Masters Program in Geospatial Technologies

Location Analysis of city sections

Socio-demographic segmentation and restaurant potentiality estimation – A case study City of Lisbon

Dejan Popović

Dissertation submitted in partial fulfilment of the requirements for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS
UNIVERSITAT JAUME·I

Location analysis of city sections

Socio-demographic segmentation and restaurant potentiality estimation – A case study City of Lisbon

Dissertation supervised by:

Professor Roberto Henriques, PhD

Professor Marco Painho, PhD

Professor Jorge Mateu, PhD

February 2016

AKNOWLEDGEMENTS

I would like to express full appreciation and gratitude to Consortium of Master's program Geospatial Technologies for enabling me to be funded throughout my studies and thus giving me the opportunity to achieve this master's degree.

Secondly, I would like to thank to my family, for support, trust and belief in my success.

In addition, I am grateful to my supervisor and co-supervisors for sharing with me an immodestly meaningful feedbacks for the improvement of the thesis.

Special thanks to PhD candidate Fernando Santa for statistical support, Yash and Abu for friendly advises.

"Know thyself" (Greek: *"Gnôthi seautón"*) - Thales of Miletus (c. 620 B.C. - c. 546 B.C.)

Location analysis of city sections

Socio-demographic segmentation and restaurant potentiality estimation – A case study City of Lisbon

ABSTRACT

One of the objectives of this study is to perform classification of socio-demographic components for the level of city section in City of Lisbon. In order to accomplish suitable platform for the restaurant potentiality map, the socio-demographic components were selected to produce a map of spatial clusters in accordance to restaurant suitability. Consequently, the second objective is to obtain potentiality map in terms of underestimation and overestimation in number of restaurants. To the best of our knowledge there has not been found identical methodology for the estimation of restaurant potentiality. The results were achieved with combination of SOM (Self-Organized Map) which provides a segmentation map and GAM (Generalized Additive Model) with spatial component for restaurant potentiality. Final results indicate that the highest influence in restaurant potentiality is given to tourist sites, spatial autocorrelation in terms of neighboring restaurants (spatial component), and tax value, where lower importance is given to household with 1 or 2 members and employed population, respectively. In addition, an important conclusion is that the most attractive market sites in Lisbon have shown no change or moderate underestimation in terms of restaurants potentiality.

KEYWORDS

Location Analysis

Exploratory Analysis

Self-Organized Maps

Spatial Weighting

Generalized Linear Model

General Additive Model

Predictive Modelling

R

ArcGIS

ACRONYMS

SOM – Self Organized Maps

GLM – Generalized Linear Model

GAM – Generalized Additive Model

EDA – Exploratory Data Analysis

ESDA – Exploratory Spatial Data Analysis

U-matrix – Unified Distance Matrix

BMU – Best Matching Unit

AIC - Akaike Information Criterion

BIC – Bayesian Information Criterion

ANOVA – Analysis of Variances

CV – Cross validation

REML - Residual Maximum Likelihood

IRLS - Iteratively Reweighted Least Squares

RMSE – Root Mean Square Error

INDEX OF TEXT

INDEX OF FIGURES

INDEX OF TABLES

1 INTRODUCTION

1.1 Overview

One of the most visited and most attractive European cities in 2015 was Lisbon [1]. In 2009 it was the 7th most visited city in Southern Europe [2]. The reason for attraction can be firstly due to its affordability when compared to other European capitals with its; beautiful antique majestic houses and buildings in downtown districts, rich history, warm climate particularly during the winter, closeness to the beaches, food and many other things [3]. Tourism as well as the service industry in general significantly contributes to the economy both at the local and state level. This is shown in the fact that it was recorded that tourism contributes approximately 330,000 direct jobs and 900,000 jobs indirectly thus representing about 20% of the total employment, which is approximately 20 billion euros of annual income in Lisbon and Portugal (Figure 1) [4].

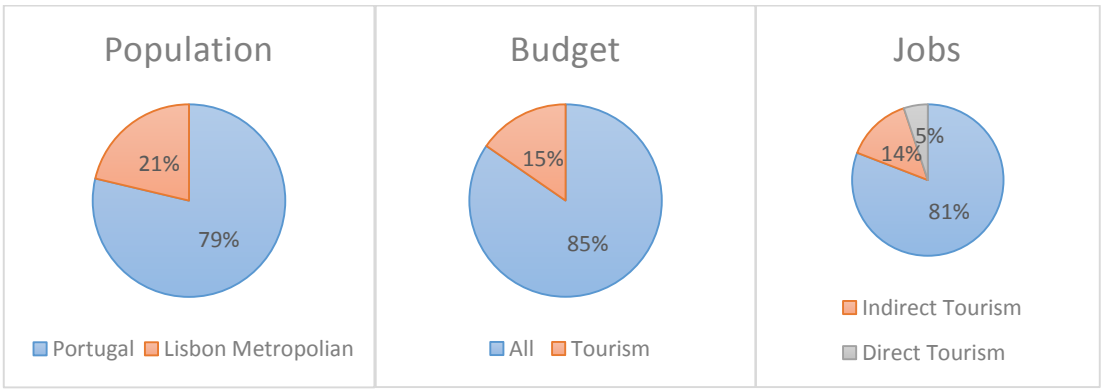


Figure 1: Population, Budget, Jobs in Portugal. Data derived from INE

In addition, population of Lisbon's metropolitan area is 2.822 million and thereabout 1/3 of the total population in Portugal; therefore the tourism industry is important in Lisbon [5]. Restaurants are one of the contributors in the hospitality and tourism industry overall [6]. The locations of restaurants with novelty and strangeness of the tourist sites are inseparable, thus they have to be considered and analysed together [7]. Therefore the primary question is - to what extent does the influence of tourist sites impacts on the location of restaurants? Certainly one of the most important factors is demography of the site itself with specific information about the age, employment status, sex, income [8]. Indeed, there are also many other non-demographic important decisions that have to be considered in order to estimate a location for restaurant such as: crime rate, visibility, area traffic, ease of access, parking, area zoning, advertisement and type of cuisine [9] [10]. Locating the most convenient site for a restaurant requires time and tedious work where stakeholders have to analyse the site and come up with a number of well justified factors with specified weights for each factor in order to determine the

most suitable location [11]. The main reason why most of the factors have not been included in further analyses is due to the lack of data. Thus the proposed solution can be useful in the scenarios where cities or urban settlements do not have much useful public data such as: purchasing index, lifestyle data, food and beverage purchasing index and so forth.

The analyses presented in this paper take advantage of the existence of secondary data about city districts and try to re-evaluate the potentiality for restaurants. Ultimately the proposed method should determine the potentiality of restaurants in the city section (Portuguese: seções). Thus in the proposed method, location analysis is conducted in the way of spatial clustering with assistance of Self-Organized Maps. The totals of tourist sites and restaurants are taken into account for the potentiality assessment on the city section level of detail. Conducted location analysis of the city section provides socio-demographic definition of the city sections and also estimates the potentiality of restaurants. The proposed method for the potentiality is applied by taking into account a spatial autocorrelation of the restaurants. The prediction methods such as Generalized Linear Model (GLM) and General Additive Model (GAM) both alone and with spatial components were tested. We propose that from the best selected method the cut-offs values of deviance residuals between the total predicted and existing restaurants can be associated with bell-shaped approximation for normal distribution. Those cut-offs values were extracted for the final product - the restaurants' potentiality map.

1.2 Objectives

The research objectives are initially based on assumptions with assistance from demographic variables from Census data 2011. Thanks to experts and examples from [9], some assumptions can be made as following:

- Age and employment might affect restaurants visits
- City sections with college educated persons indicate an increased likelihood of higher income.
- High tourist index leads to an increased number of restaurants
- People who work in the tertiary sector are more likely to visit restaurants
- Tax value defines the economic strength of city sections [12], hence it automatically influences on the suitability of restaurants
- Newer residential buildings indicate the location of younger and richer families – target group for restaurants

The assumptions helped to select suitable and appropriate variables for further analyses about the city sections. However, the census data does not contain a wealth of information about socio-demographic components in Portugal that directly relate to the suitability of restaurants, but it provides general insight about the city section (Portuguese: *seção*). Consequently the following objectives were identified:

- Undertake a segmentation of city sections with cluster analyses and identify the clusters with restaurant potential.
- Conduct the prediction analysis to determine the potentiality of city sections for the restaurants and examine the degree of contribution of covariates in the method used.

The following questions that may occur throughout analysis are: Which sites have higher potentiality of restaurants, but with non-tourist influence? What socio-demographic combination of covariates may influence suitability of restaurants? Regardless, the aim is to highlight the sections which can be potential for restaurants, but without performing an in-depth analysis about individual city section. The reason to disengage the factors from [9] provided is due to the lack of information. Hence the variables that provide *National Statistical Institute of Portugal* (INE) were analysed as an implicit or an explicit implication for restaurant suitability. In reaching these objectives, this document presents a review of the implemented methods and related work in terms of segmentation and potentiality in section 2. In the following section 3 the study area is described and an explanation of the city divisions provided. The methodology is presented in section 4. Section 5 conveys the results, in particular the segmentation analysis and results from prediction, while in section 6 were analyzed together. Finally, in section 7 the inference about the conducted analysis is stated and future improvement of the research is proposed.

2 THEORETICAL FRAMEWORK

2.1 Related work

To the best of our knowledge, limited focus has been given to the estimation of site potentiality of the restaurants on the level of city section, district, block, census track or similar subdivisions. Some of the reports are related to potentiality of the restaurant revenue by implementing machine learning methods [13]. Others are based on restaurant ratings and popularity [14]. Some experts mention the use of applying Multi Criteria Analysis (MCA) for site suitability [8]. The MCA methods can be time-consuming caused by associating weights to the specific module which requires questionnaires and meetings with catering, restaurateurs and experts. Therefore even a small change can affect the final location [15] [16].

A number of papers give analysis about fast-food restaurants and neighbourhoods [17]. However, fast-food restaurants in this analysis are deliberately avoided due to the fact that they are widely distributed all over the city. In other words a vast majority of people can afford to visit and buy a meal in fast-food restaurants. Primarily, we want to narrow down possible customers, therefore non-fast food restaurants are included in this analysis. With spatial analysis from [18] of cafe shops, a customer segmentation is considered, however the results were biased since the site selection criteria were customized for a certain brand and it relied on MCA, also. Hence the method can depend on subjective decisions and requires decision-makers to assign weighs for every factor.

2.2 Exploratory spatial data analysis and Self Organized Maps

Along with introduction into a problem, it is important to identify the components and obtain an understanding of the dataset. Consequently, exploratory data analysis were conducted for an in-depth analysis of variables in order to appropriately select those for Self-Organized Map clustering. In addition, spatial autocorrelation and spatial dependency has to be explored, since areal entities are being used.

2.2.1 Exploratory spatial data analysis and descriptive statistic tools

Exploratory data analysis (EDA) describes a set of methods for assessing features of the data and to identify patterns within a dataset. It is meaningful for problem definition since it enables one to combine graphical and numerical statistical analyses. This allows one to discuss the patterns and develop solutions to the problem [19]. Apart from EDA, exploratory spatial data analysis (ESDA) would need to be utilized to obtain a visual interpretation of features' geographical distributions, taking into account spatial proximity as well as an importance of spatial autocorrelation based on Tobler's I Law of Geography [20]. The set of dynamic graphical methods that may be applied into the dataset variables are histograms, scatterplots, boxplots, parallel coordinates, normal Q-Q plots, kernel density estimators (KDE) and

implicitly spatially related component planes and *I Moran Local* test to obtain results for spatial autocorrelation among restaurants and tourist sites. A set of thematic maps may also be used to conduct descriptive statistics.

Histograms: Widely used density estimators. The horizontal axis represents bins defined as intervals and vertical axes volume of it. It estimates probability of distribution of continuous variables [21].

Scatterplots: Prints values, typically from two variables where the horizontal axis indicates values from one variable and the vertical axis has values from another variable. The points appear as scattered; hence it gained the name scatterplot.

Boxplot: It is used for non-parametric numeral values. It depicts data with a rectangle where the bottom and upper line represents quartiles, while the horizontal line within the rectangle represents the median. Points outside the boxplot - whiskers are possible outliers [22].

Parallel coordinates: This is an alternative way to visualise multidimensional data. The horizontal axis represents each variable from the dataset and the vertical parallel axis indicates the range of values for the specific variable.

Q-Q plots: In our case these will be applied for deviance residuals distribution. It plots the quantiles of the dataset with quantiles of normal distribution. It indicates whether the scale and skewness of two datasets have similarities.

Kernel density estimators: A non-parametric estimator for probability density function (PDF) It is applied for continuous data. An important step is to choose the best bandwidth which usually depends on the number of bins, minimum and maximum value among other things. At the beginning we applied kernel density for mapping density of restaurants and tourist sites and afterwards for setting up threshold for cut-offs from deviance residuals of predicted and existing number of restaurants.

For the purpose of ESDA, provided tools were implemented in order to assist in establishing segmentation analysis of clusters associated with city districts.

2.2.2 Self-Organized Maps

SOM belongs to the family of methods from an artificial neural network. It was first presented by Finish scientist Kohonen in the early 80s [23]. The method is primarily clustering solution for multidimensionality approximation into two dimensions. The ultimate purpose of SOM is dimensionality reduction [24] [25]. Two dimensions provide more effective visualization effect of spatial data [26] [27]. The other main characteristic in regards to SOM is that it belongs to unsupervised learning. It does not require training dataset to adjust and classify data. Output networks contain nodes, so if it is two dimensional, the network is defined with a rectangular shape of nodes (Figure 2). The input dataset is directed to the output of two dimensional nodes.

For the input data the chosen winning nodes will be based on similarities. Thus, it is common to compare SOM to typical clustering method, such as *k*-means.

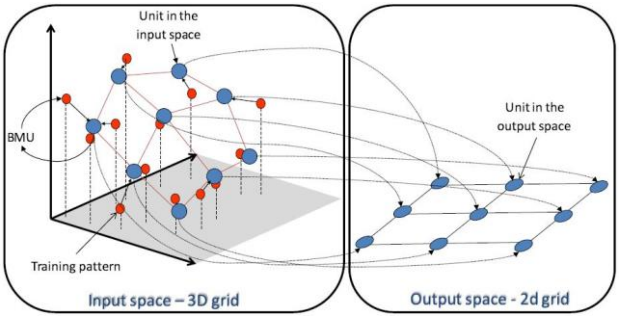


Figure 2: SOM illustration: Example of three dimensionality towards two dimensionality [25]

The importance of SOM is to preserve topological relations between interconnected elements and enables one to visualise approximated elements onto a map - in this particular case – geographic thematic map. The inventor of SOM algorithm was inspired by how the human brain works, therefore he defined neurons as a set of nodes ordered in *x* and *y* direction of rectangle shape. For instance rectangle in size of 10 x 10 will produce 100 neurons. Hereafter, the input data in multidimensional is going to be aligned to each neuron according to the training algorithm. Typically the shape of the neuron is square or hexagonal, however in most practical cases the hexagonal provides much smoother maps, since a node has potentially six (6) neuron neighbours in comparison to square which has four [25]. In addition, the bordering neurons could also have different topology. In order to preserve topological connection between neurons, cylinder and toroidal shapes help to avoid edges (Figure 3).

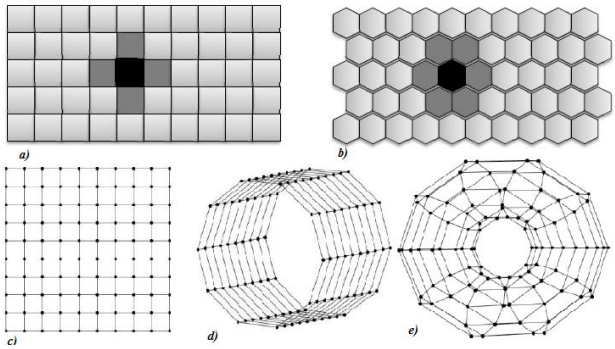


Figure 3: Typical types of topology for neighbours and borders – square (a), hexagonal (b), square net (c), cylinder (d), toroidal (e) [25]

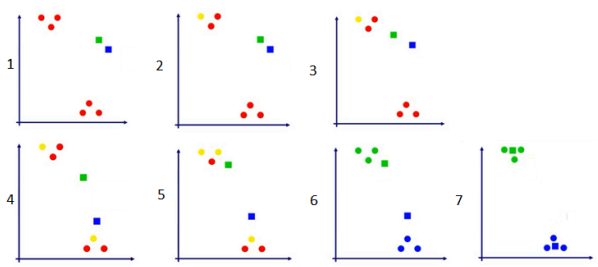


Figure 4: SOM steps - best matching unit and input dataset

The training is an iterative process. The SOM starts with basic steps explained in Figure 3 and Figure 4. In the first step the red dots represent inputs, where red and blue squares are neurons. In the second step the input value is randomly selected, that is the yellow dots. In the third step the value of the learning unit and neighbour radius is activated, therefore due to the Euclidian distance, the red node is attracted (Best matching unit). Furthermore, the weights of that node are adjusted towards the input value. Consequently, all input values were selected until the last specified iteration (step 4, 5). At the very last step nodes take position as a mean among clustered input values around and it is becomes a cluster. Formulation is presented below:

Distance calculation: $(d_{ij} ¼ ||x_k\text{-}w_{ij}||)$

Voting phase: $(w_{ij} : d_{ij} ¼ min\ (d_{mn}))$

Updating phase: $w_{ij} ¼ w_{ij} + \alpha h\ (\ w\ winner, w_{ij})\ ||x_k\text{-}w_{ij}||,$

Where:

d_{ij} – distance between weight vector and input

x_k - vector of input

w_{ij} – weight vector

α - learning unit

h – neighborhood function

In addition, if SOM is trained well, the patterns close to each other in an input space will be mapped to neurons which are close to them. In contrast, the one further away will be mapped at a longer distance [28]. The optimal number of iterations can be examined with the mean distance to the closest unit during the iteration process. Following the iterations, mean distance decreases and at a certain iteration the mean distance does not drop dramatically, thus more iterations are not necessary to be executed (Figure 5).

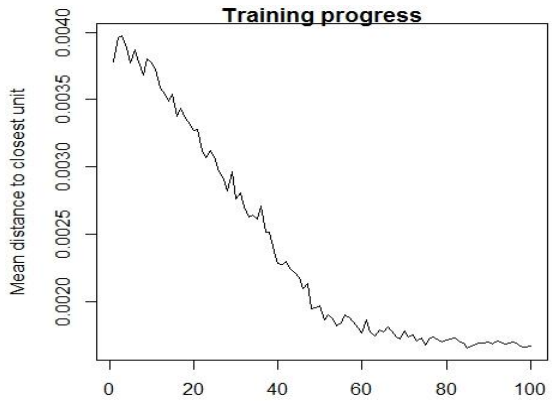


Figure 5: Training progress and mean distance to closest unit

Learning unit $\alpha\ (t)$ indicates the declination over iterations. It is linear function between [0, 1] and gives an impact on mobility of input patterns. SOM training stops when a predefined number of iterations is achieved. The neighbourhood function h, depends on radius r, and winner unit. The radius is associated with the area of influence for the winner unit.

The tools for visualization of output space includes; component planes for categorical maps, unified distance matrix or so called U-matrix for distance maps and geographic thematic maps for representation of clusters.

Hierarchical SOM clustering methodology is applied to achieve better distinction among socio-demographic clusters for the case study of Lisbon. One common way to graphically represent such hierarchical clusters is with the use of a dendogram as is shown later in this paper. Hierarchical approach is based on distance between neuron values, hence the resulting vector ordered with similar values will be creating clusters. This is applied to avoid subjective decision to divide clusters in U-matrix.

2.2.3 Spatial clustering – autocorrelation test

In order to test the existence of spatial autocorrelation of SOM clusters, Mantel test was applied. Mantel test is a tool for finding significance of statistics between correlations of two matrices. The method for correlation can be either Pearson, Spearman or Kendall. Also, it uses permutations of N rows and columns of dissimilarity matrix. The distance between points represents symmetric relationship (w - matrix), while distance between values indicates arbitrary relationship (u - matrix). The procedure is a regression approach. Null hypothesis indicates that there is no presence of spatial autocorrelation, while alternative hypothesis significates that there is a schema of spatial clustering between clusters [29].

2.2.4 SOM - related applications

For the segmentation mapping of city districts the provided dataset has aspatial features, therefore an implementation of geographical oriented SOM is not necessary. Furthermore, in the paper from [30] the distinction between standard SOM and GeoSOM was analysed for the Lisbon Metropolitan Area, however the resulting clusters derived from aspatial data were not dramatically different from the ones embedded with geographical features. As such both methods are applicable for the purpose of the thesis as the dataset from Lisbon census is identical.

With regards to ESDA, Koua [31] presented an analysis in terms of similarities for socio-demographic components between the municipalities in Netherlands. The benefit from this analysis presented, was mesh visualisation. The U-matrix nodes can be projected onto 2D and 3D dimension, while the distances between points are steadily preserved. The conclusion is that application of SOM technique on exploratory analysis supports and improves discovery of the large datasets. An inspiration for labelling the output clusters from U-matrix is related to Logo's paper [32]. The paper proposes cluster labelling for geographic maps according to the distance of input space between elements and BMU. Furthermore, the author presents and explores the use of the border between georeferenced elements. However, hierarchical clustering method helps to overcome this barrier. It is important to highlight that usually limited numbers of

variables can be mutually visualised as in large and typically complex dataset there are many mixed patterns that have to be analysed. Therefore, there is a need for linked interaction to explore and observe data from different views. SOM allows a user to extract structures by displaying patterns of the data which can be added on the map [33]. In the same report the nodes of the U-matrix are categorically coloured on the geographic maps, but in further analysis, the colours are applied for clusters. It is important to note that the number of cluster is revealed by *k*-means clustering. More precisely, with a function of the sum of within cluster difference and number of clusters.

Another example for exploration of census districts with the use of SOM is presented [34] to measure socio-economic change over time that conducted spatial-temporal analysis for 1996 - 2006 for the city of Toronto. An important inclusion they gained is standardisation of the variables, i.e. data normalisation commonly on scale 0 to 1. The raw variables themselves are not advisably to be imported into the SOM process. On the other hand data related to population usually is recommended to be converted as a decimal percentage, since the districts with higher percentage of population should have higher influence in clustering. However the authors of the paper deliberately assigned an equal number of neighbourhoods to one neuron, because they wanted to see change of one district over time, and so it was necessary to implement this idea. In further analysis, the number of city districts associated to the one neuron is not always equal. The implementation therefore does not take into account temporal components. Hence, the formulation of the number of sections per neuron is automatically determined by SOM algorithm, in regards to variables' similarity. A novel approach for urban analysis in the SOM clustering dataset was the Normalized Difference Vegetation Index (NDVI) [35]. The aim is to include NDVI values derived from bands of Landsat 7+ imagery into a dataset and later on to have a values of NDVI from achieved clusters. The values of NDVI are in the range from -1 to 1, where higher values indicate higher environmental-economical factor whereas values around 0 indicate urban, barren areas and rock. Although, the inclusion of the NDVI is not included in the study of Lisbon reported in this paper, it may be of interest for future work.

2.3 Prediction Examination

Prediction models identified in other similar cases and studies have been proven to be related to generalized linear models (GLM) as well as general additive models (GAM) [36]. Roughly the methods are similar, however GAM emphasizes smoothness of the model. In other words, each covariate used for prediction is maintained with a smooth function. Basically, once the dependent variable - response is specified, other variables – covariates were multiplied with estimated coefficients and combined together were used for response prediction. The

combination of variables which corresponds to the best prediction, quality and accuracy of the model is selected as final. Every model contains three main components:

- Response variable → Prediction
- Systematic component → Set of Explanatory Variables
- Link function → Approximates prediction into a mean. It could be by *identity, logit* (*Binominal*), *log* (*Poisson*), *Inverse* (*Gamma*)

2.3.1 Generalized Linear Model

The Generalized Linear Model is formulated with following equation:

$$g(\mu) = \beta o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \; + \; \varepsilon = x^t \beta \qquad (1)$$

Where:

μ - dependent variable

x_i - independent explanatory covariates

β_i - estimated parameters

g - a link function for the transformation response

ε - random variable or error

The response variable for both GLM and GAM follows a distribution of exponential densities. For the certain type of the response the deviance takes a form of distribution:

- Normal →Symmetric → Continuous →$\mathbb{R}$
- Binomial →Discrete → [0 or 1]
- Poisson → Discrete → $\{0\}$ U $\mathbb{N}$ → Count → Asymmetric
- Gamma → Continuous → $\mathbb{R}^+$ U $\{0\}$ → Asymmetric

In every model is assigned quantity – r degrees of freedom. It is associated with a number of independent variables to be estimated. If *n* stands for number of independent observation, then the deviance for residuals has (*n-r)* degrees of freedom.

The parameters were estimated with maximum likelihood equations with a procedure of iterative weighted least squares (IRLS) [37].

2.3.2 Poisson distribution

The selected type of the response is related to the deviance. For instance, if a return value has continuous results between 0 and 1 the suitable solution is to choose *Gamma* distribution. Since restaurants variable belongs to events per city section it will belongs to *Poisson* discrete distribution. Hence, the dependent variable is said to have *Poisson* distribution if contains integer values $y = 0, 1, 2\ldots$ with probability:

$$\Pr\{Y = y\} = \frac{e^{-\mu}\mu^y}{y!}, \text{ for } \mu > 0. \qquad (2)$$

If the assumption is that response variable mean and at the same time variance depends of explanatory variables x_i, then linear predictor can assume only real values, however *Poisson*

respects count values. The alternative is to apply *logarithm* of the mean. Thus, the link function will be:

$$\log(\mu_i) = x_i'\beta \tag{3}$$

By increasing x_i for one unit, β increases along *log* of mean. Exponentiation of the previous equation provides multiplicative effect of the linear predictor on the mean. Increasing x_j by one, the mean is multiplied by factor $exp\{\beta\}$ [38].

$$\mu_i = \exp\{x_i'\beta\} \tag{4}$$

2.3.3 Generalized Additive Model

In GAM the linear form is replaced with sum of smooth functions:

$$g(\mu) = f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + \cdots + f_k(x_k) = f_0 + \sum_{j=1}^{p} f_j(x_j) \tag{5}$$

Where:

f_j – non-parametric function

$f_j(x_j)$ – estimation using cubic spline smoother

A smoother is a way of identifying tendency of response variable as a function of covariates. Hence the estimation triggered by smoother is named smooth. The trend of the variable can be seen from the plot, thus the selection of the smoother is getting simple. The one used are presented below:

Cubic smoothing spline is one of the solution for optimization. It calculates second continuous derivatives from $f_j(x_j)$ and picks the one that minimizes penalized least square [39].

$$\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 + \lambda \int_a^b [f''(x)]^2 dx \tag{6}$$

Thin-plate regression spline is applied in multidimensional regressions. In further work is related to x_1,x_2 coordinates and simply for additional spatial component of the GAM model. Also, it minimizes least square from second derivative in two dimensions.

$$\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 + \lambda \int\int [f''(x_1, x_2)]^2 dx_1 dx_2 \tag{7}$$

The method for unbiased parameter estimation such as *restricted maximum likelihood* (REML) considers the smooth elements as random effects and produces less biased estimates than maximum likelihood [40]. For the purpose of GAM model is selected for smoothing parameter estimation.

2.3.4 Methods for spatial aggregation of neighbourhoods

In terms of census data, any information is mainly associated with some areal units. Government authorities define for example, tax, voters, or in general collects data according to some spatial entity. Thus it is important to include spatial behaviour in terms of Tobler's Law of Geography in which entities influence each other depending on a distance [41]. The closer spatial units are, they may have higher similarities. The proximate observations partially can be used for prediction of their neighbours. In this sense it is strongly encouraged to consider spatial

autocorrelation as a possible factor which can impact on the final predictive model. In order to formulate the spatial weights among neighbours within GLM, it is necessary to establish the accurate measure for spatial autocorrelation. How spatial autocorrelation can be presented among city sections, it is a crucial goal for the construction of the matrix of weights. To define the spatial weights we first have to define relationship between city sections and to calculate weights accordingly.

The spatial contiguity is given when three or more polygons can meet into a single centroid point, otherwise at least two boundary points have to be within snap distance. These two relationships are popular as *queen* and *rook* (Figure 6).

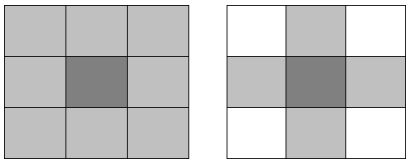


Figure 6: Queen's (left) and Rook's contiguity (right)

Other possibilities include observing relationships based on graphics. For the purpose of examining the relationship among city sections it is necessary to calculate locations of city centroids. The representatives from this group are:

- Delaunay triangulation → opposite relation to Voronoi diagrams on the same surface
- Sphere of influence → removed longer links from Delaunay triangulation
- Gabriel graph → keeps different set from Delaunay triangulation
- Relative graph neighbours →similar to Gabriel, preserves symmetry

Ultimately, the last possible group implemented is distance-based in terms of the number of centroid neighbours. The objective is to identify the closest *k* numbers of neighbours based on the minimum distance. The resultant calculation is the list of vectors of distances for the first, second, third and fourth nearest neighbours [42].

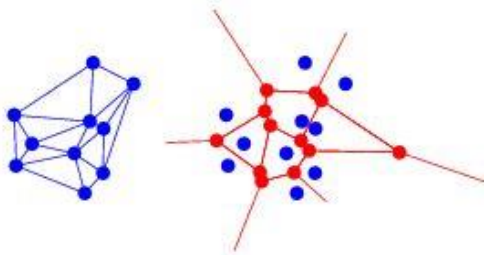


Figure 7: Delaunay (left), Voronoi (right) – Source: wolfram.mathworld[1]

Each neighbourhood criteria has been applied for the study area of Lisbon and the most appropriate criteria is selected for the spatial weights calculation. The selection method will be further discussed.

[1] http://mathworld.wolfram.com/

2.3.5 Spatial weights

Spatial weights are a list of weights associated with a list of neighbours, i.e. city sections. Weight between *i* and *j* is the *n*-th of the *i*-th weights list of city sections and *n*-th element indicates which *i*-th city section list values is equal to *j* [43].

$$\boldsymbol{W} = \begin{matrix} 0 & \dots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \dots & 0 \end{matrix}$$

Since, the locations of restaurants are at external borders of the study area, the style of the weights criteria is based on row standardization. So the weights of the areas with less neighbours is larger than those with more neighbours.

The weights of matrix **W** is extracted from one of the chosen neighbourhood criteria from the previous section. Areas without neighbours will be assigned to zero. In R the package **spdep** function **nb2listw** extracts neighbourhood list in a matrix of weights for every areal entity.

2.3.6 Spatial autocorrelation test

Once the spatial weights are determined it is advisable to do a statistical test for spatial autocorrelations. The commonly applied test is *Moran's I*. It is given as a ration of the product i.e. restaurants and its spatial lag, i.e. neighbours of restaurants with cross-product of variable restaurants and previously calculated spatial weights [36]. The formula is as follows:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y)(y_j - y)}{\sum_{i=1}^{n}(y_i - y)^2} \tag{8}$$

Where:

y_i – i-th observation

y - mean of the resturants values

w_{ij} - spatial weights between i and j neighborhoods

In terms of testing, the global test for autocorrelation as well as local test were conducted. As mentioned before for global test, it is common to work with *Moran's I*. However the *Geary C* may be included to gain more confidential test result. The outcome is a standard deviate compared with Normal distribution. The null hypothesis says that there is no spatial dependences among spatial weights. Probability values were achieved in regards to comparison of standard deviate from the test and Normal distribution.

The outcomes provided by *Moran's I* test are following:

- observed value from *I*
- expectation
- variance from *I*
- standard deviate
- *p*-value

The local test for autocorrelation is to present local behaviour between city sections and its neighbourhoods. It is meaningful test for identifying isolated observed city sections with some number of restaurants, otherwise those with a high number of restaurants. In other words, the examinations of city section is established according to four categories:

- those which have high values of the observed variable
- those which have low values of the observed variable
- neighbouring city sections with very high and simultaneously very low values
- neighbouring city sections with very low and simultaneously very high values

In addition local *Moran's I* test is explained with equation:

$$I = \frac{(y_i - y)\sum_{j=1}^{n} w_{ij}(y_i - y)}{\frac{\sum_{i=1}^{n}(y_i - y)^2}{n}} \tag{9}$$

The formulation can be interpreted as a number of elements gathered together in order to achieve global Moran's test [42]. Mostly it is reviewed with Moran scatterplot, where the horizontal axis indicates the observed variable and vertical lag or neighbourhood distances or just geographic map with labelled classification according to the above-mentioned four categories.

For the purpose of spatial component in GLM the *Moran eigenvector* approach is conducted by [44] as well. *Moran eigenvector* is a statistical approach to prove the presence of spatial independency. Moran eigenvector defines how spatial components can be incorporated into a model and it is applied for all variables from the full model.

This *brute force* method was used to find the set of eigenvectors of the spatial weight matrix (MP) which is defined as following:

$$MP = MWM \tag{10}$$

Where:

$$M = I - X(X^T X)^{-1} X^T \tag{11}$$

The above (11) equation signifies symmetric projection and unchanged element, where *W* is matrix of weights [45].

Moran eigenvector map contains two vectors. The first eigenvector is related to values of Moran's index given by spatial weight matrix, and the second vector is perpendicular to the first, although it has to maximize Moran's index. The inclusion of Moran eigenvector maps

(MEM) for a model enables one to observe spatial relationship with results (Figure 8). The more positive eigenvector values are, the more spatial phenomena is evident [46].

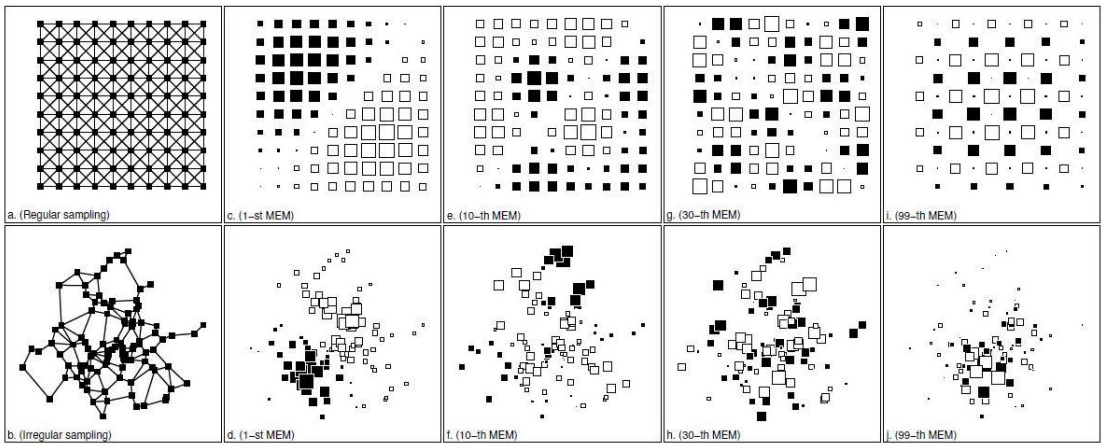


Figure 8: MEM for regular (up) and irregular (bottom) network – first in a tow have positive and last in a row have negative eigenvector values. Derived from [46]

2.3.7 Model selection

In statistical investigation it is always a concern as to whether one model is better than other. The choice for the best model requires some model selection criteria. Every model contains a score or value on which we can be based quality. Therefore the candidate which better scores should be chosen for further analysis [47].

In this analysis, the model selection is based onto further selection steps:

- Model selection in regards to the most suitable variables for both GLM and GAM
- Model selection in regards to the best GLM
- Model selection in regards to the best GAM
- Model selection in regards to the best ultimate model

The most popular methods for model selection are as follows:

- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

In the literature authors do not state a significant difference between them, however generally AIC finds the model which gives the best prediction, while BIC selects model which is considered as a "true" model [48].

Both AIC and BIC are based on maximum likelihood parameter estimates. AIC is defined to be -2 multiplied with log-likelihood with addition of 2 multiplied with number of parameters:

$$AIC = -2loglik + 2d \tag{12}$$

Unlike AIC, BIC model selection is defined as:

$$BIC = -2loglik + \log(n)d \tag{13}$$

N is number of observations. Thus, from equations BIC penalizes parameters more than AIC. When one compare *log(n)* with 2, BIC the model will be more simpler than AIC, in terms of

number of covariates. In the project analysis the idea was to have as minimum as possible variables in order to have a more robust model and more secure variables throughout. The workflow of the model selection criteria is based on BIC [49], although for the neighbourhood criteria between city sections, it is based on AICc criteria. AICc is an AIC model, however for the confined sample size. The assumption for the AICc is that the model has to follow normal distribution for residuals that is linear, hence the formulation is:

$$AICc = AIC + \frac{2d(d+1)}{n-d-1} \quad (14)$$

According to the formula it is more robust and penalizes parameters slightly higher than AIC.
For the comparison of several models, an analysis of variance (ANOVA) is computed. ANOVA provides an efficient analysis of variance in a sequential order. It compares the smaller model (with fewer variables) against the next in the sequence that is a more complex model with at least one more variable. The comparison step is examined with likelihood ration test. The output contains residual degrees of freedom and deviance from each model. For the models with a given dispersion which is Poisson, the chi-squared test is recommended [50].
The *p*- value tests the null hypothesis that groups of all populations have the identical mean. It means if *p*-value is large, then there is not enough evidence that means are different. So, the population means are all equal, thus we do not have enough evidence to reject null hypothesis. If *p*-value is smaller than 0.05, then we reject null hypothesis saying that there is statistical evidence to prove that all populations have identical means.

2.3.8 Predictive approaches - related applications

The flow for the predictive modelling is partially derived from papers related to rat sightings [51] and residential burglaries [52], and it is rearranged together for the purpose of this study. Furthermore, in proposed approach both GLM and GAM are taken into account. In the first paper [51] authors decided to find the association between aggregated observations per census track and the predefined focus points. In fact, they proved that association exists and that distance model is the most appropriate model. The model discovers even unexpected results such as that cat's feeding stations are attractive to rats. So, the distance model was better than other models and this was used for further analysis to find out indicators that have associations with focus areas. The approach related to the distance can be implemented in our example since focus areas can be established, but the other issue is lack of data in terms of factors for restaurants potentiality. On the other hand, an aim is to globally evaluate city sections; however distance model satisfies only specified sites within specified radius. In other words, the distance model is certainly a good approach to reconsider, for example focus areas could be assigned to

shopping malls, or the most popular tourist sites, and radius with influence factors can be established accordingly, but it will remain open for some future studies.

The second paper [52] presents an innovative approach for crime probability prediction with the assistance of cut-offs based on fractal skill score. Non-linearity in the covariates shares the common behaviour with covariates from our study. A one-dimensional smoother is implemented for continuous predictors and two-dimensional isotropic smoother is applied for spatial component in GAM. In addition their approach is in a way improved with temporal component, i.e. with three dimensional smoother, however our method is static, hence is not necessary.

One of the alternatives to consider is zero inflated model. A zero inflated regression model addresses with existence of zero events in response variable. In Lisbon, many city sections do not have restaurants. That would indicate that zero inflated *Poisson* (ZIP) distribution could be implemented. However it is important to differentiate structural zeros within population. Structural zeros are associated for those subgroups which can be identified from the population that have no risk at all to change behaviour. In contrast, there are random zeros or sampling zeros, which are produced by sampling variability [53]. An assumption is that certain subgroups from the population will never change behaviour, in other words they would always have associated zero counts. It can be a good idea to test for the future work, however in this research structural zeros are not considered. In this study every neighbourhood is regarded that has some degree of potentiality for restaurants either negatively or positively.

Regarding GLM model, since *Moran's I* test is a mean for detecting spatial autocorrelation, in the same way it can be utilized for embedding spatial autocorrelation into a model. In this sense, Moran eigenvector calculates two eigenvectors from the spatial lag of GLM model with spatial weights and afterwards it is fitted into the full GLM model [42].

Following theoretical framework it can be noted the importance of involving a spatial dimension both in segmentation and prediction map. SOM in many examples proved that it can be good method to cluster census data, however the novel step was to test significance of spatial clustering with Mantel test. In addition, regression models provide higher accuracy in prediction in presented examples, but spatial autocorrelation have to be evaluated and involved in equation if I Moran and Geary test point out enough evidence for spatial autocorrelation.

3 STUDY AREA

Lisbon is the capital and simultaneously the largest city of Portugal. The municipality of Lisbon (Portuguese: concelho) has approximately half a million people [54]. The municipality is divided into 24 districts (Portuguese: frequências) or small 1054 sections (Portuguese: seções). It is even further divided into smaller areas called sub-sections. For the purpose of the thesis, the level of detail selected is at a city section scale, although it is recommended as smallest as possible. However the sub-sections in terms of privacy concerns have limited a number of variables, so the city section level of detail has been chosen. On Figure 9 some sections are very small and other very large, hence the size and borders of the sections does not follow a set standard areal extent. In general, downtown areas such as *Estrela, Santa Maria Maior, Santo Antonio and Misericordia* usually have smaller sections, dense populations and are located on the south east of the map. Other city-sections such as *Ajuda, Benfica, Olivais, Belem*, have mainly larger or moderate size sections and surround the downtown area.

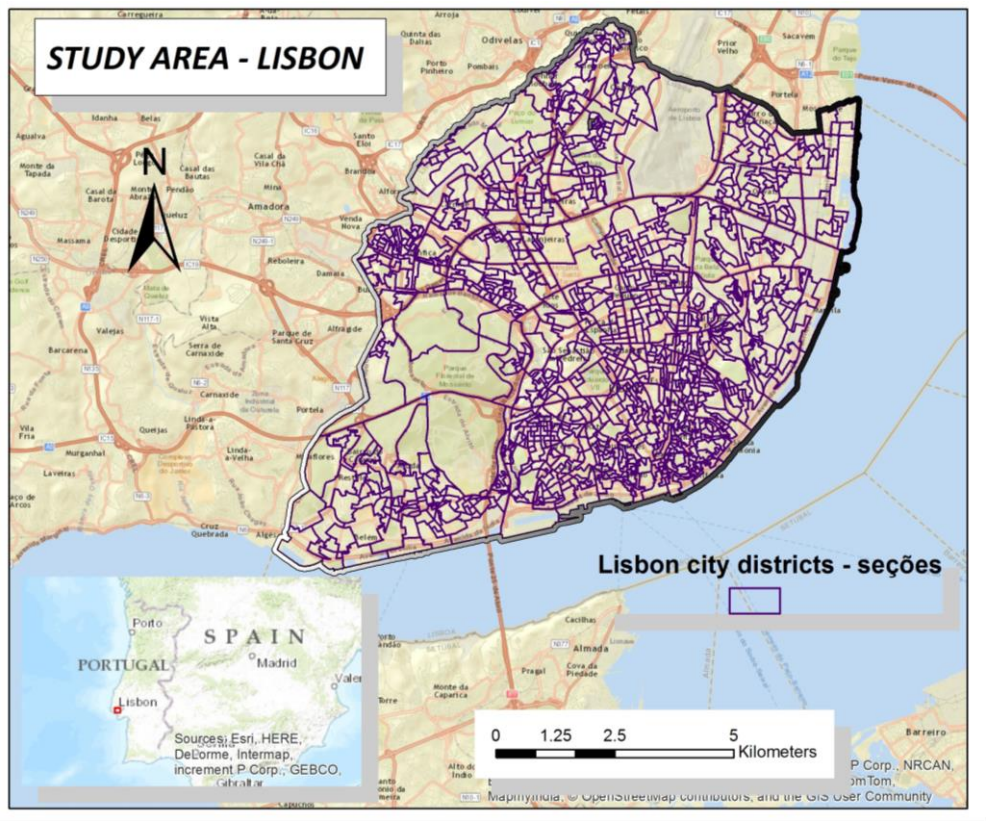


Figure 9: Study area – Lisbon

Sections of the city vary in elevation. In downtown, older areas along the water edges have lower elevation, while surrounding sections are higher. The landscape of Lisbon contains slopes ranging from small rolling hills that are widely distributed along the Tagus River to the peaks of Sintra Mountains. Thus, there are many sightseeing points towards downtown and the river and that is one of the reasons why these points attractive for both domestic and foreign tourists.

4 METHODOLOGY

In general, the proposed structural framework at Figure 10[2] contains three parallel processes, which are dependent on each other. The final step combines the visualization process and it obtains an overall visualization.

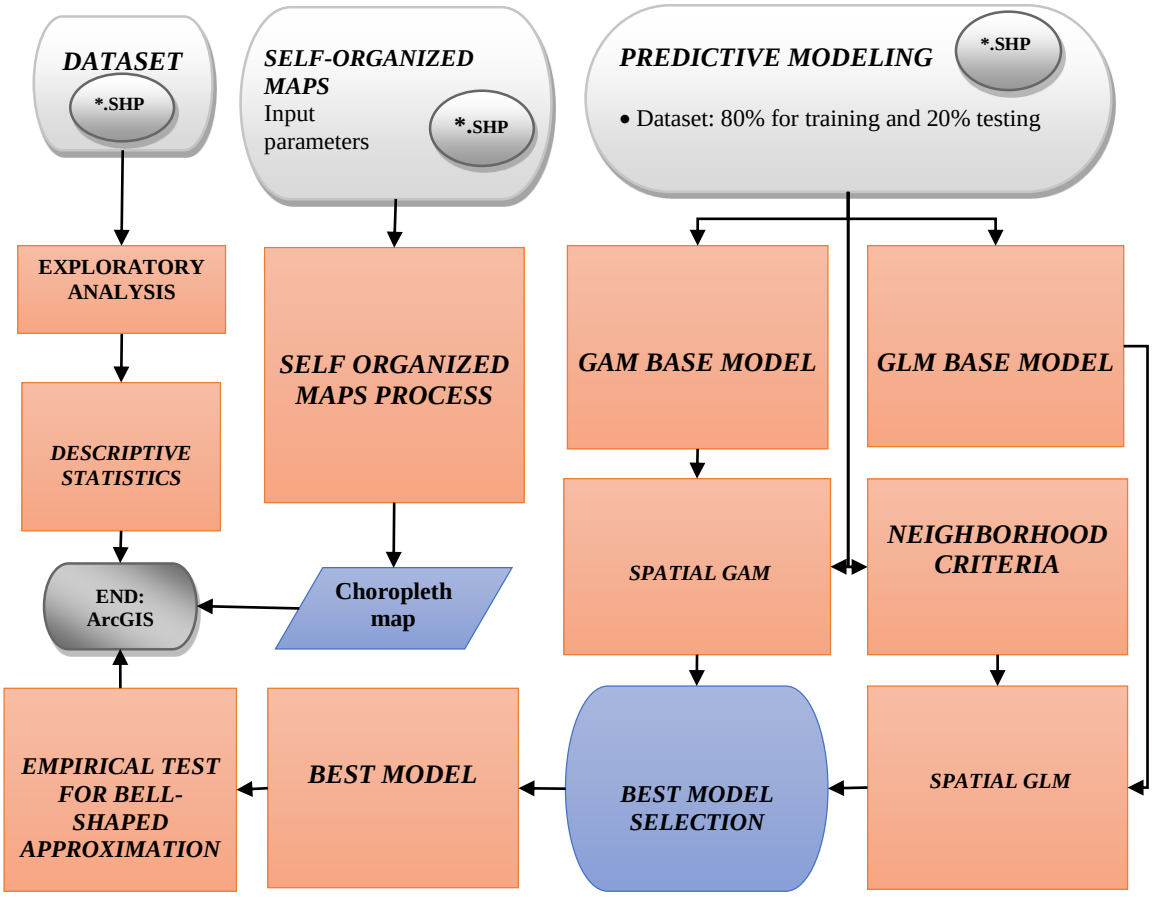


Figure 10: General Flowchart

The first step was to combine EDA and ESDA analysis in order to obtain an interpretation of variables. The second parallel process was SOM implementation with socio-demographic components provided by census data. The objective is to derive clusters which could lead to better analysis of potentiality of restaurants. Next step is the prediction of the restaurants according to the best selected models. Later on, it is important to specify cut-offs from deviance residuals and undertake classification according to potentiality.

[2] Legend for flowcharts refer to ANNEX 1

4.1 Dataset

The National Statistical Institute of Portugal provides 122 variables on the level of city section from the census 2011. Data can be divided into the following categories:

- Related to buildings (type, functionality, date of building…)
- Accommodation
- Households
- Individuals
- Education

Selected variables for the purposes of the study are presented in Table 1.[3]

	Chosen variables	Code
1	Private households with 1 or 2 people	*Hous1or2*
2	Private households with 3 or 4 people	*Hous3or4*
3	Private households with no unemployed	*HousNoUnem*
4	Resident individuals aged 20 to 24 years	*Indv20_24*
5	Resident individuals aged 25 to 64 years	*Indv25_64*
6	Men residents aged 20 to 24 years	*Men_20_24*
7	Men residents aged between 25 and 64 years	*Men_25_64*
8	Women residents aged 20 to 24 years	*Womn_20_24*
9	Women residents aged between 25 and 64 years	*Womn_25_54*
10	Men residents aged more than 64 years	*Men_64_*
11	Women residents aged more than 64 years	*Woman_64_*
12	Individuals' residents with post-secondary education	*Post_sec_e*
13	Resident individuals with a college degree	*Colle_deg*
14	Pensioners or retired individuals living	*Pensions*
15	Resident individuals without economic activity	*Res_No_act*
16	Exclusively residential areas	*Exlus_res*
17	Employed individuals resident in the secondary sector	*Work_Sec*
18	Resident Individuals employed in the tertiary sector	*Work_in_Ter*
19	Individuals employed residents	*Employed*
20	Buildings constructed from 2001	*BuildAft01*

Table 1: Selected variables from Census 2011

The variable *Tax_index*[4] is used to represent an economic indicator in terms of wealth for each city section. It was taken over from Portuguese Tributary and Customs Authority (Portuguese: *AT Autoridade Tributaria e Aduaneira*).

The number of tourist sites - *Tourst_idx* presents count per city section. Since the coordinates are known for every location, the total number per site is calculated according to spatial

[3] www.ine.pt

[4] http://www.e-financas.gov.pt/SIGIMI/default.jsp

summary. The chosen attributes that present *Tourist index* are presented in Table 2. The total number of tourist sites in Lisbon, have 720 records for the year 2015.

Selected sites - Totals
Historical monuments
Airport
Casino
Ferry terminal
Hotel
Museum
Rental car agency
Taxi stand
Tourist attraction
Tourist information
Winery shop

Table 2: Tourists sites – totals

In terms of restaurants, they were determined on the same way as tourist sites. Likewise, *Restaurant* variable indicates a number of restaurants per city section. Both X and Y locations of tourist sites and restaurants are provided from *Here Maps* with assistance from the official *Here maps* office in Lisbon.

Consequently, the final dataset presents the variables from Table 1, *Tax_index, Tourist_idx, Restaurant* arranged by city sections, thus they are stored into shapefile (including supporting files such as: **.dbf, *.shx, *.cpg, *.shp.xml, *.sbx, *.sbn*).

4.1.1 Data pre-processing

The raw data file for restaurants and tourist sites contained many redundancies. Therefore the procedures in R were conducted to erase them. Many items were previously converted from polygons to number of points and thus one restaurant would be represented four times in dataset. We required that one restaurant represent one count in order to achieve the number of restaurants within a certain city section.

The second issue was projection; reference system or simply coordinate system. The data from *Here Maps* was in *WGS-84* longitude and latitude coordinate system. However the provided dataset from census data was in local Portuguese reference system *ETRS_1989_TM06-Portugal*. Therefore restaurants and tourist sites were converted to the local Portuguese system in order to accomplish spatial summary for each city section (ANNEX 2).

4.1.2 Data normalization

In the study area, some city sections are more populated than others and varied in size. Therefore all of them have to be categorized on the same scale.

Firstly, the variables related to population were calculated in terms of percentage per unit, as well as number of buildings. On the other hand, it was not appropriate to calculate percentage counts from restaurants and tourist sites, therefore within combined dataset from all variables *Min-Max* normalization was applied.

Normalization is a process where the values from variables are approximated on a common scale. There are many methods, however *Min-Max* summarizes values in a range [0, 1] [55]. The normalization is necessary for SOM and it is computed using the following equation:

$$MM = \frac{(x-min)}{max-min} \quad (15)$$

Where:

x – field value

min – minimal value in an array

max – maximal value in an array

4.1.3 Analysis Approach

The approach taken for this project included a series of tests and evaluations in order to decide on the best methods to be applied for all aspects of this project.

In terms of the descriptive statistics, the methodology is presented as shown in Figure 11.

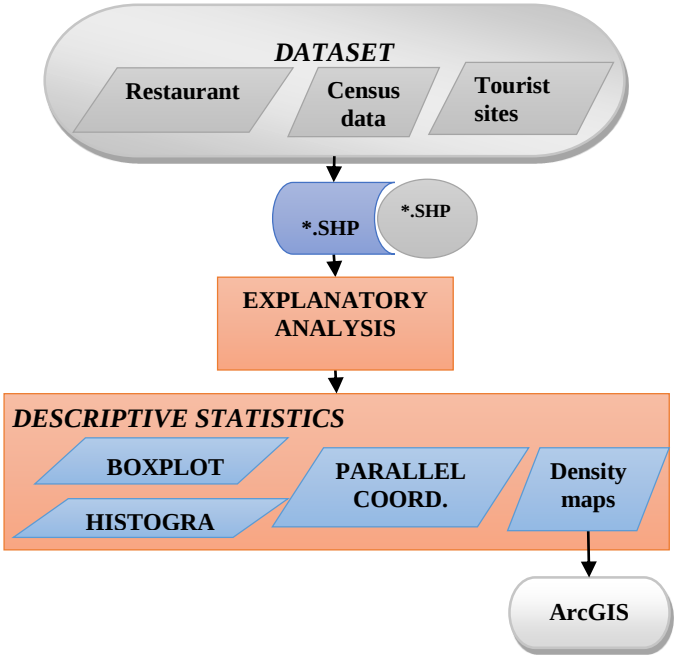


Figure 11: Descriptive statistics

In the previous section a shapefile was made and a number of graphs and plots were created in order to recognize patterns (both obvious and hidden).
It is important to have a general picture about the most important variables i.e. restaurants and tourist sites in Lisbon in order to perform further detailed analysis. For this purpose a geographic map with layers of restaurants and tourist sites was created.

It is important to find clusters that are most suitable for restaurants and to provide an explanatory spatial data analysis from socio-demographic clusters. In addition, for SOM geographical index is added so that clusters are assigned to city sections. The workflow is presented in Figure 12. In the first step it is necessary to specify input parameters from a previously normalized dataset. In the process step, from output graphs we measure quality of SOM. An important part is to analyze graphs from component planes as well as distance matrix. Vector of mean deviations helps to identify how distant are deviations from code vectors. Along with iterations the deviance is decreasing. In order to uphold more automatic decision, within SOM process, *k*-means and hierarchical clustering are executed in order to determine number of clusters.

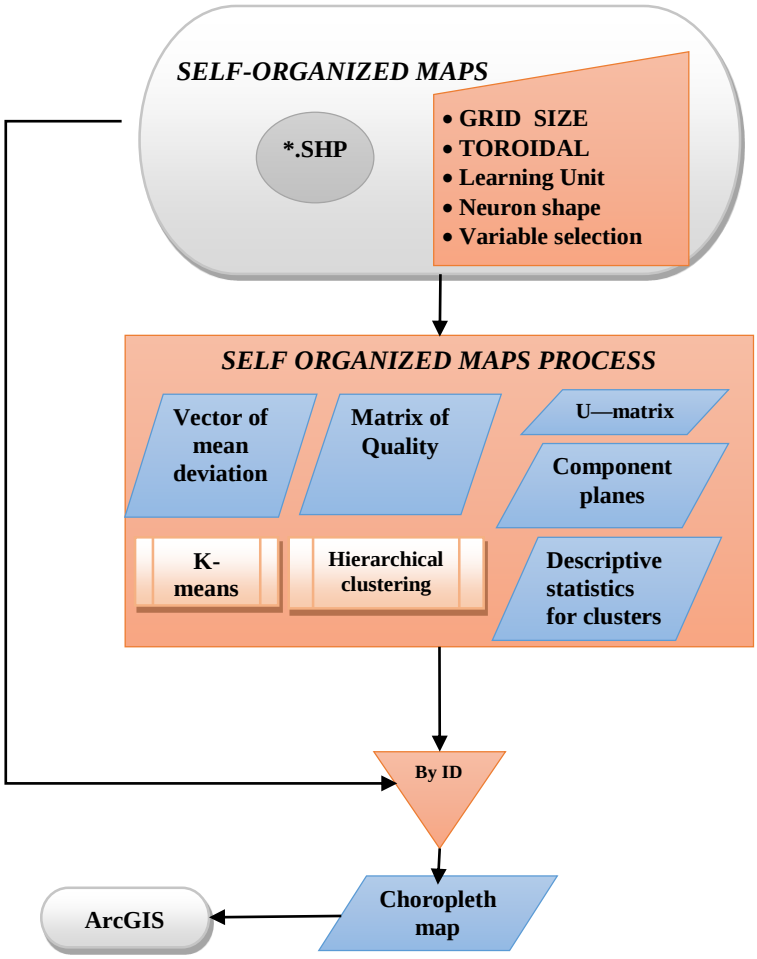


Figure 12: Flowchart - SOM

To enhance visualization and improve data analysis the decision was made to divide the dataset and develop several SOMs. The initial representatives as an input matrix are chosen from data, thus all SOMs have mutual start matrix of positions. Selected variables for each SOM are presented in Table 3.

SOM_entry	SOM_mixed	SOM_population	SOM_household
Individuals 20-24	*Household 1 or 2*	*Individuals 25-64*	*Household with 1 or 2*
College educated	*Active Population*	*College educated*	*Active population*
Residence with no activity	*Post-secondary education*	*Work in tertiary*	*Individuals 20-24*
Work in tertiary	*College education*	*Work in secondary*	*Residence with no activity*
Tax value	*Residence with no activity*	*Tax value*	*Residence 20-64*
Tourist index	*Work in tertiary*	*Tourist index*	*Work in tertiary*
Men 25-64	*Tax value*	*Woman 25-64*	*Tax value*
Women 25-64	*Tourist index*	*Woman 64*	*Tourist index*
Women 64	*Women 64*	*Men 64*	*Employed*
Men 64	*Men 64*	*Pensioners*	*Building built after 2001*
Buildings built after 2001	*Pensioners*	*Employed*	*Exclusively residential*
Exclusively residential	*Employed*		
Household with 3 or 4	*Exclusively residential*		
	Household 3 or 4		

Table 3: Selected variables for each SOM model

In accordance to predictive modelling the same datasets were used, however it had to be divided into two (2) parts.

A detailed flowchart is presented in Figure 13.

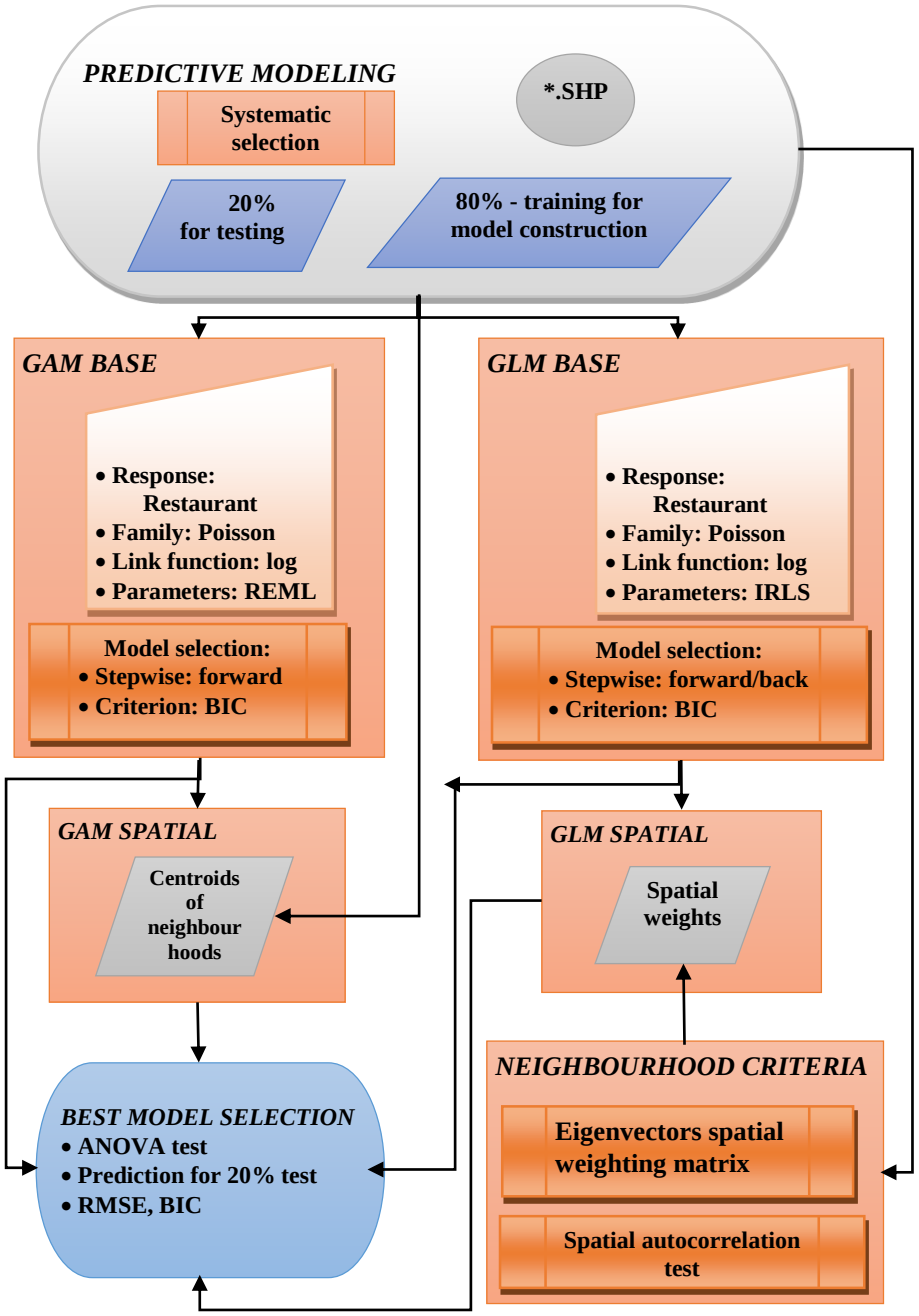


Figure 13: Flowchart - Predictive modelling

In the first step for the purpose of verifying, the dataset is divided for training 80% and testing 20%. The polygons were selected according to the systematic selection, i.e. in the sequence every fourth polygon or city section were selected for testing sample, instead of having random samples. The 20% of test data is left for the model evaluation at the end, with 80% for model construction.

4.1.4 Base model construction

Restaurants variable has a count per city section, therefore it belongs to the *Poisson* discrete link function. In other words the output belongs to a family of natural asymmetric numbers. The link function for the mean approximation is logarithmic. While base GLM and GAM were created using *Poisson loglink* function, the parameters were estimated with REML for GAM and IRLS for GLM. The components for the construction of the GLM are as follows:

- Based on formula (1) :
    a. $q(\mu)$ – response variable - *Restaurants*
    b. βk - estimated parameters based on IRLS
    c. xk - covariates – all other variables
- Link function from (3)
- Poisson: Discrete – {0}U N – Count - Asymmetric

In order to determine what is suitable the starting model is assigned for full model. The full model contains the whole set of covariates.

So, the complete GLM is presented with the following covariates:

- *BuildAft01, Colle_deg, Employed, Exlus_resi , Hous1or2, Hous3or4, HousNoUnem, Men_20_24, Men_25_64, Men_64_, Pensioners, Res_No_act, Tax_index, Tourst_idx, Woman_64_ , Womn_20_24, Womn_25_64, Work_in_Te, Work_Sec*

Having applied GLM, as an output it is possible to extract deviances residuals, estimated coefficients, standard deviation, z value as well as probability value – p value.

- *Ho* hypothesis explains that covariate is meaningful as a predictor if p – value≤ 0.
- In contrary, if p-value > 0.05, covariate should be rejected.

From the complete GLM model there are important covariates: buildings after 2001, exclusively residential areas, men aged between 20-24, tax, tourist sites and women above the age of 64.

In contrast, covariates such as working in secondary or tertiary sector, pensioners, employed etc. were not presented significant contribution for prediction. In the further step a *backward/forward* stepwise selection method was applied.

In addition, a method executes through the whole dataset of covariates back and forth in order to select or reject covariates.

Covariate selection was established via Bayesian information criterion (BIC) described in *Model Selection* section. BIC value is based on:

- the observed data,
- parameters of model,
- the number of observations,
- the number of estimated parameters and
- likelihood function

From all possible combinations of covariates and completed GLMs, according to BIC criterion, the selected model produces the base GLM with following covariates ordered by importance:

$$log(\mu) = \beta_I + \beta_1 Tour_{index} + \beta_2 Tax_{Index} + \beta_3 Hous_{3or4} + \beta_4 Exlusi_{resid} + \beta_5 Men25_{64} + \beta_6 Woman64 + \beta_7 Active \quad (16)$$

Tourst_idx, Tax_index, Hous3or4, Exlus_resi, Men_25_64, Woman_64_, HousNoUnem

One of the objectives of the thesis is to identify the city sections potentially for restaurants, hence we propose the method to calculate the deviance of the residuals between predicted and existing number of restaurants in order to determine thresholds values for cut-offs. In the flowchart (Figure 14), the first step is to derive the formula from the best selected model. However, since the best selected model contains 80% of training data, the formula is extracted and applied for the whole dataset. In the second step deviance residuals are calculated. Finally, the threshold for cut-offs are defined according to empirical test for bell-shaped approximation.

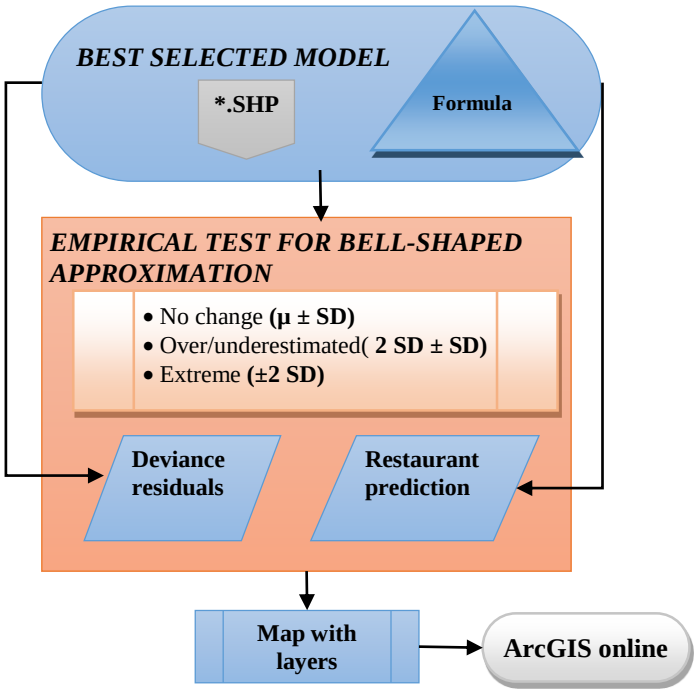


Figure 14: Flowchart - Cut-offs

5 RESULTS AND ANALYSIS

In descriptive statistics one of the steps is to present visual interpretation of the variables particularly characterized with spatial features.

In Figure 15 hotspots of tourist attractions is overlapped with restaurants. In Figure 16 a density map of restaurants with tourist sites is shown. It is visually clear that downtown areas contain higher values from both (usually around shopping malls, and trade markets) restaurants and tourist sites. However, there are locations where this patterns is not always respected. In that regard further analysis will reveal the most influential combination of variables to achieve this. Regardless, the sites where there are restaurants and non-touristic sites are important for prediction, as well.

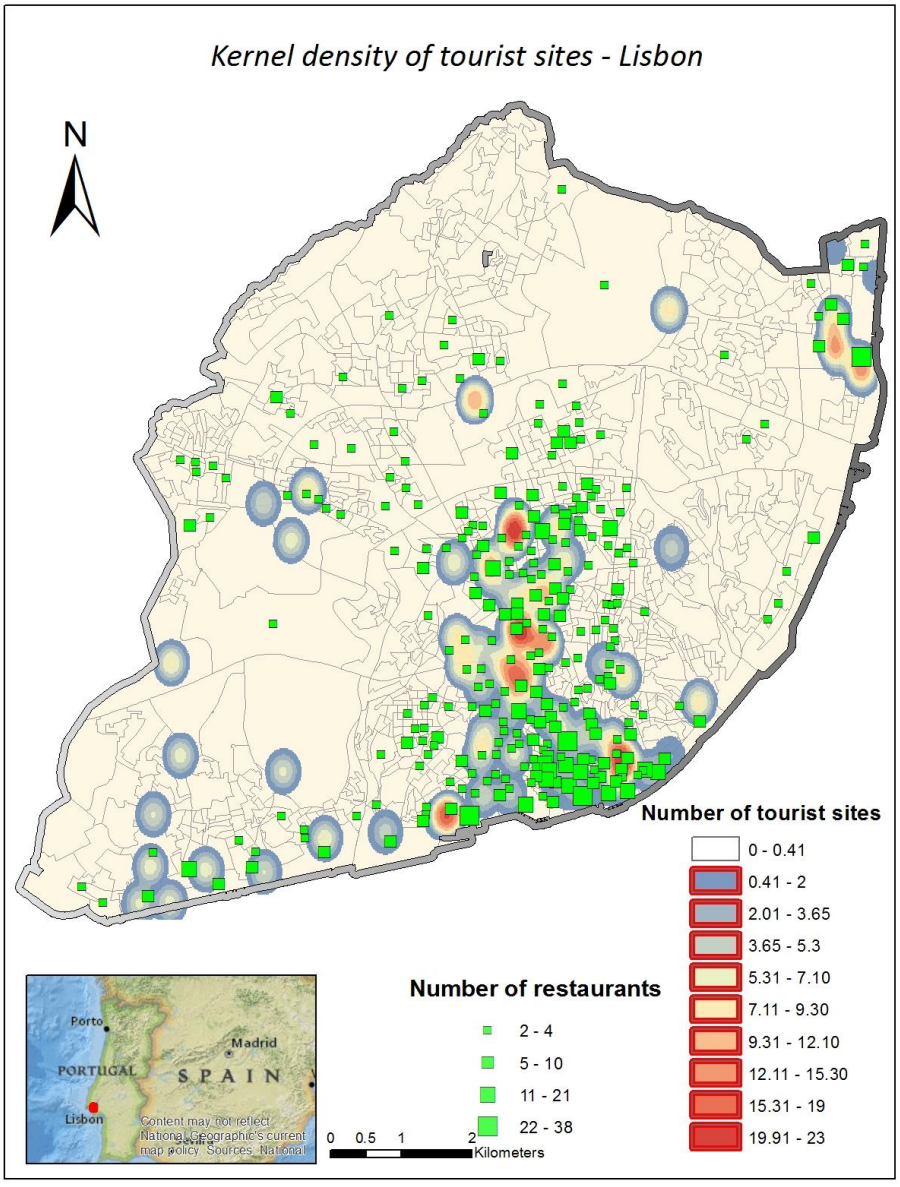


Figure 15: Density of tourist sites

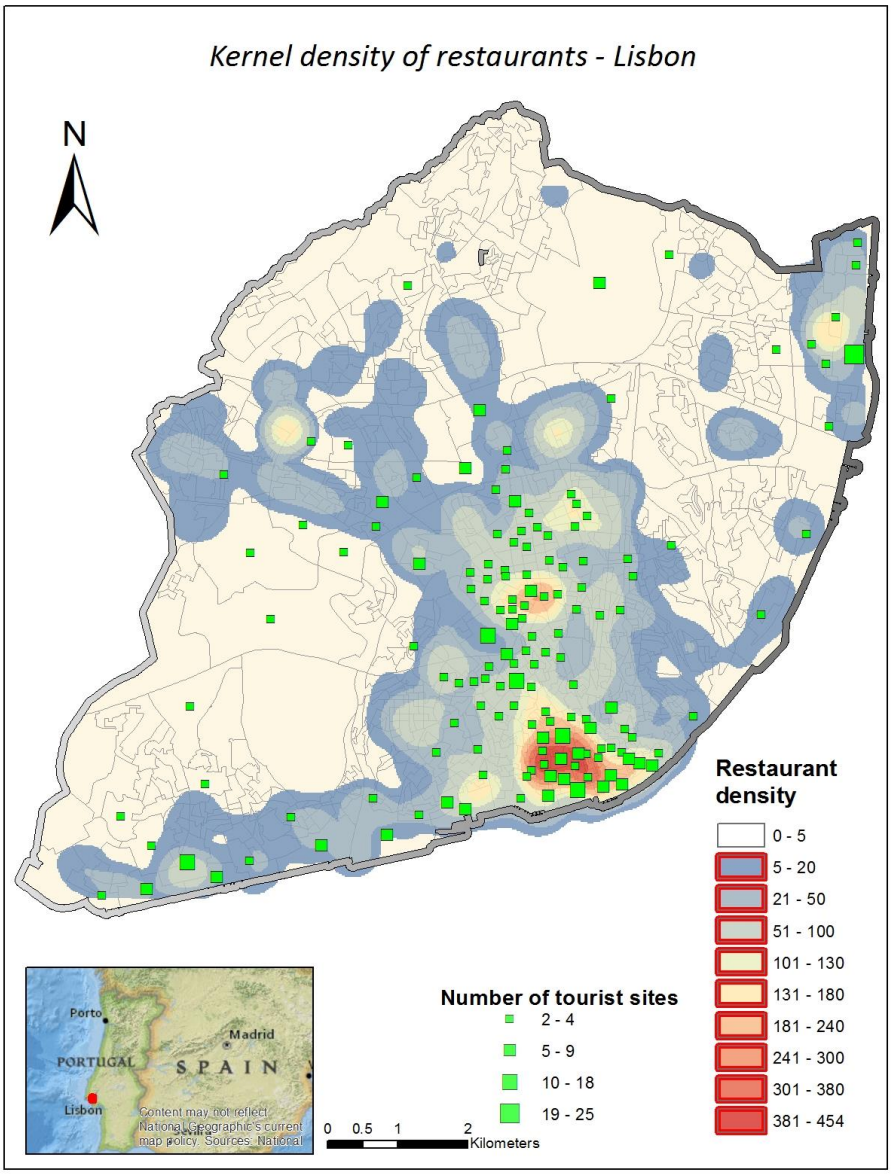


Figure 16: Density of restaurants

The larger areas where tourist sites are located without restaurants are mainly large green parks detected from imagery. With regards to non-tourist sites, the presence of restaurants identified from imagery are mainly markets and city sections on longer distance from tourist sites. One could infer, that restaurants are creating something similar to buffer zones around tourist sites. The conclusion from images is that tourist sites and restaurants in general are spatially correlated, however the important outliers could be visually identified.

A general overview about other variables was presented with parallel coordinates. An analysis was conducted by picking 100 random city sections in order to get an insight about the flowline between variables. Y-axis significates normalized range of the values, with section ID on the

left. On the X-axis are variables names. Other parallel coordinates can be found in ANNEX 3. The parallel coordinate graphs show the tendency of the variables in city sections.

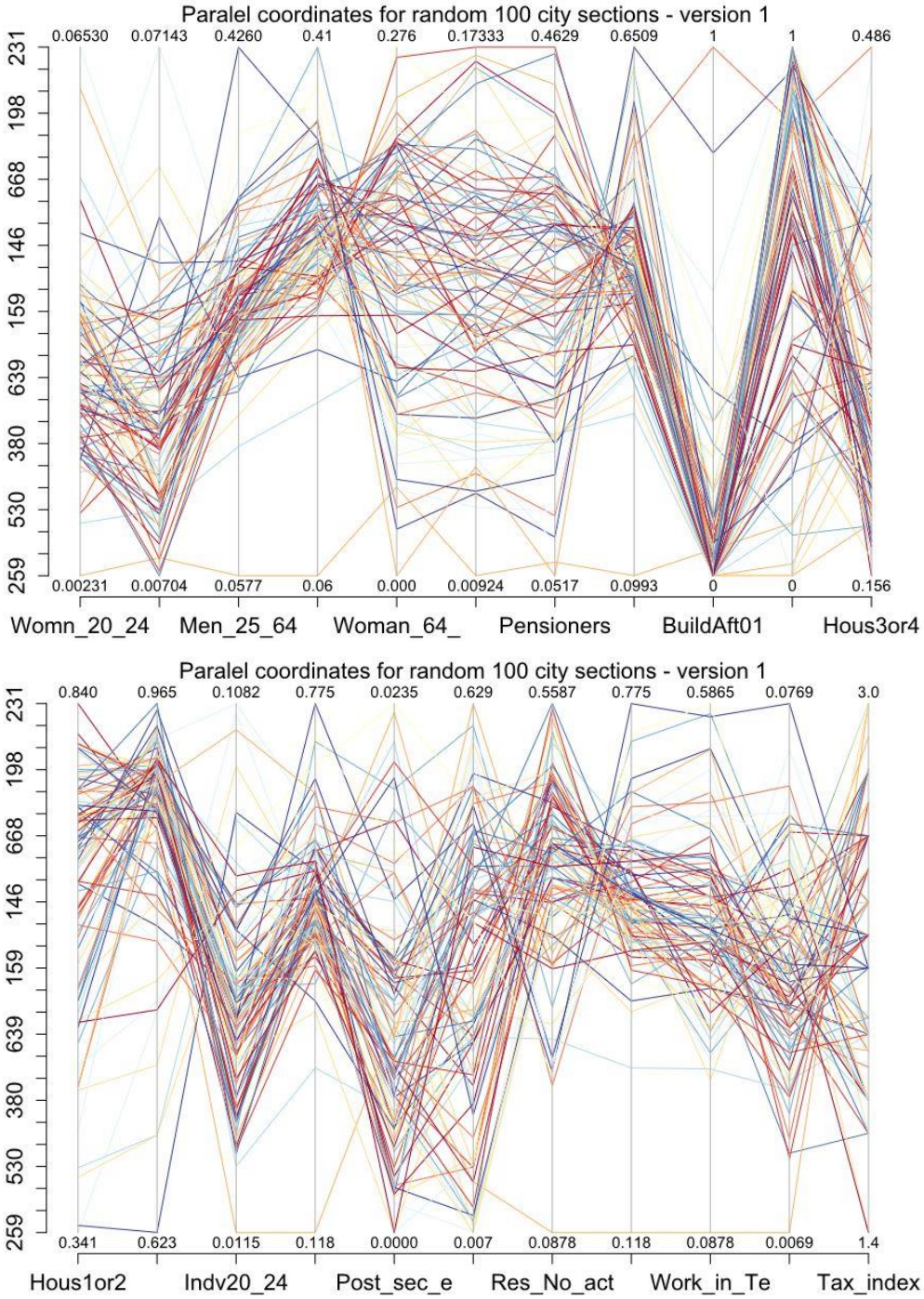


Figure 17: Parallel coordinates - all variables

The color lines have shades of red, yellow and blue, therefore the color lines close to each other would have similar shades (Figure 17). In general, according to line flow in most of the city sections, the optimal tendency can be presented in accordance to Table 4. However, taking into

consideration faded colors, we can extract many outliers such as *Buildings after 2001* which obviously describes that in Lisbon there are only two city sections with a greater number.

	Household	Employment	Education	Individuals	Others
Characteristics	70% with 1 or 2 members	80% are active	1% post-secondary	3% Women 20-24	70% exclusively residential
	25% with 3 or 4 members	55-60% employed	Very scattered with college degree	2% Men 20-24	1% buildings built after 2001
		30% works in tertiary		30% Women 25-64	
		2% works in secondary		25% Men 25-64	
				Very scattered with elder population	

Table 4: Parallel plots: an optimized city section in Lisbon

For elder population, i.e. more than 64 years and pensioners it cannot be identified as a unique pattern, although scale ranges between from ~ 10 to 40%. Likewise, population with college level education is scattered too. In other words, higher educated people can be located in every part of the city.

According to boxplots (Figure 18, Figure 19), the first impression is that distribution of restaurants and tourist sites are quite close. In other words, the median value is almost zero, and that is because from 1053 city sections, approximately one half has no identified number restaurants nor tourist sites. The other large group of sections are with one or two. In addition, there are outliers with sections containing dozens of restaurants and tourist sites. Tax values show normal distribution, because mean value and median are similar. That indicates the similar number of extremely rich sections and poor is close.

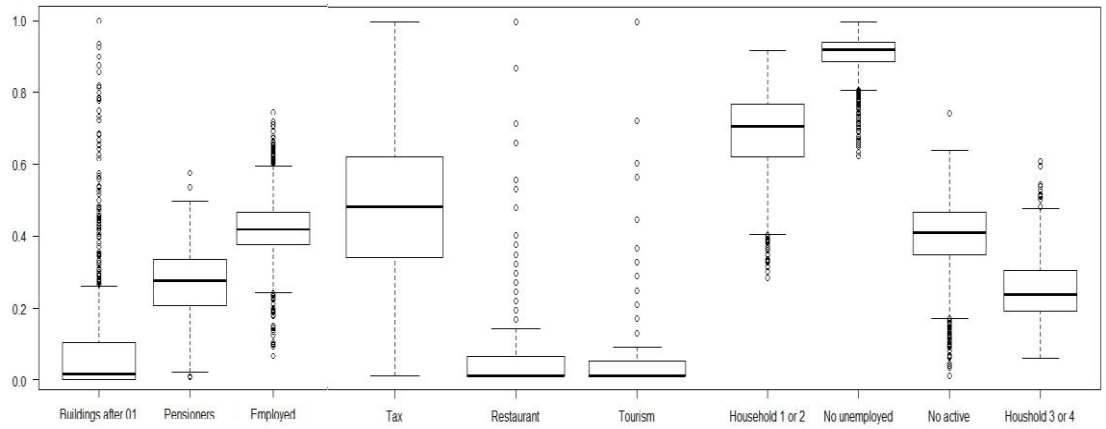


Figure 18: Boxplots

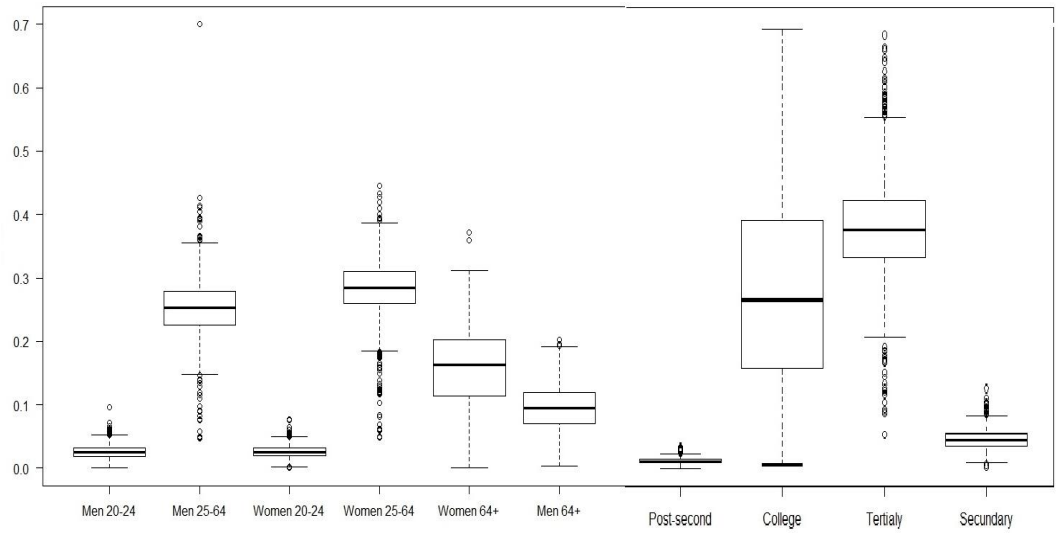


Figure 19: Boxplots

Tax values originally range between 1.2 and 3.3, and from the figures the distribution is similar to distribution of the *College* variable. This supports the assumption that the areas with higher *Tax* values are in relationship with college educated residents. In addition, it can seen that almost 40% of population works in tertiary sector. Moreover, boxplot explains that working force in sections are either mostly or rarely hired in tertiary sector, due to the presents of outliers. In regards to the number of persons in each household, boxplot indicates that Lisbon is usually a city with 1 or 2 members, since median is 70% and contrary families with 3 or 4 are around 30%.

Histograms help to analyze the frequency for each variable (Figure 20). For the *Buildings after 2001* we can see that the buildings in almost 800 city sections are older than 15 years. By analyzing *HousNoUnem* and *Employed* in the range of 80-90 % residents are active, however *Employed* are varying around 50% with many outliers. In terms of age groups, for the older population there is a higher number of women than men, and there is a link between *Houshold3or4* with *Men_64_*.

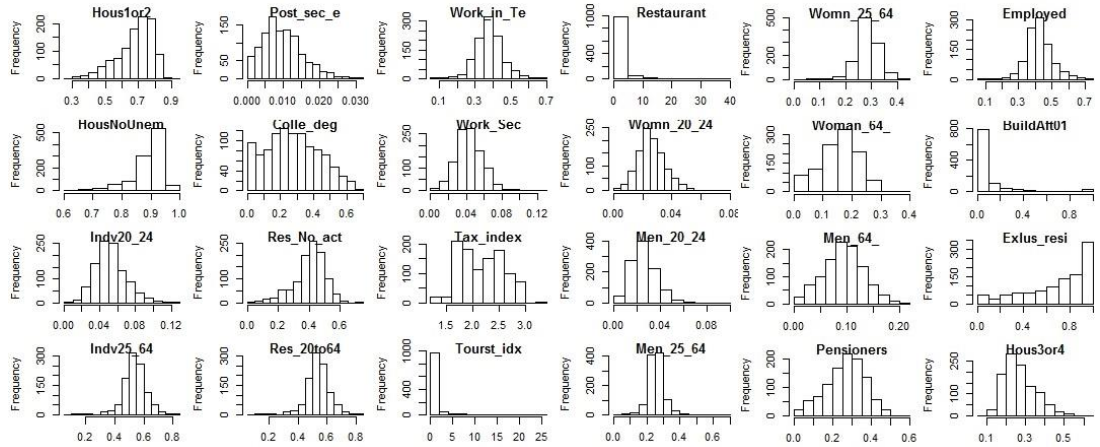


Figure 20: Histograms

In addition to the relationship between variables, Figure 21 presents two dimensional scatterplots. In terms of two dimensional correlation, the *Pearson's* linear coefficient is calculated as well as *Spearman's* correlation coefficient for non-linear parameters.

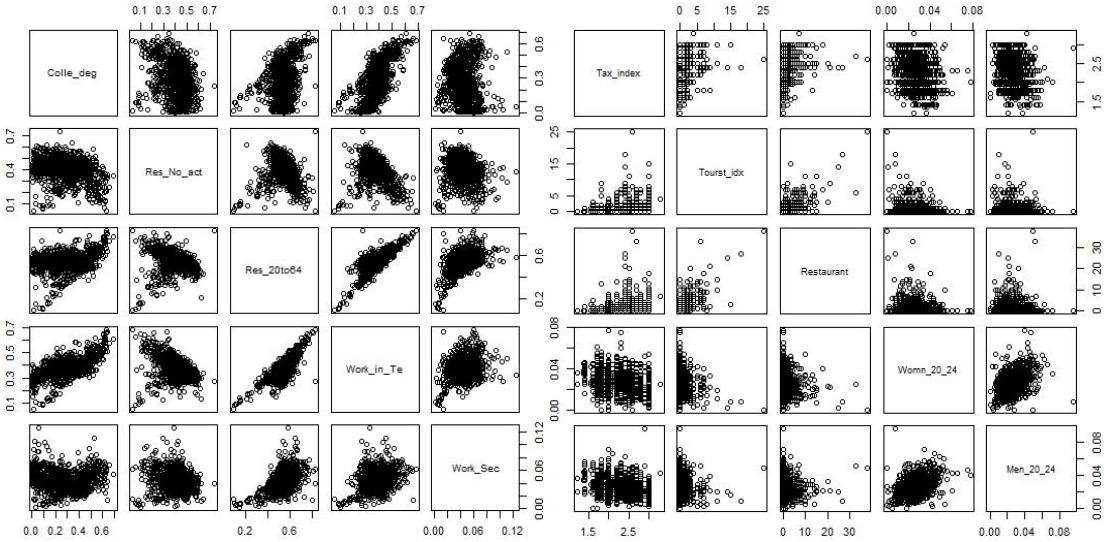


Figure 21: Scatterplots

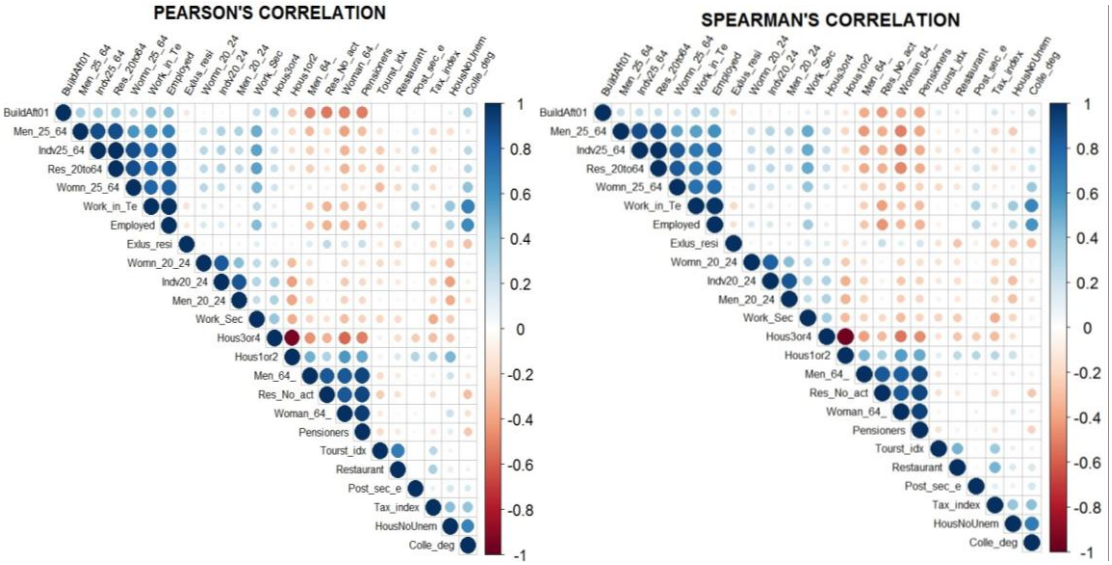


Figure 22: Coefficient - Pearson's correlation (left), Spearman's correlation (right)

In regards to *Person's* and *Spearman's* correlation (Figure 22), both of them presents similar results, although *Spearman* emphasizes negative correlations among population variables. Interesting details are:

- Tourists sites in general have no (or low negative) correlation versus other variables
- Restaurants have low positive coefficient versus employed in tertiary sector
- Active population and individuals 20-24 have moderate negative correlation
- Elderly people have a higher negative correlation with built buildings after 2001

5.1 Socio-demographic segmentation

SOMs are slightly different and they follow previous variables' analysis. Each SOM model contains different variable combination. Firstly, many variables weren't presented significant correlation, and secondly variable order has to respect restaurants potentiality. *SOM_entry* and *SOM_mixed* have mixed variables in respect to restaurants, while *SOM_population* relies more on population variables and *SOM_household* on household variables.

The results from all SOMs presented quite similar behaviour, in addition *SOM_entry* is chosen for further interpretation while others are stored in ANNEX 4, while R code can be found in ANNEX 5.

5.1.1 Input parameters

The type of SOM which is applied is hierarchical SOM based on neighbourhood distance or commonly known kohonen. Consequently, in order to accomplish results some intuitive decisions must be decided. In Table 5 is the summarized input parameters for SOMs.

SOM Parameters	Specification
Grid size	16 x 16
Iterations	100
Learning unit	0.05 - 0.01
Radius	Covers 2/3 of all unit-to-unit distances
Neuron shape	Hexagonal
Topology	Toroidal
Start	Initial matrix

Table 5: SOMs specification

Number of city sections is 1053, hence the number of neurons must be lower. Several tests with different dimensions are conducted, therefore the most suitable dimension is referred as 16 x 16, i.e. 256 neurons. In that sense, 1053 city sections are assigned on 256 neurons according to similarities. The linear unit is defined to linearly decrease along with iterations. The radius is given as default value. It will run from the given number to the negative value of that number. In order to achieve smooth topology among neighbouring neurons, the topology hexagonal with toroidal edges is selected.

5.1.2 Process

In the first step, SOM produces codebook of vectors. For all 256 neurons there were 14 vectors calculated according to input parameters. The distribution is presented at Figure 23.

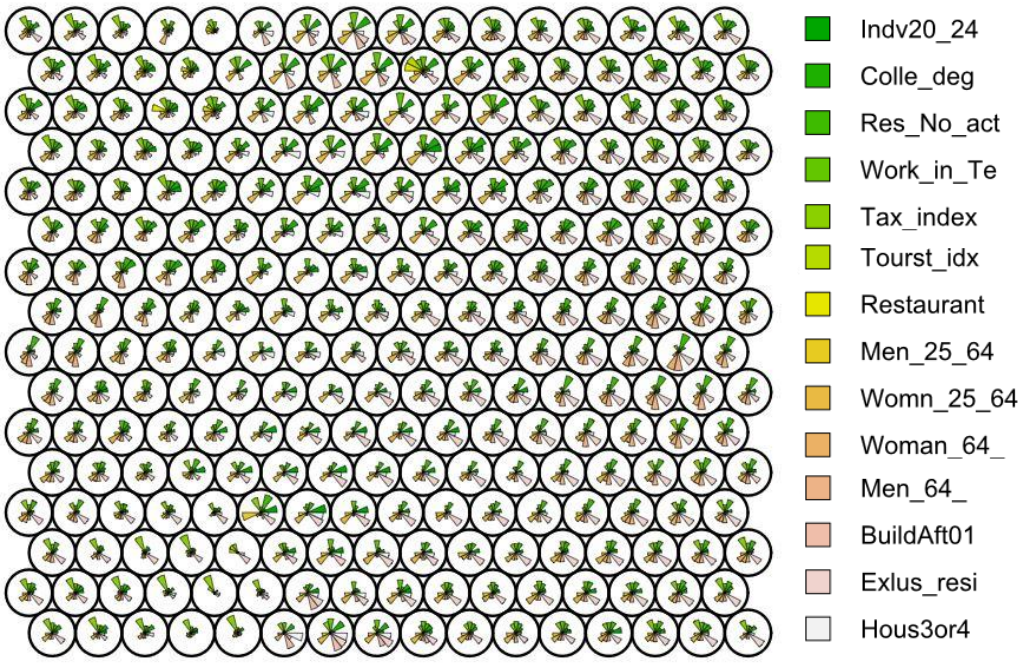


Figure 23: Codebook vectors for 256 neurons with 14 variables

Neurons are located according to distribution and distance similarities. In the further step it is necessary to analyse how many city sections are per neuron Figure 24 (left), as well as quality of codebook vectors in regards to mean distance.

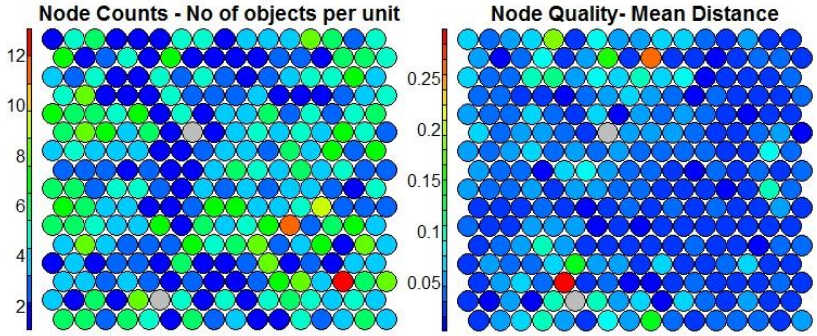


Figure 24: Number of city sections per neuron (left), Node Quality (right)

If there are less city sections per node, that indicates increased similarity. The contrary are nodes which are more independent. In this case, most of the nodes have 3, 4 city sections, but there are some with 8, 10 and more. On the right hand side *Node quality* matrix presents quality of SOM and as mean distance is lower, the deviations are negligible. The image shows the mean distance of objects mapped to a unit to the codebook vector of that unit. The smaller the

distances, the better the objects are represented by the codebook vectors. The nodes are mostly blue, which indicates good quality of codebook vector.

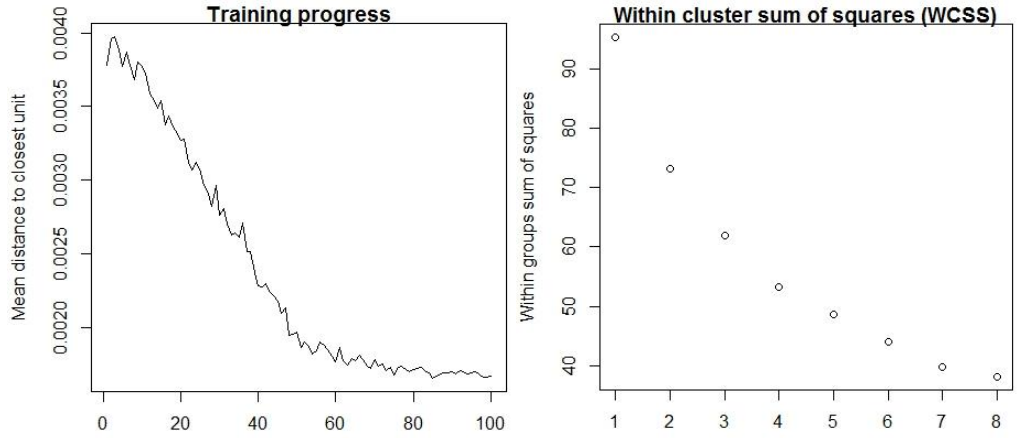


Figure 25: Number of iterations (Left), Number of clusters (Right)

The training progress is a way to visualize how much mean distance declines over iterations. According to Figure 25 (left), 100 iterations is more than enough.

In order to define number of clusters it is best to hierarchically sort codebook vectors according to distance. The dendogram allows visual estimation for number of clusters and in this case can be 7 or 8 (Figure 26), however to specify desired number of clusters it is accomplished with *k*-means.

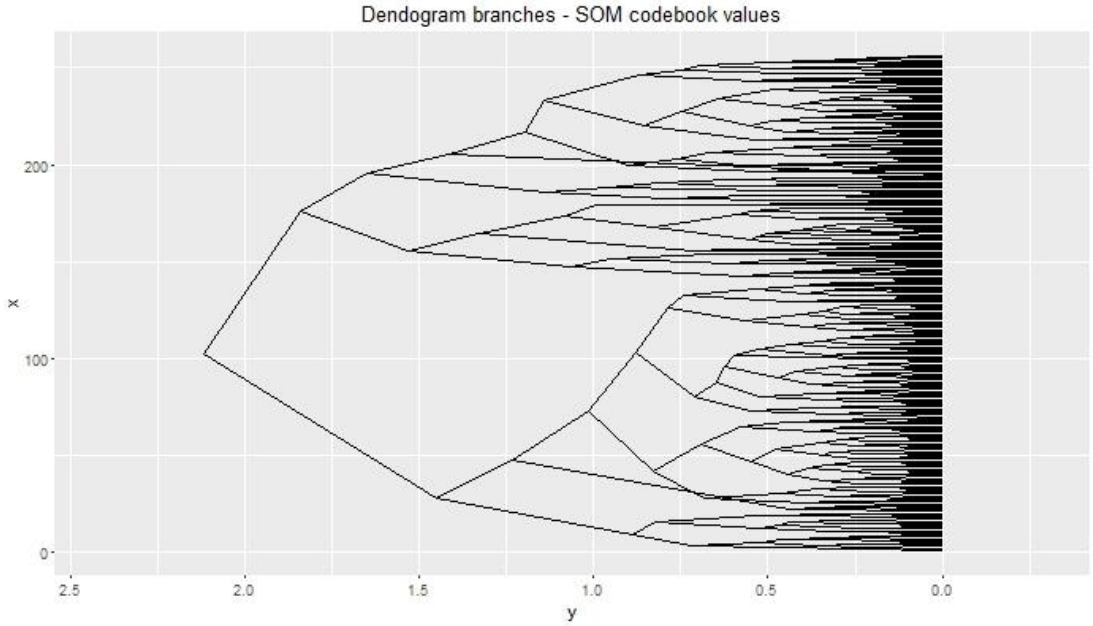


Figure 26: Dendogram

In Figure 25 (right) data given by *x*-axis are clustered by the *k*-means method, with aims to divide the codebook vector values into *k* groups such that the sum of squares from codebook

values to the assigned cluster centres is minimized. At the minimum, all cluster centres are at the mean of the set of data points which are nearest to the cluster centre.

In other words if the specified number of clusters is 1 then sum of squares is 90, with 2 clusters ~70, et cetera. The aim is that sum of squares is minimal. The suggestion is to choose the number on the curve in shape of elbow. In accordance to *Dendogram* and *WCSS* the selected number of clusters is 7.

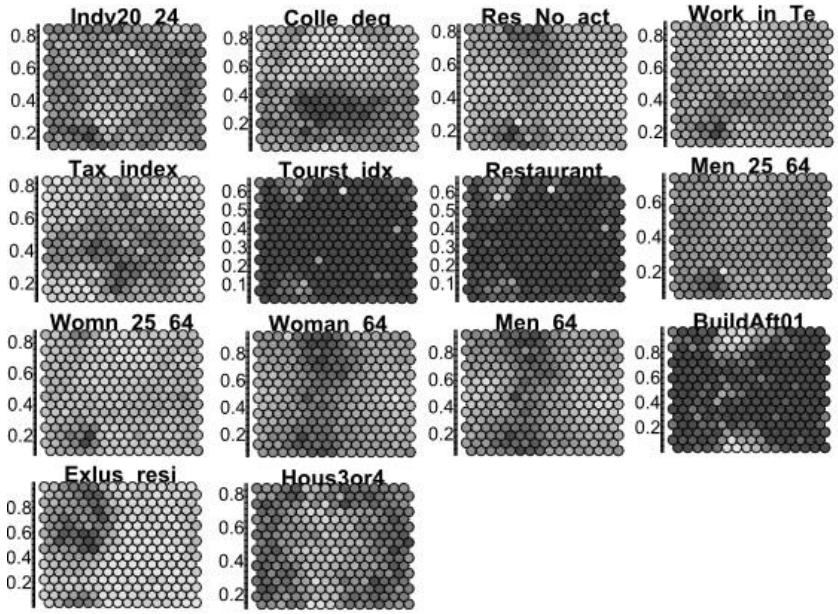


Figure 27: Component planes

Lighter neurons in component planes indicate higher values of the specified variable, therefore analysing two or more planes it is possible find related neurons between variables (Figure 27). It can be identified that there is negative correlation between *Tax index* and *Individual 20-24* as well as positive correlation between *Work in Tertiary sector* and *Buildings after 2001*.

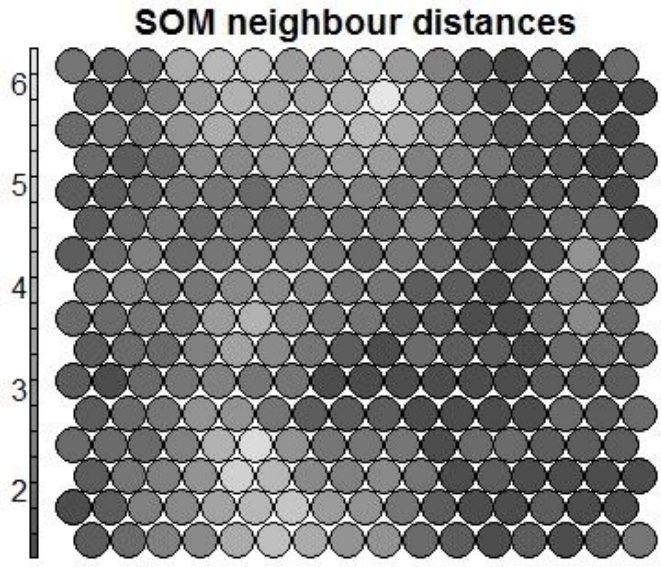


Figure 28: U-matrix

One of the outputs from SOM process is U-matrix. Whiter groups of neurons significate clusters while darker cells are borders. In assistance with hierarchical SOM, branches are divided into 7 clusters (Figure 28).

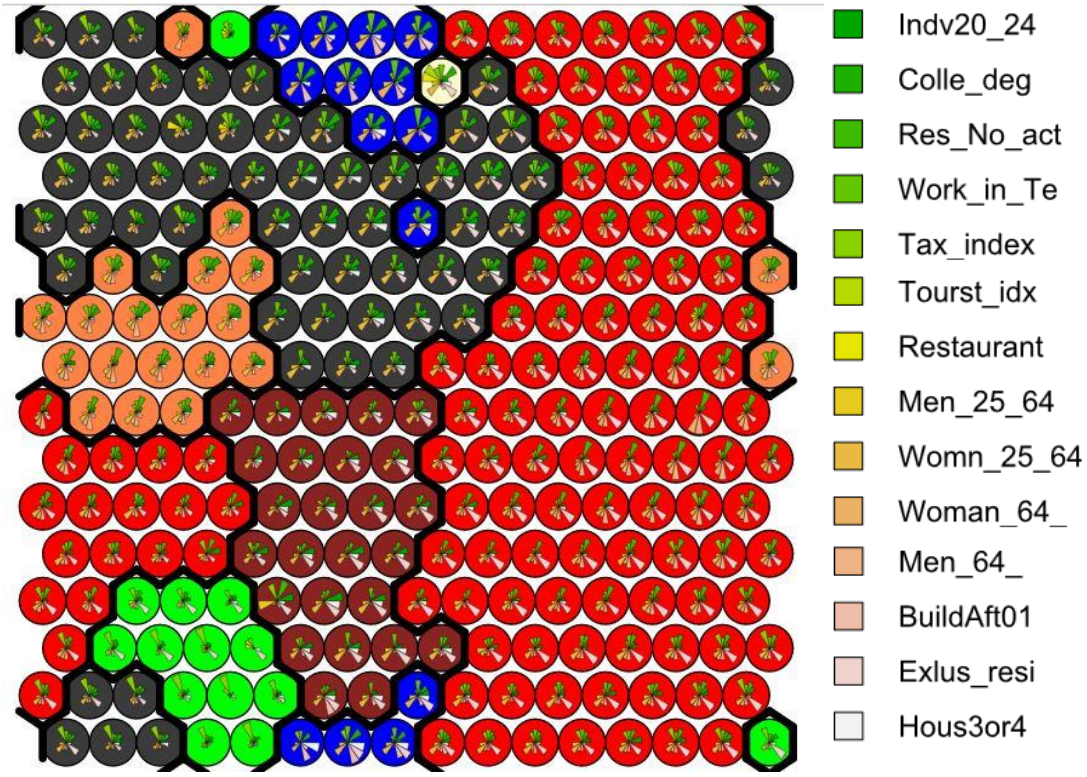


Figure 29: Clusters

Detail analysis from Figure 29 is presented in the Table 6. Description for the clusters is approximated by magnitude of variable. Legend highlights colours of variables within clusters.

Cluster / Var	Green	Blue	Red	Brown	Pink	White	Dark green
Indv20-24	.	*		*	*	.	*
College	.	***		.	***	***	***
Res_no_act	.	.			***	.	*
Work_in_tertiary	*	***		.	**	***	***
Tax_index	***	**		**	***	***	**
Tourist_idx	.	.			.	***	.
Restaurant	.	.			.	***	.
Men_25_64	*	*	*	**	*	*	*
Womn_25_64		**	**	*	***	***	***
Womn_64		.		.	**	.	.
Men_64		.		..	***	.	*
Buildaf01		***		***	.	*	.
Exlus_res	**	***	***	***	*	***	**
Hous3or4		***		**	*	*	*
	. insignificant		* low		** moderate		*** high

Table 6: Cluster analysis

5.1.3 Clusters' analysis

The description of presented SOMs are slightly different, and the geographical locations are quite similar as later explained. However, the general description for Lisbon study area cluster has common behaviour from all SOMs as follows:

- *White* cluster and the city section with "Vasco de Gama" shopping mall describes the most potential site, not even for restaurants, however in general for any kind of business. In addition, there is high level of tourist sites which is big attraction for restaurants. Persons who live there usually are employed in tertiary sector and college educated. Tax index is high and area is residential, as well. Also, according to the age group conclusion is that population is higher between 24 and 64 age.
- *Light* green cluster is also important for the further analyses. In regards to the demography, it has lower volume of age groups after 64 years. It is moderate residential area and the *Tax index* is higher.
- *Blue* cluster indicates higher level of college educated persons, newer buildings with 3 and 4 household members. The assumption is that these areas are family sites, since the sections are exclusively residential. Also, residents are mostly employed in the tertiary sector.
- *Red* cluster contains values related to unclassified cluster, thus the distribution is heterogenic. In general, sites are mainly residential with 1 or 2 household members and low 3 or 4.
- *Pink* cluster recalls the blue cluster, however the *Tax index* is higher, women from 25 to 60 and higher number of men older than 64 as well as household 3 or 4 members. Furthermore, the sections are low residential but inhabitants are mainly employed in tertiary sector.
- *Brown* cluster is with residential sections. According to demographic attributes it is less populated.
- *Dark green* cluster describes areas with higher *Tax index*, medium or higher employed, higher 1 or 2 household, low number of 20-24 ages and higher from 24-64.

5.2 Predictive modelling

5.2.1 Machine learning

In terms of predictive modelling several data mining techniques were tested to determine the best possible prediction. Some of the results are presented in Table 7.

Decision trees are extremely sensitive to small perturbations in data and a small change can result a drastically different tree. Also, they can easily overfit. This can be negated by validating methods or pruning it. However they could have problems in out-of-sample prediction (this is related to being non-smooth). The biggest problem is the lack of a principled probabilistic framework. Many methods have confidence intervals, posterior distributions etc., which give us some idea of how good a model is. A decision tree is ultimately an ad hoc heuristic, which can still be very useful (they are excellent for finding the sources of bugs in data processing), but there is the danger of people treating the output as "the" correct model. All mentioned weaknesses can be related with low accuracy for prediction in the analysis. *Decision Stump* or simply caller one R - one decision – one line is a weak learner with one level of decision tree, but if there are combined more weak learners it might have better results. Individually it has shown very low results.

J48 works on principle divide and conquer. It selects attribute, creates branch for each possible attribute and splits instances into subsets, repeats steps and eventually stops. It is one of the most popular method in decision trees, however this strategy did not present good results. Quite similar results were achieved with *REP Tree* and *LAD Tree*, fast learner and multiclass alternating decision tree, respectively.

M5P is a decision tree model with linear regression for each branch - linear patches approximate continuous function. It is executed for numeric values. It gave the first signal with implementation of linear regression rule, the accuracy can be drastically improved. Using cross validation with the full model correlation coefficient is 0.64 and with 75% of training set 0.81 accuracy.

In terms of distance learning, nearest neighbor methods are instance based. They work on principle of majority vote of its neighbours. An example is with piece-wised linear decision boundary, i.e. bunch of little linear pieces. This group of methods such as *IB1*, *IBk* and *LWL* presented in the table were not good for the prediction, since they consider that every attribute is equally important. It is important to mention that by having larger datasets the error declines to 0.

Important characteristics of probability methods is that all attributes are equally important and independently contribute to the model. *Naïve Bayes* is perfect example. The simplest function such as *Zero R* considers most likely class, most popular class and guesses it all the time. It is known as baseline accuracy. *One R* is extremely simple, it tests one particular attribute, one

branch for each value and then each branch assigns most frequent class. Finally, it chooses the attribute with the smallest error rate. In terms of implementation for restaurants, the results shown very low accuracy.

A multilayer perceptron (MLP) from artificial neural network is seen as logistic regression classifier. Weights are transformed using log-likelihood function. It transforms input data into linearly separable using non-linear transformation. This is called hidden layer. In the analysis it represented good results, i.e. 0.55 correlation coefficient or 0.81 accuracy in terms of 75% for training. However, it presented the highest RMSE error in testing. MLP was another indicator that regression rule improves the accuracy.

In addition, some ensemble learning methods were tested, as well. These methods improve performance for predictive models. *Bagging* works on principle where several training sets of the same size built a model for each one and then combine predictions by voting.

Bagging produces several models using some machine learning scheme and averages results. In other words they produce several models by using some machine learning scheme and average results. In the analysis *bagging* was applied with the best decision tree model with linear regression, i.e. *M5P*. The results were quite close to simple *M5P*.

In terms of *Randomization*, an example is decision tree method – *Random Forest*. It works on principle to select the best models and randomize them, while it bags the results to get better performance. As previously shown, the decision trees gave very low accuracy, hence *Random Forest* has shown low accuracy.

Boosting approach is an iterative process. It means that a new model is influenced from previous model. In other words *boosting* improves misclassified values and encourages the next model. Uses voting rule to bind them together. In the analysis *AdaBoostM1* is applied to improve MLP, however the results were lower accuracy than classic MLP. This is due to the reason that *AdaBoost* can be too sensitive for outliers and noisy data. It has problems with overfitting.

Linear regression methods shown the best results. As previously noted, an algorithm calculates the weights from training data and then multiply them with attributes. Weights are chosen on the way to minimize squared error on training data. They work well if there are more instances than attributes. With support vector machine (SVM) the idea is to create a line between clouds. Two closest dots, support vectors create perpendicular distance. SVM defines a boundary or plane in *n*-dimension. In real life border are not linearly separable. *SMO* and *LibSVM* are based on linear decision boundary and they are resilient to overfitting. Liner regression methods from *WEKA* presented accuracy in a range from 0.69 to 0.81 correlation coefficient, which was strong evident to obtain more sophisticated regression model by involving spatial component [56].

	Methods (Weka code names)	Accuracy		RMSE
		CV	75% training	
Decision trees	REP Tree	52.8	51.33	0.17
	J48	47.67	47.53	0.20
	LAD Tree	52.04	51.33	0.17
	Decision Stump	53.85	50.95	0.17
	M5P*[5]	0.64	0.81	2.27
Distance learning - lazy	IB1	40.17	38.78	0.23
	IBk (2)	50.33	48.29	0.20
	LWL	54.03	52.47	0.17
Probabilities	Zero R	53.94	51.33	0.17
	One R	50.05	46.39	0.21
	Naïve Bayes	38.56	38.77	0.19
	Naïve Bayes Multinomial	53.66	50.57	0.17
Regression	SMO	53.94	51.33	0.20
	SMO Regression*	0.69	0.8	2.31
	Simple Linear Regression*	0.68	0.81	2.20
	Linear Regression*	0.7	0.81	2.19
	Pace Regression*	0.69	0.8	2.20
Ensemble learning	Bagging (M5P)*	0.69	0.78	2.30
	Ada Boost (Multilayer Perceptron)	52.51	50.95	0.20
	Random Forest	53.65	49.81	0.16
	Classification Via Regression (M5P)	52.24	50.95	0.16
Neural	Multilayer Perceptron*	0.55	0.78	3.59

Table 7: Machine learning results: accuracy and RMSE

5.2.2 Regression models

The alternative for the prediction is found in the number of applications corresponding to generalized linear models (GLM) as well as general additive models (GAM)[6]. Roughly the methods are similar, however GAM emphasizes smoothness of the model. Each covariate used for prediction is maintained with smooth function. Basically, the dependent variable - response is specified for the prediction, other variables – covariates are multiplied with estimated coefficients and combined together were used for response prediction. The best combination of covariates corresponds to quality and accuracy of the model.

According to Table 8 the highest contributors to base GLM are all covariates, except households with active population which presented less influence.

[5]* Response variable is a numeric value

[6] R code can be found in ANNEX 6

BIC 2628	*Intrcpt*	*Tourst*	*Tax*	*Hous3or4*	*Exlusiv*	*Men25_64*	*Woman64*	*Active*
Estim.	-0.635	0.189	1.299	-3.739	-0.991	2.829	2.751	-2.187
p-value	0.33	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.01

Table 8: Base GLM model estimated coefficients

In regards to the GAM's covariant smoothness, the approximation of cubic regression is given to: *Tax_index, Household 1 or 2, Employed, Women above 64*. In addition for better fit in 95% of point-wise confidential area the covariates *Women between 20-24* and *Tourist index* were multiplied with square root function. The implication for the prediction can be seen from Figure 30.

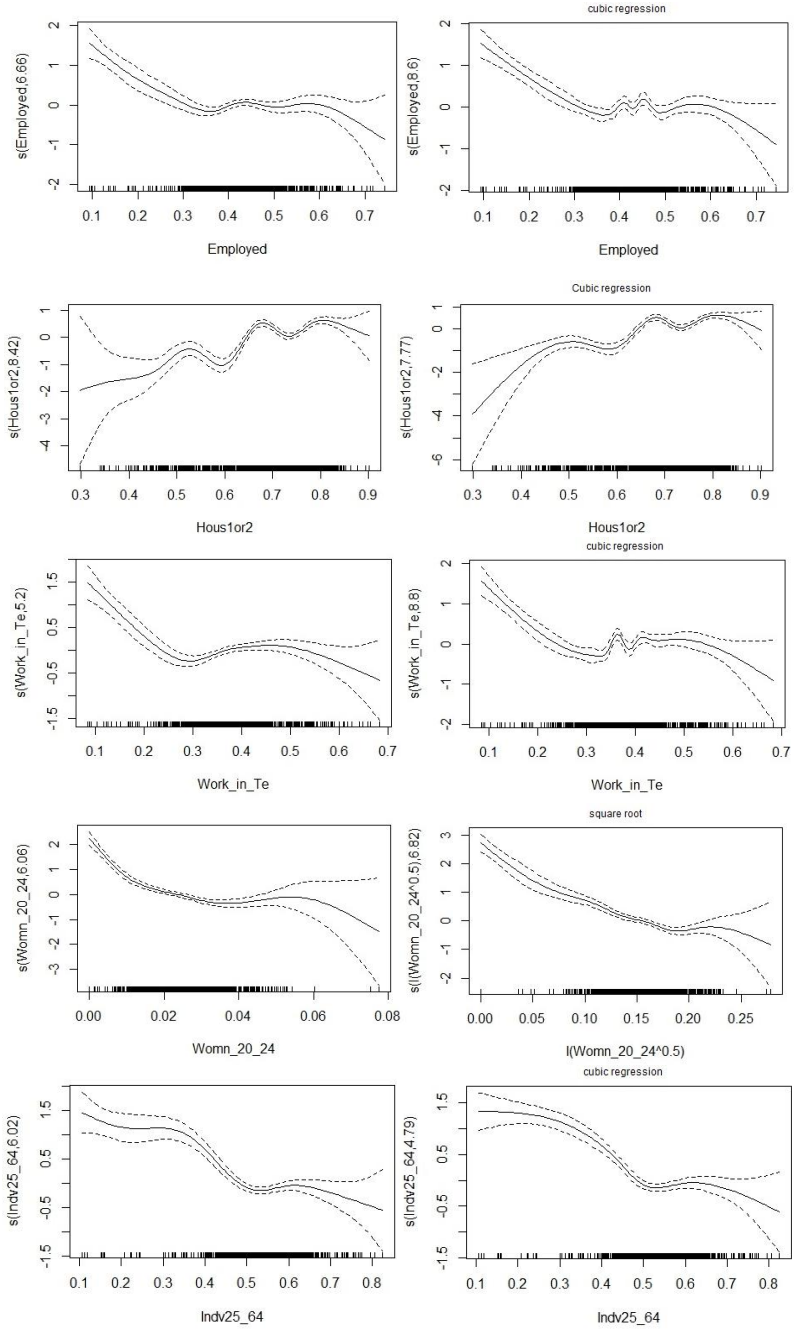


Figure 30: Compared smooth functions for covariates

The smooth functions on the right hand side are subjectively determined, since it can be seen a slight improvement of the curve in terms of 95% of confidence over optimal regression spline on the left hand side of the figure. In terms of assigning the most appropriate function to covariate, the overfitting has to be avoided and the function has to be as simplest as possible.

To recall, the components for the complete GAM are following:

- Based on formula from (5):
    a. *q(μ)* – response variable – *Restaurants*
    b. *xk* – covariates – all other variables
    c. *fi(xk)* – cubic spline smoother based on REML parameter estimation
- Link function from (3)
- Poisson: Discrete – {0}U N – Count - Asymmetric

The covariates for the full GAM model were selected in assistance to GLM as following: *Tourist index, Tax index, Household 1 or 2, Employed, Woman 20-24, College educated, Woman above 64, Residential without activity, Individuals from 25-64.*

According to the significance of smooth terms and Chi-square in the full GAM model, all covariates presented contribution to the prediction of restaurants, except *Woman in age 20-24*, and *Individuals in age 25-64.*

In terms of the model selection, from the group of 47 GAMs with various combination of covariates, based on lowest BIC stepwise forward selection the base GAM model is as follows:

$$logit(\mu) = f\left(Tourst_{idx}^{\sqrt{2}}\right) + f_1(Tax_{index}) + f_2(Hous_{1or2}) + f_3(Employed) \qquad (17)$$

Where: f – optimal regression spline

f_1, f_2, f_3 - cubic spline regression

The significance of the covariates is presented in Table 9. According to Chi-square, tourist sites have strong influence to prediction, while *Household with 1 or 2* members as well as *Employed* population share quite similar characteristics. In regards to BIC value, base GAM shows better performance. Also, the order of covariates is drastically different. In non-linear prediction *Household with 1 or 2* members as well as *Employed* population replaces older aged women, men, and *Household with 3 or 4* members in linear prediction. This is an important insight.

BIC 2519	***Intercept***	***Tourst_idx***	***Tax_index***	***Hous1or2***	***Employed***
coefficient/Chi square	-0.265	998.55	166.20	48.73	31.48
edf		7.59	6.56	2.13	2.84
df		9	9	9	9
p-value	0.0001	0.0001	0.0001	0.0001	0.0001

Table 9: Base GAM significance of smooth terms

It is important to mention that method for estimating parameters is restricted maximum likelihood (REML). The difference between maximum likelihood is that it penalizes out covariates which do not contribute to the model. An example can be seen when is applied to the full GAM model (Figure 31).

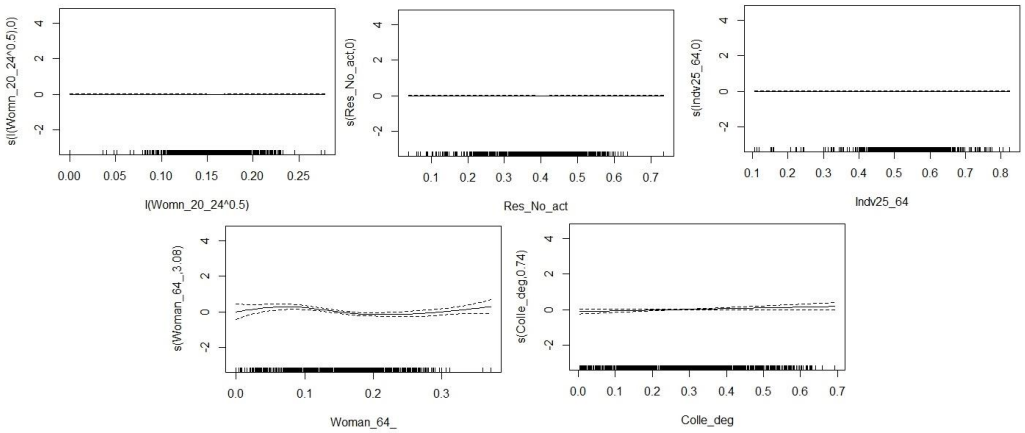


Figure 31: REML diminish effect on covariates

The effects of the covariates are presented on Figure 32. Figure *a)* indicates slow increase in tourist sites along with restaurants with a small local peaks at 2, 3.5. *Tax index b)* effect is flatten before 1.7 and after 3, and in between it slightly increases along with restaurants. The effect of the *Household 1 or 2 c)* members is flatten below 0.5, and starts after 0.8. *Employed d)* effect is non-linear with dull peak at 0.4.

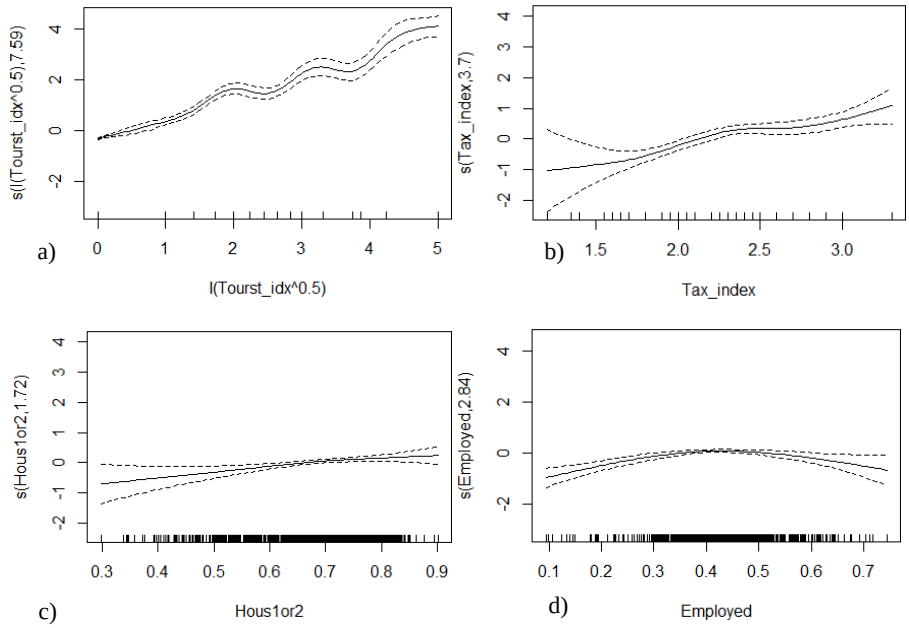


Figure 32: Effect of covariates

In order to validate goodness of fit to the base model it has to be compared with the full model. For that step the analysis of variance (ANOVA) is provided in the further analysis.

5.2.3 Neighbourhood criteria

Analysing Figure 15, Figure 16 the visual interference is that restaurants can be spatially related. Thus, the assumption is that city sections with restaurants have spatial autocorrelation. Once the assumption is proved, it is important to include this spatial relationship in a base model. The one way to incorporate this is with spatial weights. All city sections where the restaurants are located were tested with different methods. This time for the selection of the matrix based on neighbourhood is to use Akaike Information Criterion (AICc). Similar to BIC it is a mean for best model selection and it deals with the trade-off between the goodness of fit of the model and complexity of the model. The lowest AICc value indicates the most appropriate model to present spatial autocorrelation. AICc values from implemented methods for spatial aggregation (Section 2.3.4) in this study are presented in table:

	AICc values		**AICc values**
Rook	664	**Relative neighbour**	777
Queen	714	**1 nearest neighbour**	1154
Delaunay	929	**2 nearest neighbour**	297
Sphere of influence	773	**3 nearest neighbour**	553
Gabriel	809	**4 nearest neighbour**	667

Table 10: AICc values from neighbourhood matrices

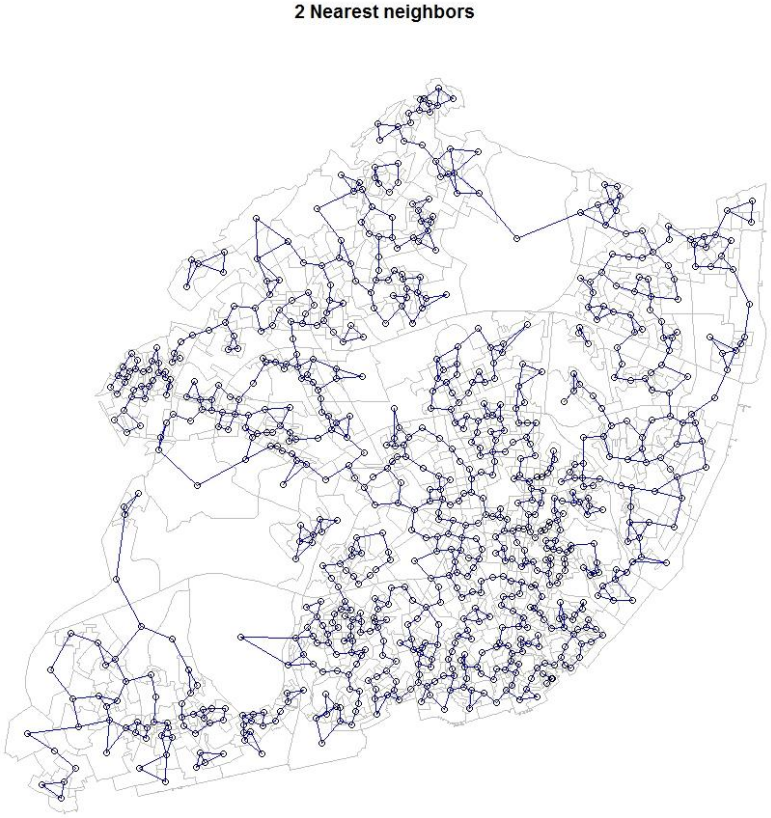


Figure 33: 2 nearest neighbour polygons relationship

The best neighbouring matrix presented by AICc value is the method with two nearest neighbours (Figure 33). The other plots can be found in ANNEX 7. In the further step the spatial weights are extracted in order to be tested for spatial autocorrelation. Global indicators for spatial autocorrelation for testing were conducted with *Geary* and *Moran* test.

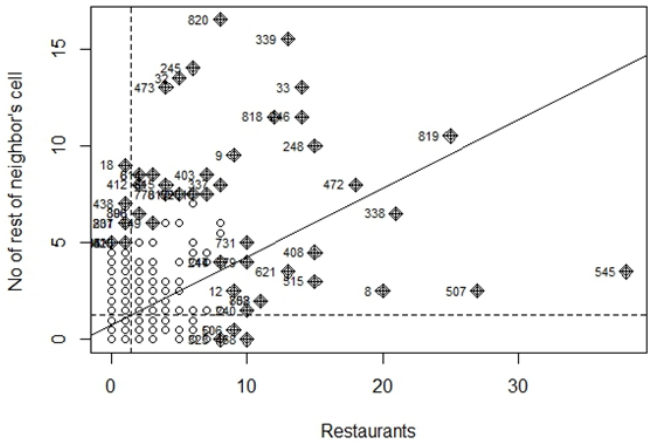


Figure 34: Moran global test for spatial autocorrelation – KNN2

Moran plot indicates visual explanation for spatial correlation. *X*-axis represents the response – restaurants and *y*-axis spatial lag, in other words number of neighbour's restaurants. Points in the upper right (or high-high) and lower left (or low-low) quadrants indicate positive spatial association. The lower right (or high-low) and upper left (or low-high) quadrants include observations that exhibit negative spatial association; that is, these observed values carry little similarity to their neighbourhoods (Figure 34, Figure 35).

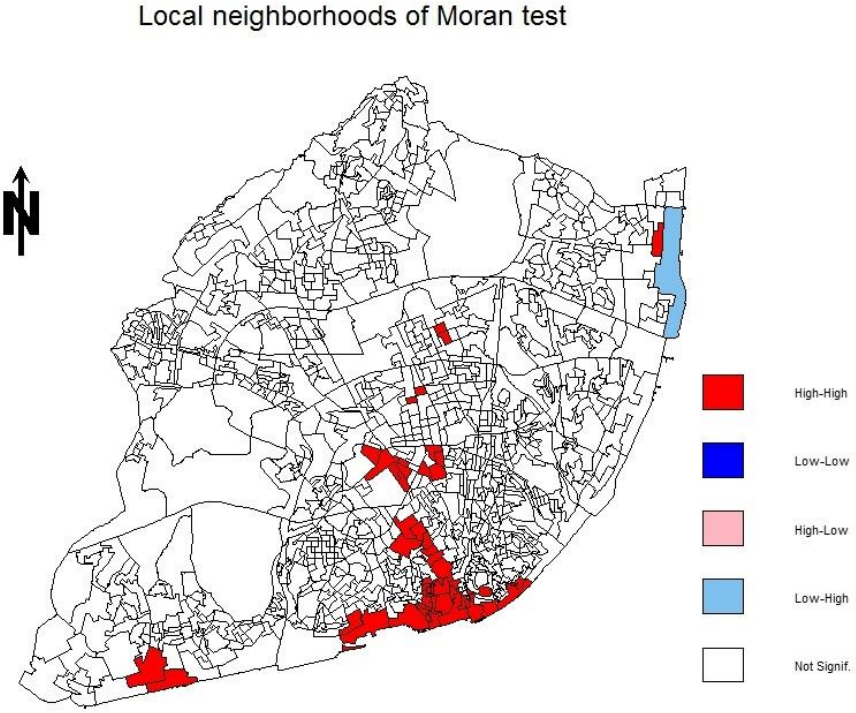


Figure 35: Local neighbourhoods of Moran I test

From the plot one can conclude that spatial autocorrelation is present. *Low-High* polygons indicate high number of restaurants with low number in neighbourhoods and red polygons, high number of restaurants both in neighbourhoods and in its own city section. In addition, Moran and Geary test showed *p*-value < 0.001 which indicates that *Ho* hypothesis is rejected, in other words spatial autocorrelation indeed exists.

Following this indication, it is proved that spatial autocorrelation exists and in the further step spatial weights were involved in the base GLM model.

5.2.4 Spatial model construction

In terms of GLM, *Moran eigenvector* is a statistical approach to prove the existence of spatial independency. *Moran eigenvector* defines how spatial weights can be incorporated into a model, which is why it is implemented to fit all variables from complete GLM model. Hence the GLM with spatial component is following:

$$log(\mu) = \beta_0 + \beta_1 Tour_{index} + \beta_2 Tax_{Index} + \beta_3 Hous_{3or4} + \beta_4 Exlusi_{resid} + \beta_5 Men25_{64} + \beta_6 Woman64 + \beta_7 Active + ME \quad (18)$$

Where:

ME – Moran eigenvectors of base GLM with spatial weights

In spatial GLM, active population is penalized out, since *p*- value is 0.13 compared to GLM. BIC value for this model is: 2459.662

As it is previously proved that spatial autocorrelation exists, the *X, Y* coordinates of centroids of the city sections with restaurants were incorporated into a base GAM with a smooth two dimensional function thin plate spline regression. This represents a spatial component for base GAM. Hence the upgraded GAM and the formula for spatial GAM is following:

$$logit(\mu) = f\left(Tourst_{idx}^{\sqrt{2}}\right) + f_1(Tax_{index}) + f_2(Hous_{1or2}) + f_3(Employed) + f_4(Y_c, X_c) \quad (19)$$

Where: f_4 – thin plate spline smother function

Yc, Xc – centroids of city section with restaurants

In Figure 36 the densities of restaurants are on several locations, but the highly dense are at *X* - 86000 and *Y* – 10000 and *X*= -88000 and *Y* = -102000.

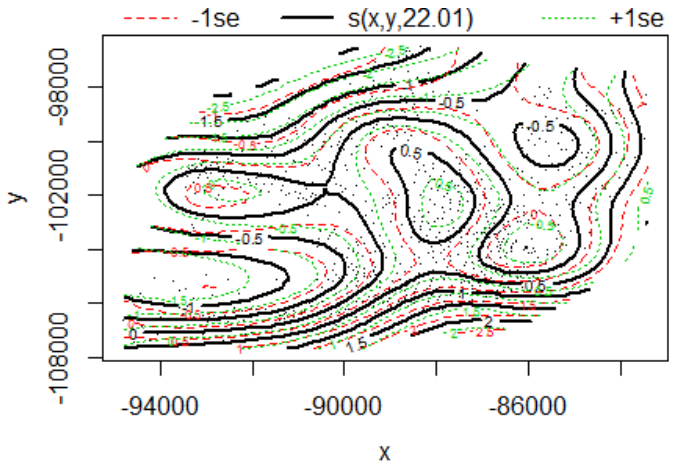


Figure 36: Effect of spatial covariate

From Table 11 Chi-square value significates the importance of spatial component. BIC value is lower, and *p*-value indicates significates of each covariate.

BIC 2453	***Intercept***	***Tourst_idx***	***Tax_index***	***Hous1or2***	***Employed***	***s(X, Y)***
coefficient/Chi square	-0.265	998.55	166.20	48.73	31.48	186.32
edf		7.54	3.67	1.72	2.17	22.01
df		9	9	9	9	29
p-value	0.0001	0.0001	0.0001	0.001	0.0001	0.0001

Table 11: Spatial GAM

5.2.5 Model selection

The final step refers to the set of tools for the best model selection. Previously created models for prediction are candidates for the best model. So the models included are:

- Base GLM
- Spatial GLM
- Base GAM
- Spatial GAM

These models are based on 80% dataset, hence it is necessary in some of the further steps to predict 20% of test. Nevertheless, the basic summary for all models is presented in Table 12:

	GLM	**spatialGLM**	**GAM**	**spatialGAM**
R^2	0.719	0.810	0.814	0.858
BIC	2668.98	2453.69	2519.52	2452.84
df	8	19	23.18	42.07

Table 12: Summary of candidates

In the following table we tested spatial GLM against GLM as well as spatial GAM against GAM with ANOVA test:

	Res. Df	Res. Dev	df	Deviance	p-value
GLM	833	1535.4			
spatialGLM	822	1286.6	11	248.78	0.0001

	Res.df	Res.dev	df	Deviance	p-value
GAM	820.89	1324.2			
spatialGAM	802.86	1130.3	18.032	193.86	0.0001

Table 13: ANOVA test

Analyzing *p*-value *Chi-square* test the conclusion is that both spatial models made significant improvement over the base models. Given the comparison between models according to BIC value over the spatial models, both spatial models have better performance than base models. Finally, both spatial models were compared with ANOVA test (Table 14).

	Res.df	Res.dev	df	Deviance	p-value
spatialGLM	822	1286.6			
spatialGAM	802.86	1130.3	19.143	156.23	0.0001

Table 14: ANOVA test for spatial models

GAM with spatial component gives better performance because both deviance and test statistic rejects spatial GLM. From Table 12, BIC shows lower value for spatial GAM, as well.

Based on ANOVA tests the best model is spatial GAM, however to make sure that right model is chosen, the models are validated with the rest of 20% testing. Therefore in the next step models were utilized to predict the testing dataset in terms of number of restaurants (Table 15).

	R^2 80%	R^2 20%	RMSE
GLM	0.719	0.585	2.18
GAM	0.813	0.54	1.78
spatialGAM	0.868	0.602	1.57

Table 15: Results from validation of 20%

GAM model predicted with *R-square* of 0.81 the number of restaurants for the construction of the model, but with *R- square* of 0.54 the model validated testing set of 20%. *Root mean square error* is 1.78. In addition, spatial GAM for the construction showed 0.86 R-square and for the testing dataset 0.60 accuracy. RMSE is 1.57. Referred to that analyses the interference is that spatial GAM presents the best qualified model for the prediction and in further analysis the whole dataset. Hence, the formula is utilized to predict the number of restaurants and to calculate the deviance residuals. In general, *Chi-square* test indicated strong rejection of null hypothesis in independency of variables towards response, which indicates significance of the formula 19. Meantime, the conclusion is that covariates: *Tourist index, Tax index, Household*

with 1 or 2, *Employed* as well as spatial component have the highest importance to the prediction of restaurants and the same time, the indicators for potentiality. Likewise, the formula extracted from 80% of dataset was used to predict restaurants from the whole dataset, as well to measure a RMSE. Figure 37 presents standard various residuals plots.

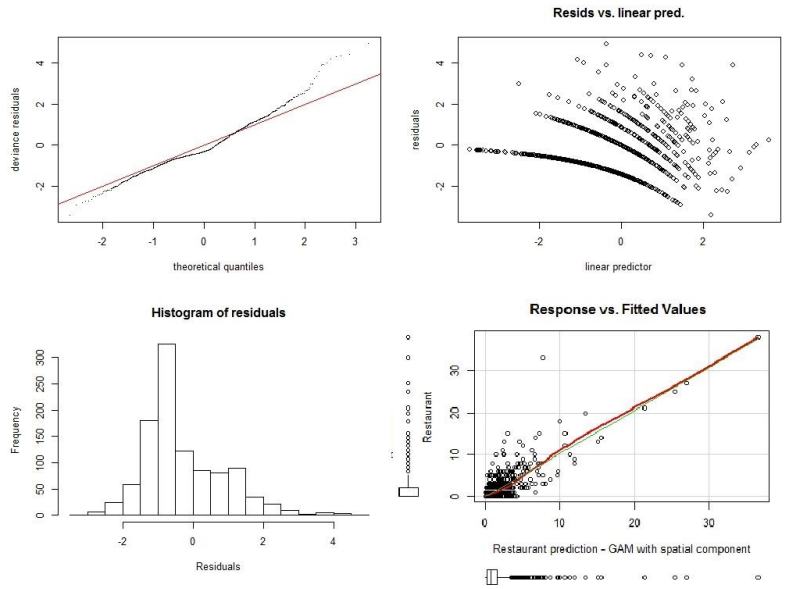


Figure 37 Diagnostic information about prediction results

Q-Q plot proves that deviance residuals against theoretical quantities according to the fitted model follows normal behavior. According to the scatterplot correlation coefficient is 0.808, however histogram of residuals slightly deviates from normal distribution.

5.3 Restaurant potentiality estimation

The deviance residuals were utilized to assess the city sections in regards to potentiality.
The summary of deviances is as follows:

Min	1st	Median	Mean	3rd	Max
-3.421	-1.011	-0.6383	-0.2897	0.3895	4.939

Table 16: Summary for deviance residuals of restaurants

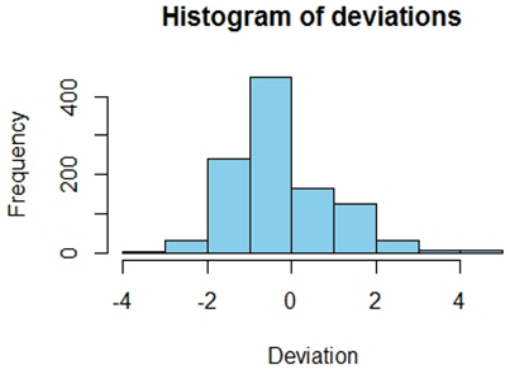


Figure 38. Histogram of deviations

According to an empirical rule (Figure 39) an approximation for bell-shaped relative frequency defines [57]:

- ~ 70% of residuals are in a range: $[(\mu - \sigma), (\mu + \sigma)]$
- ~ 95% are in a range: $[(\mu - 2\sigma), (\mu + 2\sigma)]$, and
- ~ 99.7% are in a range: $[(\mu - 3\sigma), (\mu + 3\sigma)]$

Hence, the empirical test to determine cut-off points is applied as following:

- In the first category the deviations are negligible, i.e. city sections contain *no change* around 75%
- The second category is with medium overestimation (*left tail*) and medium underestimation (*right tail*) range from 75%-95%
- The third category are deviations above 95%

In other words, threshold cut-offs values can be found in Table 17.

$\mu - 2\sigma$	$\mu - \sigma$	$\mu + \sigma$	$\mu + 2\sigma$
-2.63	-1.46	0.87	2.04

Table 17: Threshold values

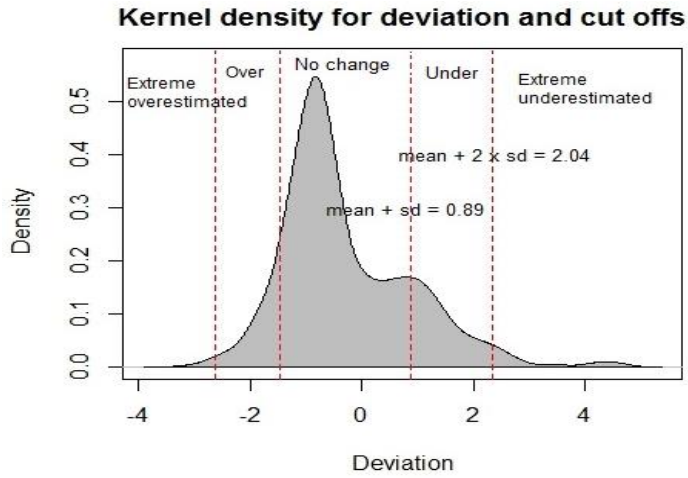


Figure 39: Deviance residuals distribution with cut-offs values

Consequently, the cut-offs values can be approximated on the geographic map and those sections can be identified, in order to visually differentiate higher potential and lower potential city sections from the once that have no change.

6 FINAL ANALYSIS AND DISCUSSION

SOM method was not highlighted values of tourist sites in clusters, therefore a density map of restaurants and tourist sites were overlapped over the cluster map. Clusters and city sections were associated on the geographic map and they were presented on Figure 40. Clusters allocation can be seen with spatial autocorrelation. For this reason, Mantel test is implemented for testing the presence of spatial autocorrelation. Distance between locations (w – vector) is associated with difference among centroids of city sections. Value (u - vector) were associated

with codebook vectors summary for each neuron (256), afterwards distance between each city section and corresponding neuron were subtracted. In addition difference between values is associated with *u*- vector in Mantel test. With *p* – value of 0.028, null hypothesis is rejected, hence it can be said that there is a scheme of spatial clustering in resulting SOM. Consequently, one can say that socio-demographic characteristic from one city section significantly depends on neighbouring city section.

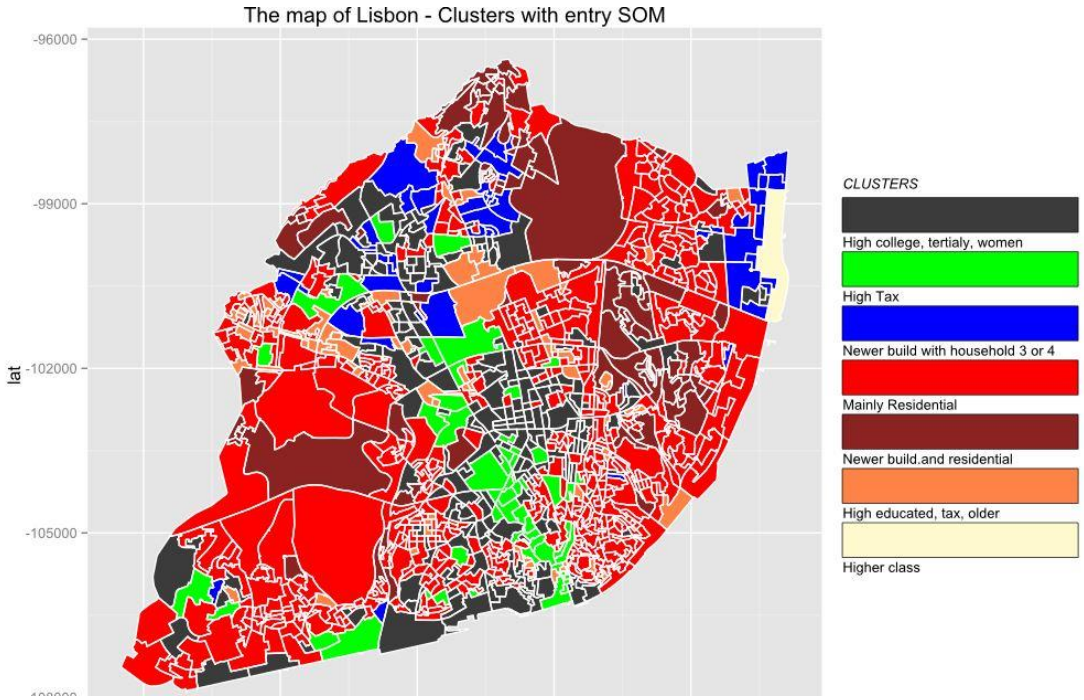


Figure 40: SOM clusters and thematic map

From the map, at a glance one can locate white clusters and high density number of restaurants. The site is as expected associated with high density both tourist and restaurant areas because at the location of the site is one of the most visited shopping malls in Lisbon. Therefore the catchment area of the shopping mall have an influence on the neighbours, i.e. specifically blue cluster. This is due to the evident local spatial autocorrelation as well. To recall, white cluster contains high tax index, high education level and high employed in tertiary sector, therefore is it not surprising to have a high number of overcrowded restaurants (Table 6).

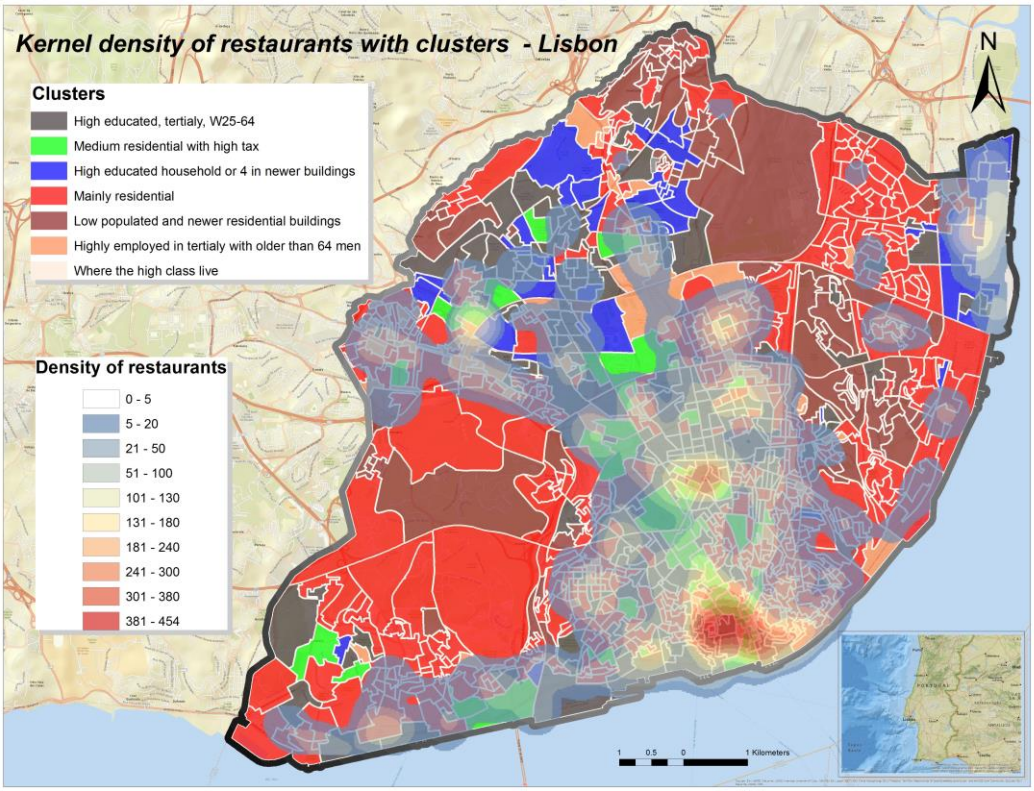


Figure 41: Cluster map with restaurants density

On the potentiality map cluster from upper-level population and city section has no change (Figure 43).

In regards to the blue cluster the assumption is that it indicates young family sites, since the number of buildings built after 2001 and household are with 3 or 4 members is high. The areas with restaurants are partially overlapped (Figure 41). On Figure 42 it can be seen tourist density overlaid on cluster map. From Figure 41 can be identified in areas on the north which are partially overlapped and the ones that are not. This is due to the fact that city sections seem strongly residential. From the imagery map the sites contain a huge parks and green open areas as well such as outdoor fields of University of Lisbon, huge parks in Luminar, parks around *Oriente*. In comparison with potentiality map, only few city sections around *Vasco de Gamma* shopping centre indicate moderate underestimated and one section overestimated, i.e. only one section could have more restaurants. However, in general blue cluster indicates no change in terms of number of restaurants.

Green clusters are also in most cases covered with restaurants density, but only in downtown area. This cluster does not have change in terms of restaurant potentiality, except for in a few sections such as *Jardim de Santos*, which indicates extreme underestimation. This is due to the

fact the location is an exception, since it reminds on industrial zone, however one small part of the section is crowded with restaurants.

As stated in section 5.1.3 light green cluster is highly enriched with tourist sites, so it is with reason attached with restaurants. The very similar results presents dark green cluster. An interesting information from the map which is related to one of the questions of the thesis, is that areas with restaurants without tourist could be identified at the dark green cluster. Red clusters have no clear description, and from the map we can see that there is not much correlation with restaurants, with the exception of sections next to the other clusters. This cluster refers to unclassified components from U-matrix, although it can be found that sections are "*Exclusively residential*".

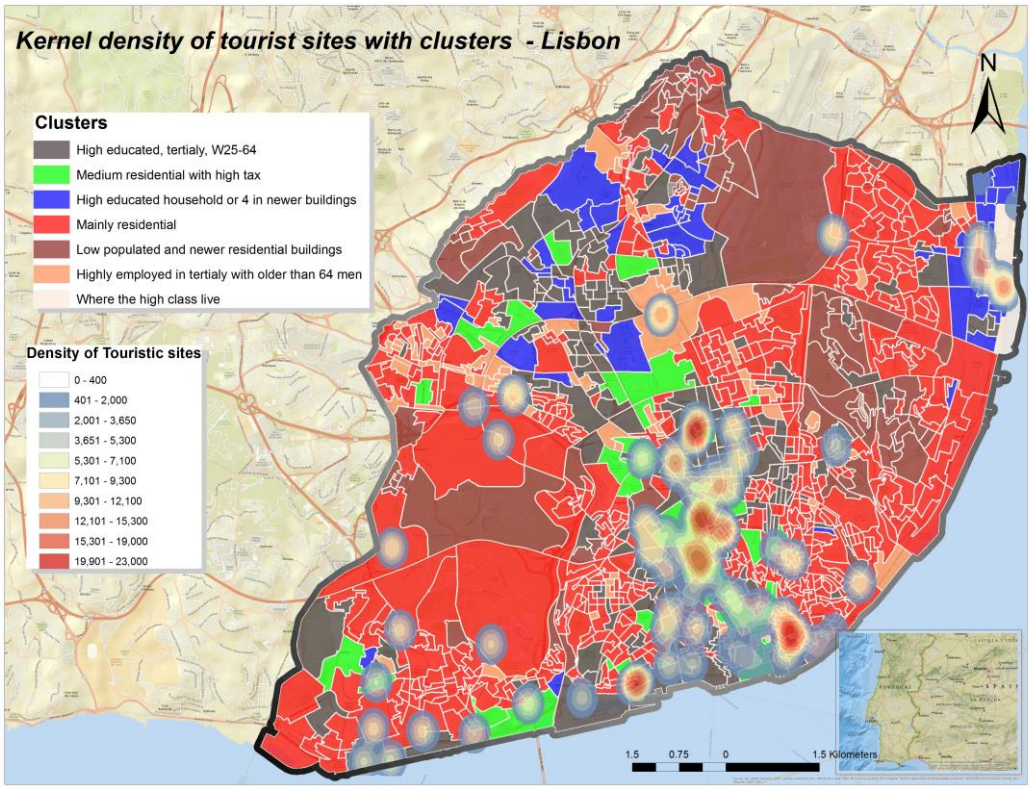


Figure 42: Cluster map with tourist site densities

The brown cluster from the map does not consist of any restaurants. So, as a conclusion less populated city sections are not attractive for restaurants and tourist sites in particular. Pink cluster shows that is almost entirely linked with restaurants, but not much with tourist sites. Here is located a bit elder population with higher *Household 3 or 4*. In addition, it has higher tax values which is opposite from the previous clusters e.g. green cluster, where high tax index indicates spatial correlation with restaurants. The combination of variables of the pink cluster do not seem visually correlated with *Tourist index*, as one can observe from the map, although the cluster is characterized with higher tax as well as *Household with 3 and 4*. The reason why

can be identified that there are high densities of restaurants is that the pink cluster is located around football stadiums of *Benfica* and *Sporting*. Here the prediction estimates moderate underestimation, in other words, there is already enough restaurants.

For the purpose of easy access and in-depth visualisation interaction of the map, the layers can be found on ArcGIS Online[7].

In addition to potentiality assessment of the restaurants, the final results according to the spatial GAM does not indicate spatial autocorrelation, since residuals are scattered around the whole map, which significates the presence of an individual independence in residual [58] (Figure 43).

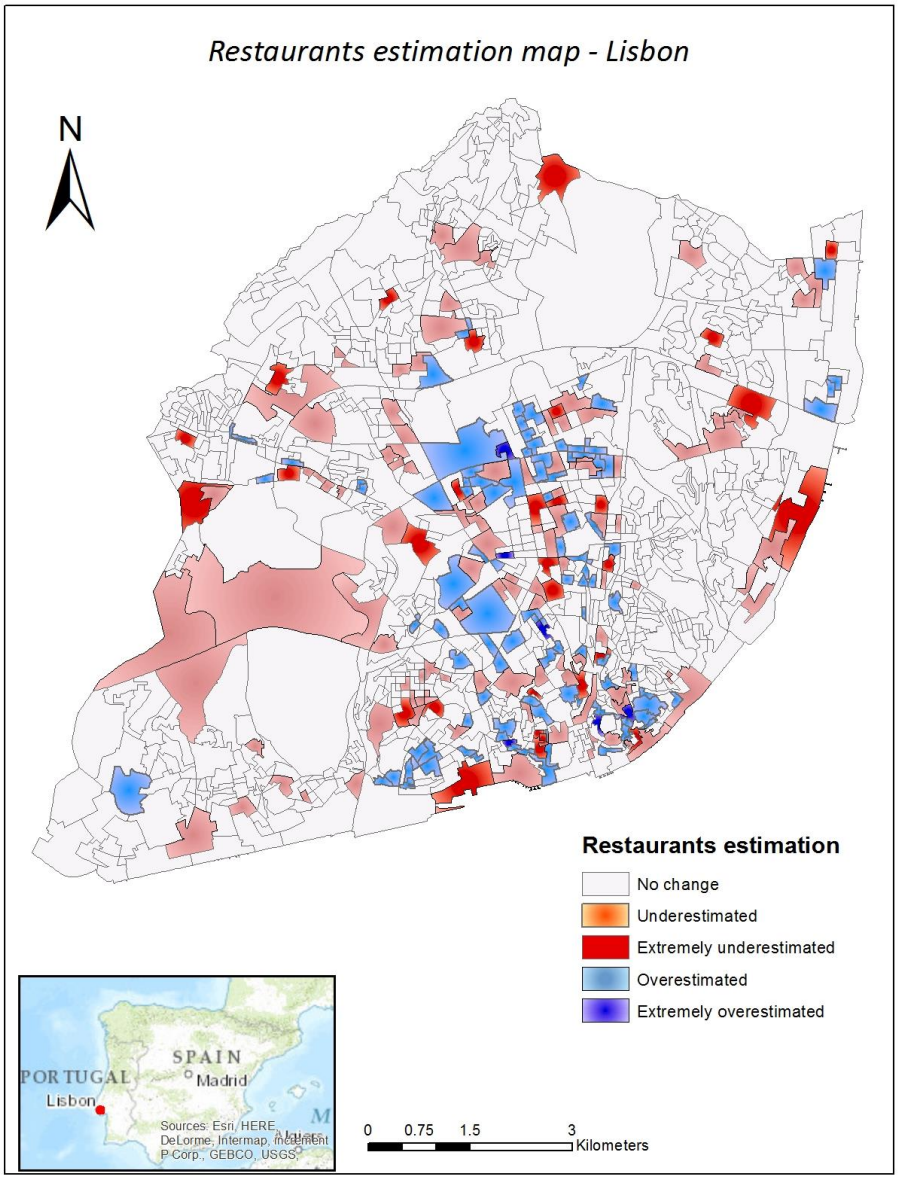


Figure 43: Restaurants potentiality map – Lisbon 2015

[7]ArcGIS Online map can be found at arcg.is/1SSTjNN

Furthermore, according to the cut-offs in terms of potentiality of restaurants:

- 825 city sections are with no change
- 194 city sections with are either moderate underestimated (101) or overestimated (93)
- 34 sections are extremely underestimated (29) and extremely overestimated (5)

From the map it can be seen that the extreme overestimated city sections are around the downtown area and these section have no restaurants, but prediction shows that are highly potential and estimation is on 5,6 restaurants. Analyzing with Imagery background map the areas are highly populated and have very low green cover. In prediction there are only a few, and it is interesting that two of them are located around *Castelo de Sao Jorge*, which is a very popular tourist site in Lisbon. Other areas are in *Bairro Alto*, which is a very popular tourist area. In addition, an extreme underestimated areas there are high numbers of restaurants but prediction estimates much less. This is due to the fact that areas are mainly industrial, or big green sites, in other words where there is a low population. These areas can be found at *Marechal Gomes da Casta, Compo de Amoreiras Monsanto*. Since the method highly depends on a tourist index, it can be identified that the sites are in the city underestimated such as *Arroios, Compo Pequeno*. Those sites already have large number of restaurants, however the method indicates that they are overcrowded. Moderate overestimated category can be seen for decision making, firstly because the sections are located all over the market areas, downtown areas and tourist areas. Prediction indicates higher moderate number, therefore areas have more potential, but not drastically. This requires an in-depth analysis of the site. In regards to moderate underestimated areas, those areas cover less populated sites, such as airport, large green sites such as; *Parque de Monsanto*, industrial zones, and very small sites with urban areas. *Santa Apolonia* railway station by the analysis is moderately underestimated, since the site is not touristic attraction, however the number of restaurants is uncorrelated.

7 CONCLUSION AND FUTURE WORK

In this project we conducted an in-depth exploratory spatial data analysis with utilization of several methods gathered around SOM in order to discover patterns and form a socio-demographic segmentation platform in regards to restaurants potentiality.

In order to improve a descriptive analysis a several predictive models were tested to determine an accurate prediction for restaurants. With 0.868 correlation coefficient, 1.57 RMSE and 2453.284 BIC value, the best performance were shown by GAM with inclusion of spatial attribute. The predictive and existing number of restaurants were used for estimation of restaurant potentiality by categorizing deviance residuals into groups of overestimated and underestimated city sections.

Based on results provided from segmentation and potentiality map the conclusion is that the most potential city section from the segmentation map do not change behavior in terms of restaurants potentiality. Those sections have high *Tax index*, high *Tourist index*, areas are residential with higher population between 24 and 64 and dense inhabited population with *Household 3 or 4*.

However in the city sections with high *Tax index* and *Household with 1 or 2* members and moderate employment rate, there were identified restaurants without tourist sites. These areas are located around downtown sites. Here is identified moderate underestimated in terms of restaurants potentiality.

In addition, according to the model the highest influence in restaurant potentiality is given to tourist sites, spatial autocorrelation in terms of neighboring restaurants (spatial component), and *Tax value*, where lower importance is given to *Household with 1 or 2* members and *Employed* population, respectively.

During analyzes it is concluded that many extreme underestimated sites contain large green sites. Furthermore, the segmentation map didn't provided clear representation about green sites, since the green sites were identified in all clusters. For the future work NDVI values can be incorporated into a SOM model in order to better differentiate urban and green sites. Also, one of disadvantages is that it was not possible to extract formula from 80% of training GLM with spatial component in order to apply for the full dataset. The main problems is that vectors of spatial weights from 80% of dataset cannot be applied for the full dataset.

Every tested predictive model and finally selected, in this analysis presented high dependency of restaurants against tourist sites. These sites are mainly around the most attractive locations in the city. For the further analysis, beside the base and spatial predictive model, distance model can be applied. Hence, shopping malls and other attractive city locations can be assigned as focus areas, while for example restaurants or other tourist sites may form a pattern with some other city features such as location of business retails.

BIBLIOGRAPHY

[1] EBD, “Top destinations in Europe,” http://www.europeanbestdestinations.com, accessed 20/12/2015, Lisbon, 2015.

[2] C. Bremner, “Euromonitor International,” Euromonitor International's Top City Destination Ranking, London, 2011.

[3] O. Raymond, “7 reasons to visit Lisbon right now,” accessed: 25-12-2015, Conde Vast Traveler, 2014.

[4] R. Turner, “Travel&Tourism: Economical Impact 2014 Portugal,” World Travel and Tourism Council, 2014.

[5] INE, “Estatísticas do Turismo 2013,” INE, Lisbon , 2013.

[6] K. W. &. J. B. Beverley Sparks, “Restaurants as a contributor to tourism,” Sustainble Tourism, Australia, 2002.

[7] N. A. Erik Cohen, “Food in Tourism, Attraction and Impediment,” *Pergamon,* pp. 755-778, 2004.

[8] M. A. Khan, Restaurant franchising, New York: John Wiley & Sons, Inc., 1999.

[9] L. B. Bill Ryan, “Downtown and Business District Market Analysis,” Wisconsin Department, 2011.

[10] L. R. L. D. Quirk Neil, “Restaurant site location,” Michigen State University, Michigen, 1978.

[11] K. Park, “Identifying site location factors,” accessed: 20-12-2015, Virginia Tech, scholar.lib.vt.edu, Blacksburg, 2001.

[12] P. G. J. I. K. R. B. P. H. W. R. D. D. J. E. Heikkila, “What happened to the CBD - distance gradient?: land value in a policentric city,” in *Enviroment and Planning A*, Los Angeles, 1989.

[13] C. Jog, “Predicting Restaurant Revenue using Ensemble Methods in Machine Learning,” Clonegy, 2016.

[14] W. Farhan, “Predicting Yelp Restaurant Reviews,” UC San Diego, La Jolla, 2014.

[15] M.-H. T. J.-J. C. S. O. Gwo-Hshiung Tzenga, “Multicriteria selection for a restaurant location,” *Pergamon,* pp. 171-187, 2002.

[16] Y. Z. Xiangyi Lin, “Multi-criteria GIS-based procedure for coffee shop location decision,” accessed: 25-11-2015, DiVA, Borås, Sweden, 2013.

[17] M. M. C. W. Hillary L. Burdette, “Neighborhood playgrounds, fast food restaurants, and crime: relationships to overweight in low-income preschool children,” in *Preventive Medicine*, Cincinatti, 2004.

[18] L. G., “Utilizing GIS Based Site Selection Analysis for Potential Customer Segmentation and Location Suitability Modeling to Determine a Suitable Location to Establish a Dunn Bros Coffee Franchise in the Twin Cities Metro, Minnesota,” University of Minnesota, Minneapolis, 2009.

[19] S. W. J. M. Robert Haining, “Exploratory Spatial Data Analysis in a Geographic Information System Environment,” in *JSTOR*, Sheffield, 1998.

[20] L. Anselin, “Interactive techniques and exploratory,” *Geographical Information Systems: Principles, Techniques, Management and Applications,* pp. 253-266, 1999.

[21] K. Pearson, Contribution of Mathematical Theory of Evolution, London: Royal Society, 1895.

[22] D. Brillinger, Explonatory Data Analysis, Berkley: Addison, 1977.

[23] T. Kohonen, "Self-Organization Formation of Topologically Correct Feature Maps," *Bibliogical Cybernetics*, pp. 59-69, 1982.

[24] A. S. Pragaya Agarwal, Self-Organising Maps, Applications in Geographic Information Science, London: Wiley, 2008.

[25] R. Henriques, "Self Organized Maps," NOVA-IMS, Lisbon, 2010.

[26] F. B. V. L. Miguel Loureiro, "Fuzzy Classification of Geodemographic Data Using," NOVAIMS, Lisbon, 2005.

[27] E. Koua, "Using Self-Organized Maps for infromation visualisation and knowledge discovery in complex geospatial dataset," Document Transformation Technologie, Enschede, 2003.

[28] V. L. P. Fernando Bacao, "The self-organizing map,the Geo-SOM,and relevant variants in geoscience," in *Computers and Geosciences*, Lisbon, 2005.

[29] G. B. R. K. Jari Oksanen, "Package 'vegan'," CRAN, accessed: 15-01-2016, 2016.

[30] R. H. Victor Lobo, "Spatial Clustering with SOM and GeoSOM," in *International Conference on Advanced Geographic Information Systems, Applications, and Services*, Lisbon, 2010.

[31] E. Koua, "Using SOM for information visualisation and knowledge discovery in complex geospatial datasets," in *International Cartography Conference*, Durban, 2003.

[32] V. L. Jorge Gorricha, "Improvements on the visualization of clusters in geo-referenced data," in *Computers &Geosciences*, Lisbon, 2012.

[33] J.-C. T. Jun Yan, "Visual Exploration of Spatial Interaction Data with Self-Organized Maps," in *Self-Organising Maps Applications in Geographic Information Science*, London, John Wiley and Sons, 2008, pp. 67-87.

[34] C. R. Andrew Lee, "Visualizing urban social change with Self-Organizing Maps: Toronto," *Habitat International*, pp. 92-98, 2015.

[35] N. S.-N. L. a. J. C. Wenxue Ju, "Application of Kohonen Self-Organizing Map for Urban Structure Analysis," in *IEEE International Conference on Granular Computing* , Atlanta, 2006.

[36] E. P. V. G. Roger Bivand, Applied Spatial Data Analysis in R, New York: Springer, 2013.

[37] W. R. Nelder J.A., "Generalized Linear Model," *Journal of the Royal Statistical Society.*, vol. III, no. 135, pp. 370-384, 1972.

[38] G. Rodriquez, "Lectures Notes on Generalized Linear Models," [accessed: 25/12/2015] www.data.princeton.edu, Princeton, 2007.

[39] H. Liu, "Generalized Additive Model," University of Minnesota Duluth, Duluth, 2008.

[40] S. Wood, "Package 'mgcv'," accessed: 20-12-2015, R-project, London, 2016.

[41] H. J. Miller, "Tobler's First Law and Spatial Analysis," Department of Geography, University of Utah, Salt Lake City, 1978.

[42] E. P. R. G. Rober Bivand, Applied Spatial Data Analysis with R, New York: Springer, 2008.

[43] D. G. B. B. M Tiefelsdorf, "A variance-stabilizing coding scheme for spatial link matrice," *Environment and Planning*, vol. I, no. 31, pp. 165-180, 1999.

[44] S. L. P. a. P.-N. P. Dray, " Spatial modeling: A comprehensive framework for principle coordinate analysis of neighbor matrices (PCNM)," in *Ecological Modelling*, 2006.

[45] R. Bivand, "Spatial Filtering - Moran Eigenvector," R-project, 2015.

[46] P. R. C. P. e. a. Dray S, "Community ecology in the age of multivariate multiscale spatial analysis," in *Ecological Monographs*, Washington, 2012.

[47] A. A. J. E., "The Bayesian information criterion: background, derivation and applications," *Month*, pp. 199-203, 1 March 2012.

[48] B. Li, "Bootstrap and Permutation," *Lecture 16*, pp. 1-6, 1 April 2015.

[49] B. Larget, "Cp, AIC, and BIC," *Statistics 333*, 7 April 2003.

[50] M. S. H. M. S. R. Sanjay Chaudhur, "Fitting Generalized Linear Models Subject to Constraints," Cran-R, Washington, 2015.

[51] J. M. I. T.-U. Carlos Ayyad, "Spatial modelling of rat sightings in relation to urban multi-source focusses," Castellon, 2016.

[52] R. v. d. M. a. S. B. Maria Mahfoud, "Spatio-temporal modelling for residential burglary: scale and cut-off selection using a neighbourhood verification approach," Amsterdam, 2016.

[53] W. T. W. W. P. C. Hua HE, "Structural zeroes and zero-inflated models," in *Shanghai Archives of Psychiatry*, Shanghai, 2014.

[54] Demographia, "Demographia World Urban Areas," Demographia World Urban Areas, Mumbai, 2015.

[55] V. L. Fernando Bacao, "Fundamentals of data preparation and pre-processing," NOVA-IMS, Lisboa, 2007.

[56] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Tauranga: Morgan Kaufman Publisher, 2005.

[57] A. Schmitz, Beginning Statistics, Creative Commons, 2012.

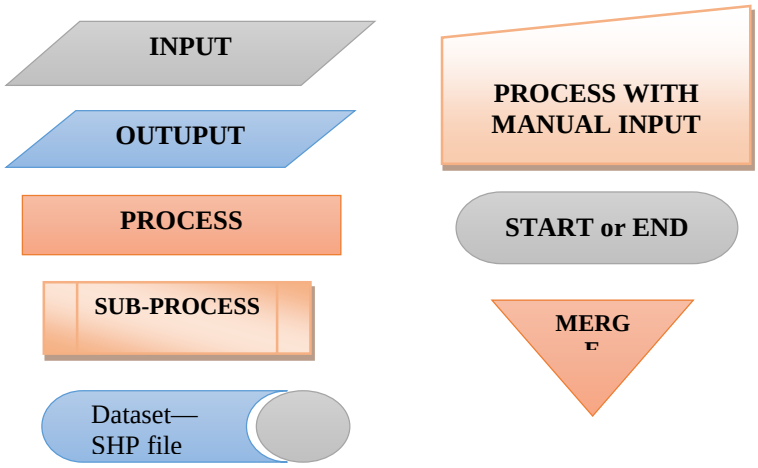
[58] Y. Chen, Spatial Autocorrelation Approaches to Testing Residuals from Least Squares Regression, Peking: Department of Geography, College of Urban and Environmental Sciences, 2015.

[59] V. L. a. F. B. Roberto Henriques, "Applications of Self Ogranized Maps," in *Spatial Clustering Using Hierarchical SOM*, Lisbon, Science, Technology and Medicine, 2012, p. Chapter 12.

[60] F. Javier, "Master thesis: Modelos Bayesianos para describir el comprotamiento del cancer gastrico en colombia en el periodo 2005 - 2012," Universidad distrital Francisco Jose de Caldas, Bogota, 2016.

ANNEX 1

Legend for flowchart



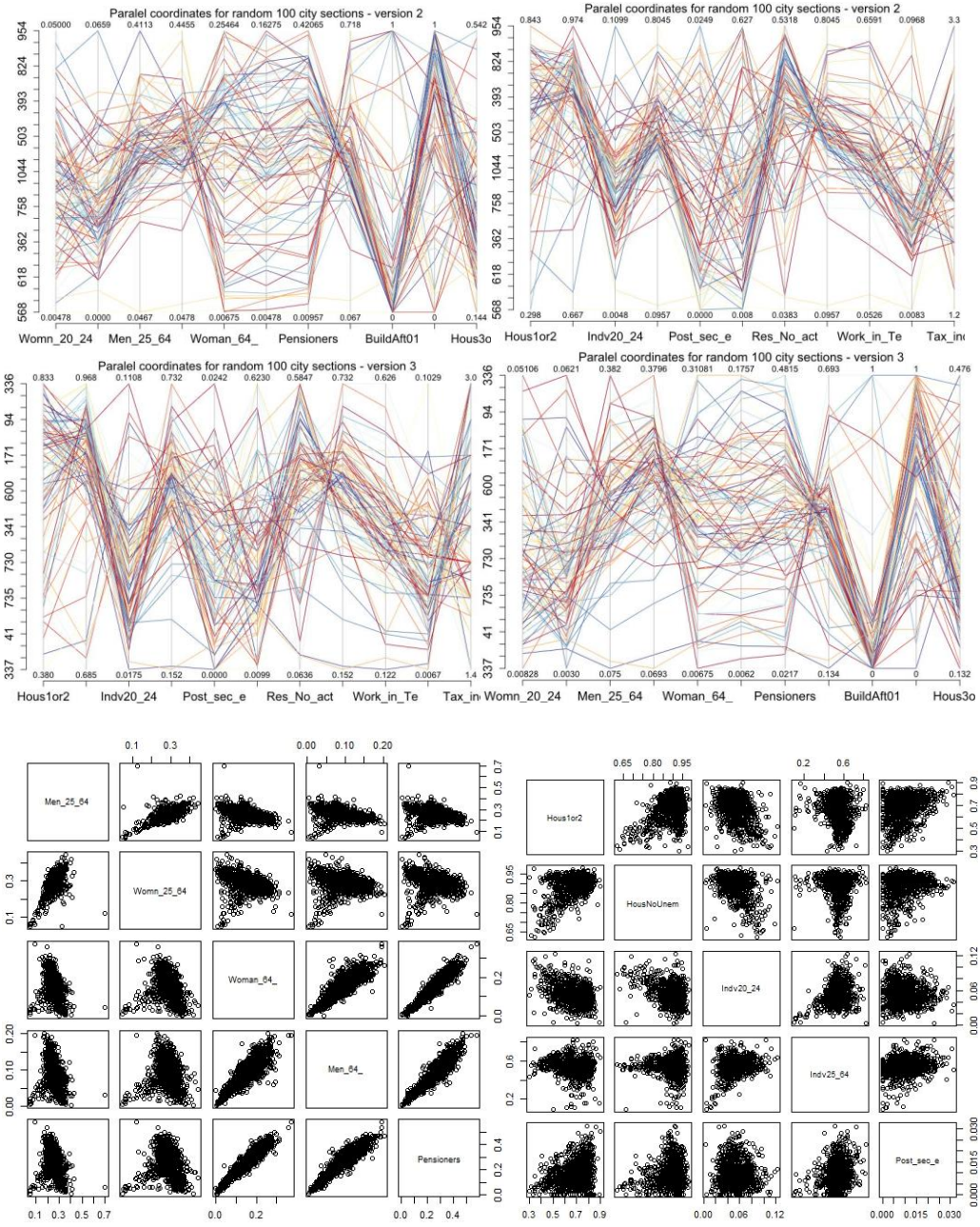
ANNEX 2

thesis_files.R

```
library (rgdal)
library (foreign)
library(maptools)
# Importing Here Maps csv tourist sites file
trst = read.csv(file = "F:/Master thesis/Data/Original_Tourist_here_maps.csv")
# calling function for cleaning duplicates
source(file = "F:/Master thesis/Data/erase_duplicate.R")
aeroport = duplicat(4581)
history = duplicat(5999)
ferry = duplicat(4482)
casino = duplicat(7985)
hotel = duplicat(7011)
museum = duplicat(8410)
rental = duplicat(7510)
taxi = duplicat(9989)
attraction = duplicat(7999)
information = duplicat(7389)
winery = duplicat(2084)
# combining all tourist sites by reducing duplicates
tourist = Reduce(function(x,y) {merge(x,y, all = TRUE)}, list(aeroport, history, ferry,casino,
hotel,museum, rental,taxi,attraction,information,winery) )
xy = cbind(tourist[,4],tourist[,3] )
# converting in UTM north and east zone 29 Datum 73 coordinates system
NE = as.data.frame(project(xy, "+proj=utm +zone=29 +north +ellps=WGS84 +datum=WGS84 +units=m +
no_defs "))
names(NE) = c("Northing", "Easting")
tourist = cbind(tourist,NE)
# saving final tourist sites file
write.csv(x = tourist,file = "F:/Master thesis/Data/Final_tourist_sites.csv")
# Importing restaurant Here Maps file
restaurants = read.csv(file = "F:/Master thesis/Data/restaurants_POI_Rest_Coffe_Lx.csv")
restaurants = duplicat1("Restaurant")
restaurants = restaurants[c("facility_type_desc","latitude","longitude","Easting","Northing")]
# Setting up shapefile for the SOM clustering Lisbon10 shp
lisbon_map = readShapePoly("F:/Master thesis/Data/New_Lisboa/Final_Lisboa10.shp")
results = data.frame(lisbon_map)
# results = results[-381,]
# Adding new variables
results[13] = results$var96/results$var67
colnames(results)[13] = "Womn_20_24"
results[23] = results$var87/results$var67
colnames(results)[23] = "Men_20_24"
results[20] = results$var89/results$var67
```

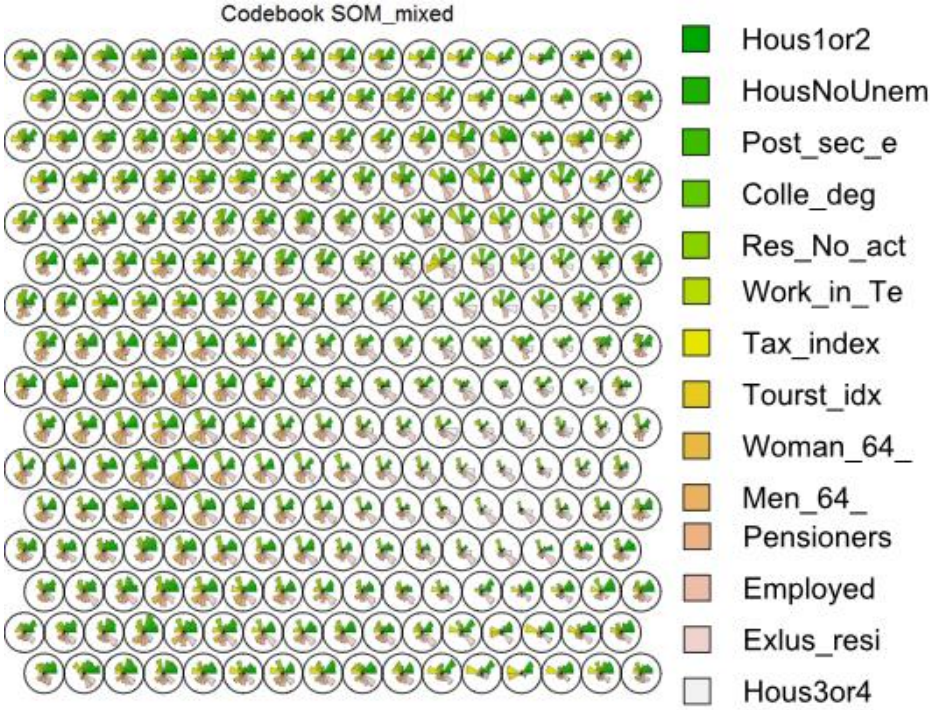
```
colnames(results)[20] = "Men_25_64"
results[21] = results$var98/results$var67
colnames(results)[21] = "Womn_25_64"
results[24] = results$var99/results$var67
colnames(results)[24] = "Womn_64_"
results[25] = results$var90/results$var67
colnames(results)[25] = "Men_64_"
results[22] = results$var116/results$var67
colnames(results)[22] = "Pensioners"
results[27] = results$var115/results$var67
colnames(results)[27] = "Empoyed"
results[26] = (results$var22_1+results$var23_1)/(results$var22_1+results$var23_1+results$var15
+results$var16+results$var17+results$var18+results$var19+results$var20+results$var21)
colnames(results)[26] = "BuildAft01"
results[,c(16,17,18,19,28,29,30,31,32,33,34,35,36)] = list(NULL)
# results[14:33] = list(NULL)
results = round(results,digits = 5)
# Output in csv
write.csv2(results,"F:/Master thesis/Data/Lisbon_Data_Final.csv")
```

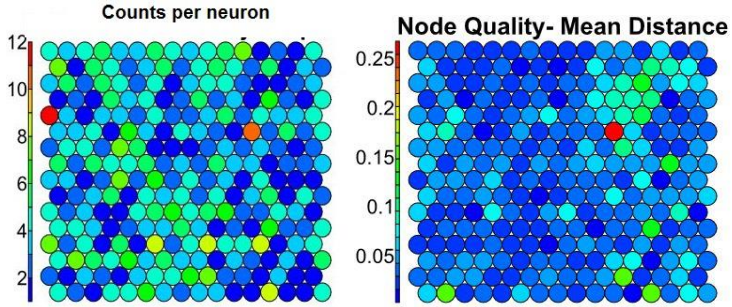
ANNEX 3

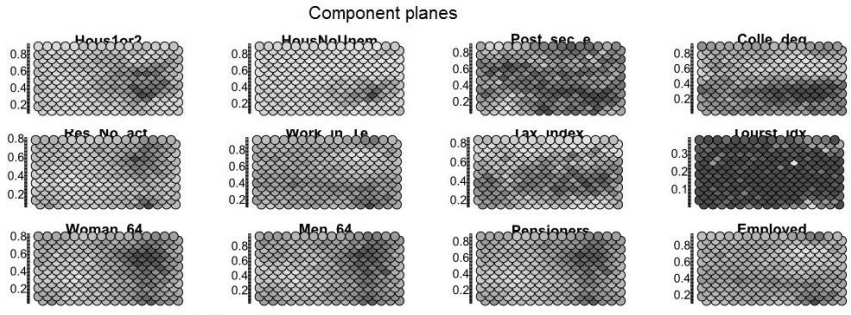


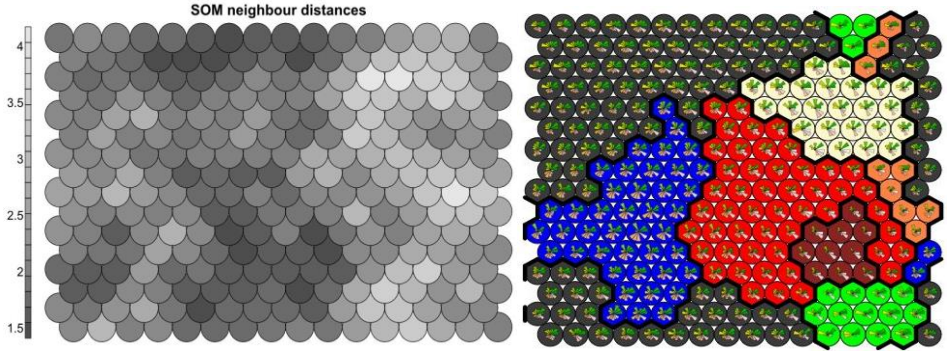
ANNEX 4

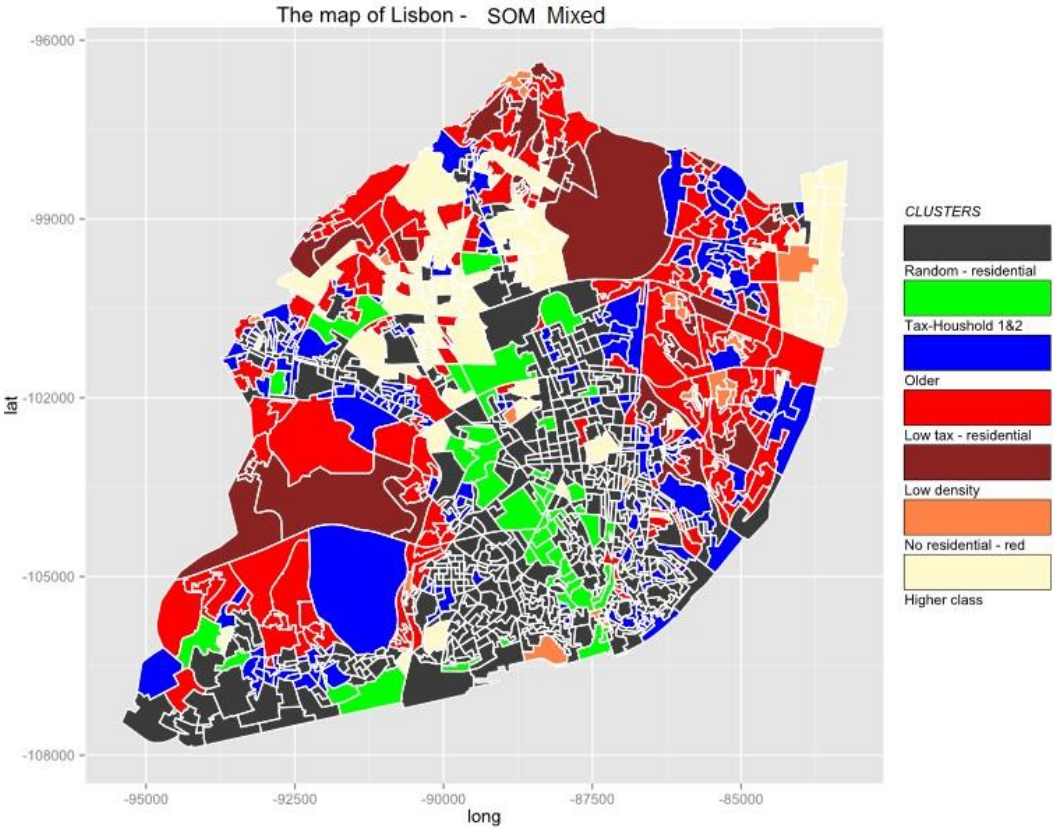
SOM_mixed





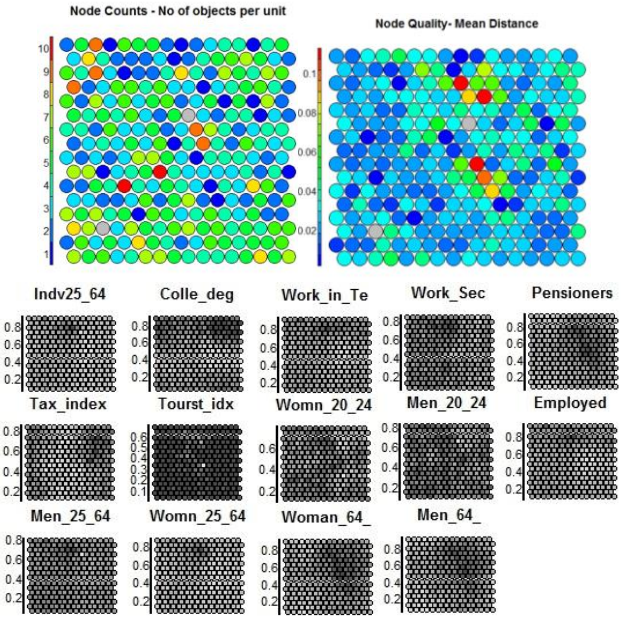




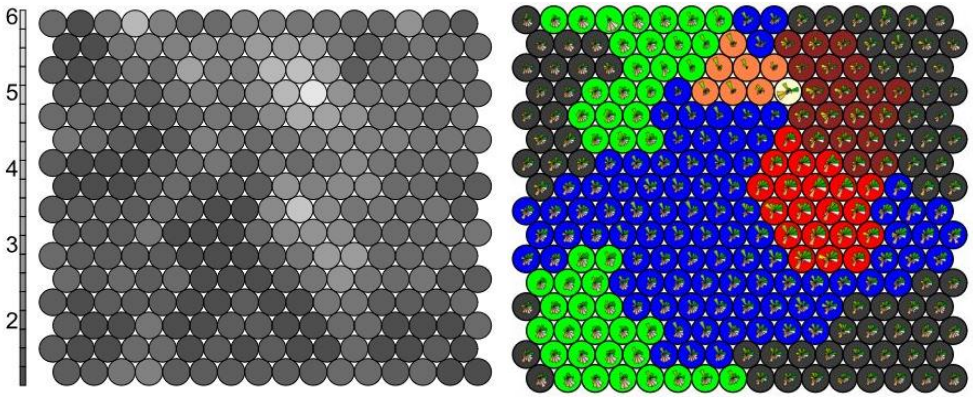


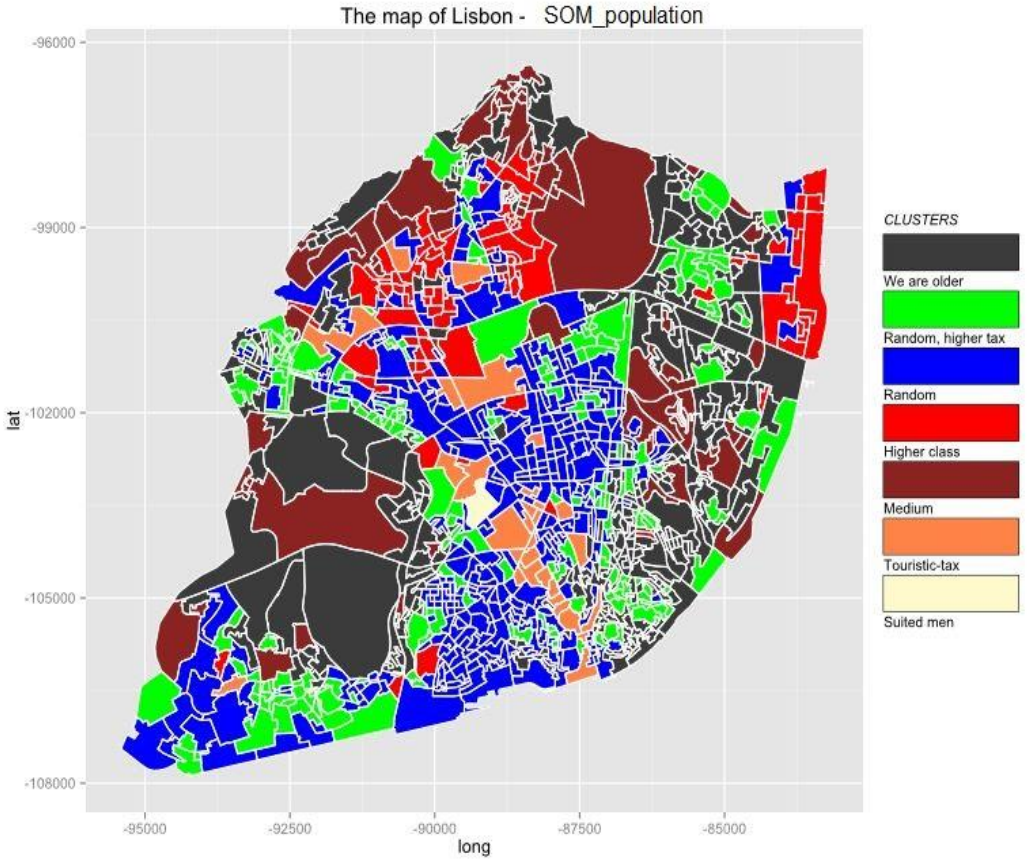
Var	Green	Blue	Red	Brown	Pink	White	Dark green
Hous1or2	***	***	*	.	close to red cluster	**	**.
NoUnempl	***	***	***	*		***	**.
Post-sec	.	.	*	.		**	*
College	.	**	*	*		***	.*
Res_no_act	.	***	*	***		**	**
Tertialy	.	*	**	.		***	
Tax	***	**	*	*		***	**.
Tourtist	./***	.	.	.		./***	**
Womn_64	*	***	*.	*		.	**
Men_64	*	***	*.	*		.	**
Pensioners	*	***	*.	*		.	**
Employed	*	**	**	*	.	***	**
Exlus_res	**	***	***	***	.	***	***
Hous3or4	.	.	**	**		**	*
	.- insignificant		* - low			** - moderate	*** - high

SOM_population



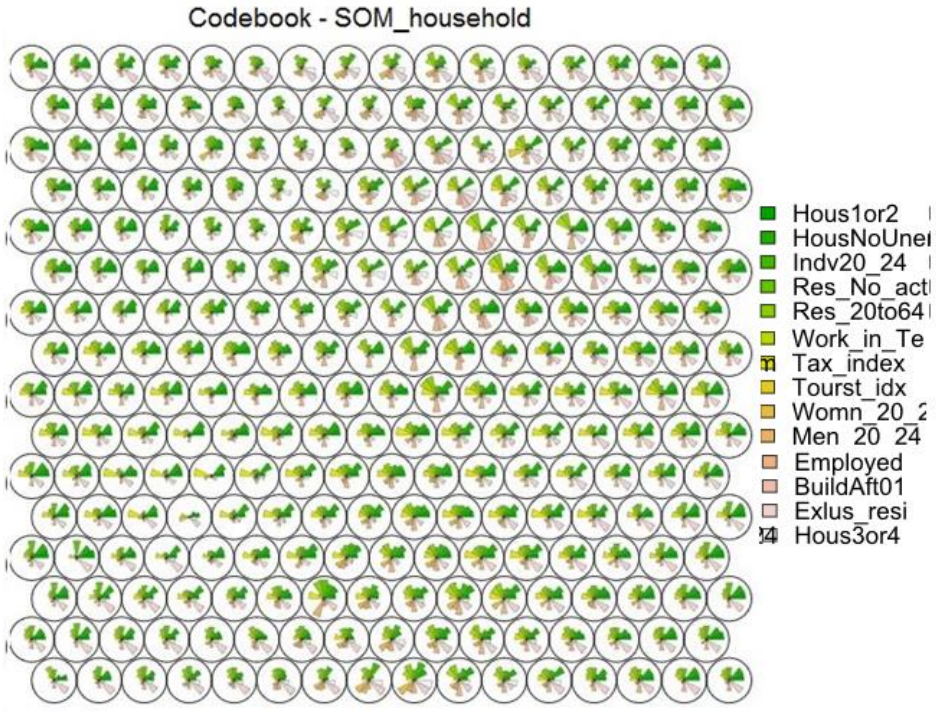
SOM neighbour distances

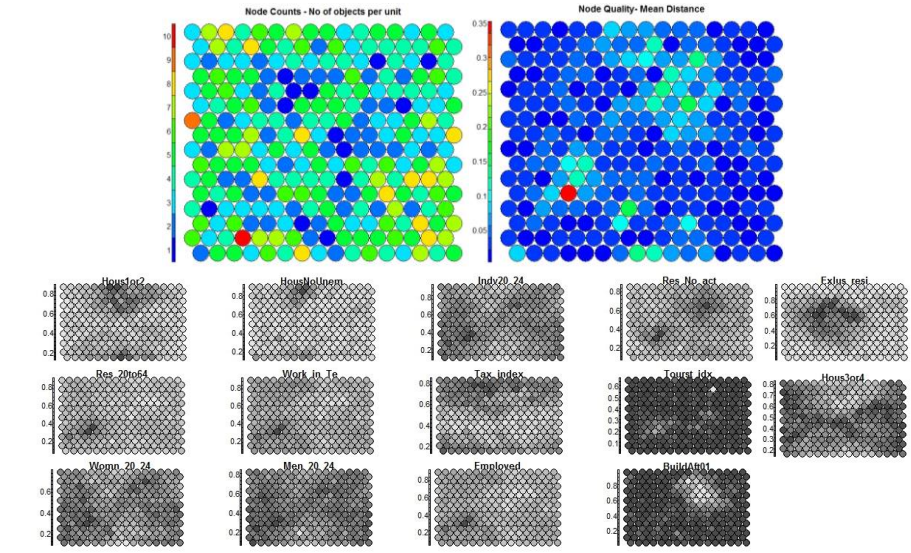


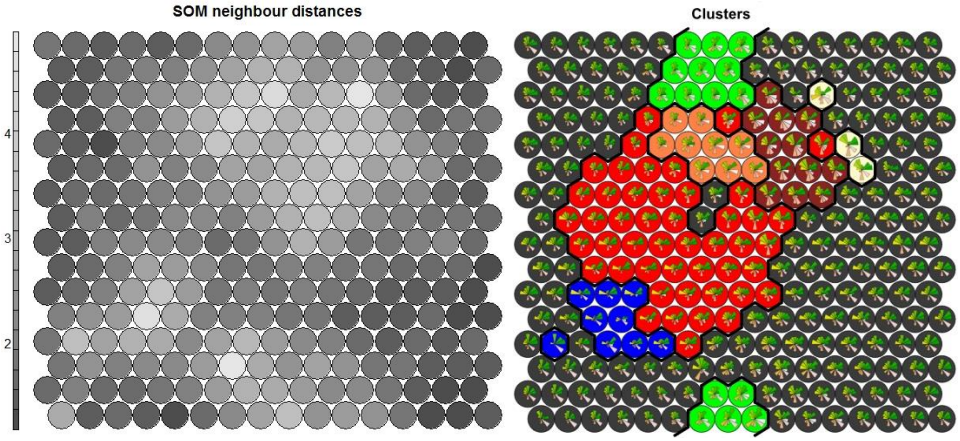


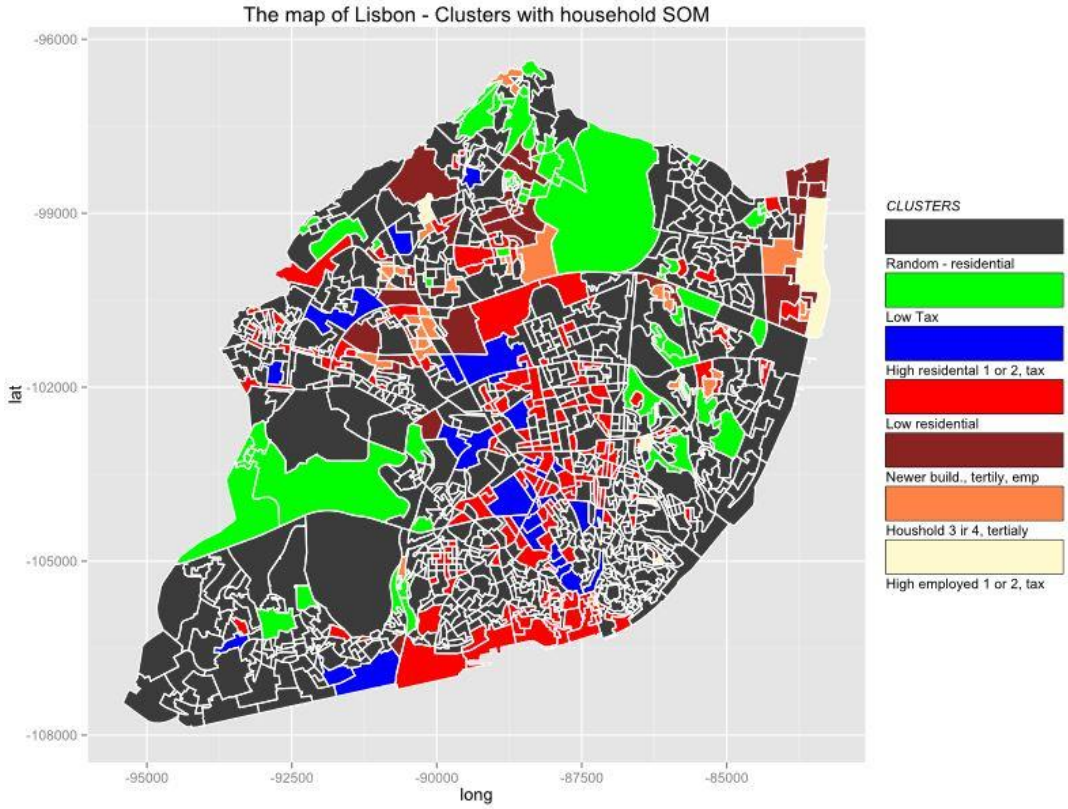
Cluster / Var	Green	Blue	Red	Brown	Pink	White	Dark green
Ind25-64	**	**	***	***		***	**
College	**	**	***	.			.
Work_in_ter	*		***	*			*
Work-sec	.		*	.			.
Tax_index	mix	***	**	.	***	***	.
Tourst_idx	.		*	.	*		*
Womn_24	.		*	**			**
Men20_24	.		*	*		***	.
Men_25_64	**		*	*		***	.
Womn25_64	*		***	***			**
Womn_64	***		.	.			**
Men_64	***		.	.			**
Pensioners	***	*	.	.			**
Employed	*	**	***	*			*
	.- insignificant		* - low		** - moderate		*** - high

SOM_household









Cluster / *Var*	**Green**	**Blue**	**Red**	**Brown**	**Pink**	**White**	**Dark green**
Hous1or2	.	***	***		*	***	***
HousNoUnem		***	***	***	***	***	**.
Indv20-24	**.	.	**	.*	*	*	
Res_no_act	**	.	*.	.	*	*	
Res20-64	***	*	***	***	***	***	
Tertial	*	*	**.	***	***	***	**
Tax_index	*	**.	***	**	**	**	**
Tourst_indx	.	.	.*	*	.	*	
W20-24	**	.	*	*	**	*	
M20-24	*	.	.	*	.	*	
Employ	*	.	**.	***	**.	***	
Build_af_01	.	.	.	***	*	.	
Exlusive_resid	***	***	.	***	.	***	***
Hous-3&4	**	.		Mix	***	.	*
	.- insignificant or 0			* - low	** - moderate		*** - high

ANNEX 5

som_entry.R[8]

```
X <- (“sp”,”class”,”kohonen”,”dummies”,”ggplot2”,”maptools”,”reshape2”,”rgeos”,”rgd
al”,”ggmap”,”leaflet”,”RColorBrewer”,”corrplot”,“ggdendro“, “vegan”)
lapply(X, require, character.only = TRUE)
source("C:/GEOTECH/NOVA/GPS/coolBlueHotRed.R")
# source(file = "F:/Master thesis/Data/thesis_files.R")
lisbon_map = readShapePoly("F:/Master thesis/Data/New_Lisboa/Final_Lisboa12.shp")
results = data.frame(lisbon_map)
# lm = lisbon_map
spearman = cor(results[-1], method = "spearman")
pearson = cor(results[-1], method = "pearson")
corrplot(spearman, type = "upper", order="hclust", tl.col="black", tl.srt=60,title
= "Spearman's correlation",mar = c(0,0,1,0), cl.cex = 1.5, tl.cex = 1 )
corrplot(pearson, type = "upper", order="hclust", tl.col="black", tl.srt=60,main =
"Pearson's correlation",mar = c(0,0,1,0),cl.cex = 1.5, tl.cex = 1)
color_palette = colors()[c(26,36,254,552,176,394,261)]
som_entry = results[c(-1:-3,-5,-6, -9,-11,-15,-16, -21,-22) ]
# data_som_matrix = scale(data_som,center = FALSE, scale = apply(data_som,2,sd, na.
rm=TRUE))
x = som_entry
data_som_matrix = apply(som_entry,2,function(x)(x-min(x))/(max(x)-min(x)))
names(data_som_matrix) <- names(som_entry)
# Execute SOM
set.seed(7)
som_model = som(data_som_matrix,
                grid=somgrid( xdim = 16, ydim = 16, topo = "hexagonal" ),
                rlen=100,
                alpha=c(0.05,0.01),toroidal = TRUE,
                n.hood = "circular",
                keep.data = TRUE,
                init = data_som_matrix[seq(4,1024,4),])
par(mar = c(2, 2, 2, 2))
plot(som_model, main = "Platform")
# list of two matrices, containing codebook vectors for X and Y, respectively(shows
the codebook vectors).
plot(som_model, type = "codes")
# shows the mean distance to the closest codebook vector during training.
plot(som_model, type = "changes")
#counts nodes
plot(som_model, type = "counts", main="Node Counts - No of objects per unit", palet
te.name=coolBlueHotRed)
plot(som_model, type="dist.neighbours", main = "SOM neighbour distances", palette.n
ame=grey.colors)
#shows the mean distance of objects mapped to a unit to the codebook vector of that
unit.# The smaller the distances, the better the objects are represented by the cod
ebook vectors.
plot(som_model, type = "quality", main="Node Quality- Mean Distance", palette.name=
coolBlueHotRed)
par(mfrow=c(4,4))
for (i in 1:14)
  plot(som_model, type = "property", property = som_model$codes[,i], main=names(som
_model$data)[i], palette.name=grey.colors)
par(mfrow=c(1,1))
#Clustering SOM - som_model$codes - values for each cell - as a result from SOM alg
orithm # wss significates within cluster sum of squares, hence first for each varia
ble (column) we are calculating variance,
mydata_clust <- som_model$codes
# within cluster sum of squares in case that is only one centroid. wcss is square s
ummary of distances from # entity point to cluster centroid
wcss = (nrow(mydata_clust)-1)*sum(apply(mydata_clust,2,var))
```

[8]R script partially derived and modified from: http://www.shanelynn.ie

```
for (i in 2:8) wcss[i] = kmeans(mydata_clust, centers=i)$tot.withinss
# adjusting plot location
par(mar=c(5.1,4.1,4.1,2.1))
par(mfcol=c(1,1), mar=c(3,4,1,0.5), oma=c(0,0,1,0))
# WCSS and number of clusters
plot(wcss, xlab="Number of Clusters", ylab="Within groups sum of squares", main="Wi
thin cluster sum of squares (WCSS)")# Form clusters on grid,use hierarchical cluste
ring to cluster the codebook vectors
som_cluster <- cutree(hclust(dist(som_model$codes)), 7)
dendo = hclust(dist(som_model$codes))
dhc <- as.dendrogram(dendo)
# Rectangular lines
ddata <- dendro_data(dhc, type = "triangle")
p1 <- ggplot(segment(ddata)) + ggtitle("Dendogram branches - SOM codebook values")+
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend)) +
  coord_flip() +
  scale_y_reverse(expand = c(0.2, 0))
ddata <- dendro_data(dhc, type = "rectangle")
p2 <- ggplot(segment(ddata)) + ggtitle("Dendogram branches - SOM codebook values")+
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend)) +
  coord_flip() +
  scale_y_reverse(expand = c(0.2, 0))
# Show the map with different colours for every cluster
color_palette = colors()[c(176,254,26,552,36,585,394,261)]
plot(som_model, type="mapping", bgcol = color_palette[som_cluster],main="Clusters")
add.cluster.boundaries(som_model, som_cluster)
#show the same plot with the codes instead of just colours
plot(som_model, type= "codes", bgcol = color_palette[som_cluster],main ="Clusters")
add.cluster.boundaries(som_model, som_cluster)
# Plot the map of Lisbon, coloured by the clusters the map to show locations
cluster_details = data.frame(id=results$seccao, cluster = som_cluster[som_model$uni
t.classif])
# Adding clusters to each city section
lisbon_map@data[26] = seq(1:1053)
cluster_details[3] = seq(1:1053)
rr2 = merge(lisbon_map@data,cluster_details, by.x = "V26", by.y = "V3")
lisbon_map@data = subset(rr2, select = c(-1,-27))
cluster_details[3]= NULL
# saving shapefile
writePolyShape(lisbon_map,"F:/Master thesis/Data/New_Lisboa/Lisboa_clusters_entry")
# This part is related to GGPLOT, we have to make data.frame from spatial data fram
e.It is possible to fortify by "cluster" and then in fill=factor(id) we would have
results of cluster, but no borders(no need to merge)
lisbon = fortify(lisbon_map, region="seccao")
# merging clusters with lisbon data.frame by id's
lisbon = merge(lisbon, cluster_details, by="id")
g = ggplot(lisbon) + aes(long, lat, group=group, fill=factor(cluster)) + geom_polyg
on() + geom_path( color = "white") + coord_equal() + scale_fill_manual(values = col
or_palette, breaks=c(1,2,3,4,5,6,7,8), labels=c("Random - residential","Tax-Houshol
d 1&2","Older", "Low tax - residential", "Low density","No residential - red", "Hig
her class", "NULL")) + ggtitle("The map of Lisbon - Clusters with entry variables")
+ theme(legend.key = element_rect(colour = "black"), legend.title = element_text(fa
ce = "italic")) + guides(fill = guide_legend(title = "CLUSTERS", title.position = "
top", label.position = "bottom"))
plot(g)
# spatial autocorrelation test for spatial clustering
l_m = readShapePoly("F:/Master thesis/Data/New_Lisboa/Lisbon_with_centroids.shp")
l_m = data.frame(l_m)
x = apply(som_model$codes,1,sum)
y = x[som_model$unit.classif]
l_m$value = y - som_model$distances
# mantel test for spatial autocorrelation
# SOM_Entry test gives 0.012 p-value
w=(dist(l_m[,c(l_m$x,l_m$y)]))
u=(dist(l_m$value))
mantel(w,u)
```

ANNEX 6

I_Final_GLM_GLMsp_GAMsp.R[9]

```
x <- c('sp','class','spdep','spacemakeR','ggplot2','maptools','reshape2','rgeos','rgdal','tripack','
mgcv','corrplot','ggdendro', 'MASS', 'Rcmdr', 'pscl', 'cluster', 'eeptools', 'stats', 'GISTools', 'n
ortest', 'hydroGOF')
lapply(x, require, character.only = TRUE)
l_m = readShapePoly("F:/Master thesis/Data/New_Lisboa/Lisbon_with_centroids.shp")
l_m_80 = readShapePoly("F:/Master thesis/Data/New_Lisboa/Buffers/Validation_Testing/80%.shp")
l_m_20 = readShapePoly("F:/Master thesis/Data/New_Lisboa/Buffers/Validation_Testing/20%.shp")
coor_rest = read.csv2(file = "restaurants_Here_Maps_Redudant.csv",sep = ",")
xy_rst = coor_rest[6:7]
dataset = data.frame(l_m)
dataset80=data.frame(l_m_80)
dataset20=data.frame(l_m_20)
xy=cbind(l_m_80$x,l_m_80$y)
plot(l_m_80,border="gray")
points(xy,pch="+",col="blue")
# for testing 20%
xy20=cbind(l_m_20$x,l_m_20$y)
# Physical contiguity criteria for irregular l_m_80
plot(l_m)
op=par(mfrow=c(1,2))
rooknb1=poly2nb(l_m_80,queen=FALSE)
plot(l_m_80,border="gray")
plot(rooknb1,xy,add=T,col="blue")
title(main="Rook")
queennb1=poly2nb(l_m_80,queen=TRUE)
plot(l_m_80,border="gray")
plot(queennb1,xy,add=T,col="blue")
title(main="Queen")
par(op)
# Criterions based on graphics
op=par(mfrow=c(2,2))
trinb=tri2nb(xy)
plot(l_m_80,border="gray")
plot(trinb,xy,add=T,col="blue")
title(main="Triangulation Delaunay")
soinb=graph2nb(soi.graph(trinb,xy))
plot(l_m_80,border="gray")
plot(soinb,xy,add=T,col="blue")
title(main="Sphere of influence")
gabrielnb=graph2nb(gabrielneigh(xy),sym=TRUE)
plot(l_m_80,border="gray")
plot(gabrielnb,xy,add=T,col="blue")
title(main="Gabriel Graphics")
relativenb=graph2nb(relativeneigh(xy),sym=TRUE)
plot(l_m_80,border="gray")
plot(relativenb,xy,add=T,col="blue")
title(main="Relative neighboors")
par(op)
# Criterions based on distance
op=par(mfrow=c(2,2))
knn1_nb=knn2nb(knearneigh(xy, k = 1))
plot(l_m_80,border="gray")
plot(knn1_nb,xy,add=T,col="blue")
title(main="Nearest neighbor")
knn2_nb=knn2nb(knearneigh(xy, k = 2))
# for the 20% testing
knn2_nb20=knn2nb(knearneigh(xy20, k = 2))
plot(l_m_80,border="gray")
plot(knn2_nb,xy,add=T,col="blue")
title(main="2 Nearest neighbors")
knn3_nb=knn2nb(knearneigh(xy, k = 3))
plot(l_m_80,border="gray")
plot(knn3_nb,xy,add=T,col="blue")
title(main="3 Nearest neighbors")
knn4_nb=knn2nb(knearneigh(xy, k = 4))
plot(l_m_80,border="gray")
plot(knn4_nb,xy,add=T,col="blue")
title(main="4 Nearest neighbors")
### Select matrix based on neighborhood by (PrincipaCoordinatesNeighMatrix)
summary(test.W(l_m_80$Restaurant,rooknb1))
```

[9] R code partially derived, but modified from [60]

```
#testing which criterion is better by AIC- select best of them with lower AIC
summary(test.W(l_m_80$Restaurant,queennb1))#physical criterion with queen
summary(test.W(l_m_80$Restaurant,trinb))#delaneuy criterion
summary(test.W(l_m_80$Restaurant,gabrielnb))#gabriel graphycs
summary(test.W(l_m_80$Restaurant,relativenb))#Relative neighbors
summary(test.W(l_m_80$Restaurant,knn1_nb))
summary(test.W(l_m_80$Restaurant,knn2_nb))
summary(test.W(l_m_80$Restaurant,knn3_nb))
summary(test.W(l_m_80$Restaurant,knn4_nb))
###Spatial weights matrices###
knn2W1=nb2listw(knn2_nb,style="W")
summary(knn2W1)
# for the 20% testing
knn2W1_20=nb2listw(knn2_nb20,style="W")
###Global indicators - spatial autocorrelation ###
par(mfrow=c(1,1))
moran.plot(as.vector(scale(l_m_80$Restaurant)),knn2W1,xlab="Restaurants",ylab="No of rest of neighbo
r's cell ",main="KNN 2")
moran.test(l_m_80$Restaurant,knn2W1,alternative="two.sided")
geary.test(l_m_80$Restaurant,knn2W1,alternative="two.sided")
###Local indicators - spatial autocorrelation ###
locm3=localmoran(l_m_80$Restaurant,knn2W1,alternative="two.sided")#local spatial autocorrelati
###MAPS###
l_m_80$sz <- scale(l_m_80$Restaurant)#sz = scale for no of restarurants - 0 is new mean and new 1 is
sd
l_m_80$lag_sz <- lag.listw(knn2W1,l_m_80$sz)#lag_sz is the same for neighbors of the number of resta
urants
l_m_80$quad_sig <- NA
l_m_80@data[(l_m_80$sz >= 0 & l_m_80$lag_sz >= 0) & (locm3[, 5] <= 0.05), "quad_sig"] <- 1
l_m_80@data[(l_m_80$sz <= 0 & l_m_80$lag_sz <= 0) & (locm3[, 5] <= 0.05), "quad_sig"] <- 2
l_m_80@data[(l_m_80$sz >= 0 & l_m_80$lag_sz <= 0) & (locm3[, 5] <= 0.05), "quad_sig"] <- 3
l_m_80@data[(l_m_80$sz >= 0 & l_m_80$lag_sz <= 0) & (locm3[, 5] <= 0.05), "quad_sig"] <- 4
l_m_80@data[(l_m_80$sz <= 0 & l_m_80$lag_sz >= 0) & (locm3[, 5] <= 0.05), "quad_sig"] <- 5
breaks=seq(1, 5, 1)
labels=c("High-High", "Low-Low", "High-Low", "Low-High", "Not Signif.")
np <- findInterval(l_m_80$quad_sig, breaks)
colors <- c("red", "blue", "lightpink", "skyblue2", "white")
plot(l_m_80, col = colors[np])#colors[np] manually set the color for each region
mtext("Local neighborhoods of Moran test", cex=1.5, side = 3, line = 1)
legend("bottomright", legend = labels, fill = colors, bty = "n",cex=0.7, inset = c(0.05,0))
###Development of (spatial)GLM and (spatial)GAM models###
###Before the models have been created, the dataset80 is split it into 80% of training and 20% of va
lidating
#20% at the end is used for testing and the best model is chosen to predict the rest of the values#F
ull GLM model
GLM.1 <- glm(Restaurant ~ BuildAft01 + Colle_deg + Employed + Exlus_resi + Hous1or2 + Hous3or4 + Hou
sNoUnem + Men_20_24 + Men_25_64 + Men_64_ + Pensioners + Res_No_act + Tax_index + Tourst_idx + Woman
_64_ + Womn_20_24 + Womn_25_64 + Work_in_Te + Work_Sec, family=poisson(log), data=dataset80)
#Stepwise model selection based on backward/forward selection
#AIC - finds the model that gives the best prediction - for model comparison
#BIC - assumes that one of the models is the true model and find the "true" model - variable selecti
on
#Creation of the base GLM model
GLM.2=stepwise(GLM.1, direction='forward/backward', criterion='BIC')
summary(GLM.2, cor=FALSE)
cor(dataset80$Restaurant,predict(GLM.2, type = "response"))
#evaluation models
BIC(GLM.2, GLM.1) # GLM.2 is better based on BIC - GLM.2 = 2628.389, GLM.1 = 2668.967
anova(GLM.2,GLM.1, test = "Chisq") #anova proves it that reduced model is good
#Developing spatial component with Moran eigenvector and spatial weights from KNN2
GLM.3 <- ME(Restaurant ~ Hous1or2+HousNoUnem+Indv20_24+Indv25_64+Post_sec_e+Colle_deg+Res_No_act+Res
_20to64+Work_in_Te+Work_Sec+Tax_index+Tourst_idx+Womn_20_24+Men_20_24+Men_25_64+Womn_25_64+Woman_64_
+Men_64_+Pensioners+Employed+BuildAft01+Hous3or4, data=dataset80, family="poisson", listw=knn2W1, al
pha=0.5)
# for 20% testing
GLM.3_20 <- ME(Restaurant ~ Hous1or2+HousNoUnem+Indv20_24+Indv25_64+Post_sec_e+Colle_deg+Res_No_act+
Res_20to64+Work_in_Te+Work_Sec+Tax_index+Tourst_idx+Womn_20_24+Men_20_24+Men_25_64+Womn_25_64+Woman_
64_+Men_64_+Pensioners+Employed+BuildAft01+Hous3or4, data=dataset20, family="poisson", listw=knn2W1_
20, alpha=0.5)
GLM.4 <- glm(Restaurant ~ Hous1or2+HousNoUnem+Indv20_24+Indv25_64+Post_sec_e+Colle_deg+Res_No_act+Re
s_20to64+Work_in_Te+Work_Sec+Tax_index+Tourst_idx+Womn_20_24+Men_20_24+Men_25_64+Womn_25_64+Woman_64
_+Men_64_+Pensioners+Employed+BuildAft01+Hous3or4+fitted(GLM.3), data=dataset80,family=poisson(log))
summary(GLM.4
cor(dataset80$Restaurant,predict(GLM.4, type = "response"))## [1] 0.8191776
#Involving spatial component into base GLM model
GLM.6 <- glm(Restaurant ~ Tourst_idx+Tax_index+Hous3or4+Exlus_resi+Men_25_64+Woman_64_+HousNoUnem+ f
itted(GLM.3),data = dataset80,family = poisson(log))
BIC(GLM.2,GLM.6)# base model with spatial component is better than base model GLM.6 = 2453.687anova(
GLM.6, GLM.2, test = "Chisq")
```

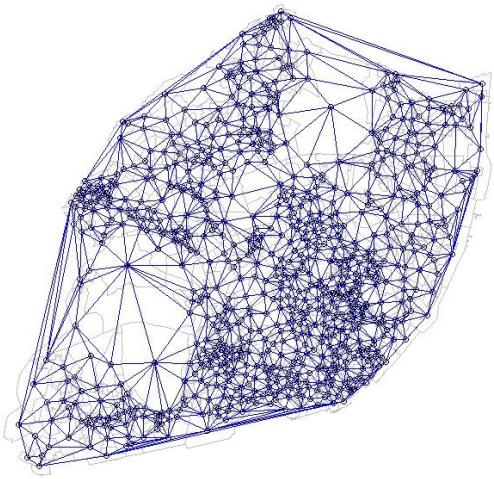
```
anova(GLM.6)
cor(dataset80$Restaurant,predict(GLM.6, type = "response"))
#Creation of base GAM model first step
lGAMp <- gam(Restaurant~s(Hous1or2, bs = "cr"),data=dataset80,family="poisson")
lGAMp1<- gam(Restaurant~s(Tax_index, bs = "cr"),data=dataset80,family="poisson")
lGAMp2<- gam(Restaurant~s(I(Tourst_idx^0.5)),data=dataset80,family="poisson")
lGAMp3<- gam(Restaurant~s(Work_in_Te, bs = "cr"),data=dataset80,family="poisson")
lGAMp4<- gam(Restaurant~s(Colle_deg),data=dataset80,family="poisson")
lGAMp5<- gam(Restaurant~s((Employed), bs = "cr"),data=dataset80,family="poisson")
lGAMp6<- gam(Restaurant~s(I(Womn_20_24^0.5)),data=dataset80,family="poisson")
lGAMp7<- gam(Restaurant~s(Res_No_act),data=dataset80,family="poisson")
lGAMp8<- gam(Restaurant~s((Woman_64_)),data=dataset80,family="poisson")
lGAMp9<- gam(Restaurant~s((Indv25_64), bs = "cr"),data=dataset80,family="poisson")
BICall = BIC(lGAMp,lGAMp1,lGAMp2,lGAMp3,lGAMp4,lGAMp5,lGAMp6,lGAMp7,lGAMp8,lGAMp9)
which.min(BICall[,2])
#Creation of base GAM model - 2nd step
lGAMp10<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Hous1or2, bs = "cr"),data=dataset80,family="poisson
", method = "REML", select = T)
lGAMp11<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr"),data=dataset80,family="poisso
n", method = "REML", select = T)
lGAMp12<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Work_in_Te, bs = "cr"),data=dataset80,family="poiss
on", method = "REML", select = T)
lGAMp13<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Colle_deg),data=dataset80,family="poisson", method
= "REML", select = T)
lGAMp14<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Employed, bs = "cr"),data=dataset80,family="poisson
", method = "REML", select = T)
lGAMp15<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(I(Womn_20_24^0.5)),data=dataset80,family="poisson",
method = "REML", select = T)
lGAMp16<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Res_No_act),data=dataset80,family="poisson", method
= "REML", select = T)
lGAMp17<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Indv25_64, bs = "cr"),data=dataset80,family="poisso
n", method = "REML", select = T)
lGAMp18<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Woman_64_),data=dataset80,family="poisson", method
= "REML", select = T)
BICall1 = BIC(lGAMp10,lGAMp11,lGAMp12,lGAMp13,lGAMp14,lGAMp15,lGAMp16,lGAMp17,lGAMp18)
which.min(BICall1[,2])
#Creation of base GAM model - 3rd step
lGAMp19<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr"),data=d
ataset80,family="poisson", method = "REML", select = T)
lGAMp21<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Colle_deg),data=dataset80,f
amily="poisson", method = "REML", select = T)
lGAMp22<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Employed, bs = "cr"),data=d
ataset80,family="poisson", method = "REML", select = T)
lGAMp23<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(I(Womn_20_24^0.5)),data=dat
aset80,family="poisson", method = "REML", select = T)
lGAMp24<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Res_No_act),data=dataset80,
family="poisson", method = "REML", select = T)
lGAMp25<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Indv25_64, bs = "cr"),data=
dataset80,family="poisson", method = "REML", select = T)
lGAMp26<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Woman_64_, bs = "cr"),data=
dataset80,family="poisson", method = "REML", select = T)
BICall2 = BIC(lGAMp19,lGAMp21,lGAMp22,lGAMp23,lGAMp24,lGAMp25,lGAMp26)
which.min(BICall2[,2])
#Creation of base GAM model - 4th step
lGAMp27<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+ s(Col
le_deg),data=dataset80,family="poisson", method = "REML", select = T)
lGAMp28<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Empl
oyed, bs = "cr"),data=dataset80,family="poisson", method = "REML", select = T)
lGAMp29<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(I(Wo
mn_20_24^0.5)),data=dataset80,family="poisson", method = "REML", select = T)
lGAMp30<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Res_
No_act),data=dataset80,family="poisson", method = "REML", select = T)
lGAMp31<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Indv
25_64, bs = "cr"),data=dataset80,family="poisson", method = "REML", select = T)
lGAMp32<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Woma
n_64_, bs = "cr"),data=dataset80,family="poisson", method = "REML", select = T)
BICall3 = BIC(lGAMp27,lGAMp28,lGAMp29,lGAMp30,lGAMp31,lGAMp32)
which.min(BICall3[,2])
lGAMp33<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Empl
oyed, bs = "cr")+ s(Colle_deg),data=dataset80,family="poisson", method = "REML", select = T)
lGAMp34<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Empl
oyed, bs = "cr")+s(I(Womn_20_24^0.5)),data=dataset80,family="poisson", method = "REML", select = T)
lGAMp35<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Empl
oyed, bs = "cr")+s(Res_No_act),data=dataset80,family="poisson", method = "REML", select = T)
lGAMp36<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Empl
oyed, bs = "cr")+s(Indv25_64, bs = "cr"),data=dataset80,family="poisson", method = "REML", select =
T)
lGAMp37<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Empl
oyed, bs = "cr")+s(Woman_64_, bs = "cr"),data=dataset80,family="poisson", method = "REML",select =T)
```

```
BICall4 = BIC(lGAMp33,lGAMp34,lGAMp35,lGAMp36,lGAMp37)
which.min(BICall4[,2])
lGAMp38<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+ s(Emp
loyed, bs = "cr") +s(I(Womn_20_24^0.5))+ s(Colle_deg),data=dataset80,family="poisson", method = "REM
L", select = T)
lGAMp39<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+ s(Emp
loyed, bs = "cr") +s(I(Womn_20_24^0.5))+ s(Res_No_act),data=dataset80,family="poisson", method = "RE
ML", select = T)
lGAMp40<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+ s(Emp
loyed, bs = "cr") +s(I(Womn_20_24^0.5))+ s(Indv25_64, bs = "cr"),data=dataset80,family="poisson", me
thod = "REML", select = T)
lGAMp41<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+ s(Emp
loyed, bs = "cr") +s(I(Womn_20_24^0.5))+ s(Woman_64_, bs = "cr"),data=dataset80,family="poisson", me
thod = "REML", select = T)
BICall5 = BIC(lGAMp38,lGAMp39,lGAMp40,lGAMp41)
which.min(BICall5[,2])
lGAMp42<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index,bs = "cr")+s(Hous1or2,bs = "cr")+ s(Emplo
yed, bs = "cr") +s(I(Womn_20_24^0.5))+s(Colle_deg)+s(Res_No_act),data=dataset80,family="poisson", me
thod = "REML", select = T)
lGAMp43<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+ s(Emp
loyed, bs = "cr") +s(I(Womn_20_24^0.5))+s(Colle_deg)+s(Indv25_64, bs = "cr"),data=dataset80,family="
poisson", method = "REML", select = T)
lGAMp44<-gam(Restaurant~s(I(Tourst_idx^0.5))+s(Tax_index,bs = "cr")+s(Hous1or2,bs ="cr")+ s(Employed
, bs = "cr") +s(I(Womn_20_24^0.5))+s(Colle_deg)+s(Woman_64_,bs="cr"),data=dataset80,family="poisson"
,method="REML",select = T)
BICall6 = BIC(lGAMp42,lGAMp43,lGAMp44)
which.min(BICall6[,2])
lGAMp45<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+ s(Emp
loyed, bs = "cr") +s(I(Womn_20_24^0.5))+s(Colle_deg)+s(Res_No_act)+s(Indv25_64, bs = "cr"),data=data
set80,family="poisson", method = "REML", select = T)
lGAMp46<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+ s(Emp
loyed, bs = "cr") +s(I(Womn_20_24^0.5))+s(Colle_deg)+s(Res_No_act)+s(Woman_64_, bs = "cr"),data=data
set80,family="poisson", method = "REML", select = T)
BICall7 = BIC(lGAMp45,lGAMp46)
which.min(BICall7[,2])
lGAMp47<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+ s(Emp
loyed, bs = "cr") +s(I(Womn_20_24^0.5))+s(Colle_deg)+s(Woman_64_, bs = "cr")+s(Res_No_act)+s(Indv25_
64, bs = "cr"),data=dataset80,family="poisson", method = "REML", select = T)
BICall8 = BIC(lGAMp,lGAMp1,lGAMp2,lGAMp3,lGAMp4,lGAMp5,lGAMp6,lGAMp7,lGAMp8,lGAMp9,lGAMp10,lGAMp11,l
GAMp12,lGAMp13,lGAMp14,lGAMp15,lGAMp16,lGAMp17,lGAMp18,lGAMp19,lGAMp21,lGAMp22,lGAMp23,lGAMp24,lGAMp
25,lGAMp26,lGAMp27,lGAMp28,lGAMp29,lGAMp30,lGAMp31,lGAMp32,lGAMp33,lGAMp34,lGAMp35,lGAMp36,lGAMp37,l
GAMp38,lGAMp39,lGAMp40,lGAMp41,lGAMp42,lGAMp43,lGAMp44,lGAMp45,lGAMp46,lGAMp47)
which.min(BICall8[,2])
#Chosen base model for GAM - lGAMp28 -
lGAMp28<- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Empl
oyed, bs = "cr"),data=dataset80,family="poisson", method = "REML", select = T)
anova(lGAMp28,lGAMp47,test = "Chisq")
cor(dataset80$Restaurant,predict(lGAMp28, type = "response"))
summary(lGAMp28)
#Base GAM model with spatial component
lGAMspat_coor80 <- gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr
")+s(Employed, bs = "cr")+ s(x,y, bs = "tp"),data=dataset80,family="poisson", method = "REML", selec
t = T)
#Comparison between GAM base and GAM with spatial component
BIC(lGAMspat_coor80, lGAMp28)
cor(dataset80$Restaurant,predict(lGAMspat_coor80, type = "response"))
summary(lGAMspat_coor80)
gam.check(lGAMspat_coor80)
plot(lGAMspat_coor80)
plot(lGAMp28)   # s(I(Womn_20_24^0.5)) is almost penalized out
anova(lGAMp28)
anova(lGAMspat_coor80,lGAMp28, test = "Chisq") # to analyse components of the model
anova(GLM.6,lGAMp28, test = "Chisq") # to analyse components of the model
anova(GLM.1,GLM.2,GLM.6,lGAMp28,lGAMspat_coor80, test = "Chisq") # to analyse components of the mode
l## codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
anova(GLM.2,GLM.6,lGAMp28,lGAMspat_coor80, test = "Chisq") # to analyse components of the mode
anova(GLM.6,lGAMp28,lGAMspat_coor80, test = "Chisq") # to analyse components of the model
anova(lGAMp28,lGAMspat_coor80, test = "Chisq") # to analyse the variance of the model
BIC(GLM.1,GLM.2,GLM.6,lGAMp28, lGAMspat_coor80)
# Inputing 20% of test data for validation - GLM and GAM with spatial component expressed the best r
esults
GLM_pred20 = predict(GLM.2,newdata = data.frame(dataset20[c(14,13,26,25,18,20,4)]), type = "response
")
GLM_pred80 = predict(GLM.2,newdata = data.frame(dataset80[c(14,13,26,25,18,20,4)]), type = "response
")
cor(GLM_pred80, dataset80$Restaurant)
cor(GLM_pred20, dataset20$Restaurant)
rmse(as.numeric(GLM_pred80), as.numeric(dataset80$Restaurant))# RMSE = 2.18
```

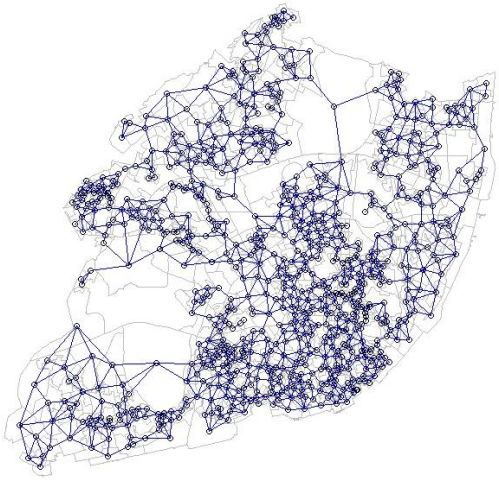
```
GAM_pred20 = predict(lGAMp28,newdata = data.frame(dataset20[c(14,13,3,23)]), type = "response”)
GAM_pred80 = predict(lGAMp28,newdata = data.frame(dataset80[c(14,13,3,23)]), type = "response")
cor(GAM_pred80, dataset80$Restaurant)#correlation of the construction
cor(GAM_pred20, dataset20$Restaurant)#correlation of the test
rmse(as.numeric(GAM_pred80), as.numeric(dataset80$Restaurant))# RMSE = 1.78
GAMsp_pred20 = predict(lGAMspat_coor80,newdata = data.frame(dataset20[c(14,13,3,23,28,29)]), type =
"response")
GAMsp_pred80 = predict(lGAMspat_coor80,newdata = data.frame(dataset80[c(14,13,3,23,28,29)]), type =
"response")
cor(GAMsp_pred80, dataset80$Restaurant)
cor(GAMsp_pred20, dataset20$Restaurant)
rmse(as.numeric(GAMsp_pred80), as.numeric(dataset80$Restaurant))# smallest RMSE 1.5757
summary(lGAMspat_coor80)
# Predicting values for the whole model
dataset$predicted_rest<-predict(lGAMspat_coor80,newdata=data.frame(dataset[c(14,13,3,23,28,29)]),typ
e="response")
GAMsp = gam(Restaurant~s(I(Tourst_idx^0.5))+ s(Tax_index, bs = "cr")+s(Hous1or2, bs = "cr")+s(Employ
ed, bs = "cr")+ s(x,y, bs = "tp"),data=dataset,family="poisson", method = "REML", select = T)
cor(dataset$predicted_rest, GAMsp$fitted.values)
dataset$deviance <-  with(dataset, residuals(GAMsp,type='deviance'))
# Cutt-offs
scatterplot(Restaurant~predicted_rest,reg.line=lm, data=dataset, spread = F,xlab = "Restaurant predi
ction - GAM with spatial component")# The Empirical RULe - left tail indicates more restaurants, lef
t less restaurants
d = dataset$deviance
Ica <- d[which(d >= mean(d)-sd(d) & d <= sd(d))]
II_intv_left <- d[which(d >= mean(d)-2*sd(d) & d < mean(d)-sd(d))]
II_intv_right <- d[which(d <= mean(d)+2*sd(d) & d > mean(d)+ sd(d))]
II_intv_righta <- d[which(d <= 2*sd(d) & d >  sd(d))]
III_intv_left <- d[which(d < mean(d)-2*sd(d))]
III_intv_right <- d[which(d > mean(d)+2*sd(d))]
III_intv_righta <- d[which(d > 2*sd(d))
a= as.data.frame((d%in%Ica)*1)
b = as.data.frame((d%in%II_intv_left)*2)
c = as.data.frame((d%in%II_intv_righta)*4)
t = as.data.frame((d%in%III_intv_left)*3)
e = as.data.frame((d%in%III_intv_righta)*5)
z = as.data.frame(c(a,b))
colnames(z) = c("I","II")
z[z==0] <- NA
z <- within(z, I <- ifelse(is.na(I), II, I))
z[2]<-c
z[z==0] <- NA
z <- within(z, I <- ifelse(is.na(I), II, I))
z[2]<-t
z[z==0] <- NA
z <- within(z, I <- ifelse(is.na(I), II, I))
z[2]<-e
z <- as.data.frame(within(z, I <- ifelse(is.na(I), II, I)))
dataset$category = as.numeric(unlist(z[1]))
# The map of estimation
par(mar = c(0,0,1.5,0))
breaks=seq(1, 5, 1)
labels=c("No change", "Overestimated", "Extremely overestimated", "Underestimated", "Extremely under
estimated" )
np <- findInterval(dataset$category, breaks)
colors <- c("white","skyblue", "blue","indianred1", "red")
plot(l_m, col = colors[np])#colors[np] manually set the color for each region
mtext("Map of Lisbon - Restaurants estimation", cex=2, side = 3, line = 0)
legend("bottomright",legend = labels,fill = colors,bty = "n",cex=0.7,inset = c(-0.27,0.07),title = "
Categories:")
SpatialPolygonsRescale(layout.scale.bar(), offset= c(-87000,-107900), scale= 2000, fill= c("transpar
ent", "black"), plot.grid= F)
text(-86000, -107200, "2KM", cex= 1)
SpatialPolygonsRescale(layout.north.arrow(1), offset= c(-95002,-99000), scale = 1300, plot.grid=F)
par(op)
h = hist(d, col = "grey", main = "Histogram of residuals")
xfit = seq(min(d), max(d), length = 30)
yfit = dnorm(xfit, mean = mean(d), sd = sd(d))
yfit = yfit*diff(h$mids[1:2])*length(d)
lines(xfit,yfit, lw = 2, col = "red")
density = density(d)
plot(density, main = "Kernel density for deviation and cut offs", xlab = "Deviation", col = "red") +
polygon(density, col = "grey") + abline(v = c(mean(d)+sd(d), mean(d)-sd(d), 2*sd(d), mean(d)-2*sd(d)
), col = "red", lty = "dashed")  + text(0.8,0.3,labels = "mean + sd = 0.89", cex = 0.8) + text(2.4,0
.4,labels = "mean + 2 x sd = 2.04", cex = 0.8)
```

ANNEX 7

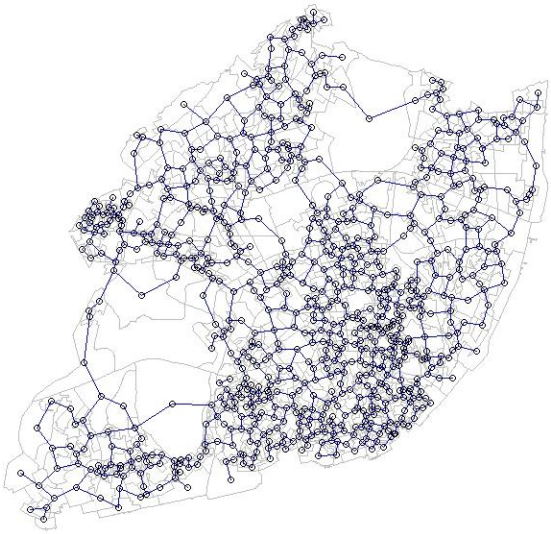
Triangulation Delaunay



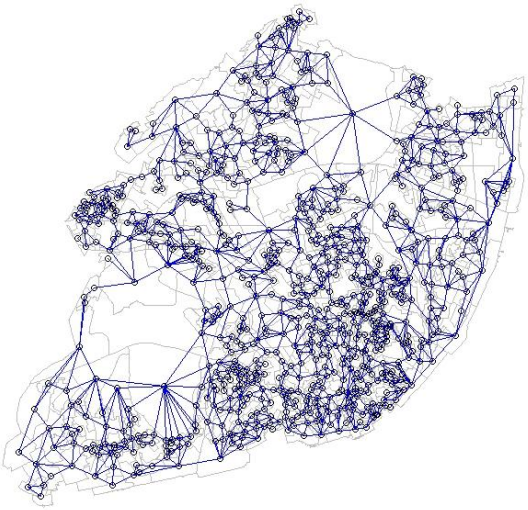
Sphere of influence



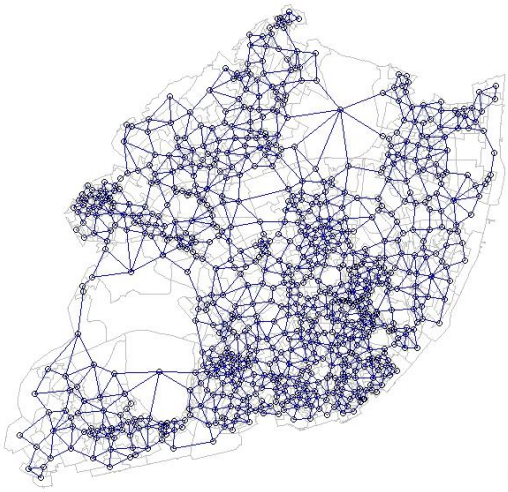
Relative neighboors



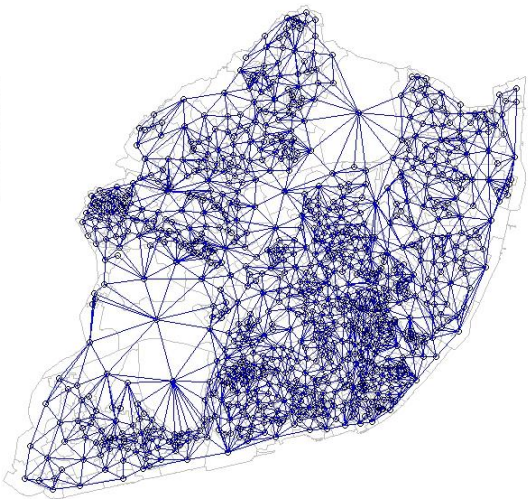
Rook



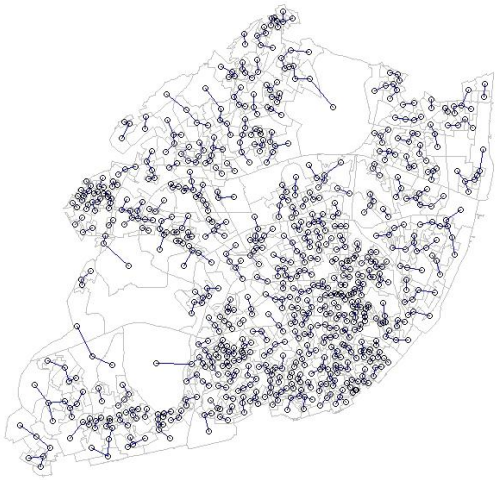
Gabriel Graphics



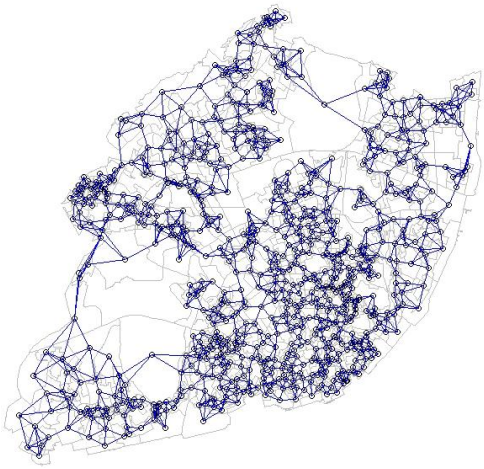
Queen



Nearest neighbor



4 Nearest neighbors



3 Nearest neighbors