Masters Program in Geospatial Technologies

# *EVALUATION OF SPATIAL DATA'S IMPACT IN MID-TERM ROOM RENT PRICE THROUGH APPLICATION OF SPATIAL ECONOMETRICS AND MACHINE LEARNING*

## *Case Study: Lisbon*

Mihail Petkov

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

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

# *EVALUATION OF SPATIAL DATA'S IMPACT IN MID-TERM ROOM RENT PRICE THROUGH APPLICATION OF SPATIAL ECONOMETRICS AND MACHINE LEARNING*

## *Case Study: Lisbon*

Dissertation supervised by:

**Professor Doutor Roberto André Pereira Henriques, PhD**
Instituto Superior de Estatística e Gestão de Informação,
Universidade NOVA de Lisboa
Lisbon, Portugal

**Joel Dinis Baptista Ferreira da Silva, PhD**
Instituto Superior de Estatística e Gestão de Informação,
Universidade NOVA de Lisboa
Lisbon, Portugal

**Professor Doutor Carlos Granell-Canut, PhD**
Institute of New Imaging Technologies,
Universitat Jaume I (UJI)
Castellón de la Plana, Spain

February 28, 2020

DECLARATION OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisbon, 28.02.2020

Mihail Petkov

# ACKNOWLEDGMENTS

An idea about making sense of economic phenomena through the lenses of geographical information have always been a research topic of interest for me. Though research and the results' aftermath are oftentimes unpredictable, the people that guided me and supported me throughout this process were fortunately not.

I want to first and foremost thank my main-supervisor Professor Dr. Roberto Henriques that was open to hearing my ideas and helping me construct them in something practical, to Professor Dr. Joel Silva who was always there to suggest improvements and guide me through new directions and doors when all I was seeing were dead ends, for Professor Dr. Carlos Granell for believing in me, and last, but not least, to Professor Dr. Marco Painho for keeping me level-headed to push and progress until I have reached the finish line.

Additionally, my greatest gratitude goes to all colleagues from the Geospatial Technologies Master as a whole, with special shout-outs to my friend and flat-mate Itza, to Damien, Carlos, Braundt, Vicente and Pablo.

Last but not least, I would like to thank all my closest family and friends in Macedonia. I am extremely grateful for all the support, patience and encouragement provided during the journey.

*EVALUATION OF SPATIAL DATA'S IMPACT IN MID-TERM ROOM RENT PRICE THROUGH APPLICATION OF SPATIAL ECONOMETRICS AND MACHINE LEARNING*

***Case Study: Lisbon***

# ABSTRACT

Household preferences is a topic whose relevance can be found to dominate the applied economics, but whereas urban economies view cities as production centers, this thesis aims to give importance to the role of consumption. Provision to PoIs might give explanation to what individuals value as an important asset for improvement of their quality of life in a chosen city. As such, understanding short-term rentals and real estate prices have induced various research to seek proof of impacting factors, but analysis of mid-term rent has faced the challenge of being an overlooked category. This thesis consists of an integrated three-steps approach to analyze spatial data's impact over the mid-term room rent, choosing Lisbon as its case study. The proposed methodology constitutes use of traditional spatial econometric models and SVR, encompassing a large set of proxies for amenities that might be recognized to hold a possible impact over rent prices. The analytical frameworks' first step is to create a suitable HPM model that captures the data well, so significant variables can be detected and analyzed as a discrete dataset. The second step applies subsets of the dataset in the creation of SVR models, in hopes of identifying the SVs influencing price variances. Finally, SOM clusters are chosen to address whether more natural order of data division exists. Results confirm the impact of proximity to various categories of amenities, but the enrichment of models with the proposed proxies of spatial data failed to corroborate attainment of model with a higher accuracy.

(Nüst et al., 2018) provides a self-assessment of the reproducibility of research, and according to the criteria given, this dissertation is evaluated as: 0, 2, 1, 2, 2 (input data, preprocessing, methods, computational environment, results).

# KEYWORDS

Predictive Modeling

Amenities

Support Vector Regression

Spatial Econometrics

Hedonic Price Modelling

Points of Interest

Mid-Term Rent

# ACRONYMS

**AIC** – Akaike Information Retrieval

**ANN –** Artificial Neural Network

**API –** Application Programming Interface

**BGRI** - Base Geográfica de Referenciação de Informação

**BMU** – Best-Matching Unit

**CCDR** – Comissão de Coordenação e Desenvolvimento Regional

**CML** – Câmara Municipal de Lisboa

**CPI** – Consumer Price Index

**DoF** – Degrees of Freedom

**DTM** – Digital Terrain Model

**GIS** – Geographic Information Systems

**HPM** – Hedonic Price Modelling

**MAE** – Mean Absolute Error

**MD** – Minimum Distance

**MLP** – Multi-Layer Perceptron

**MMH** – Maximum Margin Hyperplane

**OSM** – OpenStreetMap

**POI** – Point of Interest

**RMSE –** Root Mean Square Error

**SD –** Standard Deviation

**SDEM –** Spatial Durbin Error Model

**SLX** – Spatially Lagged X

**SOM** – Self-Organizing Map

**SQL** – Standardized Query Language

**SVM** – Support Vector Machine

**SVR** – Support Vector Regression

# INDEX OF THE TEXT

# INDEX OF TABLES

# INDEX OF FIGURES

# 1. INTRODUCTION

## 1.1 The Market and its Fluctuations

Apartment and room rent price changes are frequent and with an ever-increasing inflation of a monetary currency, terms such as "price hike" describe the phenomena of abruptly increasing prices that are often experienced in many places, be it countries' capitals, metropolitan cities or countries. As the financial capacity of the citizen oftentimes lags behind market supply's offer, rent control seems like a logical step. This is quite evident from places like San Francisco, among others, that have strict rent control to avoid exceeding the maximum allowed annual rent increase allowed. Hence, rents in these cities are limited from above by ceilings imposed by a government or a local administration. (Larsson, Rong, & Thomson, 2014).

However, rent regulation does not bring only advantages to the market, as unexpected side effects are oftentimes experienced. Although San Francisco has thorough rent regulation, it still is one of the most unaffordable cities in The United States. Evidence suggests that with a demand shock and a relatively low annual rent increase cap (60% of regional CPI rate), there is an increasing number of holders choosing to withdraw housing units from the rental market, even at cost of leaving them vacant (Asquith, 2019). Looking at The European market, Germany has also placed modest rental brakes on rent increase. Contrary to the policymakers plans, the law has at its best, no impact in the short run, and at its worst, it accelerates the price rent increase much closer to the imposed rental brakes (Kholodilin, Mense, & Michelsen, 2017).

The aforementioned markets were never so old, comprehensive and rigid as laws imposed in Portugal that were in force until recently. These laws had allowed for lifelong contracts with only minimal, inflation-linked price increase to be handed down through multiple generations. This law had been discontinued in 2012, when the government had presented a new law amending the New Urban Lease Act Law 6/2006 to ensure equally the rights and obligations of landlords and tenants. This legislation had set the five-year period of transition from the old lease contracts to a new regime of free rent, and introduced measures to broaden the conditions under which renegotiation of open-ended residential leases can take place, measures to phase out rent control mechanisms, and limits the possibility of transmitting the contract to first degree relatives (Branco & Alves, 2015). Moreover, an introduction of a golden visa scheme was introduced in 2012 that have encouraged investors to also buy Portuguese property, and with this, the real estate market had also endured a 27 percent rise within a 4-year period in the timeframe of 2014-2018 (Colectivo Marxista Lisboa, 2019).

Last, but not least, the increase of touristification creates changes to dwellings, resulting in undergoing processes of renovation in both historical and non-historical areas of the city (President & Marques, 2018). Bearing this in mind, the regeneration of the inner city then favors the movement of affluent, middle-class residents into these areas, triggering gentrification (Gravari-Barbas & Guinand, 2017). Citizens have not taken these trends lightly, as the process of gentrification in neighborhoods like Alfama and

have made Lisbon's old town rent market exceed the financial capacities of the average Lisbon's worker (C.A. Brebbia & J.J. Sendra, 2017). As a response to these changes there has been a wave of anti-tourism marches that has slowly spread across multiple cities in Europe including Lisbon and Barcelona (Hughes, 2018).

## 1.2 Factors Driving a Market – Predictive Modelling

The analysis of rent is focused mainly on more tangible variables that pertain to a household. These can be very few, and include an apartment's square meter area, street, and the year the apartment and building has been constructed, or otherwise, include as many variables as are publicly available to be assessed with. Various research works bring forward price prediction models that use data from Airbnb (Kakar et al., 2016), Twitter (Bing, Chan, & Ou, 2014) and other local websites like Sofang (Li, Ye, Lee, Gong, & Qin, 2017). However, literature does often not consider important spatial modules like proximity of buildings with important cultural heritage, distance of a point of interest (public transport stops, closest university, green areas library or co-working space), or location being rated as undesirable, oftentimes as a consequence of spatial factors.

It is of high importance to create a clear distinction between a real estate market, apartment rent market and the market of renting a room. Research made through applying correlations and Granger causality tests on the available data of Lisbon had given results that the housing prices are a good indicator that has causal implications over apartment's rent. Nevertheless, when re-validated using National Data, neither of the variables has been proven significant to refer to it as solid proof of being the cause of the other (Pereira, 2017). Another long-term temporal study of the Singaporean real estate market using co-integration analysis had also failed to establish a relationship between the two on nationally estimated indexes (Baltagi & Li, 2015). Therefore, a study covering the real estate market does not necessarily offer proof to be important in the price of the rent of the same apartment in question. Subsequently, room rent is a third different term that denotes an entity whose specifications of the outside of the room might become of secondary interest if they are to be shared with strangers, or landlords (and their family). Hence, research that has offered great findings of the real estate market, rent of an apartment, or even short-term rents does not directly compare to outcomes of this case study which is evaluating the impact of the mid-term rent prices which correspond to a monthly duration of service they are intended for.

Today, a lot of data is available to assess and use as a contributing factor in building a model, but oftentimes the standardization of this information lacks uniformity and proper integration. Moreover, a study that captured data from 4 years ago might have had different findings as the market endures a constant shift, both by internalized (economy, laws, wage, inflation) and external factors (the area wherein it's located might increase in desirability as new spatial proxies appear in the vicinities).

The prediction of price represents fitting a continuous range of numeric variables through other explanatory data. Hence, it is a regression problem and it requires a linear

or non-linear model using exogenous variables at disposal to yield an output able to explain the target variable of interest.

Many intelligent algorithms have been developed in recent years that make use of Machine Learning (ML) to solve regression problems. Some of them have a black-box design that restrains the user from fully understanding the model built as is the case with Multi-Layer-Perceptron (MLP). Others, however produce and outline information importance, as is the case of Support Vector Regression. Other, more traditional approaches include Hedonic Price Modeling (HPM), although various slight modifications of the algorithm exist to account for the spatial dependence of the data.

However, mid-term rental as a research topic has been a highly under-represented case, and even then, spatial variables, usually play a secondary part in the analysis. Subsequently, the idea behind the thesis was to incorporate methodologically constructed spatial data proxies in predictive models and analyze the possible impact they might have through tuning their configuration mechanisms and analyzing the outputs.

## 1.3 Research Objectives

The thesis aims to evaluate the influence and significance of spatial variables and corroborate proof of whether these data provide aid in the construction of reliable prediction models of mid-term room rental prices in the specified study area of Lisbon. The methodological approach applied is to be based on spatial econometrics and machine learning.

To achieve the aim given, specific objectives were defined, as follows:

- Analyze available points of interests within the study area and divide them into generalized categories from which knowledge can be extracted
- Review and apply suitable spatial econometric models of HPM to evaluate the significance of the variables provided
- Design and implementation of SVR for creation of a model that best fits the data. Evaluation and review of the accuracy metrics with, or without the new proposed variables
- Application of the Self-Organizing Map (SOM) algorithm through the use of GeoSOM software to check for correspondence of more natural clusters of the listings

## 1.4 Dissertation Structure

The dissertation is organized in six chapters. Chapter 1 gives introductory remarks, aims and objectives. Chapter 2 expands on the background and related works. The 3rd Chapter describes the contextual geographical location of the study area chosen, the dataset used, as well as the methodology undertaken to give spatial context to variables created from spatial features. Then, Chapter 4 comprises results and discussions thereafter. Finally, Chapter 5 summarizes the research work by portraying conclusions.

# 2. THEORETICAL FRAMEWORK AND RELATED WORKS

This chapter makes a brief introduction of the theoretical concept and defines the frequently used terminologies in the dissertation

## 2.1 Theoretical background

### 2.1.1 Artificial Intelligence and Machine Learning

*Artificial Intelligence* mimics the intelligence of a human brain, and is applied to, and subsequently, demonstrated by machines. The field studies intelligent agents. Hence, it utilizes the detected information through computational techniques to understand patterns and to learn how to solve given problems. Artificial Intelligence can be relevant to any intellectual task, and most prominent ones include medical diagnosis, self-driving cars, and image and speech recognition.

*Machine Learning* is a subdomain of AI wherein machines learn to perform a task without explicitly being given instruction, but by identifying existing patterns and gaining knowledge about continuous decision-making. Various algorithms have come to be developed throughout the years, including SVR, MLP and RF. Today, machine learning is having widespread application as it brings a large value to the ever-growing heaps of data being generated in a technology driven world. Its high computational abilities has allowed it to protrude in various sectors including finance, banking, insurance and internet fraud detection amongst others.

### 2.1.2 Spatial Econometrics

*Spatial Econometrics* is the field that interweaves econometrics and spatial analysis. Instead of understanding data observations as independent observations, new models are put into place that also account for the rule of geography wherein closer things are more similar (Tobler, 2004), and give meaning to neighbourhood effects and spatial spillovers.

*Hedonic Price Theory* suggests that a property can be viewed as an aggregation of many individual components or attributes. Looking from the consumer perspective, people are primarily purchasing goods that are embodying many attributes that maximize their underlying utility functions (Rosen, 1974). Hence, this technique had grown to be present in giving value to various commodities that include, but are not limited to residential amenities (Bourassa, Cantoni, & Hoesli, 2007), benefits of environmental improvements (Harrison & Rubinfeld, 1978; Kong, Yin, & Nakagoshi, 2007), and wildlife recreation resources (Messonier & Luzar, 1990).

*The Hedonic Price Model* in its simpler nature describes regression wherein the target attribute is a price of an amenity. For example, one can expect that attributes giving value to the price of a house should include land size, age of house, types of rooms inside and their count.

*Ordinary Least Squares (OLS)* is a simple HPM regression that takes the form of equation **(1)**:

$$\boldsymbol{y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i} \quad (\mathbf{1})$$

Where $y_i$ would be the *i*th observation of the dependent variable, x is a vector of all explanatory variables, and α is an intercept that holds the value of **y** when all explanatory variables are equal to 0. $\beta$ is the vector of the corresponding coefficients of the estimates predictors and $\boldsymbol{\varepsilon}$ is the random error.
When dealing with given data, the OLS model has certain assumptions, whose violation would lead to an inappropriate model. One of the assumptions specifies that finding spatial autocorrelation in the residuals of the predictions shows that there is a spatial pattern in the data under analysis and more appropriate models exist that can account for it.

*Spatial Lagged X (SLX)* portrays a spatial regression model that extends to include explanatory variables observed on neighboring cross-sectional units terms. In order to improve the model, a spatial exogenous autoregressive lag is introduced (Elhorst & Halleck Vega, 2017). The equation **(2)** would then take the form of:

$$\boldsymbol{y_{it} = \mathbf{W} x_{it} + \mathbf{X}_{it}\beta + \varepsilon_{it}} \quad (\mathbf{2})$$

Then, $\mathbf{X}_{it}$ represents the autoregressive coefficient that produces the global spatial spillover effect. If this comes to be equal to zero, the model simplifies to a conventional linear regression model (Lesage, 2008). Nevertheless, this approach has its drawbacks because assumptions are made that all the spatial dependence among residuals is caused by variables one can measure.

*Spatial Durbin Model* is an extension of SAR that includes the average of the variables from the neighboring houses, but it extends to allow for global diffusion of shocks to the model disturbances. This means that the spillover can be felt over multiple regions, but the decay parameter λ controls that higher-order neighbors (indirect ones) to receive less impact than the direct ones (Lesage, 2014). Additionally the *Error* addition within this model creates the SDEM model that accounts for spatial dependence among the error terms and redistributes this error among the sample. The equation in vector form takes the equation **(3)** (LeSage & Pace, 2009):

$$\boldsymbol{y_{it} = \mathbf{D}_{it}\delta + \mathbf{W} y_{it}\rho + \mathbf{X}_{it}\beta + W\theta Z u} \quad (\mathbf{3})$$

$$\boldsymbol{u = \lambda W u + \epsilon}$$

*Dummy Variables* are binary variables that equal '1' if the target variable is described to either, contain something (e.g. terrace with seaside view), or it belongs somewhere (e.g. an apartment belongs to the parish of Alfama). '0' then describes the opposite case.

Each of the variables can have a positive or a negative relationship with the price. In an ideal case, minimum distance to an amenity makes a price rise. Nevertheless, the market can behave unpredictably or interact with other variables to yield a negative relationship. For example, a rent of room in an apartment where the ratio of number of

bedrooms divided by number of bathrooms is high might yield a negative effect. Additionally, having a minimum distance of clubs below 20 meters might dissuade potential tenants away from a property.

*Direct effect*: Change of a particular entity of a listing (e.g. renovation of apartment from 1979 to 2019) changes the price target variable.

*Indirect effect*: Measures the impact on the price of a target variable from changing an exogenous variable in another data point.

*Weight Matrix* - A matrix W is a matrix of order $n$ x $n$ where $n$ is the number of regions. Non-zero elements in row $i$ and column $j$ in the matrix then represent the element $n_i$ is a neighbor of element $n_j$. Rook and Queen each refer to two common ways to calculate statistics for focal cells, and these are known as Moore's and Neumann's neighborhoods. These spatial weights are calculated, such that each element in a matrix represents a spatial weight between two neighbors. The spatial weights $\boldsymbol{W_{ij}}$ are non-zero when $i$ and $j$ are neighbors, or zero otherwise. The difference between the two types of neighborhood is that in Rook contiguity, common vertices do not make two polygons - neighbors. The use of queen neighborhood was the chosen approach in order to reflect real-life phenomena where if two neighborhoods meet at a common corner, they would still cause influence between one-another.

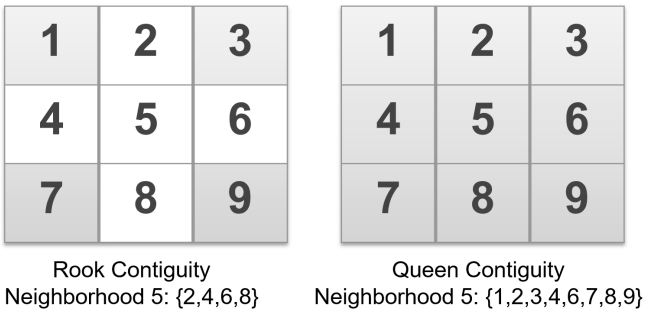


**Figure 2.1** - Types of Weighted Neighbourhoods

### 2.1.3 Support Vector Regression

SVM originated from an algorithm implementation from 1995 by Cortes and Vapnick (Cortes & Vapnik, 1995). The algorithm had been first introduced in order to solve pattern recognition problems. SVM is described as an algorithm that makes use of a hyperplane constructed in a higher dimensional input space that separates the data in this high dimensional space optimally (Suykens & Vandewalle, 1999). Nevertheless, what began as a classification problem, rose to application of solving function estimation problems. These findings proceeded to additional inclusion of optional parameters like epsilon's loss function and introduction of different types of kernel functions in which data can be more easily divided. This way, there could be different solutions produced with different complexities and different thresholds of errors in order to build the most cost-efficient model that could be deemed a good enough fit for the purpose it had been built for.

There are infinite possible ways to construct a hyperplane, and in classification, the best hyperplane can be determined as the one that holds the largest distance between hyperplanes and the support vectors. Subsequently, detection of SVs and Lagrange Multipliers demonstrate the records important for building the model (Witten, Pal, & Fourth, 2017).

*Epsilon within SVR*

With given training data $\{(x_1, y_1) \ldots. (x_e, y_e)\}$ where $X \subset X \times R$, where *X* denotes the space of the input patterns, SVR is looking for a function *f(x)* that would have at most ε deviation from the actual targets for all training data, whilst maintaining the largest flatness (Smola & Schölkopf, 2004). A linear fit of a regression function is presented in Figure 2.2:

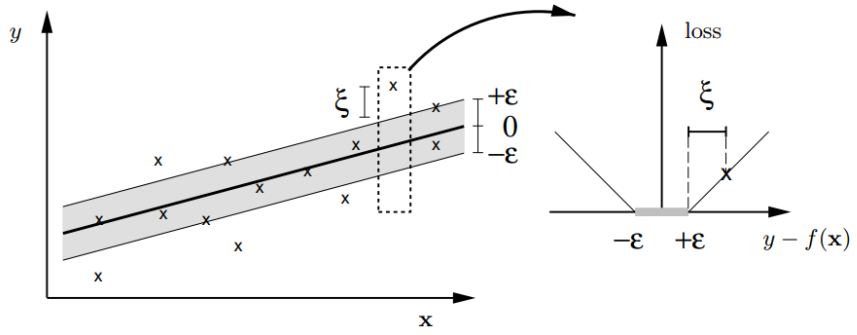


**Figure 2.2** - Tube with radius ε (Schölkopf, 2002)

If all data points fit within the tube presented, the fitted model then has an error 0. In this way, a model can be built with an infinitely high value ε. This would lead to no penalty given to any data point and an error of 0, but this would be a meaningless accuracy measure.

In a linear case the SVR function takes the form of Equation **(4)**

$$X = b + \sum i \, \alpha \, a(i) \cdot a \tag{4}$$

The dot notion can be substituted by various kernel functions if a non-linear problem is presented. Most common kernels used are:

- Linear kernel $X_i$ – Equation (4)

$$K(x_i, x) = x^T x_i \tag{5}$$

- Polynomial kernel – Equation (5)

$$K(x_i, x) = (x^T x_i + 1)^d \tag{6}$$

- Radial Basis Function kernel – Equation (6)

$$K(x_i, x) = \exp\left(-\gamma\left|\left|x_i - x\right|\right|^2\right), where\ \gamma = \frac{1}{2\sigma^2} \quad (7)$$

Only the support vectors are important for the model – the deletion of all other rows bears a coefficient 0 and hence, does not change the outcome of the predictive model.

Another variable whose value plays importance in building the model is the regularization parameter C that denotes the limit of the absolute value of the coefficients $\alpha_i$, which gives the limit of how much it is needed to avoid misclassifying a training example. The limit cases then become the following:

I. the larger C becomes, the closer the function can fit the data in the hyperplane and smaller-margin hyperplane will be needed
II. an ε of 0 creates a least-absolute-error regression with constraint C
III. A large value of ε just outputs the flattest tube that encloses all data

Gamma ($\gamma$) is a specific parameter for radial basis kernel that defines how far an influence of a single training example reaches. A large gamma would correspond to more support vectors having influence on the hyperplane.

### 2.1.4 Self-Organizing Map and GeoSOM

*The Self-Organizing* Maps are a clustering algorithm introduced in the 1980s (Kohonen, 1982). The main idea is to map high-dimensional data, into dimensions from which the human eye can understand and extract patterns. The units by themselves can be loosely or closely connected to each other, and the deciding clusters are decided by the user. SOM algorithm can have different distance metrics, but most often makes use of the *Euclidean distance* as a measure of closeness between two arbitrary units. Once a SOM has been trained with a given pattern, all the units move towards the best matching unit (BMU). At the end, patterns that are similar in the input space should also carry this behaviour in the output space.

*GeoSOM* is an adaptation of SOM to account for the specificity of spatial data. The search of the BMU takes part in two cycles. Instead of searching for a BMU throughout the whole dataset, it tries to find a neighbouring one that is limited in the search radius with the parameter *k.* The output space in SOM usually takes the form of a 2-dimensional space (Kohonen, 2001) as the easiest one to visualize.

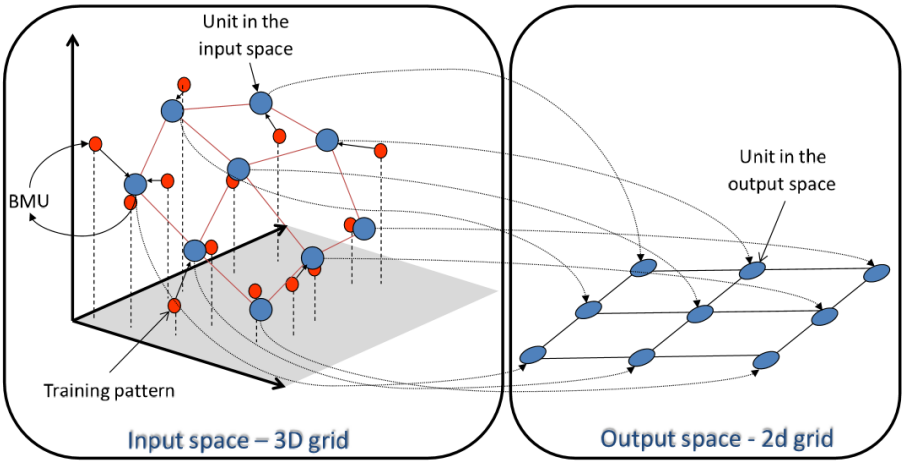


**Figure 2.3** - Self-Organizing Map - I/O space (Henriques, Bação, & Lobo, 2009)

There are a number of important parameters that need to be tuned when training a dataset with SOM. These include learning rate, radius, a number of iterations, and they need to be chosen accordingly to minimize a model's topological error.

## 2.2 Related Work

Throughout this subsection, references and links to studies regarding price prediction are mentioned to gain a general overview of what research has been applied and the correspondent findings follow. For a clearer structure, the related works are divided into three sections, one denoting Spatial Econometrics' related studies, a second for SVR related studies, and a third for giving briefing of other algorithms' usage.

### 2.2.1 Research through Application of Spatial Econometrics

Research shows indisputable evidence of the impact of the cultural heritage on the real estate prices in the city. A case study of the city of Lisbon finds proof that a protected zone can produce a negative impact on the house value but it also produced findings that when accounting for the heterogeneity of the areas under question, the effects had seemingly disappeared (Franco, Macdonald, Franco, & Macdonald, 2016).

Another study has revealed that overall, historic amenities generally give rise to dwelling prices in order of 4.2%, but once the radius broadened, a high concentration of these historical amenities started to yield a slightly negative effect of 0.1% (Franco & Macdonald, 2016).

When analyzed, these papers bring about conclusions that living in a proximity of an amenity with some degree of cultural importance gives exposure and identity of an apartment being in the center of a local cluster (a desirable residential area). Living in close proximity to many of these objects on the other hand had shown a correlation to

it being a highly touristic area, which in turn is valued as a very desirable dwelling for one's day-to-day life.

Other research has also found tax exemptions to drive positive spillovers to nearby properties (van Duijn and Rouwendal 2012, Ahlfeldt, Holman, and Wendland 2012, Coulson and Lahr 2005).

What's more, using a HPM on real estate prices have found Lisbon's 'Metro Station Proxy' variable coefficient to give between 3.49% and 5.18% rise in prices with rated positive accessibility to a single metro line, or between 4.62% and 6.17% for accessibility to two metro lines and even larger for the rail accessibility (Martínez & Viegas, 2009).

Additionally research have also used additive hedonic regression models to account for the heterogeneity in the rent market (Brunauer, Lang, Wechselberger, & Bienert, 2010) for real estate prices in Vienna. This had led to conclusions about discoveries of submarkets which have as a significant factor a 'Belonging' of listings in the specific city's districts.

Other sources (Moro, Mayor, & Lyons, 2013) include a study over the city of Dublin as a case study where heritage sites within the city were categorized as factors using a limited number of categories. This study had given a small but significant positive score to certain types of categories (historic buildings and memorials) as a spillover effect to apartments in the vicinity, although other categories (archeological sites) had produced a negative one. The study had been made through producing dummy variables as flags of whether a certain type of building was in a radius of a chosen amenity. Variables pertaining to metros and bus stop as means of public transport had also been used and it had been noted that living within 100 meters of a metro station resulted in a positive correlation with a price increase of an apartment (Franco et al., 2016).

(Poort, 2015) have moreover argued that the presence of cultural heritage attracts highly educated households using their case study of the Netherlands. This factor had also caused a spillover to industries in the country investing and preferring to reside where the elite households of the country reside (Marlet and Woerkens, 2005).

Last, but not least, other study findings include cultural heritage sites itself carrying a multiplier effect to other amenities including restaurants and shops (van Duijn, 2013). This gives an idea of linear dependencies between exogenous variables could clutter or bias an analysis and make for an overtly complex model.

These studies, for the most part, deconstruct many spatial variables segregated and analyzed separately (as objective of understanding single variable's importance) rather than fitting one model using many. The majority of them use Euclidean distance to calculate proximity, wherein a PoI can reside anywhere within a proposed radius threshold of an amenity in question.

### 2.2.2 Research through Applications of SVR

Finding using the SVR include (Chen, 2010) that discuss that stratification of Shanghai's market into more homogeneous subsets provides considerable benefits as contrast to taking the aggregate of the whole Shanghai's rent market. The approach used a hedonic appraisal model as a preprocessing step of choosing the right variables before building the SVR. Another paper has built a real estate forecasting model based on particle swarm optimization (PSO) before applying the SVR (Wang, Wen, Zhang, & Wang, 2014). Optimization through HPM seems to be a recurring theme when using SVR and SVM. In this way, the significant variables need to be detected first before building the model so the time complexity also lowers substantially, whilst also making sure the model's accuracy does not suffer from the multicolinearity problem.

### 2.2.3 Research through Usage of Other Algorithms

A recent study of real-estate price evaluation was done through the usage of both NN and HPM as a way of comparing the two algorithms. Nevertheless, this research was concentrated on producing minimum MAE accuracy model for predictions and does not give further info about the singular influence of unique variables in building the model (Safronov, 2017). This is because of the "black box" nature of ANN in which multiple neurons are exchanging information and are updating their weights, but to the outer world not a lot of substantial information is given through which the observer can get many conclusions. Another study was done using ANN on the island of Cabo Verde using 1092 data points with a yearly temporal scale between 2009 and 2014. This study also came about as a comparison between using the algorithm of ANN and Random Forest and draws conclusions of MLP producing higher errors than the latter algorithm (Ester & Martins, 2016). The most important variables there were found with the model fit, and among them, the location of the apartment and the square meter of the apartment seemed to cause both highest significant and positive influence, with the closeness of public institutions and an existing balcony in the apartment computed as ones with least importance. ML studies often have a lot of research concentrated solely on finding the best model with the least error (Cristina and Teixeira 2009, McCluskey et al. 2013, Limsombunchai et al. 2004) with using NN and MLP's as ML standards that manage to build models that capture the data's behavior the best, producing the smallest error. One can also note usage and comparison of Random Forest in real estate predictions in Ljubljana (Kilibarda, 2018), where findings suggest in the study area under evaluation, it performed significantly better than other ML algorithms. These are just a few small references of the vast usage of many different algorithms within the ML field that could be used to assess this type of data, although it must be noted that every algorithm varies in the output information and hence, the type of analysis it allows for.

# 3. DATA AND METHODOLOGY

This chapter explains in detail the undertaken steps to produce this dissertation. First, a geographical context is given in Section 3.1 wherein general information about the encompassing study area is introduced. In Section 3.2 a brief overview of the proposed architecture. The chapter continues with Section 3.3 thereupon giving a synopsis of the software and algorithms that are imperative for producing new spatially consistent data. Section 3.4 then describes the data sources and variables choices. The final, Section 3.5 portrays how the different undertaken parts of the methodology come together to encapsulate the analysis.

## 3.1 Geographical Context

Lisbon is the capital and largest city of Portugal, with a bounding box describing a latitude range between -9.2379 and -9.0863 and a longitude range of 38.6800 and 38.7967. The administrative area of the city covers approximately 100.05km$^2$, and the city's last census had recorded 505,526 citizens. A clear distinction is made between Lisbon's administrative area and the urban area that extends far beyond these limits. The positive elevation goes from 0 at its minimum up to 227 meters of altitude.

The city has a circular physical configuration covering approximately 12 kilometers both east to west and north to south. Official divisions divide it into 24 parishes, or in 3623 subsections at the lowest mapping level. Maps of the parishes' polygons and the level 4 subsections are shown in Figure 3.1 below.

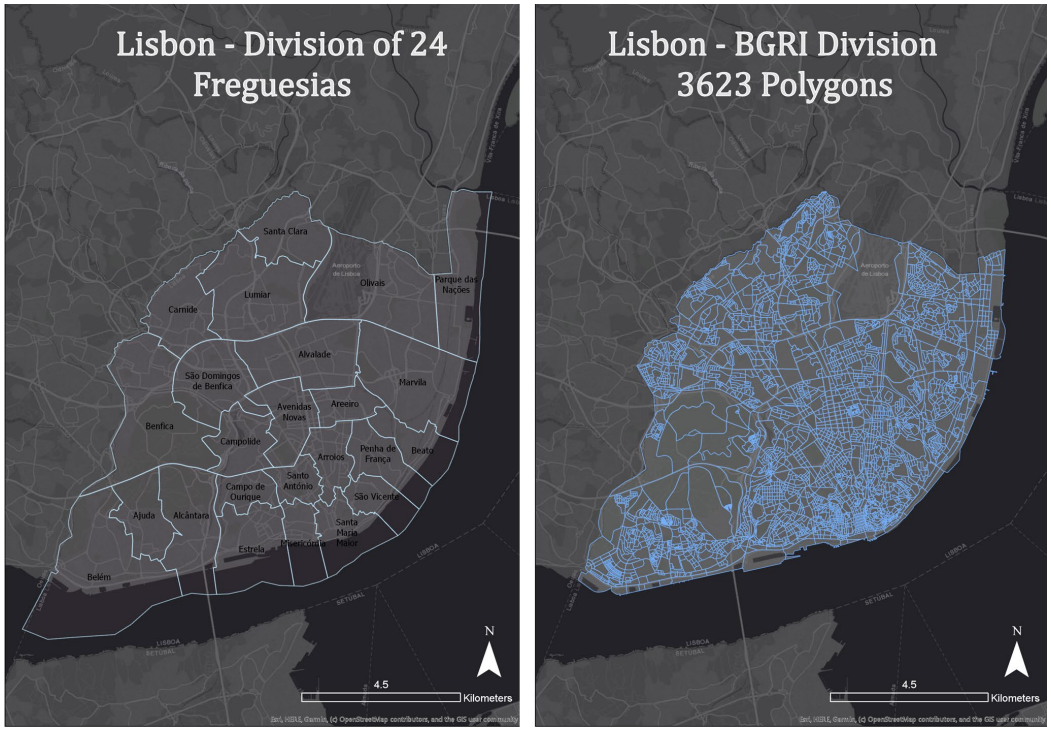


**Figure 3.1** - Study Area Divisions: Parishes versus Level 4 Census Blocks

## 3.2 Proposed Architecture

The purpose of the methodology is to try to incorporate many different sources, according to the availability of the data and the categories of data points they offer. The policy of transparency of the EU gives space to use open data freely which helps in the process of gathering an abundance of data that could be inter-checked for reliability between the different sources and gives the choice between using sources with the biggest quantity, biggest quality or integrations. Nevertheless, this also brings about weight and complexity for computations needed for replicating the analysis. For easier processes, all data is imported to a PostgreSQL database, from where it could be easily exported and read in both QGIS and RStudio. This integration of environments provides a robust bridge for easier data conversion, manipulation and additional parallelization of computational processes. The proposed architecture is found in Figure 3.2.

The Figure shows little detail of the R analysis, although the packages used are to be mentioned in the following section.

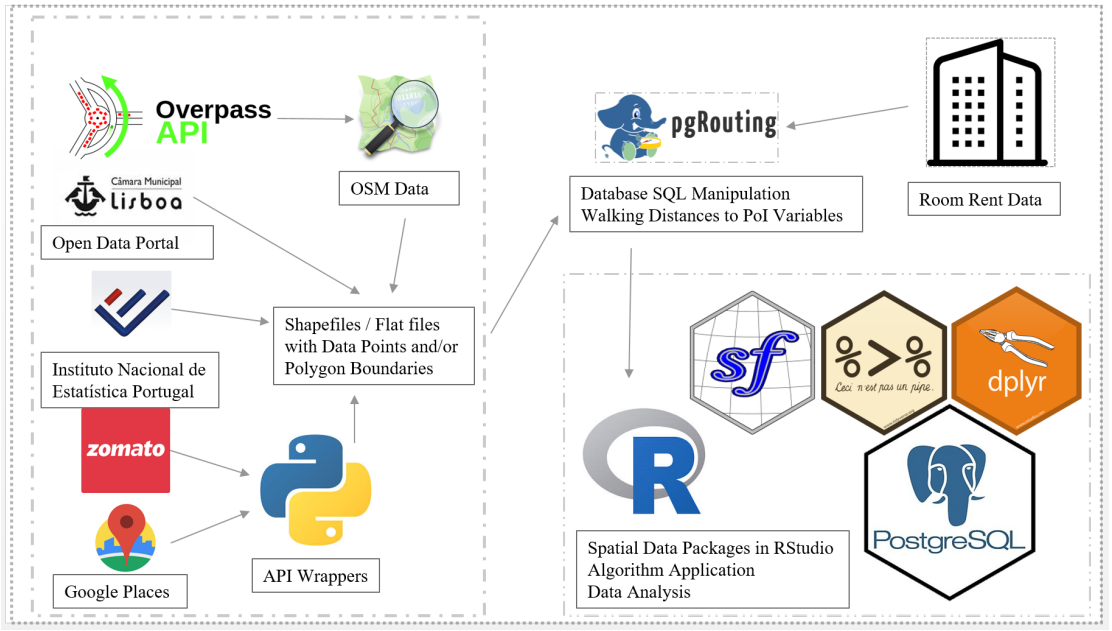


**Figure 3.2 –** Proposed Architecture

## 3.3 Hardware and Software

Concerning *hardware*, the research used an Intel Core i7-7700HQ with 16 GB of RAM capacity, under a Windows 10 operating system.
Regarding *software*, the majority of the undertaken analysis was done within PostgreSQL and R. Additionally ancillary software was used for visualization including QGIS, ArcGIS Pro, as well as minor checks and flat files of the dataset for versioning history as excel files.

### 3.3.1 PostgreSQL

This is a free and open-source relational database management system that extends the SQL language where users can safely store and scale complicated data workloads (The PostgreSQL Global Development Group., 2014). When dealing with spatial-data, the

database capabilities can be accompanied by additional add-ons. *pgAdmin4* version is used, alongside the following extensions:

- **PostGIS**: Allows for geographic location queries to be run in SQL (PostgreSQL Global Development Group, 2020).
- **pgRouting**: Enables a road network analysis to be performed; It is used to calculate the shortest path to places through *pgr_drivingDistance* function (OSGeo Foundation, 2018)

### 3.3.2 R

R is a system for statistical computation and graphics. It provides, among other things, a programming language, high-level graphics, interfaces to other languages and debugging facilities (R Core Team, 2019). The language is able to be executed on many operating systems free of charge. The *RStudio,* subsequently, represents an interface for the execution of R code. Packages that make the analysis possible are listed as follows:

- **sf:** Package that allows for reading, writing, manipulating and visualizing spatial data and geometries, to and from R (Edzer Pebesma, 2019)
- **RPostgreSQL:** Database interface and 'PostgreSQL' driver for R (Conway et al., 2019). With this package direct connection to the PostgreSQL database is made, and tables and queries of the database can be directly imported and saved as variables within the RStudio environment
- **dplyr:** A tool used for easier manipulation of data frames (Wickham, 2019)
- **magrittr:** Operator that forwards a results of an expression to the next function that comes after (Stefan Milton Bache and Hadley Wickham, 2014)
- **caret:** A comprehensive framework for machine learning models (Kuhn, 2020)
- **e1071:** A package used for intuitive application and analysis of the SVR algorithm, one among many probabilistic clustering and regression models (Meyer, 2019)
- **spatialreg:** A collection of estimation functions for spatial cross-sectional models (Roger Bivand 2019). Library used for applying HPM on the data
- **spdep:** A collection of functions to create spatial weights matrix objects from polygon 'contiguities' by distance and tessellations (Roger Bivand 2019). This package is used for applying functions in the HPM spatial regression analysis.

### 3.3.3 Python

Python is an interpreted, high-level, general-purpose programming language. The language can be used for any computational purposes, although for the means of the dissertation, it was used for finding specific-purpose API Wrappers. The API Wrapper then provides an interface between an API and a programming language, so all provided API methods for data collection can be made through the desired programming language's function calls and wrappings.

## 3.4 Data Description and Resources

In the following section, all data used for this study is to be briefly described.

### 3.4.1 Room Rental Sources

Listed below are the two data sources used. The data collected was used only for the means of this research, and no personal data was collected that could expose an individual's identity.

***Uniplace*s** is an innovative service, and a global brand primarily used for mid-term accommodation, which thus puts the average target of accommodation between 6 months and 1 year of contract binding, targeted at students. Nevertheless, the service is not necessarily discriminative towards anybody that can potentially apply for booking accommodation at one of the listed rooms and apartments (Uniplaces, 2019). There are room listing for many cities and information taken for the case study contains the variables shown in Table 3.1

| Variable Type | Variables |
|---|---|
| **Numerical** | Number of bedrooms, Number of Bathrooms, Price of Room, Room Square Meters, Apartment Square Meters, Minimum Stay |
| **Binary** | Overnight guests, Full-time employees (Allowed), Pets, Smoking, Domestic Students, Type of Bed, Preferred Gender, Private Bathroom, Landlord Gender, Outdoor Area, Elevator |
| **Text Attribute** | Description of Room/Apartment, Review |
| **Categorical** | Bills Inclusion, Cleaning Frequency |

**Table 3.1** – Variables Types and Names

Most of the binary variables can have an N/A value depicting missing information (the exceptions being Type of Bed, Type of Booking and Private Bathroom). Moreover, Room Square Meters and Apartment Square Meters are not mandatory variables to be input by the host.

***BQUARTO*** (BQUARTO, 2010) is the second web source that offers similar services as Uniplaces, but its only target market is the country of Portugal. The variables at disposal portray a similar pool, however there are no reviews, and the items listed as available within the room are far scarcer to be mentioned. The two datasets are merged and the variables unified to get a total of 2149 unique rooms.

### 3.4.2 Zomato API Restaurant Data Collection

***Zomato*** is a food delivery start-up founded by Deepinder Goyal in 2018. It contains information, menus, reviews and delivery options from many restaurants throughout the world (Zomato, 2019). Zomato's API allows for easy retrieval of restaurants within a city, and it offers multitude of variables including restaurant types, ratings and price range.
This website includes an API that allows 1000 free requests a day, and the city of Lisbon bears the id of "82". Conversely, once a call has been made, the server returns only 20 most important PoIs from which all related information can be extracted. To be able to

extract all PoIs, a common approach would be to search within the radius of a limited distance. To start, an arbitrary radius of 1km is chosen as a starting point, with results sorted by distance. If the requests yield less than 20 restaurants, then this grid size covers all existing restaurants within the proposed radius of that sample point. If not, the area is further divided within a smaller level grid to collect new data. This behaviour can be repeated, as diving deeper into smaller tessellating square grids covers smaller areas until all records are extracted.
A library that provides ready wrapper APIs for Zomato is "pyzomato", and only a key from Zomato which has not exceeded the daily allowance of requests is needed (MIT, 2016). The data collected from Zomato counts 8629 unique restaurants.

### 3.4.3 Google Places API Data Collection

*Google Places API* has a similar way of functioning, where a single API request would target points of interest near a coordinate, ordered by importance. There is a "next page" redirect, which returns the following page with 20 new results. Still, the pages have a maximum of 3 redirects. This behaviour needs a denser tessellated grids, as it also contains a denser count of amenities within a chosen radius since PoIs can be obtained from any commercial business of interest, as well as buildings as objects (Google-Cloud, 2019). Google Places provides a division of 97 types of PoIs, and the outputs contain exact coordinates of the records, and are thus suitable for further analysis.


**Figure 3.3** - Tessellation of Study Area for Maximization of PoIs coverage

### 3.4.4 OpenStreetMap Data

*OpenStreetMap* (OSM) is a project that aims for a free geographic database of the world, with a specific target of having a recently updated version of the available spatial features on earth (Haklay & Weber, 2008). The idea is based on a crowdsourced

gathering of information from people that want to contribute towards this collection. The statistics have recently risen to count 5.7 million users registered on its platform (OpenStreetMap, 2020). A common middle ground of users that are producing content stands at 20%. This, however, describes a common statistic for applications that rely on users bringing in its value (Haklay & Weber, 2008).

*OSM PoIs* are downloaded through a German geofabrik server that stores OSM data about various points of interests divided by categories (Geofabrik, 2019). Partial usage of categories from the PoIs is available, and the choices are explained further in Section 3.5.2.

*OSM Road Network* – The creation of routes from an origin to a destination is commonly referred to as the shortest path problem. Each node can have many edges, through which one or multiple other nodes can be reached. This describes the graph theory which lies behind the construction, visualizations and computations within road networks of a city. The OSM data for the proposed network is created through an import tool called *osm2pgrouting*. The tool allows for the creation of a network where two tables are present that carry the weight of nodes and the vertices through which cars, pedestrians and/or bikes can subsequently traverse from an origin to a destination at a certain cost (given in *meters)*. The configuration of the network lies in a configuration file (mapconfig.xml). Highways are removed, as the proposed places of interest should be reachable to commuting pedestrians, rather than vehicles.

*OSM Data Quality Concerns* - Although Google and Zomato keep a constant flow of updates of their database to remain a competitive force in their respective field on the web, OSM's data quality portrays a unique issue with crowd-sourced contributions wherein no quality control procedure to guarantee the safeness of its use is present (Haklay & Weber, 2008).
One common quality of data is its completeness, but the specificity of spatial data makes this an unfeasible task since things can change vastly within temporal scale events and spatial point features, could turn to ruins through a natural disaster. What's more, the data completeness of OSM varies from country to country, and undertaken studies of OSM coverage can lead to different findings according to the features under analysis (e.g. road network, point of interest, administrative boundaries) (Girres & Touya, 2010; Hochmair, Juhász, & Cvetojevic, 2018; Neis & Zipf, 2012). Nevertheless, the studies also point out that the major cities coverage is vastly better than the rural areas. The latest findings suggest that today, upwards of 83% completeness of the real-life data is present within a 95% confidence interval. (Barrington-Leigh & Millard-Ball, 2017).
Another type of accuracy of interest is positional accuracy. As a user-generated data platform, OSM relies on GPS devices and skilled contributors to carry this weight, which in turn might produce data with accuracy errors in areas wherein high buildings reside or where a device's signal is prone to flutter.
Because of these reasons, when it comes to spatial features, if a PoIs categories' quantity was either comparable or higher from other sources (Google Places, CML), the OSM PoIs records from the category in question were dismissed.

### 3.4.5 Census Data

The Census Data from 2011 is a realization from the National Statistics Institute from Portugal, on the day of 21 of March 2011. The study counts 10 562 178 individuals and gives information of aggregate statistics for both individuals and buildings within the proposed census block level of analysis (National Office for Statistics, 2011). The levels of the analysis bears the 4$^{th}$ level, as shown in Table 3.2

| Level | Scale | Count |
|---|---|---|
| **1** | Municipality | 1 |
| **2** | Parish | 53 |
| **3** | Section | 1054 |
| **4** | Subsection | 3623 |

**Table 3.2** - Census Data – Sections Divisions

Level 4 was chosen for the purposes of this study, as it represents the smallest aggregation of census block statistics. It can be noted that the parishes count differs from the one shown in Figure 3.1, as new parish divisions of Lisbon were presented in 2012, a year after the last Portuguese census at the time of this writing.

### 3.4.6 Ancillary Data

Ancillary data from Open Data sources further includes these selected variables, as shown in Table 3.3

| Data Source | Variable | Type |
|---|---|---|
| **CML** | Parks and Gardens | Polygons |
| | Public Transport (Bus, Train, Metro, Lifts) | Points |
| | Parks and Gardens | |
| | Hotels | |
| | Health Centres, Pharmacies | |
| | Sport Facilities | |
| | Education (1º, 2 º, 3 º, Vocational) | |
| | Fitness Centres | |
| | Religious and Cultural Heritage | |
| | Art Galleries | |
| | Statues | |
| | National Monuments | |
| | Cemeteries | |
| | Theatres | |
| | Museums | |
| | Commercial Centres | |
| **CCDR** | PM2.5 Particles | |
| **Open Data Lisbon** | DTM - Altitude | Raster |

**Table 3.3** - Ancillary Data – Sources and Variables

All of the collected data from Zomato, Google Places, OSM PoIs, Road Network and its configuration file, as well as Open Data shapefiles with an accompanying timestamp of September 2019 can be shared upon requests to the writer.

### 3.4.7 Other Variables Importance – Access Limitations

Some studies show a very clear negative coefficient index for apartment prices wherein crime rates were higher (Ceccato & Wilhelmsson, 2016). Others present findings that noise negatively influencing a buyers' decision to purchase a property. There is also some evidence of urban traffic noise having adverse economic consequences (ERF, 2004), and this comes from noise pollution proxy which is described as Noise Depreciation Sensitivity Index. In this way, hedonic estimates with continuous noise data were analyzed on the city of Zurich and its flight regime to estimate the effect of aircraft noise on rental rates. Following the study, it is mentioned that under the Swiss protection law, landlords are able to receive compensation for the decrease in price as a consequence of noise. The paper follows to present findings that the apartments that experienced an increase of more than three decibels of noise within 3 years, showed a decrease of 0.5% within its rental price for each decibel. However, semi-parametric analysis revealed that the relationship between aircraft noise and rental rates did not satisfy functional forms throughout the entire noise distribution, and consequently, the noise was seen as redundant unless the noise levels were at least medium (Boes & Nüesch, 2011). Unfortunately, the data mentioned in this section might have a possible significance, but was not available at the desired levels of the case study.

## 3.5 Methodology

This section describes data cleaning and integration when multiple sources contain spatial features providing the same category of PoIs. Preprocessing steps needed to be undertaken for the creation of a uniform dataset. The Methodology can be seen in Figure 3.4

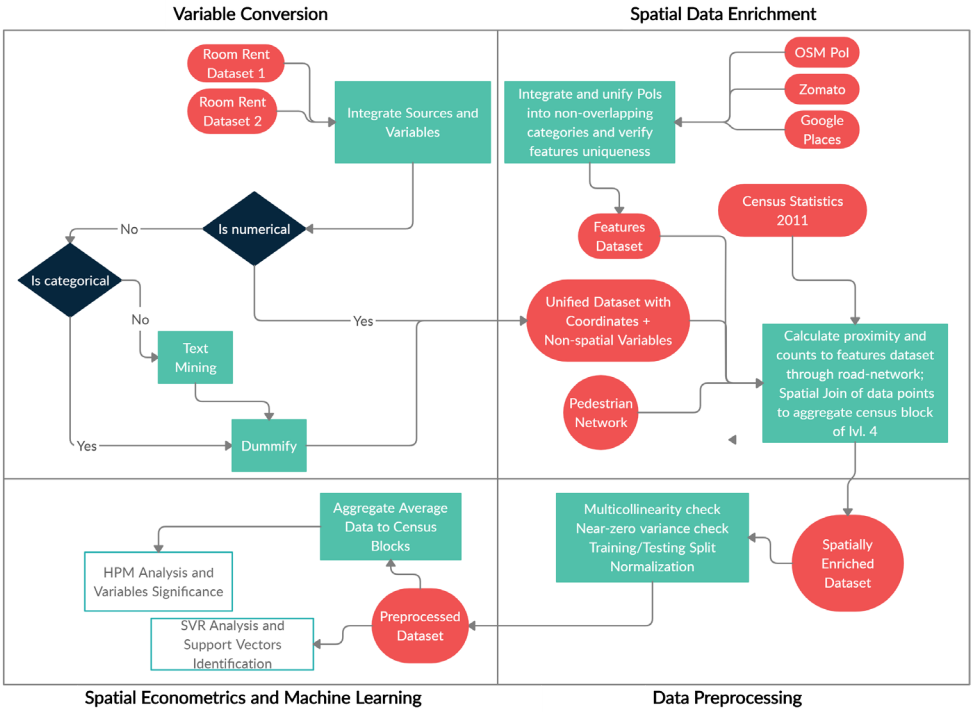


**Figure 3.4** - Methodology

### 3.5.1 Variables Conversion

Variable conversion is the process wherein the variables are converted into meaningful data to be fed to the algorithms. For this, all non-numerical variables need to undergo conversion. According to the types shown in Table 3.3:

- All numerical variables take the same form in the unified dataset
- Each mandatory binary variable takes the form of 1 if the room contains the amenity or 0 the lack thereof
- All non-mandatory binary variables that contain a large amount of N/A's mean this variable contains a third value of N/A which cannot be quantified as a 0, since it signifies a lack of information, not a lack of existence. In these cases, instead of putting arbitrary third value 2 (or any number that bears no significance), creation of three dummy variables (e.g. outdoor_areaF, outdoor_areaT, outdoor_areaNA) is used. This is one of many ways to impute missing data values, although there is proof that this type of imputation could lead to bias of coefficients in regression analysis (Allison, 2002). Because of this, when variables are noted to have many N/A values (>30%) the variable is deleted. This is where both the size of the room and size of the apartment are removed since above 84% of the data entries are missing
- Categorical variables follow the creation of N amount of dummy variables equal to the count of their unique categories
- Text mining is performed on the short description text, using unigram n-gram bagging. This, in turn produces a specific word binary variable (e.g. a variable named "cosy" would contain 1 in each data record that contains this word in the text description, or 0 otherwise. Connecting words (in Portuguese and English), or description that would be true for all records are not taken into consideration (e.g. "Lisbon", "and", "de")

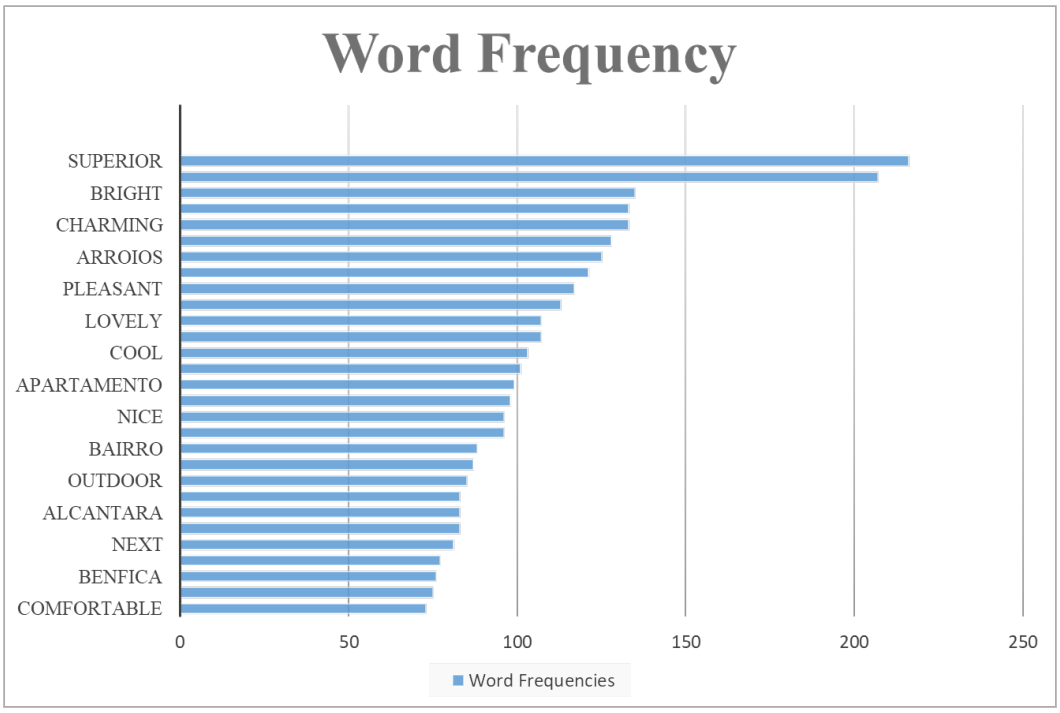


**Figure 3.5 –** Top 20 Word Frequency in Room Descriptions

Nevertheless, the word frequency counts were small with only two words with more than 10% appearance among the records describing "superior" and "bright. The word

dummy variables were, in turn, not included. The most frequent words are shown in Figure 3.5 above.

### 3.5.2 Spatial Data Enrichment

Data from many of the sources overlaps. Therefore, decisions for unification of these data into one source were undertaken.

The data from Zomato holds the largest quantity of restaurants, and contains information pertaining to reviews and the costliness of the restaurants. Hence, it was decided to use it as the only source for food-related PoIs. Other PoI categories from a single source (CML) were merged depending on whether the differentiation between them failed to appear as a response in similar studies of this type (e.g. private and public schools of first, second degree and pre-cycle are merged). Medical practitioners were subsequently merged with independent doctors. Nevertheless, hospitals and medical centres are left as a separate category.

If two used sources are found to hold a substantial amount of different unique data, *ArcGIS Pro Analysis – "Generate Near Features"* function is used for detection of all features between the two layers that are within 2 meters one other. These were presumed to represent the same data point and were deleted from the second source before merging.

The categories of amenities chosen are listed in Figure 3.6 and Figure 3.7 below.

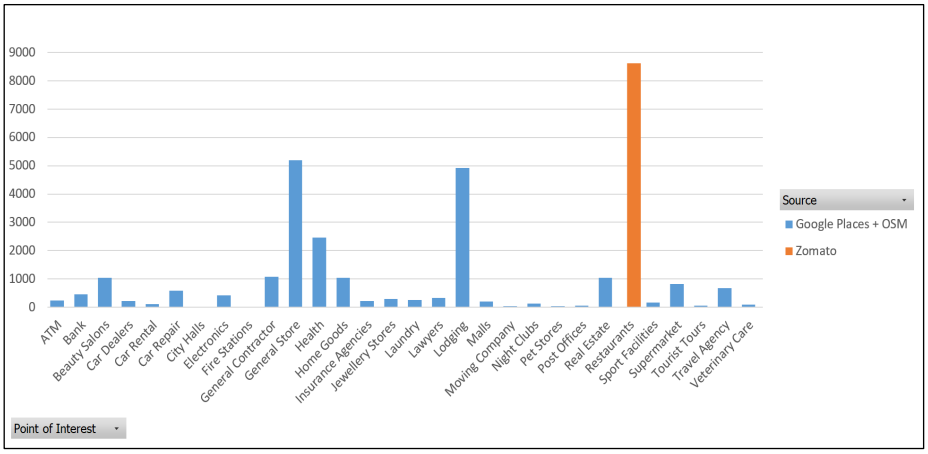


**Figure 3.6** - Counts of Features per Category A

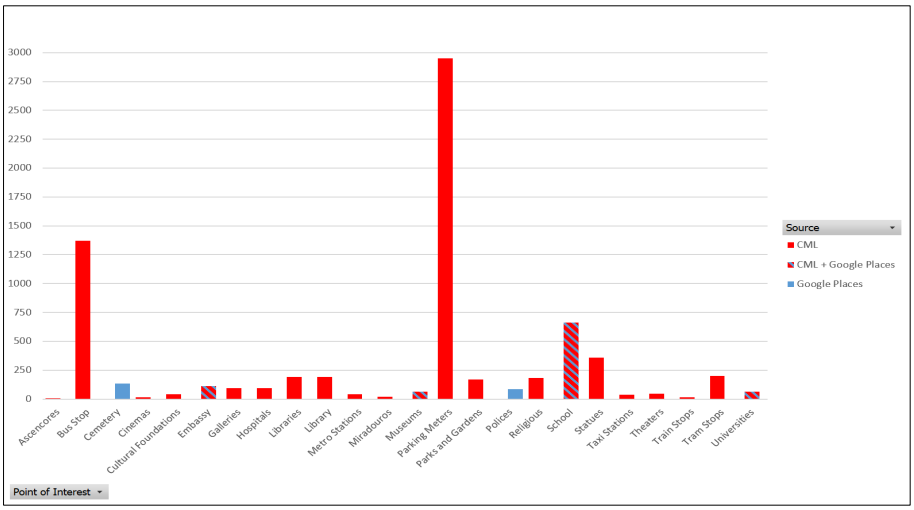


**Figure 3.7** - Counts of Features per Category B

Fifty-five new variables thus exist; 54 of them are shown in the Figures above, and the last one denotes elevation extracted from a DTM (*Lisboa - Digital Terrain Model*, 2010).
The spatial data enrichment then makes use of the function *pgr_drivingDistance()* for detection of reachable amenities from a record's listing.

```
pgr_drivingDistance('SELECT id, source_osm, target_osm, cost_s, FROM ways',
starting_vertices, distance, directed := false)
```

To be able to execute this command, a listing needs its node origin point, and a destination point. In the case of the rooms, computations generate the closest node to each room as well as the closest nodes to each of the destinations. Figure 3.8 below shows sample tables with generated closest nodes' field using the *ways* table.


**Figure 3.8** - Origin (Room), Destination (PoI) and Edge Network DB Schema

For less costly computation, an additional intermediary table *catchment_area* is created, which describes all potentially reachable nodes from all origin nodes of the rooms (given a desired meters upper threshold).

```
WITH
nodes AS (
SELECT array_agg(node_id) AS nodes from apt_room.node_id)

SELECT from_v as start_node, node as end_node, agg_cost as cost from nodes,
pgr_drivingdistance(
        'SELECT gid as id, source as source, target as target, cost_s as
cost, directed FROM public.ways'::text, nodes, 2000, false)
```

This code, in turn, produced a table that holds all the reachable nodes within 2km. A visualization of the reachable area from a sample point is shown in Figure 3.9 below.


**Figure 3.9** - Road Network Accessibility from a Sample Room

The proxies for each category of PoI then constitute of:

- Minimum distance (in meters) from the room to the closest amenity of a predefined type
- Count per category of PoIs within a distance threshold. Five distinct thresholds are proposed as follows: 2km, 1.5km, 1km, 0.5km, and 0.25km.

Two variables that were specific in their segregation process were the 'Parks', many of which have multiple closest nodes as being 0 meters away from its entrance, and the 'Restaurants' that have a very high quantity and density throughout the whole study area. Therefore, a separate table for parks was constructed that holds all nodes less than 10 meters away from the park of interest, as possible destination nodes from an origin point. Unique proxies are constructed for the restaurants according to their expensiveness (1-4 stars), and additionally, according to their average review rate (starting from 0 as no-reviews, 1 as bad, up to 5 as excellent rate).

Figure 3.10 below, presents an example with an interpolated surface of the minimum distance needed to walk from a sample room to the first registered statue in the dataset.

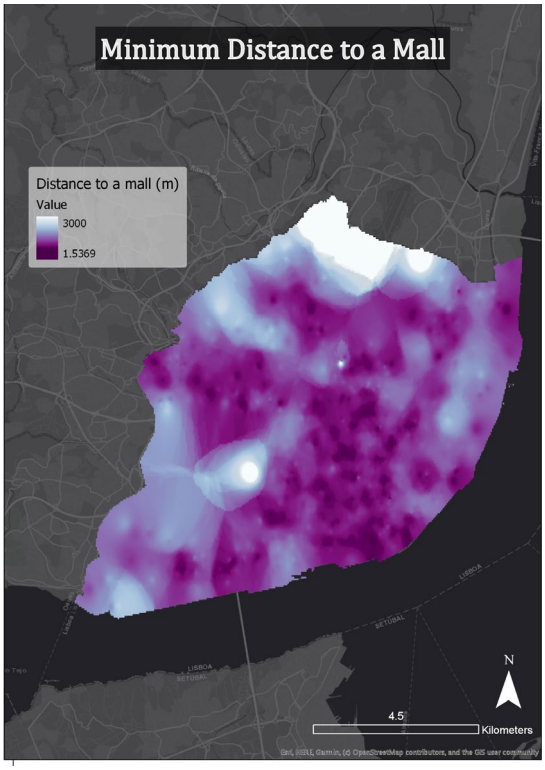


**Figure 3.10** - Example: Minimum Distance to a Mall (in meters)

### 3.5.3 Data Preprocessing

The goal of this section is to provide an overview of the descriptive data statistics, as well as some preprocessing decisions of the integrated dataset and the reasoning behind them.

***Outlier detection*** – A recommended approach when dealing with datasets is to remove outliers. These can represent erroneous data and in this specific case, false entries can be part of overpriced offers if a room's vacancy is far in the future, or additional negotiations might be present between customer-client, which cannot be accounted for. The statistics for the endogenous variable are shown in Table 3.4 below:

| Statistic | Value |
|---|---|
| Mean | 473 |
| Standard Error | 3.72 |
| Median | 420 |
| Mode | 400 |
| Standard Deviation | 173 |
| Range | 900 |
| Minimum | 100 |
| Maximum | 1000 |
| **Count** | **2149** |

**Table 3.4** – Descriptive Statistics of Dependent Variable

The interquartile range finds all values from 870 above as outliers and no outliers below the minimum value, which amounts to a total of 105 outliers outside of the 1.5 interquartile ranges below the first and above the third quartile. Nevertheless, this data

and scenario depicts a real-life Lisbon rent market at a specific temporal instance in September 2019. Hence, it is decided against the removal of these outliers.

***The Cluster and Outlier Analysis*** was run in order to identify clusters and spatial outliers. False Discovery Rate Correction (FDR) parameter is checked, to reduce the critical *p-values* to account for multiple testing and spatial dependence. It is noticeable that there are many clusters adjoining high-priced rooms with low-priced rooms nearby, as well as low-priced rooms with many high-priced rooms nearby. The high-low price clusters tend to exhibit a pattern from central to the northern borders of the study area, whereas the low pricings with many high-priced rooms tend to have a generalized localization near the coast.

| **Cluster Type** | **Count** |
|---|---|
| High-High | 185 |
| High-Low | 64 |
| Low-High | 86 |
| Low-Low | 232 |
| Non-Significant | 542 |
| **Total** | **1109** |

**Table 3.5 –** Counts of Significant Hot-Spot Analysis Clusters

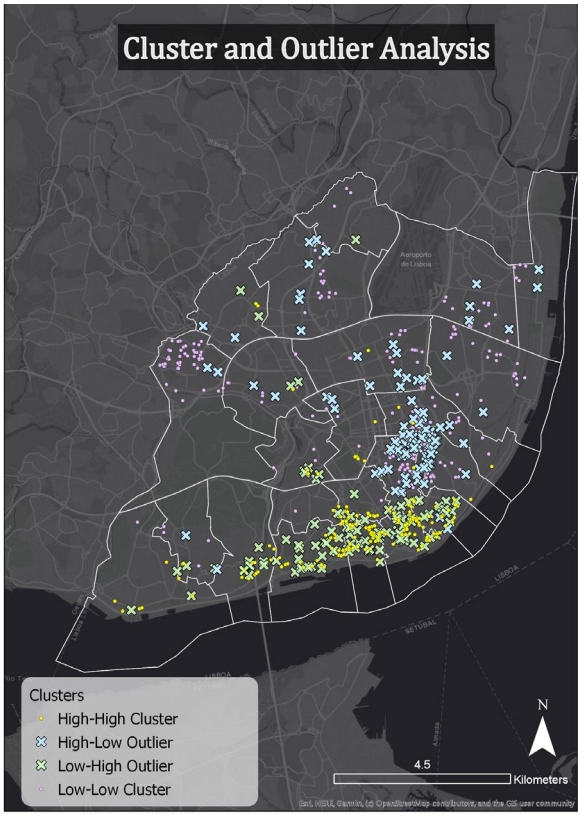


**Figure 3.11** - Anselin Local Moran' I

When it comes to numbers of rooms available per Parish, most of the available rooms are found in parishes where the significant clusters dominate. As such, Avenidas Novas, Areeiro, Alvalade, Santa Maria Maior and Misericórdia hold the largest quantities. Additionally, a high count of bedrooms predominates the central parishes of Lisbon, as

the bedrooms tend to decrease rapidly towards the north borders and decrease marginally towards the southwest and southeast borders.
Following this approach, a few binary variables were put under inquiry and results of a Hot Spot Analysis are subsequently shown in Table 3.6.

| Variable | Parish – Hotspot | Parish – Coldspot |
|---|---|---|
| **Male Landlord** | Lumiar, Santa Clara, Benfica, Alvalade, Beato, Penha de França, Belém | Santo António, Arroios |
| **Female Landlord** | Carnide, Alvalade, Lumiar, Santa Clara, Parque das Nações | São Domingos de Benfica, Areeiro |
| **No Landlord** | Avenidas Novas, Campolide, Arroios, Santo António | Alvalade, Lumiar, Benfica |
| **Has Outdoor Area** | Ajuda, Marvila, Olivais, Estrela, Campo de Ourique | Avenidas Novas |
| **No Outdoor Area** | Alcántara, Santa Maria Maior, Marvila, Misericórdia, São Vicente, Areeiro | Lumiar, Alvalade, Olivais, São Domingos de Benfica |
| **Outdoor Area N/A** | Campolide, Carnide, Lumiar, Alvalade, Avenidas Novas | Olivais, Marvila, Belém |

**Table 3.6** – Hot-Spots and Cold-Spots of Parishes

These hotspots can indirectly invoke knowledge about the pattern – and it can be seen that most rooms that are rented in an apartment with a host appear in all parishes but the central ones. This signifies the area in question is likely to be comprised of student-dominated apartments and it includes the parishes of Alvalade, Areeiro, Avenidas Novas and Santo Antônio. 'Male Landlord' hot-spots are found both in the Parishes in the north and on the river coast, but 'Female Landlords' do not appear as hotspots on any coastal parish. An existing 'Outdoor Area' hot-spots appear frequently in Ajuda and Alcántara, but hot-spots of apartment without it appear frequently in all parishes alongside the coast (in which Ajuda and Alcántara also belong). Rooms belonging in apartment with 'Non-specified Outdoor Area' also show a few hotspots in the most central parishes.

***Identification of Correlated Predictors and Linear Dependencies -*** The dataset contains many variables, and a pairwise check is done to remove variables that might bias the regression. All predictors that are correlated with above 85% are thus removed.
***Data Normalization –*** The dataset largely contains data in very different metrics (meters, quantities and binary variables) – the min-max range is applied so there exists a common scale that would not lead to any distorting differences in the range of values.
***Data Aggregation*** – As the spatial application of hedonic price modelling through the *'spatialreg'* R package required polygons neighbours for calculating the lagged influence of the exogenous variables, different dataset is created for applying HPM. A spatial join is performed using the layer of the Census Blocks at the $4^{th}$ level with the layer of rooms and its features. For each polygon, the averages statistics of the listings belonging in the mapping unit is calculated. The resulting shapefile then contains 987 polygons that denote only census blocks for which there are listings present.

| Count of Dwellings per Block | # of Census Blocks | Total Rooms |
|---|---|---|
| **1** | 503 | 503 |
| **2** | 218 | 436 |
| **3** | 113 | 339 |
| **4** | 63 | 252 |
| **5** | 36 | 180 |
| **6** | 15 | 90 |
| **7** | 12 | 84 |
| **8** | 11 | 88 |
| **9** | 7 | 63 |
| **11** | 3 | 33 |
| **14** | 2 | 28 |
| **10** | 1 | 10 |
| **12** | 1 | 12 |
| **13** | 1 | 13 |
| **18** | 1 | 18 |
| | **987** | **2149** |

**Table 3.7** – Count of Aggregated Dwellings in Blocks

***Training and Testing Split –*** The HPM is to be built to measure how well the data can fit accounting for the error of the residuals through the spatial-specific chosen models, and the whole dataset is used, as the knowledge of the neighbour polygons variables is needed.
The SVR model, on the other hand, needs to be first trained to be able to test the performance on a different unseen set. In order to avoid the overfitting problem (Zang, Berardi, & Reitermanová, 2010), a separate testing dataset is allocated to check whether the algorithm performs well only because it has memorized the patterns on the training set too well.
The process of overfitting can also be waned by K-Fold cross-validation, which can be invoked as a contributing parameter within the 'e1071' library with which the SVR model is built. The choice of *k* is usually 5 or 10, although a formal best-value of K does not exist. In general, as *k* gets larger, the difference in size between the training set and the resampling subsets gets smaller. As this difference decreases, the bias of the technique becomes smaller (Kuhn & Johnson, 2013).

### 3.5.4 Spatial Econometrics and Machine Learning

As a last step in the methodology, different subsets of the clean dataset are to be tested for comparisons of variables' influence, and the assessment will be done with the datasets containing:

- All variables (applied to HPM and SVR)
- Only non-spatial variables (applied to HPM, SVR, SOM)
- Only spatial variables (applied to HPM and SVR)
- All significant variables from HPM (applied to SVR and SOM)

# 4. RESULTS AND DISCUSSION

For comparison of the different regression models applied, Adjusted $R^2$ and Akaike Information Criterion (AICs) metrics are used.

## 4.1 Spatial Dependence

OLS was performed on the dataset, and on that model, the Global Moran I statistic was computed over the residuals. The null hypothesis assumes no spatial dependence, and the results show the following:

- *Standard deviation* = 1.751
- *p-value* = 0.03583

Alternative hypothesis: *GREATER*
Sample Estimates:

| Observed Moran Index | Expected Index | Variance |
|---|---|---|
| **-0.0624** | -0.1157 | 0.0008 |

**Table 4.1** - Local Moran's I

The p-value below 0.01 presents a statistically significant rejection of the initial null hypothesis. The data under analysis leads to spatial residuals, and spatial models are deemed appropriate in order to make the effects of spatially dependent errors disappear.

*The LaGrange Multiplier test* diagnostics additionally gives directions of whether there is a significant improvement of the model fit if spatial models were applied. Results can be seen in Table 4.2 below:

| Model | *p-value* |
|---|---|
| **SAR** | 0.9457 |
| **SDEM** | 0.0042 |
| **SARMA** | 0.0166 |

**Table 4.2** – Models Improvement

The *p-value* suggests either Spatial Durbin Error Model (SDEM) or Spatial Autoregressive Moving Average (SARMA) can bring a significant improvement, although SAR is deemed unfit (insignificant p-value). According to Luc Anselin, the best approach is to continue with the one with the lowest *p-value,* which in this case was the SDEM (Anselin, Bera, Florax, & Yoon, 1996). Another model introduced is SLX, in order to be able to compare models either with local (SLX) or global (SDEM) spillovers in the disturbances.

## 4.2 Spatial Lagged X

When applying an HPM regression, the contributing influence of individual variables is computed. The number of stars assigned next to a variable indicates its significance. Counts of how many of each significance type is found for the model that includes all available variables is shown in Table 4.3.

| Code | Probability (>t\|) | Count of Variables |
|---|---|---|
| *** | >0.001 | 3 |
| ** | >0.01 | 6 |
| * | >0.05 | 38 |
| X | >0.1 | 41 |

**Table 4.3 –** Significance Percentage

The results have a residual standard error of 114.7€ which amounts to a MAPE of 15% on predicting a room price within census blocks that has aggregate descriptive statistics of maximum price of 950€ and a minimum of 200€ among them.
The variables that are denoted as important excluding the X intercept amount to 97 variables. The most significant *direct effect* ones (above 95% significance) are present in Table 4.4 below:

| Variable | Estimate | SE | t-value | Pr(>\|t\|) | Sig. |
|---|---|---|---|---|---|
| **C. of Bedrooms** | -199.908 | 54.823 | -3.646 | 0.00001 | *** |
| **Male Landlord** | -135.44 | 20.312 | -6.668 | 0.00031 | *** |
| **MD to Club** | 20.344 | 7.799 | -2.609 | 0.00958 | ** |
| **MD to Viewpoint** | 89.951 | 33.764 | 2.664 | 0.00817 | ** |
| **MD to Theater** | 51.507 | 18.864 | 2.73 | 0.00672 | ** |
| **Lag CentralHeat_F** | -271.162 | 97.123 | -2792 | 0.00560 | ** |
| **Lag C. of Bedrooms** | -126.196 | 46.517 | -2.713 | 0.00708 | ** |
| **Lag Elevator N/A** | 350.502 | 159.086 | 2.203 | 0.00283 | ** |

**Table 4.4 –** Significant Variables – Model using All Variables

From these variables, it is inferred that the existence of a male landlord living inside the apartment makes the price of the room be a subject to negative influence. This may result from the fact that the rooms are mostly used between 6 months and 1 year, and as such are perfect for either students or people that have recently started working. In this sense, foreigners that want to experience Lisbon may tend to either gravitate towards an apartment that there is no landlord living in, to experience more freedom during their stay – or they statistically value a female person living or taking care of an apartment much more than a male person. Still, on average rooms without landlords are priced higher. This significance can be reinforced by studies that confirm that young adulthood has become a distinct new life phase (Berthoud, Gershuny, & British Household Panel Survey., 2000). During this phase, people value and choose to live in peer-shared households and quasi-communes with people likely to be their own age, rather than living with a landlord (Heath, 2004).

However, similar studies that measure a host's gender significance in the prices of short-term stays applied on an Airbnb dataset corroborate no impact on it, as opposed to detection of significant racial discrimination (Kakar, Voelz, & Wu, 2017).
The number of bedrooms variable impact in decline of price comes also as generally, apartments tend to have one kitchen to use and thus, the spaces would be subject to shared use among people that are not necessarily familiar with each other.
From the spatial ones, the most significant ones denote the minimum distance from a viewpoint as a positive coefficient. This means apartments closer to a viewpoint would be lower-priced. This might be a result as these viewpoints are often off-grid of where most daily activities happen, and living close to a viewpoint can be often connected to an upward slope when looking at the room as a destination. Counts of universities within 500 meters and trams within 1km also all yield positively influencing variables (Troy & Grove, 2008). The transportations' effects are not unique for mid-term rentals as it has been proven they are found in residential property value indices within the same study area, and further they can be used to forecast possible upward price changes if an area is a subject to new transportation investments (Martínez and Viegas, 2009).
A lag of the dummy variable signifying the existence of an elevator within the buildings of the apartments listed within a census area close to the one being rented would seem farfetched to have an influence in predicting price – but it might be that the first rule of nearby things being more similar than far ones comes to play here – and buildings with no elevators might behave in a clustered way. However, the influencing steep coefficient of 350 raises a questionable impact.

From the variables with less significance, the following counts are present:

| Type of variable | Significance "*" | Significance "X" |
|---|---|---|
| **Spatial** | 14 | 13 |
| **Spatial Lagged** | 8 | 19 |
| **Census** | 5 | 2 |
| **Census Lagged** | 3 | 3 |
| **Non-Spatial** | 2 | 1 |
| **Non-Spatial Lagged** | 3 | 2 |
| **Total** | **35** | **39** |

**Table 4.5 -** Count of Significant Variables in SLX

Adding each of the variables' effect combined with its lagged counterpart's indirect effect yields the total significance of variable. In this way, a statistically insignificant variable with its statistically insignificant lagged counterpart might yield a statistically significant variable for the model. The differences between the two models include a diminishing amount of variables being significant, but spatial ones that appear to be important in the model include a university within 0.5km, a tourist tour (counts within both 1km and 0.5km), and a non-spatial positive one denoting that the host has disclosed whether the apartment has central heating. This, however, might lead to different knowledge – apartments with more missing data might be less desirable to rent.
When subsets of the variables are applied the results differ marginally, but the AIC's metric does not improve. It can be noted that the variables significant with confidence above 95% stay the same in both the non-spatial model and the spatially enriched one.

The spatial model yields a similar error with a residual error of 116 euros on 300 DoF, and the non-spatial model contains a residual standard error of 124 euros on 922 DoF.

| Metric | All Var | Spatial Var | Non-Spatial Var |
|---|---|---|---|
| **RSE (€)** | 116 | 139 | 124 |
| **Multiple R-Squared** | 0.845 | 0.731 | 0.456 |
| **Adjusted R-squared** | 0.490 | 0.269 | 0.418 |
| **DoF** | 300 | 362 | 922 |
| **p-value** | *** | *** | *** |

**Table 4.6 –** Model Subsets Statistics

Cross-referencing the models according to their goodness of fit and accuracy measurement statistics, the residual error is found to be smallest on the model with all variables included. Although the adjusted R-Squared has a parameter that punishes overuse of variables, the model containing proxies to amenities still manages to outperform the non-spatial model, capturing more of the variability. The significant variables are shown in Figure 4.1 below:

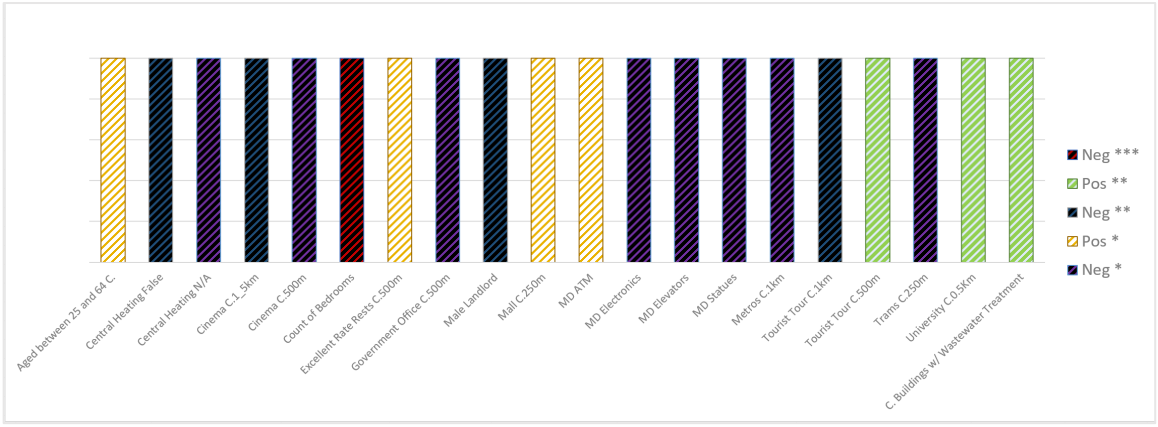


**Figure 4.1** - Significant Variables from Model using All Variables

In order to understand the underlying structure of the most significant variable computed, the variable “Count of Bedrooms” is further mapped in Figure 4.2

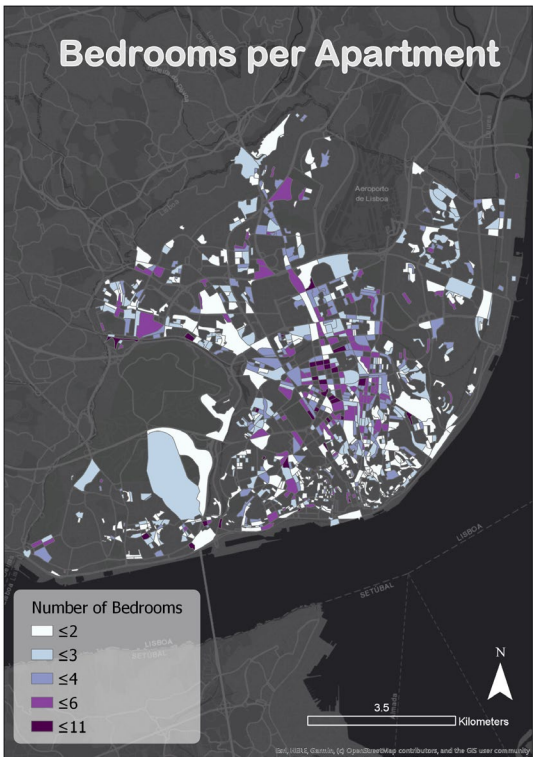


**Figure 4.2** - Count of Bedrooms in Apartment wherein the Room Resides

Evaluating the spatial pattern of the room prices in Lisbon, the highest prices tend to be concentrated in a pattern from north-east towards mid-south west with an all-around high price concentration near the river coast. Looking at the Figure above, this does not seem to be the case when it comes to the number of bedrooms which exhibits a central pattern with highest counts centered in the neighborhood of Arroios, and a dwindling number following towards the south reaching the lowest counts near the riverside.
Another significant variable, the dummy "Central Heating - True" exhibits a seemingly random spatial pattern with multiple direct neighboring census blocks holding opposite values. This, however, is normal to influence a client's decision for a place as people in mid-term rents are not a target audience for investment in a place where they are not expected to live for a long period.

When analyzing the spatial variables influence, "Count of Government Offices" within 500 meters, and "Malls" within 250 meters are some of the variables found. The first one shows a high concentration within census blocks with direct proximity to Campo Mártires da Pátria (parish of Arroios), and a slightly lower concentration near the census blocks bordering "Praça do Comerçio" (Santa Maria Maior). Malls within 250 meters is a variable whose very small distance threshold makes the value of the variable fairly small throughout the study data, but the listings that are exposed to a small amount of this category seem to boost the target variable.

## 4.3 Spatial Durbin Error Model

Most models trying to fit a linear equation onto the data will have a high likelihood to find a similar range of variables to influence a target variable at the highest confidence level (>99%). However, as there are many variables in the dataset, the models tend to calculate a different range of variables having a less significant influence. The table below shows the most significant variables (above 0.01 level) as computed by SDEM.

| Variable | Type | Coefficient |
|---|---|---|
| **Number of Bedrooms** | Non-Spatial | -177.587 |
| **Cultural C.500m** | Spatial | -140.265 |
| **MD Theater** | Spatial | 63.903 |
| **Electronics C.1km** | Spatial | 113.234 |
| **Landlord – Male** | Non-Spatial | -127.266 |
| **Minimum Stay** | Non-Spatial | 20.856 |
| **Central Heating - False** | Non-Spatial | -90.369 |
| **MD Insurance Agency** | Spatial | -23.345 |
| **Mall C.250m** | Spatial | 68.298 |
| **MD Travel Agency** | Spatial | 16.445 |
| **MD Viewpoint** | Spatial | 84.069 |
| **Veterinary C.1km** | Spatial | -81.822 |
| **Insurance Agency C.250m** | Spatial | 70.937 |
| **Theater C.250m** | Spatial | 145.481 |
| **Theater C.500m** | Spatial | 102.380 |
| **MD Beauty Salon** | Spatial | 13.229 |
| **Car Dealer C.250m** | Spatial | -69.283 |
| **MD Mall** | Spatial | 25.427 |
| **Taxis C.500m** | Spatial | -83.621 |
| **City Hall C.500m** | Spatial | 192.767 |
| **Statue C.250m** | Spatial | 58.900 |
| **Outdoor Area - True** | Non-Spatial | 128.305 |

**Table 4.7 –** Top 20 Significant Variables – SEM Model with All Variables

The non-spatial variables are not dominating the model in significance here, but they still have the largest coefficients adjoining to them. The count of significant variables seems to rise as the model finds 118 significant variables (intercept inclusive), although variables like minimum distance to a salon have a very small contributing coefficient compared to most non-spatial ones.

It can be seen that distance metrics can sometimes be found to contribute negatively, as it can be seen with counts of "Veterinary Stores" within 1km, and this suggests the model might be learning from unnecessary noise to fit the data perfectly.

Another thing noticeable from both models is that there is almost no significance to be found in any variables above 1000m of distance with very few exceptions (both minimum and counts). Concerning the multiple restaurant categories, it gives findings of the counts of the two least expensive categories of restaurants within 500m produce significant variables with negative coefficients. This might also give knowledge about the neighborhood wherein high concentrations of cheap amenities of this kind signifies areas that consist of room rents in mainly older apartments, or an altogether undesirable neighborhood.

The comparison of the SDEM applied to the three data subsets is then shown in Table 4.8

| Metrics | All | Spatial | Non-Spatial |
|---|---|---|---|
| **z-value** | -6.8807 | 0.399 | 4.3151 |
| **Log Likelihood** | -5891 | -6090 | -6159 |
| **Total Residual (€)** | 8781 | 13398 | 15291 |

**Table 4.8** – Performance Evaluation in SDEM

The findings suggest the best accuracy to be present in the model using all the variables. To further understand how well these models capture the variability of the data and how they compare, the *MAE* and *R Squared* metrics are both calculated and shown in Table 4.9

| Model | Variables | $R^2$ | MAE (€) |
|---|---|---|---|
| **SLX** | All | 0.843 | 50.21 |
| | Only spatial | 0.714 | 67.34 |
| | Non-spatial | 0.451 | 94.14 |
| **SDEM** | All | 0.921 | 35.69 |
| | Spatial | 0.760 | 61.97 |
| | Non-spatial | 0.459 | 93.64 |

**Table 4.9** – MAE and R Squared – Performance Evaluation

The best performing model according to the provided accuracy metrics is SDEM with all variables included.

Additionally, *Studentized Breusch-Pagan* test for heteroscedasticity, as well as *sensitivity analyses* are performed. A non-significant| *p-value* of 0.509 rejected the null hypothesis of the presence of heteroscedasticity. As for the sensitivity analyses, variants excluding some of the (less significant) variables are tested, and a high fluctuations in the coefficients of the significant variables proves the model to not be reasonably robust. Nevertheless, the MAE differences were marginal and the chosen range of the most significant variables stayed consistent.

For a visual comparison, side by side maps of SDEM predictions and one with the original prices are shown in Figure 4.3. The symbology divides the census blocks into five classes with quantile distribution breaks preserving the frequency of the price in each range according to the original listings.

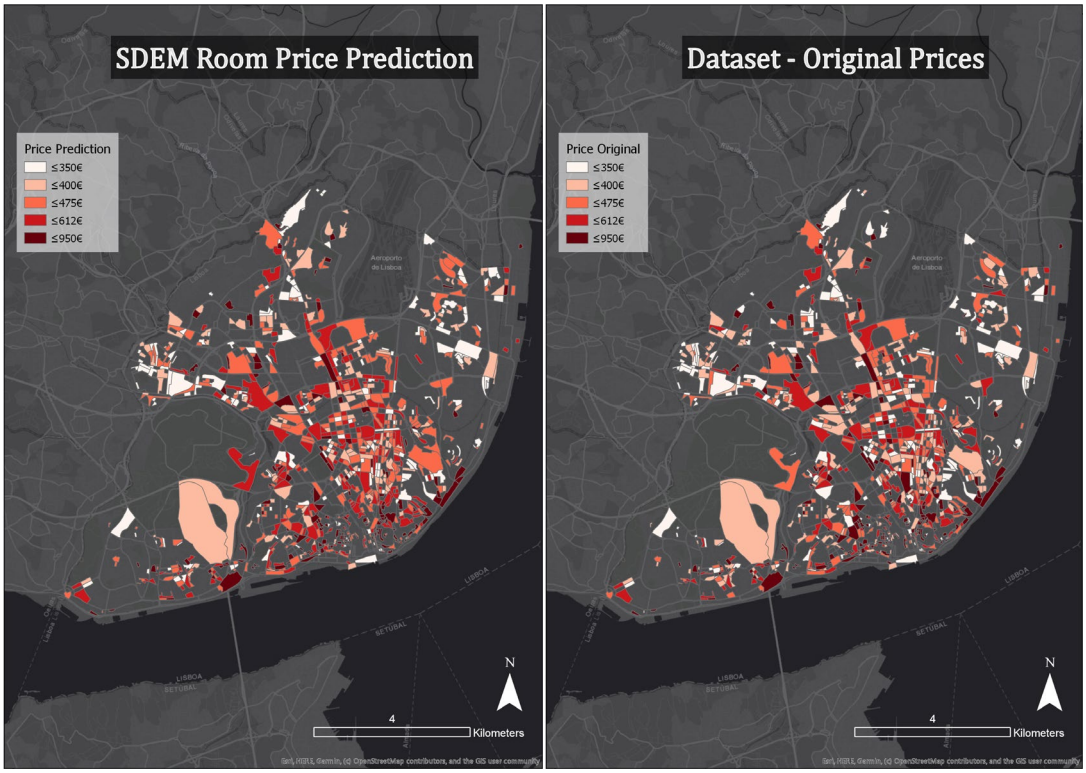


**Figure 4.3** - Side by Side – Original Price versus SDEM Prediction

The figures demonstrate that the faults in the model lie in predicting the average prices of the census blocks adjacent to clusters of high prices. In the central area of the study, most rooms exhibit similar high counts of amenities nearby and a high probability of low minimum distance to the closest amenity of any type. Hence, they share a much closer likeness between each other as opposed to ones in the outskirts where there seem to exist localized centers reachable only from certain locations. This, however, might also be a matter of differences in the area of the polygons under question since the polygons with the smallest area dominate the south coastal area of the study region.

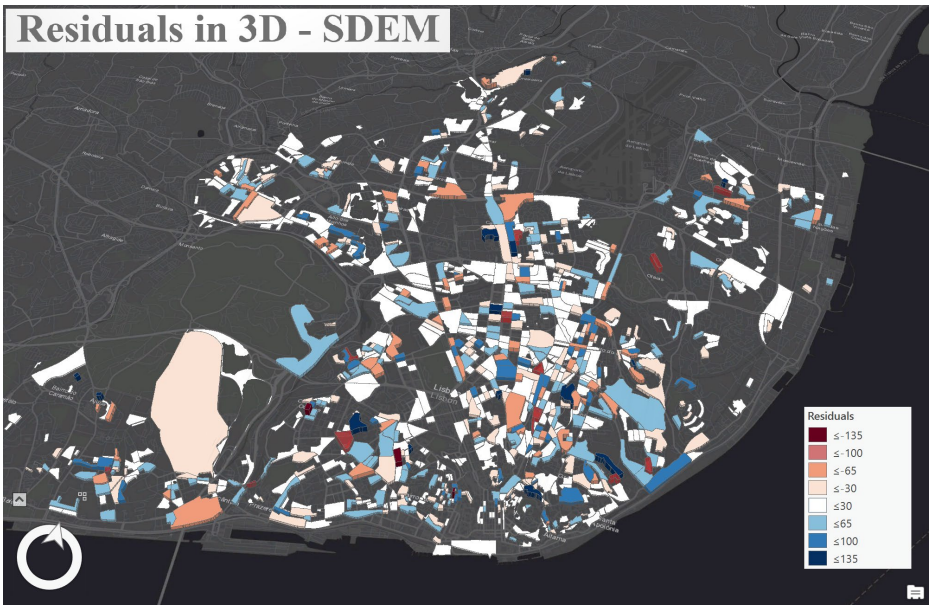


**Figure 4.4 –** Residuals in SDEM

Moreover, from mapping the residuals in Figure 4.4, one can notice the highest error prediction lies in a single neighbor-less census block near Olaias. This suggests that missing data from a neighbor's lagged variables might give a better prediction for this region. Additionally, blue predominating color in the map (positive missing residual) suggests the model tends to underestimate the prices (or it cannot make a distinction of what makes a higher-valued room price). The following statistics are derived from the model as seen in Table 4.10

| | |
|---|---|
| **Price within limits of SD (€)** | 688 |
| **Overestimated price beyond limits of SD (€)** | 156 |
| **Underestimated price beyond limits of SD (€)** | 143 |
| **Total (Census Blocks)** | **987** |

**Table 4.10** – Statistics of model predictions

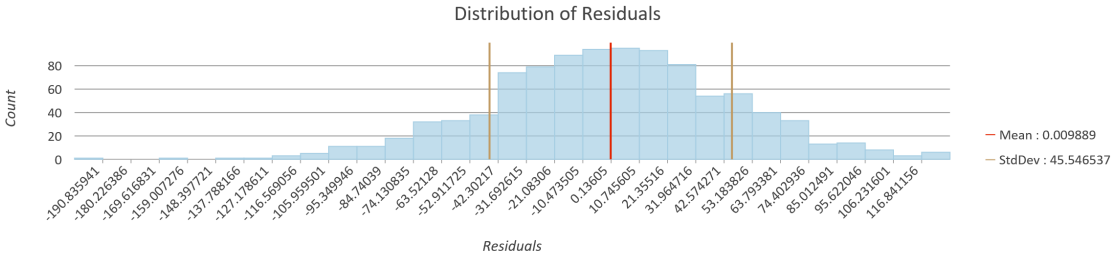


**Figure 4.5 -** Residuals Distribution

The scatterplot produced between the variables price and the residuals show a rather uniform variance throughout its range.

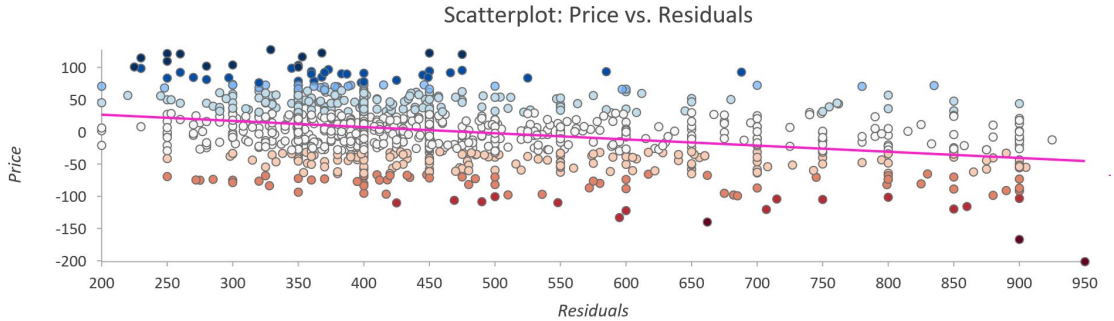


**Figure 4.6 -** Price vs. Residuals Scatterplot

The scatterplot in Figure 4.6 shows a clear pattern of overestimation in the lowest priced rooms, as opposed to an underestimation of the biggest priced rooms. One fault of this is that also, data points exhibiting the right extreme limits of the range are scarce.
Another problem could be a dynamic pricing behavior of the listing, which in turn might have the host put expensive prices for their listing if it is not available for rent in the

foreseeable future, and later revise the room to a less expensive price when there is less time to find a suitable tenant for the proposed location. This is a strategy that is commonly used in short-term rentals by more experienced hosts that manage more listings as it helps in maximizing profits (Gibbs, Guttentag, Gretzel, Yao, & Morton, 2018).
These dynamics could be overcome if there is a multi-temporal registering on price modifications over the study area taken throughout different months within the year.

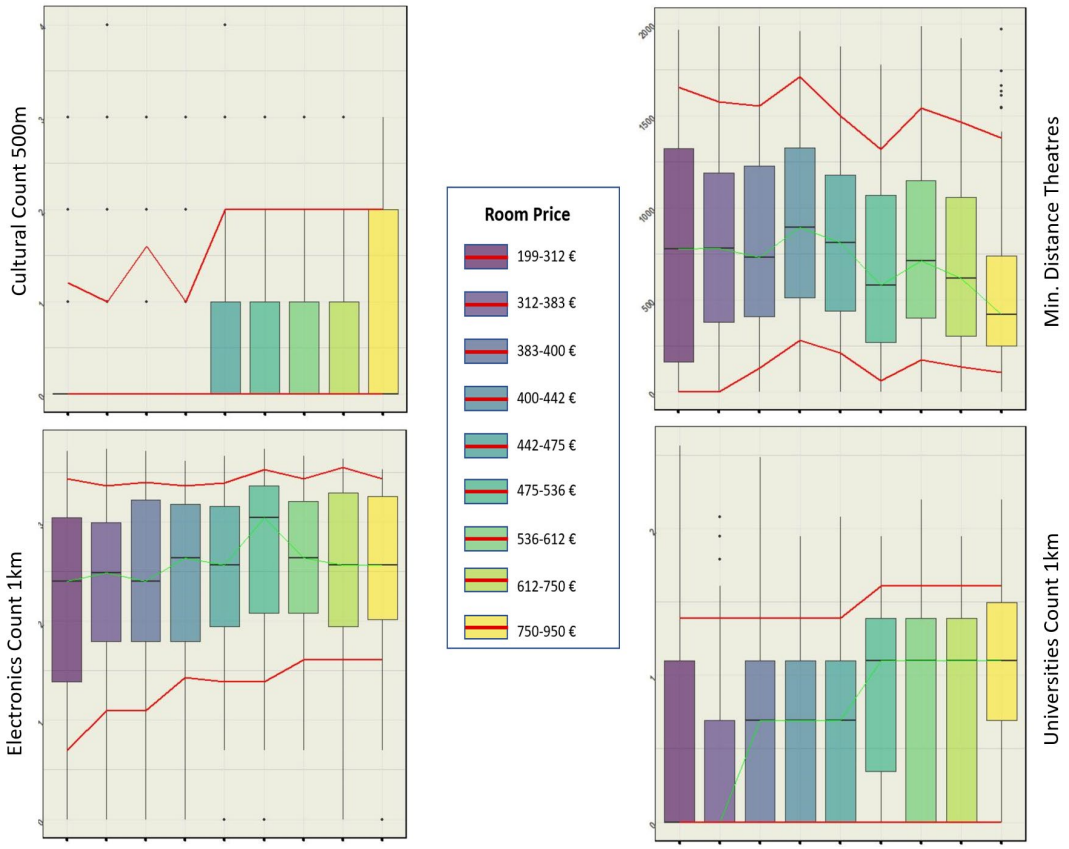


**Figure 4.7** - Boxplots of 4 Highly Significant Proxies - SDEM

Additionally, boxplots for significant variables are shown in Figure 4.7 above. These boxplots represent aggregations of ranges of the price on the X-axis , and a significant variable of choice on the Y-axis, with the red line denoting the $10^{th}$ and $90^{th}$ percentile of its values within the given price range.
From the boxplots, it can be noted that count of universities within 1km have a slight tendency to correlate to a raise in a listing's price. As is, this is a desirable geographic location for a student. As late findings suggest the housing market in Lisbon is exhibiting an accommodation shortage and housing crisis (Cocola-Gant and Gago, 2018), the hosts might feel the higher price will not impede them from finding a tenant. This finding follows and reaffirms studentification literature claims that stress proximity to campuses as a main driving force in choice processes of student clusters within a city (Sage, Smith, & Hubbard, 2012).

From the other variables, it is found that cheaper rooms have very few, if at all, cultural facilities around them. The minimum distance to a theater also records a slight downward spiral with the listings with highest prices residing closest to this amenity. Another variable that follows a similar structure is the count of electronic stores within 1km. Nevertheless, with 427 electronics stores throughout the study area, it is likely that most of them have clustered hotspots in central parishes where this influence stems from. The count of electronic stores within 1 km of the listing distributed among the parishes of the study area are shown on a choropleth map in Figure 4.8 below and that assumption can be empirically confirmed.

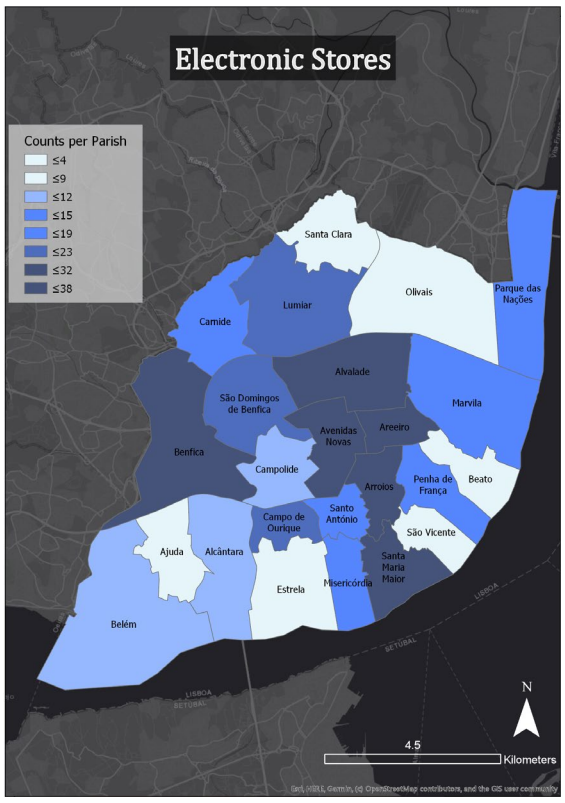


**Figure 4.8 –** Electronic Stores per Parish

To continue the analysis, a dataset containing only the significant variables as denoted by SDEM is extracted and further applied to creating an SVR model.

## 4.4 Support Vector Regression

Before applying SVR, the datasets are divided into a training (85%) and testing dataset (15%) with randomized rows taken as training samples. The outputs models' errors range between 80 and 150, although a MAE that differs a lot between a training and testing set points to an over fitted model, so results need to be analysed carefully and parameter choices subsequently tuned.

In Table 4.11, a grid search of the hyperparameter optimization is shown. The minimal value of the accuracy metric of RMSE as given by *e1071* finds the model that best fits the data.

| Parameter | Search Grid Values | Count |
|---|---|---|
| **C** | 0.001,0.01,0.1,1,5,10,50,250,500 | 9 |
| **ε** | 0.001, 0.01, 0.1, 0.3, 0.5, 1.5 | 6 |
| **$\gamma$** | .0001,.001,.01,.1,1, 2 | 6 |
| | **Models Tested** | 324 |

**Table 4.11 –** Hyperparameter Optimization

| Var Inclusion | C | ε | $\gamma$ | SV | SV% | RMSE (€) |
|---|---|---|---|---|---|---|
| **RBF All (M1)** | 10 | 0.3 | 0.001 | 1109 | 65% | 133 |
| **RBF Spatial (M2)** | 1 | 0.5 | 0.01 | 909 | 52% | 147 |
| **RBF Non-Spa (M3)** | 5 | 0.3 | 0.01ssd | 1079 | 62% | 130 |
| **RBF Sign (M4)** | 5 | 0.5 | 0.1 | 888 | 67% | 143 |
| **Lin All (M5)** | 0.1 | 0.5 | / | 907 | 52% | 155 |
| **Lin Spatial (M6)** | 0.001 | 0.5 | / | 906 | 52% | 157 |
| **Lin Non-Spa (M7)** | 0.1 | 0.5 | / | 905 | 52% | 146 |
| **Lin Sign (M8)** | 0.01 | 0.5 | / | 888 | 52% | 141 |

**Table 4.12 –** Best Performing Models

From the proposed hyperparameter optimization used on the SVR models (Table 4.11) and the resulting RMSE and SVs (Table 4.12), it can be noted that all models require a substantial amount of records to be used (upward of 50%) to define the margins of the hyperplanes. This is a contributing proof there is not much regularity in the dataset. Moreover, a small sample of 2149 data points over 312 variables might be deemed as too little training data to discern a meaningful hyperplane.

The best performing model still shows the non-spatial inclusive as the one that best fits the data. Nevertheless, the usage of spatial variables in addition seems to produce a model with a comparable behavior in both the counts of support vectors and RMSE. Moreover, the linear kernel performs comparatively worse for all the data subset choices.

In Table 4.13, the MAE metric of the best models is shown (see Table 4.12 for model names). In model M1, it can be noted that the train and test performance differ substantially. Additional grid search over the same dataset with a lower value of the parameter *C* (0.1) was done, and a new appropriate model is calculated. In Table 4.13, this model is denoted as M9:

| Model | MAE Train (€) | MAE – Test (€) | $R^2$ – Train | $R^2$ - Test |
|---|---|---|---|---|
| **M1** | 60.78 | 102.93 | 0.80 | 0.44 |
| **M2** | 84.95 | 114.46 | 0.66 | 0.32 |
| **M3** | 83.96 | 98.96 | 0.55 | 0.37 |
| **M4** | 104.78 | 110.15 | 0.42 | 0.26 |
| **M9** | 115.08 | 114.35 | 0.33 | 0.23 |

**Table 4.13 –** MAE on Best RBF Kernel Models

The new training and testing errors in model M9 are much more comparable although they produce a much worse fit. The variability of the data is not captured well, nor in the training, nor in the testing set.
Model M3 using only the non-spatial variables is still the best predictor and a difference of 16 euros between training and testing data provides proof that this model does not overfit the data.
The largest variables' contribution towards models M3 and M4 are shown in Table 4.14 below.

| M3 Variable Importance | Contribution (%) |
|---|---|
| **Number of bedrooms** | 18.9 |
| **Number of bathrooms** | 17.3 |
| **No Preferred Gender** | 5.50 |
| **Double Bed - Dummy** | 5.20 |
| **Not specified - Elevator** | 4.90 |
| **Central Heating – True** | 4.90 |
| **Outdoor Area – True** | 4.40 |
| **Single Bed- Dummy** | 4.30 |
| **Male Landlord** | 3.50 |
| **Female Landlord** | 2.40 |

**Table 4.14 –** Variables Contributions; Model M3 and M4

| M4 – Variable Importance | Contribution (%) |
|---|---|
| **Number of bedrooms** | 6.45 |
| **Bedroom - Single** | 3.26 |
| **Park Counts – 250m** | 2.34 |
| **Room Squared – Not Specified** | 2.17 |
| **Elevation** | 1.95 |
| **MD - Lodging** | 1.94 |
| **Tourist Tour Counts – 500m** | 1.91 |
| **Viewpoint Counts – 500m** | 1.80 |
| **MD – Real Estate** | 1.79 |
| **Supermarket – 250m** | 1.74 |

Results suggest that Model M3 gives little influence to unique variables, although it uses a large range of them. Nevertheless, when the variables are separated into spatial, and non-spatial, it is computed that the models gains contributions constitute only 25% from the non-spatial related variables, and 75% importance from variables that are pertaining to the surroundings of the area. As the spatial variables are only supposed to better the prediction or close the error gap enough to justify their use, it seems this model has learnt to separate the data using them as the primary exogenous variables.

Extraction of the *Lagrange Multiplier*'s coefficients of the SV of the models did not benefit the study as counts of 852 and 601 SV's coefficients in model M3 and M4 accordingly are given the maximum multiplier's value as the force required to enforce the constraints of the models.

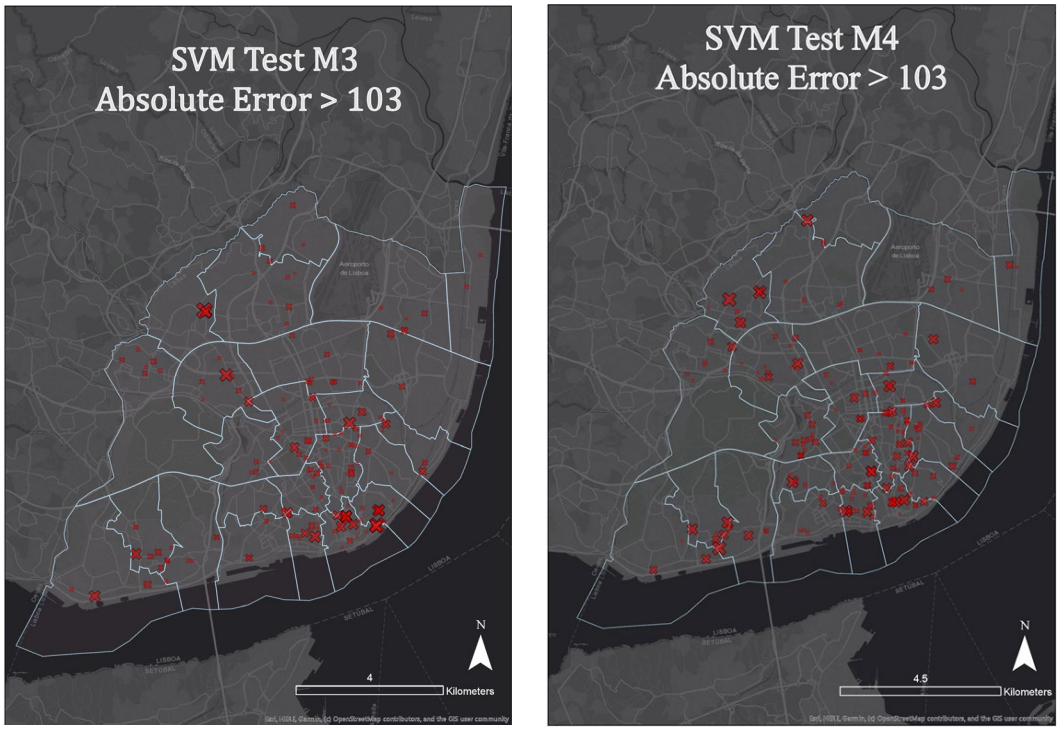


**Figure 4.9 –** SVM Regression – Residual Errors above SD Threshold

From mapping the high residuals in Figure 4.9, the two models demonstrate a different error behavior, wherein model M3 in central Lisbon and follow a high residual pattern towards the south-west borders of the area, whereas M4 seems to have its errors spread throughout most parishes. Moreover, the model performs quite well in certain freguesias in the north-eastern regions of the case study wherein no data points with large errors are found. However, the central areas experience versatile prices for which a margin cannot be found – the spatial variables are thus not helpful, but they are creating closer clusters in the most central parishes wherein large counts to many amenities are situated. To have a better overview of the faults of each model, the two models are compared through scatterplots in Figure 4.10 and Figure 4.11 below:


**Figure 4.10** – Model Comparison in Residual Error

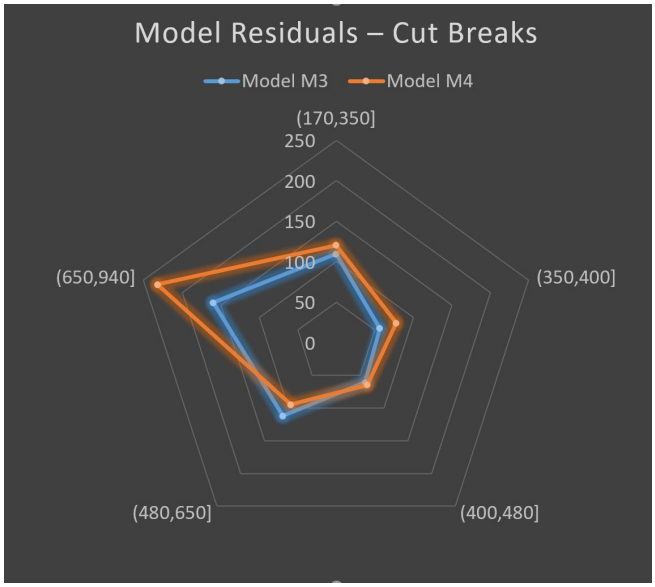


**Figure 4.11** – Comparison of Errors through Breaks' Aggregations

Model M3 has a more stabilized error behavior throughout the ranges when compared to M4, according to the brakes given. Model M4, on the other hand, has a very large absolute difference between the predicted price and the actual price of the room for rent in the breaks denoting the higher ranges. However, both models fail to find reasoning for prices rising towards the upper range in general (and this is where the largest residuals are produced). The maximum absolute errors are registered in model M3 on the positive scale (overestimation of 395€) on an average priced room (450€), and M1 with -426€ underestimation on a high priced listing of 900€.

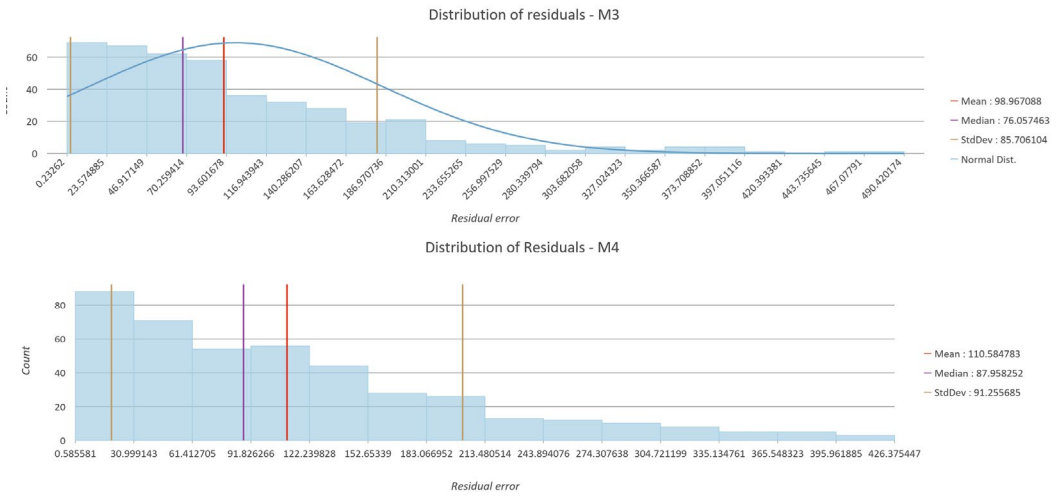


**Figure 4.12 –** Residual Error Comparison

## 4.5 Natural Data Clusters using SOM and GeoSOM

Both SOM and GeoSOM clustering of the datasets were applied on Models M3 and M4, to check whether the structure of the variables would make similar-priced rooms join together, or it would just force the order of natural geographic coordinates to overrule the non-spatial variables impact. Different sizes of neurons are verified as follows: 10x1, 15x1 and15x10. Manual separation of clusters are elected visually using the GeoSOM software, with anticipations of clearer separation of the data whilst also maintaining a small topological error.

From the descriptive statistics of each cluster, understandings of a variable's aggregate behavior is to be outlined. The SOM algorithm clusters of M4 were chosen with a topological error at a value of 0.01 and 14 number of clusters. The clusters show a very sparse behavior throughout the study area. Even though Cluster 3 neurons on the U-Matrix showed a similarity between these records, very little spatial pattern was detected when dividing them between the parishes.

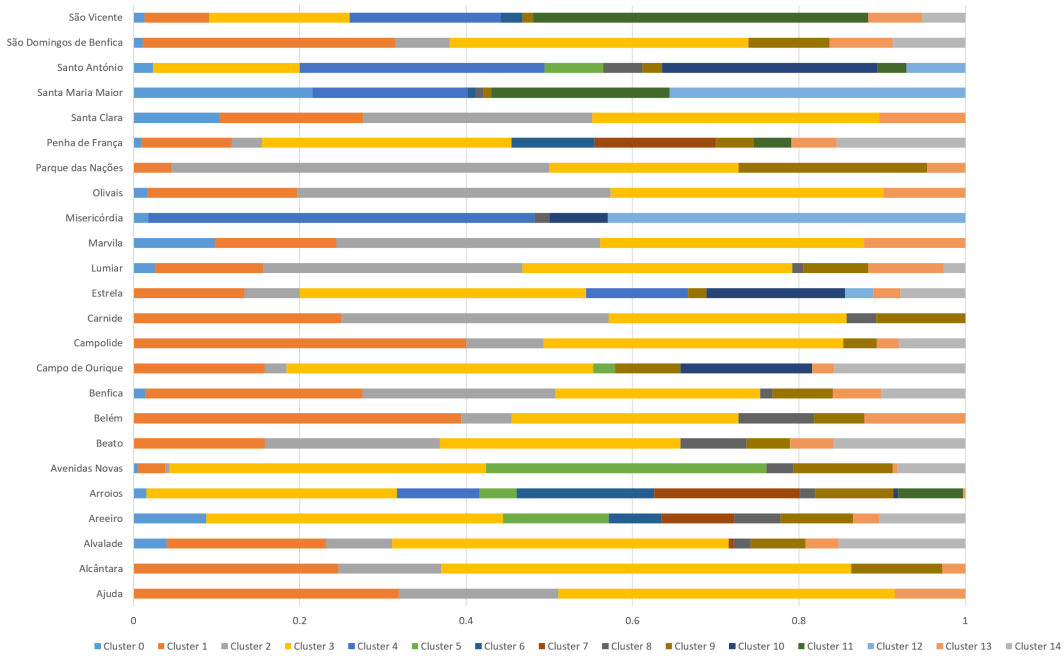


**Figure 4.13** – SOM Clusters – Model M4

When it comes to the price variable, the chosen clusters' minimum prices ranged from a minimum average of 398€ in Cluster 2, up to a maximum average price of 607€ in Cluster 10. Both of these clusters are found partially in the majority of the parishes.

When GeoSOM is used, more natural order of geographic clusters appear, as imposed by the limiting radius parameter. In Figure 4.14 below, the binary variables from the GeoSOM cluster of M3 are presented on parallel coordinate plots.

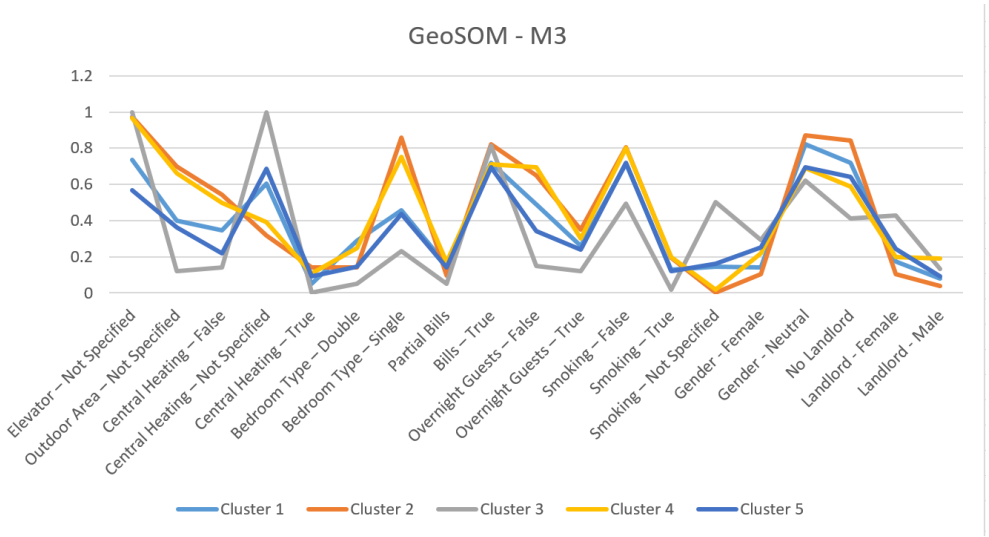


**Figure 4.14 –** GeoSOM – M3 Binary Variables

| Variable/Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **C. of Bedrooms** | 3.08 | 4.55 | 2.78 | 3.36 | 3.07 |
| **C. of Bathrooms** | 1.27 | 1.58 | 1.21 | 1.26 | 1.29 |
| **Minimum Stay** | 43.01 | 75.48 | 21.55 | 75.45 | 43.59 |
| **Average Price** | 506 | 454 | 403 | 434 | 420 |
| **Minimum Price** | 110 | 200 | 200 | 225 | 100 |
| **Maximum Price** | 1000 | 940 | 890 | 900 | 920 |

**Table 4.15 –** Cluster Statistics of Numeric Variables in GeoSOM M3

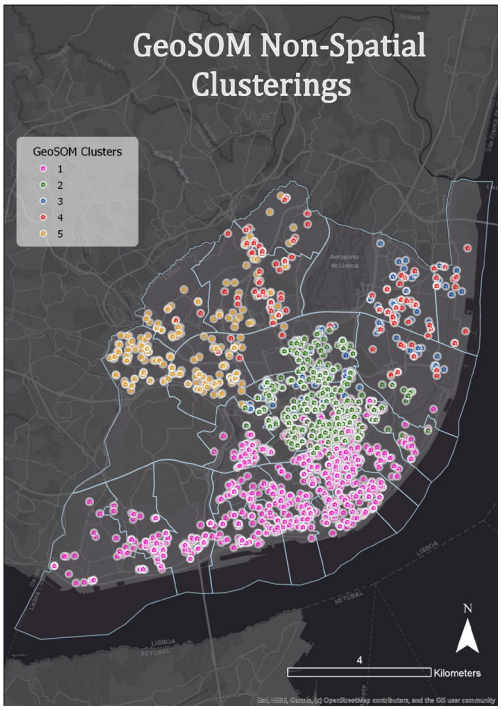


**Figure 4.15 –** GeoSOM – Model M3

Listings in Cluster 3 are relatively indiscernible with those from Cluster 4 in the north-east boundaries as shown in Figure 4.15, and they also blend with Cluster 2 in the central area. A huge discrepancy can be noted in the 'Minimum-Stay' variable, with a substantially low value in Cluster 3 denoting that there is no minimum-stay rule for these listings. This cluster also describes a high chance of a landlord living in the vicinity of the apartment, correlates with a smaller number of bedrooms, and missing information about the existence of central heating. Although geographically these rooms tie in together with the two other clusters, this cluster denotes an area where it is more likely the owner is subletting a room of their own living space to increase their living income. Unsurprisingly, the room price here is the lowest, although the largest difference between the minimum averaged room price Cluster and the maximum is only 103 euros (12% of the price range). Cluster 2 describes rooms which are not gender-assigned, and contain the largest number of bedrooms within one apartment. Cluster 1 acquires the coastal parishes geographically, but non-spatial variable wise the average of the variables are more comparable to Cluster 2 with a slightly smaller percentage of listings without landlords, and a smaller average of rooms within the apartment. Cluster 4 seems to have a high probability of double beds within the room, and about an equal probability of a landlord possibly living in the listing. Cluster 5 is the one with the second-lowest price average and a large lack of information about the type of bed. This variable can presumably be corrected by image processing/analysis of existing images of the room, to detect the bed size – an area that could be further explored.

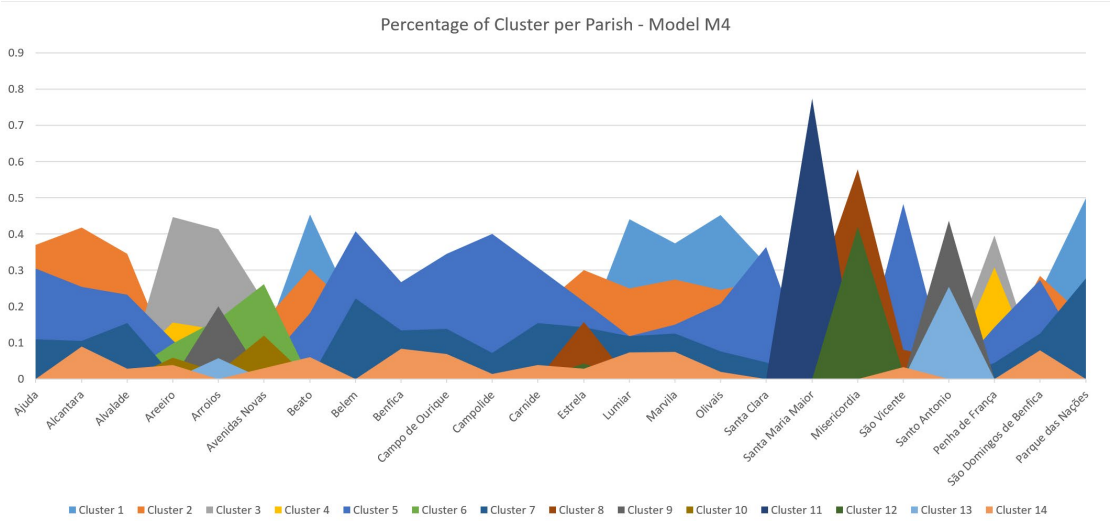


**Figure 4.16** – SOM, Model M4

At the end, a SOM division using the SDEM significant variables is made, with a choice of 14 Clusters for easier comparison with Model M3. Although there is a large percentage of some clusters being constituted of one parish it is still hard to perceive a great division, as many clusters have a heterogeneity that spreads throughout the whole study area (e.g Cluster 2 and Cluster 7). The Misericórdia area is the purest parish described by only 2 Clusters (8 and 12), followed by Santa Maria Maior with 3 Clusters (8 and 11 with multiple listings and 1 listing in Cluster 5). Cluster 11 and 12 describe

the most expensive listings from two coastal parishes: Santa Maria Maior and Misericórdia, with averages of 685€ and 721€ respectively. Cluster 8 contains cheaper listings from the two parishes mixed together. Cluster 9 describes the cheapest listings (averaged), completely exclusive to listings from the parishes of Arroios and Santo Antonio. These parishes by themselves, however are not cheap, as listings with a wide variety of prices from the two parishes can be found in 7 and 8 clusters of the proposed 14, accordingly.

## 4.6 Overview of Given Analysis

The dataset was analyzed using spatial econometrics, and two types of HPM models were tested (SDEM and SLX) as judged by the most significant p-value from the Lagrange Multiplier test diagnostics. The spatial specifications of the models made use of data aggregation with 987 resulting polygons describing averages prices within the blocks of the study area that contain at least one listing.

The most significant variable in both SDEM and SLX, and further in SVR was the count of bedrooms – and it bears a negative coefficient. The dummy variable 'Male Landlord' also leveled a confidence of above 99% bearing a negative coefficient. The SDEM model found 118 variables to bear significance opposed to 44 from the SLX model. Variables denoting minimum distance and counts of amenities up to 250m have a high frequency in the more significant levels. 500m and 1km share lower, but also similar frequencies between one another. The importance of variables denoting a proximity to amenity above 1000m are rarely a case of interest, and only appear at significance levels 'X' and '*'.

The data variability within the best SDEM model was captured with an R Squared of 0.92 denoting 35.7€ error. This presents 8% improvement over the non-spatially enriched model. Usage of a high amount of variables raised question to model's data overfitting. Sensitivity analysis illustrated vast changes within the coefficients adjoined to the variables, though the variables significance yielded marginal differences.

Hyperparametrization of the SVR parameters allowed for creation of a model that fits the original data points with an averaged percentage error of 11% of the price range. The spatially enriched model resulted in worse performance with a 12% mean absolute error. All SVR models needed more than 50% of the data as SVs, and the maximum Lagrange Multiplier coefficient was given to the majority of the SVs.

The biggest faults in building this predictive model were the bigger priced listings which are found to be vastly underestimated. The models thus fail to find a discernable pattern of what allows a higher price tag in a record.

A SOM model with all the significant variables (as denoted by SDEM) demonstrated some geographically understandable clusters – as two rooms belonging to the same node (or one very close) would share very similar proxies statistics. Nevertheless, the biggest denominators in the cluster divisions were still the non-spatial variables which were illustrated on a corresponding PCP figure with a short discussion thereafter.

# 5. CONCLUSIONS

This dissertation focused on the themes of economic valuation of numerous types of spatial data as contributing factors in the mid-term rental market of Lisbon. In the introduction, several research questions were presented and a proposed methodology was carried out to provide an answer.

Various proxies were introduced to dissolve the categories into meaningful variables that one can assess influence from – and given minimum distances to PoI as well counts to amenities that were differing by 500m from 2km to 500m, and a smallest one denoting a distance of 250m, the willingness-to-pay for a room was attached mostly to proxies either describing minimum distance to amenities, or counts of distances up to 1km with very few exceptions from the larger thresholds.

The findings made clear that the success of a model prediction does not depend exclusively on non-spatial variables, but it can also depend on the amenities.

From the spatial variables, viewpoints, park counts, supermarket, universities, theaters, cultural amenities, malls and public transports (mainly metros and trams) were appearing as bearing significance in the model. The census statistics did not prove to be beneficial, as the exposure to these data seemed to bear little to no influence.

However, any and all impact from the amenities mentioned did not justify their use when evaluating the accuracy metrics in the predictive models.

Using all of the created variables, the data was found to best fit with a spatial variant of HPM, the SDEM - with a MAE of 35€,. Nevertheless, this presents a complex model that observed 312 variables as a learning platform. A 91% of captured variability and a substantially changing coefficients as shown through a sensitivity analysis makes overfitting of the data and a lack of robustness of the model a highly-probable scenario.

The best performing SVR model was a non-spatial variable one, and it outperformed the spatially-enriched one with a 11% mean error over the testing sample. The best spatially-enriched model performed 1% worse at 12%, although the variability captured in the testing set down performed by 10%. Hence, complication of a model and going through a methodological procedure that requires time and resources does not bring the value to the model.

The use of GeoSOM was resourceful, as it demonstrated the data could be clustered in meaningful ways, although the prices themselves were seldom the contributing factor. A discussion of the averaged statistics over the chosen clusters found the differences in the probabilities of a landlord within the apartment, the number of rooms, type of bed and the minimum stay allowed for booking to be some of the deciding factors. None of these are inherently spatial variables, although it was found through Local Moran's I that there are hot-spots of male hosts on the coastal regions of the study area as opposed to the central area wherein records increase in number of bedrooms.

In this sense, introducing a spatial factor in the analysis invoked knowledge about the area, an idea of spatial patterns of generalized room and apartment characterization in relation to the study area, and even gives rise to new questions about gender-based analysis of hosts, but the inclusion of amenities' proxies in the predictive models remained questionable.

A limitation of this study is that mid-term rent is rarely a case of interest in similar research – where estate prices predominate this field with prices that are largely driven

by the square meter area of the listings. Moreover, missing variables, imputation of data and market fluctuations create unpredictability in the dataset and might lead to a model that might prove faulty to predict unseen data, as was shown with the adjusted R results in the testing sample.

Last, but not least, usage of spatial proxies within these types of studies comes with a presumption that the people in the demand market have knowledge of the spatial amenities' existence as to create rise in value as a response to listings with desirable proximities to amenities. Nevertheless, as the mid-term rent could be overrepresented by foreign exchange or domestic students that might primarily look only for a good connectivity with their choice of university, this assumption might lack base.

# 6. BIBLIOGRAPHIC REFERENCES


AHLFELDT, G. M., HOLMAN, N., & WENDLAND, N. (2012). An assessment of the effects of conservation areas on value. *Final Report Commissioned by English Heritage*, (May), 1–155.

ALLISON, P. D. (2002). *Missing data*. SAGE.

ANSELIN, L., BERA, A. K., FLORAX, R., & YOON, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*. https://doi.org/10.1016/0166-0462(95)02111-6

ASQUITH, B. J. (2019). *Do Rent Increases Reduce the Housing Supply under Rent Control ? Evidence from Evictions in San Francisco*. 8–31.

BALTAGI, B. H., & LI, J. (2015). Cointegration of matched home purchases and rental price indexes - Evidence from Singapore. *Regional Science and Urban Economics*, *55*, 80–88. https://doi.org/10.1016/j.regsciurbeco.2015.10.001

BARRINGTON-LEIGH, C., & MILLARD-BALL, A. (2017). The world's user-generated road map is more than 80% complete. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0180698

BERTHOUD, R., GERSHUNY, J., & BRITISH HOUSEHOLD PANEL SURVEY. (2000). *Seven years in the lives of British families : evidence on the dynamics of social change from the British Household Panel Survey*. Policy Press.

BING, L., CHAN, K. C. C., & OU, C. (2014). Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements. *2014 IEEE 11th International Conference on E-Business Engineering*, 232–239. https://doi.org/10.1109/ICEBE.2014.47

BOES, S., & NÜESCH, S. (2011). Quasi-experimental evidence on the effect of aircraft noise on apartment rents. *Journal of Urban Economics*, *69*(2), 196–204. https://doi.org/10.1016/j.jue.2010.09.007

BOURASSA, S. C., CANTONI, E., & HOESLI, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *Journal of Real Estate Finance and Economics*. https://doi.org/10.1007/s11146-007-9036-8

BQUARTO. (2010). Quartos e Apartamentos Lisboa Porto Coimbra | Portugal - BQUARTO. Retrieved February 13, 2020, from https://www.bquarto.pt/

BRANCO, R., & ALVES, S. (2015). Affordable housing and urban regeneration in Portugal: A troubled tryst? *European Network for Housing Research*.

BRUNAUER, W. A., LANG, S., WECHSELBERGER, P., & BIENERT, S. (2010). Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna. *Journal of Real Estate Finance and Economics*, *41*(4), 390–411. https://doi.org/10.1007/s11146-009-9177-z

C.A. BREBBIA, & J.J. SENDRA. (2017). *The Sustainable City XII*. Retrieved from

https://books.google.de/books?hl=en&lr=&id=-IJDDwAAQBAJ&oi=fnd&pg=PP1&dq=+Sustainability+and+the+City+jj+brebbia+sendra&ots=g9u7bNDEcw&sig=-lazM4wxNL7tV6Y9YdWdQtgecSE#v=onepage

CECCATO, V., & WILHELMSSON, M. (2016). *The impact of crime on apartment prices : evidence from stockholm , sweden evidence from Stockholm , Sweden*. *3684*. https://doi.org/10.1111/j.1468-0467.2011.00362.x

CHEN, J. (2010). Submarket, Heterogeneity and Hedonic Prediction Accuracy of Real Estate Prices: Evidence from Shanghai. *International Real Estate Review*, *13*(2), 190–217.

COCOLA-GANT, A., & GAGO, A. (2019). *Airbnb, investimento imobiliário e a crise de habitacão em Lisboa*. Retrieved from http://www.ceg.ulisboa.pt/smartour/

COLECTIVO MARXISTA LISBOA. (2019). Portugal: housing crisis spreading beyond city centres. Retrieved September 29, 2019, from https://www.marxist.com/portugal-housing-crisis-spreading-beyond-city-centres.htm

CONWAY, J. . E. D. . N. T. . P. S. K. . & T. N. (2019). RPostgreSQL: R Interface to the 'PostgreSQL' Database System. Retrieved January 22, 2020, from https://cran.r-project.org/web/packages/RPostgreSQL/index.html

CORTES, C., & VAPNIK, V. (1995). Support-vector networks. *Machine Learning*. https://doi.org/10.1007/bf00994018

COULSON, N. E., & LAHR, M. L. (2005). Gracing the land of elvis and beale street: Historic designation and property values in Memphis. *Real Estate Economics*, *33*(3), 487–507. https://doi.org/10.1111/j.1540-6229.2005.00127.x

CRISTINA, M., & TEIXEIRA, C. (2009). *Palavras chave:*

EDZER PEBESMA. (2019). CRAN - Package sf: Simple Features for R. Retrieved January 22, 2020, from https://cran.r-project.org/web/packages/sf/index.html

ELHORST, P., & HALLECK VEGA, S. (2017). The SLX model : Extensions and the sensitivity of spatial spillovers to W. *Papeles de Economía Española*.

ERF. (2004). *Road Traffic Noise The Road Sector's Perspective*. (May), 1–6.

ESTER, E., & MARTINS, S. (2016). *O Uso de Redes Neuronais Artificiais na Estimação do Preço das Habitações na Ilha do Sal, em Cabo Verde*.

FRANCO, S. F., & MACDONALD, J. (2016). The Effects of Cultural Heritage on Residential Property Values: Evidence from Lisbon, Portugal. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2776207

FRANCO, S. F., MACDONALD, J. L., FRANCO, S. F., & MACDONALD, J. L. (2016). *The Effects of Cultural Heritage on Residential Property Values : Evidence from Lisbon , 2016*.

GEOFABRIK. (2019). Geofabrik Download Server. Retrieved January 23, 2020, from

https://download.geofabrik.de/europe/portugal.html

GIBBS, C., GUTTENTAG, D., GRETZEL, U., YAO, L., & MORTON, J. (2018). Use of dynamic pricing strategies by Airbnb hosts. *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/IJCHM-09-2016-0540

GIRRES, J. F., & TOUYA, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*. https://doi.org/10.1111/j.1467-9671.2010.01203.x

GOOGLE-CLOUD. (2019). Places | Google Maps Platform | Google Cloud. Retrieved January 23, 2020, from https://cloud.google.com/maps-platform/places/

GRAVARI-BARBAS, M., & GUINAND, S. (2017). Tourism and gentrification in contemporary metropolises: International perspectives. In *Tourism and Gentrification in Contemporary Metropolises: International Perspectives*. https://doi.org/10.4324/9781315629759

HAKLAY, M., & WEBER, P. (2008). OpenStreet map: User-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

HARRISON, D., & RUBINFELD, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*. https://doi.org/10.1016/0095-0696(78)90006-2

HEATH, S. (2004). Peer-Shared Households, Quasi-Communes and Neo-Tribes. *Current Sociology*. https://doi.org/10.1177/0011392104041799

HENRIQUES, R., BAÇÃO, F., & LOBO, V. (2009). GeoSOM suite: A tool for spatial clustering. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5592 LNCS*(PART 1), 453–466. https://doi.org/10.1007/978-3-642-02454-2_32

HOCHMAIR, H. H., JUHÁSZ, L., & CVETOJEVIC, S. (2018). Data quality of points of interest in selected mapping and social media platforms. *Lecture Notes in Geoinformation and Cartography*. https://doi.org/10.1007/978-3-319-71470-7_15

HUGHES, N. (2018). 'Tourists go home': anti-tourism industry protest in Barcelona. *Social Movement Studies*, *17*(4), 471–477. https://doi.org/10.1080/14742837.2018.1468244

KAKAR, V., FRANCO, J., VOELZ, J., WU, J., KAKAR, V., FRANCO, J., … WU, J. (2016). *Effects of Host Race Information on Airbnb Listing Prices in San Francisco*. Retrieved from https://econpapers.repec.org/paper/pramprapa/69974.htm

KAKAR, V., VOELZ, J., & WU, J. (2017). *Munich Personal RePEc Archive The Visible Host : Does Race guide Airbnb rental rates in San Francisco ? The Visible Host : Does Race guide Airbnb rental rates in San Francisco ?* (78275).

KHOLODILIN, K. A., MENSE, A., & MICHELSEN, C. (2017). Market Break or Simply Fake? Empirics on the Causal Effects of Rent Controls in Germany. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2793723

KILIBARDA, M. (2018). *Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments*. (Ml). https://doi.org/10.3390/ijgi7050168

KOHONEN, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288

KOHONEN, T. (2001). *The Basic SOM*. https://doi.org/10.1007/978-3-642-56927-2_3

KONG, F., YIN, H., & NAKAGOSHI, N. (2007). Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China. *Landscape and Urban Planning*. https://doi.org/10.1016/j.landurbplan.2006.02.013

KUHN, M. (2020). *Classification and Regression Training [R package caret version 6.0-85]*.

KUHN, M., & JOHNSON, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

LARSSON, B., RONG, K., & THOMSON, W. (2014). *Non-manipulable house allocation with rent control b*. *82*(2), 507–539. https://doi.org/10.3982/ECTA10893

LESAGE, J. P. (2008). *AN INTRODUCTION TO SPATIAL ECONOMETRICS*. 19–44.

LESAGE, J. P. (2014). What regional scientists need to know about spatial econometrics. *Texas State University-San Marcos*, 394–396.

LESAGE, J., & PACE, R. K. (2009). Introduction to spatial econometrics. In *Introduction to Spatial Econometrics*. https://doi.org/10.1111/j.1467-985x.2010.00681_13.x

LI, S., YE, X., LEE, J., GONG, J., & QIN, C. (2017). Spatiotemporal Analysis of Housing Prices in China: A Big Data Perspective. *Applied Spatial Analysis and Policy*, *10*(3), 421–433. https://doi.org/10.1007/s12061-016-9185-3

LIMSOMBUNCHAI, V., GAN, C., LEE, M., BOX, P. O., & ZEALAND, N. (2004). *House Price Prediction : Hedonic Price Model vs . Artificial Neural Network House Price Prediction : Hedonic Price Model vs . Artificial Neural Network*. (March). https://doi.org/10.3844/ajassp.2004.193.201

LISBOA - DIGITAL TERRAIN MODEL. (2010). Retrieved February 1, 2020, from https://www.arcgis.com/home/item.html?id=d6b2c1b4ccab4e0899d70d9565f89cd6

MARLET, G., & WOERKENS, C. VAN. (2005). *Tolerance , aesthetics , amenities or jobs ? Dutch city attraction to the creative class*. 5–33.

MARTÍNEZ, L. M., & VIEGAS, J. M. (2009a). Effects of Transportation Accessibility on Residential Property Values. *Transportation Research Record: Journal of the Transportation Research Board*, *2115*(1), 127–137. https://doi.org/10.3141/2115-16

MARTÍNEZ, L. M., & VIEGAS, J. M. (2009b). *Effects of Transportation Accessibility on Residential Property Values Hedonic Price Model in the Lisbon , Portugal , Metropolitan Area*. 37–39. https://doi.org/10.3141/2115-16

MCCLUSKEY, W. J., MCCORD, M., DAVIS, P. T., HARAN, M., & MCILHATTON, D. (2013).

Prediction accuracy in mass appraisal: A comparison of modern approaches. *Journal of Property Research*, *30*(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

MESSONIER, M. L., & LUZAR, E. J. (1990). A Hedonic Analysis of Private Hunting Land Attributes Using an Alternative Functional Form. *Journal of Agricultural and Applied Economics*, *22*(2), 129–135. https://doi.org/10.1017/s1074070800001887

MEYER, D. (2019). *Misc Functions of the Department of Statistics, Probability Theory Group*.

MIT. (2016). GitHub - Python Zomato API Wrapper. Retrieved January 23, 2020, from https://github.com/fatihsucu/pyzomato

MONTERO, J. M., MÍNGUEZ, R., & FERNÁNDEZ-AVILÉS, G. (2018). Housing price prediction: parametric versus semi-parametric spatial hedonic models. *Journal of Geographical Systems*, *20*(1), 27–55. https://doi.org/10.1007/s10109-017-0257-y

MORO, M., MAYOR, K., & LYONS, S. (2013). *Does the housing market reflect cultural heritage ? A case study of Greater Dublin*. *45*(1), 2884–2903. https://doi.org/10.1068/a45524

NATIONAL OFFICE FOR STATISTICS. (2011). *Census 2011 - General Report for Portugal*.

NEIS, P., & ZIPF, A. (2012). Analyzing the contributor activity of a volunteered geographic information project - The case of OpenStreetMap. *ISPRS International Journal of Geo-Information*. https://doi.org/10.3390/ijgi1020146

NÜST, D., GRANELL, C., HOFER, B., KONKOL, M., OSTERMANN, F. O., SILERYTE, R., & CERUTTI, V. (2018). Reproducible research and GIScience: An evaluation using AGILE conference papers. *PeerJ*, *2018*(7), e5072. https://doi.org/10.7717/peerj.5072

OPENSTREETMAP, C. (2020). OpenStreetMap Statistics. Retrieved January 23, 2020, from https://www.openstreetmap.org/stats/data_stats.html

OSGEO FOUNDATION. (2018). *pgRouting Manual Release v2.6*.

PEREIRA, S. M. DA C. (2017). *Exploring the Relationship between Residential Rents and House Prices in the Portuguese Residential Estate Market*.

POORT, J. (2015). *Cultuur en creativiteit naar waarde geschat*. (September).

POSTGRESQL GLOBAL DEVELOPMENT GROUP. (2020). PostGIS 3.0.1dev Manual. Retrieved January 22, 2020, from https://postgis.net/docs/index.html

PRESIDENT, M. M., & MARQUES, P. (2018). *THE GLOBAL RENT GAP OF LISBON ' S HISTORIC*. *13*(4), 683–694. https://doi.org/10.2495/SDP-V13-N4-683-694

R CORE TEAM. (2019). R Language. Retrieved January 22, 2020, from https://cran.r-project.org/doc/manuals/r-release/R-lang.html

ROGER BIVAND. (2019a). CRAN - Package spdep. Retrieved January 22, 2020, from https://cran.r-project.org/web/packages/spdep/index.html

ROGER BIVAND. (2019b). Spatial Regression Analysis for R. Retrieved January 22, 2020, from https://cran.r-project.org/web/packages/spatialreg/index.html

ROSEN, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34–55. https://doi.org/10.1086/260169

SAFRONOV, O. (2017). *NON-RESIDENTIAL REAL ESTATE PRICE EVALUATION OF LISBON , PORTUGAL NON-RESIDENTIAL REAL ESTATE PRICE EVALUATION OF LISBON , PORTUGAL*.

SAGE, J., SMITH, D., & HUBBARD, P. (2012). The rapidity of studentification and population change: There goes the (Student)hood. *Population, Space and Place*. https://doi.org/10.1002/psp.690

SCHÖLKOPF, B. (2002). Learning with kernels. *Proceedings of 2002 International Conference on Machine Learning and Cybernetics*. https://doi.org/10.7551/mitpress/4175.001.0001

SMOLA, A. J., & SCHÖLKOPF, B. (2004). A tutorial on support vector regression. *Statistics and Computing*. https://doi.org/10.1023/B:STCO.0000035301.49549.88

STEFAN MILTON BACHE AND HADLEY WICKHAM. (2014). magrittr: A Forward-Pipe Operator for R. Retrieved January 22, 2020, from https://cran.r-project.org/web/packages/magrittr/index.html

SUYKENS, J. A. K., & VANDEWALLE, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*. https://doi.org/10.1023/A:1018628609742

THE POSTGRESQL GLOBAL DEVELOPMENT GROUP. (2014). PostgreSQL: The world's most advanced open source database. Retrieved January 22, 2020, from https://www.postgresql.org/

TOBLER, W. (2004). On the first law of geography: A reply. *Annals of the Association of American Geographers*. https://doi.org/10.1111/j.1467-8306.2004.09402009.x

TROY, A., & GROVE, J. M. (2008). Property values, parks, and crime: A hedonic analysis in Baltimore, MD. *Landscape and Urban Planning*, *87*(3), 233–245. https://doi.org/10.1016/j.landurbplan.2008.06.005

UNIPLACES. (2019). Uniplaces - General. Retrieved January 25, 2020, from https://www.uniplaces.com/about-us

VAN DUIJN, M. (2013). *Location Choice , Cultural Heritage and House Prices*.

VAN DUIJN, M., & ROUWENDAL, J. (2012). Cultural heritage and the location choice of Dutch households in a residential sorting model. *Journal of Economic Geography*, *13*(3), 473–500. https://doi.org/10.1093/jeg/lbs028

WANG, X., WEN, J., ZHANG, Y., & WANG, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik*, *125*(3), 1439–1443. https://doi.org/10.1016/j.ijleo.2013.09.017

WICKHAM, H. . & M. K. (2019). *R Database Interface [R package DBI version 1.1.0]*.

WITTEN, I. H., PAL, C. J., & FOURTH, M. (2017). *Maximum Margin Hyperplane Extending instance-*

*based and linear models*.

ZANG, G., BERARDI, V., & REITERMANOVÁ, Z. (2010). Time series forecasting with neural network ensembles: An application for exchange rate prediction. *The Journal of the Operational Research Society*.

ZOMATO. (2019). Zomato API. Retrieved January 23, 2020, from https://developers.zomato.com/api

# 7. ANNEX

- Sample Query – Creation of all reachable nodes from the listings within 5KM of their vicinity using the OSM Network and pgRouting

```
CREATE table catchment_a as WITH nodes AS (
  SELECT
    array_agg(n_id) AS nodes
  from
    house_market.closest_h_node
)
SELECT
  from_v as start_node,
  node as end_node,
  agg_cost as cost
from
  nodes,
  pgr_drivingdistance(
    'SELECT gid as id, source as source, target as target, cost_s as cost FROM public.ways' :: text,
    nodes,
    5000,
    false
  )
```

- Sample Query - Generating Closest Nodes to Parks

```
CREATE TABLE poi_park_nodes AS
SELECT
  poi_parks.id as parks_id,
  nodes.id as nodes_id
FROM
  (
    SELECT
      DISTINCT ON (id, geom) *
    FROM
      poi_data.poi_parks
    WHERE
      id IS NOT NULL
    ORDER BY
      RANDOM()
  ) AS poi_parks CROSS
  JOIN LATERAL (
    SELECT
      id,
      the_geom
    FROM
      public.ways_vertices_pgr
    ORDER BY
      poi_parks.geom <-> the_geom
    limit
      CEIL(
        ST_AREA(poi_parks.geom :: geography)+ 15000
      )/ 20000
  ) AS nodes
WHERE
  ST_Distance(
    geography(poi_parks.geom),
    geography(nodes.the_geom)
  )< 30
  AND ST_DISTANCE(
    geography(poi_parks.geom),
    geography(nodes.the_geom)
  )<= 30
```

- Sample Query – Creation of Variable: Counts of Trams within 2KM

```
INSERT INTO
  poi_data.interm_min_costs (
    select
      '15' as poi_type,
      start_node,
      min(cost) as cost
    from
      (
        select
          fee.start_node,
          fee.end_node,
          cost
        from
          public.catchm_a as fee
        WHERE
          (start_node, end_node) in (
            select
              node_id,
              pe.n_id
            from
              house_market.rent_market,
              transport.tram_nodes as pe
          )
      ) d
    GROUP BY
      d.start_node
  );


ALTER TABLE
  house_market.rent_market
ADD
  column tram_min double precision;
UPDATE
  house_market.rent_market
set
  tram_min = (
    SELECT
      cost
    from
      (
        SELECT
          poi_type,
          start_node,
          cost
        from
          poi_data.interm_min_costs
        WHERE
          poi_type = '15'
      ) f
    WHERE
      f.start_node = rent_market.node_id
  );
```

- Sample Query – Creation of Variable: Minimum Distance to Trams

```
INSERT INTO
  poi_data.interm_counts_2KM (
    select
      '15' as poi_type,
      f.start_node,
      count(f.start_node) as cnts
    from
      (
        select
          distinct fee.start_node,
          fee.end_node,
          cost
        from
          public.catchm_a as fee
        WHERE
          (start_node, end_node) in (
            select
              node_id,
              n_id
            from
              house_market.rent_market,
              transport.tram_nodes
          )
      ) f
    group by
      start_node
  );


ALTER TABLE
  house_market.rent_market
ADD
  column trams_2KM bigint;
UPDATE
  house_market.rent_market
SET
  trams_2km = (
    SELECT
      cnts
    from
      (
        SELECT
          poi_type,
          start_node,
          cnts
        from
          poi_data.interm_counts_2KM
        WHERE
          poi_type = '15'
      ) f
    WHERE
      f.start_node = rent_market.node_id
  );
```

# Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS