NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Modelling Abstention Rate using Spatial Regression

Afonso Manita Santos Mota

Dissertation presented as partial requirement for obtaining the Master's degree in Statistics and Informa

ion Management, with specialization in Information Analysis and Management.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

2018 Modelling Abstention Rate using Spatial Regression Afonso Manita Santos Mota MEGI

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

MODELLING ABSTENTION RATE USING SPATIAL REGRESSION

by

Afonso Manita Santos Mota

Dissertation presented as

partial requirement for obt

aining the Master's degree in Statistics and Information Management, with specialization in Information Analysis and Management.

Advisor Ana Cristina Marinho da Costa, PhD

November 2018

ABSTRACT

During the last few elections that were held in Portugal, there have been very low percentages of voter turnout. This will obviously impact the result of those elections and can maybe be related to the general disenchantment of the population regarding the country's recent political environment.

This study aims to contribute to a better understanding of the patterns in the abstention rate of the last elections in Portugal. Sociological and economic variables such as age, unemployment rate, education level and many others will be used in trying to find out if they influence the abstention rate. It is logical to assume that the abstention rate in a certain municipality will be related to the abstention in neighboring municipalities. Therefore, the study also investigates if there is spatial autocorrelation in the abstention rates.

Modeling a phenomenon like this with a simple linear regression model, estimated by Ordinary Least Squares (OLS), will render less efficient and biased results because of the spatial correlation of the observations and possible spatial clustering of values. Spatial regression methods have been proposed to overcome these drawbacks, particularly the

Geographically Weighted Regression (GWR). This method will take into account possible local influences, allowing the coefficients of the model to vary depending on the geographic location, possibly obtaining a more appropriate fit. Many different OLS and GWR models were investigated by considering different combinations of explanatory variables and diagnosing their results through statistical tests and goodness-of-fit measures.

Results show that indeed the data exhibits a non-random spatial pattern, and that a GWR model is a better approach in modeling abstention rates, when compared to an OLS model. Hence, the percentage of voter turnout in a municipality is likely to be better modelled taking into account its geographic location.

KEYWORDS

Voter turnout; Abstention rate; Sociological Variables; Economic Variables; Spatial Analysis; Geographically Weighted Regression; Spatial Non-Stationarity.

INDEX

1. Introduction 1
   1.1. Study Relevance and Importance 1
   1.2. Study Objectives 1
2. Literature review 3
   2.1. Voter Turnout 3
   2.2. Theoretical Framework 5
      2.2.1. Exploratory Spatial Data Analysis 5
      2.2.2. Ordinary Least Squares Regression 5
      2.2.3. Geographically Weighted Regression 6
3. Methodology 9
   3.1. Data Collection 9
   3.2. Exploratory Spatial Data Analysis 11
      3.2.1. Global Moran's I statistic 11
      3.2.2. Local Moran's I statistic 12
      3.2.3. Getis-Ord General G statistic 13
      3.2.4. Getis-Ord Gi* statistic 14
   3.3. Ordinary Least Squares Models 15
   3.4. Geographically Weighted Regression Model 16
   3.5. Comparing Models 19
4. Results and discussion 20
   4.1. Exploratory Spatial Data Analysis 20
   4.2. Ordinary Least Squares Models 23
   4.3. Geographically Weighted Regression Model 25
   4.4. Comparing Models 31
5. Conclusion 33
   5.1. Limitations 33
   5.2. Future work 34
6. REFERENCES 35

LIST OF TABLES

Table 1 – Initial variables of the study 11
Table 2 – Global Moran's Analysis 21
Table 3 – Getis-Ord General G Analysis 21
Table 4 – Explanatory variables used for the best GWR models for each number of variables 25
Table 5 – Results for the GWR models 26

LIST OF FIGURES

Figure 1 – Distribution of abstention rate 20
Figure 2 – Local Moran's I 22
Figure 3 – Getis-Ord Gi* 22
Figure 4 – Standardized Residuals of the GWR model 27
Figure 5 – Local R^2 28
Figure 6 – Coefficients for Ind_env 29
Figure 7 – Standard Error of the Coefficients for Ind_env 29
Figure 8 – Coefficients for Superior 29
Figure 9 - Standard Error of the Coefficients for Superior 29
Figure 10 – Predicted values 31
Figure 11 – Actual values 31

LIST OF ABBREVIATIONS AND ACRONYMS

AIC	Akaike Information Criterion
AICc	Adjusted Akaike Information Criterion
ESDA	Exploratory Spatial Data Analysis
GWR	Geographically Weighted Regression
OLS	Ordinary Least Squares
VIF	Variance Inflation Factor

1. INTRODUCTION

1.1. STUDY RELEVANCE AND IMPORTANCE

The study of voter turnout in political elections is a problem that has been progressively more taken into account throughout the recent years (Cancela & Geys, 2016; Hooghe & Kern, 2017). Although there are many older studies on this matter (Geys, 2006), there are many new studies every year, since the causes of voter turnout can vary dramatically through time and across the territorial scope of the election (André Blais, 2006).

On the year 2000, Burden made a study about the studies of voter turnout. The objective of that study was to understand the increasingly severe problem of the decrease of voter turnout throughout the last few years in the USA. In fact, voter turnout has been linearly decreasing, and in some cases, the decrease is higher than 15% when comparing to the 1960's. A number of impressive voter turnout declines were also recorded in newly consolidated democracies such as Portugal, El Salvador, South Korea and Romenia (Kostelka, 2017).

In Portugal, voter turnout exhibits this trend in every kind of election. Either in the presidential or municipal elections, the decreases have been in some cases, close to 20%, between consecutive elections. This problem is even more severe nowadays because the last few elections have had voter turnouts very close to only 50%. The effects of socioeconomic variables on voter turnout have been studied in different countries, including Portugal (Freire & Magalhães, 2002; Martins & Veiga, 2013), but, to the best of our knowledge, spatial regression models have never been used to investigate why some locations, such as municipalities, exhibit higher or lower voting turnouts.

1.2. STUDY OBJECTIVES

The main objective of this study is to understand how the explanatory power of different economic and sociological variables affecting the abstention rates varies from municipality to municipality in continental Portugal, particularly the abstention rate of the 2013 Portuguese municipal elections. We will investigate the possible spatial correlation between the abstention rates in municipalities that are close to one another. We hope this research

allows for developing a regression model that is able to explain the abstention rate in a certain municipality using information from neighboring municipalities.

During this study, we will have several secondary objectives:

- To gather reliable data for the dependent and potential explanatory variables, in the chosen study region;
- To find explanatory variables that are statistically significant for our model;
- To investigate patterns that suggest spatial correlation in the data;
- To estimate a somewhat reliable linear regression model using Ordinary Least Squares (OLS);
- To estimate a Geographically Weighted Regression (GWR) model;
- To draw a conclusion regarding the comparison of the models' results.

In summary, in this study we will try to find the answer for the following main questions:

1. *Are spatial regression models more appropriate than classical regression models to model abstention rates?*
2. *Which economic and sociological variables affect more the abstention rates of each municipality?*

2. LITERATURE REVIEW

2.1. VOTER TURNOUT

During the last few years, all throughout Europe, the problem of low percentage of voter turnout has been studied extensively (Cancela & Geys, 2016; Hooghe & Kern, 2017; Lahtinen, Mattila, Wass, & Martikainen, 2017; Sundström & Stockemer, 2015). Some causes of variations in turnout are consistently supported by empirical evidence, but others remain ambiguous (André Blais, 2006). Moreover, Freire & Magalhães (2002) discuss conflicting results of different studies on Portuguese elections, which may suggest that the main "influencers" of voter turnout vary from municipality to municipality.

Many studies have been made to try and find out the determinants of voter turnout. From these studies many theories have emerged regarding this problem. Sociological variables, especially age and education, have been consistently referenced as important explanatory factors (André Blais, Pilet, Van der Straeten, Laslier, & Héroux-Legault, 2014; Franklin, 2004; Ley, 2017; Lijphart, 2007). We also know that there is a difference in those determinants from country to country. Similar studies show that different Institutional arrangements related to electoral laws provide different turnout outcomes (Jackman, 1987; Jackman & Miller, 1995; Powell, 1986). These studies usually state that more competitive political environments will be related to higher voter turnouts.

On the other hand, the effect of the economy on turnout has been studied and rendered contradictory results. There have been many studies that prove that there is in fact a relation between various economic variables and the way a person votes (Nannestad & Paldam, 1994; Paldam, 2008). However, studies about the way it affects voter turnout do not have any kind of consensus, although there are signs that show that more economically advanced countries have a higher turnout (Blais & Dobrzynska, 2009; Fornos, Power, & Garand, 2004; Norris, 2004).

In Portugal, these issues have also been studied (Freire & Magalhães, 2002; Martins & Veiga, 2013). The impact of the economy on voter turnout at a municipal level has been observed to be higher in times when the economy state of the country is very good or very bad. These

studies investigate the effects of standard sociological, demographical, institutional, and economic variables.

Freire & Magalhães (2002) point out several advantages and disadvantages of using electoral polls, such as the Eurobarometer (http://ec.europa.eu/commfrontoffice/publicopinion/), to study electoral behaviours. These authors also discuss the results of several multiple linear regressions of average abstention in the 70s and 90s in European democracies, by analysing the impact of institutional variables (e.g., average number of political parties), and socioeconomic variables (e.g., percentage of population with higher education, unemployment rate, GDP per capita at purchasing power parity). Freire & Magalhães (2002) also used logistic regressions to investigate individual and contextual determinants of abstention in Portugal and Europe. Moreover, they used multivariate logistic regressions to study socioeconomic causes of abstention in the 1999 parliamentary elections, and also in the 2001 presidential elections, in Portugal. Similarly to other authors, Freire & Magalhães (2002) concluded that age is a very relevant factor, among other institutional and political determinants.

Martins & Veiga (2013) used an autoregressive model using a set of(local and national) economic explanatory variables, and, to try and control non-economic factors, a set of non-economic explanatory variables, such as size of the municipality, number of political parties, number of consecutive terms in office for the same party and a Boolean variable that indicates if the previous Mayor is running for another term. This model considered several different electoral periods which rendered a fairly trustworthy result. One of the most important conclusions of this study is that local governments seem to held accountable for times when the unemployment rate is higher or lower than average. This effect is even more noticeable when the local government is tied to the national governing party.

One aspect of this topic that has not been fully studied is the possibility of using spatial regression to investigate why some locations (e.g., municipalities) exhibit higher or lower voting turnouts.

2.2. THEORETICAL FRAMEWORK

2.2.1. Exploratory Spatial Data Analysis

According to Anselin (1998), Exploratory Spatial Data Analysis (ESDA) is a collection of techniques to:

- Describe and visualize spatial distributions, by mapping the values of each variable and examining the patterns it displays;
- Identify spatial outliers and discover patterns of spatial association such as local cluster. Which can be done by analyzing the values of the Local Moran's I statistic;
- Identify variables that exhibit high or low value clusters (Hotspot Analysis), by examining the value of the Getis-ord Gi* statistic;
- Suggest spatial regimes or other forms of spatial heterogeneity which happens when a variable has distinct distributions for different geographic sub regions.

The objective of this analysis is to get an initial visualization of how the variables are distributed throughout the study region and hopefully identifying spatial autocorrelation in the dataset. Spatial autocorrelation measures if a variable is correlated with itself in neighboring locations. If it has high positive values, similar values occur close to one another, if it has high negative values, we observe very different values occurring close to one another.

2.2.2. Ordinary Least Squares Regression

Simple linear regression is the most used regression method. This method helps investigate bivariate and multivariate relationships between variables, where we hypothesize that there is one predicted variable that depends on a combination of other variables. This model creates an estimate for the coefficients of each variable using the Ordinary Least Squares (OLS) estimator, which is proven to be the best linear unbiased estimator given the following assumptions as explained by Poole & O'Farrell, 1971 and by Hayashi, 2000:

1. Linearity (i.e. the model is correctly specified);

2. Random sampling (i.e. observations are independent of each other, thus no autocorrelation of the residuals);
3. Strict exogeneity (expected value of the residuals equal to zero);
4. No multicollinearity;
5. Spherical error variance (homoscedasticity of the residuals).
6. Normally distributed residuals;

To test if these assumptions are met in a given dataset, several diagnostics have to be made:

- Examine the scatter plots of the dependent variable with each explanatory variable (Assumption 1);
- Durbin-Watson test to assess the temporal autocorrelation of the residuals (Assumption 2);
- Student's T-test (Assumption 3);
- Examine the Variance Inflation Factor (Assumption 4);
- Examine the plot of residuals versus predicted values, and test homoscedasticity for example with the Breush-Pagan test (Assumption 5)
- Test residuals normality, for example with the Shapiro-Wilk's test or Jarque-Bera test (Assumption 6)

2.2.3. Geographically Weighted Regression

When the residuals of an OLS model have spatial autocorrelation or exhibit spatial heterogeneity such as clusters (i.e. spatial non-stationarity), spatial regression models should be used.

The Geographically Weighted Regression (GWR) models are frequently used in geographical analysis. GWR was developed initially by Fotheringham et al. (1997) and more fully detailed again by Fotheringham et al. (2002). In these papers, GWR is explained as a method to create local summary statistics from geographically weighted point data. Afterwards these statistics are mapped and used to identify possible variations in the distribution of the variable of interest from location to location. This method has been used to model various situations from various different areas, such as Agriculture (Xu & Lin, 2017), Health,

Environment (Wu, Yang, Guo, & Han, 2017), Economy (Benassi & Naccarato, 2017) and Transports (Chiou, Jou, & Yang, 2015).

GWR is a spatial regression method that is particularly used when spatial non-stationarity is found in the study data. Since stable variance in the data is required in order to assume that the Ordinary Least Squares Regression provides the best linear unbiased estimator, the GWR model will differ from the OLS model by estimating coefficients that depend on the spatial location of each observation. Therefore, unlike the OLS model, GWR is a local model.

This method takes into account the variability of the data throughout the study area. And, to do so, it estimates a set of coefficients, one for each variable, at any given location (point or polygon, generally referred to as feature). GWR makes a point-wise calibration of the coefficients by assuming that observations which are closer to the regression feature will have a bigger influence in estimating that set of coefficients when compared to the observations that are farther away (Brunsdon, Fotheringham, Charlton, Brunsdont, & Chariton, 1998). Basically, GWR models relationships around each location in the data set, estimating the regression coefficients by weighted least squares using a spatial weights matrix. For each location, the data will be weighted differently so that the results of any one calibration are unique to a particular location.

The Weighting matrix can be calculated with many different techniques, but the one which is most used is the 'Gaussian-like' kernel method. The kernel bandwidth is a parameter that needs to be specified either by a fixed number of nearest neighbours or by a fixed distance. When these values are the same for all features in the data set the procedure is named as fixed spatial kernel.

Fixed spatial kernels have a few potential drawbacks. Where data points are sparse the local models might be calibrated on very few data points, thus the coefficient estimates are less reliable (i.e. have large standard errors). Where data points are dense the coefficient estimates are more likely to be biased, because there is more scope for examining changes in relationships over relatively small distances and such changes might be missed with larger kernels.

If the features are reasonably regularly spaced in the study region, then a fixed spatial kernel is appropriate for modelling. However, according to the previous discussion, the use of adaptive spatial kernels is recommended for irregularly sampled data (A. S. S. Fotheringham et al., 2002). Adaptive kernels increase the bandwidth size when the sample points are sparser and decrease its size when the sample points are denser.

The choice of distance metric is important to study these phenomena, and usually GWR uses Euclidean Distance to measure the "geographical proximity" of two different observations. On the other hand, there have been several attempts to use non-Euclidean Distance such as: a modified ward-to-ward distance matrix (Shuttleworth & Lloyd, 2005), or a spatial-temporal distance that takes into account not only the geographical location of an observation but also the time at which it was recorded. Besides this, Longley et al. showed in 2005 that a distance metric can depend on a number of different factors such as the presence of rivers between two locations, the quality of the road infrastructure, presence of notorious public spaces, etc. In conclusion, in order to choose a distance matrix, we need to take into account the study area's spatial context to see if a non-Euclidean Distance is appropriate.

There are also a few problems associated with this model. First of all, if there is global multicollinearity in the data, as it could be expected, both OLS and GWR models will not be able to produce reliable estimates. More likely than this is to exist local multicollinearity in the data, which is also a problem when trying to implement a GWR model. This characteristic in the spatial data can prevent the Akaike's Information Criterion (AIC) and Cross-Validation (CV Bandwidth) method in ArcGIS software from discovering the optimal distance or number of neighbours for the bandwidth, which may result in a wrong interpretation of the spatial patterns in the data.

3. METHODOLOGY

To study the percentage of voter turnout in municipalities of Portugal, we will use the ArcGIS software to produce the analyses of the following stages:

1. Data Collection;
2. ESDA;
3. OLS models;
4. GWR model;
5. Comparing results.

Each of these stages is detailed in the following sections.

3.1. DATA COLLECTION

The dependant variable of our study will be the abstention rate of the 2013 Portuguese municipal elections.

Taking into account different studies on this matter, we chose variables that have been shown to have some explanatory value regarding voter turnout and / or abstention rates. Therefore, using the websites PORDATA (http://www.pordata.pt) and Statistics Portugal (https://www.ine.pt), which contain data from various different subjects from several levels (European, Portuguese and Municipality), we were able to get all the variables that will be necessary to conduct the study. Taking into account that not all of the variables will be chosen for the final model, the initial set of variables that were chosen for the initial Exploratory Spatial Data Analysis are described in Table 1. All data have been collected at municipality level for continental Portugal.

Variable	Unit Measure	Description
Tx_abst	%	Abstention rate
Dens_pop	hab/km2	Population density
Tx_desemp	%	Unemployment rate

Variable	Unit Measure	Description
Crimes	#/1000hab	Number of crimes commited by 1000 residents
Sal_med	€	Average Salary
Dim_fam	#	Average family size
Per_Div	%	Divorce rate
Inactivos	%	Percentage of inactive residents compared to the active residents
Inactivos_hom	%	Percentage of inactive male residents compared to the active male residents
Inactivos_mulh	%	Percentage of inactive female residents compared to the active female residents
Ind_env	%	Age index (Number of people over 65 for every 100 people under 15)
Aloj_km2	#/km2	Average number of inhabited houses per km2
Poder_compra	%	Purchase power of each municipality regarding the country
Estr_UE	#/km2	Average number of EU foreigners per km2
Estr_outro	#/km2	Average number of non-EU foreigners per km2
Homens	#/km2	Average number of men per km2
Mulheres	#/km2	Average number of women per km2
Receitas	%	Revenue of the municipality divided by the expenses
Sem_escolaridade	%	Percentage of residents without any education
Ciclo1	%	Percentage of residents with a Primary Education
Ciclo2	%	Percentage of residents that completed the 2nd Cycle
Ciclo3	%	Percentage of residents with a Basic Education
Secundario	%	Percentage of residents with a Secundary Education
Superior	%	Percentage of residents with a Higher Education
Sect1	%	Percentage of residents working on the 1st Sector
Sect2	%	Percentage of residents working on the 2nd Sector
Sect3	%	Percentage of residents working on the 3rdSector
Hab18_29	%	Percentage of residents over 18 and under 29
Hab30_49	%	Percentage of residents over 30 and under 49

Variable	Unit Measure	Description
Hab50_69	%	Percentage of residents over 50 and under 69
Hab70_	%	Percentage of residents over 70

Table 1 – Initial variables of the study

3.2. EXPLORATORY SPATIAL DATA ANALYSIS

After converting our data in order for it to be compatible with the ArcGIS formats we make an ESDA. This analysis starts with correlation analysis and graphical and visual methods, such as using mapped data to try to find patterns, outliers, or clusters in some region of the map. Afterwards, statistics like the Global and Local Moran's I statistics, and the Getis-Ord General G and Getis-or Gi* statistic are used to evaluate if there is spatial autocorrelation and non-stationarity in the considered variables.

3.2.1. Global Moran's I statistic

The Global Moran's I statistic is used to measure spatial autocorrelation. Given a set of features associated to a certain attribute, it will help determine if the pattern in the data is clustered, dispersed or random.

In the software used to perform this study, there is a "Spatial Autocorrelation (Global Moran's I" tool that will calculate this statistic in the following manner:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{i,j} z_i z_j}{\sum_{i=1}^{n} {z_i}^2}$$

With z_i being the distance from the value of a variable for a given location i to its mean, $\omega_{i,j}$ being the spatial weight between location i and j, n being the total number of locations in the data and S_0 being the sum of all spatial weights.

This tool will also calculate the Moran's Index p-value and test statistic (z_I-score) as:

$$z_I = \frac{I - E(I)}{\sqrt{V(I)}}$$

With $E(I) = -1/(n-1)$ and $V(I) = E(I^2) - E(I)^2$.

Afterwards we will interpret, for each variable, the p-values and z_I-score with the following criteria, considering a 5% significance level:

- If the p-value is not statistically significant we cannot reject the null hypothesis. So it is quite possible that this variable has a random spatial distribution;
- If the p-value is statistically significant and the z_I-score is positive, then we reject the null hypothesis and can state that the data exhibits a clustered spatial distribution. This means that high [low] values of the given variable in a certain location, will be associated with high [low] values on neighbouring locations;
- If the p-value is statistically significant and the z_I-score is negative, then we reject the null hypothesis and can state that the data exhibits a dispersed spatial distribution. This means that high values of the given variable in a certain location, will be associated with low values on neighbouring locations, or vice-versa.

3.2.2. Local Moran's I statistic

The Local Moran's I statistic is similar to the Global Moran's I statistic in the sense that both assess spatial autocorrelation in the data. While the Global Moran's I statistic allows drawing a conclusion for the spatial pattern of the whole study area, the Local Moran's I statistic evaluates local patterns. Hence, the Local Moran's I statistic identifies spatial clusters where variables have high or low values (positive spatial autocorrelation), and spatial outliers where high values correlate with low neighboring values and vice versa (negative spatial autocorrelation).

In the software used to perform this study, there is a "Cluster and Outlier Analysis" tool that will calculate this statistic in the following manner:

$$I_i = \frac{x_i - \bar{X}}{{S_i}^2} \sum_{j=1, j \neq i}^{n} \omega_{i,j}(x_j - \bar{X})$$

With x_i being the value of a variable for a given location i, $\bar{X}$ being the mean of said variable, $\omega_{i,j}$ being the spatial weight between location i and j, n being the total number of locations in the data and S_i^2 being:

$$S_i^2 = \frac{\sum_{j=1, j\neq i}^{n}(x_j - \bar{X})^2}{n-1}$$

This tool will also calculate the local Moran's Index p-value and test statistic (z_I-score) as:

$$z_I = \frac{I_i - E(I_i)}{\sqrt{V(I_i)}}$$

With $E(I_i) = -\frac{\sum_{j=1, j\neq i}^{n}\omega_{i,j}}{n-1}$ and $V(I_i) = E\left(I_i^2\right) - E(I_i)^2$.

Afterwards we will interpret, for each variable, the p-values and z_I-score with the following criteria:

- If the p-value is not statistically significant we cannot reject the null hypothesis. So it is quite possible that this location is neither an outlier nor part of a cluster;
- If the p-value is statistically significant and the z_I-score is positive, then we reject the null hypothesis and can state that location i is part of a cluster of either high or low values;
- If the p-value is statistically significant and the z_I-score is negative, then we reject the null hypothesis and can state that location i is a spatial outlier (i.e. dissimilar values cluster together).

3.2.3. Getis-Ord General G statistic

The Getis-Ord General G statistic is used to measure the concentration of high or low values in a dataset for a given variable. Given a set of features associated to a certain attribute, it will help determine if the pattern in the data has cluster of high or low relative values.

In the software used to perform this study, there is a "High/Low Clustering (Getis-Ord General G)" tool that will calculate this statistic in the following manner:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{i,j}\,x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n}x_i x_j} \quad , \quad \forall j \neq i$$

With x_i and x_j being the value of a variable in locations i and j, $\omega_{i,j}$ being the spatial weight between location i and j, n being the total number of locations in the data.

This tool will also calculate the Getis-Ord General G p-value and test statistic (z_G-score) as:

$$z_G = \frac{G - E(G)}{\sqrt{V(G)}}$$

With $E(G) = \frac{\sum_{j=1, j\neq i}^{n}\omega_{i,j}}{n(n-1)}$ and $V(G) = E(G^2) - E(G)^2$.

Afterwards we will interpret, for each variable, the p-values and z_G-score with the following criteria:

- If the p-value is not statistically significant we cannot reject the null hypothesis. So it is quite possible that we cannot distinguish the possible clusters of these variables from being of high or low values;
- If the p-value is statistically significant and the z_G-score is positive, then we reject the null hypothesis and can state that this variable exhibits a clustered spatial distribution of high values;
- If the p-value is statistically significant and the z_G-score is negative, then we reject the null hypothesis and can state that this variable exhibits a clustered spatial distribution of low values.

3.2.4. Getis-Ord Gi* statistic

The Getis-Ord Gi* statistic is used to find a given location in the dataset belongs to a high or low value cluster or hotspot.

In the software used to perform this study, there is a “Hot Spot Analysis” tool that will calculate this statistic in the following manner:

$$G_i^* = \frac{\sum_{j=1}^{n} \omega_{i,j} x_j - \bar{X} \sum_{j=1}^{n} \omega_{i,j}}{S \sqrt{\frac{[n \sum_{j=1}^{n} \omega_{i,j}{}^2 - (\sum_{j=1}^{n} \omega_{i,j})^2]}{n-1}}}$$

With x_j being the value of a variable for a given location j, $\bar{X}$ being the mean of said variable, $\omega_{i,j}$ being the spatial weight between location i and j, n being the total number of locations in the data and S being:

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j{}^2}{n} - \bar{X}^2}$$

The G_i^* statistic is a z-score so no further calculations are required.

Afterwards we will interpret, for each variable, the p-values and G_i^* statistic value with the following criteria:

- If the p-value is not statistically significant we cannot reject the null hypothesis. So it is quite possible that this location does not belong in a cluster;
- If the p-value is statistically significant and the G_i^* statistic is positive, then we reject the null hypothesis and can state that location i is part of a cluster of high values;
- If the p-value is statistically significant and the G_i^* statistic is negative, then we reject the null hypothesis and can state that location i is part of a cluster of low values.

3.3. Ordinary Least Squares Models

After the ESDA, we will, iteratively, use different subsets of the explanatory variables to find the OLS model that will best fit the data. We need to take into account that, due to the nature of the data, it is expected to exist spatial autocorrelation in the data. Moreover, taking into consideration other OLS assumptions, we will perform the following tests to diagnose the models:

- Jarque-Bera test to test the normality of the residuals (Bai & Ng, 2005)

- Koenker test to test the heteroscedasticity of the residuals (Koenker & Bassett, 1982)
- Variance Inflation Factor (VIF) to find multicollinearity in the data (O'Brien, 2007); VIF values under 7,5 indicate no multicollinearity between the model's explanatory variables
- Robust t-tests to assess the significance of the explanatory variables, instead of the usual t-tests that may not be trustworthy due to heteroscedasticity (provided by the ArcGIS software)
- Wald test to assess the significance of the model (Wald, 1943), instead of the usual F-test that may not be trustworthy due to heteroscedasticity

After testing some models, we choose the explanatory variables subset that had the highest Adjusted Akaike's Information Criterion (AICc) while passing the tests previously stated. This subset will be used to perform the GWR model.

3.4. GEOGRAPHICALLY WEIGHTED REGRESSION MODEL

In this step, we will estimate a GWR model using ArcGIS. This software will construct a separate equation for every observation in the dataset incorporating the dependent and explanatory variables of observations falling within the bandwidth of each target observation.

First of all, the dependent variable is modeled as a linear function of a set of explanatory variables:

$$y_i = \alpha_0 + \sum_{k=1}^{m} \alpha_k x_{ik} + \varepsilon_i$$

In this case, y_i is the i-th value of the dependent variable, x_{ik} is the i-th value of the k-th explanatory variable, ε_i is the residual associated to the i-th value which should be normally distributed with mean 0 and constant variance, α_0 is the intercept of the model and α_k is the regression coefficient of the model corresponding to variable k (i.e. the OLS estimator). Considering $\hat{\alpha}$ to be a vector of the k+1 coefficients of the model,

$$\hat{\alpha} = (X^T X)^{-1} X^T Y$$

With Y being the vector of the observations of the dependent variable and X being the matrix with i rows and k columns which will correspond to the i values of the k explanatory variables. These estimates can be seen as the "change rate" between each explanatory variable and the dependent variable. This means that a change of value δ in variable k will impact the dependent variable by α_k.

Afterwards we can define the basic GWR model as follows (Brunsdon et al., 1998; A. S. S. Fotheringham et al., 2002):

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i$$

With y_i being the dependent variable at geographic location i, x_{ik} being the k-th independent variable at the same location i of the dependent variable, β_{ik} being the regression coefficient of the kth variable at location i, with β_{i0} being the intercept. Lastly, ε_i is the random residual at location i.

To estimate the different weights for each variable at each given location, the following expression is used:

$$\hat{\beta}_i = (X^T W_i X)^{-1} X^T W_i y$$

With X being the matrix of the independent variables with the first column being filled with 1s for the intercept, y being the dependent variable vector, $\hat{\beta}_i$ being the vector of $m + 1$ local coefficients (with m being the number of dependent variables) and W_i being the weight matrix which will be a diagonal matrix denoting the geographical weighting of each observation for location i associated to every other location in the study area.

The weighting matrix is determined by a 'Gaussian-like' kernel method, which will calculate the proximities between location i and all the other data points using the following expression,

$$w_{ij} = e^{-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2}$$

In this case, we use d_{ij} as the distance between location i and j which will be calculated with the Euclidean Distance between the centroids of both locations, and b as the kernel bandwidth. We used an adaptive kernel as recommended, and the optimal number of neighbours was determined based on the results of the Adjusted Akaike's Information Criterion (AICc) obtained using cross-validation. While the Akaike's Information Criterion (AIC) only measures the model's precision, the AICc also accounts for the trade-off between prediction accuracy and complexity which is commonly referred to as the model parsimony.

In 1998, Hurvich, Simonoff, & Tsai extended the AIC by making it also a function of the sample's size. This new criterion was named Adjusted Akaike's Information Criterion (AIC_C). In the GWR model the AIC_C is found by the following expression:

$$AIC_c(b) = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left(\frac{tr(S)}{n-2-tr(S)}\right)$$

With n being the sample size, $\hat{\sigma}$ being the estimated standard deviation of the residuals and S being the projection matrix from the observed values (y) of the dependent variable to the fitted values $(\hat{y})$. For the case of GWR, each row of this matrix is calculated as,

$$r_i = X_i(X^T W_i X)^{-1} X^T W_i$$

With X_i being the i-th row of matrix X.

GWR has only a limited number of diagnosing tools, because it is unclear what statistical tests can reliably diagnose model problems (Páez, Farber, & Wheeler, 2011). Local multicollinearity will be assessed using the Condition Number provided by ArcGIS. If all municipalities have this metric's value below 30, then there is evidence of no local multicollinearity. Low values of the coefficient standard errors provide evidence of a good reliability of parameter estimates.

Besides from the AICc we will also compute the following statistics to find the best GWR model:

- Sum of Residual Squares (the lower the value, the better the fit of the model)

$$\sum_{i=1}^{n} \varepsilon_i^{\,2}$$

- Sigma (the lower the estimated standard deviation of the residuals, the better the fit of the model)

$$\hat{\sigma}$$

- Adjusted R^2 (the higher the value, the better the fit of the model relatively to other models)

In the end, we also examine the residuals obtained from the GWR model keeping in mind that they should exhibit a random pattern. Moreover, the Global Moran's I statistic of the residuals is computed and if the value of this statistic is non-significant, then we can conclude that the residuals exhibit a random pattern.

3.5. COMPARING MODELS

To compare the results of both OLS and GWR models with the aim of finding which one provided the better fit for our data, we use the Adjusted R^2 or the Adjusted Akaike Information Criterion (AICc).

4. RESULTS AND DISCUSSION

4.1. EXPLORATORY SPATIAL DATA ANALYSIS

Figure 1 shows how the abstention rate (Tx_abst) in the 2013 Portuguese municipal elections varied from municipality to municipality. The highest rates are located in municipalities closer to the ocean, whereas the lowest values were recorded in the interior.

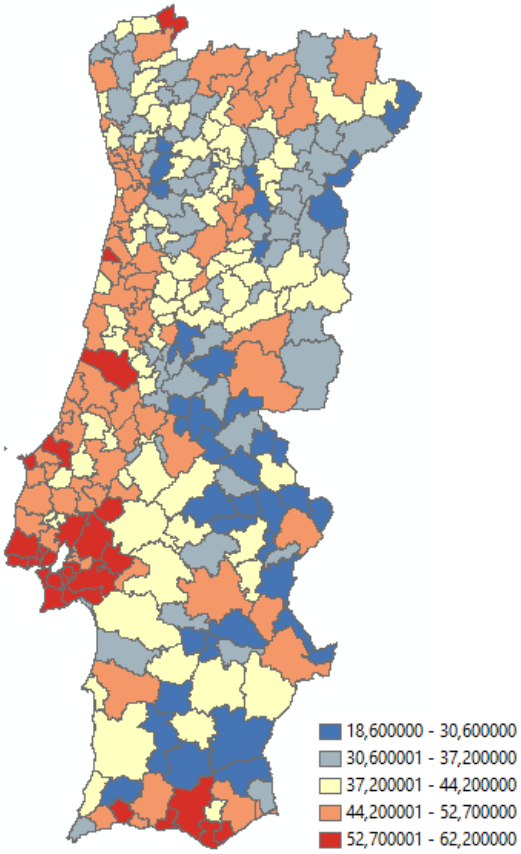


Figure 1 – Distribution of abstention rate

We calculated the correlation matrix of the variables in our dataset (Appendix A) and got the following results:

- Variables Aloj_km2, Homens and Mulheres are heavily correlated with Dens_pop, so we will not include them in our further analysis and we will keep Dens_pop;
- Variables Inativos, Inativos_hom and Inativos_mulh, Ind_env and Sem_escolaridade are heavily correlated with Hab70_, so we will not include them in our further analysis and we will keep Hab70_ (percentage of residents over 70);
- Variable Secundario is heavily correlated with Ciclo1 and since it has a higher correlation with our dependent variable, we will keep variable Secundario and exclude Ciclo1 from our further analysis;

- Variable Dim_fam is heavily correlated with Ciclo2 and since it has a higher correlation with our dependent variable, we will keep variable Dim_fam and exclude Ciclo2 from our further analysis;

Afterwards we computed the Global Morans' I and the Getis-Ord General G statistics which rendered the following p-values and z-scores:

Variables	Z-score	P-value
Tx_abst	24,9	< 0,0001
Dens_pop	34,2	< 0,0001
Tx_desemp	18,0	< 0,0001
Crimes	13,1	< 0,0001
Sal_Med	23,2	< 0,0001
Dim_fam	52,7	< 0,0001
Per_Div	10,2	< 0,0001
Poder_compra	24,4	< 0,0001
Estr_UE	20,8	< 0,0001
Estr_outro	34,1	< 0,0001
Receitas	9,5	< 0,0001
Ciclo3	27,1	< 0,0001
Secundario	37,8	< 0,0001
Superior	18,6	< 0,0001
Sect1	30,6	< 0,0001
Sect2	41,8	< 0,0001
Sect3	40,2	< 0,0001
Hab18_29	27,2	< 0,0001
Hab30_49	27,6	< 0,0001
Hab50_69	27,1	< 0,0001
Hab70_	25,1	< 0,0001

Table 2 – Global Moran's Analysis

Variables	Z-score	P-value
Tx_abst	2,0	0,0492
Dens_pop	25,4	< 0,0001
Tx_desemp	1,3	0,1810
Crimes	-4,8	< 0,0001
Sal_Med	2,5	0,0134
Dim_fam	10,8	< 0,0001
Per_Div	-0,8	0,3977
Poder_compra	20,4	< 0,0001
Estr_UE	19,2	< 0,0001
Estr_outro	34,2	< 0,0001
Receitas	3,7	0,0002
Ciclo3	0,0	0,9905
Secundario	-2,5	0,0133
Superior	5,0	< 0,0001
Sect1	2,7	0,0069
Sect2	13,0	< 0,0001
Sect3	-5,8	< 0,0001
Hab18_29	21,1	< 0,0001
Hab30_49	21,2	< 0,0001
Hab50_69	20,3	< 0,0001
Hab70_	18,1	< 0,0001

Table 3 – Getis-Ord General G Analysis

From these results we can draw the following conclusions:

- For the Global Moran's I statistics, since all p-values are statistically significant, we can consider with a very high level of certainty that all the variables do not exhibit a random spatial distribution. And, because all z-scores are positive, we may conclude that the spatial distribution of high values and/or low values in all variables is spatially clustered.

- For the Getis-Ord General G statistics, we can conclude with a confidence of 95% that nearly every variable (every one with a p-value lower than 5%) exhibits a spatially clustered pattern with clusters being of either high or low vlaues. Actually, we cannot

reject the null hypothesis only for Tx_desemp, Per_Div and Ciclo 3. For all other variables, those that have positive z-scores have clusters of high values and, on the other hand, those that have negative z-scores have clusters of low values.

Now, we will examine for each municipality what kind of pattern is exhibited in each value's neighborhood. In order to do that, we computed the Local Moran's I Index and the Getis-Ord Gi* statistic for every municipality in our dataset. For the dependent variable Tx_abst, we had the following results:

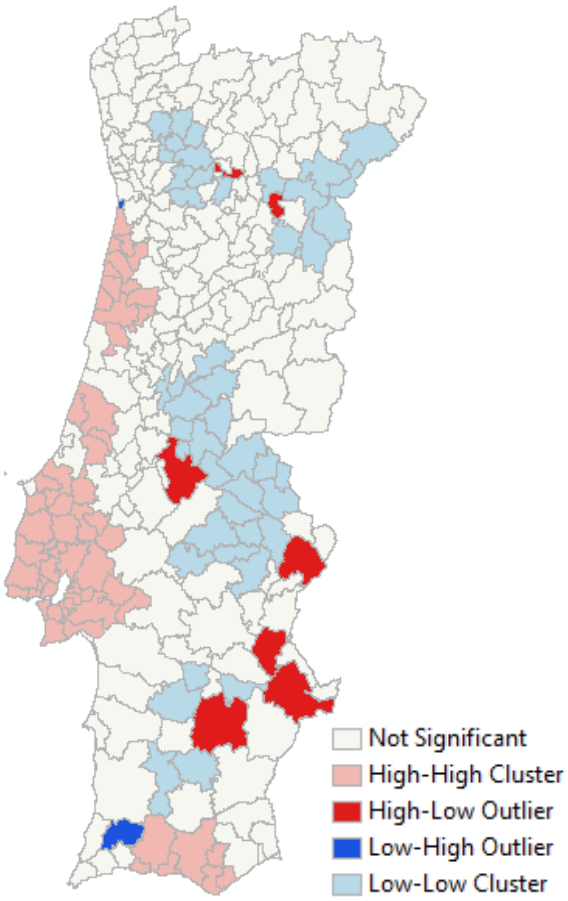


Figure 2 – Local Moran's I of abstention rate

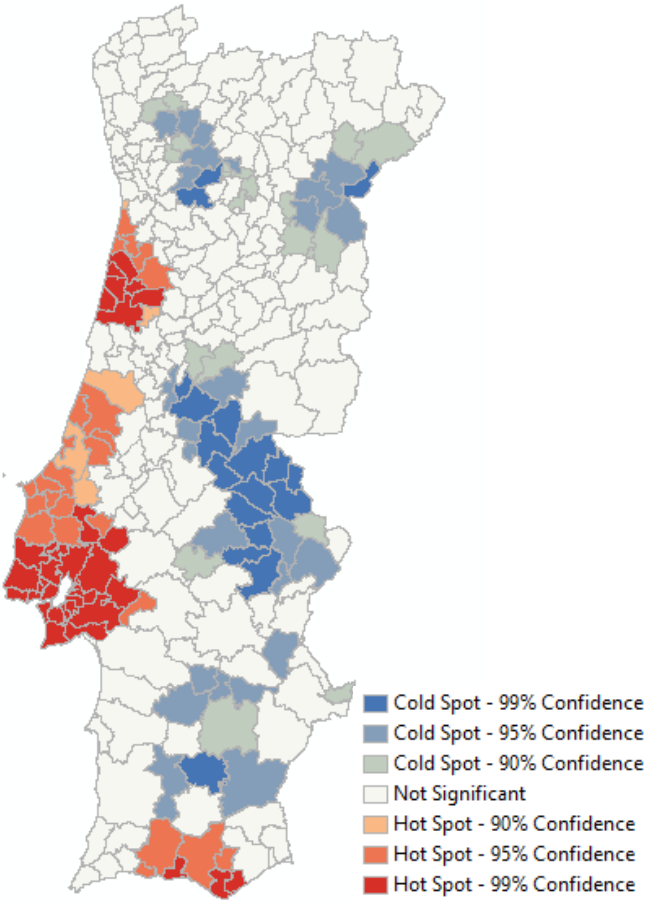


Figure 3 – Getis-Ord Gi* of abstention rate

So, taking into account that colored municipalities are the ones that returned a statistically significant p-value for the corresponding test, we can conclude the following:

- From the Local Moran's I, we can conclude that the brightly red colored municipalities (Abrantes, Beja, Elvas, Moura, Penedono, Peso da Régua and Reguengos de Monsaraz) correspond to spatial outliers of high abstention rates surrounded by municipalities of low rates (significant negative spatial

autocorrelation). We also conclude that the blue brightly colored municipalities (Espinho and Monchique) correspond to spatial outliers of low abstention rates surrounded by municipalities of high rates (significant negative spatial autocorrelation). Municipalities with soft red or soft blue colors are spatial clusters, thus having a local pattern of significant positive spatial autocorrelation. Coastal municipalities colored with soft red have a high value surrounded primarily by high values, whereas those inland ones colored with soft blue have a low value surrounded primarily by low values.

- From the Getis-Ord Gi*, we conclude that municipalities that are colored in orange/red (Aveiro, Faro, Lisboa and Setúbal), are *hotspots* of high abstention rates. On the other hand, municipalities that are colored blue (Beja, Bragança, Castelo Branco, Guarda, Portalegre and Porto) are *coldspots* of low abstention rates.

These analyses helped us to have a better understanding of the spatial patterns in the dataset, which will be helpful in the next steps of the study and the interpretation of future results.

4.2. ORDINARY LEAST SQUARES MODELS

In this step, we made exploratory regressions using the "Exploratory Regression" tool in ArcGIS. Using an iterative process, we chose subsets of variables in order to find which variables would provide a better model, not only considering the model's comparison metric (AICc), but also the results of the diagnosing tests as described in the Methodology section.

The variables that had better explanatory power regarding the dependent variable were:

Positively significant:

- Percentage of residents with a Secundary Education (Secundario)
- Number of crimes commited by 1000 residents (Crimes)
- Revenue of the municipality divided by the expenses (Receitas)
- Percentage of residents with a Higher Education (Superior)
- Percentage of residents over 50 and under 69 (Hab50_69)
- Divorce rate (Per_Div)

- Average number of non-EU foreigners per km^2 (Estr_outro);

Negatively significant:

- Age index: number of people over 65 for every 100 people under 15 (Ind_env)
- Percentage of residents without any education (Sem_escolaridade)

This means that, for most of the models created during each iteration of the exploratory analysis, an increase in the values of the first subset of variables would cause a significant increase in the abstention rate. On the other hand, an increase in the values of the negatively significant variables would cause a significant decrease in the abstention rate.

Of all the models that were tested, one stood out regarding the criteria that were stated before. The model was the following:

$$Tx\_abst = 10{,}84 + 0{,}14 * Crimes + 0{,}03 * Per\_Div - 0{,}02 * Ind\_env + 0{,}09 * Receitas +$$

$$+ 0{,}91 * Secundário + 0{,}45 * Superior$$

- Explanatory variables:
  - Number of crimes commited by 1000 residents (Crimes)
  - Divorce rate (Per_Div)
  - Age index: number of people over 65 for every 100 people under 15 (Ind_env)
  - Revenue of the municipality divided by the expenses (Receitas)
  - Percentage of residents with a Secundary Education (Secundario)
  - Percentage of residents with a Higher Education (Superior)

- Statistic/p-values:
  - Adjusted R^2: 50%
  - Corrected Akaike´s Information Criterion (AICc): 1.837,45
  - Variance Inflation Factor (VIF): 2,98
  - Jarque-Bera p-value: 0,37
  - Koenker (BP) statistic p-value: 0,07
  - Global Moran's I p-value: 0,00

Since the VIF value was smaller than 7,5 there is no multicollinearity. The Jarque-Bera test allows to conclude that there is evidence that the residuals of this model are normally distributed, for the usual significance levels. The p-value of the Koenker (BP) statistic indicates that there is evidence of heteroscedasticity at significance levels greater than 7%. The Global Moran's I test allows to conclude that the residuals have spatial autocorrelation, at any significance level.

So, following this analysis, since we cannot trust the results of OLS due to the spatial autocorrelation and non-stationarity in our data, we will investigate the GWR model that uses these explanatory variables.

4.3. GEOGRAPHICALLY WEIGHTED REGRESSION MODEL

In this step of the study, we used the variables from the best OLS model found and, iteratively, tried every possible subset of variables as the set of explanatory variables for a GWR model. After all possible combinations, we registered in Table 4 the six best models for each possible number of explanatory variables (from 1 to 6), and their results are detailed in Table 5.

Explanatory Variables	Crimes	Per_Div	Ind_env	Receitas	Secundário	Superior
Model 1						**X**
Model 2			**X**			**X**
Model 3	**X**		**X**			**X**
Model 4	**X**	**X**	**X**			**X**
Model 5	**X**	**X**	**X**		**X**	**X**
Model 6	**X**	**X**	**X**	**X**	**X**	**X**

Table 4 – Explanatory variables used for the best GWR models for each number of variables

Statistics	Neighbours	Sum of Residual Squares	Effective Number of Coefficients	Sigma	AICc	Adjusted R^2
Model 1	31	6.558	57	5,45	1.774,58	64%
Model 2	35	5.742	67	5,21	1.760,62	67%
Model 3	58	6.349	57	5,35	1.766,36	66%
Model 4	98	7.178	43	5,53	1.771,12	63%
Model 5	272	10.278	12	6,20	1.815,27	53%
Model 6	Local Multicolinearity Error					

Table 5 – Results for the GWR models

So, first of all we can conclude that variable Receitas exhibits Local Multicollinearity, which makes it impossible to create a GWR model using this variable. Then, examining the AICc of the models we see that the model which is a better fit for our data is Model 2 with just 2 explanatory variables (Ind_env and Superior). As expected this model also has the lowest Sum of Residual Squares and Sigma, and the highest Adjusted R^2.

To analyse the Standardized Residuals of this model, we started by mapping them:

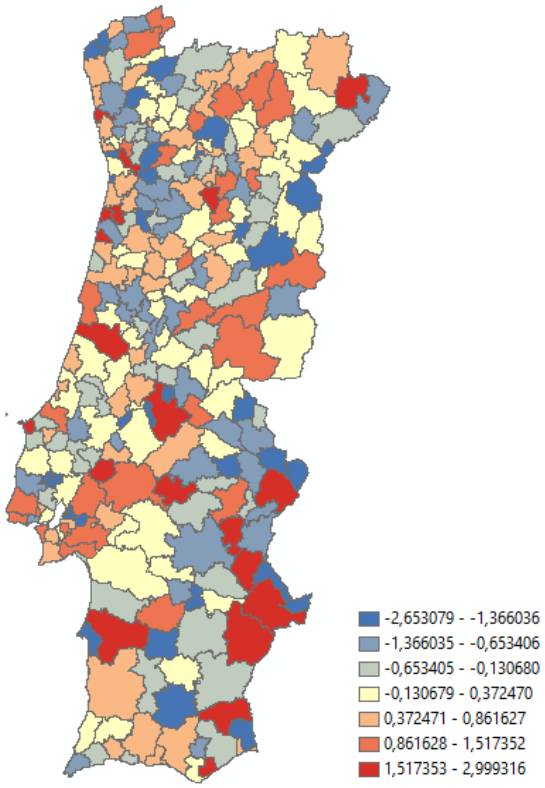


Figure 4 – Standardized Residuals of the GWR model

As we can see, there are no clear areas with clustered high or low values, so there is evidence that the residuals have a random pattern, thus the model is well-specified and it is not missing any key explanatory variables. To make sure that this conclusion was valid, we measured the Global Moran's statistic for the residuals and had the following results:

- Moran's Index: 0,0132
- Z-score: 0,58
- P-value: 0,56

So we can in fact say that the residuals exhibit a random spatial pattern throughout the data which suggests that this model is able to predict the spatial pattern of the dependent variable.

Figure 5, shows the Local R^2 for each municipality, which is a measure of the variability of the abstention rate that is explained by the local model. We can use this to understand where the model will render more explanatory power.

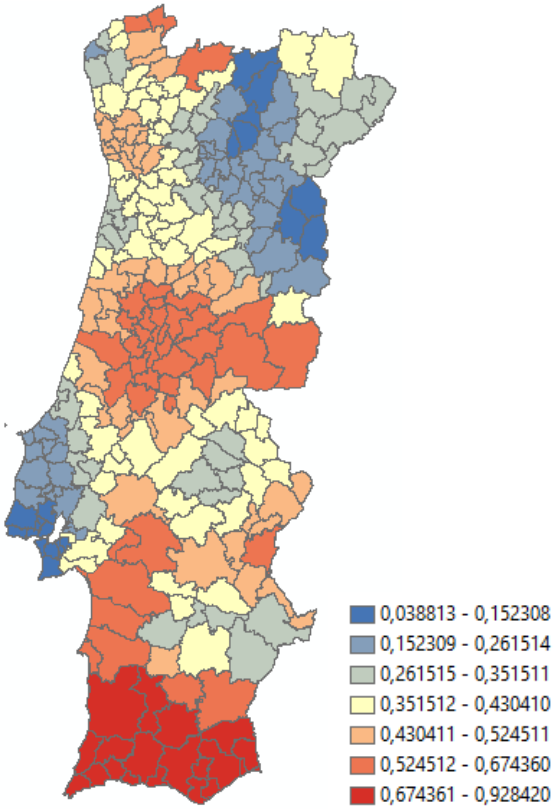


Figure 5 – Local R^2

So, we can conclude that for the regions of Algarve and Alentejo, and for districts such as Castelo Branco, Coimbra, Leiria, Porto, Braga and Viana do Castelo, the local models will better explain the value of the abstention rate. On the other hand, the variability of the abstention rate explained by the local models in municipalities such as Lisbon, Guarda, Viseu, Vila Real and Bragança is much smaller.

Now, we will examine how each explanatory variables influences the abstention rate by mapping the regression coefficients for each variable:

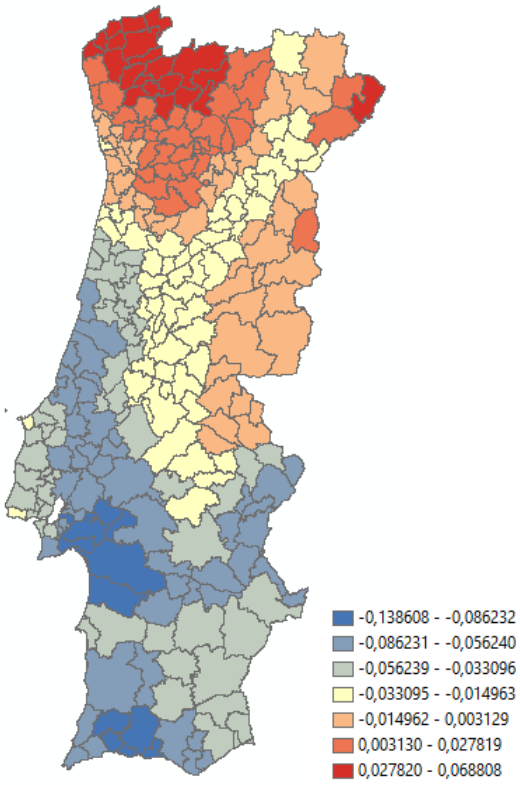


Figure 6 – Coefficients for Ind_env

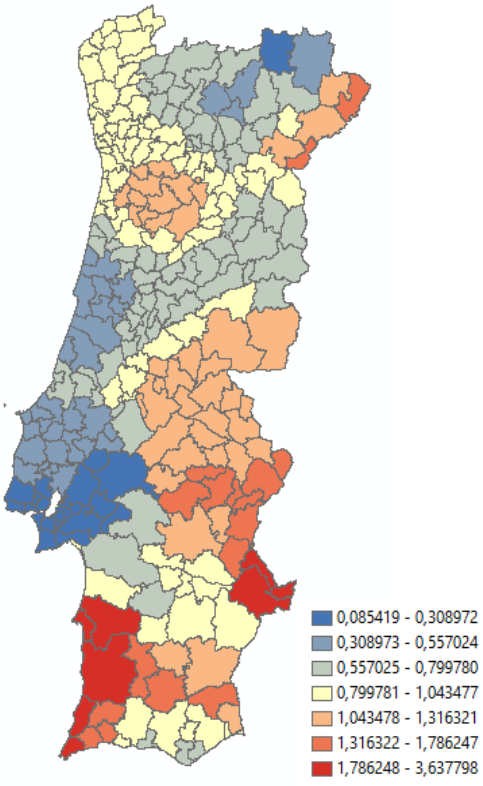


Figure 8 – Coefficients for Superior

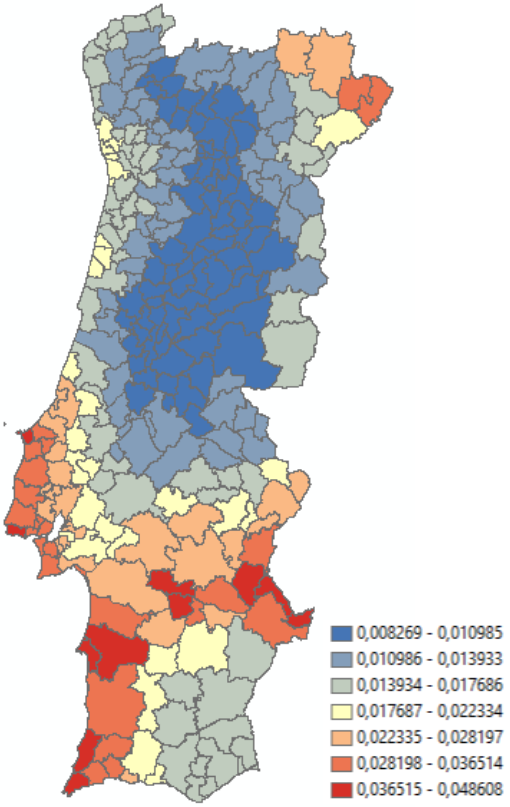


Figure 7 – Standard Error of the Coefficients for Ind_env

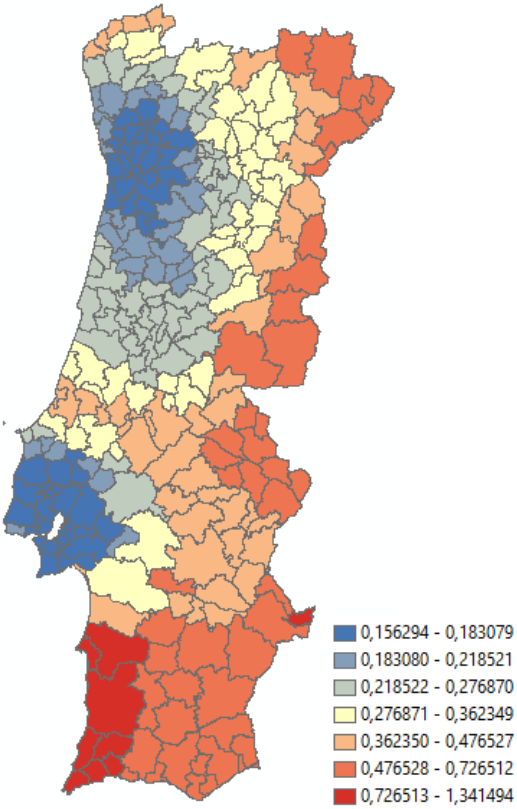


Figure 9 - Standard Error of the Coefficients for Superior

First of all, looking at the coefficients for Ind_env (Age index: number of people over 65 for every 100 people under 15), we can see that in the northern districts of Portugal, excluding Coimbra, an older population will result in a higher abstention rate. On the other hand, in the rest of the country, and especially in the districts of Setúbal and Faro, an older population will result in a lower abstention rate. Adding to this, we can say that, according to the Standard error of the coefficients for Ind_env, parameter estimates for locations to the north of Lisbon will be more reliable than those for the south.

Looking at the coefficients for Superior (Percentage of residents with a Higher Education), we can see that in a higher percentage of the population with a University degree will generally result in a higher abstention rate, particularly in the districts of Beja, Castelo Branco, Évora, Faro, Portalegre and Viseu. We can also say that, looking at the Standard error of the coefficients for Superior, parameter estimates for municipalities close to Lisbon and Porto will be more reliable when compared to other municipalities.

Finally, we mapped the values for the abstention rate predicted by the model and compare them to the actual values to evaluate the goodness of fit of the model (Figures 6 and 7).

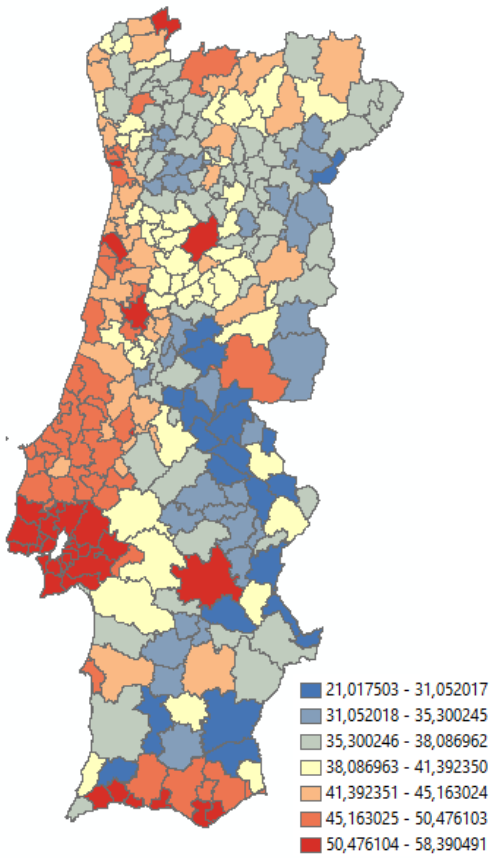


Figure 10 – Predicted values

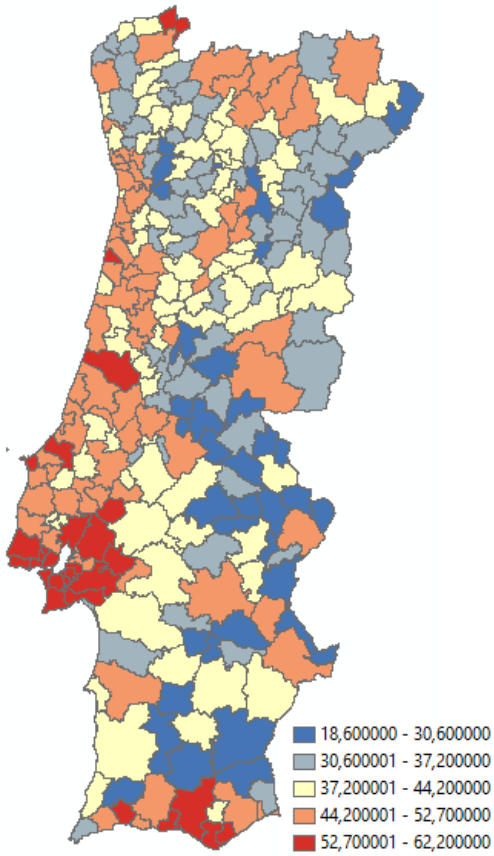


Figure 11 – Actual values

As we can see, in both maps we have a similar spatial pattern and similar values for the abstention rate, which is not surprising considering the spatial pattern of the residuals (Figure 3).

4.4. Comparing Models

Now, we will compare the results of our best OLS model to our best GWR model, based on the AICc and the Adjusted R^2. The best OLS model had an Adjusted R^2 of 50%, and the AICc was equal to 1.837,45. Considering the best GWR model, the Adjusted R^2 was equal to 67% and the AICc was equal to 1.760,62.

Accordingly, the GWR model rendered far better results than the OLS model, since it had a much higher Adjusted R^2 and lower AICc, which indicates a real improvement in model performance. Moreover, the OLS model is unreliable since its residuals are spatially autocorrelated, and there is evidence of non-stationarity. Therefore, we can conclude that

the GWR model not only is more appropriate than the OLS model, but it also provides a far better fit regarding the abstention rate.

5. CONCLUSION

We started this study with a research question aiming to find out if we should use spatial regression models such as GWR, instead of classical regression models such as OLS, to model the abstention rate of the last presidential elections in Portugal.

In fact, we found that the abstention rate exhibited a non-random spatial distribution and that a GWR model is, in fact, a better approach to model this behavior when compared to a regular OLS regression model. Moreover, both metrics commonly used to compare the goodness of fit of models, namely the AICc and the Adjusted R^2, provided better results for the GWR model than for the best OLS regression model found.

From all the possible explanatory variables of the study, we found that variables related to the population education level, age and income had a significant impact in the estimation of the abstention rate for each municipality using OLS.

We concluded that the best model to estimate the abstention rate is a GWR model with the percentage of individuals with a higher education (Superior), and the number of individuals over 65 years old for every 100 individuals under 15 (Ind_env) as the explanatory variables. Using a logistic regression model, Freire & Magalhães (2002, p. 149) concluded that age was the most relevant factor affecting the probability of an elector to vote in the presidential elections of 2001 in Portugal, but the degree of education was not statistically significant. However, it is important to note that variables affecting voting turnout might not be *exactly* the same as those affecting the abstention rate. Our results show that the aging index (Ind_env) will have a higher explanatory power regarding the abstention rate in regions to the north of Lisbon especially in areas further away from the ocean. On the other hand, the percentage of individuals with a higher education will render more reliable estimates in municipalities around higher populated districts such as Lisbon and Porto.

5.1. LIMITATIONS

Examining other studies that tried to model the abstention rate in Portuguese elections, we were able to identify several limitations with our study. The main limitation is the lack of

available explanatory variables, that were proven to have a high explanatory value regarding the voting turnout in previous studies (e.g. Freire & Magalhães, 2002 and Martins & Veiga, 2013). Variables such as the degree of confidence in democratic institutions, and other political variables were not found at a municipality level, and therefore were not used in this study.

5.2. FUTURE WORK

Future work regarding this study might include a wider range of explanatory variables. Mainly, the inclusion of political values at a municipality level, such as the number of political parties, the degree of confidence in democratic institutions, the number of consecutive terms for the governing political party and others. These variables were previously shown to have some explanatory variable regarding the voting turnout (Martins & Veiga, 2013) and therefore might have enriched this study.

6. REFERENCES

Anselin, L. (1998). Exploratory spatial data analysis in a geocomputational environment. In *GeoComputation* (pp. 77–94).

Bai, J., & Ng, S. (2005). Tests for skewness, kurtosis, and normality for time series data. *Journal of Business and Economic Statistics*. https://doi.org/10.1198/073500104000000271

Benassi, F., & Naccarato, A. (2017). Households in potential economic distress. A geographically weighted regression model for Italy, 2001–2011. *Spatial Statistics*, 1–15. https://doi.org/10.1016/j.spasta.2017.03.002

Blais, A. (2006). WHAT AFFECTS VOTER TURNOUT? *Annual Review of Political Science*. https://doi.org/10.1146/annurev.polisci.9.070204.105121

Blais, A., & Dobrzynska, A. (2009). Turnout in electoral democracies revisited. In *Activating the Citizen: Dilemmas of Participation in Europe and Canada* (pp. 63–82). https://doi.org/10.1057/9780230240902_4

Blais, A., Pilet, J. B., Van der Straeten, K., Laslier, J. F., & Héroux-Legault, M. (2014). To vote or to abstain? An experimental test of rational calculus in first past the post and PR elections. *Electoral Studies*, *36*, 39–50. https://doi.org/10.1016/j.electstud.2014.07.001

Brunsdon, C., Fotheringham, S., Charlton, M., Brunsdont, C., & Chariton, M. (1998). Geographically Weighted Regression-Modelling Spatial Non-Stationarity. *Source Journal of the Royal Statistical Society. Series D (The Statistician) Journal of the Royal Statistical Society. Series D The Statistician*, *47*(3), 431–443. https://doi.org/10.1111/1467-9884.00145

Burden, B. C. (2000). Voter Turnout and the National Election Studies. *Political Analysis*, *8*(4), 389–398. https://doi.org/10.1093/oxfordjournals.pan.a029823

Cancela, J., & Geys, B. (2016). Explaining voter turnout: A meta-analysis of national and subnational elections. *Electoral Studies*. https://doi.org/10.1016/j.electstud.2016.03.005

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

Fornos, C. A., Power, T. J., & Garand, J. C. (2004). Explaining voter turnout in Latin America, 1980 to 2000. *Comparative Political Studies*. https://doi.org/10.1177/0010414004267981

Fotheringham, A. S., Charlton, M., & Brunsdon, C. (1997). Two techniques for exploring non-stationarity in geographical data. *Geographical Systems*, *4*(1), 59–82. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0001952078&partnerID=40&md5=9bc05012e62750994ea12d29c80df58d

Fotheringham, A. S. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. *Notes*. https://doi.org/10.1111/j.1538-4632.2003.tb01114.x

Franklin, M. N. (2004). *Voter turnout and the dynamics of electoral competition in established democracies since 1945*. *Voter Turnout and the Dynamics of Electoral Competition in Established Democracies since 1945*. https://doi.org/10.1017/CBO9780511616884

Freire, A., & Magalhães, P. (2002). A abstenção eleitoral em Portugal.

Geys, B. (2006). Explaining voter turnout: A review of aggregate-level research. *Electoral Studies*. https://doi.org/10.1016/j.electstud.2005.09.002

Hayashi, F. (2000). Econometrics. 2000. *Princeton University Press. Section*.

Hooghe, M., & Kern, A. (2017). The tipping point between stability and decline: Trends in voter turnout, 1950-1980-2012. *European Political Science*. https://doi.org/10.1057/s41304-016-0021-7

Hurvich, C. M., Simonoff, J. S., & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. https://doi.org/10.1111/1467-9868.00125

Jackman, R. W. (1987). Political Institutions and Voter Turnout in the Industrial Democracies. *The American Political Science Review*, *81*(2), 405. https://doi.org/10.2307/1961959

Jackman, R. W., & Miller, R. A. (1995). Voter turnout in the industrial democracies during the 1980s. *Comparative Political Studies*, *27*(4), 467–492. https://doi.org/10.1177/0010414095027004001

Koenker, R., & Bassett, G. (1982). Robust Tests for Heteroscedasticity Based on Regression Quantiles. *Econometrica*. https://doi.org/10.2307/1912528

Kostelka, F. (2017). Does democratic consolidation lead to a decline in voter turnout? Global evidence since 1939. *American Political Science Review*. https://doi.org/10.1017/S0003055417000259

Lahtinen, H., Mattila, M., Wass, H., & Martikainen, P. (2017). Explaining Social Class Inequality in Voter Turnout: The Contribution of Income and Health. *Scandinavian Political Studies*. https://doi.org/10.1111/1467-9477.12095

Ley, S. (2017). To Vote or Not to Vote. *Journal of Conflict Resolution*, 002200271770860. https://doi.org/10.1177/0022002717708600

Lijphart, A. (2007). Unequal participation: Democracy's unresolved dilemma. In *Thinking about Democracy: Power Sharing and Majority Rule in Theory and Practice* (pp. 201–231). https://doi.org/10.4324/9780203934685

Martins, R., & Veiga, F. J. (2013). Economic performance and turnout at national and local elections. *Public Choice*, *157*(3–4), 429–448. https://doi.org/10.1007/s11127-012-0047-5

Nannestad, P., & Paldam, M. (1994). The VP-function: A survey of the literature on vote and popularity functions after 25 years. *Public Choice*, *79*(3–4), 213–245. https://doi.org/10.1007/BF01047771

Norris, P. (2004). *Electoral engineering: Voting rules and political behavior*. *Electoral Engineering: Voting Rules and Political Behavior*. https://doi.org/10.1017/CBO9780511790980

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*. https://doi.org/10.1007/s11135-006-9018-6

Páez, A., Farber, S., & Wheeler, D. (2011). A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A*. https://doi.org/10.1068/a44111

Paldam, M. (2008). Vote and popularity functions. In *Readings in Public Choice and Constitutional Political Economy* (pp. 533–550). https://doi.org/10.1007/978-0-387-75870-1_29

Poole, M. A., & O'Farrell, P. N. (1971). The Assumptions of the Linear Regression Model. *Transactions and Papers, The Institute of British Geographers*. https://doi.org/10.2307/621706

Powell, G. B. (1986). American Voter Turnout in Comparative Perspective. *American Political Science Review*, *80*(01), 17–43. https://doi.org/10.2307/1957082

Shuttleworth, I. G., & Lloyd, C. D. (2005). Analysing average travel-to-work distances in Northern Ireland using the 1991 Census of Population: The effects of locality, social composition, and religion. *Regional Studies*. https://doi.org/10.1080/00343400500289895

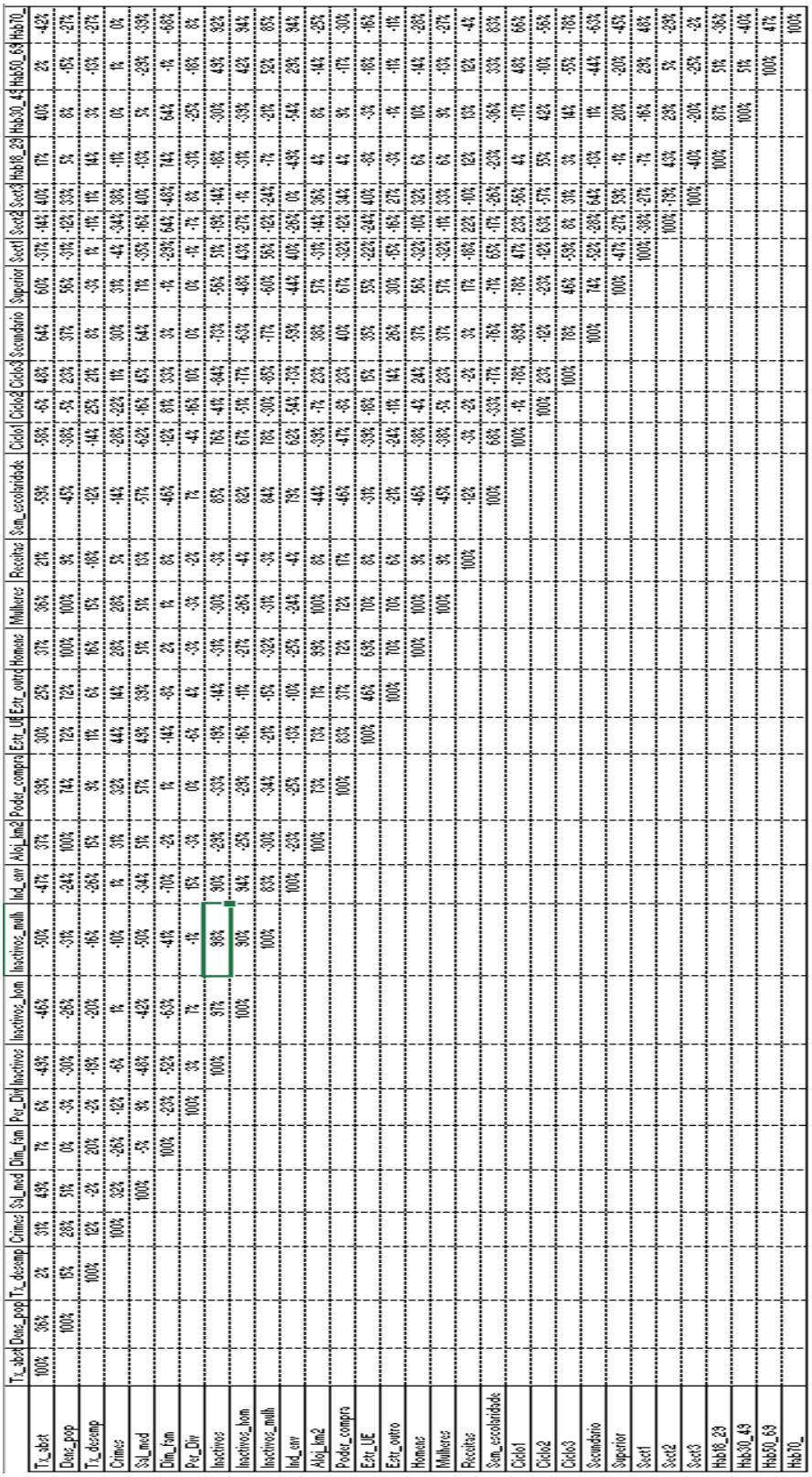
Sundström, A., & Stockemer, D. (2015). Regional variation in voter turnout in Europe: The impact of corruption perceptions. *Electoral Studies*. https://doi.org/10.1016/j.electstud.2015.08.006

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*. https://doi.org/10.1090/S0002-9947-1943-0012401-3

Wu, S.-S., Yang, H., Guo, F., & Han, R.-M. (2017). Spatial patterns and origins of heavy metals in Sheyang River catchment in Jiangsu, China based on geographically weighted regression. *Science of the Total Environment*, *580*, 1518–1529. https://doi.org/10.1016/j.scitotenv.2016.12.137

Xu, B., & Lin, B. (2017). Factors affecting CO2 emissions in China's agriculture sector: Evidence from geographically weighted regression model. *Energy Policy*, *104*(July 2016), 404–414. https://doi.org/10.1016/j.enpol.2017.02.011

7. APPENDIX



Appendix A – Correlation Matrix