes	Dill et al. (2013)^{3}, Chakour and Eluru (2016)$^{1/2}$	Chakour and Eluru (2016)^{2}
Highway length	Chakour and Eluru (2016)^{2}, Rahman et al. (2021)^{1}	Chakour and Eluru (2016)$^{1/2}$
Metro stations	Chakour and Eluru (2016)$^{1/2}$	Chakour and Eluru (2016)^{1}
Park area	Chakour and Eluru (2016)$^{1/2}$	Dill et al. (2013)^{3}
Land use (education)	Frei and Mahmassani (2013)^{1}, Marques and Pitombo (2021b)^{1}	Frei and Mahmassani (2013)^{1}
Rail transit	Dill et al. (2013)^{3}	Frei and Mahmassani (2013)^{1}; Rahman et al. (2021)^{2}
Headway	Marques and Pitombo (2022)^{2***}	Ryan and Frank (2009)^{3***}, Dill et al.

(*continued on next page*)

Table 2 (*continued*)

Predictor	Positive sign	Negative sign
		(2013)[3], Kerkman et al. (2015)[3***]; Chakour and Eluru (2016)[1/2], Mucci and Erhardt (2018)[3***], Rahman et al. (2019)[1/2], Rahman et al. (2021)[1/2], Cui et al. (2022)[1***], Marques and Pitombo (2022)[2***]
Land use (commercial)	Johnson (2003)[1], Pulugurtha and Agurla (2012)[1], Dill et al. (2013)[3], Chakour and Eluru (2016)[1/2], Marques and Pitombo (2021c)[1], Marques and Pitombo (2022)[1]	Marques and Pitombo (2022)[1]
Income	Marques and Pitombo (2022)[1]	Ryan and Frank (2009)[3], Pulugurtha and Agurla (2012)[1], Dill et al. (2013)[3], Kerkman et al. (2015)[3], Mucci and Erhardt (2018)[3], Rahman et al. (2019)[1/2], Rahman et al. (2021)[1/2], Cui et al. (2022)[1], Chu (2004)[1], Marques and Pitombo (2022)[1/2]
Distance to the nearest station	Marques and Pitombo (2022)[1/2]	Marques and Pitombo (2021c)[1], Marques and Pitombo (2022)[1/2]
Land use (institutional)	Pulugurtha and Agurla (2012)[1], Chakour and Eluru (2016)[2], Marques and Pitombo (2021b)[1]	
Land use (industrial)		Dill et al. (2013)[3], Frei and Mahmassani (2013)[1], Chakour and Eluru (2016)[1/2]
White (%)		Ryan and Frank (2009)[3], Dill et al. (2013)[3], Chu (2004)[1]
Land use mix	Johnson (2003)[1], Dill et al. (2013)[3], Rahman et al. (2019)[1/2], Rahman et al. (2021)[1/2]	
Shelter	Lanza and Durand (2021)[1], Shi et al. (2021)[1], Rahman et al. (2021)[1/2]	
Education level		Dill et al. (2013)[3], Rahman et al. (2019)[1/2]
Street connectivity	Dill et al. (2013)[3], Frei and Mahmassani (2013)[1****], Marques and Pitombo (2021b)[1], Cui et al. (2022)[1****]	
Bus station	Kerkman et al. (2015)[3], Mucci and Erhardt (2018)[3]	
Female (%)	Ryan and Frank (2009)[3], Chu (2004)[1]	

[1, 2, 3] represent the cases where the dependent variable was Boarding, Alighting and Transit Ridership, respectively. [1/2] are the cases where the authors addressed both Boarding and Alighting, but separately. [*, **, ***, ****] the reference predictor was represented by the number of vehicles (with a negative sign), housing density, frequency (with a positive sign) and Walk Score, respectively.

(Chakour and Eluru, 2016). Even though some opposite signs may be due to different time periods or another dependent variable, various cases with reverse effects of the same predictor come from different geographic regions. Therefore, in addition to the characteristics previously listed, we hypothesize that the effect of predictor data on transit ridership can vary spatially, a feature called spatial heterogeneity.

Two main observations arise from the bibliographic review carried out: 1) transit ridership and its relationships with the built environment and transport system variables are fundamental to transit-oriented development and bus network planning. However, there is a lack of stop-level boarding and alighting information, especially in developing countries, due to the high costs for collecting these data. Within the scope of our literature review, only the research from Marques and Pitombo (Marques and Pitombo, 2021a,b, 2022) provides case studies in a developing country (Brazil). However, they mainly focus on predicting the variables of interest well, lacking a more explanatory point of view of the phenomenon and identification of the main variables affecting transit ridership in a developing country context. 2) Spatial dependence and spatial heterogeneity have been overlooked in previous bus ridership studies. Since the spatial heterogeneity theory assumes different effects of explanatory variables along the database bus stops, geographically weighted regressions can be proposed as a solution to the lack of stop-level ridership data, as all bus stops (sampled and non-sampled) receive a different parameter value and a ridership estimate, following the concept of spatial dependence.

Therefore, the main objective of the present article is to model transit ridership using Geographically Weighted Negative Binomial Regression (GWNBR) with distances along the bus network. As specific objectives, we aim to compare the results of this model to the traditional case, in which Euclidean distances are used. GWNBR will also be compared to the global Negative Binomial Regression, and the statistical significance of the assumed spatially varying relationships will be attested. In addition, to explore the potential of local models to generate accurate parameter and ridership estimates in situations of missing data, a set of scenarios will be stablished, from which we aim to analyze the following topics: parameters' sensitivity, sensitivity of the model's prediction power and sensitivity to the use of network distances.

This paper has five sections. The next one presents a brief review on the application of spatial and local models in the scope of travel demand modeling, considering the use of network distances. Section 3 describes the case study dataset and the sequence of steps from the proposed method. Section 4 displays and discusses the results, and Section 5 provides a synthesis of the conclusions achieved, in addition to suggestions for future research.

2. Spatial models and network distances in travel demand modeling

Recognizing the existence of spatial dependence in travel demand variables, many researchers started to apply spatial and local models in their case studies. In this context, spatial models can be found in Annual Average Daily Traffic (AADT) studies (Eom et al., 2006; Mathew and Pulugurtha, 2021; Selby and Kockelman, 2013), urban travel demand by traffic analysis zones (Chiou et al., 2015; Ma et al., 2018; Tu et al., 2018), stations (Cardozo et al., 2012; Liu et al., 2018; Zhu et al., 2019), bus stops (Marques and Pitombo, 2021a, 2022; Rahman et al., 2021) and at the pedestrian level (Kim et al., 2019). Spatial dependence is traditionally incorporated into the modeling by means of a spatial weights matrix, which is commonly based on the distance between the database points, or by a semivariogram function, which also depends on the distance between pairs of points in the database. However, both modeling approaches, encompassing the Spatial Lags Model, Spatial Error Model and Kriging, adopt a single predictor parameter for the entire database (known as global models).

The emergence of geographically weighted regressions (Brunsdon et al., 1996) addressed the concept of spatial heterogeneity, from which each geographic unit in the database has its own parameter estimate, obtained from neighbor weighted data. Therefore, the closer a neighbor is from the point where the model is being estimated, the bigger the weight assigned to it. In case studies using geographically weighted regressions, they have proven to provide a better understanding of the impact of factors affecting the travel demand at a local level, in addition to achieving more accurate estimates, when compared to non-spatial models (Liu et al., 2018; Marques and Pitombo, 2022; Pulugurtha and Mathew, 2021; Tu et al., 2018; Zhu et al., 2019). Although geographically weighted regressions have been consistently used to analyze transit ridership data at the station level (Blainey and Mulley, 2013; Blainey and Preston, 2010; Cardozo et al., 2012; Liu et al., 2018; Zhu et al., 2019), applications of local models in the context of bus stops are still rare.

In short, three main factors remain underexplored: 1) Most research does not attest the statistical significance of the parameters' spatial variation. If the effect of a predictor does not have significant spatial variation, public policies may be applied equally to the whole system under analysis, resulting in a simpler decision-making. 2) Previous studies do not consider missing data situations. 3) Most spatial models found in the travel demand literature use, as a proximity measure, the Euclidean distance between the database points.

To verify the best suitability of network distance over the Euclidean one in the travel demand modeling, scholars have compared the prediction accuracy of spatial models using both types of distance. The performance improvement provided by network distances was not remarkable in some case studies (Marques and Pitombo, 2021b,c; Sarlas and Axhausen, 2015; Selby and Kockelman, 2013; Zhang and Wang, 2014). Nevertheless, Wong and Kwon (2021) showed that network distances' results are, in fact, better than those of Euclidean distances. However, all cited studies used network distances only in a global model, and local models found so far are restricted to the Euclidean distance approach.

2.1. Contributions, research questions and hypotheses

Based on the comments made before, the contributions of this article reach not only the scope of bus ridership, but also the broader context of the travel demand modeling. In short, the objectives are to evaluate the importance of distance type, the potential of GWNBR, and the impact of missing data. Aligned to these objectives are the following research questions and associated hypotheses.

a) *What are the factors affecting stop-level transit ridership?*
b) *Is the spatial variation of predictors' effects statistically significant?*
c) *Is GWNBR capable of providing good predictions of stop-level ridership in missing data scenarios?*

Associated hypothesis 1: taking the complete database model (the base model) as the result closest to reality, the higher the number of available data, the closer the predictors' effects in missing data scenarios will be to the base model.

Associated hypothesis 2: increasing the percentage of missing data reduces the GWNBR prediction accuracy, compared to the base model.

d) *Can network distances improve the prediction accuracy of GWNBR compared to the traditional Euclidean distances?*

Associated hypothesis 3: since the transit ridership variable occurs along a bus network, network distances may better represent the phenomenon under analysis, providing more accurate predictions than Euclidean distances.

Associated hypothesis 4: by increasing the percentage of missing data, the difference between network and Euclidean distances also increases. Therefore, higher levels of missing data will lead to more significant differences between results from network and Euclidean GWNBR, with a better performance of the network case.

3. Materials and method

São Paulo – SP is the most populated city in Brazil, having an estimated population of 12 million in 2021, according to the Brazilian Institute of Geography and Statistics (IBGE, 2021). The 2017 Origin and Destination Survey revealed that bus transit remains as the most used travel mode among the public ones, sharing a percentage of 21% of all produced trips in the city (Metrô, 2019).

General Transit Feed Specification (GTFS) data, provided by the manager of the São Paulo bus system (*SPTrans*), points out that the bus network comprised a total of 1355 lines (each line corresponding to a round trip) and 20,006 bus stops in 2017. However, as the São Paulo buses do not have APCs yet, the boarding and alighting data used hereby was collected through a boarding and alighting count survey. The survey was performed in 2012 only along eight lines, and, based on smart card data, *SPTrans* carried out an extrapolation of the survey results for 2017.

3.1. Dependent variable

The database used corresponds to the 2017 boarding and alighting data for two bus lines: line 856R-10-2 (outbound trip of line 856R-10) and line 6913-10-1 (inbound trip of line 6913–10), which, together, have 97 bus stops in the same direction (from south to north). These bus lines have an overlapping section with 16 bus stops. *SPTrans* also provided information on boarding and alighting for the other six lines, but not all of them present spatial contiguity with the others. Therefore, solely a few points from one line would be likely to affect the calibration of a local model in stops from another line. Another reason for choosing only lines 856R-10 and 6913–10 for the case study refers to the limitation present in the transit ridership data from *SPTrans*: only passengers boarding and alighting from the respective line were counted, while the ideal scenario would be the one in which the survey covers all passengers boarding and alighting at each bus stop. Since each bus line has its own characteristics, the transit ridership measurement is not the same from one line to another. Therefore, the resulting variable of interest is likely to be different for each line. As lines 856R-10 and line 6913–10 have route sections close to each other in space and belong to the same category (radial/regional), the limitation regarding the variable of interest was relaxed, so we could increase the number of points available for calibrating the models and their spatial coverage. Hence, the case study only covers these two bus lines. However, the proposed method can also be replicated to the remaining bus lines, separately. The variable of interest is the total transit ridership (boardings plus alightings) for the 97 bus stops of the case study lines in a weekday (from 5h to 23h59). Fig. 1 shows the case study lines, and the other bus lines for which information on boarding and alighting was also available.

3.2. Independent variables

Based on Table 2, predictor data were collected from both the catchment area and the bus stop itself. In the case of the catchment area, a network distance of 400 m (Zhao et al., 2003) from the bus stop was considered, using the Open Street Map (OSM) road network. Table 3 consolidates all the information obtained, and respective sources.

An overlapping analysis was also conducted, seeking to calculate the percentage of area from neighboring catchment areas that overlap with the catchment area of the bus stop of reference (Peng et al., 1997). As multiple catchment areas can overlap with a single one, this variable could range from 0 to a ratio higher than 1 (percentage higher than 100%). In addition, area shapefiles went through aerial interpolation for extracting only the data inside the bus stops' catchment area.

Potential multicollinearity in the models was prevented by an exploratory analysis of linear correlation. If a pair of predictors exhibited high correlation (Pearson linear correlation coefficient higher than 0.60) (Profillidis and Botzoris, 2019), the predictor with the lowest correlation with the variable of interest was discarded. After completing the database with the transit ridership variable and potential predictors, the modeling step was started.

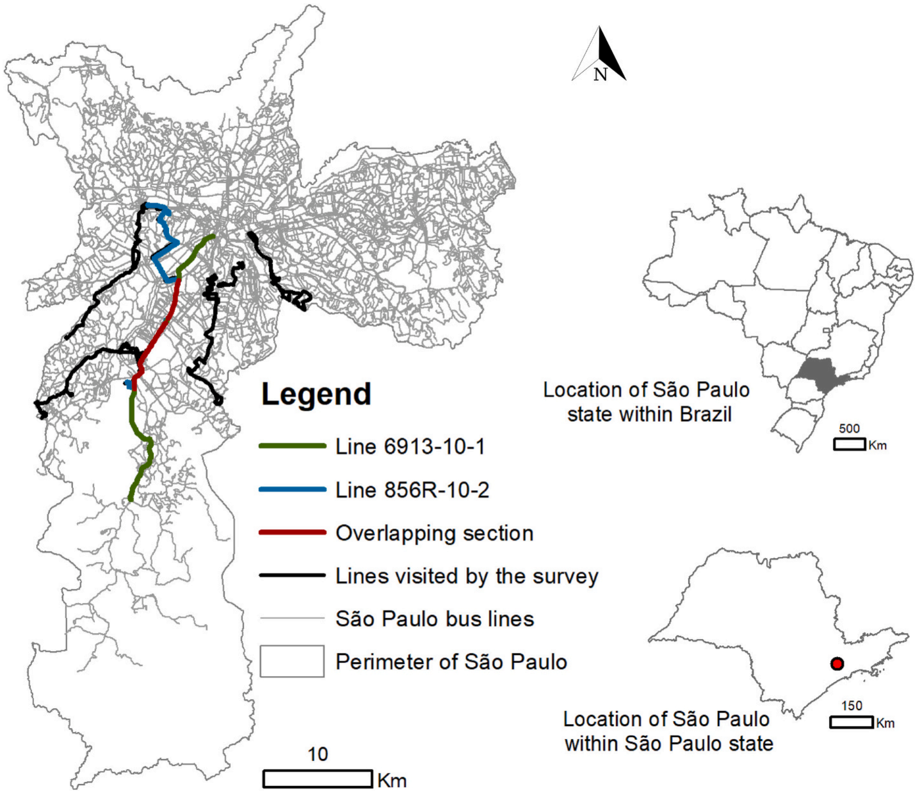

N
Legend
Line 6913-10-1
Line 856R-10-2
Overlapping section
Lines visited by the survey
São Paulo bus lines
Perimeter of São Paulo
10
Km
Location of São Paulo state within Brazil
500
Km
Location of São Paulo within São Paulo state
150
Km

Fig. 1. Case study lines within the São Paulo bus network.

3.3. Modeling

The modeling procedure consisted of the following steps: 1) Poisson Regression (PR); 2) Overdispersion test in the PR model; 3) Negative Binomial Regression (NBR); 4) Spatial dependence and dispersion tests in the NBR model; 5) Geographically Weighted Negative Binomial Regression (GWNBR); 6) Stationarity test in the GWNBR parameters; 7) Spatial dependence test in GWNBR residuals; and 8) Sensitivity analysis and validation. Fig. 2 illustrates the flowchart of the method steps, which are described in the next four subsections.

3.3.1. Global count data models: Poisson and negative binomial regressions

In the first modeling step, the predictor data kept from subsection 3.2 was used to model transit ridership by means of Poisson Regression (PR). After that, the overdispersion test of Cameron and Trivedi (Cameron and Trivedi, 1990) was applied to the resulting Poisson model. Having confirmed the presence of overdispersion in the Poisson Regression, a Negative Binomial Regression (NBR, Eq. 1) (da Silva and Rodrigues, 2014; Hilbe, 2014) was calibrated using the same set of predictors from PR. In order to achieve a parsimonious model, only the explanatory variables with statistically significant coefficients ($p < 0.10$) were kept in the final NBR model.

$$y_j \sim NB\left[t_j exp\left(\sum_k \beta_k x_{jk}\right), \alpha\right] \tag{1}$$

Where y_j is the dependent variable for point j; NB expresses Negative Binomial distribution; t_j is an offset variable; x_{jk} stands out for the explanatory variable k with associated parameter β; and α is the overdispersion parameter. As both Poisson and Negative Binomial regressions use, traditionally, an exponential link function, PR is also expressed by Eq. 1, except for the assumed probability distribution, which is the Poisson one in the case of PR (Hilbe, 2014), and the overdispersion parameter.

The overdispersion test of Cameron and Trivedi is commonly used only in the Poisson Regression context. However, a dispersion test described by Hilbe (2014) was applied to both PR and NBR models. The statistic of this test is based on the Pearson residuals, which account for the variance function imposed by the model. In the PR case, the variance equals the expected mean, while the variance modeled by NBR increases as a squared function of the mean. Assuming that this statistic follows a Chi-square distribution (Hilbe, 2014), one can say whether the variance accounted for by the chosen model appropriately fits the data variance. Rejecting the null hypothesis of dispersion equal to 1 indicates the presence of overdispersion in the Poisson model, if the test statistic is higher than 1.

Spatial dependence in the NBR model was attested by applying the Moran Index (Moran, 1948) to its residuals. In this step, the spatial weights matrix for calculating the Moran's I was based on the network distances between bus stops, using, as the network, the route of line 856R-10-2 and line 6913-10-1.

3.3.2. Local count data model incorporating overdispersion: Geographically weighted negative binomial regression

The next step was to calibrate the Geographically Weighted Negative Binomial Regression (GWNBR, Eq. 2) (da Silva and Rodrigues, 2014), using, as explanatory variables, the same data from NBR. GWNBR, in addition to accounting for the overdispersion of transit ridership, can address spatial dependence and spatial heterogeneity of estimated parameters. As a local model, it consists of calibrating a NBR model at each point of the database, by using weighted neighbor data, both of dependent and independent variables. However, the information from the point where the model is being calibrated is omitted.

$$y_j \sim NB\left[t_j exp\left(\sum_k \beta_k\left(u_j, v_j\right) x_{jk}\right), \alpha\left(u_j, v_j\right)\right] \tag{2}$$

Where (u_j, v_j) are the geographic coordinates of point j.

The weighting scheme of geographically weighted regressions, called kernels, is based on the distance between point j at which the model is

Table 3
Potential predictor collection.

Predictor(s)	Originated from	Source
Network distance, in meters, to the city center *Praça Sé* (Sé Square)[1]	bus stop	GeoSampa[2] shapefile and Open Street Map
Network distance, in meters, to the nearest bus terminal		
Network distance, in meters, to the nearest metro station		
Network distance, in meters, to the nearest train station		
Number of bus lines passing by the bus stop, except the line(s) of interest		2017 GTFS data provided by SPTrans
Average frequency, in trips per hour, of the bus lines passing by the bus stop, except the line(s) of interest		
Number of shelters		SPTrans
Presence/absence (1/0) of shelter and seat		
Population, in inhabitants	catchment area	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 (MHSHR); low standard vertical residential; medium/high standard vertical residential; commerce and services; industry and warehouses; residential, commerce and services; residential, industry and warehouses; commerce, services, industry and warehouses; public facilities; schools; empty land; and land with no predominant use		2016 GeoSampa shapefile, given in blocks
Entropy index (Song et al., 2013)		–
Employment		2018 Annual List of Social Information
Average household income		Household data from the 2017 Origin and Destination survey (Metrô, 2019)
Female (%)		
Population with a bachelor's degree or higher (%)		
Households with no private vehicles (%)		
Youth (up to 17 years) (%)		
Elderly (65+) (%)		
Park area, in hectares		GeoSampa shapefiles
Sidewalk length, in meters		
Arterial length, in meters		
Bus lanes length, in meters		
Bicycle path length, in meters		
Number of intersections		Open Street Map
Overlapping area ratio		–

[1] *Praça Sé* is a public space considered as the geographic center of São Paulo.
[2] http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx [Accessed in June 2022].

being calibrated and its neighbors *i*. Eq. 3 (Fotheringham et al., 2003) shows the bisquare kernel, the weighting function adopted for the present study.

$$W_j(i) = \begin{cases} \left(1 - d_{ij}^2 \big/ b^2\right)^2, if\ d_{ij} \le b \\ 0, if\ d_{ij} > b \end{cases} \quad j = 1, 2, \dots n, \tag{3}$$

Where *W* is the weight assigned to neighbor *i* for calibrating GWNBR in *j*; d_{ij} is the network or Euclidean distance between *i* and *j*; and *b* is the bandwidth distance. Fig. 3 illustrates the importance of using network distances in the current case study.

The Euclidean distance between bus stops 2 and 3 is smaller than between 2 and 1, but 2 and 1 belong to the same bus line (line 1). Following the concept of spatial dependence, points close to each other belonging to the same bus line are assumed to be more related than points from different bus lines. However, according to Eq. 3, if we use Euclidean distances, point 3 from line 2 will have more influence than point 1 on the calibration of GWNBR in point 2, but this inconvenience is prevented when network distances are applied.

Eq. 3 also shows that, unlike NBR and PR, GWNBR does not use all points for calibration. Instead, only points spatially correlated to a point where the model is being calibrated participate in the calibration process, and the point itself is not included to avoid overfitting. The autocorrelation range is controlled by the bandwidth, which can be either a fixed or an adaptive distance. In cases where the distance between database points varies significantly, the adaptive case is preferable, as it uses the same number of neighbors for all points.

To achieve the model with best results, this bandwidth can be optimized by meeting some well-defined criterion, such as to minimize the cross-validation error (CV, (Fotheringham et al., 2003)) or the corrected Akaike Information Criterion (AICc). In the case of GWNBR, AICc depends on the effective number of parameters due to β and α. As the effective number of parameters due to α remains undefined (Gomes et al., 2017), the authors opted to minimize the CV (Eq. 4) as the criterion to obtain the optimum bandwidth.

$$CV = \sum_{j=1}^{n} \left[y_j - \widehat{y}_{\neq j}(b) \right]^2 \tag{4}$$

Where $\hat{y}_{\neq j}$ is the estimated value of *y* using the bandwidth *b*, and *n* is the total number of points (96). Subscript $\neq j$ recalls that the real value of *y* in *j* is not used in the estimation process.

3.3.3. Stationarity test

Finally, the better fit of GWNBR over global models was attested by two means: 1) applying the Moran's I to the residuals of GWNBR to verify if the model was able to address the spatial dependence found in NBR; 2) by a stationarity test, to assess whether or not the hypothesized spatially varying relationships are statistically significant. In this context, the test hypotheses are outlined as follows (Leung et al., 2000):

$$\mathrm{H}_0 : \beta_{1k} = \beta_{2k} = \dots = \beta_{nk}, for\ a\ given\ k,$$

$$\mathrm{H}_1 : not\ all\ \beta_{jk}\ (j = 1, 2, \dots, n)\ are\ equal.$$

999 permutations of the original database were carried out, by replacing the geographic coordinates of the 96 bus stops for the coordinates of another point of the database. Therefore, from the 999 GWNBR models generated, a measure of the βs variance could be calculated, in which the test statistic is based. The statistic follows an *F*-distribution, which provides the critical value for rejecting or accepting H_0, given an adopted significance level (Leung et al., 2000). GWNBR modeling, including the search for optimum bandwidth, the stationarity test, and Moran's I verification, were carried out for both types of distance (network and Euclidean), and therefore their results could be compared.

3.3.4. Sensitivity analysis and validation

In the sensitivity analysis, the GWNBR model was calibrated for five scenarios with an increasing percentage of missing data: 15%, 30%, 45%, 60% and 75%. From scenario 0, in which all 97 bus stops are used, five calibration samples were stablished containing the following number of bus stops: 82, 68, 53, 39 and 24. Hence, the validation samples (missing data) had 15, 29, 44, 58 and 73 bus stops, respectively.

The sample selection was based on the density of points in the original database. A kernel density function is calculated for each point

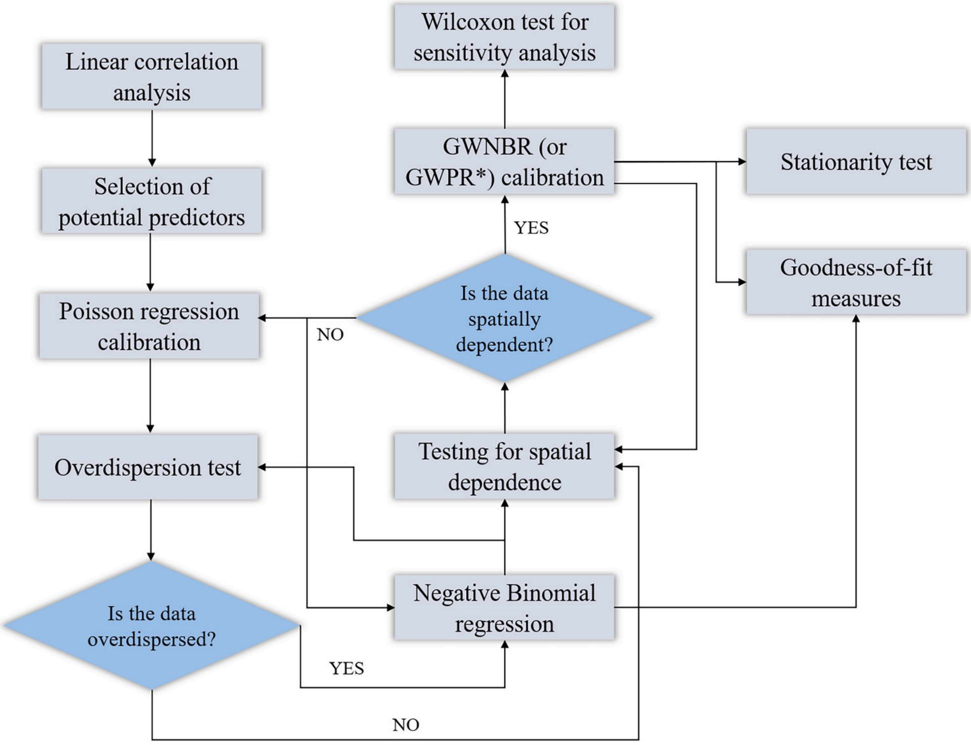

Linear correlation analysis
Selection of potential predictors
Poisson regression calibration
Overdispersion test
Is the data overdispersed?
YES
NO
Negative Binomial regression
Testing for spatial dependence
Is the data spatially dependent?
NO
YES
GWNBR (or GWPR*) calibration
Wilcoxon test for sensitivity analysis
Stationarity test
Goodness-of-fit measures

Fig. 2. Method flowchart. * Geographically Weighted Poisson Regression.

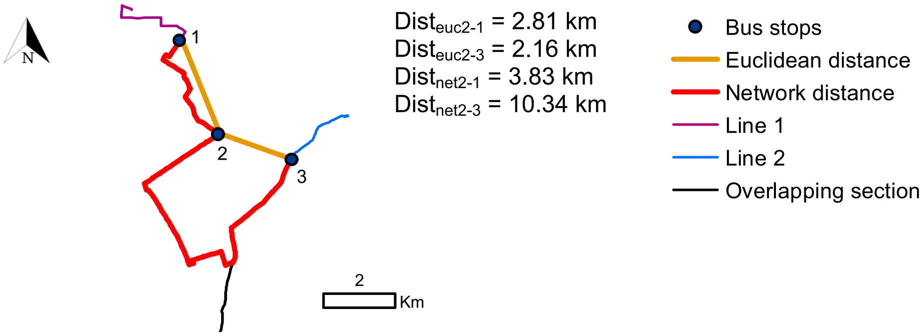

N
1
2
3
Dist$_{euc2-1}$ = 2.81 km
Dist$_{euc2-3}$ = 2.16 km
Dist$_{net2-1}$ = 3.83 km
Dist$_{net2-3}$ = 10.34 km
● Bus stops
Euclidean distance
Network distance
Line 1
Line 2
Overlapping section
2
Km

Fig. 3. Comparison between network and Euclidean distances.

in the database. Higher values of the function are assigned to points located in high-point-density areas. From this method, points with higher values of the density function have a higher probability of being selected. We assumed that regions with high density of stops are served by a higher number of bus lines. Therefore, the higher the number of lines, the higher the probability of having information on boarding and alighting at bus stops located within these regions.

As geographically weighted regressions can obtain a different model for all geographic units in the database, the GWNBR model was calibrated for the 97 bus stops in all scenarios using the same set of predictors defined in 3.3.1 and the two types of distance. Therefore, for each scenario, all bus stops had estimates of both β parameters and the variable of interest. Taking the GWNBR results for the entire database as the estimates closest to reality, the βs' sensitivity to missing data was assessed by applying the non-parametric Wilcoxon test (Wilcoxon, 1945) to the following pair of parameters: βs from scenario 0 versus βs from the scenario with missing data. The analyses were carried out for the network and Euclidean distances separately, and results from calibration and validation samples were also analyzed separately. Afterward, the Wilcoxon test was also applied to comparing β estimates from GWNBR using network and Euclidean distances in all the scenarios stablished.

The prediction power's sensitivity to the increasing percentage of missing data was assessed by the following goodness-of-fit measures: Median of Absolute Percentage Error, Root Mean Squared Error and Mean Absolute Error (Hollander and Liu, 2008). Once more, calibration and validation samples were analyzed separately.

3.3.5. *Computational tools*

The Poisson Regression, overdispersion test, Negative Binomial Regression, and the Moran Index were calculated using the open source programming language R (Kleiber and Zeileis, 2008; Paradis et al., 2004; R Core Team, 2021; Venables and Ripley, 2002). The GWNBR calibration, search for the optimum bandwidth and stationarity test were carried out through the code provided by da Silva and Rodrigues (2016) in the Statistical Analysis Software (SAS® 9.4). The original routine was slightly modified to accommodate the network distances, which were obtained in GRASS GIS 7.4.0, using, as the network, the route of line 856R-10-2 and line 6913-10-1. Euclidean distances were obtained in QGIS 2.18.12.

The calibration samples for the sensitivity analyses were randomly selected from the original database using the R library “spatialEco” (Evans, 2021). The non-parametric Wilcoxon test for matched pairs was run in the software IBM SPSS Statistics 22.

4. Results and discussion

Fig. 4 shows the spatial variation of transit ridership (boardings plus alightings from 5h to 23h59 of 2017-11-07) along the case study lines, together with some characteristics of the surrounding area. The location of the bus lines within São Paulo is also shown.

The maximum value of transit ridership occurs in the first stop of line 6913-10-1, the *Varginha* Terminal, situated in a predominantly residential area. Travelling from south to north, the case study lines approach areas with high employment densities, which might be the destination of most users by the morning and their origin point at the end of the day. Therefore, high volumes of passengers can be seen even at bus stops outside the overlapping section, as they are closer to the city center *Praça Sé* (Sé Square), which concentrates the highest employment densities in São Paulo.

Table 4 shows the measures of central tendency and dispersion for the dependent variable and independent variables associated to statistically significant parameters ($p < 0.10$) kept from Negative Binomial Regression, which include some features presented in Fig. 4.

As expected, the median of transit ridership is lower than its mean, causing the distribution of the data to be skewed to the right. In addition, the standard deviation is much higher than the mean, pointing out to the wide range of variation and overdispersion of the dependent variable. Due to the relatively small size of the catchment areas, variables such as the MHSHR area and area with no predominant land use are not present in the influence region of all bus stops. However, the first two variables show significant variation over the case study bus lines. The remaining variables also show great amplitude of variation. The overlapping variable, for example, reveals that most bus stops have neighboring bus stops located within a distance smaller than 400 m along the road network. As the catchment area depends on the street layout, some of them can be very small compared to the Euclidean-distance buffer. In the current case study, catchment area overlapping reaches a percentage of almost 200%.

For the sake of brevity, Poisson Regression results will not be shown. However, the Cameron and Trivedi’s test rejected the null hypothesis of equidispersion in the Poisson model in favor of the alternative hypothesis of true dispersion >1 (p-value = 0.000062). Following the proposed method, the NBR model was calibrated after this. The next subsection presents the results from NBR and GWNBR.

4.1. Factors affecting the stop-level transit ridership

Estimated coefficients of NBR and GWNBR are presented in Table 5. In turn, Fig. 5 shows the spatially varying parameters of GWNBR. Maps

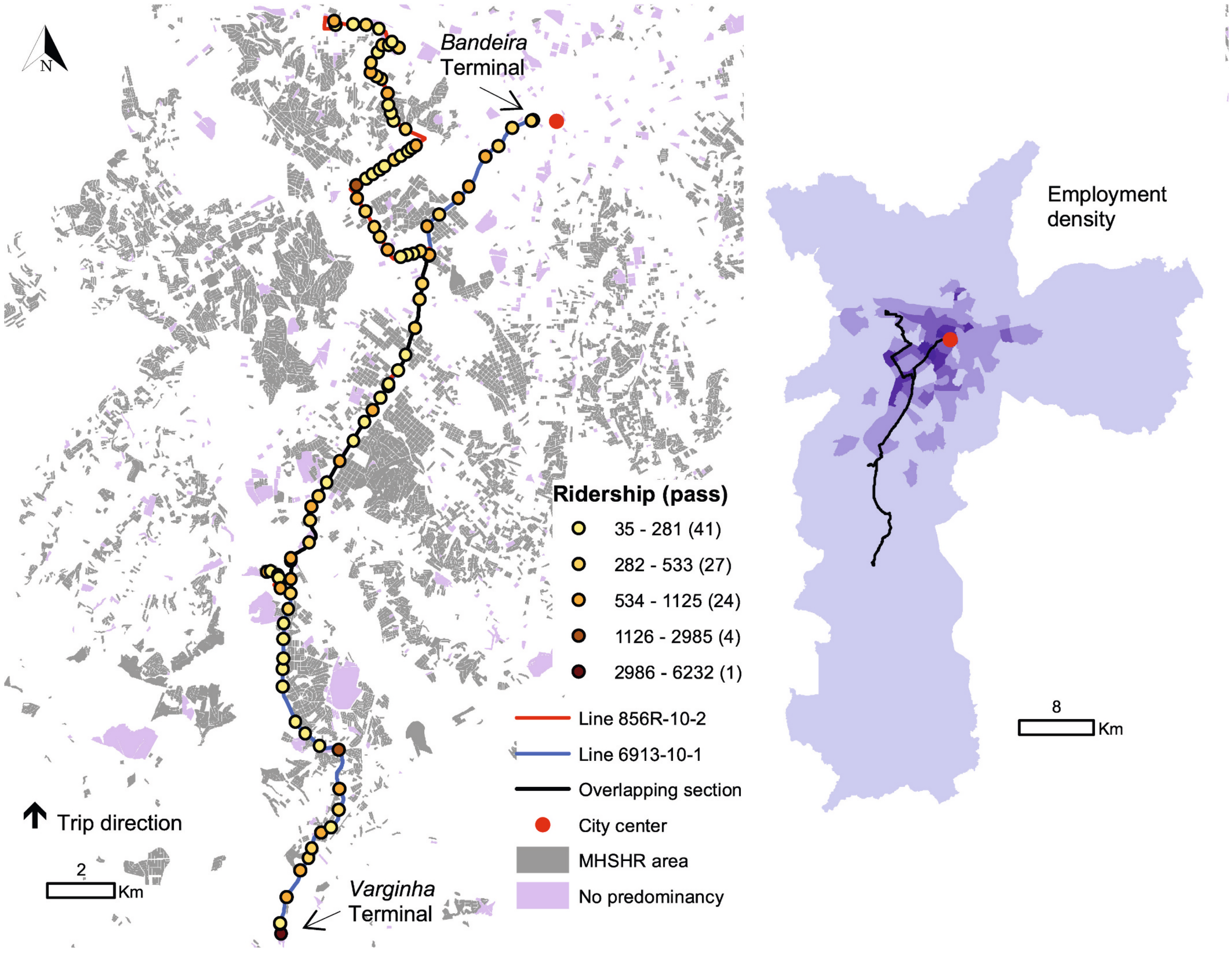


Fig. 4. Map of transit ridership along the case study lines and their location within São Paulo.

Table 4
Descriptive statistics of dependent and independent variables ($N = 97$).

Variable	Mean	Std. Dev.	Min.	25%	50%	75%	Max.
Transit ridership (pass)	495.103	702.495	35.000	200.500	356.000	553.000	6232.000
MHSHR* area (ha)	1.928	2.562	0.000	0.000	0.655	3.347	10.317
No predominance area (ha)	0.749	1.079	0.000	0.000	0.093	1.221	4.611
Overlapping area ratio	0.716	0.572	0.000	0.224	0.578	1.148	1.966
Intersections	2639.340	1446.784	576.000	1750.000	2326.000	2970.000	8232.000
Frequency (trips/h)	3.306	1.232	0.000	2.915	3.081	3.521	9.344

* Medium/High Standard Horizontal Residential.

Table 5
Summary of results for NBR and GWBR.

Variable	NBR	Network GWNBR					Euclidean GWNBR				
		Min	25%	50%	75%	Max	Min	25%	50%	75%	Max
Intercept	4.96	4.33	4.36	5.75	5.90	6.10	4.31	4.34	5.82	5.87	5.91
MHSHR area	−0.08	−0.10	−0.09	−0.07	−0.06	−0.05	−0.08	−0.07	−0.07	−0.06	−0.05
No predominant land use	0.11	0.06	0.07	0.18	0.22	0.22	0.06	0.07	0.19	0.19	0.20
Overlapping area ratio	−0.21	−0.50	−0.50	−0.43	0.06	0.10	−0.42	−0.41	−0.40	0.05	0.08
Intersections (10e-3)	0.27	0.26	0.28	0.29	0.30	0.30	0.26	0.27	0.27	0.30	0.31
Frequency	0.17	−0.20	−0.10	−0.05	0.28	0.31	−0.11	−0.10	−0.09	0.28	0.30
Alpha	0.35	0.23	0.24	0.27	0.31	0.34	0.27	0.28	0.28	0.30	0.33
AIC	1330.60	1312.26*					1313.18*				
MedAPE	42.99%	36.22%					36.56%				
RMSE	687.20	271.96					285.01				
MAE	289.09	184.59					186.86				

Note: MHSHR, AIC, MedAPE, RMSE and MAE stands for Medium/High Standard Horizontal Residential, Akaike Information Criterion, Median of Absolute Percentage Error, Root Mean Squared Error and Mean Absolute Error; * considering the number of effective parameters due to α as 1.

of coefficients from both types of distance are shown, highlighting the non-statistically significant ones for a 90% confidence level.

Five predictors explained the transit ridership variable better: medium/high standard horizontal residential area, area with no predominant land use, overlapping, intersections and frequency. Table 5 confirms the better fit of GWNBR over NBR. The best model is the network case of GWNBR.

The optimum bandwidth obtained by minimizing the CV error was 77 and 83 neighbors, considering network and Euclidean distances, respectively. Fig. 5 shows that all variables had a majority of coefficients with a sign equal to the NBR result. However, there are clear differences between the results from NBR, network GWNBR and Euclidean GWNBR. As the MHSHR area refers to families with medium/high income, the prevailing negative effects are understandable. Conversely, areas with a higher mixture in the land use have a strong tendency to contribute to increasing transit ridership at bus stops.

In the case of the overlapping area variable, the bus stops in the north part of Fig. 5 are under the highest rates of expected decrease in transit ridership if new bus stops are added to their surrounding area. Considering only statistically significant coefficients and network distances, the GWNBR results state that a 1-unit increase in the overlapping area ratio is associated with a decrease varying from −31.20% to −39.49% if the other attributes are held constant. However, this effect is underestimated to −7.69% in the global Negative Binomial Regression.

Together with the intercept, all coefficients from the intersection variable were statistically significant ($p < 0.10$). We recall that the greater the number of intersections, the greater the number of possible paths to be taken, which affects the walking distance to the bus stop and to possible users' destinations. In addition, in the linear correlation analysis (subsection 3.2), the intersection variable was positively correlated (with statistical significance) to the number of jobs, percentage of no vehicle households, and length of sidewalks, arterial roads, bus lanes and bike lanes. Therefore, neighborhoods with high density of intersections are richly supplied by the transport infrastructure, of both motorized and active modes.

Regarding the frequency variable, Euclidean results yielded only positive statistically significant coefficients, whose effects vary from 27.07% to 35.57% if the average frequency of lines passing though the bus stop increase by 1 trip/h, *ceteris paribus*. However, in the network approach, this range is from −18.37% to 35.72%. While positive impacts refer to potential intramodal integration, the negative ones may occur in areas with high bus frequency, but relative low transit ridership. Table 6 consolidates a comparison between the results found in the present case study and Table 2.

Except for the frequency variable, all explanatory variables showed only positive or negative statistically significant coefficients in the GWNBR models, which coincides with most results found in the literature. This suggests that, in spite of showing great spatial variability, the explanatory variables of transit ridership tend to show only one type of impact: positive or negative. In turn, negative statistically significant effects of frequency have already been seen in Marques and Pitombo (2022). A predictor correlated to frequency, the number of lines through stops, was also negatively related to boardings and alightings in Chakour and Eluru (2016). However, in local models, coefficients' extreme signs may be caused by the criterion for finding the optimum bandwidth. Minimizing the cross-validation error can yield small bandwidths, which usually provide the highest levels of spatial variability in the estimated parameters (Farber and Páez, 2007).

4.2. Statistical significance of the spatial variation

Table 7 shows the *p*-values for each estimated parameter in the stationarity test, highlighting, in bold and italic, the statistically significant ones ($p < 0.10$).

By adopting a 90% level of confidence, the intercept, overlapping area ratio and frequency had a statistically significant spatial variation. The conclusion is that the estimated parameters showed enough evidence to reject the null hypothesis of the same value throughout all bus stops, as imposed by global models.

4.3. Impact of missing data

The spatial distribution of the calibration samples defined in subsection 3.3.4 are shown in Fig. 6. Six scenarios were analyzed: scenario

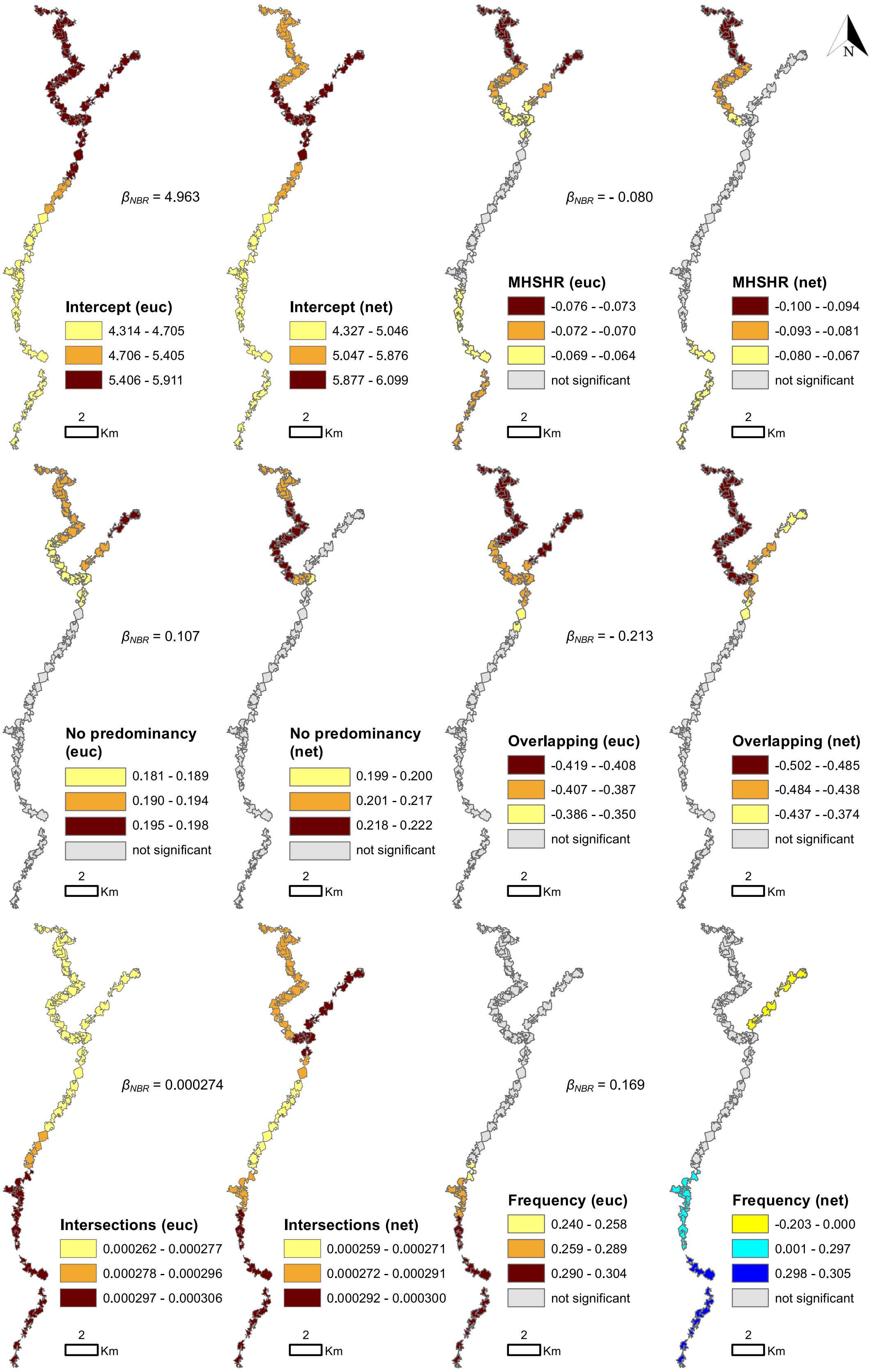


Fig. 5. NBR global coefficients and maps of estimated GWNBR coefficients (distance in parenthesis).

Table 6
Comparison with previous studies.

Predictor	Equivalent predictor in Table 2	Number of previous studies from Table 2		Sign of statistically significant coefficients in the current study	
		Positive effect	Negative effect	Euclidean	Network
MHSHR* area	Income	1	10	−	−
No predominant land use	Land use mix	4	0	+	+
Overlapping area ratio	Bus stops in a buffer	1	8	−	−
Intersections	Street connectivity	9	1	+	+
Frequency	Headway (inverse)	4	0	+	+/−

* Medium/High Standard Horizontal Residential.

Table 7
Stationarity test p-values.

Parameter	P-value	
	Euclidean	Network
Intercept	***0.006***	***0.011***
MHSHR* area	0.846	0.629
No predominant land use	0.177	0.173
Overlapping area ratio	***0.026***	***0.026***
Intersections	0.870	0.989
Frequency	***0.003***	***0.001***
Alpha	0.780	0.503

* Medium/High Standard Horizontal Residential.

0, consisting of the complete database with no missing data ($N = 97$). Scenarios 1, 2, 3, 4 and 5 have an increasing percentage of missing data (from 15% to 75% with intervals of 15%).

Table 8 summarizes the Wilcoxon-test p-values for the comparison between β parameters from the scenario 0 GWNBR model and the scenarios with missing data. Network cases with missing data are compared to the GWNBR model with network distances from scenario 0, and Euclidean cases with missing data are compared to the complete GWNBR model with Euclidean distances. The null hypothesis of the Wilcoxon test for matched pairs stablishes that the median of the differences between the distributions being compared equals zero. Therefore, Table 8 highlights the cases where the test identified no statistically significant differences between the β parameters for a 90% confidence level.

Table 8 shows that the β parameters of the intersections, overlapping and frequency variables of some models with missing data had no statistically significant differences from the βs in the complete GWNBR model. Surprisingly, the calibration and validation samples from the scenario with the highest percentage of missing data (scenario 5) were the only cases where two explanatory variables (overlapping and frequency) had effects similar to those of the original GWNBR model, considering both types of distance. This result may be due to the fact that, despite being the scenario with the least number of bus stops, their spatial distribution was able to replicate the same spatial pattern of the scenario 0, regarding these two variables. Noting that overlapping and frequency showed statistically significant spatial variation, this outcome suggests that local models with missing data can reproduce the whole spatial pattern of impacts from some factors on transit ridership.

Therefore, we cannot confirm the hypothesis that the results worsen as the percentage of missing data increases, at least for some predictors. However, there is a strong impact of missing data in the predictors' effects. Most cases analyzed had statistically significant differences between the scenarios with missing data and the complete database model. This might be due, among other factors, to the bandwidth obtained for each analysis. The Euclidean bandwidth in scenarios 0, 1, 2, 3, 4 and 5 was 83, 78, 35, 22, 37 and 23 neighbors, respectively, while in the network case it was 77, 73, 27, 52, 37 and 23. The bandwidth is strongly associated to the degree of spatial dependence found in each database (Fotheringham et al., 2022), and this factor was modified in the missing data scenarios, causing the β parameters to vary from one scenario to another. Despite that, variables such as frequency still had coefficients similar to the complete GWNBR model in almost 50% of the cases analyzed.

The prediction power's sensitivity to missing data and type of distance is shown in Fig. 7, in which three goodness-of-fit measures are presented: Median of Absolute Percentage Error, Root Mean Squared Error and Mean Absolute Error. Results from calibration and validation samples are displayed on the left and right, respectively.

As expected, errors are smaller in the calibration samples than in the validation ones, pointing to a negative impact of missing data on the prediction accuracy. However, errors do not increase monotonically by increasing the percentage of missing data, as previously hypothesized. Scenario 5 tended to present the worst validation estimates, followed by scenarios 2 and 3 (30% and 45% of missing data, respectively). Conversely, scenario 4 (60% of missing data) showed consistently good estimates in both calibration and validation samples.

Considering the MedAPE results, errors are limited to 40% in the calibration samples, and to 50% in three of the five validation scenarios stablished. This assures that GWNBR is capable of estimating the transit ridership for uncounted locations with relative accuracy. Mucci and Erhardt (2018) found a validation error of −11% comparing the total observed transit ridership (for a year other than the one used for calibration) with the value estimated by their model. In our case study, this error metric for the network approach was 31.15%, −19.01%, 1.45%, 3.58% and − 9.23%, from the first to the fifth scenario.

4.4. *Impact of network distances*

From Fig. 7, network distances performed better than the Euclidean ones in the calibration sample of scenarios 0, 1, 2 and 4. This could partially confirm the hypothesis of better performance from network distances as the percentage of missing data increases. Considering the validation samples, the network case yields better results only until scenario 3. Higher differences between goodness-of-fit measures from network and Euclidean distances occur in scenarios 2 and 3.

Similar outcomes from network and Euclidean distances might be due the kernel formulation (Eq. 3), responsible for assigning weights to the neighboring data: when the network distance between the database points is divided by the bandwidth network distance, the resulting number is similar to the one obtained by Euclidean distances, given that the bandwidth is also a distance. In addition, as the biggest weights are commonly assigned to the nearest bus stops, the difference between network and Euclidean distances for these pairs of points might not be great enough to cause significant disparities in outcomes from both types of distance. However, network distances could, in fact, provide better transit ridership estimates in some cases from Fig. 7. Table 9 presents the results of the comparison between coefficients from the network and Euclidean approaches.

Statistically significant differences were observed in most cases shown in Table 9. On the other hand, for each explanatory variable, at least three cases had similar effects from network and Euclidean distances. Regarding the frequency variable, network and Euclidean coefficients were alike in more than half of the conditions analyzed. In these cases, using network distances would yield parameters with no

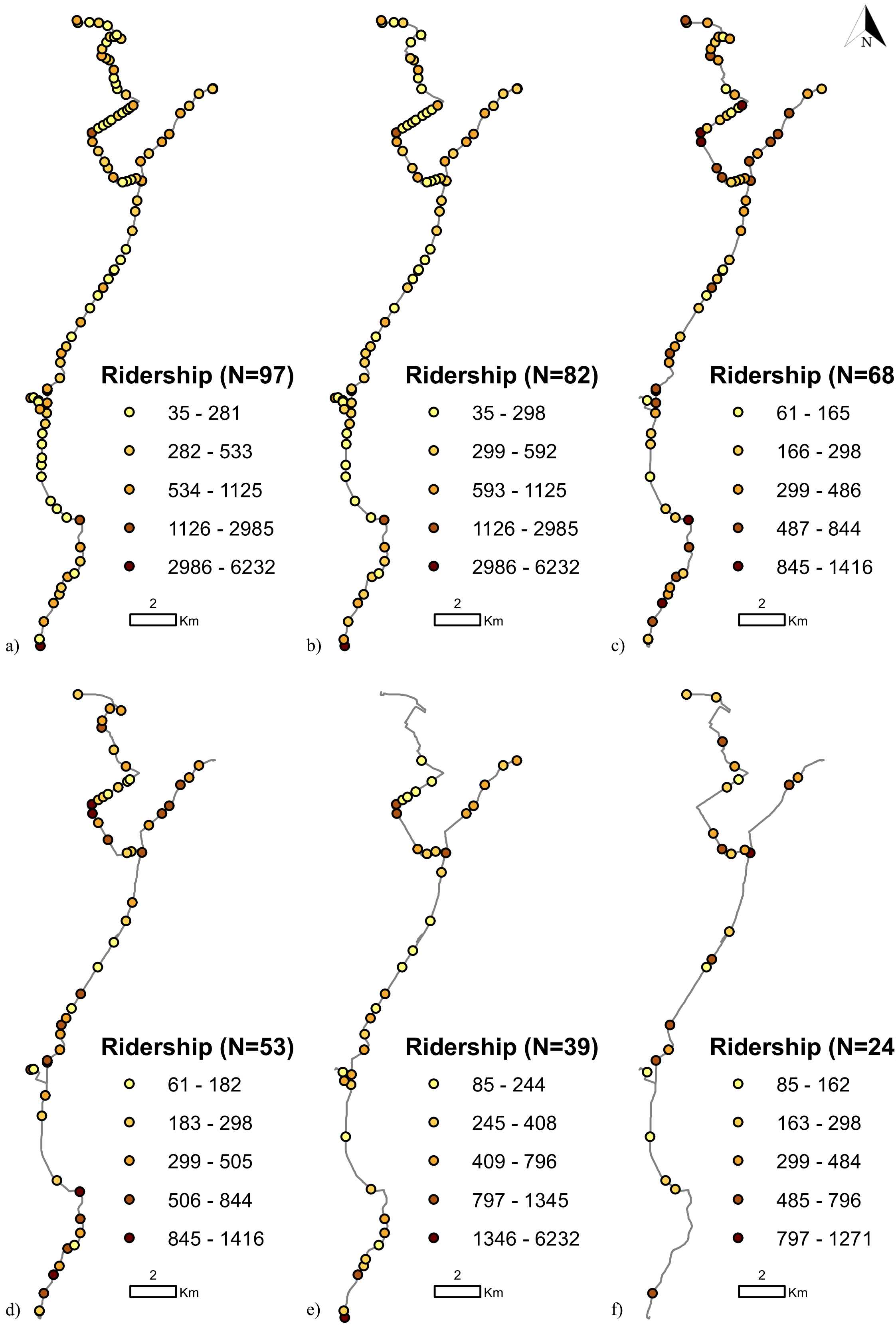


Fig. 6. Calibration samples for sensitivity analysis: scenarios 0 (a), 1 (b), 2 (c), 3 (d), 4 (e) and 5 (f).

Table 8
Coefficients' sensitivity to missing data.

Scenario	Sample	Dist.	MHSHR	No predominant land use	Intersections	Overlapping	Frequency
1	Calibration	Euc	0.000	0.000	***0.574***	0.000	***0.713***
		Net	0.000	0.000	0.037	0.000	***0.766***
	Validation	Euc	0.011	0.001	***0.733***	0.011	***0.865***
		Net	0.011	0.001	***0.733***	0.011	0.003
2	Calibration	Euc	0.000	0.000	0.000	0.000	0.004
		Net	0.000	0.008	***0.951***	0.000	0.053
	Validation	Euc	0.000	0.005	0.000	0.000	***0.405***
		Net	0.000	0.016	***0.393***	0.000	***0.315***
3	Calibration	Euc	0.087	0.051	***0.611***	0.032	0.000
		Net	0.000	0.000	0.001	0.044	0.000
	Validation	Euc	0.004	***0.552***	0.018	***0.815***	0.000
		Net	0.000	0.000	0.001	***0.852***	0.000
4	Calibration	Euc	***0.812***	0.000	***0.686***	0.015	0.001
		Net	0.006	0.000	***0.900***	0.000	0.089
	Validation	Euc	0.005	0.000	0.005	0.000	0.000
		Net	0.000	0.000	0.010	0.000	0.000
5	Calibration	Euc	0.000	0.007	0.000	***0.841***	***0.819***
		Net	0.000	0.015	0.000	***0.648***	***0.648***
	Validation	Euc	0.000	0.000	0.000	***0.453***	***0.266***
		Net	0.000	0.000	0.000	***0.547***	***0.858***

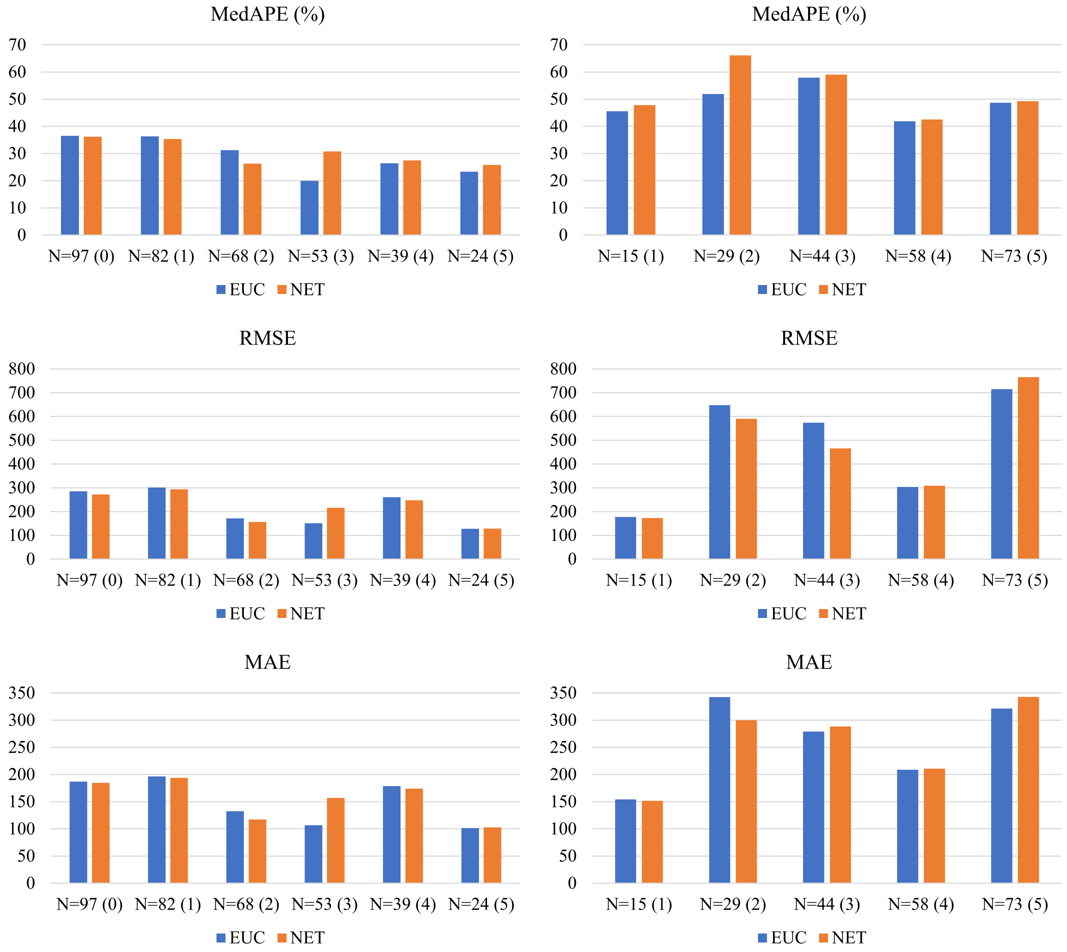


Fig. 7. Prediction power's sensitivity to missing data and type of distance (scenario in parenthesis). Calibration results on the left and validation on the right.

statistically significant differences from the Euclidean ones.

Recalling the results shown in Fig. 7, the better suitability of network or Euclidean distances may depend on the explanatory variable considered. As stated by Ver Hoef (2018), some spatial relationships are linked to processes that operate more in the Euclidean space than in the linear one, and vice-versa. Therefore, the best modeling option for these cases would be a geographically weighted model with parameter-specific distance metrics (Lu et al., 2017).

4.5. Summary of results

Fig. 8 illustrates how the limitation each model addresses is identified. Results and conclusions from the dispersion and spatial dependence tests are also provided, guiding the decision for adopting another

Table 9
Coefficients' sensitivity to the type of distance.

Scenario	Sample	MHSHR	No predominant land use	Intersections	Overlapping	Frequency
0	Calibration	0.001	0.000	0.000	0.000	***0.140***
	Validation	–	–	–	–	–
1	Calibration	***0.389***	0.000	0.072	0.000	0.000
	Validation	0.011	0.001	***0.776***	0.001	***0.427***
2	Calibration	0.006	***0.413***	0.006	0.000	***0.271***
	Validation	0.002	0.018	0.045	0.000	***0.991***
3	Calibration	0.001	0.000	***0.787***	***0.794***	0.008
	Validation	0.000	0.000	0.030	***0.161***	0.023
4	Calibration	0.000	0.081	0.012	0.000	***0.706***
	Validation	0.000	0.016	***0.187***	0.001	0.001
5	Calibration	***0.110***	***0.977***	0.009	***0.092***	***0.179***
	Validation	***0.140***	***0.179***	0.000	0.006	0.080

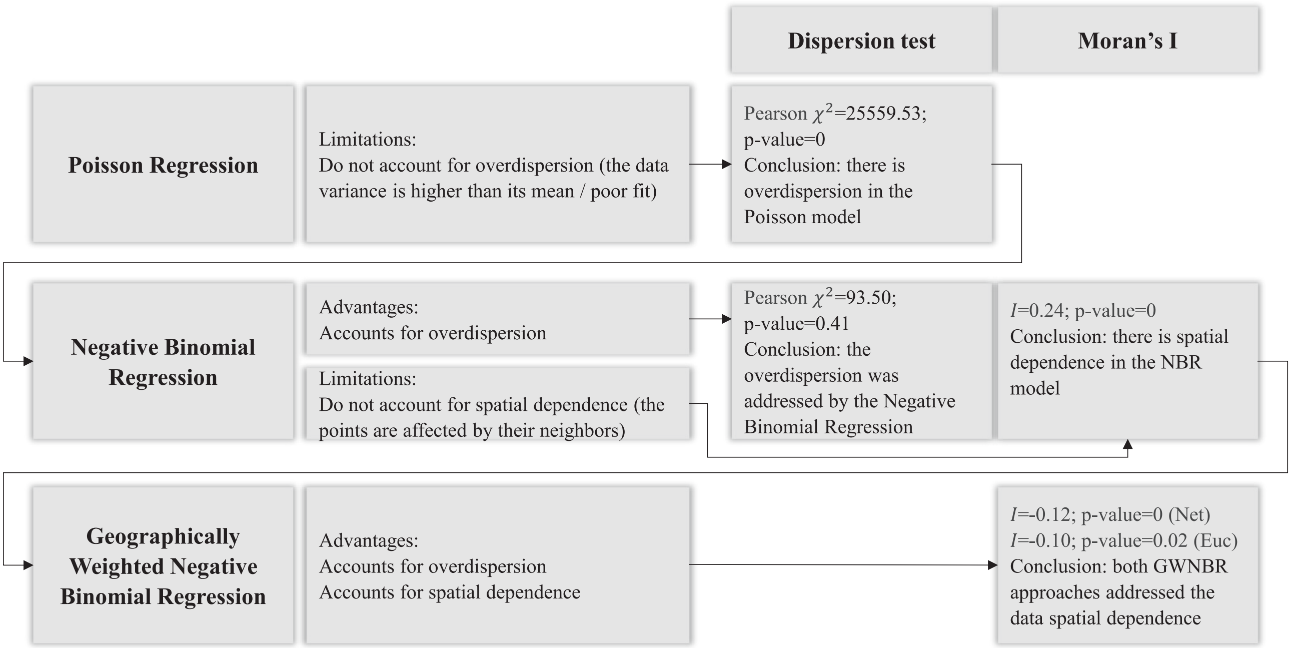

Dispersion test
Moran's I
Poisson Regression
Limitations:
Do not account for overdispersion (the data variance is higher than its mean / poor fit)
Pearson χ^2=25559.53; p-value=0
Conclusion: there is overdispersion in the Poisson model
Negative Binomial Regression
Advantages:
Accounts for overdispersion
Limitations:
Do not account for spatial dependence (the points are affected by their neighbors)
Pearson χ^2=93.50; p-value=0.41
Conclusion: the overdispersion was addressed by the Negative Binomial Regression
I=0.24; p-value=0
Conclusion: there is spatial dependence in the NBR model
Geographically Weighted Negative Binomial Regression
Advantages:
Accounts for overdispersion
Accounts for spatial dependence
I=-0.12; p-value=0 (Net)
I=-0.10; p-value=0.02 (Euc)
Conclusion: both GWNBR approaches addressed the data spatial dependence

Fig. 8. Flowchart of results from addressing each model limitation.

Table 10
Synthesis of the results from each method step.

Method step	Results
Overdispersion test on Poisson Regression	The data is overdispersed. Therefore, the Negative Binomial distribution is more appropriate than the Poisson distribution to model the transit ridership variable
Negative Binomial Regression	Five predictors had statistical significance: medium/high standard horizontal residential area, area with no predominant land use, overlapping area ratio, intersections, and frequency
Dispersion test on Negative Binomial Regression	The NBR model properly accounted for the overdispersion found in the Poisson Regression
Moran's I on residuals from Negative Binomial Regression	Statistically significant spatial dependence detected in the NBR model
Moran's I on residuals from Geographically Weighted Negative Binomial Regression	The spatial dependence was completely incorporated in both GWNBR approaches.
Comparison between GWNBR and NBR	Goodness-of-fit results from GWNBR were better than those of NBR. GWNBR had a MedAPE of 36%, meaning that half of the database underestimated or overestimated from 0% up to 36% the real number of riders per day. The remaining points underestimated or overestimated the real number of riders per day in a percentage higher than 36%. MedAPE for NBR was 42%. Comparing these two percentages, the improvement is of 14%.
Stationarity test on GWNBR results	Parameters from the intercept, overlapping area ratio and frequency variables had statistically significant spatial variation
Parameters' sensitivity to missing data	Intersections and frequency were the variables with more prevailing coefficients. No pattern of sensitivity was identified based on the number of available points for calibration. The spatial sampling strategy may be fundamental to the parameter consistency under missing data conditions.
Prediction power sensitivity to missing data	Some scenarios with high percentage of missing data presented errors close to the results from the model using the entire database
Comparison between network and Euclidean distances	Frequency was the variable with more occurrences of similar coefficients between network and Euclidean approaches. No pattern of the best approach was found based on goodness-of fit measures. The more appropriate distance metric may vary from one explanatory variable to another

modeling approach.

Table 10 summarizes all results achieved.

5. Conclusions and final considerations

To tackle the lack of transit ridership data and some other issues, the present paper proposed the application of Geographically Weighted Negative Binomial Regression to six data scenarios, in which five of them had an increasing percentage of missing data. GWNBR can provide an adequate modeling routine for transit ridership at the bus stop level by accounting for some characteristics that have been overlooked in previous studies: asymmetry, overdispersion, spatial dependence, spatial heterogeneity, and network distances. The case study covered two bus lines from the city of São Paulo – SP (Brazil).

Results showed that the GWNBR model was able to address the spatial dependence found in the global Negative Binomial Regression. In addition, by confirming the hypothesis of overdispersion, NB distribution is considered a better modeling option for ridership data.

Four questions were addressed. The following topics summarize the questions and respective conclusions.

a) *What are the factors affecting stop-level transit ridership?*
   Five variables were found to better explain transit ridership in this Brazilian context: medium/high standard horizontal residential area, area with no predominant land use, intersections, overlapping area ratio and frequency.
b) *Is the spatial variation of predictors' effects statistically significant?*
   The spatially varying relationships from overlapping area ratio and frequency proved to be statistically significant, in both distance approaches. The stationarity test can be used for municipalities to decide whether the spatial variation of predictors' effects on transit ridership justifies the application of urban policies at the local level. In our case study, solutions regarding the land use and intersections variables could be developed based on the global model results, if simpler approaches are to be prioritized.
c) *Is GWNBR capable of providing good predictions of stop-level ridership in missing data scenarios?*
   GWNBR is an effective tool to the transit ridership estimation in uncounted bus stops, even when the availability of data is low. This may facilitate the transit ridership modeling in cities whose bus lines do not have Automatic Passenger Counters (APCs) and the ones with APCs installed only in a limited number of lines. As expected, missing data negatively affects the prediction accuracy and parameter estimates. However, based on our case study, we cannot confirm the existence of a clear relationship between percentage of missing data and the results consistency.
d) *Can network distances improve the prediction accuracy of GWNBR compared to the traditional Euclidean distances?*
   Positive effects of network distances on the prediction accuracy were seen mostly in the calibration results. Increasing the percentage of missing data led to higher differences between prediction results from network and Euclidean distances in the calibration cases. However, this pattern could not be confirmed for the validation samples. Network distances also had an impact on the coefficients estimates. A better approach would be to adopt parameter-specific distance metrics.

Although the consistency of parameters' estimates was not as good as the prediction power consistency, GWNBR in missing data scenarios could, in fact, reproduce the spatial pattern of effects shown in the complete database model, for some explanatory variables. The similarity between coefficients might be conditioned by the spatial arrangement of bus stops, which must be able to replicate the same spatial dependence pattern of the whole system under analysis. Thus, the sampling strategy may play a fundamental role in the sensitivity of parameter estimates and should be further investigated in future studies.

The increasing availability of the GWR routines in free software provides analysts with enough material to apply the local modeling in a municipal project. The GWmodel library in R (Gollini et al., 2015; Lu et al., 2014b; R Core Team, 2021) has computational routines for various possibilities of geographically weighted analyses. The free software GWR4.0 is an example of interface that needs no programming knowledge for running the GWR modeling. Therefore, consultants have free and easy access to carry out analyses based on GWR. We hope that forthcoming studies, provided with robust databases, can help to consolidate the conclusions achieved. In addition, as modeling routines are increasingly evolving over time, future alternatives to transit ridership modeling would be a mixed GWNBR model, allowing some coefficients to vary spatially and others to be fixed; a GWNBR model with parameter-specific distance metrics; and geographically and temporally weighted regressions, for the cases where panel data is available, and it exhibits temporal nonstationarity.

Availability of data and material

The transit ridership dataset analyzed during the current study is not publicly available due to the fact that the data is held by *SPTrans* but can be requested from it through the Electronic Citizen Information System (http://esic.prefeitura.sp.gov.br/Account/Login.aspx). The datasets used to collect the potential predictors can be found on the GeoSampa website (http://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/).

Funding

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

CRediT authorship contribution statement

Samuel de França Marques: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Cira Souza Pitombo:** Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data in the "Availability of data and material" section

Acknowledgements

The authors would like to thank *SPTrans* for the Boarding/Alighting survey data used in this study, and André Borgato Morelli, who provided the employment shapefile for the modeling step.

References

Blainey, S., Mulley, C., 2013. Using geographically weighted regression to forecast rail demand in the Sydney region. In: Australasian Transport Research Forum 2013 Proceedings. Brisbane, Australia.

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

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr. Anal. 28, 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x.

Cameron, A.C., Trivedi, P.K., 1990. Regression-based tests for overdispersion in the Poisson model. J. Econ. 46, 347–364. https://doi.org/10.1016/0304-4076(90)90014-K.

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. Appl. Geogr. 34, 548–558. https://doi.org/10.1016/j.apgeog.2012.01.005.

Ceder, A., 2007. Public Transit Planning and Operation: Modeling, Practice and Behavior. CRC press.

Cervero, R., 2006. Alternative approaches to modeling the travel-demand impacts of smart growth. J. Am. Plan. Assoc. 72, 285–295. https://doi.org/10.1080/01944360608976751.

Chakour, V., Eluru, N., 2016. Examining the influence of stop level infrastructure and built environment on bus ridership in Montreal. J. Transp. Geogr. 51, 205–217. https://doi.org/10.1016/j.jtrangeo.2016.01.007.

Chiou, Y.C., Jou, R.C., Yang, C.H., 2015. Factors affecting public transportation usage rate: geographically weighted regression. Transp. Res. Part A Policy Pract. 78, 161–177. https://doi.org/10.1016/j.tra.2015.05.016.

Chu, X., 2004. Ridership Models at the Stop Level. National Center for Transit Research, University of South Florida.

Cui, B., DeWeese, J., Wu, H., King, D.A., Levinson, D., El-Geneidy, A., 2022. All ridership is local: accessibility, competition, and stop-level determinants of daily bus boardings in Portland. Oregon. J. Transp. Geogr. 99, 103294 https://doi.org/10.1016/j.jtrangeo.2022.103294.

da Silva, A.R., Rodrigues, T.C.V., 2014. Geographically weighted negative binomial regression—incorporating overdispersion. Stat. Comput. 24, 769–783. https://doi.org/10.1007/s11222-013-9401-9.

da Silva, A.R., Rodrigues, T.C.V., 2016. A SAS® Macro for Geographically Weighted Negative Binomial Regression, in: Proceedings of the SAS® Global Forum 2016 Conference. SAS Institute Inc., Las Vegas, USA.

Marques, S.F., Pitombo, C.S., 2021a. Applying multivariate Geostatistics for transit ridership modeling at the bus stop level. Bol. Ciências Geodésicas 27. https://doi.org/10.1590/1982-2170-2020-0069.

Marques, S., Pitombo, C.S., 2021b. Spatial modeling of transit ridership along bus lines with overlapping sections. In: Anais Do 35º Congresso de Pesquisa e Ensino Em Transportes. Associação Nacional de Pesquisa e Ensino em Transportes, 100% Virtual, pp. 1568–1580.

Marques, S.F., Pitombo, C.S., 2021c. Ridership estimation along bus transit lines based on kriging: comparative analysis between network and Euclidean distances. J. Geovisualizat. Spat. Anal. 5, 7. https://doi.org/10.1007/s41651-021-00075-w.

Marques, S.F., Pitombo, C.S., 2022. Transit Ridership Modeling at the Bus Stop Level: Comparison of Approaches Focusing on Count and Spatially Dependent Data. Appl. Spat. Anal. Policy. https://doi.org/10.1007/s12061-022-09482-y.

Dill, J., Schlossberg, M., Ma, L., Meyer, C., 2013. Predicting transit ridership at the stop level: The role of service and urban form. In: Transportation Research Board 92nd Annual Meeting. Washington DC, United States, pp. 1–19.

Eom, J.K., Park, M.S., Heo, T.-Y., Huntsinger, L.F., 2006. Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. Transp. Res. Rec. 22–29 https://doi.org/10.3141/1968-03.

Erhardt, G.D., Lock, O., Arcaute, E., Batty, M., 2017. In: Thakuriah, P. Vonu, Tilahun, N., Zellner, M. (Eds.), A Big Data Mashing Tool for Measuring Transit System Performance BT - Seeing Cities through Big Data: Research, Methods and Applications in Urban Informatics. Springer International Publishing, Cham, pp. 257–278. https://doi.org/10.1007/978-3-319-40902-3_15.

Evans, J.S., 2021. _spatialEco_. R package version 1.3-6. URL. https://github.com/jeffreyevans/spatialEco.

Farber, S., Páez, A., 2007. A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. J. Geogr. Syst. 9, 371–396. https://doi.org/10.1007/s10109-007-0051-3.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2003. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons.

Fotheringham, A.S., Yu, H., Wolf, L.J., Oshan, T.M., Li, Z., 2022. On the notion of ‘bandwidth’ in geographically weighted regression models of spatially varying processes. Int. J. Geogr. Inf. Sci. 36, 1485–1502. https://doi.org/10.1080/13658816.2022.2034829.

Frei, C., Mahmassani, H.S., 2013. Riding more frequently: estimating disaggregate ridership elasticity for a large urban bus transit network. Transp. Res. Rec. 2350, 65–71. https://doi.org/10.3141/2350-08.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P., 2015. GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Softw. 63, 1–50. https://doi.org/10.18637/jss.v063.i17.

Gomes, M.J.T.L., Cunto, F., da Silva, A.R., 2017. Geographically weighted negative binomial regression applied to zonal level safety performance models. Accid. Anal. Prev. 106, 254–261. https://doi.org/10.1016/j.aap.2017.06.011.

Hilbe, J.M., 2014. Modeling Count Data. Cambridge Univeersity Press, Cambridge. https://doi.org/10.1017/CBO9781139236065.

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

IBGE, 2021. São Paulo (IBGE cities).

Johnson, A., 2003. Bus transit and land use: illuminating the interaction. J. Public Transp. 6, 21–39. https://doi.org/10.5038/2375-0901.6.4.2.

Kerkman, K., Martens, K., Meurs, H., 2015. Factors influencing stop-level transit ridership in Arnhem–Nijmegen City Region, Netherlands. Transp. Res. Rec. 2537, 23–32. https://doi.org/10.3141/2537-03.

Kim, S., Park, S., Jang, K., 2019. Spatially-varying effects of built environment determinants on walking. Transp. Res. Part A Policy Pract. 123, 188–199. https://doi.org/10.1016/j.tra.2019.02.003.

Kleiber, C., Zeileis, A., 2008. Applied Econometrics with R. Springer-Verlag, New York.

Lanza, K., Durand, C.P., 2021. Heat-moderating effects of bus stop shelters and tree shade on public transport ridership. Int. J. Environ. Res. Public Health. https://doi.org/10.3390/ijerph18020463.

Leung, Y., Mei, C.-L., Zhang, W.-X., 2000. Statistical tests for spatial nonstationarity based on the geographically weighted regression model. Environ. Plan. A Econ. Sp. 32, 9–32. https://doi.org/10.1068/a3162.

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, 4684.

Lu, B., Charlton, M., Harris, P., Fotheringham, A.S., 2014a. Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. Int. J. Geogr. Inf. Sci. 28, 660–681. https://doi.org/10.1080/13658816.2013.865739.

Lu, B., Harris, P., Charlton, M., Brunsdon, C., 2014b. The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. Geo-spatial Inf. Sci. 17, 85–101. https://doi.org/10.1080/10095020.2014.917453.

Lu, B., Brunsdon, C., Charlton, M., Harris, P., 2017. Geographically weighted regression with parameter-specific distance metrics. Int. J. Geogr. Inf. Sci. 31, 982–998. https://doi.org/10.1080/13658816.2016.1263731.

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. Comput. Environ. Urban. Syst. 70, 113–124. https://doi.org/10.1016/j.compenvurbsys.2018.03.001.

Mathew, S., Pulugurtha, S.S., 2021. Comparative assessment of geospatial and statistical methods to estimate local road annual average daily traffic. J. Transp. Eng. Part A Syst. 147, 04021035. https://doi.org/10.1061/jtepbs.0000542.

Metrô, 2019. Origin and Destination Survey. Obtained from. https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino.

Miao, Q., Welch, E.W., Sriraj, P.S., 2019. Extreme weather, public transport ridership and moderating effect of bus stop shelters. J. Transp. Geogr. 74, 125–133. https://doi.org/10.1016/j.jtrangeo.2018.11.007.

Moran, P.A.P., 1948. The interpretation of statistical maps. J. R. Stat. Soc. Ser. B 10, 243–251.

Mucci, R.A., Erhardt, G.D., 2018. Evaluating the ability of transit direct ridership models to forecast medium-term ridership changes: evidence from San Francisco. Transp. Res. Rec. 2672, 21–30. https://doi.org/10.1177/0361198118758632.

Ngo, N.S., 2019. Urban bus ridership, income, and extreme weather events. Transp. Res. Part D Transp. Environ. 77, 464–475. https://doi.org/10.1016/j.trd.2019.03.009.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of Phylogenetics and evolution in R language. Bioinformatics 20, 289–290. https://doi.org/10.1093/bioinformatics/btg412.

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 (Amst). 24, 159–181. https://doi.org/10.1023/A:1017951902308.

Profillidis, V.A., Botzoris, G.N., 2019. Statistical methods for transport demand modeling. In: Romer, B. (Ed.), Modeling of Transport Demand. Elsevier, pp. 163–224. 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. J. Public Transp. 15, 33–52.

Pulugurtha, S.S., Mathew, S., 2021. Modeling AADT on local functionally classified roads using land use, road density, and nearest nonlocal road data. J. Transp. Geogr. 93, 103071 https://doi.org/10.1016/j.jtrangeo.2021.103071.

R Core Team, 2021. R: A Language and Environment for Statistical Computing.

Rahman, M., Yasmin, S., Eluru, N., 2019. Evaluating the impact of a newly added commuter rail system on bus ridership: a grouped ordered logit model approach. Transp. A Transp. Sci. 15, 1081–1101. https://doi.org/10.1080/23249935.2018.1564800.

Rahman, M., Yasmin, S., Faghih-Imani, A., Eluru, N., 2021. Examining the bus ridership demand: application of Spatio-temporal panel models. J. Adv. Transp. 2021, 8844743. https://doi.org/10.1155/2021/8844743.

Ryan, S., Frank, L., 2009. Pedestrian environments and transit ridership. J. Public Transp. 12, 39–57. https://doi.org/10.5038/2375-0901.12.1.3.

São Paulo, 2018. Anexo VI - Investimentos e Responsabilidades. São Paulo, Brazil. Obtained from. https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/transportes/edital2018/001_ESTRUTURAL/ANEXO-VI_INVESTIMENTOS/ANEXO-VI_INVESTIMENTOS_RESPONSABILIDADES.pdf.

São Paulo, 2022. Centro de Operações da SPTrans ganha projeto de modernização. São Paulo, Brazil. Obtained from. https://www.capital.sp.gov.br/noticia/centro-de-operacoes-da-sptrans-ganha-projeto-de-modernizacao.

Sarlas, G., Axhausen, K.W., 2015. Prediction of AADT on a nationwide network based on an accessibility-weighted centrality measure. In: Arbeitsberichte Verkehrs- und Raumplan, p. 1094. https://doi.org/10.3929/ethz-b-000102909.

Selby, B., Kockelman, K.M., 2013. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. J. Transp. Geogr. 29, 24–32. https://doi.org/10.1016/j.jtrangeo.2012.12.009.

Shi, X., Moudon, A.V., Hurvitz, P.M., Mooney, S.J., Zhou, C., Saelens, B.E., 2021. Does improving stop amenities help increase bus rapid transit ridership? Findings based on a quasi-experiment. Transp. Res. Interdiscip. Perspect. 10, 100323 https://doi.org/10.1016/j.trip.2021.100323.

Song, Y., Merlin, L., Rodriguez, D., 2013. Comparing measures of urban land use mix. Comput. Environ. Urban. Syst. 42, 1–13. https://doi.org/10.1016/j.compenvurbsys.2013.08.001.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234–240. https://doi.org/10.2307/143141.

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. J. Transp. Geogr. 69, 45–57. https://doi.org/10.1016/j.jtrangeo.2018.04.013.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4th ed. Springer, New York. https://doi.org/10.1007/978-0-387-21706-2.

Ver Hoef, J.M., 2018. Kriging models for linear networks and non-Euclidean distances: cautions and solutions. Methods Ecol. Evol. 0 https://doi.org/10.1111/2041-210X.12979.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1, 80–83. https://doi.org/10.2307/3001968.

Wong, A.H., Kwon, T.J., 2021. Advances in regression kriging-based methods for estimating statewide winter weather collisions: an empirical investigation. Futur. Transp. 1, 570–589. https://doi.org/10.3390/futuretransp1030030.

Yan, X., Su, X.G., 2009. Linear Regression Analysis: Theory and Computing. World Scientific.

Zhang, D., Wang, X.C., 2014. Transit ridership estimation with network kriging: a case study of second avenue Subway. NYC. J. Transp. Geogr. 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. Transp. Res. Rec. 1835, 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 (Amst). 46, 2269–2289. https://doi.org/10.1007/s11116-018-9928-x.

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

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.

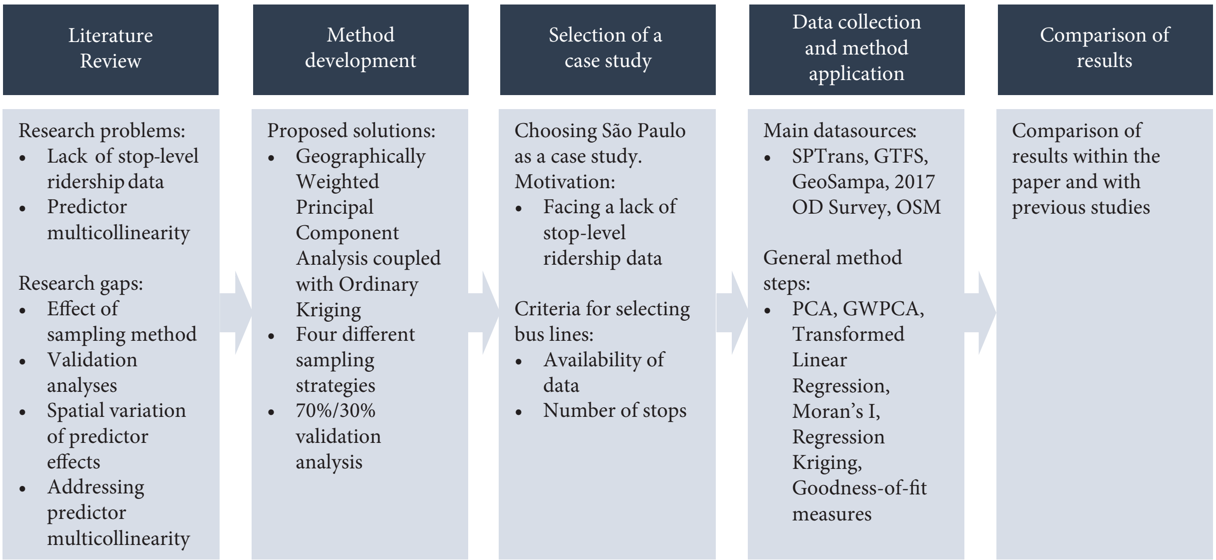

Literature Review
Method development
Selection of a case study
Data collection and method application
Comparison of results
Research problems:
• Lack of stop-level ridership data
• Predictor multicollinearity
Research gaps:
• Effect of sampling method
• Validation analyses
• Spatial variation of predictor effects
• Addressing predictor multicollinearity
Proposed solutions:
• Geographically Weighted Principal Component Analysis coupled with Ordinary Kriging
• Four different sampling strategies
• 70%/30% validation analysis
Choosing São Paulo as a case study.
Motivation:
• Facing a lack of stop-level ridership data
Criteria for selecting bus lines:
• Availability of data
• Number of stops
Main datasources:
• SPTrans, GTFS, GeoSampa, 2017 OD Survey, OSM
General method steps:
• PCA, GWPCA, Transformed Linear Regression, Moran's I, Regression Kriging, Goodness-of-fit measures
Comparison of results within the paper and with previous studies

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

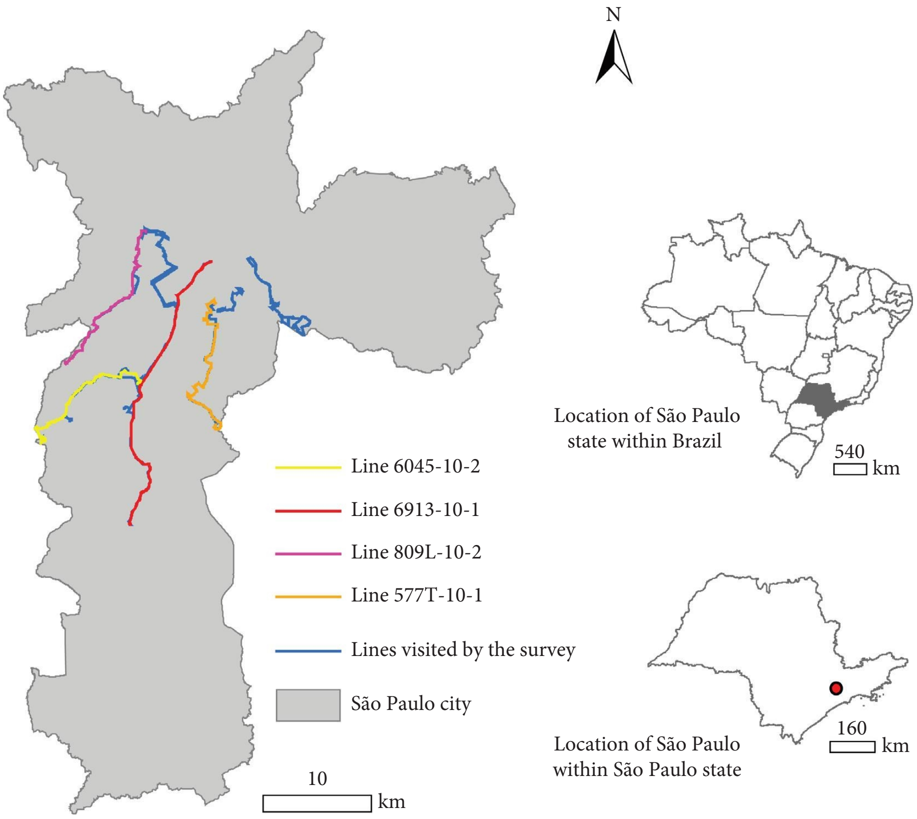


Figure 2: Case study databases.

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]:

$$\sum (u, v) = X' W (u, v) X, \qquad (1)$$

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].

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

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.

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

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:

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

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

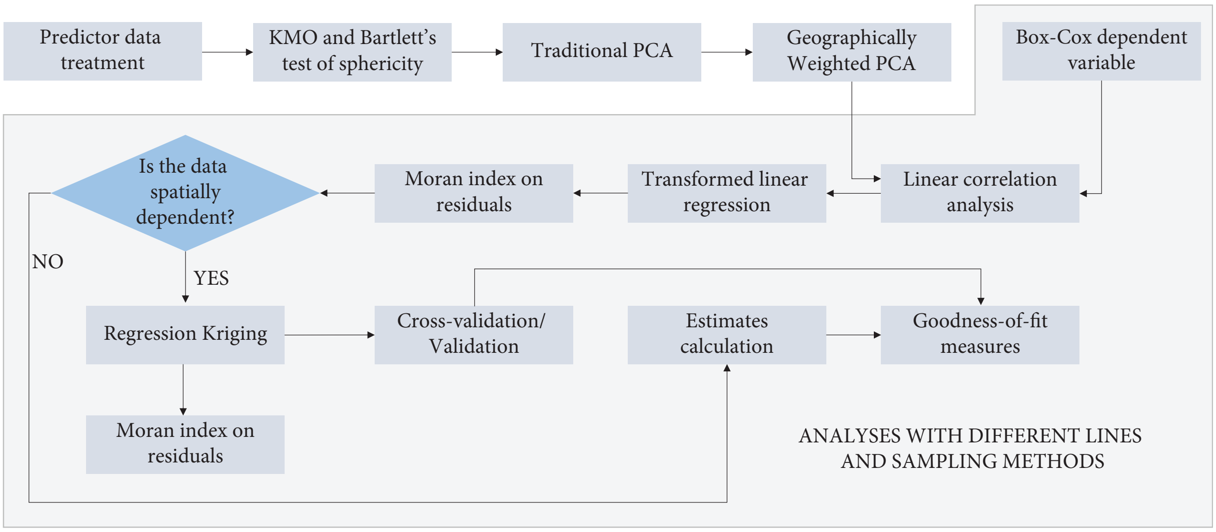

Predictor data treatment
KMO and Bartlett's test of sphericity
Traditional PCA
Geographically Weighted PCA
Box-Cox dependent variable
Is the data spatially dependent?
Moran index on residuals
Transformed linear regression
Linear correlation analysis
NO
YES
Regression Kriging
Cross-validation/ Validation
Estimates calculation
Goodness-of-fit measures
Moran index on residuals
ANALYSES WITH DIFFERENT LINES AND SAMPLING METHODS

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
Bartlett test of sphericity	Approx. χ^2	257281.195
	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

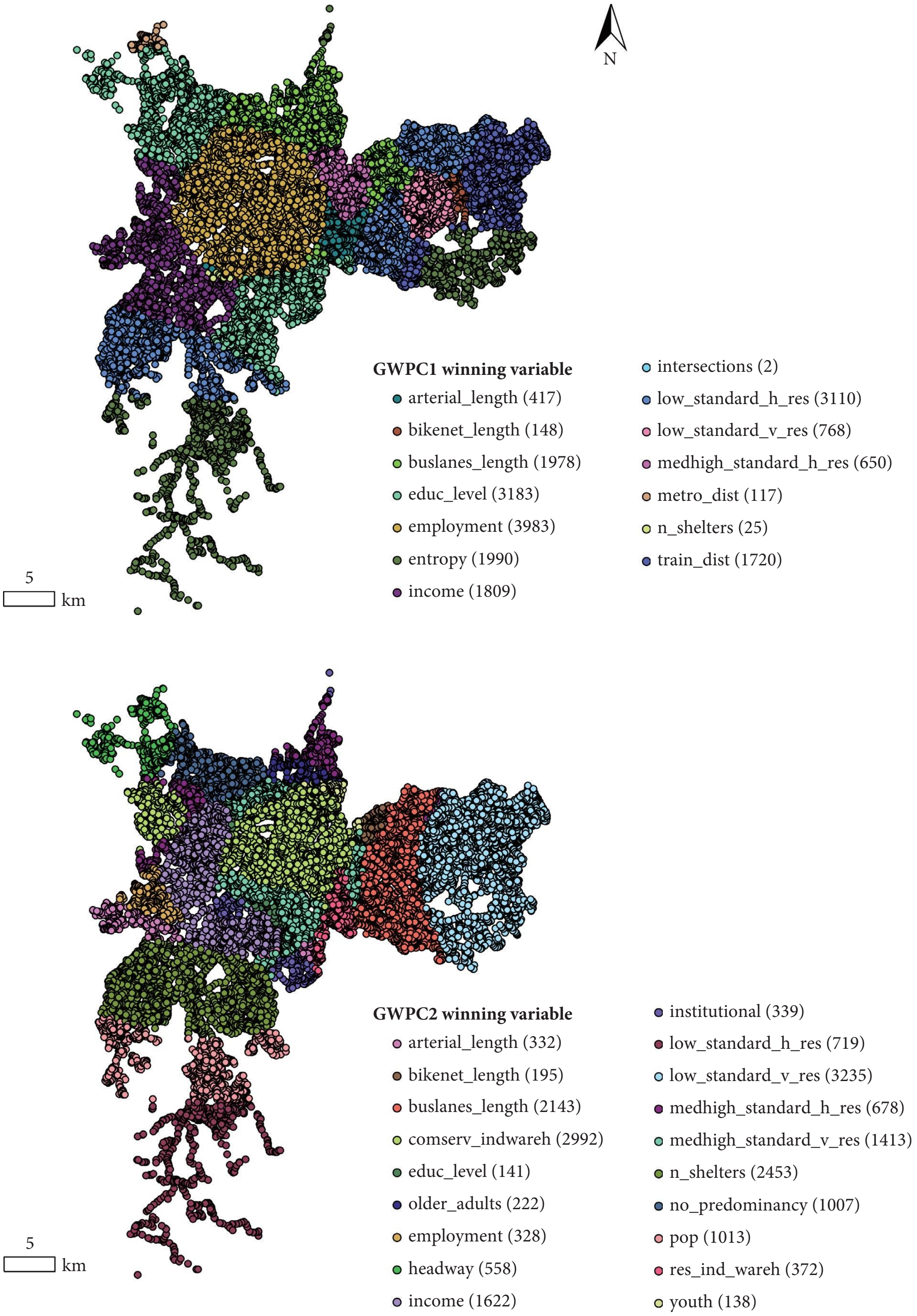


Figure 4: GWPC1 and GWPC2 winning variables ($N = 19{,}900$).

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

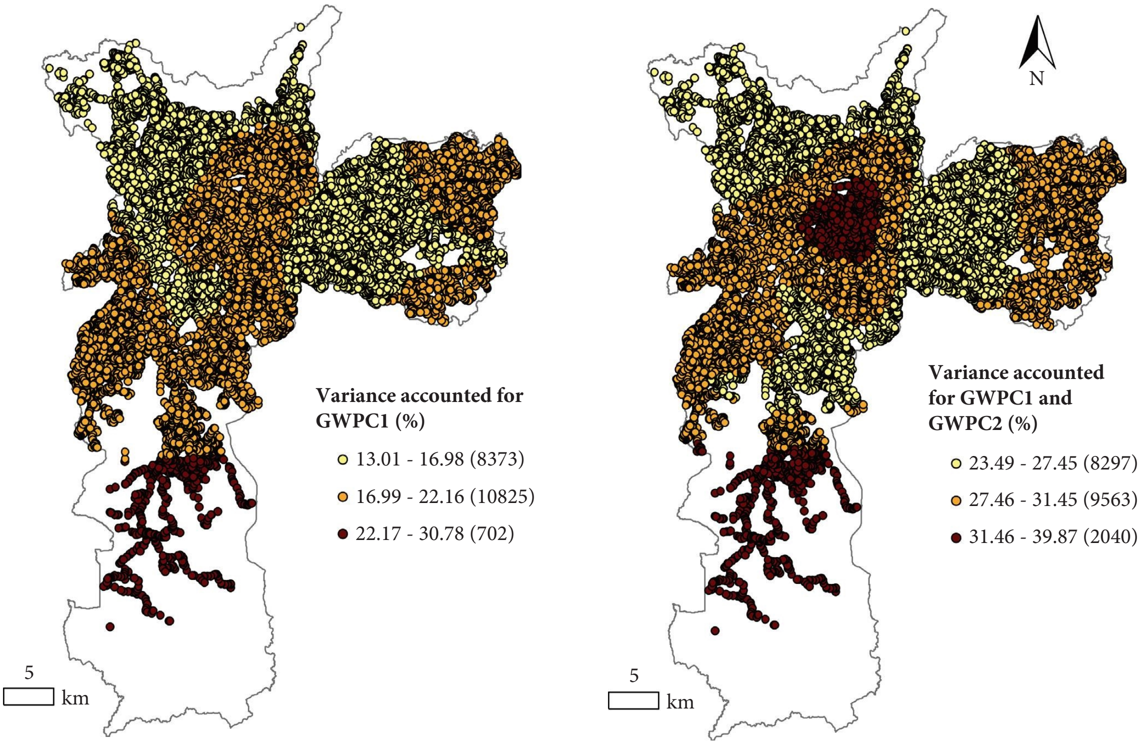


FIGURE 5: Local percent of variance ($N = 19{,}900$).

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

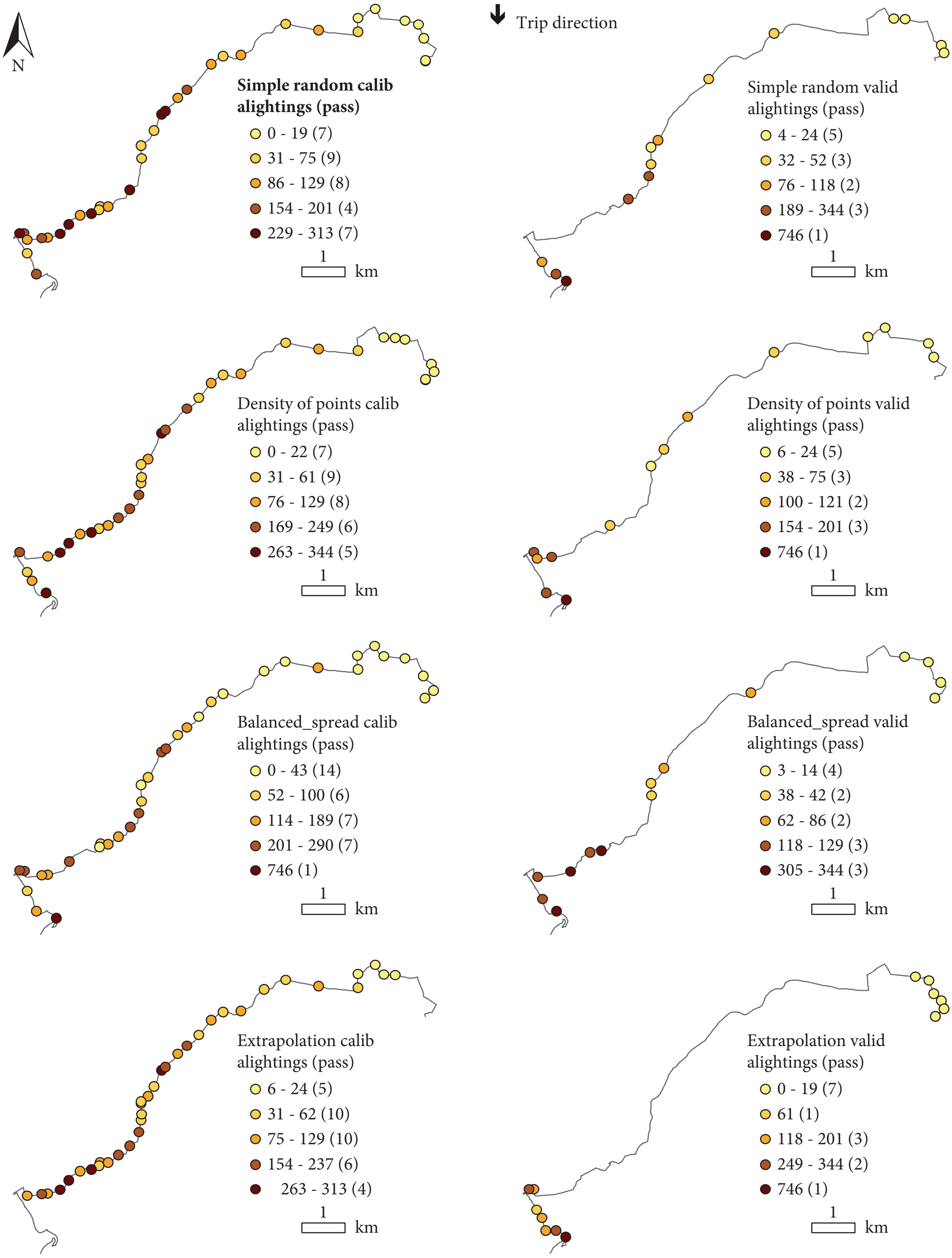


Figure 6: Alightings along calibration and validation samples of line 6045-10-2.

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

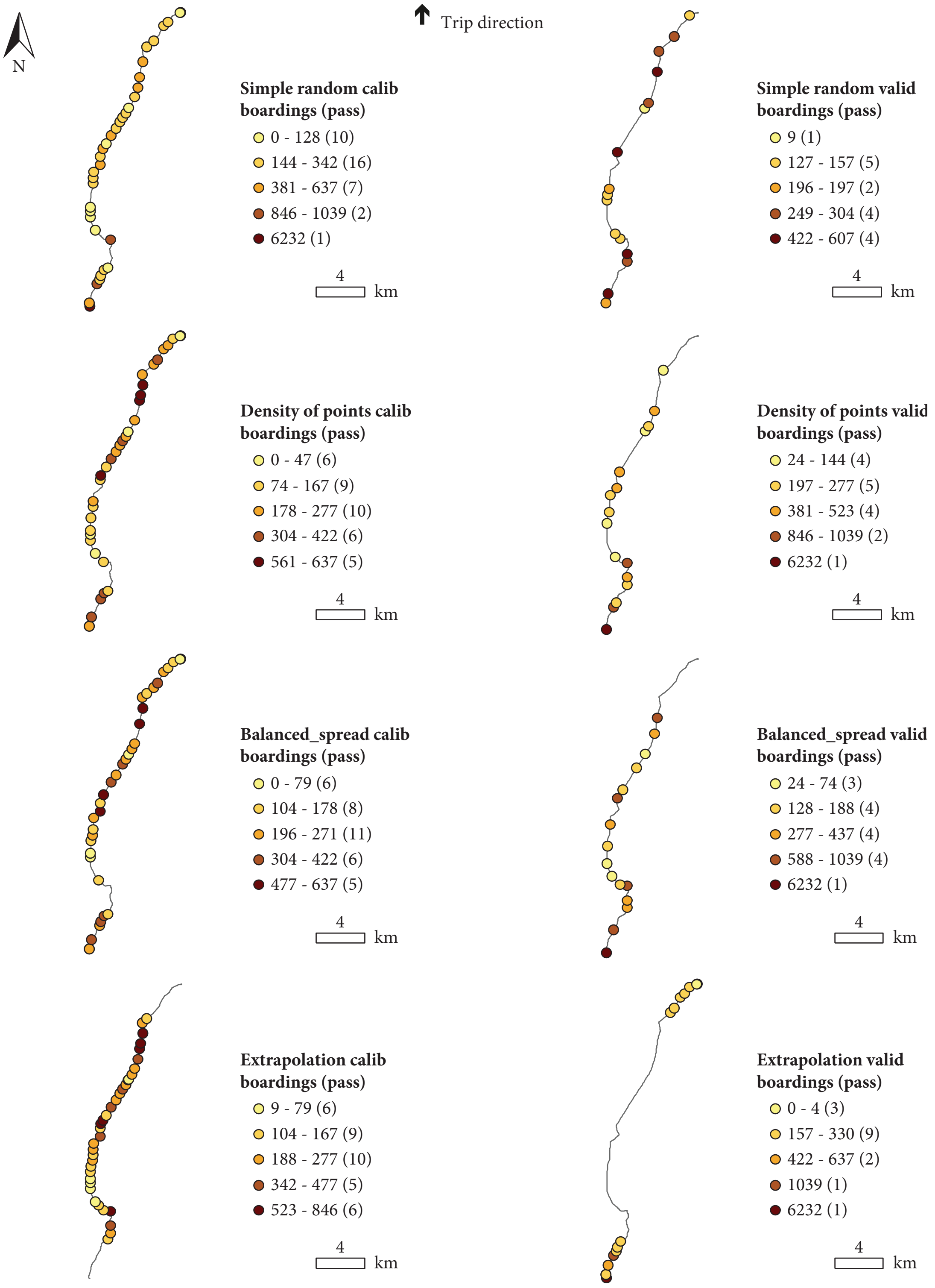


Figure 7: Boardings along calibration and validation samples of line 6913-10-1.

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

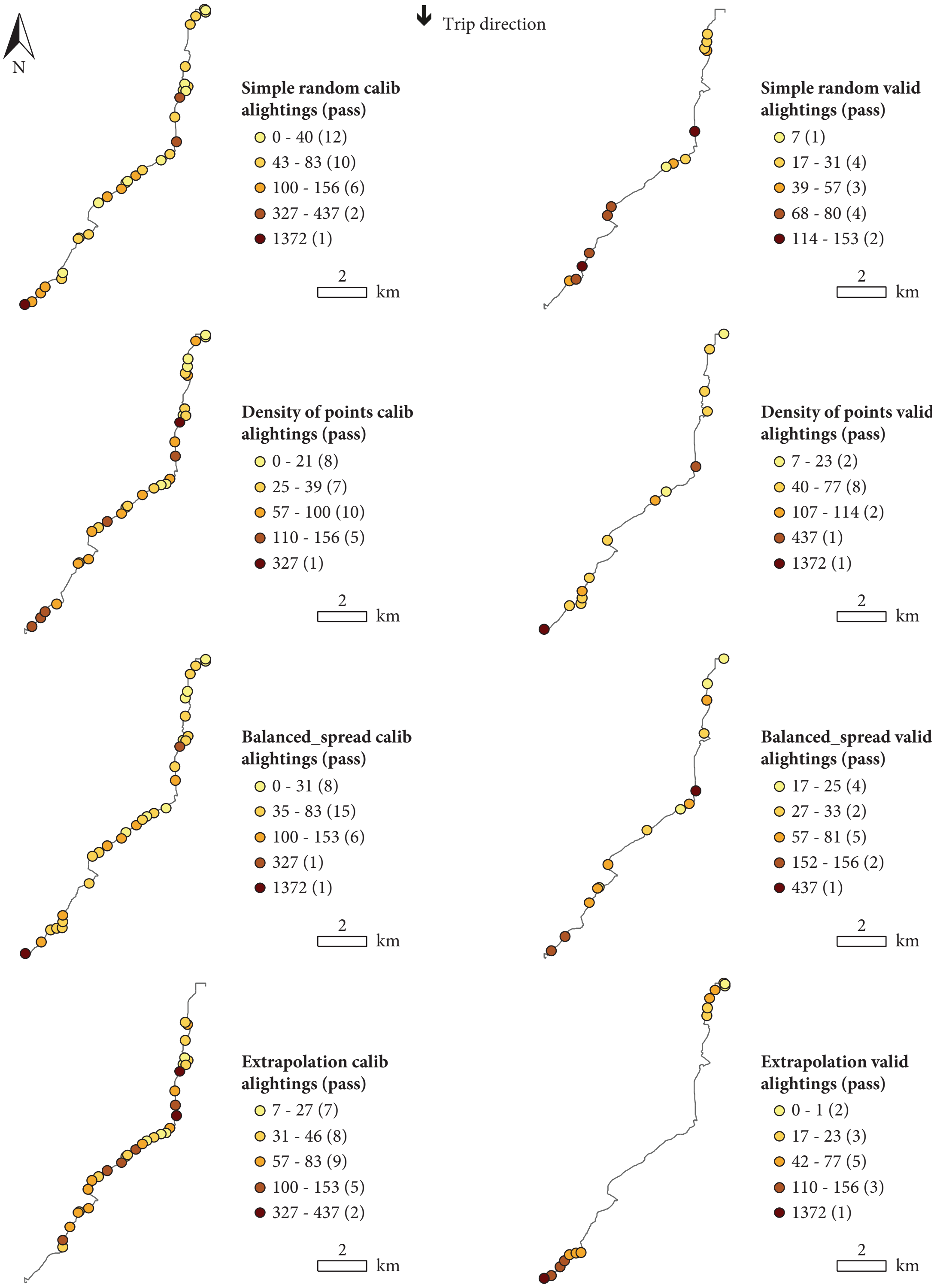


Figure 8: Alightings along calibration and validation samples of line 809L-10-2.

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

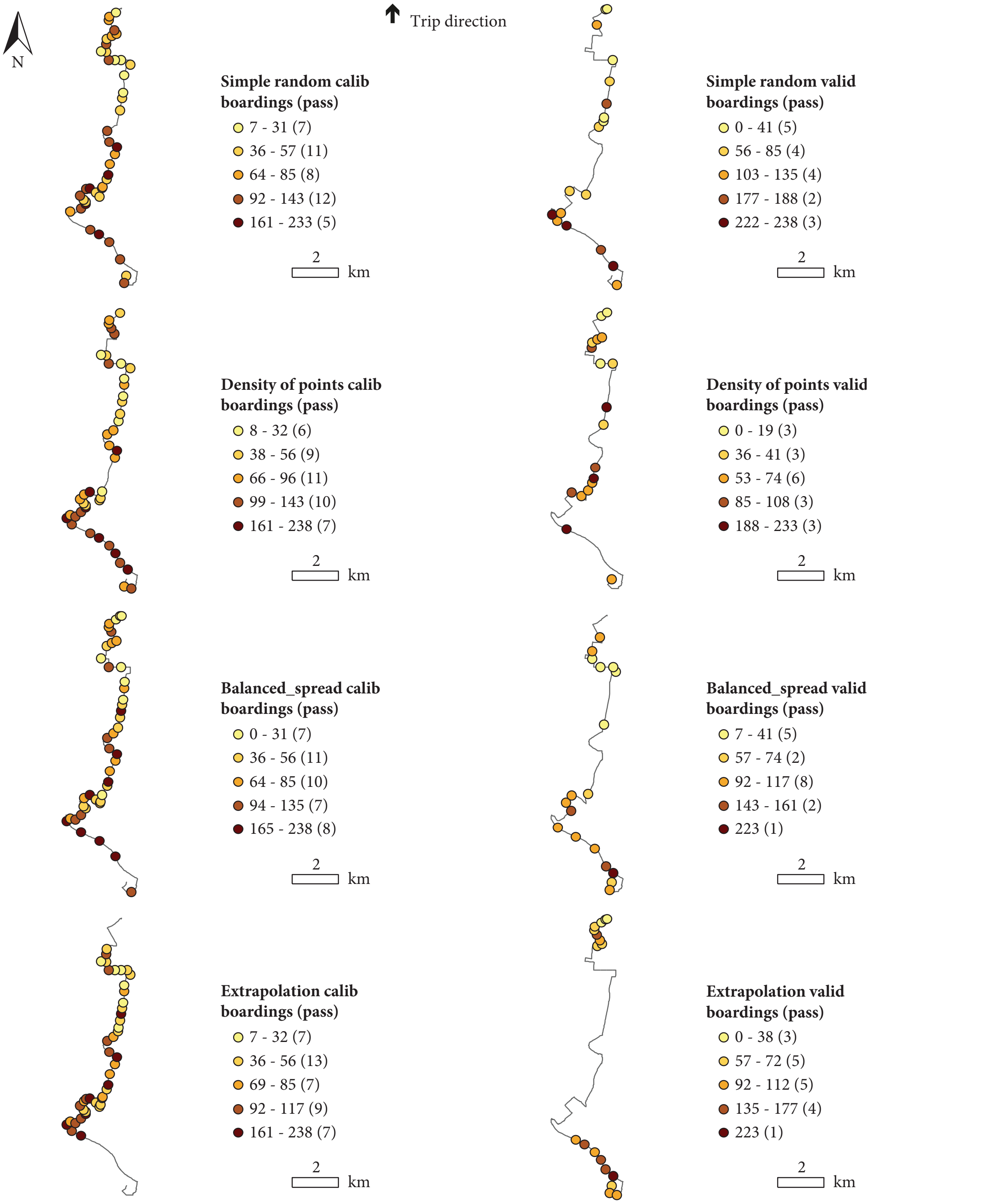


Figure 9: Boardings along calibration and validation samples of line 577T-10-1.

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.

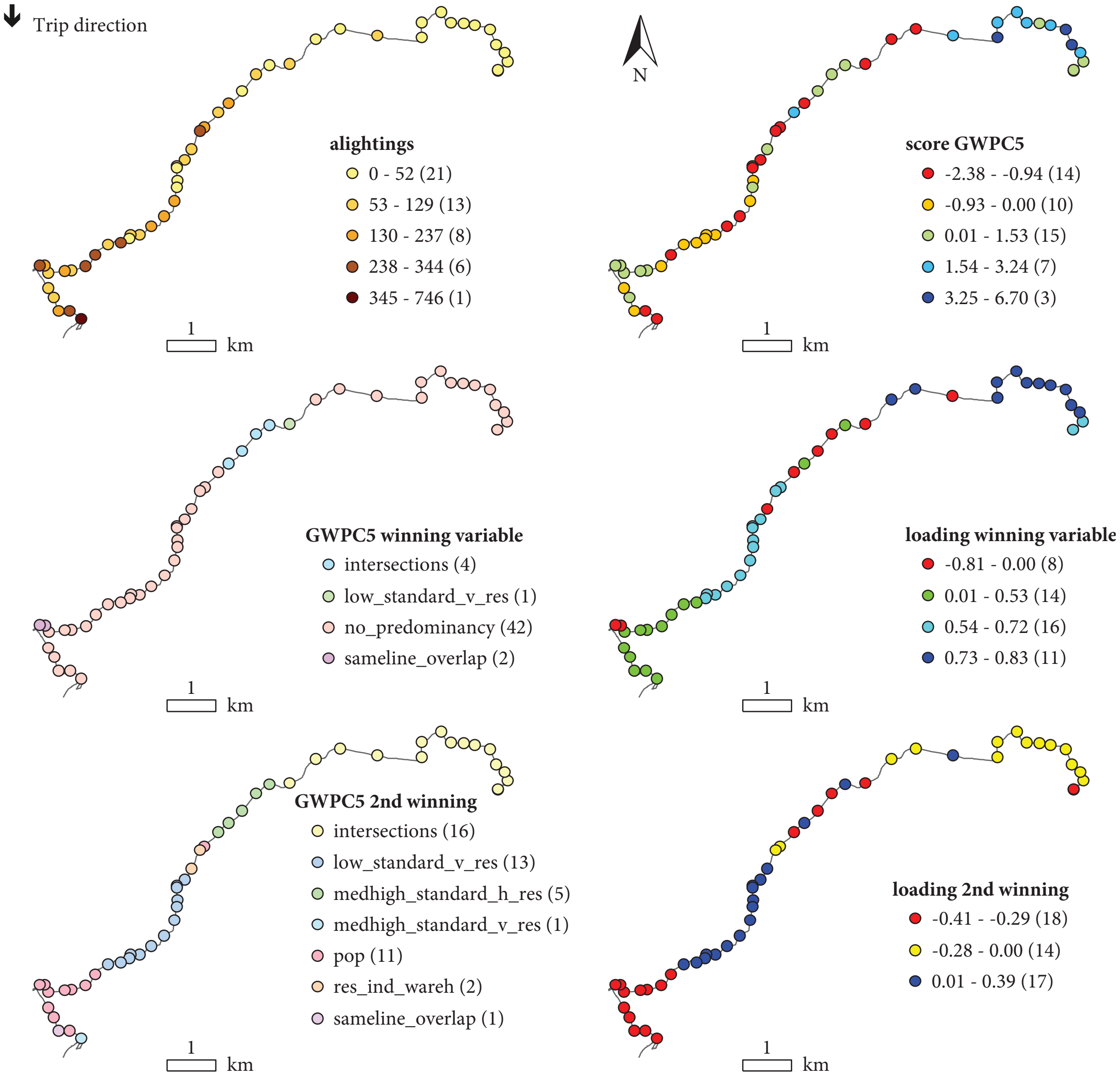


Figure 10: Spatial pattern of alightings and GWPC5 along line 6045-10-2.

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

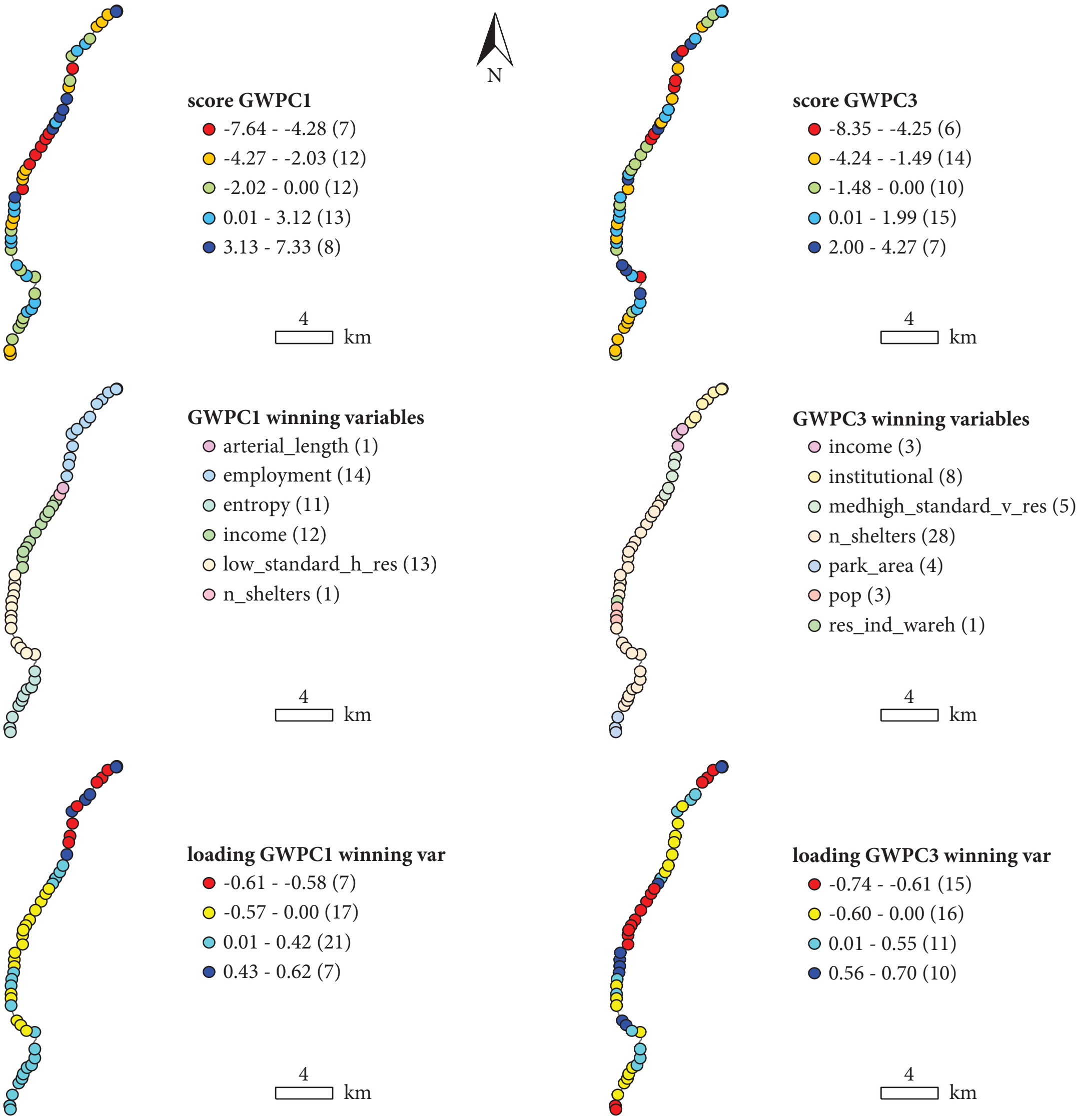


Figure 11: Spatial pattern of GWPC1 and GWPC3 along line 6913-10-1.

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.

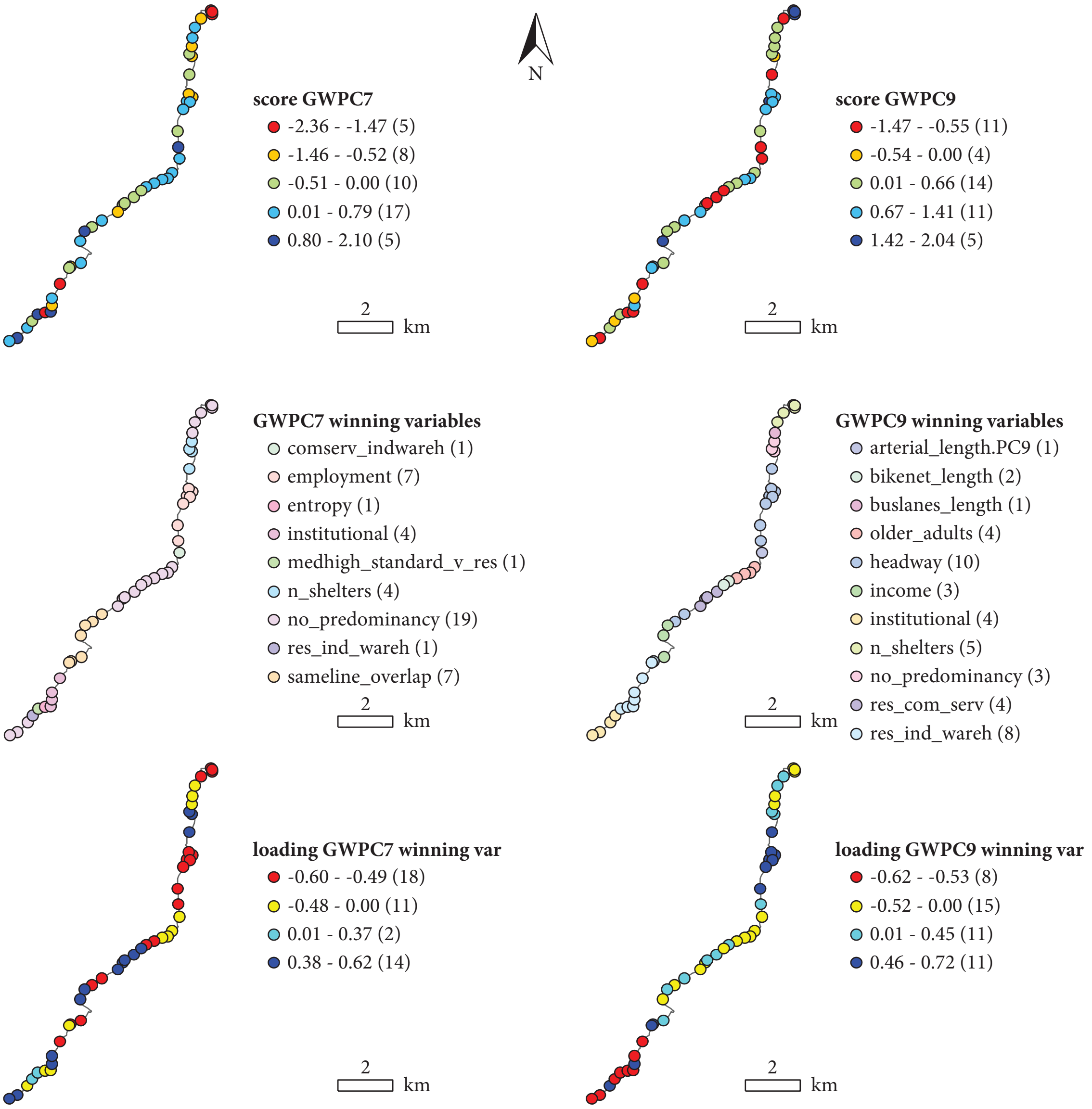


FIGURE 12: Spatial pattern of GWPC7 and GWPC9 along line 809L-10-2.

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

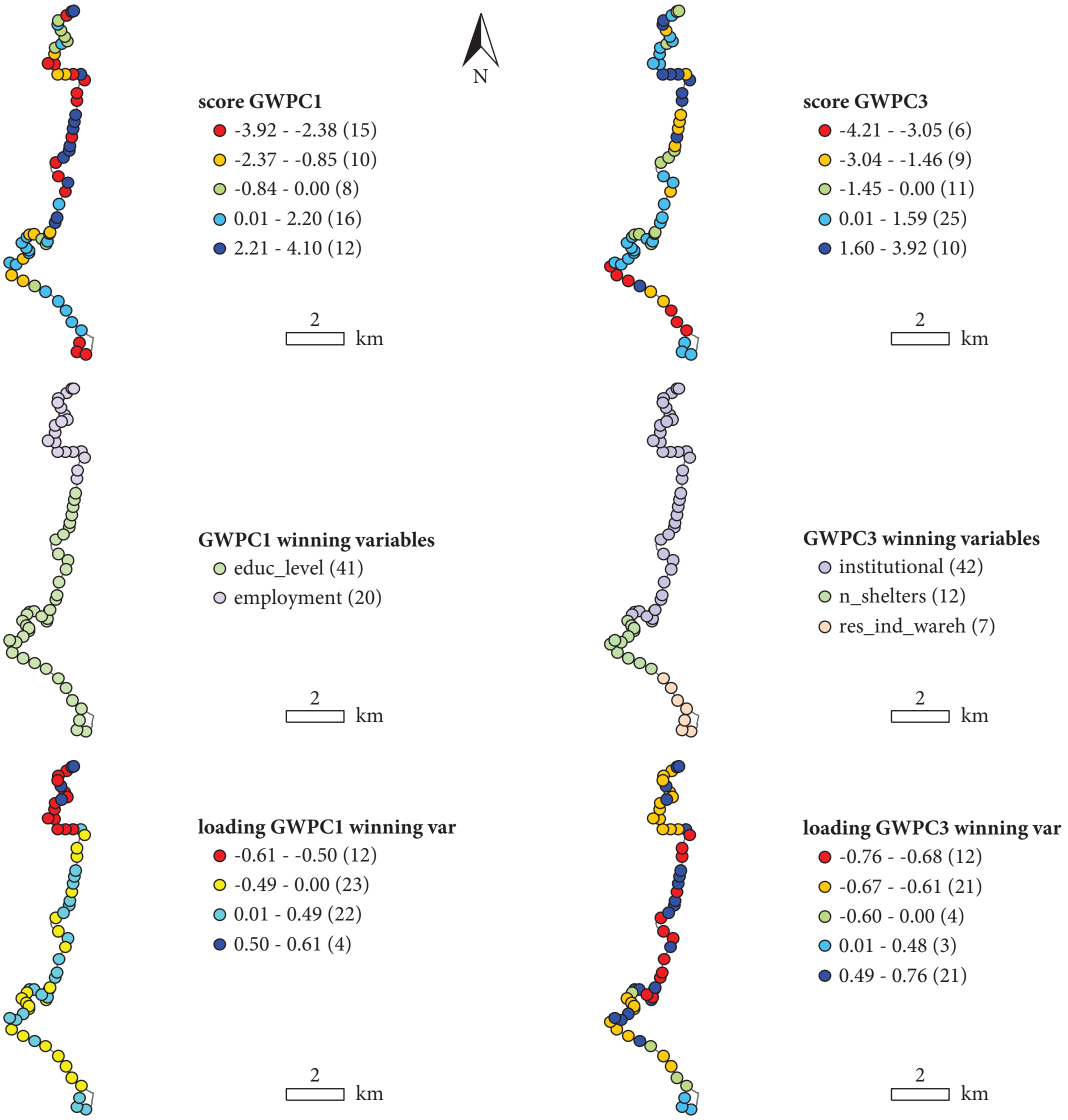


FIGURE 13: Spatial pattern of GWPC1 and GWPC3 along line 577T-10-1.

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," *Geo-spatial 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.

6 COMPLEMENTARY RESEARCH

“*Curiouser and curiouser!*”
Alice in Wonderland by Lewis Carroll

This section presents additional research, developed during the doctoral period, which had an impact on the thesis conclusions. Four complementary papers were published, addressing some of the thesis objectives and other variables of interest. Following the ordering of the main articles, the additional papers are listed as follows and briefly explained in the next paragraphs.

V. **Marques, S. de F.**, and Pitombo, C. S. (2020) Intersecting Geostatistics with transport demand modeling: a bibliographic survey. *Revista Brasileira de Cartografia*, *72*, 1028-1050. doi: 10.14393/rbcv72nespecial50anos-56467

VI. **Marques, S. de F.**, and Pitombo, C. S. (2021b) Spatial modeling of transit ridership along bus lines with overlapping sections. *Proceedings of the 35º Annual Conference of the Brazilian National Association for Transportation Research and Teaching* (p. 1568–1580). Brazilian National Association for Transportation Research and Teaching, 100% Virtual. Available at: https://www.researchgate.net/publication/357517939_SPATIAL_MODELING_OF_TRANSIT_RIDERSHIP_ALONG_BUS_LINES_WITH_OVERLAPPING_SECTIONS. Accessed April 2023.

VII. **Marques, S. de F.**, and Pitombo, C. S. (2022) Spatially varying relationships of transit ridership at the bus stop level. *Proceedings of the 36º Annual Conference of the Brazilian National Association for Transportation Research and Teaching* (p. 152517). Brazilian National Association for Transportation Research and Teaching, Fortaleza, Brazil. Available at: https://proceedings.science/anpet-2022/trabalhos/spatially-varying-relationships-of-transit-ridership-at-the-bus-stop-level. Accessed April 2023.

VIII. **Marques, S. de F.**, Favero, R., and Pitombo, C. S. (2023) Should we account for network distances or anisotropy in the spatial estimation of missing traffic data? *Transportes*, *31*(1), e2822. doi:10.58922/transportes.v31i1.2822

Article V presents the state-of-art of Geostatistics applications to the travel demand modeling along four geographic units of analysis: Traffic Analysis Zones; road segments; stations, bus stops and bus lines segments; and households/individuals. This bibliographic

review identified research gaps in applying Ordinary Kriging and Universal Kriging in the context of bus stops. In addition, using network distances was considered as an underexplored topic in the bus stop and road segment levels.

Article VI analyzes two main issues in the stop-level ridership modeling: overlapping of bus stops' catchment areas and network distances. In this case, dummy variables characterized the overlapping problem in a regression modeling that accounted for the spatial dependence of residuals though Ordinary Kriging, comparing Euclidean and network distances. The variable of interest was the number of boardings from 20h to 23h59 over lines 856R-10-2 and 6913-10-1, the same lines from Article III. Results showed that addressing the overlapping issue significantly contributed to improve the models' goodness-of-fit measures, but few differences were seen between the network and Euclidean approaches.

Article VII is a preliminary version of Article III proposing the application of Geographically Weighted Negative Binomial Regression only in a complete database case. Missing data analyses and comparisons between network and Euclidean distances were not carried out in this paper. The variable of interest was the number of boardings plus alightings from 20h to 23h59 along lines 856R-10-1 and 6913-10-2, the opposite direction of lines in the Article III. Nine predictors explained the transit ridership variable better: the Medium/High Standard Horizontal Residential (MHSHR) area, no predominancy area, employment, park area, overlapping area ratio, distance to the center, distance to the nearest train station, number of lines through stops and number of shelters. Of them, only the MHSHR area, employment, and number of shelters did not have significant spatial variation in their coefficients.

The absence of significant differences between the interpolation results from network and Euclidean distances in the Article VI moved the author to explore another travel demand variable (the traffic volume along road segments) that occurs in a broader context than the urban variables analyzed so far (stop-level ridership). In article VIII, the entire state of São Paulo was used as a case study and two scenarios were considered: the first one with a low density of data and the second with a higher number of collected samples. Improvements provided by network distances were significant only in the first scenario, which, by having a low density of sampled data, reached higher differences between network and Euclidean distances. When the availability of data was increased (the second scenario), network and Euclidean distances started to be similar, approximating the interpolation results from both types of distances.

The next chapter summarizes the conclusions achieved on each specific objective of the thesis. The comments made in Section 7 include all articles from I to VIII

7 CONCLUSIONS AND FINAL CONSIDERATIONS FROM THE THESIS

> "*Perfection is not just about control. It's also about letting go. Surprise yourself so you can surprise the audience. Transcendence!*
> *[…] I felt it…it was perfect…*"
> Black Swan by Darren Aronofsky

This chapter synthetizes the conclusions provided by articles I to VIII on the specific objectives outlined in Subsection 1.3. The specific objectives are presented as questions in the next four subsections. A summary of conclusions is provided in subsection 7.5. The last subsection (7.6) outlines limitations and suggestions for future research. The articles are referred to in citation form, as listed in Table 7.1.

Table 7.1 – Article reference

Article	Reference
I	Marques and Pitombo (2021a)
II	Marques and Pitombo (2023a)
III	Marques and Pitombo (2023b)
IV	Marques et al. (2024)
V	Marques and Pitombo (2020)
VI	Marques and Pitombo (2021b)
VII	Marques and Pitombo (2022)
VIII	Marques et al. (2023)

7.1 WHAT ARE THE INTERVENING FACTORS OF THE STOP-LEVEL TRANSIT RIDERSHIP?

This question reproduces the specific objective (a) "To investigate what factors affect the stop-level transit ridership". The transit ridership predictors whose parameters appeared with statistical significance in different case studies are summarized in Table 7.2.

Table 7.2 – Factors affecting the stop-level ridership throughout the thesis articles

Predictor	Dependent variable(s)	Positive sign	Negative sign
Population	Boardings / Alightings	Marques and Pitombo (2021a)[1], Marques and Pitombo (2023a)[1/2]	Marques and Pitombo (2023a)[1/2], Marques and Pitombo (2021b)[1]
Commerce and service area	Boardings	Marques and Pitombo (2023a)[1]	Marques and Pitombo (2023a)[1]
Distance to the nearest metro station	Boardings	Marques and Pitombo (2021a)[1]	Marques and Pitombo (2021b)[1]
Distance to the nearest metro or train station	Boardings / Alightings	Marques and Pitombo (2023a)[1/2]	Marques and Pitombo (2023a)[2]
Distance to the nearest bus terminal, metro or train station	Boardings	Marques and Pitombo (2023a)[1]	Marques and Pitombo (2023a)[1]
Income	Boardings / Alightings	Marques and Pitombo (2023a)[1]	Marques and Pitombo (2023a)[1/2]
Frequency	Boardings / Alightings / Boardings plus alightings	Marques and Pitombo (2023a)[2], Marques and Pitombo (2021b)[1], Marques and Pitombo (2023b)[3]	Marques and Pitombo (2023a)[2], Marques and Pitombo (2023b)[3]
Lines through stops	Boardings / Boardings plus alightings	Marques and Pitombo (2021b)[1]	Marques and Pitombo (2022)[3]
Residential, commerce and service area	Boardings	Marques and Pitombo (2021a)[1]	
School area	Boardings	Marques and Pitombo (2021b)[1]	
MHSHR area	Boardings / Boardings plus alightings		Marques and Pitombo (2021b)[1], Marques and Pitombo (2022)[3], Marques and Pitombo (2023b)[3]
Intersections	Boardings / Boardings plus alightings	Marques and Pitombo (2021b)[1], Marques and Pitombo (2023b)[3]	
No car households (%)	Boardings	Marques and Pitombo (2021b)[1]	
Bus stops within buffer (overlapping)	Boardings / Boardings plus alightings		Marques and Pitombo (2021b)[1], Marques and Pitombo (2022)[3], Marques and Pitombo (2023b)[3]
No predominant land use	Boardings plus alightings	Marques and Pitombo (2022)[3], Marques and Pitombo (2023b)[3]	
Employment	Boardings plus alightings	Marques and Pitombo (2022)[3]	
Park area	Boardings plus alightings		Marques and Pitombo (2022)[3]
Distance to the center	Boardings plus alightings	Marques and Pitombo (2022)[3]	
Distance to the nearest train station	Boardings plus alightings	Marques and Pitombo (2022)[3]	
Number of shelters	Boardings plus alightings	Marques and Pitombo (2022)[3]	

[1, 2, 3] represent the cases where the dependent variable was Boarding, Alighting and Transit Ridership, respectively. [1/2] are the cases where the authors addressed both Boarding and Alighting, but separately.

Population was the explanatory variable most visualized throughout the case studies, followed by frequency, Medium-High Standard Horizontal Residential (MHSHR) area and an overlapping measure. While population is a demographic indicator, the MHSHR area represents the combined influence of population and income on the bus patronage, characterizing the socioeconomic condition around bus stops.

In turn, frequency is a service measure, often associated with the transit accessibility. The overlapping variable reflects the competition between stops, but, together with frequency, they reflect the importance of the bus network coverage to increasing the passenger demand.

As the regression modeling results from Article IV are based on the Geographically Weighted Principal Component Analysis and not on the isolated predictors, they were not included in Table 7.2. However, this paper pointed to a strong contribution of intersections, which characterizes the walkability in the bus stops' vicinity, and the area with no predominant land use, a measure of land use mix.

The first eight predictors in Table 7.2 (from population to lines through stops) had reverse signs in different case studies. A statistically significant spatial variation was detected in coefficients from the following variables: overlapping, frequency, no predominancy area, park area, center distance, distance to the nearest train station and lines through stops.

In short, results show that urban and transport planners should concentrate their efforts on four groups of factors: sociodemographic, bus network coverage, street layout and land use. Accounting for local characteristics when making a decision on these factors has been shown to be necessary to optimize their impacts on the sustainable transport.

7.2 WHAT IS THE BEST MODELING APPROACH FOR STOP-LEVEL RIDERSHIP DATA?

Specific objectives (b) "To assess the improvements in the estimates provided by the inclusion of explanatory variables in the geostatistical modeling of bus ridership"; and (c) "To compare the performance of spatial and local models of bus ridership with traditional approaches" are addressed by this question. In addition, this question is associated to two of the hypotheses described in Subsection 1.1: 1) Spatial approaches of ridership modeling can yield better estimates than non-spatial models; and 2) Ridership models considering the asymmetry of travel demand variables are more adequate than those based on the normality assumption. The first hypothesis was confirmed in the articles I, II, III, IV, VI and VII, while the second one was mainly attested by articles II and III.

Spatial dependence of transit ridership was consistently detected throughout the research papers. Therefore, spatial approaches should be preferred over the non-spatial ones as they provide better ridership predictions and a correct interpretation of the effects from intervening factors. In addition, including explanatory variables proved to significantly improve goodness-of-fit measures of stop-level ridership modeling as the related dependent variables usually have a wide range of variation that cannot be fully accounted for by univariate approaches.

Results have shown that checking for asymmetry and overdispersion is also an important step in the stop-level ridership modeling. Addressing these factors, together with the spatial dependence of boardings and alightings, proved to have a positive impact on the models' outcomes.

An instigating question would be whether prioritize a local spatial approach (the Geographically Weighted Regressions) over the spatial methods with a single coefficient for each explanatory variables (the multivariate interpolators). The answer to this question may depend on the purpose of the modeling: if the intention is focused on achieving the best ridership estimates with a low number of predictors, the interpolators might be preferred; however, if the main concern is about the effect and spatial heterogeneity of explanatory variables, a GWR approach has shown to be a better alternative.

On the other hand, the combination of GWR and kriging interpolation (Marques et al., 2024) could meet both requirements: good predictions and spatially varying relationships. A drawback of this method is perhaps the need for a larger predictor dataset.

7.3 WHAT IS THE BEST TYPE OF DISTANCE FOR STOP-LEVEL RIDERSHIP MODELING?

The current subsection is dedicated to the specific objective (d) “To compare the performance of spatial approaches of bus ridership using network distances and Euclidean distances”. The title question is linked to the following hypothesis: the distances along the bus lines can yield better estimates than with the traditional Euclidean distance. Conclusions on this issue are not simple.

Throughout the doctoral period, the comparison between spatial approaches using network and Euclidean distances was gradually evolved. In the first analysis, the two types of distances were used in the Ordinary Kriging of residuals from a regression model (Marques and Pitombo, 2021b); a more comprehensive investigation was provided by considering different densities of sampled data, which could increase the difference between network and Euclidean distances (Marques and Pitombo, 2023b). Finally, a third piece of research (Marques et al., 2023) covered another variable of interest, the traffic volume, along a rural road network significantly larger than the previous ones, which were based on urban bus lines.

Regarding the prediction power of proposed models, the two first articles reported a better performance of the network approach in some cases. However, in general, results from

network and Euclidean distances were comparable. When addressing the traffic volume over the entire state of São Paulo, network distances yielded better results only in a scenario with a low density of collected data.

However, a consensual conclusion from the three analyses is that, by considering the real path taken by the traveler, estimated parameters from the network approach are more appropriate than those from the Euclidean distance. In the case of spatial interpolators, this parameter refers to the autocorrelation range, while for GWR, the parameters refer to the coefficients from each explanatory variable and the bandwidth. The sensitivity analysis carried out in Marques and Pitombo (2023b) showed that, although some predictors may be represented similarly by both types of distance, the network approach is more appropriate for others.

Nevertheless, when dealing with large datasets, the use of network distances may become computationally expensive. If there are no resources available to calculate the network distances, an alternative approach would be to incorporate the data anisotropy into the modeling to improve the prediction accuracy (Marques et al., 2023). However, this solution can be applied only in the context of geostatistical interpolators, and for cases with availability of sampled data in multiple spatial directions.

7.4 ARE THE SPATIAL MODELS CAPABLE OF PROVIDING GOOD RIDERSHIP ESTIMATES IN UNSAMPLED STOPS?

This question analyzes the last two specific objectives: (e) “To evaluate the performance of spatial and local models on predicting stop-level ridership data in unsampled stops”; and (f) “To analyze the effect of the sampling strategy on the prediction accuracy of stop-level ridership models”. The thesis hypothesis serving as a basis to this question is: spatial modeling of ridership data can overcome the problem of data scarcity, regarding boarding and alighting per bus stop. Articles III and IV confirmed this hypothesis.

The detailed missing data analysis of Marques and Pitombo (2023b) proved that a local spatial model is able to achieve satisfactory performance when predicting a stop-level ridership data in unsampled stops. Some missing data samples had goodness-of-fit measures close to the calibration samples. Best results were seen in scenarios 1 and 4, with validation samples equal to 15% and 60% of the dataset, respectively. Scenario 4 showed the lowest Median Absolute Percentage Error (MedAPE): 40%, approximately.

The prediction power of the spatial interpolation was attested in the fourth article. The best result refers to a MedAPE of 33% on a 30% validation sample. The contribution of spatial interpolation was higher in estimating missing data than calibration data.

Confirming the hypothesis mentioned above has implications to the field collection of stop-level ridership data. Transit agencies are provided with evidence that allows them to carry out the counting of boardings and alightings only for a pre-defined number of stops and then apply a spatial model to predict the ridership data in the remaining bus stops. This may encourage municipalities that face a lack of stop-level ridership data, mainly due to budget constraints, to perform the collection of boarding and alighting, supporting an optimized bus network planning and promoting the transit-oriented development. The first approach to the possible sampling strategies is shown in Marques et al. (2024), pointing to a better performance of the balanced sampling with geographical spreading.

7.5 SUMMARY OF CONCLUSIONS

Table 7.3 summarizes the main conclusions on the specific objectives and hypotheses outlined in Section 1.

Table 7.3 – Summary of the main conclusions

Specific objectives	Hypotheses	Main conclusions
(a) “To investigate what factors affect the stop-level transit ridership”		Four groups of factors gather the most important predictors: sociodemographic, bus network coverage, street design and land use
(b) “To assess the improvements in the estimates provided by the inclusion of explanatory variables in the geostatistical modeling of bus ridership”		Multivariate models should be preferred over the univariate ones
(c) “To compare the performance of spatial and local models of bus ridership with traditional approaches”	1) Spatial approaches of ridership modeling can yield better estimates than non-spatial models; 2) Ridership models considering the asymmetry of travel demand variables are more adequate than those based on the normality assumption	Models that consider asymmetry and spatial dependence should be prioritized over the ones that overlook these characteristics

Specific objectives	Hypotheses	Main conclusions
(d) "To compare the performance of spatial approaches of bus ridership using network distances and Euclidean distances"	The distances along the bus lines can yield better estimates than with the traditional Euclidean distance	Despite the network distance approach not contributing significantly to improving the models' prediction power, this type of distance may better represent the relationship between the transit ridership and its intervening factors. Prioritizing the use of network distances in the spatial modeling of boardings and alightings is recommended
(e) "To evaluate the performance of spatial and local models on predicting stop-level ridership data in unsampled stops"	Spatial modeling of ridership data can overcome the problem of data scarcity, regarding boarding and alighting per bus stop	The spatial models proved to be able to estimate the volume of boardings and alightings in unsampled points accurately, even in some cases with low availability of calibration data
(f) "To analyze the effect of the sampling strategy on the prediction accuracy of stop-level ridership models"		The balanced sampling with geographical spreading showed the best performance in estimating a transit ridership variable in an unsampled bus stop

7.6 LIMITATIONS AND FUTURE DEVELOPMENTS

The main limitation of the thesis is probably the low coverage of the stop-level ridership database. In addition, for bus stops serving each sampled line, the boarding and alighting survey counted only passengers using that specific line. Therefore, the transit ridership measurement was not the same for all lines visited during the survey. These facts prevented the possibility of including the 16 unidirectional lines in a case study, which would allow a higher spatial representativity of the dependent variables.

However, this observation confirmed the urgency of proposing solutions to the lack of stop-level ridership data, which was one of the main motivations of the thesis. Therefore, other municipalities facing this problem can benefit from the conclusions achieved. Despite the limitations, we believe that the research carried out was successful in extracting all the information available in an efficient way to assure strong contributions.

Nevertheless, there are still non-explored topics, which can provide excellent material for forthcoming studies. One of them refers to the concept of loss function in the geostatistical framework. It consists of a way to penalize the prediction errors of kriging and is based on the squared differences between real and estimated values. However, when dealing with stop-level ridership data, the common loss function embedded in kriging might not be the best representation of the errors' shortcomings. Calculating a more appropriate loss function and adjusting the kriging weights from it is an exciting path for future research.

This suggestion converges with exploring the sequential simulation methods for stop-level ridership prediction. This type of modeling relaxes the smoothed pattern of kriging predictions, allowing the wide variation of boardings and alightings to be better captured in the estimation process.

Analyzing the sensitivity to missing data of different sampling strategies is also an interesting topic to be investigated. It is possible that the best sampling method varies according to the percentage of points in the calibration samples. Testing several sampling methods in the context of Geographically Weighted Regressions is also a recommended topic.

Although we have investigated only the variables of boardings and alightings, an alternative research line could focus on the loading variable, which is given at the route-segment level. The loading information (number of passengers inside the bus) is derived from boardings and alightings and its maximum value is often applied for an optimized fleet sizing. Analyzing this variable from a spatial perspective can provide useful insights to the transit-oriented development.

One final suggestion is to address the potential anisotropy of stop-level ridership data in cases where sampled bus stops are available for multiple spatial directions in the city. As shown by Marques et al. (2023), this approach is a lower cost alternative to using network distances when the data on the variable of interest is densely distributed over the spatial field under analysis.

REFERENCES

Blainey, S., and Mulley, C. (2013) Using geographically weighted regression to forecast rail demand in the Sydney region. *Australasian Transport Research Forum 2013 Proceedings*. Brisbane, Australia. Available at https://australasiantransportresearchforum.org.au/wp-content/uploads/2022/03/2013_blainey_mulley.pdf. Accessed in May 2023.

Blainey, S., and Preston, J. (2010) A geographically weighted regression based analysis of rail commuting around Cardiff, South Wales. *12th World Conference on Transport Research*. Lisbon, Portugal.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996) Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, *28*(4), 281–298. doi:10.1111/j.1538-4632.1996.tb00936.x

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

Ceder, A. (2007) *Public transit planning and operation: modeling, practice and behavior*. CRC press.

Cervero, R. (2006) Alternative Approaches to Modeling the Travel-Demand Impacts of Smart Growth. *Journal of the American Planning Association*, *72*(3), 285–295. doi:10.1080/01944360608976751

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

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

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

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

Dill, J., Schlossberg, M., Ma, L., and Meyer, C. (2013) Predicting transit ridership at the stop level: The role of service and urban form. *Transportation Research Board 92nd Annual Meeting* (p. 1–19). Washington DC, United States. 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. Accessed in May 2023.

Eom, J. K., Park, M. S., Heo, T.-Y., and Huntsinger, L. F. (2006) Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. *Transportation Research Record*, (1968), 22–29. doi:10.3141/1968-03

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

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

IBGE. (2021) São Paulo. *IBGE cities*. Brazilian Institute of Geography and Statistics. Available at https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama

Kerkman, K., Martens, K., and Meurs, H. (2015) Factors Influencing Stop-Level Transit Ridership in Arnhem–Nijmegen City Region, Netherlands. *Transportation Research Record*, *2537*(1), 23–32. doi:10.3141/2537-03

Liu, Y., Ji, Y., Shi, Z., and 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. doi:10.3390/su10124684

Marques, S. de F., Favero, R., and Pitombo, C. S. (2023) Should we account for network distances or anisotropy in the spatial estimation of missing traffic data? *Transportes*, *31*(1), e2822. doi:10.58922/transportes.v31i1.2822

Marques, S. de F., and Pitombo, C. S. (2020) Intersecting Geostatistics with Transport Demand Modeling: a Bibliographic Survey. *Revista Brasileira de Cartografia*, *72*, 1028–1050. doi:10.14393/rbcv72nespecial50anos-56467

Marques, S. de F., and Pitombo, C. S. (2021a) Applying multivariate Geostatistics for transit ridership modeling at the bus stop level. *Bulletin of Geodetic Sciences*, *27*(2). doi:10.1590/1982-2170-2020-0069

Marques, S. de F., and Pitombo, C. S. (2021b) Spatial modeling of transit ridership along bus lines with overlapping sections. *Proceedings of the 35º Annual Conference of the Brazilian National Association for Transportation Research and Teaching* (p. 1568–1580). Brazilian National Association for Transportation Research and Teaching, 100% Virtual. Available at https://www.researchgate.net/publication/357517939_SPATIAL_MODELING_OF_TRANSIT_RIDERSHIP_ALONG_BUS_LINES_WITH_OVERLAPPING_SECTIONS. Accessed in May 2023.

Marques, S. de F., and Pitombo, C. S. (2022) Spatially varying relationships of transit ridership at the bus stop level. *Proceedings of the 36º Annual Conference of the Brazilian National Association for Transportation Research and Teaching* (p. 152517). Brazilian National Association for Transportation Research and Teaching, Fortaleza, Brazil. Available at https://proceedings.science/anpet-2022/trabalhos/spatially-varying-relationships-of-transit-ridership-at-the-bus-stop-level?lang=pt-br. Accessed in May 2023.

Marques, S. de F., and Pitombo, C. S. (2023a) Transit Ridership Modeling at the Bus Stop Level: Comparison of Approaches Focusing on Count and Spatially Dependent Data. *Applied Spatial Analysis and Policy*, *16*(1), 277–313. doi:10.1007/s12061-022-09482-y

Marques, S. de F., and Pitombo, C. S. (2023b) Local modeling as a solution to the lack of stop-level ridership data. *Journal of Transport Geography*, *112*, 103682.

doi:https://doi.org/10.1016/j.jtrangeo.2023.103682

Marques, S. de F., Pitombo, C. S., and Gómez-Hernández, J. J. (2024) Spatial modeling of travel demand accounting for multicollinearity and different sampling strategies: a stop-level case study. *Journal of Advanced Transportation*, *2024*, 7967141. doi:10.1155/2024/7967141

Matheron, G. (1963) Principles of geostatistics. *Economic Geology*, *58*(8), 1246–1266. doi:10.2113/gsecongeo.58.8.1246

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

Moran, P. A. P. (1948) The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 243–251. doi:10.1111/j.2517-6161.1948.tb00012.x

Pulugurtha, S. S., and Agurla, M. (2012) Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. *Journal of Public Transportation*, *15*(1), 33–52. Available at https://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1095&context=jpt. Accessed in May 2023.

Rodrigue, J. P., Comtois, C., and Slack, B. (2016) *The geography of transport systems*. *The Geography of Transport Systems*. doi:10.4324/9781315618159

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

Taylor, B. D., and Fink, C. N. Y. (2013) Explaining transit ridership: What has the evidence shown? *Transportation Letters*, *5*(1), 15–26. doi:10.1179/1942786712Z.0000000003

Tobler, W. R. (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234–240. doi:10.2307/143141

Vuchic, V. R. (2005) *Urban Transit: Operations, Planning, and Economics*. John Wiley & Sons.

Wackernagel, H. (2003) *Multivariate Geostatistics: an introduction with applications*. (3rd.). Kluwer Aeademic Publishers, Berlin. doi:10.1007/978-3-662-05294-5

Wang, X., and Kockelman, K. (2009) Forecasting Network Data. *Transportation Research Record: Journal of the Transportation Research Board*, *2105*, 100–108. doi:10.3141/2105-13

Zhao, F., Chow, L.-F., Li, M.-T., Ubaka, I., and Gan, A. (2003) Forecasting Transit Walk Accessibility: Regression Model Alternative to Buffer Method. *Transportation Research Record*, *1835*(1), 34–41.

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

EESC • USP