Journal of Transport Geography 118 (2024) 103906

Contents lists available at ScienceDirect

Journal of Transport Geography

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

A spatial statistical approach to estimate bus stop demand using GIS-processed data

Yaiza Montero-Lamas [a,*], Rubén Fernández-Casal [b], Francisco-Alberto Varela-García [c], Alfonso Orro [a], Margarita Novales [a]

[a] Universidade da Coruña, Group of Railways and Transportation Engineering, ETS Ingenieros de Caminos, Canales y Puertos, Campus de Elviña, 15071 A Coruña, Spain
[b] Departamento de Matemáticas, CITIC, Facultad de Informática, Universidade da Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain
[c] cartoLAB, Grupo de Visualización Avanzada e Cartografía, Departamento de Ingeniería Civil, E.T.S. Ingeniería de Caminos, Canales y Puertos, Universidade da Coruña, A Coruña, Spain

ARTICLE INFO

Keywords:
Geospatial analysis
Spatial dependence
GIS
Generalized additive model
Bus stop demand estimation
Transit planning

ABSTRACT

This study integrates the fields of geography, urban transit planning, and statistical learning to develop a sophisticated methodology for predicting bus demand at the stop level. It uses a Generalized Additive Model that captures non-linear relationships and incorporates spatial dependence, improving traditional methods. It showcases a high predictive capacity with a pseudo R-squared of 0.79 during its validation, ensuring substantial explanatory power for new observations. A large number of variables, including land-use characteristics, socioeconomic factors, and transit supply, are analysed. These widely available predictors facilitate the transferability of the methodology to other urban areas. Transit supply predictor considers the number of annual trips per stop and area as well as the location of stops along the lines that serve them. GIS processing of the data allows the calculation of variables within the areas of influence of each stop, obtained by following the walkable street network. For the case study, the presence of universities, hospitals, and lodgings areas, as well as inhabitants and ratio of bus trips show a positive impact on bus demand. This geo-analysis process employs accurate disaggregated data, such as information on uses in each building, as well as methods for assigning socioeconomic information from local areas to residential buildings. This study highlights the complex relationship between the location of transit network stops, both along the bus line and in terms of geographical proximity, their transit supply, and its surrounding factors. The results indicate that there is spatial dependence for stops less than 1.15 km apart. The developed methodology provides reliable information to transit network planners for decision making. Specifically, this proposed methodology can contribute to designing new routes, optimizing stop locations, and estimating the impact of changes in the transit network or urban planning on bus demand. All these improvement measures promote sustainable urban mobility, consequently fostering environmental and social benefits.

1. Introduction

Public transportation is key to urban mobility, playing a vital role in shaping modern cities. As cities grow and evolve, understanding and optimizing transit systems become essential for improving quality of life, reducing traffic congestion, and minimizing environmental impacts. A robust urban transit system is not only an investment in urban sustainability but also a commitment to social inclusion, fostering inclusive and connected communities. To achieve these benefits, urban transit demand assessment, which explores the factors that influence passengers' decisions in the urban environment, is imperative.

Effective decision-making is fundamental in the field of urban transit, where transit system managers and planners deal with complex challenges and usually limited budgets. These challenges range from optimizing bus stop locations to creating new transit routes and assessing the impact of urban developments on transit ridership. To address these intricate demands, our research aims to develop a versatile tool that empowers transit planners with indispensable information.

In this context, examining transit demand at the bus stop level is particularly enlightening, as bus stops represent the front lines of urban

* Corresponding author.
E-mail address: y.montero@udc.es (Y. Montero-Lamas).

https://doi.org/10.1016/j.jtrangeo.2024.103906
Received 7 December 2023; Received in revised form 1 May 2024; Accepted 4 June 2024
Available online 10 June 2024

transit. Bus boardings depend on a multitude of factors beyond just transit supply. Understanding why and where individuals use transit services at these specific locations can provide essential insights for transit planners. Such insights can enable the enhancement of transit networks to better serve to the different needs of the city's population. Our research aims to address this challenge by developing a novel model that considers the socioeconomic and land-use characteristics of urban areas within the bus stops' surroundings along with transit supply data, to predict bus demand. This model must account for potential similarities in the response of nearby stops, thereby addressing the need to consider spatial dependence in the observations.

The subsequent sections of this paper are organized as follows: Section 2 reviews related previous studies, while Section 3 introduces the case study and data sources. In Section 4, we detail GIS data processing, and Section 5 focuses on the development of advanced statistical models. Section 6 presents results and discussions, and finally, Section 7 concludes with findings and recommendations for future works.

2. Literature review

Understanding transit demand is fundamental in the fields of urban planning and transit management, presenting one of the main current challenges for local governments and transit agencies. Transit demand research delves into the complexities of why, when, and where individuals use transit services, shedding light on the intricate relationship between urban environments and mobility patterns. By comprehending the users' perspectives behind transit demand, we can adapt and optimize transit networks to better serve communities, adjust budgets, alleviate traffic congestion, reduce environmental impacts, and ultimately improve the quality of life in urban areas.

In this regard, analysing transit demand at the bus stop level proves particularly insightful, as it reveals a complex interplay of factors that influence passenger decisions. The transit demand at the bus stops level could be influenced by a variety of factors, such as population density, level of incomes, land use characteristics, accessibility, and transit supply. Analysing and quantifying how these different factors can impact bus boardings at the stops level is a key objective of transit demand research. Some analyses aim to estimate bus demand when changes in nearby bus-related factors are planned or during the planning of new public transport routes (Ryan and Frank, 2009; Amoroso et al., 2010; Kerkman et al., 2015). Furthermore, other studies aimed to examine the impact on bus demand at the stop level under special conditions, such as the restrictions imposed during the COVID-19 pandemic, in order to identify city areas that should be reinforced in the event of a future exceptional situation (Hu and Chen, 2021; Montero-Lamas et al., 2022).

In the context of analysing transit boardings to explain bus patronage in relation to surrounding socioeconomic factors and built environment, it is essential to delineate the influence area of bus stops. Various studies approach this task differently, some aligning bus stop characteristics with the attributes of the census tract in which they are located (Chakrabarti, 2015; Ngo, 2019). Others define this area based on user accessibility, often considering a walking time or, more commonly, a walking distance criterion. The ideal spatial proximity between a bus stop and its area of influence depends on factors such as the city's size, dispersion, population behaviour, and urban and transit planning. Typically, these studies adapt the walking distance metric to fit the city under examination, often calculating it as a straight-line distance rather than tracing the walkable street network (Johnson, 2003; Chakour and Eluru, 2013, 2016). Marques and Pitombo (2021a, 2023a) performed a bus ridership estimation comparing both criteria. The factors within each delineated area of influence or census section are analysed to estimate bus demand at the bus stop level.

In addition to selecting the factors that may impact bus boardings and defining the area of influence, it is fundamental to choose an appropriate model for performing the estimations. Evaluating such impacts with the traditional four-step process (see e.g. Ortúzar and Willumsen, 2011) is not suitable due to the size of the analysis zones typically employed (Chu, 2004). In the field of transit planning, the literature offers a range of statistical methods and approaches. Direct models, in contrast to the four-step modelling approach, have the capability to estimate transit boardings at individual bus stops (Cervero, 2006). These models encompass techniques ranging from Multiple Linear Regression (MLR) models and Generalized Linear Models (GLM) to categorical and dynamic models. Some studies have applied, evaluated, and even compared some of these models to verify their effectiveness in several contexts (Pulugurtha and Agurla, 2012; Marques and Pitombo, 2021b, 2023b).

Cervero et al. (2010) studied the bus rapid transit (BRT) ridership in Southern California, employing ordinary least squares (OLS) regression and hierarchical linear model (HLM). They predicted the average daily boardings for 69 bus stops, considering a range of predictors such as daily buses, the presence of perpendicular daily feeder bus lines, daily rail feeder trains, distance to the nearest BRT stop, population density, park-and-ride lot capacity, and total density (population and employment). Their results showed a positive influence of all these variables on BRT ridership. Notably, the study highlighted the significance of service intensity, both BRT system and the feeder services. Furthermore, high levels of intermodal connections were found to be relevant for attracting BRT users. Besides, the primary neighbourhood attribute influencing BRT ridership was the population density within a 1/2-mile radius of a bus stop. Metro Rapid BRT stops located in areas with higher residential densities exhibited greater ridership.

Johnson (2003) estimated weekday bus boardings by analysing 3362 stops within Sector 5 of Minneapolis-St. Paul, USA. To achieve this, he employed an MLR model and considered transit service, socioeconomic characteristics, and land uses as predictors. The results revealed that the most significant variables, all with a positive impact on demand, were population density, adjacent vertical mixed-use, and retail-commercial use within a 1/8-mile radius. Cui et al. (2022) employed MLR models to forecast daily bus boardings, primarily examining job accessibility alongside variables like population, household income, daily stop departures, and stops serving the same routes. To evaluate competition among close stops, they examined job accessibility overlaps in a 1/4 mile (about 400 m) service area. They discovered that a 1% rise in 30-min job accessibility would increase bus boardings by 0.21%. Using ensemble machine learning, they found accessibility to be the most influential variable in their models, when not accounting for the daily departures variable.

Pulugurtha and Agurla (2012) performed a comparison among Linear, Gamma, Poisson, and Negative Binomial Regression (NBR) models to estimate the average daily bus transit ridership at the bus stop level. NBR, which incorporates overdispersion of transit ridership, performed better than the rest. The authors considered various spatial attributes, including land use, road network, demographic, and socioeconomic characteristics at a census section level. The research concluded that the Spatial Proximity Method (SPM) with a 0.25-mile buffer radius provides superior goodness-of-fit statistics for estimating bus stop ridership. The study reveals that factors such as the presence of institutional areas, light and heavy commercial zones contribute positively to ridership, while mean household income and certain residential and light industrial areas show a negative influence. The study highlights the importance of walking distance as a critical factor in ridership estimation.

NBR models have also been used by other authors in this field. Ngo (2019) used an NBR model to determine the effect of weather events on bus ridership and its variation based on the annual household income per census tract and destination (municipal parks and commercial zones). Chakrabarti (2015) examined the influence of bus service reliability on stop level boardings in the Los Angeles Metro bus system using an NBR model as well. The analysis included factors like on-time

performance, density, median income, and the position of bus stops in relation to line ends. They found that boardings were higher in census tracts with dense populations and employment.

Chakour and Eluru (2016) assessed the impact of built-environment factors on bus ridership (boardings and alightings) at the bus stop level in Montreal, considering four daily time periods. Their study involved the delineation of influence areas around bus stops using straight-line buffers. The analysis examined factors in proximity to each bus stop, encompassing variables related to transit operations, transit accessibility metrics, attributes of transportation infrastructure, and measurements of land use characteristics. To study the influence of these factors on bus ridership, they employed an ordered response probit (ORP) model. Their findings indicated that the presence of transit facilities and nearby parks positively impacted on bus ridership.

When choosing a model to predict bus demand at stop level, few studies have considered that the bus stop location and proximity might vary the influence of different factors on bus boardings. To address this, a geospatial analysis should be performed. Marques and Pitombo (2023a) modelled bus ridership at the stop level in Sao Paolo, Brazil, belonging to a developing country. The lack of available data is the main challenge faced by the study, with 97 stops analysed from two bus lines. They performed a Geographically Weighted Negative Binomial Regression (GWNBR; da Silva and Rodrigues, 2014) model after problems of overdispersion and spatial dependence were found with previous models. Their results revealed that medium/high standard residential areas, areas without a predominant land use, intersections, overlapping area ratios, and frequency better explain transit ridership. They concluded that GWNBR model is a useful tool for estimating bus ridership at unregistered stops, even in data-limited scenarios, similar to the case study. These same authors have used Ordinary Kriging (OK) as a spatial interpolation technique in a previous study (Marques and Pitombo, 2021a) to estimate bus ridership at the stop level (and sections). They additionally performed a comparative evaluation using Universal Kriging (UK), OK, and OLS regression (Marques and Pitombo, 2021b). Subsequently, Marques and Pitombo (2023b) conducted a comparison of different approaches for modelling transit ridership at the bus stop level, including UK, Linear, Poisson, Geographically Weighted (GW), and Geographically Weighted Poisson (GWP) regressions.

Rahman et al. (2021) performed two forms of spatial autocorrelation models to study bus ridership at the stop level in Greater Orlando region: spatial autoregressive model (SAR) and spatial error model (SEM). SEM captures spatial interactions by introducing a spatial autocorrelation component within the error term, whereas SAR model considers spatial dependence in the dependent variable, by assuming that an observation's value might be influenced by nearby observations. They considered several buffer distances from bus stops to create the predictors. Land uses were considered per census tract where the stop is located. The results confirm the impact of spatial effects on bus ridership.

After reviewing the related works, it is evident that many models have been used so far to estimate bus demand at the stop level based on its surroundings data. However, few of them have addressed spatial dependence. This study applies a Generalized Additive Model (GAM) with spatial correlation to estimate bus boardings at the bus stop level, a novel approach not previously performed in bus ridership prediction studies, thereby contributing to this field of research. This statistical model allows for the consideration of both linear and non-linear relationships of variables, being a generalized additive model, and it also addresses spatial dependence by considering spatial correlated errors. In Section 5, we will delve into the development and characteristics of this model.

The input variables are based on open data, widely available in most cities. These data are processed and refined using a Geographic Information System (GIS) software. Based on census section data, we estimate the population and their socioeconomic characteristics per building, all geolocated as building's centroids. Therefore, data within the same census section vary significantly depending on the presence and type of buildings, the number of dwellings, and geolocated land uses. This level of detail, reproducible in other cities, has not been seen in other authors' works and contributes to the reliability of the study. Additionally, considering the actual walking network from the bus stop also brings us closer to the reality we aim to analyse. Another contribution is the creation of a transit supply variable that considers, on one hand, the position of the bus stop along the lines serving it, as attempted Chakrabarti (2015) with a categorical variable. On the other hand, this variable also addresses the transit supply competition between all stops within each influence area. While these two elements have been separately analysed in some research, this study creates a variable that accounts for both transit aspects.

The innovative methodology proposed in this study, from the collection and processing of variables to the modelling of bus boardings and their validation, can be transferred to other urban areas, enabling transit planners to make informed decisions regarding route design, stop optimisation, and the impact of changes in the transit network or urban planning on bus demand. This research harnesses the power of statistical learning and contributes to the understanding of the complex relationships between bus stop location, transit supply and surrounding factors. Overall, this study contributes to the field of urban bus demand assessment and provides a valuable tool for improving transit systems in growing and evolving cities.

3. Case study and data collection

This study was conducted in the city of A Coruña, Spain, which has the densest urban bus network in the country and a population of 244,700 inhabitants. The operating company of the urban bus service records all the boardings at each bus stop and line, among other data. The total annual boardings in 2019 per bus stop have been selected as the dependent variable we will use to build a model that allows to eventually estimate the annual transit demand at each stop of the city in a different situation. Regarding the independent variables, we have considered socioeconomic characteristics, land use, and transit supply variables, as all these factors can influence the use of urban transit.

Detailed land use data for A Coruña has been downloaded from the Cadastral website. More than 200 land-use types and 280,000 data points from the city and its surroundings were initially obtained, which were filtered and grouped. Each of these data points is georeferenced and associated with its corresponding area. After analysing the data, land uses related to health, education, offices, retail, hospitality, cultural, religious, and sports centres, among others, have been preselected to be processed using a GIS software.

Regarding the socioeconomic variables, public available information has been downloaded from the National Institute of Statistics (INE) and the Galician Institute of Statistics (IGE) at the census section level. These data include population information related to age, income, occupation, and country of origin, among others.

We have also considered the transit supply to explain the bus demand at stops level. The independent variable used to do so is the number of annual trips per stop (recorded by the operating company), multiplied by a position-specific coefficient that considers the number of bus lines passing through each stop and the stop relative location along the lines. This annual trips variable, as well as socioeconomic data, is also processed using a GIS software as it is explained in detail in the next section.

4. GIS data processing

To comprehend the impact of the aforementioned variables on bus stop demand, it is essential to use data that are not only refined but also enriched with accurate spatial information. The data must be exactly georeferenced to reflect the current reality within the area of influence of each bus stop, thereby enhancing the study.

4.1. Spatial socioeconomic data refinement

The socioeconomic data associated with the resident population, initially obtained at the census section level (the most detailed level provided by INE and IGE), has been allocated to each residential building in the city based on the inhabitants within each section and the number of dwellings in each edifice. The census sections used in our case study range between 691 and 2793 inhabitants (minimum and maximum values, respectively), based on data from 2019. The allocation within each census section has two main purposes. Firstly, it distinguishes areas close to bus stops without buildings, those consisting only of non-residential uses, and buildings with residential use. Secondly, within residential buildings, it considers the number of dwellings in each edifice close to bus stops. To perform this task, the following process was applied in Quantum GIS (QGIS 3.28.3):

i. Geolocated building data is downloaded from the Cadastral website with the corresponding land use and number of dwellings. The centroid of the polygon representing the geolocated buildings is generated. The centroid is created to facilitate the assignment of residents to each building.
ii. Only buildings with dwellings are selected, and the total number of dwellings in each census section is added. By dividing the total population in each census section by the total number of dwellings, the average number of inhabitants per dwelling in each census section is calculated.
iii. This value is associated with each building (centroid), and it is then multiplied by the number of dwellings within each building to obtain the estimated number of inhabitants per building. Fig. 1 shows this value across the city (along with the AdjTrips per stop variable, that will be explained in Section 4.3).

The remaining socioeconomic variables also consider the existence

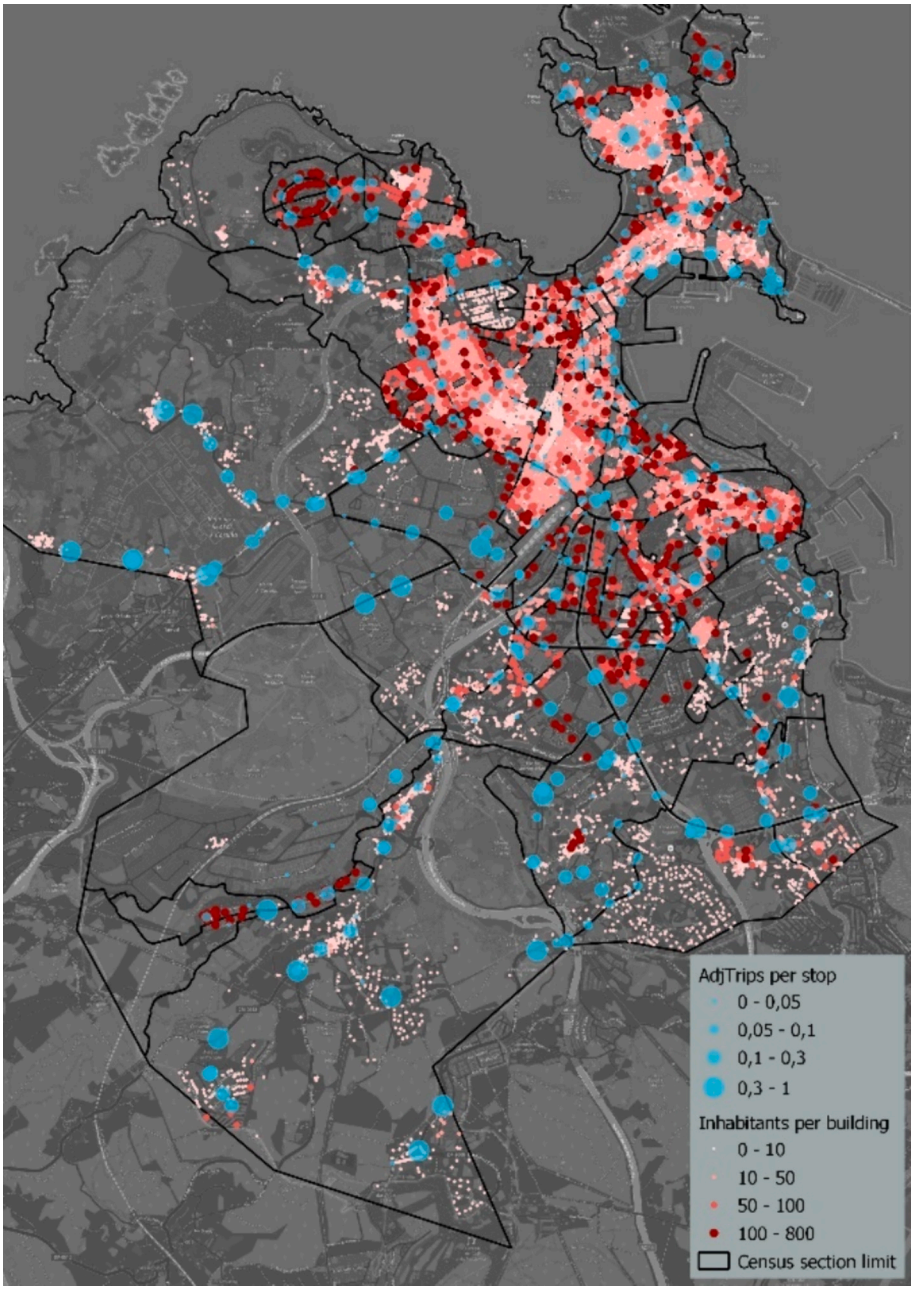

Fig. 1. Map of the entire city studied including inhabitants and stops information.

and number of dwellings of each residential building within the same census section. Variables related to the number of inhabitants meeting specific criteria (age range, occupation, or origin) are calculated as percentages within the census section and multiplied by the inhabitants of each building within that section. Other socioeconomic variables, such as index, average age, or incomes, are directly multiplied by the number of inhabitants per building according to the census section in which they are located. In a subsequent step, the total values of these variables within the influence area of the bus stop are aggregated and then divided by the total population in that area. This process allows us to calculate a weighted average of these variables for each stop, considering buildings belonging to different census sections within the area of influence of the stop. This step is not necessary for land use variables, as they are already geolocated in the corresponding premises with their assigned area.

4.2. Influence area of bus stops

In previous studies, a straight-line radial distance from the bus stop has been considered to define its influence area (George, 1999; Noh et al., 2021; Montero-Lamas et al., 2022). However, this study conducts a notable improvement to avoid erroneous assumptions and over-estimations associated with conventional radial influence areas (El-Geneidy et al., 2014). To achieve this goal, the existing street network has been considered to determine the influence area of each stop within a real walking distance. The value of 500 m has been selected based on its common use in cities with similar characteristic (Talavera-Garcia and Valenzuela-Montes, 2018). This approach ensures that inaccessible areas for pedestrians are excluded and aligns the analysis as closely as possible with the reality and actual pedestrian routes taken by bus users. This alignment with reality enhances the credibility of the analysis and ensures that the influence area reflects the practical considerations of urban transportation dynamics. The socioeconomic and land-use characteristics within this defined area are considered to explain the factors driving the number of passenger boardings at each bus stop, for generated or attracted trips, respectively.

The procedure for obtaining the characteristics within the area of influence was developed using a custom specific model in QGIS that is summarized in Fig. 2. Based on the walkable street network (derived from Open Street Map, OSM, information), the bus stops layer of the city, and the 500 m distance as inputs, the model first creates the service area from each bus stop following the line that indicates the street axis (purple colour in Fig. 3). Then, a 30 m buffer is applied to ensure that all calculated geometries contain relevant information. This step prevents anomalies resulting from road network configurations that might create linear areas of influence not containing the buildings' centroids. This step is followed by the generation of the corresponding minimum bounding geometry. This geometry represents the influence area of the bus stops. Fig. 3 shows the area of influence of a stop (blue area) near the railway station following this procedure, highlighting the contrast between considering the walking distance (yellow lines) and the straight-line radial distance (dark grey circle) to which a 30 m buffer has also been added to perform the comparison. This procedure is particularly relevant for stops located near physical obstacles where pedestrians cannot pass through. In Fig. 3, the buildings considered for calculations within the influence area are outlined with a thicker polygonal border. However, buildings are treated as points generated at their centroid.

The subsequent step involves transferring the land use and socio-economic data within each area of influence to the corresponding bus stop. The total area of each land use category for all buildings within the area of influence is calculated and associated with its respective bus stop. Socioeconomic variables are also added based on the information calculated for each of the buildings within the area of influence, which are represented as geographic points. In addition, for indexes and income-related variables, it is necessary to divide them by the total population of the corresponding area of influence to obtain their

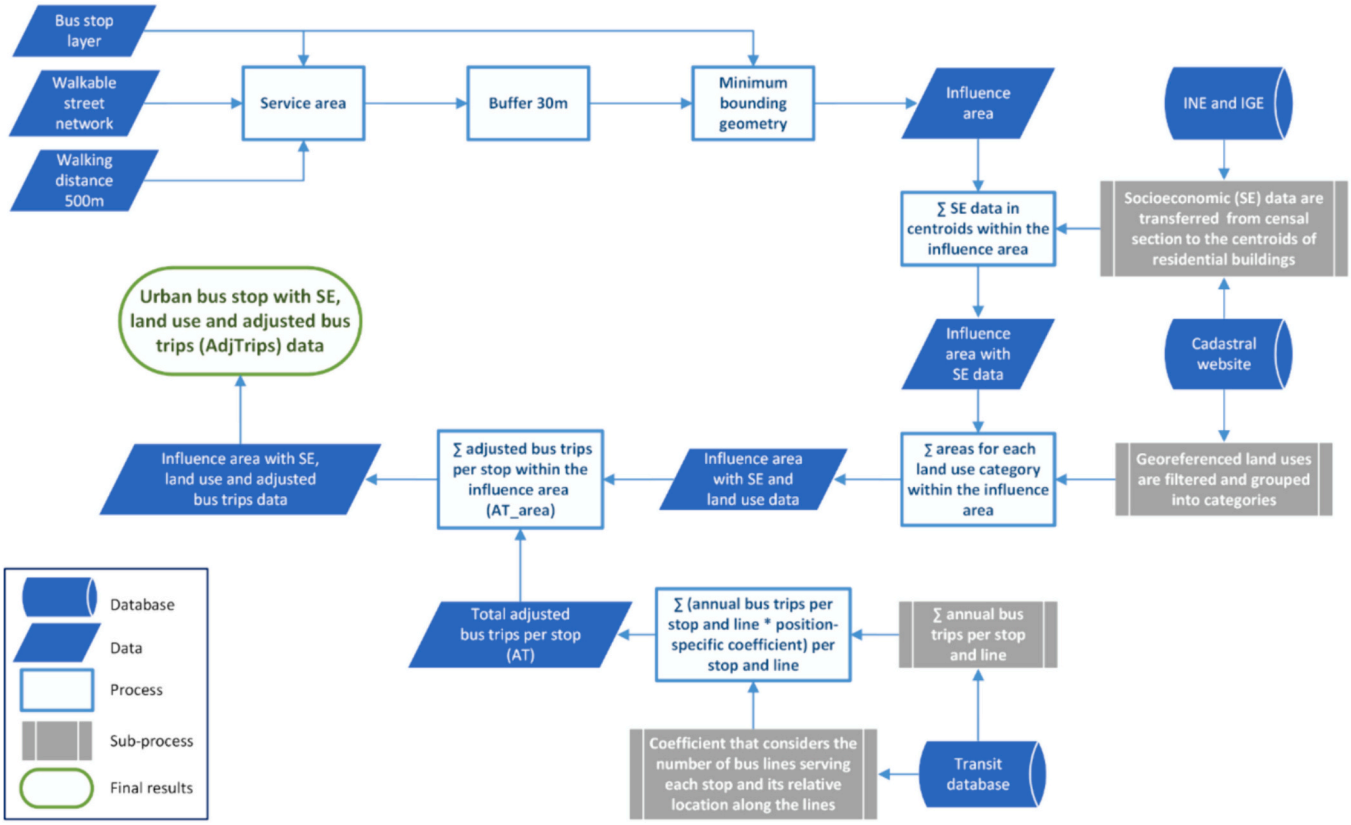

Fig. 2. Flowchart illustrating the procedure for calculating the socioeconomic, land use, and transit data within the respective influence area of each bus stop.

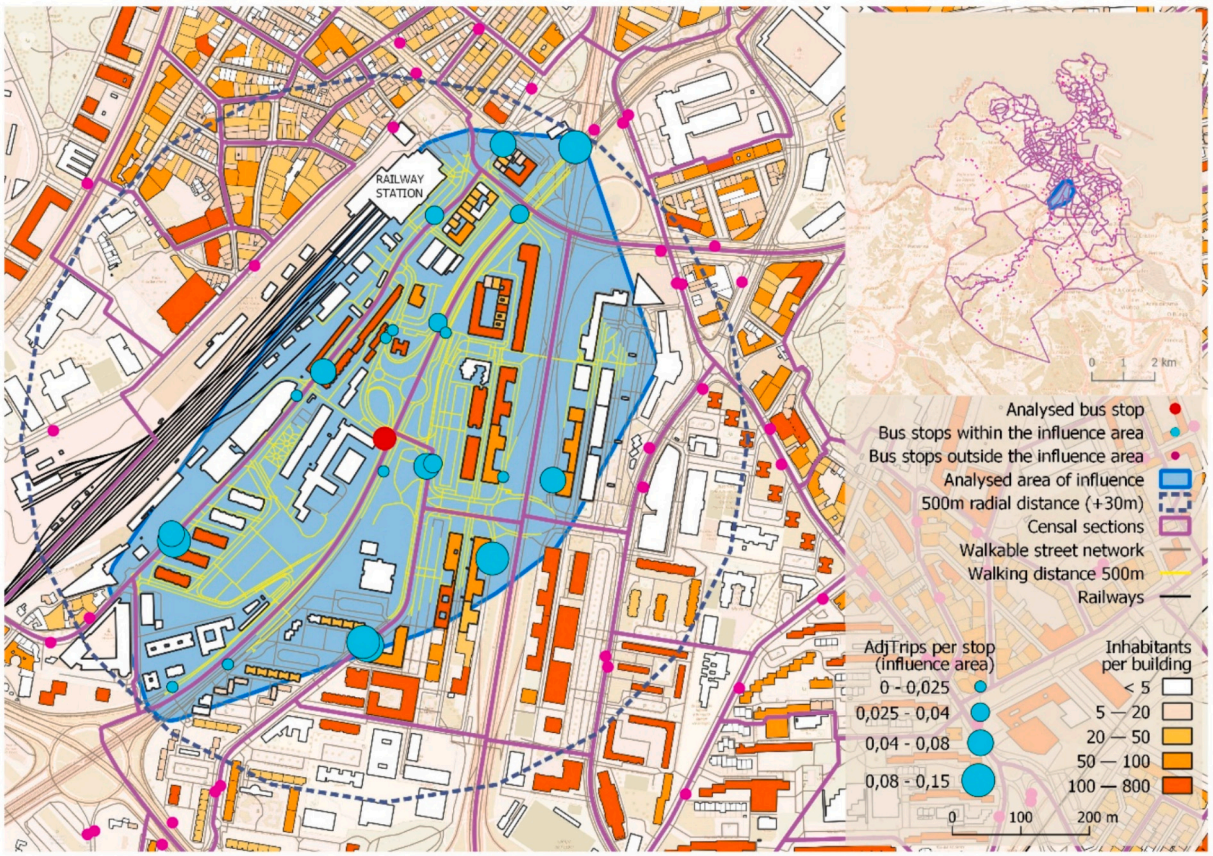

Fig. 3. Area of influence of a bus stop considering a walking distance of 500 m.

weighted average value as explained in the previous section.

In addition, we need to characterise the public transport supply in the area of influence. To achieve this, it is essential to consider the total number of bus trips from all the bus stops in each area of influence. It is important to note that the number of boardings along a bus line varies significantly depending on the location of the stop along the line, with alightings prevailing over boardings at the final stops. Chakrabarti (2015) considered the stop location along the lines by introducing a categorical variable labelled as "stop near line end". This variable was determined by whether the stop sequence number was within the upper 25% of all stops served by the line.

In our model, to account for this boardings variation along lines, a weighting coefficient is applied to each bus trip, considering the position of the stop along the bus line(s) serving it (Eq. 1). For example, the outbound bus line 7, with 18 stops, would have a coefficient of 1 for the initial stop, 0.94 for the next stop, 0.88 for the following one, and so on, until the last two stops with coefficients of 0.06 and 0 for the end-of-line stops. Once the weighted coefficients are calculated, they are multiplied by the bus trips of each stop to obtain the adjusted bus trips for that stop. In cases where multiple bus lines serve a single stop, the bus trips for each line are multiplied by their respective coefficients. The adjusted final bus trips at that stop are then obtained by aggregating all the adjusted bus trips from each line (Eq. 2).

$$C(l,i) = \frac{n_l - i}{n_l - 1}; i = 1, \ldots, n_l \quad (1)$$

$$AT(i) = \sum_{l=1}^{N_i} T(l,i)^{*}C(l,i) \quad (2)$$

where:

l = bus line; i = analysed bus stop; n_l = total number of bus stops for bus line l; N_i = total number of bus lines serving stop i; $C(l,i)$ = position-specific coefficient for bus line l at bus stop i; $T(l,i)$ = annual bus trips for bus line l at bus stop i; $AT(i)$ = total adjusted bus trips at bus stop i, considering all bus lines serving the stop.

Finally, following the same procedure used for the other variables in QGIS, the total number of adjusted bus trips within the area of influence of a stop is calculated by aggregating $AT(i)$ from all the bus stops in that area ($AT\_area(i)$). This total value is then transferred to the corresponding bus stop, providing us with the data for the adjusted bus trips at that particular stop and the total adjusted bus trips within its area of influence.

4.3. Variable processing

Once all urban bus stops have the information regarding the characteristics of their area of influence, it is necessary to adjust the variables before applying the models to be able to compare the results and the impact of each variable on bus demand. Fig. 4 presents an overview of the variables that will finally be analysed in Section 6.1.

Our goal is to estimate bus demand at the stop level. In order for these values to be comparable, the final dependent variable is considered as the annual boardings per stop divided by the size of its area of influence (boardings/m^2). Regarding the predictors, land-use variables are calculated as the ratio of the area of influence assigned to each category. Socioeconomic variables that are not related to income or indexes are considered as the percentage of the total population in the area meeting a criterion, ranging from 0 to 1. The total inhabitants will be divided by the size of the area of influence (inh/m^2).

Regarding the transit supply variable, in the previous section it has been explained how to obtain the adjusted bus trips per stop (Eq. 2) and the total within its area of influence (AT_area). The number of stops and trips within an area of influence is relevant as it reflects the competition

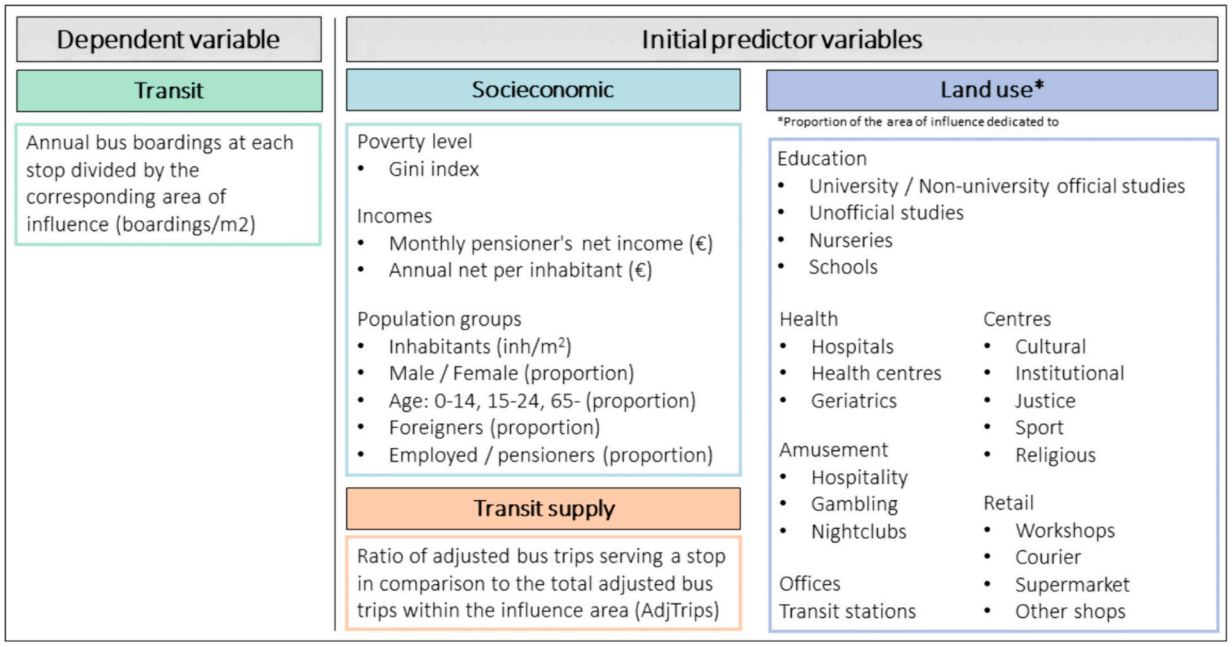

Fig. 4. Summary of variables initially analysed in the statistical models.

among transit stops and lines, which is typically more pronounced in areas closer to the city centre. For this reason, some studies have considered the number of stops within the buffer of each stop as a predictor (Chu, 2004; Chakour and Eluru, 2016). The competition effect between nearby stops was also studied by Cui et al. (2022), differentiating whether these stops serve different or same routes and analysing their destination and accessibility overlaps.

To account for this competition between stops, the lines serving them, and their trips, we calculate the transit supply variable for each stop by dividing its adjusted bus trips by the total adjusted bus trips within its area of influence (Eq. 3). If there is only one stop, this variable will have the same value as the adjusted bus trips for that stop. Therefore, we calculate the transit supply variable as the proportion of adjusted bus trips serving a stop in relation to the total adjusted bus trips in the area of influence (Eq. 3). Hereafter, we will refer to this variable as adjusted trips (or AdjTrips, represented in Fig. 1 across the city).

$$AdjTrips(i) = \frac{AT(i)}{AT\_area(i)} \; \epsilon \; (0,1] \tag{3}$$

where:

$AT\_area(i)$= sum of the adjusted bus trips for all the stops within the area of influence of stop i; $AdjTrips(i)$ = ratio of adjusted bus trips serving the stop i relative to the total adjusted bus trips within its area of influence.

The proposed approach, which combines frequency with the number of other stops in the area instead of treating them as isolated variables, enables considering together the effect of the number of nearby stops and the actual supply offered at each of them. The frequency of each line is heavily determined by its demand, and for that reason, the number of services in the analysed stop has not been considered as an explanatory variable on its own.

5. Statistical model with spatial dependence

The purpose of this study is to develop a model that allows to estimate the demand for a bus stop, considering the city's evolving demographics and land-use characteristics and distribution. To achieve this goal, we explore sophisticated techniques beyond traditional linear models, as we observed during the analysis non-linear effects of predictors as well as spatial dependence. This section presents a detailed step-by-step explanation and description of the model, providing insights into its application and performance.

During the modelling process, a wide variety of models were fitted to the data, including MLR, Partial Least Squares Regression (PLSR), Projection Pursuit Regression (PPR), Multivariate Adaptive Regression Splines (MARS) and Random Forest (RF) (see e.g. James et al., 2013). The best results were obtained with a Generalized Additive Model (GAM; see e.g. Hastie and Tibshirani, 1990; Wood, 2017). This type of model allows for both linear and non-linear effects of the predictors, with the added advantage of being easily interpretable. Specifically, we will consider that the i-th observation (Y_i, X_i), for $i = 1,\ldots,n$, is modelled by an expression of the form:

$$Y_i = \beta_0 + f_1(X_{1i}) + f_2(X_{2i}) + \cdots + f_p(X_{p_i}) + \varepsilon_i$$

being $f_j, j = 1,\ldots,p$, arbitrary smooth functions. It is important to highlight that the linear model would be a particular case. In particular, if $f_j(x) = \beta_j x$ the effect of the variable X_j would be linear.

Furthermore, in the analysis of urban public transport stop data, observations may exhibit spatial correlation. Such correlation, if not considered, can lead to poor performance in classical statistical inference models (Cressie, 1993, Section 1.3). In general, data points that are close in space are correlated, tending to have similar values, and the correlation decreases as the separation between them increases (known as the first law of geography; Tobler, 1970). Therefore, we will also assume that there is potential spatial dependence. Specifically, if we denote by $s_i = (s_{1i}, s_{2i})$ the coordinates of the i-th observation, we will assume that ε is a stationary spatial process with an exponential covariance function:

$$Cov(\varepsilon_i, \varepsilon_j) = C_\theta(s_i - s_j)$$

being

$$C_\theta(h) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ \sigma^2(1 - c_0)exp\left(-\frac{3\|h\|}{a}\right) & \text{if } h \neq 0 \end{cases}$$

Where $C_\theta(0) = \sigma^2$ is the variance, c_0 is the proportion of nugget effect ($\sigma^2 c_0$ is the uncorrelated variability) and a is the correlation range

(Cressie, 1993, Section 2.3).

Therefore, the considered model is a Generalized Additive Model (GAM) with spatially correlated errors, denoted as SGAM. The fitting of these types of models in practice can be performed using the gamm() function from the mgcv package (Wood, 2023) within the statistical environment (R Core Team, 2023). Generalized Additive Mixed Models (GAMM) represent an extension of GAM that, in addition to allowing for non-linear effects of predictors, enhance their capabilities by enabling the incorporation of random effects. Through these random effects, spatial dependence can be modelled (SGAM could be considered as a particular case of GAMM, although we prefer to distinguish between both models).

In these following paragraphs, we describe the steps to follow the methodology proposed in this paper. At first, an initial univariate descriptive analysis and a subsequent data cleaning are performed. During the year, there are often provisional, seasonal, or irregular bus stops, as well as some affected by works, which should be removed before modelling. Ensuring the dependent variable's normal distribution often requires a logarithmic transformation due to the typical skewness in transit data.

In the subsequent step, a multivariate descriptive analysis is conducted, focusing on filtering predictors by examining linear and non-linear relationships through scatter plots and Pearson correlation coefficients. Variables with high correlations (>0.7) are scrutinized. The most highly correlated variable with the response is incorporated into the model, while the others are discarded. This step avoids potential collinearity problems (see comments in Section 6.3), as well as potentially reducing the computational time in the next step.

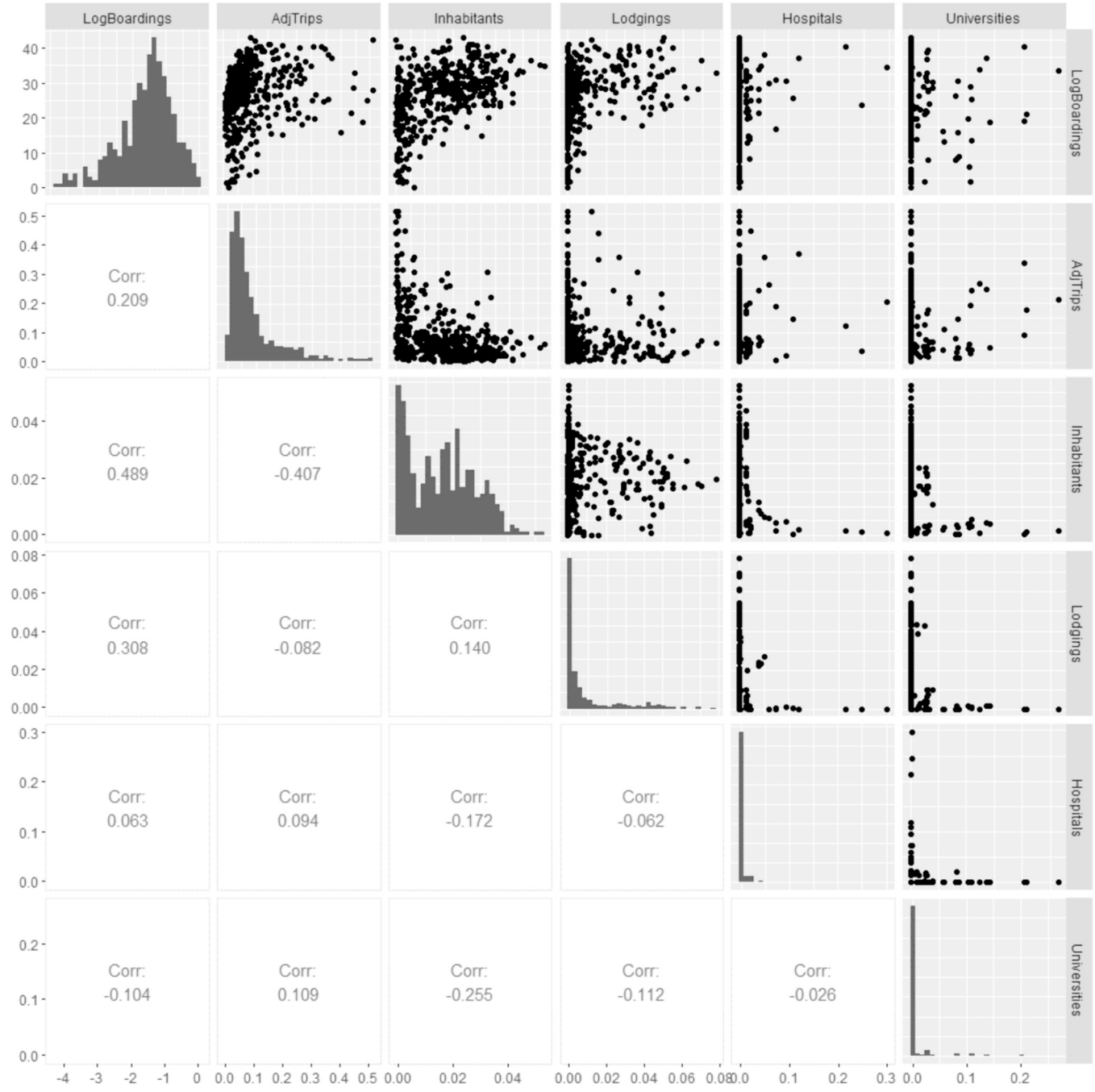

Fig. 5. Matrix plot with histograms (diagonal), scatter plots (upper diagonal) and correlations (lower diagonal), for the response and the final predictors.

Final predictor selection uses a backward stepwise method. Each step evaluates whether there is no significant loss when removing each of the included predictors. If any predictor is identified for elimination (otherwise, the process is stopped), the one that results in the least non-significant reduction in the proportion of explained variability of the response is eliminated, and the model is fitted again. Once a final model is reached, number of degrees of freedom and partial effect plots are examined to identify the variables exhibiting linear relationships. The fitted model is then evaluated through leave-one-out cross-validation, identifying influential data points by comparing predictions with actual values. Influential observations are removed, and the model is refitted to ensure accuracy. This diagnostic process and model validation are detailed further in Section 6.3.

6. Results and discussion

6.1. Exploratory analysis of variables

In this section, we present the outcomes of the predictor selection process, following the methodology described in Section 5. Under-serviced bus stops were removed, including temporary, seasonal, and line-extension stops, as well as those serving only a few trips per day, alongside company bus depots, totalling 72 (14%). Furthermore, stops with fewer than 20 boardings per year were excluded (2%). The annual boardings variable (boardings/m^2) exhibits a positive skew, requiring a logarithmic transformation (refer to the top left plot in Fig. 5 for the distribution of the transformed variable). It will be shown in the model validation Section 6.3 that this transformation renders the residuals' distribution of the final model approximately symmetric.

Initially, 40 predictors were considered, but the number was reduced from 40 to 28 after removing those with high correlation, as illustrated in Fig. 5. Due to space constraints, only the predictors used in the final model are displayed. The final selection of predictors was made using a backward stepwise method, and after identifying two influential outliers and repeating the process, the final fitted model included 430 observations.

The predictors included in the fitted model are detailed in Table 1, including descriptive statistics, also for the response. It could be noted that the mean and median of the variables Hospitals, Lodgings, and Universities are close to zero, indicating that these land uses are non-existent in most of the areas of influence of the stops. Fig. 5 shows a matrix plot with the variable distributions on the diagonal and scatter plots and correlations off the diagonal. In this figure, the distributions reflect the large number of zeros of some predictors, as well as the symmetry in the distribution of the dependent variable. In the first row, outside the diagonal, the relationships of the predictors of the model with the response are shown and reflect the presence of apparently non-linear relationships with Inhabitants, Lodgings and AdjTrips. No strong correlation between predictors is observed in this figure. This will be discussed in more detail in Section 6.3.

6.2. SGAM results

The final model obtained was as follows:

$$\begin{aligned} log_{10}(Boardings) = & -1.664 + f_1(AdjTrips) + f_2(Inhabitants) \\ & + f_3(Lodgings) + 3.24 Universities + 5.93 Hospitals + \varepsilon \end{aligned} \quad (4)$$

The variables Universities and Hospitals have a linear effect, and the variables AdjTrips, Inhabitants, and Lodgings have a non-linear effect, shown in Fig. 6, all the predictors have a positive effect.

An increment of 0.1 in the Universities ratio is expected to increase LogBoardings by 0.324, which corresponds to a multiplicative increase of 2.11 in annual bus boardings/m^2. Following the same reasoning to the ratio of hospitals, it would be expected that if this ratio is increased by 0.1, LogBoardings would be multiplied by 3.92. What we observed in our results are in line with those obtained in Frei and Mahmassani (2013). The results of their predictive model also highlight educational and medical land use. For the two time periods analysed in their study, medical land use has a positive impact with the natural logarithm of bus boardings. However, the coefficient of educational land use is positive in October and negative in May, which could be attributed to the fact that some Chicago schools conclude their academic year during that month.

For the variable AdjTrips (top-left of Fig. 6), a pronounced and almost linear increasing effect is shown at lower values. In this range, even slight changes exert a relevant increase of LogBoardings. However, beyond a certain point (around 0.15 AdjTrips), the effect becomes flatter. Note that there are fewer observations in that interval and the estimate of the predictor's effect has higher variability.

Regarding the variable Inhabitants, a similar behaviour is observed (top-right of Fig. 6). Until a value of 0.02, a large positive linear effect is noticeable, and beyond that value, the effect appears to be much flatter. These two variables have the most significant effect on the log of annual bus boardings. This aligns with existing literature that establishes population density as the leading built-environment factor for forecasting transit ridership (Cervero et al., 2004). In the study performed by Johnson (2003), population density was the most significant socioeconomic variable for the prediction of weekday bus boardings. Cui et al. (2022) also found a significant association between larger populations and higher bus boardings at public transit stops, up to a certain threshold associated with dense, highly walkable areas. However, distinct from our study, their analysis did not incorporate spatial dependence, and dis not flexibly estimate non-linearity for all the predictors but rather posits it a priori as squared or logarithmic for some variables.

Lastly, in the bottom-left graph of Fig. 6, the effect of the variable Lodgings ratio is represented. A nearly quadratic effect is observed, although it is very close to zero. This aligns with the finding that this variable proved to be one of the least useful in explaining the response.

The dependence parameters estimates were $\widehat{\sigma}^2 = 0.232$, $\widehat{c_0} = 0.714$, and $\widehat{a} = 1153.95$. These values suggest that approximately 71% of the variability is independent, while the remaining 29% is attributed to spatial dependence. The practical range is approximately 1.15 km. This suggests that we can assume dependence between stops separated by less than this distance.

6.3. Model diagnosis and validation

To validate the fitted model, several checks have been performed.

Table 1
Descriptive statistics of the response and the final predictors.

Variable	Description	Min	Mean	Median	Max	SD
LogBoardings	Logarithm of annual boardings ratio ($log_{10}(boardings/m^2)$)	−4.271	−1.584	−1.441	0.056	0.825
AdjTrips	Adjusted trips ratio	0.005	0.089	0.057	0.510	0.090
Inhabitants	Total inhabitants ratio (inh/m^2)	0.000	0.016	0.016	0.052	0.012
Lodgings	Lodgings ratio (m^2/m^2)	0.000	0.008	0.001	0.077	0.015
Hospitals	Hospitals ratio (m^2/m^2)	0.000	0.005	0.000	0.297	0.024
Universities	Universities ratio (m^2/m^2)	0.000	0.008	0.000	0.265	0.029

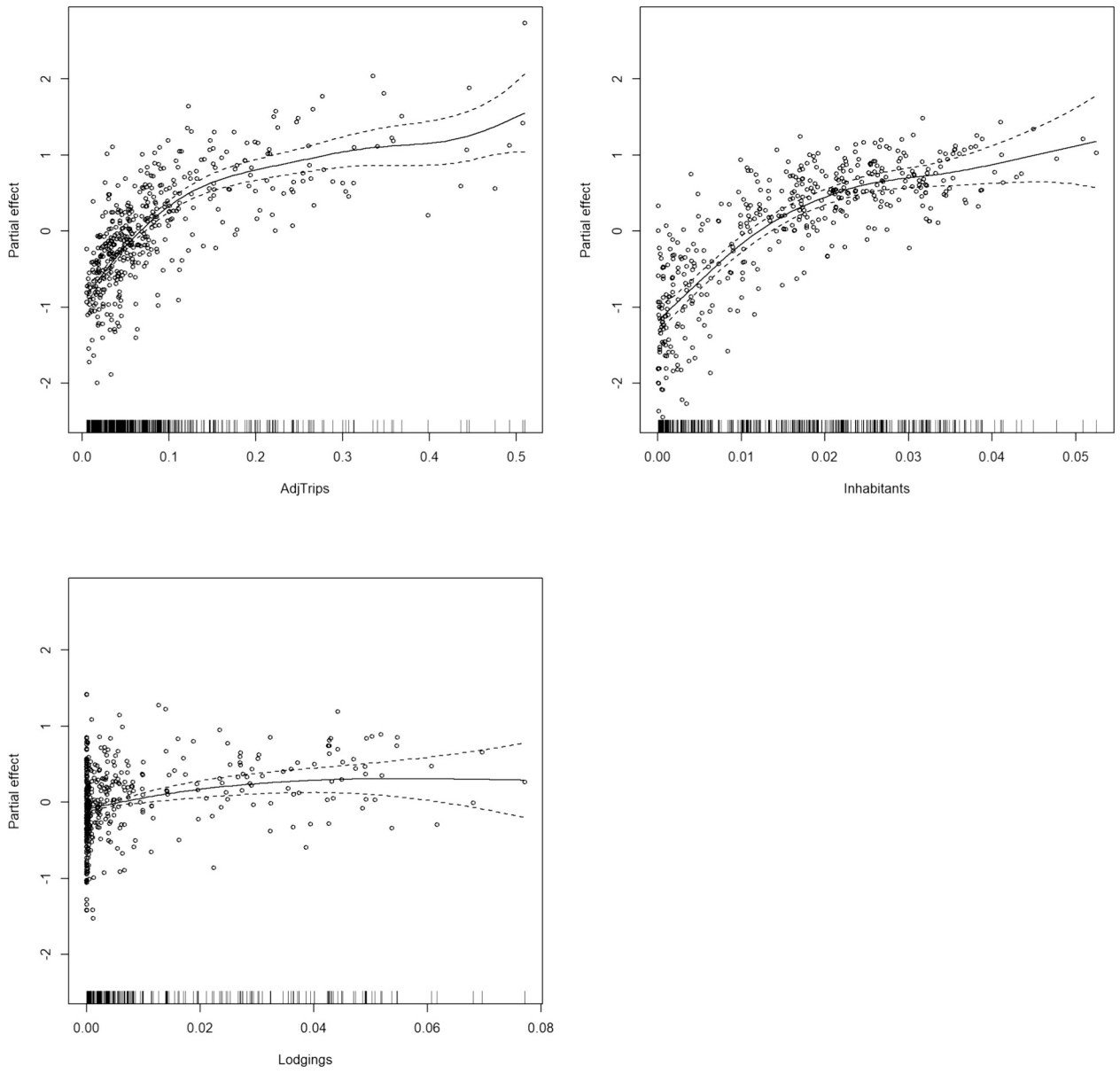

Fig. 6. Estimates of non-linear partial effects (solid line), confidence intervals (dashed lines; 2 standard errors above and below the smooth estimate), and corresponding partial residuals.

First, we diagnosed whether there were issues of collinearity. As can be seen in Fig. 5, the linear correlation between predictors is not very strong, with −0.41 being the highest value according to the Pearson's correlation coefficient.

In the case of non-linear effects, the generalization of collinearity is called concurvity. Concurvity occurs when one non-linear component can be approximated by combining one or more of the other terms in the model. The concurvity() function from the mgcv package allows the computation of concurvity measures, which can be interpreted as the proportion of variability in a component that can be explained by the rest (Wood, 2008). Its values are bounded between 0 and 1, with 0 indicating no problem and 1 indicating a total lack of discernibility. The maximum estimated value obtained was 0.23, corresponding to component $f_2(\textit{Inhabitants})$. Therefore, we consider that there are no concurvity issues.

The remaining structural hypotheses of the model were examined based on the residuals, and no issues were detected. For instance, the histogram of the residuals is displayed on the left side of Fig. 7, where it can be observed that residuals' distribution is approximately symmetric and centred around zero. Thus, it would be reasonable to assume normality.

Furthermore, cross-validation residuals were computed using a leave-one-out approach and were employed to identify potential influential observations. These residuals versus predictions are plotted in the right side of Fig. 7. A similar graph is obtained when using standardised residuals. From this graph, it is inferred that there are apparently no influential data points. The plot exhibits a pattern that could be interpreted as a funnel shape, suggesting that there may be greater variability in low prediction values. However, this observation does not compromise the overall validity or the interpretative strength of our model's predictions. Instead, it highlights an area for further refinement and investigation in future research, where methods specifically designed to accommodate or correct for such variance patterns could be explored to enhance the accuracy of the model.

Cross-validation is a widely accepted technique for assessing the model's predictive capabilities. By using cross-validation residuals, we obtain a pseudo R-squared of 0.79, which can be interpreted as the proportion of variability in the response of new observations explained

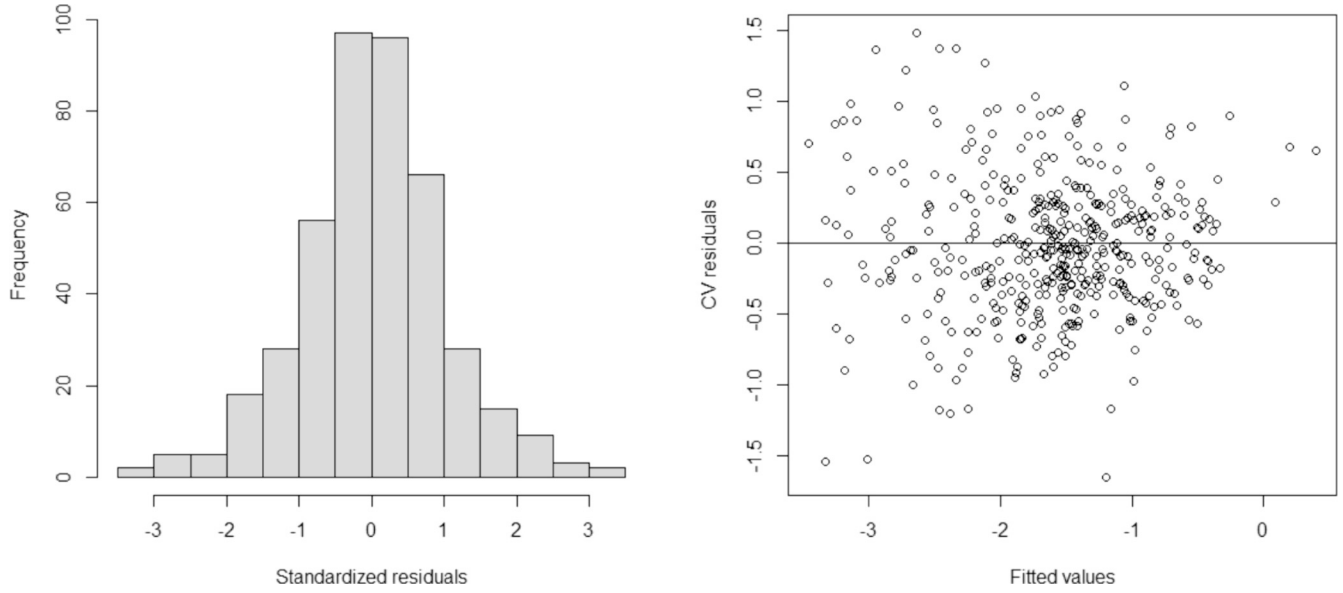

Fig. 7. Distribution of the standardised residuals (left) and scatterplot of the cross-validation residuals versus the predicted values (right).

by the model. This level of explanatory power is considered notably high and provides confidence in the model's ability to make accurate predictions. This contrasts with the 0.67 pseudo R-squared obtained from an MLR model, underscoring the relevance of incorporating both linear and non-linear relationships and spatial dependencies for enhanced prediction accuracy.

7. Conclusions

This study bridges interdisciplinary perspectives, considering the synergy between geography, urban transit planning, and statistics in addressing complex urban challenges. Our research has developed a refined model for predicting bus demand at stop level by analysing land-use characteristics, socioeconomic factors, and transit supply variables, all while accounting for spatial dependence. Notably, this approach provides a more direct alternative to predict bus demand compared to traditional methods, such as the four-stage approach, demanding fewer data that need to be collected specifically and calibration efforts. The adaptability of this model facilitates its transferability and implementation in other cities.

Capturing the reality of pedestrian movements in urban environments is intricate, demanding a detailed analysis. To address this complexity, we have developed a methodology that incorporates actual walking distances from bus stops using street networks, as opposed to simplistic straight-line radial distances. While straight-line radial distances yield similar results in some areas, the benefits of this approach become notably pronounced when examining bus stops located near physical barriers, such as long edifices, shopping centres, or transit stations. In such cases, the differences between both distance metrics are substantial, highlighting the practical importance of our method in these specific urban contexts. In addition to considering the street network from the bus stop, we have conducted detailed variable processing within QGIS to ensure that the areas surrounding the stops have the most accurate and realistic information.

The utilisation of a SGAM highlights the importance of considering linear and non-linear relationships as well as potential spatial dependence. The comparison of our model with a traditional MLR model reveals a notable improvement in predictive performance. Furthermore, our findings suggest that it is not reasonable to assume independence between stops situated less than 1.15 km apart. Thus, according to our case study, spatial dependence should be considered to accurately estimate the influence of variables within geographical proximity, aligning with the spatial nature of urban transit systems.

Out of the 40 initially considered explanatory variables, in our case study, the presence of dedicated areas for hospitals and universities near bus stops has a significant, linear, and positive effect on boardings, with hospitals exerting the greatest influence. This may be due to the fact that both types of land uses attract users who are more likely to use public transit, such as elder individuals and university students. The higher presence of inhabitants/m^2 near the bus stop and bus transit supply has a positive and non-linear effect on the number of boardings, which is logically consistent. The presence of lodgings near the bus stop also exhibits a positive and non-linear effect, possibly because they are in proximity to economic activity areas where tourists tend to stay and do not commonly use private vehicles.

This research serves as a powerful tool for predicting the impacts of changes in bus stop locations, route designs, or urban developments (socioeconomic and land uses) on bus transit planning. This tool provides information to optimize the positioning of new bus stops based on estimated demand. The validation process supports the reliability of this model for bus-demand prediction in urban areas. By providing predictions of potential passenger boardings at each bus stop, this model supplies decision-makers with valuable knowledge to contribute and support sustainable urban mobility solutions. While this research brings valuable insights, future research could conduct sensitivity analyses on area of influence length.

Despite the strengths of our research, certain limitations should be considered. As there are no available data on the number of inhabitants per building, we estimate it using census section information (smallest geographic units with socioeconomic data) and the number of dwellings per building. Although this estimation is an improvement on existing studies, this approach does not account for empty dwelling or variations in household size. Furthermore, some buildings span multiple census sections. Given the absence of detailed data regarding the internal distribution of dwellings within buildings, both in terms of layout and vertical positioning, and the absence of unit division by portals for buildings with multiple entrances, we chose to use a single point as the building's representation (the centroid) and assign it the total number of dwellings. It would be interesting to have the entrance door data of each building for future work to make the analysis more accurate, instead of using the centroid of the buildings' geometry. If any city had this data, the 30 m buffer stablished could be adapted to that case. The 30 m buffer is on the side of caution to include stops and centroids but can be adapted for cities with a different configuration of roadways, sidewalks,

stop placements, etc., if experts deemed it appropriate. On the other hand, information about jobs location is not available in our case, but it would be interesting to explore the possibility of considering it for other case studies. Finally, other ways of considering competition among stops could be explored, refining the AdjTrips variable definition.

Our study underscores the importance of considering spatial dependence, non-linear relationships, and advanced statistical techniques in estimating bus demand. By contributing to a deeper understanding of the complex interplay between spatial factors, stop surroundings' characteristics and transit bus demand, we aim to foster more informed decisions in urban transit planning. This endeavour ultimately contributes to the enhancement of urban well-being and sustainability.

CRediT authorship contribution statement

Yaiza Montero-Lamas: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rubén Fernández-Casal:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Francisco-Alberto Varela-García:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alfonso Orro:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Margarita Novales:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

None.

Data availability

The authors do not have permission to share bus demand data. The remaining data used is open data.

Acknowledgments

The authors would like to thank Compañía de Tranvías de La Coruña and Concello da Coruña for providing the data required to prepare this paper. This work was funded by grants PID2021-128255OB-I00 and PRE2019-089651, funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU and ESF/EU.

The research of Rubén Fernández-Casal has also been supported by Grant PID2020-113578RB-I00, funded by MCIN/AEI/10.13039/501100011033, by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020/14) and by CITIC that is supported by Xunta de Galicia, convenio de colaboración entre la Consellería de Cultura, Educación, Formación Profesional e Universidades y las universidades gallegas para el refuerzo de los centros de investigación del Sistema Universitario de Galicia (CIGUS).

Funding for open access charge: Universidade da Coruña/CISUG.

References

Amoroso, S., Migliore, M., Catalano, M., Galatioto, F., 2010. A demand-based methodology for planning the bus network of a small or medium town. Eur. Transp. 44, 41–56.

Cervero, R., 2006. Alternative approaches to modeling the travel-demand impacts of smart growth. J. Am. Plan. Assoc. 72 (3), 285–295. https://doi.org/10.1080/01944360608976751.

Cervero, R., Murphy, S., Ferrell, C., Goguts, N., Tsai, Y.-H., Arrington, G.B., Boroski, J., Smith-Heimer, J., Golem, R., Peninger, P., Nakajima, E., Chui, E., Dunphy, R., Myers, M., McKay, S., 2004. Transit-oriented development in the United States: experiences, challenges, and prospects. In: Transit Cooperative Research Program (TCRP) Report 102, published by the Transportation Research Board, Washington. https://www.worldtransitresearch.info/research/3066.

Cervero, R., Murakami, J., Miller, M., 2010. Direct ridership model of bus rapid transit in Los Angeles County, California. Transp. Res. Rec. 2145 (1), 1–7. https://doi.org/10.3141/2145-01.

Chakour, V., Eluru, N., 2013. Examining the influence of urban form and land use on bus ridership in Montreal. Procedia Soc. Behav. Sci. 104, 875–884. https://doi.org/10.1016/J.SBSPRO.2013.11.182.

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.

Chakrabarti, S., 2015. The demand for reliable transit service: new evidence using stop level data from the Los Angeles metro bus system. J. Transp. Geogr. 48, 154–164. https://doi.org/10.1016/J.JTRANGEO.2015.09.006.

Chu, X., 2004. Ridership Models at the Stop Level. University of South Florida, National Center for Transit Research, University of Florida. https://rosap.ntl.bts.gov/view/dot/64250.

Cressie, N.A.C., 1993. Statistics for Spatial Data, 1st ed. John Wiley & Sons. https://doi.org/10.1002/9781119115151.

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 (5), 769–783. https://doi.org/10.1007/s11222-013-9401-9.

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., Surprenant-Legault, J., 2014. New evidence on walking distances to transit stops: identifying redundancies and gaps using variable service areas. Transportation 41 (1), 193–210. https://doi.org/10.1007/s11116-013-9508-z.

Frei, C., Mahmassani, H.S., 2013. Riding more frequently: estimating disaggregate ridership elasticity for a large urban bus transit network. Transp. Res. Rec. 2350 (1), 65–71. https://doi.org/10.3141/2350-08.

George, K.A., 1999. Transportation compatible land uses and bus-stop location. WIT Trans. Built. Environ. 44, 459–468.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models, 1st ed. Routledge. https://doi.org/10.1201/9780203753781.

Hu, S., Chen, P., 2021. Who left riding transit? Examining socioeconomic disparities in the impact of COVID-19 on ridership. Transp. Res. Part D Transp. Environ. 90, 102654 https://doi.org/10.1016/j.trd.2020.102654.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer, New York.

Johnson, A., 2003. Bus transit and land use: illuminating the interaction. J. Public Transp. 6 (4), 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 (1), 23–32. https://doi.org/10.3141/2537-03.

Marques, S. de F., Pitombo, C.S., 2021a. Ridership estimation along bus transit lines based on kriging: comparative analysis between network and Euclidean distances. J. Geovis. Spat. Anal. 5 (1), 7. https://doi.org/10.1007/s41651-021-00075-w.

Marques, S. de F., Pitombo, C.S., 2021b. APPLYING MULTIVARIATE GEOSTATISTICS FOR TRANSIT RIDERSHIP MODELING AT THE BUS STOP LEVEL. Bol. Ciências Geodésicas 27 (2). https://doi.org/10.1590/1982-2170-2020-0069.

Marques, S. de F., Pitombo, C.S., 2023a. Local modeling as a solution to the lack of stop-level ridership data. J. Transp. Geogr. 112, 103682 https://doi.org/10.1016/j.jtrangeo.2023.103682.

Marques, S. de F., Pitombo, C.S., 2023b. Transit ridership modeling at the bus stop level: comparison of approaches focusing on count and spatially dependent data. Appl. Spat. Anal. Policy 16, 277–313. https://doi.org/10.1007/s12061-022-09482-y.

Montero-Lamas, Y., Orro, A., Novales, M., Varela-García, F.-A., 2022. Analysis of the relationship between the characteristics of the areas of influence of bus stops and the decrease in ridership during COVID-19 lockdowns. Sustainability 14 (7), 4248. https://doi.org/10.3390/SU14074248.

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.

Noh, N.M., Mohamad, D., Hamid, A.H.A., 2021. Acceptable walking distance accessible to the nearest bus stop considering the service coverage. Int. Congress Adv. Technol. Eng. (ICOTEN 2021) 1–7. https://doi.org/10.1109/ICOTEN52080.2021.9493435.

Ortúzar, J. de D., Willumsen, L.G., 2011. Modelling transport. In: Modelling Transport, 4th edition. John Wiley and Sons. https://doi.org/10.1002/9781119993308.

Pulugurtha, S.S., Agurla, M., 2012. Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. J. Public Transp. 15 (1), 33–52. https://doi.org/10.5038/2375-0901.15.1.3.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/.

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.F., 2009. Pedestrian environments and transit ridership. J. Public Transp. 12 (1), 39–57. https://doi.org/10.5038/2375-0901.12.1.3.

Talavera-Garcia, R., Valenzuela-Montes, L.M., 2018. Análisis conceptual de la distancia peatonal al transporte público: hacia un enfoque más integrador. Archit. City Environ. 13 (37), 183–204. https://doi.org/10.5821/ACE.13.37.5134.

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.

Wood, S.N., 2008. Fast stable direct fitting and smoothness selection for generalized additive models. J. R. Stat. Soc. Ser. B Stat Methodol. 70 (3), 495–518. https://doi.org/10.1111/J.1467-9868.2007.00646.X.

Wood, S.N., 2017. Generalized Additive Models: An Introduction with R, 2nd ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279.

Wood, S.N., 2023. mgcv: mixed GAM computation vehicle with automatic smoothness estimation. R package version 1.9–0. https://cran.r-project.org/package=mgcv.