DOI 10.1590/s1982-21702025000100001

*ORIGINAL ARTICLE*

# Analysis of spatial effects for a mode choice problem: a test proposal for spatial variability evaluation

Pedro Henrique Caldeira Caliari[1] - ORCID: 0000-0002-0585-529X

Cira Souza Pitombo[1] - ORCID: 0000-0001-9864-3175

[1] São Carlos School of Engineering – University of São Paulo (EESC-USP), Department of Transport Engineering, São Carlos - SP, Brazil.

E-mail: pedro.caliari@usp.br, cirapitombo@usp.br

Received in 28th May 2024

Accepted in 09th December 2024

**Abstract:**

Spatial effects are intrinsic to studies on travel behavior, including modal choice problems. Among spatial dependence and heterogeneity, the resulting spatial variation of estimated coefficients is one of the most significant gains in local spatial models. Thus, this article aims to analyze a modal choice problem, considering spatial dependence and variability simultaneously. Additionally, a test is proposed to evaluate and validate the spatial variability, obtaining disaggregated results. A database adapted from the household origin-destination survey was used, which was carried out in 2007/2008 in the city of São Carlos – São Paulo, Brazil. The proposed spatial variability test uses the estimated parameters of the GWLR model (main database and 200 spatially randomized databases), compared to the confidence intervals of the coefficients in the non-spatial logit model for the spatial variability hypothesis. The results of the proposed test are similar to the reference test in the case study. The disaggregated results can be used to verify if there are certain subgroups that are more likely to be spatially stationary and if these groups exhibit a spatial pattern. Moreover, it can be observed that the local spatial model provides a better fit and estimates when compared to the non-spatial model.

**Keywords**: Spatial dependence; Spatial heterogeneity; Geographically weighted logistic regression model; Modal choice.

**How to cite this article:** CALIARI PHC, PITOMBO CS. Analysis of spatial effects for a mode choice problem: a test proposal for spatial variability evaluation. *Bulletin of Geodetic Sciences*. 31: e2025001, 2025.

# 1. Introduction

Traditionally, modal choice problems are analyzed using discrete choice models, based on random utility functions (Ben-Akiva and Lerman, 1985). However, these models are typically unable to handle the spatial dimension, either because they do not explicitly incorporate such data, or because spatial effects tend to violate assumptions in non-spatial models (Anselin, 1988; LeSage, 1999). It is also known that variables associated with travel behavior have intrinsic spatial effects (Paez, 2006; Ibeas et al., 2011; Yang et al., 2020), as well as the modal choice (Lindner and Pitombo, 2018; Nkeki and Asikhia, 2019; Tao et al., 2019; Assirati and Pitombo, 2021; Rajamani et al., 2023).

Spatial effects have been addressed by spatial econometrics, which defines them as spatial dependence and heterogeneity, which can be present simultaneously (Anselin, 1988). Spatial dependence can be defined as the multidirectional interaction between observations (Anselin, 1988; LeSage, 1999). In other words, in the presence of spatial dependence, a variable in a given observation is also influenced by the values of neighboring observations. Meanwhile, spatial heterogeneity can be understood as a structural instability in the response variables (Anselin, 1988). The presence of spatial heterogeneity implies that the real parameter variable is a surface of values that vary regionally. Despite the greater complexity of obtaining local results, they enable a more diverse and credible analysis that is not available with single estimators (global or stationary). In this study, the term spatial variability is used as a synonym for spatial heterogeneity. Considering that modal choice is a phenomenon with the intrinsic presence of spatial effects (Rajamani et al., 2003; Lindner and Pitombo, 2018; Nkeki and Asikhia, 2019; Tao et al., 2019; Assirati and Pitombo, 2021; Mondal and Bhat, 2022), this article explores local spatial models and spatial effect tests, testing spatial dependence and variability.

Testing the hypothesis of spatial variability of a variable is an important step that must be taken to properly analyze the results of local models. Some authors have proposed different tests throughout the literature related to the spatial analysis of variables (Brunsdon et al., 1996; Fotheringham et al., 2002; Nakaya et al., 2014; Oshan et al., 2019). In Nakaya et al. (2014) it is suggested comparison tests among the metrics of the GWR models and non-spatial models. They are the difference between the corrected Akaike Information Criterion (AICc), which is -2 or less, the difference between the pseudo $R^2$ (Nakaya et al., 2014) and the difference in degrees of freedom. However, interpreting these differences is either arbitrary or questionable because it does not clearly consider the scale of the estimates.

Another test, proposed by Brunsdon et al. (1996) and Fotheringham et al. (2002), is implemented in the Multiscale Geographically Weighted Regression (MGWR) software (Oshan et al., 2019). After executing the original Geographically Weighted Regression (GWR) model, permutations are made to the coordinates, creating spatially random databases. Once this has been done, the GWR model is applied to each of the simulations. In addition to permuting the coordinates, given the expectation of spatial randomness (Fotheringham et al., 2002), each simulation receives a new bandwidth. After that, the standard deviation of the reference model is compared to the standard deviations of the simulated models. The pseudo p-value of the test is the position of the standard deviation of the original model in this rank. For example, if a model variable has the third smallest standard deviation out of a thousand simulations, its pseudo p-value is 0.003. This test was selected as a reference to corroborate the hypothesis of spatial variability, associated with the proposed test. The test proposed by Leung et al. (2000) was also found and later adapted by Fotheringham et al. (2002). Leung et al. (2000) performed an aggregate F test based on the model variance estimators to verify whether there is a gain in using local estimators instead of stationary estimators. Its biggest impediment for application in this study is the limitation to dependent variables with normally distributed residuals. Another important aspect is the lack of advantages compared to the Monte Carlo test proposed by Brunsdon et al. (1996) and Fotheringham et al. (2002). A less common tool for testing spatial variability is to compare the confidence interval of non-spatial model parameters with local model parameters (Fotheringham et al., 2002; Propastin, 2009).

The present article aims to analyze a modal choice problem simultaneously considering spatial dependence and variability. Moreover, a test is proposed to evaluate and validate the spatial variability, obtaining aggregated and disaggregated results. The proposed test includes simulations in confidence interval comparisons. A stationary coefficient model (non-spatial logit model in this study), a set of permuted local spatial models and a non-permuted local spatial model (reference model) were used. As an additional step, thematic maps and available local tests were analyzed to complement the proposed test, since aggregated results, even within spatial econometrics, may not accurately represent the actual local context (Anselin, 1988; Anselin, 1995; Getis and Ord, 1992).

Within the scope of the literature review of this article, most of the papers found that address spatial effects in modal choice presented the estimated coefficients as spatially stationary (Lindner and Pitombo, 2018; Assirati and Pitombo, 2021; Mondal and Bhat, 2022). Also, in Nkeki and Asikhia (2019), not all variables in the model had their spatial variability corroborated, but a new semi-parametric GWLR model was not presented. Properly considering spatially stationary variables as global could bring more realistic results (Nakaya et al., 2014). Thus, this significant gap can be observed, related to the more detailed analysis of spatial variability in modal choice problems.

The second contribution is the proposal of a spatial variability test. The proposed test provides an aggregate result for each variable, as in all tests found (Leung et al., 2000, Fotheringham et al., 2002, Nakaya et al., 2014). This study also presents disaggregated results that can be used for local spatial variability analyses. Thus, this paper aims to contribute to a methodological procedure that incorporates the analysis of spatial dependence and variability simultaneously, focusing on modal choice. Table 1 presents a summary of the main articles consulted, their topics, contributions, and relevant gaps. Additionally, we highlighted the gaps and contributions addressed in this article.

**Table 1:** Summary of studies consulted research gaps and gaps and contributions addressed in this article.

| Study | Main topics | Contributions | Gaps | Gap addressed in this study |
|---|---|---|---|---|
| Ben-Akiva and Lerman (1985) | Modeling modal choice by utility functions. | Pioneering work for modeling modal choice. | It does not consider the geographic aspect in the model structure. | Simultaneous analysis of spatial dependence and variability for modal choice. |
| Assirati and Pitombo (2021) | Spatial dependence model for modal choice. | Spatial dependence by ESDA and global spatial logit model. | The chosen model does not incorporate spatial variability. | |
| Nkeki and Asikhia (2019) | GWLR model for modal choice | Incorporates both spatial effects by a GWLR model. | Not all variables have spatial variability, yet a semi-parametric GWLR is not applied. | |
| Leung et al. (2000) | Developing statistical tests for GWR models. | Tests for GWR gain compared to non-spatial models. | Only aggregated test results. Limited to normally distributed dependent variables. | Proposal for testing spatial variability, with aggregated and disaggregated results. |
| Nakaya et al. (2014) | Semi-parametric GWR models and spatial variability test. | Simple and easy, non-parametric test. | Only aggregated test results. Arbitrary comparisons of model fit metrics. | |
| Fotheringham et al. (2002) | Fundamentals of GWR models. | Non-parametric test, not arbitrary, easy to understand. | Only aggregated test results. | |

The database in this paper refers to a region in the city of São Carlos – São Paulo, Brazil, from the household origin destination survey, carried out in 2007/2008 (Rodrigues da Silva, 2008). It contains basic information on modal choice, such as vehicle ownership, number of trips, age, whether the person works or studies (or both), gender and level of education.

This article is organized into four sections, including this introduction. Section 2 describes the data used and the methodological procedure adopted. Section 3 presents the main results and discussions. Finally, Section 4 draws the main conclusions considering the results, methodological limitations and suggestions for future research.

# 2. Materials and method

## 2.1 Dataset

The database used originated from the household origin-destination survey carried out in São Carlos, São Paulo, Brazil, in 2007/2008 (Rodrigues da Silva, 2008) and corresponds to a region in the city of São Carlos, with 86 valid observations. Figure 1 shows the location of the state of São Paulo in Brazil (a), the location of the city of São Carlos in São Paulo (b) and the study region, located in São Carlos city (c).

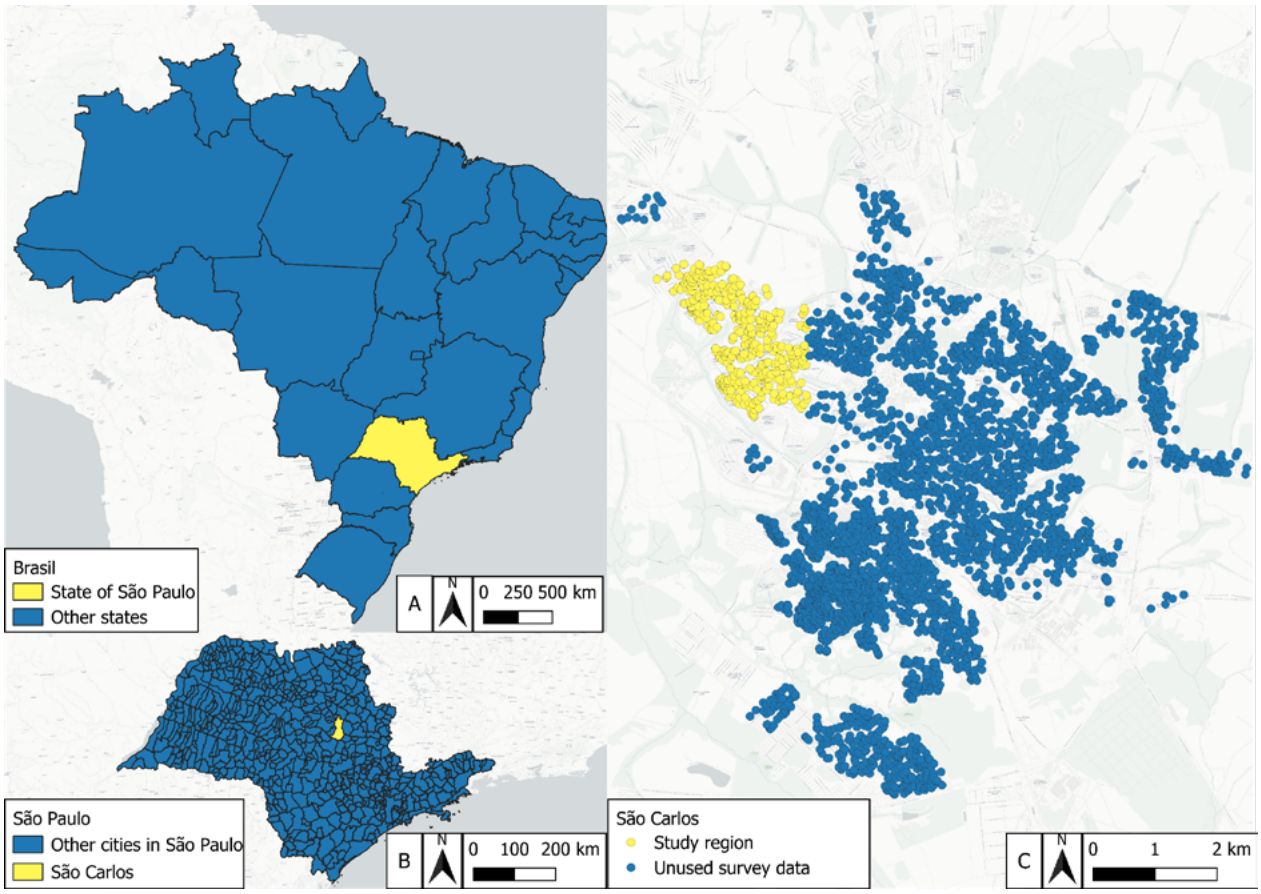


**Figure 1:** Original database and the region used (Rodrigues da Silva, 2008). Location of the state of São Paulo in Brazil (A), Location of the city of São Carlos in São Paulo (B) and the study region located in São Carlos city (C).

This region was chosen because previous studies corroborated the spatial dependence of the concerned variables (Pitombo et al., 2015; Pitombo et al., 2015b; Gomes et al., 2016). The household origin-destination survey used a Stratified random sampling of households regarding the population of the Traffic Analysis Zones (TAZs). The sample size was 6% of the city population. As the entire region did not have a spatial dependence, we chose a subset of the available data, following the procedure proposed by Pitombo et al. (2015b) and Costa (2013).

The number of valid observations available in the region, highlighted in yellow in Figure 1C, was not considered sufficient to calibrate the local spatial model. Thus, observations were created in random locations, close to the original ones. The values of the variables in the new observations were attributed by simple inverse distance weighting interpolation from the obtained surface to the new coordinates. Considering this change, the database, in this article, has 601 observations with four variables: modal choice (binary, see Figure 2), number of daily trips, age, and number of cars in the household. Table 2 describes the quantitative variables used. Table 3 shows the available binary variables. The dependent variable of modal choice has a value of 0 for public transport (frequency of 47.6%), and a value of 1 for car use (frequency of 52.4%).

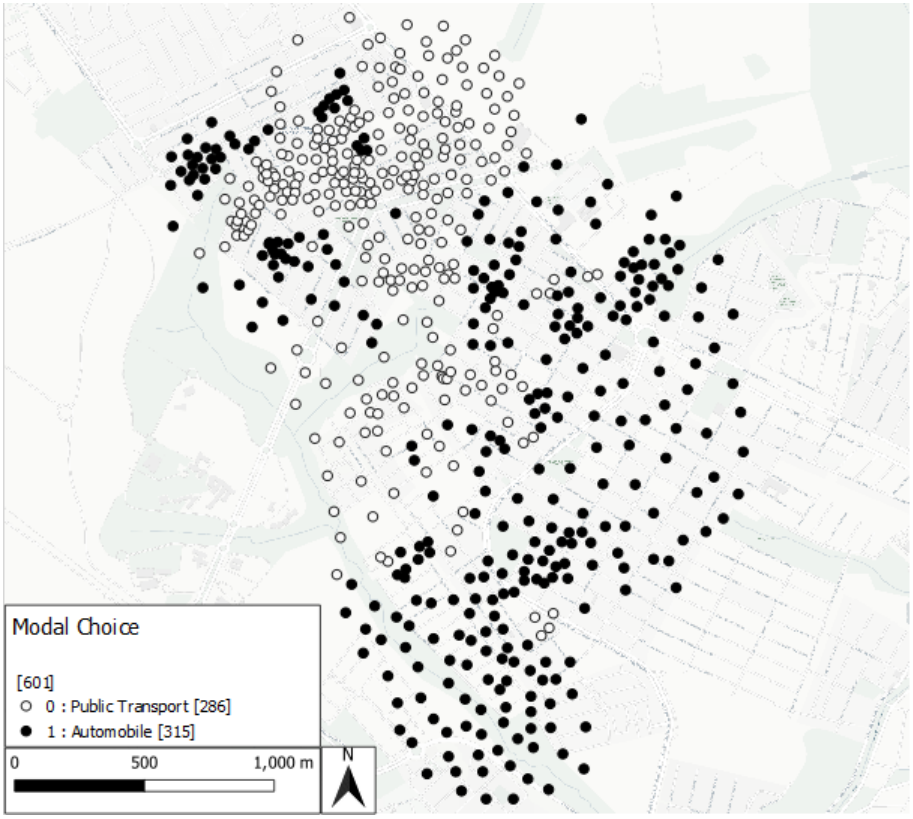


**Figure 2:** Spatial distribution of the dependent variable.

**Table 2:** Basic descriptive measures from the quantitative variable database (601 observations).

| Variable | Average | Median | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| Age | 40.6 | 41 | 6.70 | 20 | 78 |
| Motorcycle[1] | 0.115 | 0 | 0.329 | 0 | 2 |
| Trips[2] | 2.05 | 2 | 0.317 | 1 | 4 |
| Car[1] | 0.943 | 1 | 0.387 | 0 | 3 |

Motorcycle/Car[1] - quantity of vehicles per household; Daily trips[2] – daily trips per household.

**Table 3:** Percentages of observations of the qualitative variable categories (601 observations).

| Variable | Category frequency 0 | Category frequency 1 |
|---|---|---|
| Modal choice[1] | 0.476 | 0.524 |
| Driver's license[2] | 0.198 | 0.802 |
| Study[3] | 0.070 | 0.930 |
| Work[4] | 0.084 | 0.906 |
| Gender[5] | 0.210 | 0.790 |
| Level of education[6] | 0.494 | 0.506 |

Modal choice[1] - public transport (0), and car (1); Driver's license[2] - does not hold one (0), holds one (1); Study[3] - no (0), yes (1); Work[4] - no (0), yes (1). Gender[5] - male (0), female (1); Level of Education[6] - incomplete high school or less (0), complete high school, higher education and postgraduate studies (1).

Statistical tests were carried out to compare the original sample (86 observations) with the synthetic database, originating from the original (601 observations). Table 4 and 5 present the comparative tests of each available variable and the respective p-values. For numerical variables, the Wilcoxon-Mann-Whitney test (Hollander et al, 2013) was applied and for binary variables, the Chi-square test (Pearson, 1900) was used. According to Table 4, only the motorcycle ownership variable had a statistically different mean. However, this variable was not used in subsequent modeling. For binary variables, similarity was observed between almost all pairs of variables that comprise the original sample (86 observations) and the synthetic sample (601 observations). Only modal choice and work don't replicate the same distribution of the original dataset.

**Table 4:** Wilcoxon-Mann-Whitney test for comparisons between the numerical variables that comprise the original and synthetic samples.

| Variable | W[1] | P-value |
|---|---|---|
| Age | 25438 | 0.814 |
| Motorcycle | 23266 | 0.009 |
| Trips | 24867 | 0.186 |
| Car | 27090 | 0.271 |

Ho: The numeric variable is equal for the two samples. [1]: The statistic of the Wilcoxon-Mann-Whitney test. For details, see Hollander et al. (2013).

**Table 5:** Chi-square test for comparisons between the binary variables that comprise the original and synthetic samples.

| Variable | DF[1] | Q2[2] | X2[3] | P-value |
|---|---|---|---|---|
| Modal choice | 1 | 5.2813 | 0.0039 | 0.02156 |
| Driver's license | 1 | 0.01344 | 0.0039 | 0.9077 |
| Studies | 1 | 2.287E-30 | 0.0039 | 1 |
| Work | 1 | 13.472 | 0.0039 | 0.0002421 |
| Gender | 1 | 2.208E-29 | 0.0039 | 1 |
| Level of Education | 1 | 3.024E-30 | 0.0039 | 1 |

Ho: The binary variable is equal for the two samples. [1]: Degrees of Freedom, [2]: Chi-square test statistic, [3]: Chi-square critical value. For more details, see Pearson (1900).

## 2.2 General methodological procedure

After data processing and obtaining a synthetic sample after spatial interpolation (601 observations), the calibration of the non-spatial logit model is initially carried out. The set of independent variables chosen to calibrate the non-spatial logit model was trips, age and number of cars in the household (car). The modal choice (0 – Public transport, 1 – Car) was the dependent variable used.

Application of ESDA is the first spatial step and serves to better understand the data before a model is applied. Its importance in the proposed procedure is to analyze the presence of spatial effects for each variable and, if there are any, what the local patterns look like. Therefore, tools with local and global results were selected. It should be noted that the application of these analyses depends on the type of variable. For quantitative variables, Univariate Local Moran's I (Anselin, 1995) is appropriate. For binary variables, the Local Join Count (Anselin and Li, 2019) is recommended, which only provides local results. Two bandwidths were used: the first one considers the 20 nearest neighbors to assess the presence of spatial dependence. The second, used *a posteriori*, has the same value as the bandwidth used in the GWLR model, so that the ESDA results are viewed in the same scale as the spatial variability. Once the ESDA is carried out, if the results corroborate the spatial dependence, the GWLR model is calibrated

To calibrate the GWLR, a bi-square adaptive kernel was adopted with the nearest neighbor bandwidth with the smallest K possible that properly calibrates the model and minimizes its AICc. After the model's calibration, a spatial variability test is required to validate the model structure and local analysis of the estimated parameters. If the spatial variabilities of all the variables are confirmed, the GWLR model is the most appropriate. If part of the variables is spatially stationary, semi-parametric GWLR (Nakaya et al., 2014) or Multiscale GWLR (Fotheringham et al., 2017) might yield better results. If all analyzed variables are spatially stationary, a spatial global regression model could be the most suitable tool. A flowchart of the general procedure is present in Figure 3.

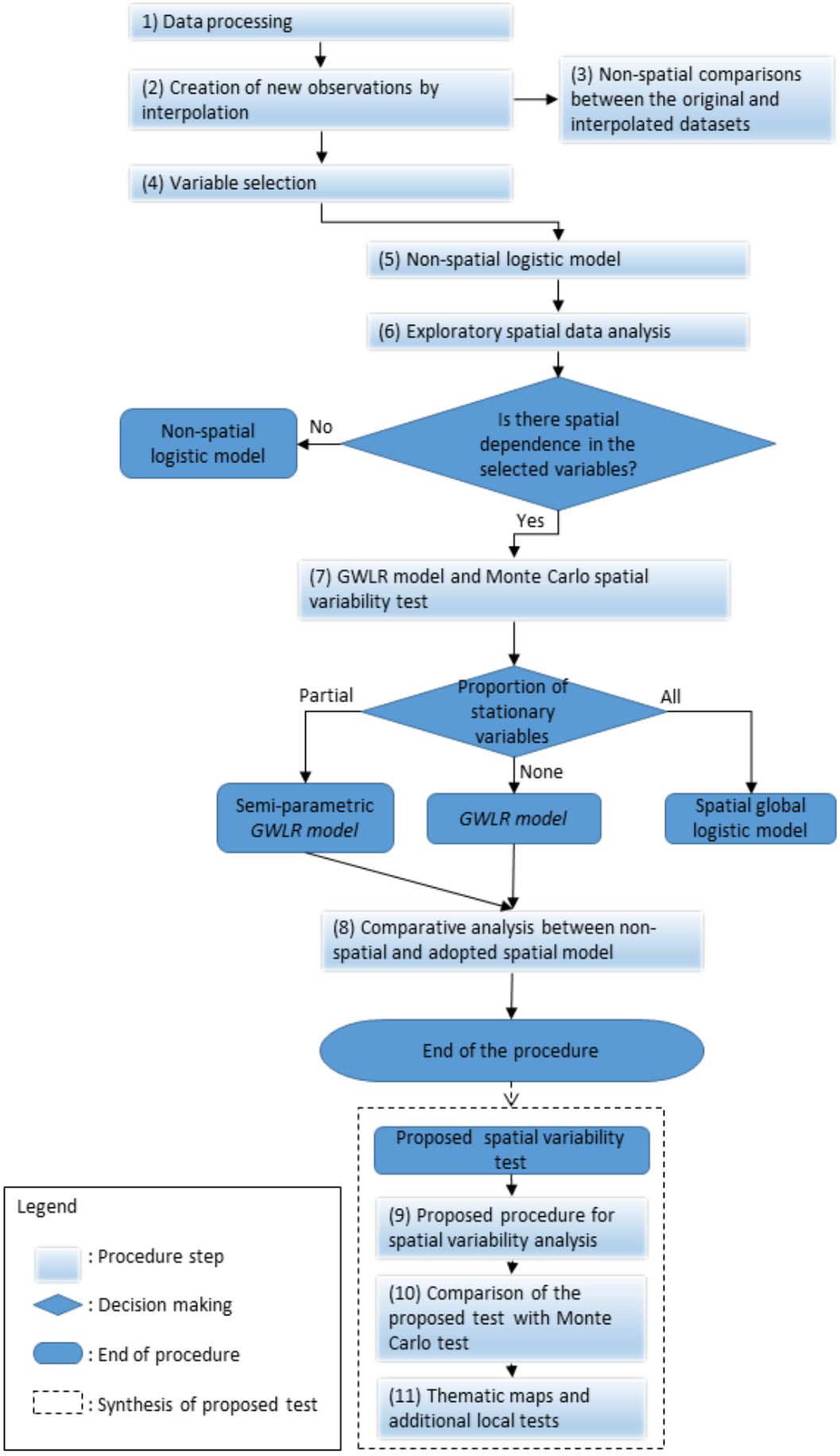


**Figure 3:** Flowchart of the general methodological procedure. GWLR: Geographically weighted logistic regression.

## 2.3 Procedure of the proposed spatial variability test

For the proposed test, 200 permutations were created of the set of variables from the original database for calibrating GWLR models under the assumption of spatial randomness. In this context, permutation means randomly changing only the coordinates of the observations, without repetition. Figure 4 illustrates the procedure for carrying out the proposed test. This test has the following requirements: (i) the parameters estimated in the non-spatial model are significant; (ii) presence of spatial dependence is corroborated by Exploratory Spatial Data Analysis for at least the dependent variable and one independent variable; (iii) the coefficients obtained by the local spatial model do not have very discrepant orders of magnitude between them (outliers are allowed); (iv) there is no considerable multicollinearity between the estimated parameters.

The null hypothesis of the test is spatial randomness of the estimated coefficients of the local model. Thus, the alternative hypothesis corroborates spatial variability. The following subsections describe each step.

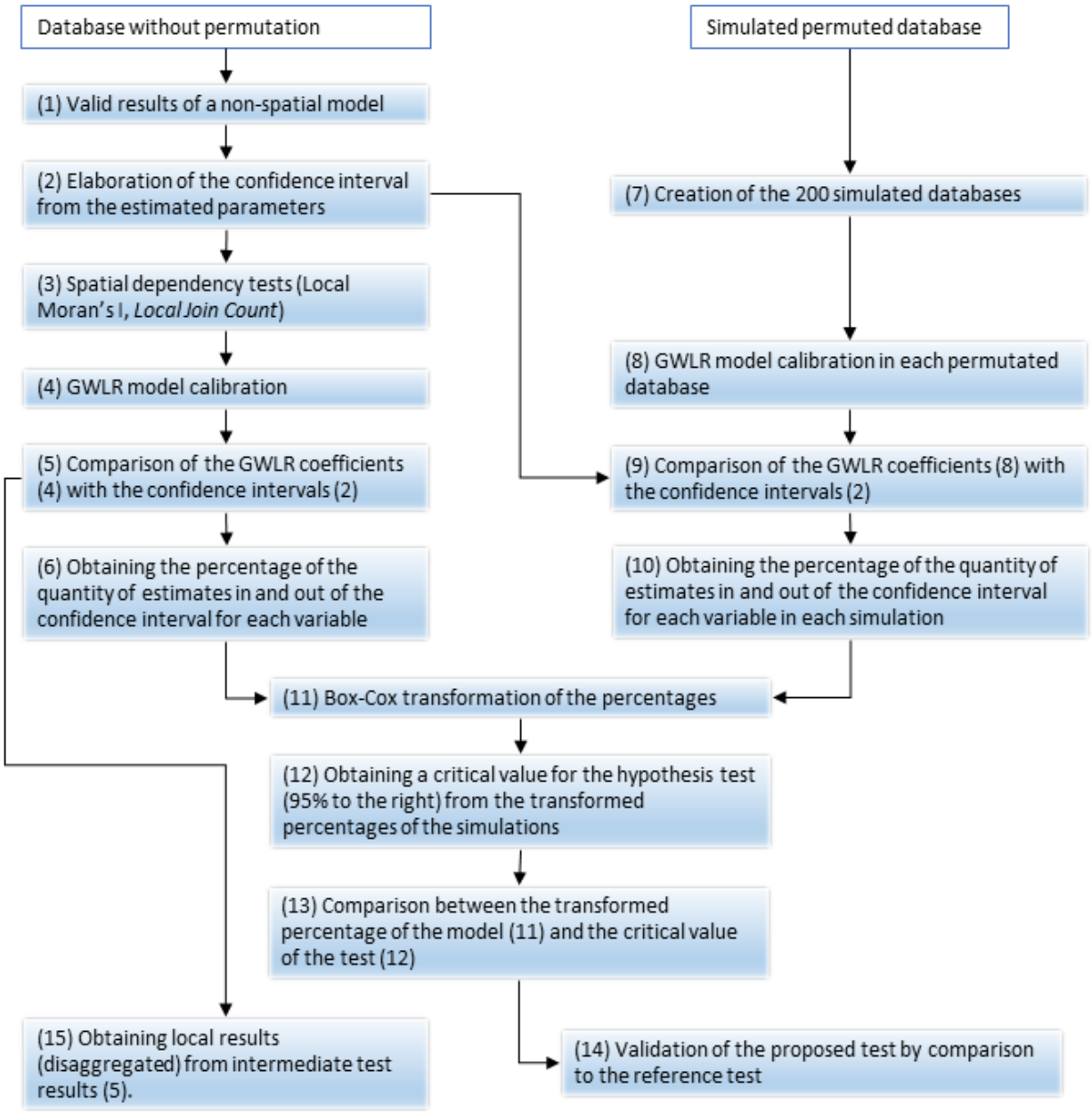


**Figure 4:** Procedure for carrying out the proposed test. GWLR: Geographically Weighted Logistic Regression.

The criteria of selection of the independent variables in the non-spatial model were: relevance of the variable to modal choice according to the literature and significance of the associated parameter. After calibrating the non-spatial model, a multicollinearity test of the independent variables is also performed. The confidence intervals for the parameters of each variable are then obtained, considering standard error values and normal distribution of the estimated parameters.

For the second requirement, exploratory analyses of univariate spatial data are performed to test spatial dependence. If the dependent variable and at least one independent variable have spatial dependence, the second test requirement is met. Otherwise, the most appropriate option is to adopt a non-spatial model.

Afterward, the GWLR model is applied at a bandwidth suitable for the case study. The smallest bandwidth, capable of calibrating the model, is recommended, and if the density of observations is heterogeneous, the use of the k nearest neighbors (KNN) criterion is suggested. If coefficients of the same variable are of very different orders of magnitude, the test indicator in the local spatial model or the critical value obtained may not be reliable. In such cases a bigger bandwidth is recommended. Since the variables are tested separately by their estimated parameters, the multicollinearity in the model coefficients (except simulated ones) needs to be evaluated as well. If there is no evidence of significant multicollinearity, the last prerequisite for applying the test is met. In this paper, the VIF - Variance Inflation Factor (Wheeler, 2007) metric was adopted to evaluate multicollinearity.

After obtaining the results of the GWLR model, the estimated coefficients are compared with the confidence intervals of the non-spatial model. Binary values inside and outside the range of estimated coefficients are also summarized in percentages, one for each variable in the model.

Two hundred new databases are simulated, with the coordinate pairs permuted (without repetition) from the original database. The local spatial model is calibrated for each new database, obtaining models in which the absence of spatial effects is expected. The only difference between processing the GWLR model in the original database and the GWLR model in the permuted databases is that the bandwidth was chosen by maximizing the AICc, leaving it variable between the bandwidth of the reference model and the maximum nearest neighbors (N-1). This is necessary because, in databases with spatial randomness, the bandwidth optimized by AICc in GWR models is expected to be very large (Fotheringham et al., 2022). Subsequently, the estimated local coefficients of all variables and all simulations are compared with the confidence interval of the non-spatial models, creating a percentage of estimated coefficients outside the confidence interval.

After creating a percentage of estimated coefficients outside the confidence interval for the GWLR models, a distribution of percentages in each variable is obtained, each with 200 elements. After that, the Box Cox transformation for normality is applied (Box and Cox, 1964). Then, the cutoff value for the significance level of 5% to the right is obtained. It is a one-tailed test because, as the confidence interval is created by coefficients estimated from a non-spatial model, a small percentage of parameters estimated outside the confidence interval does not corroborate spatial non-randomness.

Thus, the value of coefficients outside the Confidence Interval obtained by the reference model is compared to the cutoff value obtained by simulations. If the model's reference value is greater than the cutoff value, the null hypothesis of the test is rejected, consequently corroborating the spatial variability of the variable.

To verify the quality of the proposed test, the spatial variability test available in the MGWR program (Oshan et al., 2019) was applied. This test performs a thousand simulations (by default) per permutation in the coordinates. Each permuted database is modeled by selecting a new bandwidth, storing the standard error of the variables' coefficients. These standard errors are placed in ascending order, and the p-value of the test is the position of the reference model in this rank. Note that it is not necessary for the dependent variable to be binary, nor for the model used to be the GWLR. Any local model can be adopted. However, spatial variability is specific to the set of independent variables chosen from the reference model.

Furthermore, it has been shown that representing spatial effects by aggregated metrics can be misleading (Getis and Ord, 1992). Thus, more reliable results can surface from the application of ESDA tools to analyze the spatial non-randomness of the estimators by intermediate results of the proposed test. The thematic map of the local coefficients of the reference model inside and outside the Confidence Interval of the non-spatial model, and the Local Join Count test on this map are recommended to evaluate the spatial randomness of the generated result. Table 6 summarizes the computational tools used in each methodological step of this article.

**Table 6:** Tools used in the methodological procedure and proposed test.

| Step | Tool used | Authorship |
|---|---|---|
| Simple interpolation | ArcGIS | ESRI (2019) |
| Non-spatial models | R, *car package* | R Core Team (2023), Fox and Weisberg (2018) |
| Spatial exploratory analysis | GeoDa | Anselin et al. (2009) |
| Thematic maps | GeoDa, QGIS | Anselin et al. (2009), QGIS.org (2022) |
| *GWR models* | MGWR | Oshan et al. (2019) |
| Running the proposed test | R, *car package* | R Core Team (2023), Fox and Weisberg (2018) |

# 3. Results and discussion

## 3.1 Non-spatial modeling and obtaining parameter confidence intervals

Table 7 shows the estimated parameters obtained in the non-spatial model.

**Table 7:** Estimated parameters of the non-spatial model.

| Variable | Coefficient | Standard error | P-value | *VIF*[1] | Confidence interval |
|---|---|---|---|---|---|
| Intercept | -5.177 | 0.828 | 0.000 | - | -6.80, -3.55 |
| Trips | 5.735 | 1.774 | 0.001 | 1.01 | 2.26, 9.21 |
| Age | 3.163 | 0.870 | 0.000 | 1.06 | 1.46, 4.87 |
| Cars | 6.889 | 1.115 | 0.000 | 1.06 | 4.70, 9.08 |

VIF[1] – Variance Inflation Factor, metric for multicollinearity analysis (Wheeler, 2007)

All coefficient values obtained are in accordance with expectations, that is, the results of the estimated parameters are consistent with the literature, proving a positive relationship between car ownership, age and number of trips with the choice of car (Ben-Akiva et al., 1993; De Palma and Rochat, 2000; Ibrahim, 2003). Furthermore, as the VIF are less than 5, there is no evidence of multicollinearity (Wheeler, 2007). However, if spatial dependence is corroborated for most of the variables adopted, it is plausible that the non-spatial logit model is not the most appropriate as the observations may not be spatially independent of each other.

## 3.2 Univariate tests of spatial dependence

Local Moran's I (Anselin, 1995) was applied to the independent variables, see Table 8 and Figure 5. The initial bandwidth value was 20 nearest neighbors, to test whether spatial dependence was present, with positive results. After calibrating the GWLR model, these tests were rerun using the same model bandwidth, to maintain the analysis scale and show how spatial dependence is present in the variables. Although the values of the global indicators are relatively low (local average), the p-values, the number of significant local indicators and, mainly, the clear definition of the clusters, shown in Figure 5 corroborate the presence of spatial dependence in the three independent variables of the non-spatial model. Figure 5.A represents the number of trips in the household, Figure 5.B represents the age and Figure 5.C number of vehicles in the household. Figure 6 presents the Local Join Count test on the travel choice variable, car option, also corroborating the spatial dependence.

**Table 8:** Summary of Moran's I local for the independent variables of the non-spatial model.

| Variable | Percentage of significant local Moran's I[1] | Global Moran's I | P-value of global Moran's I |
|---|---|---|---|
| Trips | 63.73 | 0.022 | 0.00001 |
| Age | 92.35 | 0.105 | 0.00001 |
| Car | 34.44 | 0.009 | 0.0052 |

Percentage of significant local Moran's I[1] - Considering local p-value less than 0.05.

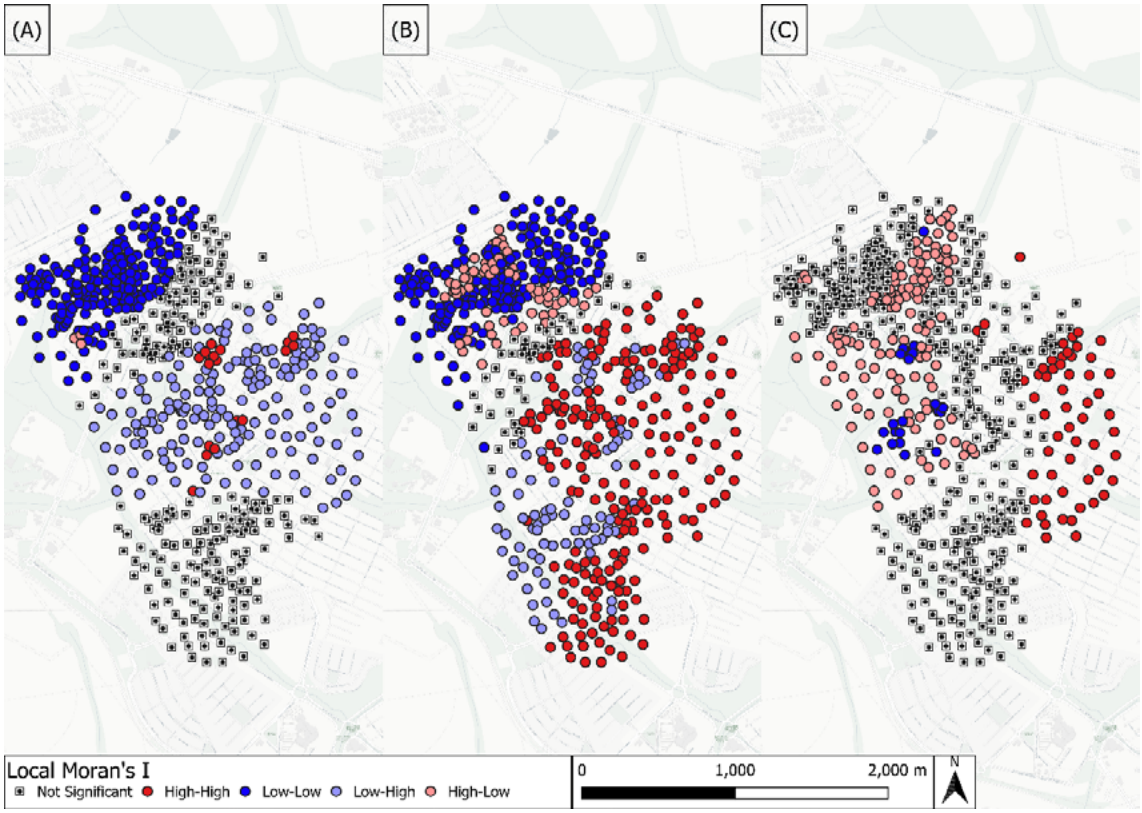


**Figure 5:** Local Moran's I clusters of the independent variables of the non-spatial model. Figure 5.A: Number of trips in the household, Figure 5.B: age, Figure 5.C: Number of vehicles in the household.

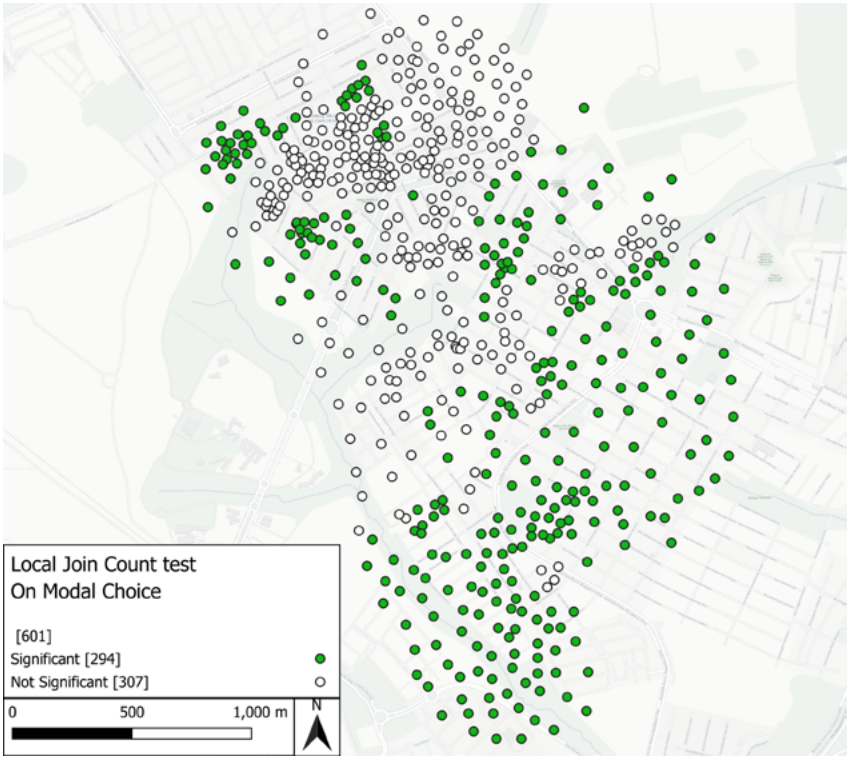


**Figure 6:** Local Join Count results for the modal choice.

## 3.3 GWLR model without permutations

Table 9 reports the medians and standard deviations of the estimated parameters of the GWLR model (considering only significant coefficients for a t-value of 1.96), with bandwidth of 232 nearest neighbors. Table 9 also shows the p-values of the reference spatial variability test. The VIF of the local model coefficients were also analyzed for multicollinearity. There is no considerable evidence of multicollinearity (Wheeler, 2007; Oshan et al., 2019). Thus, all test assumptions were met. Figure 7 shows the maps of the estimated coefficients of the GWLR model, with the legend values considering only the significant estimated coefficients of the variables.

**Table 9:** Summary of estimated parameters of the GWLR model.

| Variable | Minimum | Maximum | Median | Standard deviation | Significant coefficients (%)[1] | Maximum *VIF*[2] | Monte Carlo test |
|---|---|---|---|---|---|---|---|
| Intercept | -27.69 | -2.81 | -6.76 | 8.931 | 14.98 | - | 0.001 |
| Trips | 3.87 | 12.17 | 6.90 | 14.242 | 9.32 | 3.34 | 0.000 |
| Age | -12.19 | 28.58 | 7.27 | 8.562 | 46.09 | 1.34 | 0.000 |
| Cars | 2.73 | 18.65 | 9.73 | 3.642 | 93.18 | 3.27 | 0.019 |

Significant coefficients (%)[1] – Percentage of significant local coefficients considering p-value lower than 0.05; Maximum *VIF*[2] - *Variance inflation factor,* metric for multicollinearity analysis (Wheeler, 2007).

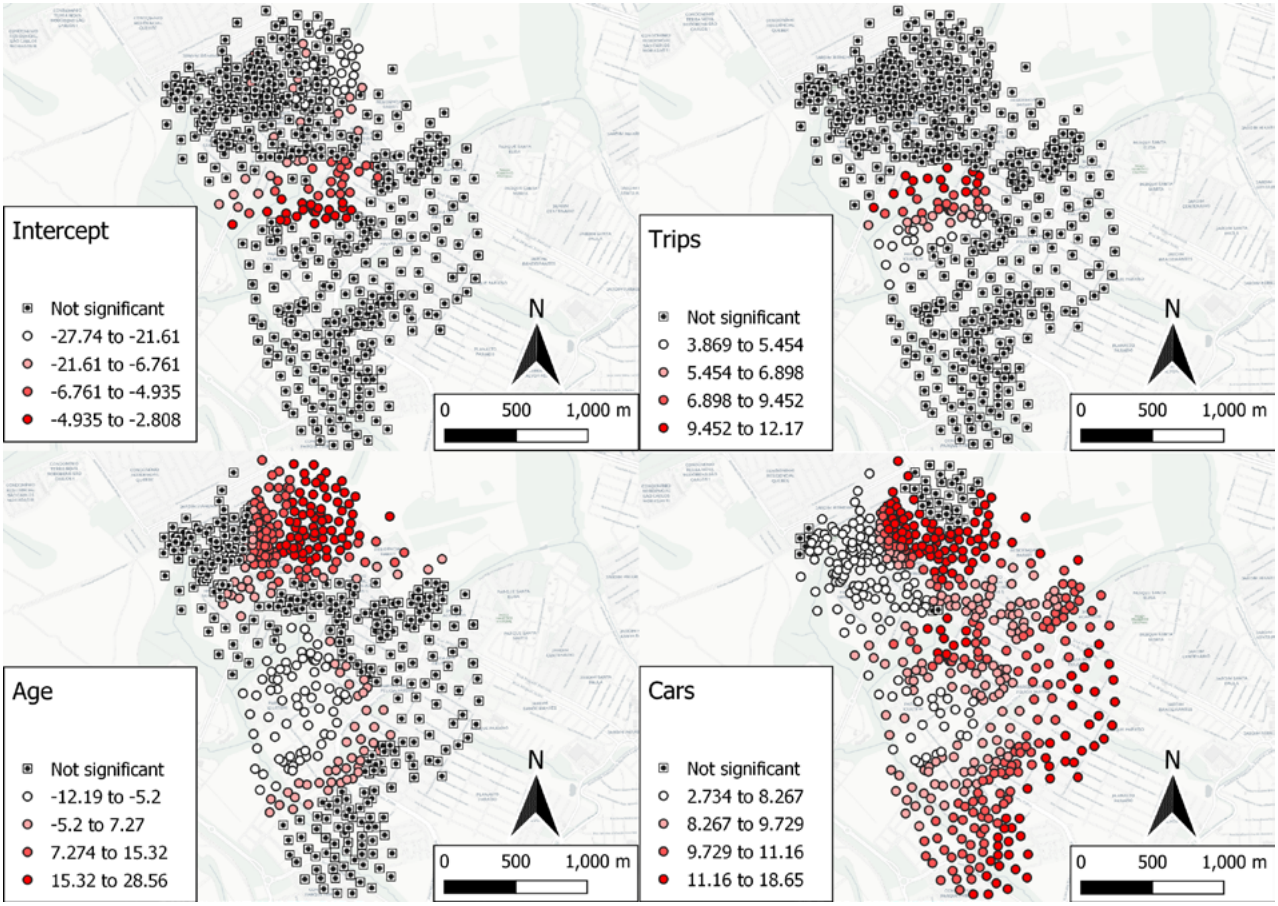


**Figure 7:** Estimated values for the intercepts, estimated coefficients of the variables “trips”, “age” and “number of vehicles in the household”.

For the intercept, the negative sign across the region is interpreted by the tendency that in the absence of influence from other factors, public transport has more utility. The estimators for “trips” and “car” variables are strictly positive, but with different regional trends and both with great variation in regional influence.

Another notable distinction between trip and car maps lies in the quantity of local estimators that are significant at a p-value of 0.05. While the “car ownership” estimators are not significant only in a northern region and part of the center, the “trip” estimators are only valid in the central-western region. The “age” variable, in turn, has parameters that, in addition to strong numerical variability, are negative or positive according to the region (with well-defined clusters), which may indicate a regional variation in the residents’ modal choice profile and the lack of other pertinent socioeconomic information (Ben-Akiva and Lerman, 1985). It might also indicate the need for a more detailed analysis on the elderly age group (Paez et al., 2007).

## 3.4 Comparisons between non-spatial and local spatial models

In the comparison between the non-spatial logit model and the GWLR model, the latter had a better fit in all the metrics considered (AICc, Log-Likelihood, adjusted pseudo $R^2$ and global accuracy rate). Table 10 describes the fit metrics that were compared, while Table 11 shows the confusion matrix for each model.

**Table 10:** Fit metrics for the non-spatial logit and GWLR models.

| Metric | Non-spatial logit model | GWLR model |
|---|---|---|
| *AICc* | 743.262 | 425.206 |
| *Log-Likelihood* | -367.598 | -191.491 |
| Adjusted $R^2$ pseudo | 0.112 | 0.523 |
| Global accuracy rate | 64.22 % | 85.19 % |

**Table 11:** Confusion matrix for the non-spatial logit and GWLR models.

| Observed value[1] | Non-spatial logit model | | GWLR model | |
|---|---|---|---|---|
| | **0** | **1** | **0** | **1** |
| 0 | 21.96 % | 25.62 % | 44.09 % | 3.49 % |
| 1 | 10.15 % | 42.26 % | 11.31 % | 41.1 % |

Observed value[1] - 0: public transport, 1: private car.

Table 11 shows that the largest gain in accuracy in the GWLR model is in the public transport mode with an increase of approximately 22% in correct choice estimates. Consequently, the error of the GWLR model considering only the public transport mode was 7.3%. Similarly, considering only the automobile mode, there was a slight increase in error, with 19.4% error in the non-spatial model and 21.6% in the GWLR model.

## 3.5 Proposed test

The geographic coordinates were randomly permuted 200 times, obtaining 200 randomly selected databases where there should be spatial randomness. Afterward, the GWLR model was calibrated on each of the new permuted databases, and its estimated coefficients were compared to the confidence interval of the non-spatial logit model, obtaining a binary vector. The value of 0 is given if the estimated coefficient is within the confidence interval, and 1 otherwise.

Table 12 summarizes the critical values and those of the model without permutation. In all variables, the hypothesis of spatial variability is validated by the proposed test, in line with the reference test used (see p-values in Table 9).

**Table 12:** Results of the proposed test.

| Variable | Model value without permutation [1] | Test cut-off value [1] | Proposed test result | Reference test result |
|---|---|---|---|---|
| Intercept | -0.126 | -0.809 | Spatial variability | Spatial variability |
| Trips | -0.130 | -1.064 | Spatial variability | Spatial variability |
| Age | -0.142 | -3.088 | Spatial variability | Spatial variability |
| Car | -0.310 | -2.359 | Spatial variability | Spatial variability |

Model value without permutation [1]/Test cut-off value [1] - Percentages of local coefficients outside the confidence interval of the non-spatial model coefficients; transformed by Box-Cox (1964).

## 3.6 Additional thematic maps and local tests

Figure 8 shows that the four analyzed variables are classified by comparing whether the local parameters estimated from the non-permutation GWLR model are outside or within the Confidence Interval of the non-spatial model.

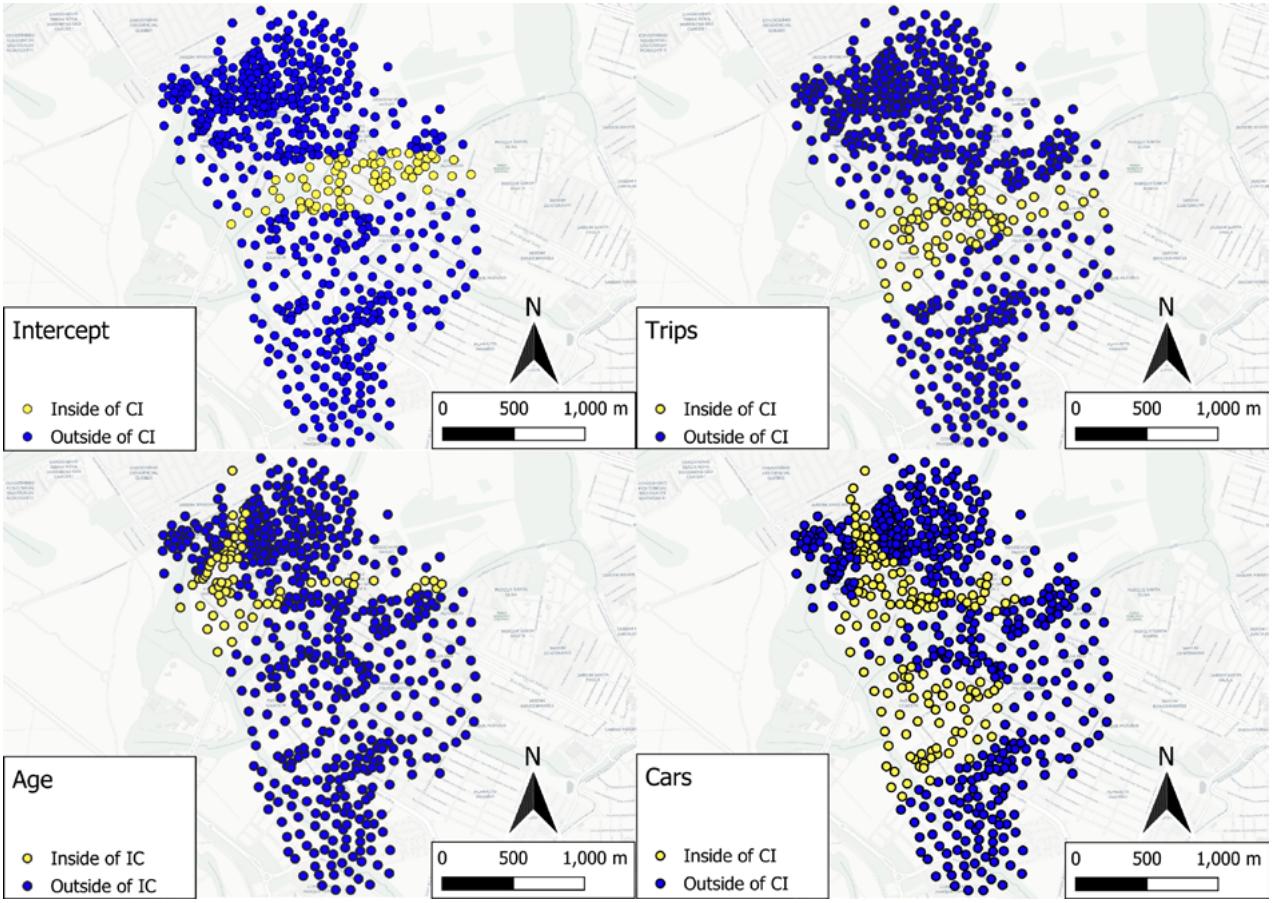


**Figure 8:** Thematic maps of local coefficients inside and outside the Confidence Interval (GWLR without coordinate permutations).

From the maps produced, clear patterns of division between local coefficients outside (blue dots) and inside (yellow dots) the confidence interval of the non-spatial model can be seen. Thus, in addition to the overall positive result, it is also clear that the patterns formed are not spatially random, providing an additional indication of the presence of spatial variability. Observing Figure 8 the regions where the local coefficient is within the confidence interval can be explored, where spatial variability may be weaker, irrelevant or absent.

Regarding the map values in Figure 8 the Local Join Count can be applied to analyze the local spatial dependence in the binary values formed. As suggested by (Anselin and Li, 2019), the value 1 of the variable is the one with the lowest occurrence (coefficients within the Confidence Interval). Figure 9 shows the spatial distribution of observations with local pseudo p-values below 0.05 (black dots) by the Local Join Count test.

Given the inversion of values, the clusters of black dots are regions with spatial dependence of the coefficients within the Confidence Interval of the non-spatial model. In other words, the blank regions are those with corroborated spatial variability. Once more, there are clear regional patterns in each variable and in the intercept, very similar to Figure 9. However, the Local Join Count has the clear advantage of being a test and its interpretation is more reliable.

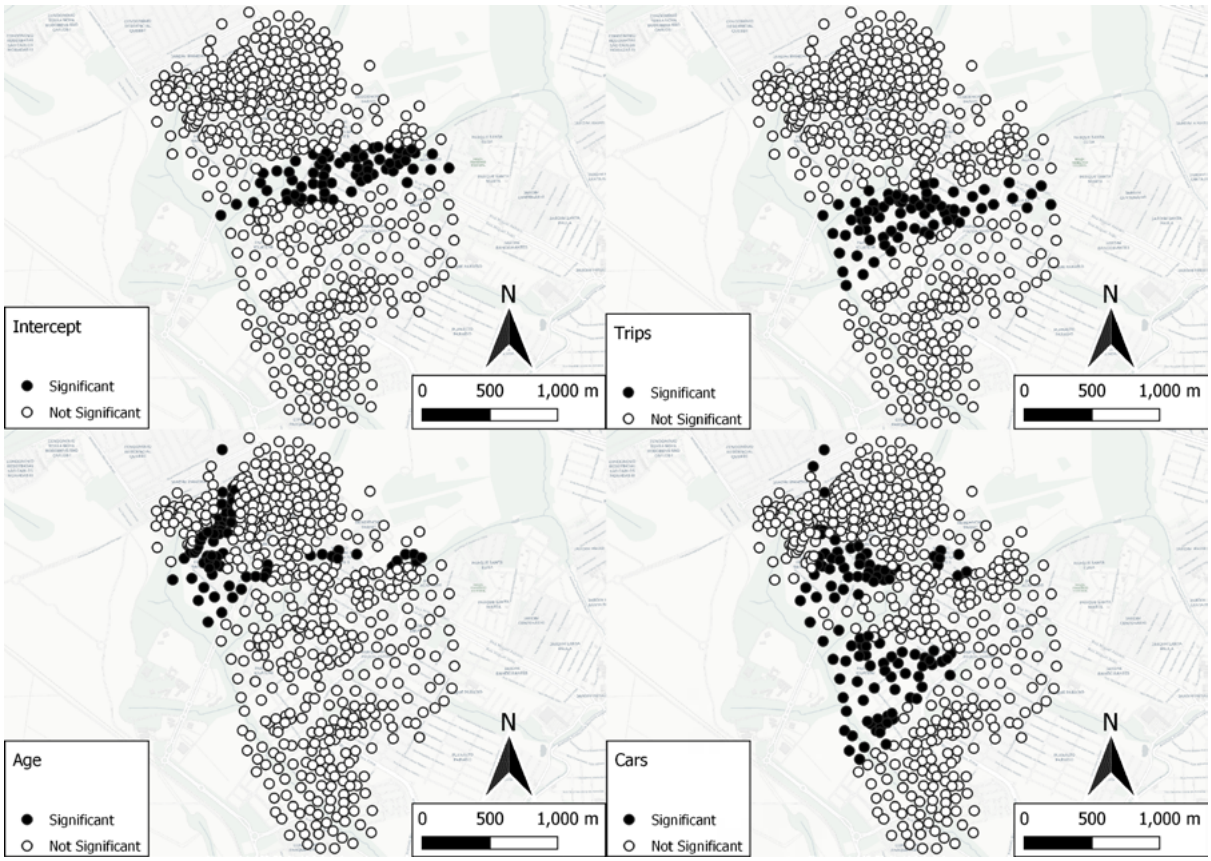


**Figure 9:** Local Join Count test when comparing local coefficients within the Confidence Interval.

# 4. CONCLUSIONS, LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

This paper has two objectives. The first is to analyze a modal choice problem simultaneously, considering spatial dependence and variability. The second is to propose a spatial variability test, using results from both local and non-spatial models, through simulations and comparisons with confidence intervals. In the first objective, a methodological sequence was used starting from a non-spatial logit model, followed by ESDA and ending with the GWLR model. The order of these steps was designed to meet the conceptual requirements of both spatial effects, the necessary assumptions of the GWLR model, highlight the best fit of the models and qualitative gains in local results and also guide researchers to better understand the case study, given the lack of work focusing on GWLR models in the modal choice literature.

The values of the estimators obtained are in line with what was expected given the literature consulted. The cohesion of values in each region also stands out, forming concise clusters. The results presented corroborate the feasibility of applying a GWLR model to study modal choice, and also provide strong evidence that the available local results expand the interpretation of results beyond what would be possible with stationary coefficients. However, depending on the variables selected and the literature consulted, it is more plausible that they are considered stationary. In these contexts, semi-parametric GWLR models may outperform the classic GWLR model (Fotheringham et al., 2002).

Furthermore, the results of the article must be seen with some limitations in mind. First, the lack of a wider quantity of observations for the model required the creation of new observations by interpolation, which limited more in-depth discussions of the results, including the spatial variability. Secondly, the unavailability of variables considered important in the literature, such as the alternative attributes. Moreover, these limitations refer only to the case study, not the proposed method.

Regarding the second objective, the test application was considered successful. Its results are easy to interpret, they are in accordance with the reference test for this case study, and the additional local results allow an in-depth insight to how the spatial variability is present in each variable. It is noteworthy that, in the future, the test can also be applied with a spatial global model instead of a non-spatial one to create the confidence interval of the variables. Spatial global models can incorporate spatial dependence, but their estimated parameters are stationary. Thus, in this case, there is greater guarantee that the statistical difference between the local estimators and the adopted confidence interval is only due to spatial variability.

According to Getis and Ord (1992) and Fotheringham et al. (2002), aggregated metrics, even if intrinsically spatial, can hide important local trends. Therefore, an additional step proposed was to use local, intermediate test results or easily developed ones, to corroborate the proposed test and evaluate the gain in exploratory potential. For this, direct comparison maps of local parameters were used with the Confidence Interval of the non-spatial model and a Local Join Count test on the values used. This proved to be viable for analyzing spatial variability and facilitate the understanding and exploration of the case study. Using these maps, the regional segregation between spatially stationary estimate and spatially varying estimate groups and their internal cohesion could be evaluated. This allows us to analyze if certain subgroups are more likely to be spatially stationary, and if these groups exhibit a spatial pattern (internal cohesion and external segregation). This is not an available result in the observed literature and can yield interesting conclusions if there is interest in verifying where spatial variability might not be valid. Thus, it benefits a wide array of study areas, including mode choice, which has received growing interest to expand in spatial analysis in recent years.

Finally, future research is suggested using entirely real data and with more available observations to better evaluate the proposed test and possibly observe cases where there are also stationary variables.

## ACKNOWLEDGEMENT

We extend our thanks for the financial support from the National Council for Scientific and Technological Development (CNPq) - Process 305973/2023-1 and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 617463/2021-00.

## AUTHOR´S CONTRIBUTION

Author 1: Conceptualization, methodology, data analysis, writing (draft and revision), editing, final review; Author 2: Conceptualization, methodology, writing (revision), proofreading, supervision.

## REFERENCES

Anselin, L. 1988. *Spatial Econometrics: Methods and Models* (Vol. 4). Springer Science & Business Media.

Anselin, L. 1995. Local indicators of spatial association—LISA. *Geographical analysis*, 27(2), 93-115.

Anselin, L. and Li, X. 2019. Operational local join count statistics for cluster detection. *Journal of geographical systems,* 21, 189-210.

Anselin, L. Syabri, I. and Kho, Y. 2009. GeoDa: an introduction to spatial data analysis. In *Handbook of applied spatial analysis: Software tools, methods and applications* (pp. 73-89). Berlin, Heidelberg: Springer Berlin Heidelberg.

Assirati, L. and Pitombo, C. S. 2021. Modeling spatial effect on travel mode choice using a synthetic spatially correlated data set. *Bulletin of Geodetic Sciences*, 27, e2021008.

Ben-Akiva, M. Bolduc, D. and Bradley, M. 1993. Estimation of travel choice models with randomly distributed values of time. *Transportation Research Record*, 1413, 88-97.

Ben-Akiva, M. E. and Lerman, S. R. 1985. *Discrete choice analysis: theory and application to travel demand*. MIT press.

Box, G. E. and Cox, D. R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2), 211-243.

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.

Costa, A. S. G. D. 2013. *Proposta de um método para estimação de escolha modal através da geoestatística* [Mastership dissertation, Universidade Federal da Bahia]. Repositório Institucional da UFBA. https://repositorio.ufba.br/handle/ri/18709

ESRI 2019. *ArcGIS Desktop 10.8* (Version 10.8). [online] Available at: <https://desktop.arcgis.com/en/arcmap/latest/get-started/installation-guide/installing-on-your-computer.htm>. [Acessed 26 April 2024].

Fotheringham, A. S. Brunsdon, C. and Charlton, M. 2002. *Geographically weighted regression: the analysis of spatially varying relationships.* John Wiley & Sons.

Fotheringham, A. S. Yang, W. and Kang, W. 2017. Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers*, 107(6), 1247-1265.

Fotheringham, A. S. Yu, H. Wolf, L. J. Oshan, T. M. and Li, Z. 2022. On the notion of 'bandwidth'in geographically weighted regression models of spatially varying processes. *International Journal of Geographical Information Science*, 36(8), 1485-1502.

Fox, J., and Weisberg, S. 2018. *An R companion to applied regression*. Sage publications.

Getis, A. and Ord, J. K. 1992. The analysis of spatial association by use of distance statistics. *Geographical analysis,* 24(3), 189-206.

Gomes, V. Pitombo, C. Rocha, S. and Salgueiro, A. 2016 Kriging Geostatistical Methods for Travel Mode Choice: A Spatial Data Analysis to Travel Demand Forecasting. *Open Journal of Statistics,* 6, 514-527.

Hollander, M. Wolfe, D. A. and Chicken, E. 2013. *Nonparametric statistical methods.* John Wiley & Sons.

Ibeas, Á. Cordera, R. dell'Olio, L. and Moura, J. L. 2011. Modelling demand in restricted parking zones. *Transportation Research Part A: Policy and Practice*, 45(6), 485-498.

Ibrahim, M. F. 2003. Car ownership and attitudes towards transport modes for shopping purposes in Singapore. *Transportation*, 30, 435-457.

LeSage, J. P. 1999. *The theory and practice of spatial econometrics.* [pdf] University of Toledo. Available at: <http://spatial-econometrics.com/html/sbook.pdf> [Acessed 26 April 2024].

Leung, Y. Mei, C. L. and Zhang, W. X. 2000. Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning A,* 32(1), 9-32.

Lindner, A. and Pitombo, C. S. 2018. A conjoint approach of spatial statistics and a traditional method for travel mode choice issues. *Journal of Geovisualization and Spatial Analysis*, 2, 1-13.

Mondal, A. and Bhat, C. R. 2022. A spatial rank-ordered probit model with an application to travel mode choice. *Transportation Research Part B: Methodological*, 155, 374-393.

Nakaya, T. Charlton, M. Lewis, P. Brunsdon, C. Yao, J. and Fotheringham, A. S. 2014. GWR4 user manual. *Windows Application for Geographically Weighted Regression Modelling.* Ritsumeikan University, Kyoto, Japan. [online] Available at: <https://gwr.maynoothuniversity.ie/wp-content/uploads/2013/04/GWR4_Manual.pdf> [Acessed 26 April 2024]

Nkeki, F. N. and Asikhia, M. O. 2019. Geographically weighted logistic regression approach to explore the spatial variability in travel behaviour and built environment interactions: Accounting simultaneously for demographic and socioeconomic characteristics. *Applied geography*, 108, 47-63.

Oshan, T. M. Li, Z. Kang, W. Wolf, L. J. and Fotheringham, A. S. 2019. mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information,* 8(6), 269.

Paez, A. 2006. Exploring contextual variations in land use and transport analysis using a probit model with geographical weights. *Journal of Transport Geography,* 14(3), 167-176.

Paez, A. Scott, D. Potoglou, D. Kanaroglou, P. and Newbold, K. B. 2007. Elderly mobility: demographic and spatial analysis of trip making in the Hamilton CMA, Canada. *Urban Studies,* 44(1), 123-146.

De Palma, A. and Rochat, D. 2000. Mode choices for trips to work in Geneva: an empirical analysis. *Journal of Transport Geography*, 8(1), 43-51.

Pearson, K, 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157-175.

Pitombo, C. S. Salgueiro, A. R. da Costa, A. S. G. and Isler, C. A. 2015a. A two-step method for mode choice estimation with socioeconomic and spatial information. *Spatial Statistics*, 11, 45-64.

Pitombo, C. S. Costa, A. S. G. D. and Salgueiro, A. R. 2015b. Proposal of a sequential method for spatial interpolation of mode choice. *Bulletin of Geodetic Sciences,* 21(2), 274-289.

Propastin, P. A. 2009. Spatial non-stationarity and scale-dependency of prediction accuracy in the remote estimation of LAI over a tropical rainforest in Sulawesi, Indonesia. *Remote Sensing of Environment*, 113(10), 2234-2242.

QGIS.org 2022. *QGIS Geographic Information System. QGIS Association*. [online] Available at: <qgis.org> [Acessed 26 April 2024].

R Core Team 2023. *R: A language and environment for statistical computing.* [online] Available at: <https://www.r-project.org/> [Acessed 26 April 2024]

Rajamani, J. Bhat, C. R. Handy, S. Knaap, G. and Song, Y. 2003. Assessing impact of urban form measures on nonwork trip mode choice after controlling for demographic and level-of-service effects. *Transportation research record,* 1831(1), 158-165.

Rodrigues da Silva, A. N. 2008. Elaboração de um Banco de Dados de Viagem para Auxílio ao Desenvolvimento de Pesquisas na Área de Planejamento dos Transportes. Universidade de São Paulo, Escola de Engenharia de São Carlos, Relatório FAPESP, Processo, n. 04/15843, p. 4.

Tao, X., Fu, Z., and Comber, A. J. 2019. An analysis of modes of commuting in urban and rural areas. *Applied Spatial Analysis and Policy,* 12, 831-845.

Wheeler, D. C. 2007. Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and planning A,* 39(10), 2464-2481.

Yang, H. Zhang, Y. Zhong, L. Zhang, X. and Ling, Z. 2020. Exploring spatial variation of bike sharing trip production and attraction: A study based on Chicago's Divvy system. *Applied geography,* 115, 102130.