Lisbon School of Economics & Management
Universidade de Lisboa

MASTERS IN

MANAGEMENT

MASTER'S FINAL WORK

PROJECT

ANALYSIS OF GEOGRAPHICAL FACTORS ON HOUSE SHARING: LISBON CASE STUDY

ALESSANDRO TORELLA DI ROMAGNANO

SUPERVISION:

CARLOS J. COSTA

CO-SUPERVISION:

JOAO T. APARICIO

MARCH - 2023

ACKNOWLEDGMENTS

I would like express my sincere gratitude to professor Carlos J. Costa and João T. Aparicio for their patience, guidance and consistent support throughout the research project. In particular I would like to thank professor Costa for the motivation conveyed during the lectures which revived in me the interest toward data science, and João for the friendly chats at the end of our meetings, which many times uplifted and helped me carry on the research.

To my family, who I know is always with me and supports me in every decision. For their unconditional love and for always being by my side in every step of my life.

To my girlfriend, for the love and emotional support given to me throughout this thesis. For understanding me, for cheering results with me and for endure through difficult times with me, for always being by my side.

To my university colleagues, who I must call friends, with whom I spent unforgettable memories during these 2 years of study and I am sure and hope our paths will cross again in the future.

ABSTRACT

House-sharing and peer-to-peer accommodation expansion in Lisbon transformed the real estate business and the touristic industry of the city. In this context, Airbnb is the main driver of such transformation with rising popularity among both private property owners and professional businesses seeking to monetize their dwellings. Among the consequences of the concentration of many properties in certain neighborhoods, gentrification is the one which requires the most attention and needs for regulations. House affordability and rent gaps in the central locations is a key theme explored extensively by the literature in social and geographical studies. The purpose of this study is to analyze the Airbnb pricing distribution in Lisbon by identifying the main factors contributing to gentrification in order to help regulatory agencies reach targets 11.1 and 11.2 of the Sustainable Development Goal 11[1], namely providing a safe affordable housing and a sustainable transportation system. The methodology implemented encompasses traditional spatial analysis tools combined with innovative approaches through topic modelling and natural language processing techniques. The results obtained confirm, on one hand, the presence of high prices hotspots identified by the literature as well as uncovering new ones and, on the other, show a very close relationship between the subway network and key transport hubs of Lisbon and the pricing of the peer-to-peer accommodations, indicating that clustering of high rental values around these spots promote gentrification and low house affordability phenomenon. The study contributes, on one side, to enrich the current literature on spatial analysis and gentrification in the context of house sharing and, on the other hand, provides regulatory agencies with insights to monitor and efficiently control areas at risk.

[1] https://www.globalgoals.org/goals/11-sustainable-cities-and-communities/

RESUMO

A expansão da partilha de casas e do alojamento peer-to-peer em Lisboa transformou o negócio imobiliário e a indústria turística da cidade. Nesse contexto, o Airbnb é o principal impulsionador dessa transformação, com popularidade crescente entre proprietários privados e empresas profissionais que buscam monetizar suas residências. Dentre as consequências da concentração de muitos imóveis em determinados bairros, a gentrificação é a que mais requer atenção e necessidade de regulamentação. A acessibilidade das casas e as lacunas de aluguel nos locais centrais é um tema-chave explorado extensivamente pela literatura em estudos sociais e geográficos. O objetivo deste estudo é analisar a distribuição de preços do Airbnb em Lisboa, identificando os principais fatores que contribuem para a gentrificação, a fim de ajudar as agências reguladoras a atingir as metas 11.1 e 11.2 do Objetivo de Desenvolvimento Sustentável 11[2], ou seja, fornecer uma habitação segura e acessível e um sistema de transporte sustentável. A metodologia implementada engloba ferramentas tradicionais de análise espacial combinadas com abordagens inovadoras por meio de modelagem de tópicos e técnicas de processamento de linguagem natural. Os resultados obtidos confirmam, por um lado, a presença de hotspots de preços elevados identificados pela literatura, bem como a descoberta de novos e, por outro, mostram uma relação muito estreita entre a rede de metro e os principais centros de transporte de Lisboa e a tarifação do acomodações peer-to-peer, indicando que o agrupamento de altos valores de aluguel em torno desses pontos promove a gentrificação e o fenômeno de baixa acessibilidade de casas. O estudo contribui, por um lado, para enriquecer a literatura atual sobre análise espacial e gentrificação no contexto do compartilhamento de casas e, por outro lado, fornece às agências reguladoras insights para monitorar e controlar com eficiência as áreas de risco.

Keywords— spatial analysis, gentrification, house sharing, sharing economy, tourism

[2]https://www.globalgoals.org/goals/11-sustainable-cities-and-communities/

Contents

List of Figures

List of Tables

1 INTRODUCTION

Airbnb is a short-term rental portal allowing people to rent their properties, whether apartments or single rooms, over the internet. Since its launch in 2007, Airbnb has established itself as a prominent platform and a notable representative of the sharing economy. As of 2023, more than 7 million properties are listed on the platform, of which a stunning 4 million are in the European area. Among the capitals of Europe, Lisbon is the one with the highest percentage of Airbnb listings per 1000 inhabitants, therefore making it one of the most interesting cases to analyze. The city, as of 2023, has roughly 13 thousand listings published on the website[3] as of the time of writing. As Mendes (2021) argues, the effect of Covid-19 only lasted momentarily and didn't have any negative consequence in the upward trend of the platform in the Portuguese capital.

The rising number of Airbnb offerings is coupled with high demand due to the steady increase in tourism Lisbon has experienced over the past decade, with a significant boost in the last two years. Phenomena like digital nomads' affluence from Europe and overseas as well as a rising establishment of new companies in the territory are benefiting the already positively trending tourism economy.

The analysis of house sharing in Lisbon contributes to SDG 11[4], which focuses on the development of sustainable cities and communities by efforts in making cities and human settlements more inclusive, safe, resilient, and sustainable. The research also contributes specifically to target 11.1, safe and affordable housing, and target 11.2, affordable and sustainable transportation system. By addressing the rising pricing issue of short-term rentals in Lisbon, which is the main cause promoting gentrification of specific neighborhoods of the historical center (Lestegás et al., 2019), the thesis aims at providing insights for the public administration to address the problem. Gentrification, the displacement of local inhabitants due to the massive affluence of foreign people in the area, and house affordability in Lisbon are in fact ongoing problems that the municipality is trying to address by implementing regulations. This work, therefore, aims at bringing additional knowledge through the application of traditional and innovative spatial analysis techniques to better address these issues.

By studying the geographical factors that influence house sharing in Lisbon, researchers can identify ways to promote sustainable urbanization, such as reducing urban sprawl and promoting compact, mixed-use development (Lee et al., 2016). These findings can then be used by municipalities to take decisions regarding the implementation of regulatory processes to short-term rentals spread (Pereira, 2020). The phenomenon of gentrification is strictly related to the study field of spatial analysis and of urban geography as it is shown in many studies like the one of Mendes (2021) and Gant (2016). The purpose of the thesis is to give concrete answers to the above by implementing techniques of geospatial data analysis and exploration combined

[3] http://insideairbnb.com/lisbon

[4] https://www.globalgoals.org/goals/11-sustainable-cities-and-communities/

with Natural Language Processing (Natural Language Processing (NLP)) tools in order to give structure to the phenomenon based on quantitative data, so to explain the relations between Airbnb offerings and their surroundings promoting gentrification. In particular, it focuses on understanding which are the elements influencing the rise of house pricing in certain areas of Lisbon by finding critical areas of high values.

The methodology adopted to provide such insights also contributes to enriching the bag of techniques used to explore this phenomenon around the world, using Lisbon as a case study given its peculiarity. Previous research on the accommodation field is in fact extensive and dates back to the 1970s, during the period quantitative studies emerged in the context of the housing market and real estate as well as in the hotel industry, with the need to quantify the variables associated with the accommodation industry. The current research is focused on re-adapting the bag of techniques developed in the last 5 decades to the short-term rental accommodation industry like Airbnb.

The method followed to carry out the analysis starts with an Exploratory Spatial Data Analysis framework on the available data and ends with the design and implementation of a new method of analysis as an artifact. A tool dedicated to spatial analysis is developed, and its use in this case study is discussed.

2 LITERATURE REVIEW

This chapter begins with a thorough analysis of the literature on urban studies and Airbnb phenomena in Lisbon. Afterward, it provides a deep dive into the literature regarding specific approaches and techniques used in the context of the accommodation industry, considering spatial analysis studies performed in the traditional housing business and their revisitation in the modern era of short-term rentals like Airbnb.

2.1 Airbnb Phenomenon: the Lisbon case

The decision to study the Airbnb phenomenon and its characteristics in Lisbon is highly justified due to the unique nature of the case. Lisbon, the capital of Portugal, has experienced a substantial surge in tourist influx to the point where it can be described as over-tourism According to Costa et al. (2019), a strict relationship exists between over-tourism and short-term accommodation in Lisbon. While providing interesting figures regarding the expansion of the platform in the city as well as emphasizing the touristic wave phenomenon, the study also shows how the growth of Airbnb and Tourism progresses at the same pace, but cannot significantly prove the causality between one and the other for which it suggests the use of more advanced econometric models with the necessity to take into consideration the specialty measures. Short-term rentals (STR) are known to reshape and displace the housing markets. Studies like the one of Cocola-Gant and Gago (2021), show how this phenomenon directly applies to the city of Lisbon and, in particular to the Alfama district. By analyzing the Airbnb business from an investor perspective, it shows how the platform offers a new way of financing through the housing market. Through the professionalization of the hosts, Gant argues how a gentrification displacement is caused by mainly 2 reasons 1) homeowners are incentivized to sell at profitable prices their houses to professional Airbnbers, 2) the residents can't keep up with the rise in prices and leave the area. Not only gentrification but also the rent gap is a consequence of the expansion of Airbnb in every area of the city. As Wachsmuth and Weisler (2018) points out, in fact, the data reveals that short-term rentals (STR) are on the rise in areas that have not traditionally been popular for accommodations such as hotels, resulting in unexpectedly high rents.

Short-term rents in Portugal are known as 'Alojamento Local', AL in short, which is the denomination given to any peer-to-peer (P2P) accommodation in Portugal, including apartments, villas, guesthouses, and hostels.

Enlisting a property to the registry of AL was streamlined and eased in 2014 by Portuguese legislation, with the intention to ride the tourist wave by meeting the high demand for accommodation. As a direct consequence of the easy application process individuals seeking to monetize their dwellings found the situation particularly appealing. Enlisting personal properties in the registry is, in fact, completely free and the 'license' acquired does not have an expiring date. To give quantitative meaning to the phenomenon, the number of guests who stayed at 'Alojamento Local' properties doubled in the period 2014-2019, going

from almost 14 million to 26 million. (Jover and Cocola-Gant, 2022) Portuguese municipalities, especially the ones of Lisbon and Porto, together with the south region of the Algarve, certainly benefited from this implementation. The positive effects are numerous, from the revitalization of overlooked areas to the rehabilitation of old inhabitable properties. Many dwellings in Lisbon are in fact in need of restructuring and the certain profit via STR is incentivizing investors to restore many of these buildings, indirectly helping the municipality in the effort of re-modernizing the city.

As pointed out before, STR are also responsible for a series of negative consequences, and current residents are the ones who suffer from them in the first place. In certain areas of the city of Lisbon, in fact, the expansion of AL has provoked many complaints from the Portuguese citizens, especially in terms of identity and gentrification of the territory and, secondly, because of the irrational increase in the pricing of the rent. To express these phenomena, Costa et al. (2019) points out that Lisbon has a capacity of around 50 guests per 1000 inhabitants, and this limit is exceeded in many of the parishes with the particular case of Misericordia and Santa Maria Maior, the two most central zones, where the number of visitors in 'AL' is higher than the one of residents. Social studies like the one of Lestegás et al. (2019), report an increase in prices of more than 30% in just two years and additionally depict an increasing rent gap phenomenon in specific areas of the city, damaging the quality of the residents' living conditions. The study provides concrete evidence of the commodification of the housing market, where the exchange value of a property surpasses its use value. In an effort to mitigate the rapid expansion of AL in critical areas, therefore, Portuguese government legislated law 61/2018, essentially allowing municipalities to regulate how many new properties could be granted the AL license in certain areas of their choice. Nevertheless, according to Peralta et al. (2020), there has been a rush to register the dwellings in the days preceding the implementation of the law, essentially boosting the number of possible STR in a very short period. The same study also concentrates on the price effect of this measure but concludes that the rent values are not affected by the law.

Although the consequences of such limitations are very interesting to be analyzed, the literature about them is cautious and doesn't present accurate findings whether the problems of gentrification and tensions among residents are rising or remain the same. Only Costa et al. (2019) suggests that the reform has had little effect on the growth trend of Airbnb in critical areas.

2.2 Spatial Analysis

This section aims to provide a comprehensive understanding of the current state of the research regarding spatial analysis in the accommodation industry, specifically focusing on the spatial distribution of Airbnb listings. The review highlights the common objective of explaining the distribution of observations in the study area, which involves finding determinants shaping the territorial distribution of Airbnb, exploring the characteristics and factors related to its expansion, and investigating the influence of neighborhood

environment and points of interest on the spatial distribution of house sharing. The review also shows that most of the studies employ regression-based techniques to study such patterns and distributions. The review contributes to the understanding of how Airbnb listings are distributed in urban areas and their potential impacts on tourism and the economy, emphasizing the need for further research in this area.

Geospatial data analysis encompasses all the techniques used to gain insights from geospatial data as the one analyzed in this work. Geospatial Datasets are tables including geographic information of the observations contained in them, usually location specific, like coordinates of a specific point or area of a more extensive site. More formally, according to Fischer (2011), to perform spatial analysis there needs to be information regarding the attributes and the location of an observation. Moreover, he adds, to be able to assert such analysis is performed one has to include the spatial relationship between observations, otherwise it would be a simple data analysis, even if its is underlined the spatial distribution of the variables. It is therefore essential to compare observations with one another in spatial terms, whether with distance or by area covered.

Spatial analysis in the housing market and real estate as well as in the hotel industry dates back to the 1970s. Palm (1978) segments the urban housing market through correlation measures between the neighborhood and the dwellings in the area, Basu and Thibodeau (1998) makes extensive use of spatial regression to assert the existence of spatial autocorrelation in the house price using the structural characteristics of the property and the location in which they are built.
What interested the researchers was finding the dynamics of the housing prices and the hotel prices in different neighborhoods as well as assessing the most significant factors influencing them. It is safe to say that most of the approaches provided the ground for the methodologies still in use nowadays and many of the insights found are still valid for today's cases.
Although these studies are extensive, their applicability to the present context may be limited, as they were conducted in a different setting, that is the one of the housing market or the hotel accommodation industry, very different from the environment in which this analysis resides. Airbnb, as stated above, has drastically changed the accommodation industry on which those models were built, that's why current researchers have the need to adapt or find new ways of explaining accommodation dynamics in this new context. As mentioned above some study methods are still applicable to house-sharing scenarios while for some cases there's a need to introduce new elements and variables to better give answers to the research questions.

The current literature for spatial analysis in the accommodation industry has the common objective of explaining the distribution of the observations in the study area. More precisely, examples of questions arising from this kind of research are: finding determinants shaping the territorial distribution of Airbnb,

finding the characteristics of the spatial distribution of Airbnb and the factors related to its expansion, or exploring the neighborhood environment's influence on the spatial distribution of house sharing. As pointed out by Sun et al. (2021), most of the studies regarding spatial distribution take into consideration external factors in the study area to find a relationship with Airbnb listings. Xu et al. (2019) concentrate on the spatial relation between the offerings and Points of Interest (POI) of the city of London as well as the distance to the city center and the transportation network influence showing how the distribution of Airbnb is uneven and the proximity to some of the POIs increases the number of listings. Singh and Botelho Azevedo (2021) introduces demographic elements in an attempt to find if Airbnb benefits tourism by expanding accommodations also in the periphery of the study area. The research proves that in both the study areas, Lisbon and Barcelona, offerings are concentrated in the already popular spots and, rather than benefiting the rural areas by bringing more visitors, they increase the inequalities.

As seen in the aforementioned study, the question of whether Airbnb could potentially benefit rural areas is relevant also from a business point of view, Zhang and Chen (2019) introduces the concept of geographic convenience. In essence, the question posed is whether, as for hotel accommodation, the proximity to the city center positively affects the reputation of Airbnb. Analyzing the cities of Chicago, Los Angeles, and New York, the paper shows how only in New York Airbnb is predominant in the center of all three cities, with the exception of New York which presents a slightly higher concentration of offerings in the peripheral areas when compared to the other two cities. The aforementioned study confirms the finding of Singh and Botelho Azevedo (2021), strengthening the fact that Airbnb is not bringing any touristic benefit to peripheries. Distance to the city center has also been addressed by Gyódi and Łukasz Nawaro (2021), showing how the prices of the offerings are positively affected by the proximity to the center of the city for the case of Rome. Lastly, the centrality importance is explored by Celata et al. (2020), explicitly stating that central offerings shadow the peripheral ones, which are not even considered, according to their conclusion, by the users when consulting the platform.

To study such patterns and distribution mentioned above the literature explored presents an extensive use of regression techniques, the majority of the sources addressing the questions related to spatial dynamics and spatial distribution contains or conclude with a regression-based approach. For example, (Xu et al., 2019) first apply and then re-adapt an OLS regression to explore the density of Airbnb listings in the city of London. In the analysis, the Ordinary Least Squares method is adapted to a Geographically Weighted regression model (GWR), where the main purpose is to give weight to the listings based on the parish they are located in by constructing a matrix containing all the offerings and their respective weight, to finally introduce this matrix into the OLS regression model. Hong and Yoo (2020), similarly begins with an OLS regression approach but takes it further than the GWR by modeling a Multiscale GWR to account for spatial variance within the study area. Gyódi and Łukasz Nawaro (2021) uses a full econometrics approach that entirely bases the results on regression techniques.

The literature regarding spatial regression and its application in this field of study is very extensive, this work concentrates therefore on exploring different approaches which bring insights to explain the distribution of Airbnbs around main metropolitan cities from another perspective than the one of spatial regression.

2.3 Relevant statistical methods

Describe the area using local indicators of spatial association (LISA)

According to Oxoli (2019), among the Local Indicators of Spatial Association (LISA), Moran's I is the most popular statistic to represent patterns of spatially distributed variables. Spatial association is a concept that leaves space for interoperability but Tobler (1970) summarises it efficiently stating that "*everything is related to everything else, but near things are more related than distant things*". With this rule in mind, it is useful to introduce the concept of Spatial Weight Matrix, where each row and column value corresponds to an observation pair (Fischer and Wang, 2011). The value representing the observation pair is strictly related to the spatial relation, or spatial lag, between the values. Formally, there are two ways of representing such relation:

1. If a common boundary is present

$$W(i,j) = \begin{cases} 1, & \text{if there exists a boundary between the 2 points} \\ 0, & \text{otherwise} \end{cases}$$

2. If the distance between the two observations is not higher than a certain threshold

$$W(i,j) = \begin{cases} 1, & \text{d(i,j) > d, (d>0)} \\ 0, & \text{otherwise} \end{cases}$$

Other ways of calculating such values include clustering algorithms like K-NN. The concept of Spatial Weights Matrix is the key to understanding Moran's statistic, created in 1950 by Patrick Moran to express spatial association.

Spatial association statistics take into account the scope and magnitude of the analysis. In fact, when all the observations in the study area are taken into account the statistic is referred to as Global, while it is called Local if the calculation focuses on a single point or a limited area. Generally, the baseline for the statistic is a cross product between the Weights Matrix W defined above and a matrix R denoting attribute similarity between two locations i and j as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} W_{i,j}R_{i,j}; i \neq j$$

The baseline hypothesis for this statistic is the following:

- H0: Complete randomness, spatiality has no importance
- H1: Spatial relation is present for the variable of interest

The statistics allow for both univariate and bivariate analysis. Anselin (2019), warns regarding the interpretation of the bivariate case, in which just one variable's values are encoded in the spatial weight matrix. Therefore it does not take into account the correlation between two values at the same location.

2.4 Relevant definitions

As the analysis is carried out in a Geographic Information Science (GIS) setting it is necessary to define the ad-hoc terminology used throughout this work. In essence, to indicate geographical elements, we will use the following terms:

- Study area: limited region in which the analysis is carried out and where all the observation lie, in this case the municipality of Lisbon, depicted in Figure 1
- Neighbourhood or Parish: the terms are used interchangeably and refer to the districts within the study area In this case the ones that the Camara Municipal de Lisboa defines for the city of Lisbon. In total there are 25 parishes. According to Peralta et al. (2020), neighborhoods are a point of reference for tourists looking for where to stay in a city.
- Listing or Offering: a data point within the parishes referring to a single Airbnb.

A study area suitable for spatial data analysis must contain the following elements:

- Polygons: defining geographical areas
- Points: defining geographical coordinates.

In the terminology of the programming language used and the associated library, Polygons are defined as *MultiPolygon objects* and points as *Points objects*.
In the context of this work, the MultiPolygons objects define the Parishes, while the Points objects define each of the Airbnb Listings.

NLP, as the denomination suggests, is the ensemble of methods aiming at understanding human written communication. As any other form of communication, language is in fact very complex and, to be understood, need to be analyzed in all its component. NLP comprises of automatic summarization, translation

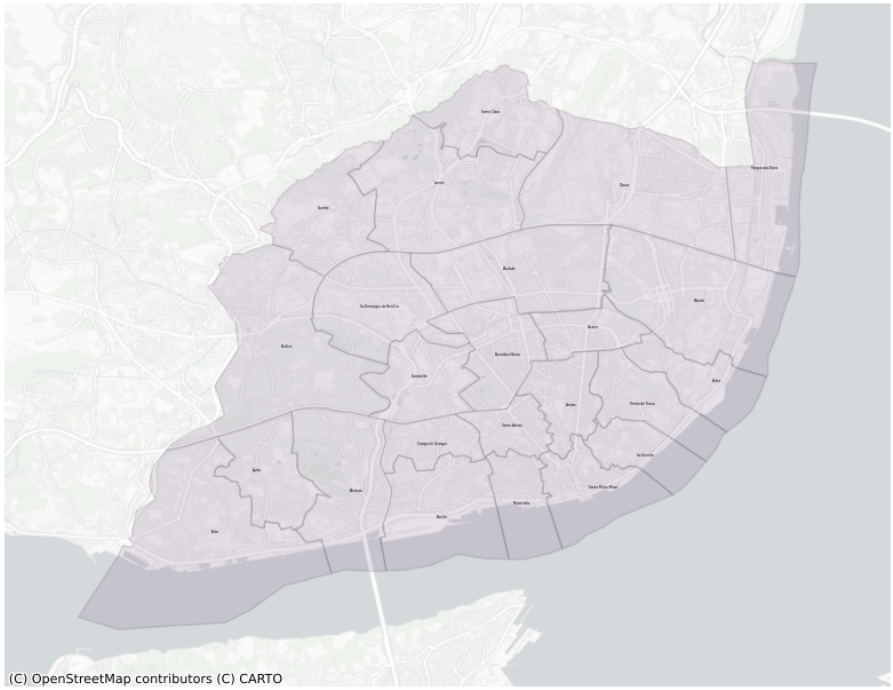


FIGURE 1: Overview of the study area

and segmentation techniques and finds its application in any field involving human communication systems. In the context of this work techniques related to Information retrieval will be applied, with the goal of understanding the relative importance of the words in the description of Airbnb listings based on their frequency, relative and absolute. A technique involving term frequency analysis will be carried out with the goal of quantifying the topic the hosts wants to focus on when describing the offering.

3 METHODOLOGY

The presented methodology aims, through the implementation of traditional and innovative approaches, at defining the spatial relations that govern the pricing of short-term rentals in Lisbon and the consequent gentrification.

OBJECTIVES	METHODS
IDENTIFY CLUSTERS OF SHORT-TERM RENTALS OFFERINGS IN LISBON	1. SPATIAL STATISTICS - Global Morans I - Local Morans I
IDENTIFY FACTORS & TOPOLOGICAL FEATURES PROMOTING HIGHER PRICES	2. TOPIC MODELLING - Text Analysis - Terminology Clustering - Term frequency analysis - Sentiment Analysis

TABLE I: Methods' summary

By first individuating hot spots for high rental values through spatial statistics, it then proceeds to analyze in detail the description of each listing, characterizing how it influences the perspective of the guests and how the pricing is justified by topological features. Consequently, it links the pricing hotspots identified with the terminology clusters by means of spatial statistic and finally interprets the relationship between the accessibility to the subway system through correlation analysis.

The first step of the methodology implemented consisted of the calculation of the Global Moran's I statistics to assess the presence of clusters among the observations in the study area. As stated by Oxoli (2019), the statistic, among the Local Indicators of Spatial Association (LISA), is the most informative for identifying spatial relations. The calculation of this statistic is useful in confirming the presence of clusters in the Portuguese capital, which is a result stated in numerous research analyzed. Since the Global statistic doesn't inform of the specific location of the clusters, a local version of the Moran's I was calculated in order to obtain information about where exactly the spatially related clusters of high and low values are. This approach is commonly used in the literature ((Adamiak et al., 2019),(Basu and Thibodeau, 1998),(Singh and Botelho Azevedo, 2021)) and is a very effective and intuitive tool to describe a study area that presents a tendency to clusters. By putting together the spatial statistics it was possible to define the spatial distribution of pricing in Lisbon, which provided the base ground on which the following steps were carried on. It also provided an understanding of the hot spots and cold spots for investors and property owners who want to develop an Airbnb business in Lisbon.

A text analysis of each Airbnb description was then carried out through the implementation of Natural Language processing techniques. NLP has been employed to extract meaningful insights from the textual data by identifying the ten most common terms used in the Airbnb listings' descriptions for each neighborhood, a grouping by topic was then performed based on the contextual relevance of the terms in order to give structure to the findings.

Prior to implementing the retrieval of the most common terminology used, an extensive pre-processing of the description has been carried out encompassing the following steps: 1) removal of HTML elements as '
' or backlashes 2) conversion to lowercase 3) removal of English and Portuguese stopwords[5]. To account for the completeness of the insights the method embeds a semi-automatic translation functionality if frequent terms appear in both languages. For example, 'restaurant' in English, and 'restaurante' in Portuguese, have the same relevance but would count twice in the set of most common words, therefore making the analysis biased.

The text analysis then allowed for a representation of the most influencing terms by area, therefore pro-

[5]stopwords are those terms with highest frequency like 'and' or 'is' which don't bring additional insights to a text analysis

viding a high-level view of what hosts mostly value when describing their property. After the discovery of terminology clusters, that is to say, parishes or sets of parishes where a set of words was particularly relevant, it was performed a deep dive into each topic frequency per description. Topic modeling was then performed on the set of most common words overall. The term frequency analysis consisted of calculating the occurrence of the words belonging to each of the previously defined topics in each Airbnb description, with the goal of providing a sentiment-alike analysis identifying the most relevant topic of the description. The results provided by the term frequency analysis contributed to specifically define the clusters identified by the aforementioned high-level view.
To allow for future and ad-hoc research a tool exploring terms in each neighborhood was developed, among its functionalities it allows the user to tailor its search by, for instance, defining a threshold on how many top words to look for.

In order to link the results obtained from the statistics which allowed to define clusters in the study area in the first step (i.e. Local Morans I), a bivariate form of the same was applied, considering the price and the term frequency per topic per description. The result presented allows to link the value of the property not only with its location but also with the terminology used in describing it. The versatility of the methodology for each topic allows for replication in many different business contexts.

The results from the previous steps suggested a particular focus on the linkage between the subway system and the Airbnb listings, which was carried out by means of descriptive and correlation analysis in order to understand the relation between the two environment

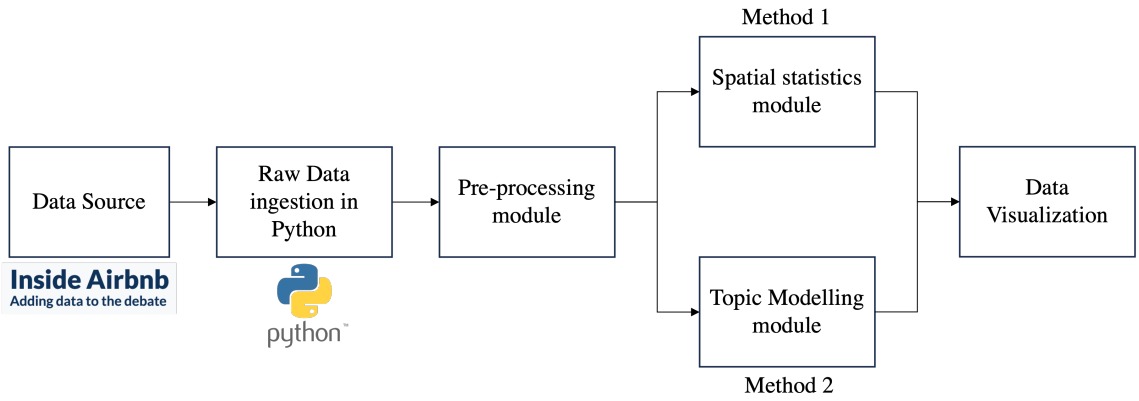


FIGURE 2: Methodology pipeline

Figure 2 shows how the methodology described is implemented. Each 'module' is a package of python functions allowing for the calculations needed for this analysis.

The analysis was carried out following steps of the POST-DS (Process Organization and scheduling Elected Tools for Data Science) methodology for Data Science (Costa and Aparicio, 2020), which incorporates the business objectives in the analysis, making it coherent with the purpose of this thesis, which is, not only analyzing the spatial distribution and dynamics of Airbnb context in Lisbon but also giving the results a contextual meaning and justification. It also allows for a very precise scheduling of the work, an essential component of such kind of work.

3.1 Data collection and Tools

Airbnb

The data collected to carry out the analysis comes from insideairbnb.com[6], a website collecting the latest data from the Airbnb website through web scraping. It is worth mentioning that the main purpose and reason of this web source is providing *data and advocacy about Airbnb's impact on residential communities*. The main goal is in fact to provide communities with the necessary information to *understand, decide, and control the role of renting residential homes to tourists*[7].

After careful distinction, the variables in the dataset have been summarized in table II.

HOST	HOOD	PROPERTY ATTRIBUTES	REVIEWS	DATES AND NIGHTS	OTHERS
id	neighbourhood	property_type	number_of_reviews	minimum_nights	calculated_host_listings_count
host_url	neighbourhood_cleansed	room_type	number_of_reviews_ltm	maximum_nights	calculated_host_listings_count_entire_homes
host_since	neighbourhood_group_cleansed	accommodates	number_of_reviews_l30d	minimum_minimum_nights	calculated_host_listings_count_private_rooms
host_location	neighborhood_overview	bathrooms	first_review	maximum_minimum_nights	calculated_host_listings_count_shared_rooms
host_about		bathrooms_text	last_review	minimum_maximum_nights	listing_url
host_response_time		bedrooms	review_scores_rating	maximum_maximum_nights	last_scraped
host_response_rate		beds	review_scores_accuracy	minimum_nights_avg_ntm	source
host_acceptance_rate		amenities	review_scores_cleanliness	maximum_nights_avg_ntm	description
host_is_superhost		price	review_scores_checkin	calendar_updated	picture_url
host_thumbnail_url			review_scores_communication	has_availability	latitude
host_picture_url			review_scores_location	availability_30	longitude
host_neighbourhood			review_scores_value	availability_60	geometry]
host_listings_count			instant_bookable	availability_90	
host_total_listings_count			reviews_per_month	availability_365	
host_verifications				calendar_last_scraped	
host_has_profile_pic					
host_identity_verified					

TABLE II: Airbnb dataset features overview

The dataset contains 13,138 data points, each corresponding to an Airbnb listing in the city. Variables information comprises of hosts' attributes, especially relevant among these are the super host status, known for its positive influence on guests' perception (Abrate et al., 2022) and the longevity of the host (host_since). Neighborhood attributes are useful because they already capture the high-level distribution of the offerings and are easy to use in combination with the dataset containing a geographical structure of the city of Lisbon. Property attribute contains 2 of the most important attributes used in the analysis, which are,

[6] http://insideairbnb.com
[7] http://insideairbnb.com/about/

respectively, the price and the room type. Reviews attributes, and in particular scores for location, values, and cleanliness have been proven to be very explicative in an analysis of this kind and will be broadly used in this work, especially for what concerns the review of the location. Dates and Nights attributes were discarded during the analysis as they are not strictly related to the main objective. Among the aforementioned Property attributes, description is certainly the one that brought the most insights into this analysis. Lastly, geometries (latitude and longitude) attributes are the key element to give spatiality context to the work.

Metro Network

Regarding the analysis of the underground system in Lisbon, data related to the location of each metro station was retrieved from the public administration open data of Portugal[8]. In order to make it suitable for a network analysis the downloaded data has been transformed manually and analyzed through the Python library networkx[9] to find the relevant topological features.

Tools

The chosen tool to extract information from the collected data is Python[10], which is useful for its versatility regarding both ease of use and tailoring of the analysis.
This being a geographical analysis dealing with tabular data the library chosen is Geopandas[11]. The main characteristic of this tool is that it replicates the same functionalities and usage of its 'parent' library Pandas, with the caveat of allowing for geographic data (i.e. coordinates) to be treated, therefore being a hands-on tool for those who are already acquainted with tabular analysis in Python.
To perform spatial statistics, the library of choice is ESDA[12], a subpackage of PySal (Python for Spatial Data Analysis).
Regarding Network analysis, the library of reference is networkx[13], chosen for its completeness and flexibility.

3.2 Data Description

Figure 3 shows the high density of Airbnb listings in the center of the city, as argued extensively by the literature explored, visually confirming the high ratio between the inhabitants and the count of offerings. In Figure 4 it is possible to notice the uneven distribution of the median pricing by neighborhood, such

[8] https://dados.gov.pt/pt/
[9] https://networkx.org
[10] https://www.python.org
[11] https://geopandas.org/en/stable/
[12] https://pysal.org/esda/
[13] https://networkx.org

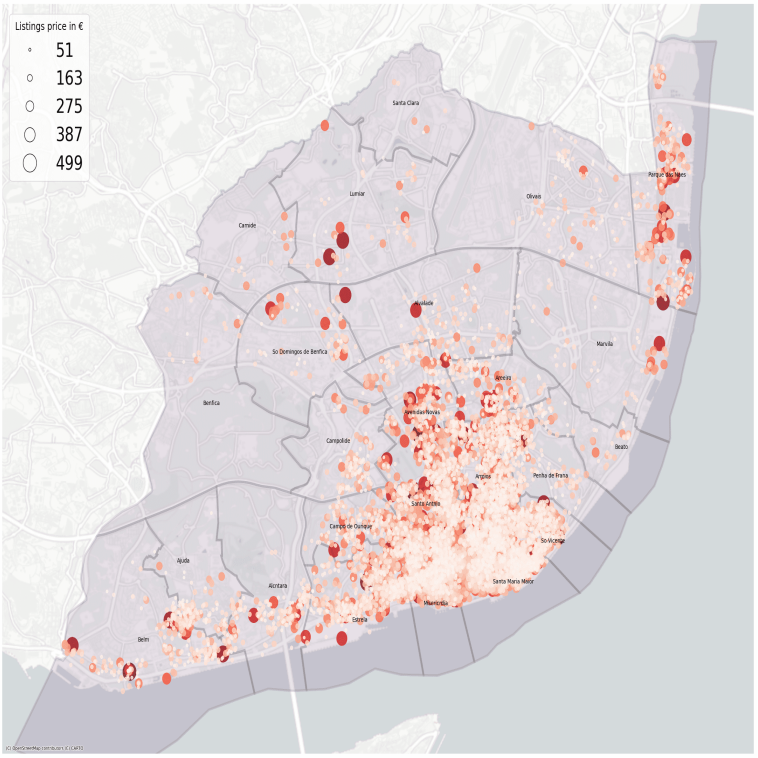


FIGURE 3: Price distribution of Airbnb in Lisbon

unevenness suggests the need for a closer look, through statistical spatial methods, in order to obtain a more precise clustering of values, whether high or low.

The Metro Network is built in the following way: Metro stations and their location are considered the nodes of the graph, while the distances between each of them are the edges of such graph. Since, generally, metro lines travel the same distance in both directions, the network is built as an undirected graph. Metro systems, as noted by Garrison and Marble (1962), is a mean of transport that can be well represented through graph theory. The topological features of a network apply to the metro system, in particular with

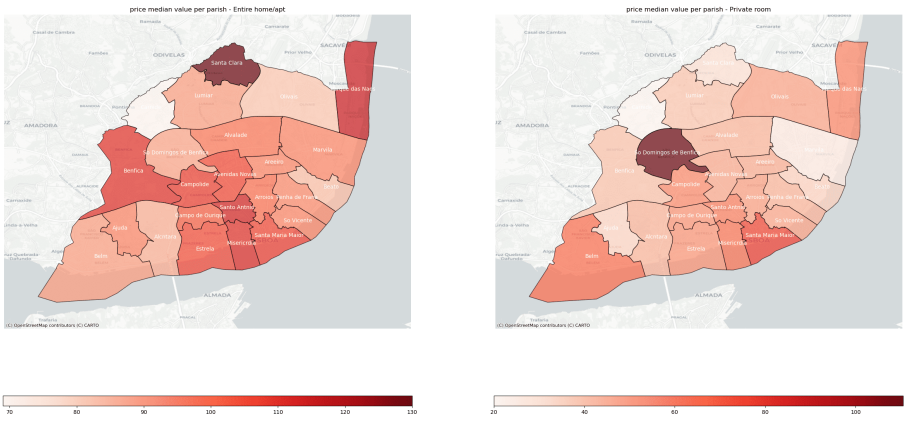


FIGURE 4: Median price distribution of Airbnb by parish

regard to centrality measures, studies like the one of Aparicio et al. (2022) and Derrible (2012) focus on assessment robustness and resilience making extensive use of such measures. A key centrality measure of the metro network is betweenness, as Derrible (2012) clearly states, it indicates a metro stop's importance based on how well it connects two other stations. Connectedness is in fact particularly important to users since they'll value the ability of the station to be a transfer point to many locations.

4 RESULTS

In this section, the methodologies outlined in section 3 are followed and the main results of each step are shown. To get a sense of the importance of each neighborhood it's worth looking at the count of Airbnb listing per parish, presented in table III.

Neighbourhood	Entire home/apt	Private room	% Entire home/apt	% Private room
Ajuda	154	21	88%	12%
Alcntara	193	19	91%	9%
Alvalade	99	72	58%	42%
Areeiro	116	141	45%	55%
Arroios	1064	580	65%	35%
Avenidas Novas	272	179	60%	40%
Beato	65	17	79%	21%
Belm	246	63	80%	20%
Benfica	26	22	54%	46%
Campo de Ourique	212	40	84%	16%
Campolide	117	43	73%	27%
Carnide	22	8	73%	27%
Estrela	747	142	84%	16%
Lumiar	56	14	80%	20%
Marvila	51	21	71%	29%
Misericordia	1961	305	87%	13%
Olivais	71	73	49%	51%
Parque das Naçoes	221	36	86%	14%
Penha de França	300	135	69%	31%
Santa Clara	5	14	26%	74%
Santa Maria Maior	2715	304	90%	10%
Santo Antnio	819	239	77%	23%
So Domingos de Benfica	62	36	63%	37%
So Vicente	909	111	89%	11%

TABLE III: Count by room type in each Neighborhood

This representation allows discovering which are the densest areas and allows for a first understanding of the distribution of the listings.

It is important to notice here the fact that for the Parishes nearest to the center of the city, namely Misericordia, Santa Maria Maior, and Santo Antonio, the percentage of Entire Apartments is the highest, while private rooms have a higher share in peripheral areas.

Another key metric and functionality of Airbnb is the reviewing system, guests are in fact incentivized to give reviews at the end of the service issued. But how does the metric behaves in a spatial context?

Among the types of reviews, the location score is worth exploring to find out if there is a relation between the price of an offering and its location. When depicting the distribution of the location reviews a skewness is identified, in fact, most of the scores lie between 4.0 and 5.0, it being the maximum.

Figure 5 shows clearly how the distribution is almost flat before reaching the mentioned range. To see how the distribution behaves spatially, the median for each parish is presented in Figure 6. It is noticeable how the central areas present higher review scores, as expected, as well as the parish of Parque des Naçoes and Belèm. This central areas review scores for the location is in line with the pricing distribution depicted in Figure 3 for the same areas and justifies, at least theoretically, the higher prices. Interestingly, the pricing of Parque des Naçoes and Belèm doesn't follow the same trend, since it doesn't reflect the high location reviews scores. This result is particularly relevant because it is not justified by the distance to the center, and leaves space for further research into its causes. Table IX, in the appendix, contains correlation scores between location reviews and pricing for each of the parishes. Correlation scores average at 0.017 and none of them is particularly relevant to answer the initial question of whether there exists a relation between the price and the location review score of an offering. Such value is possibly given by the skewness of the distribution, as noted before in Figure 5.

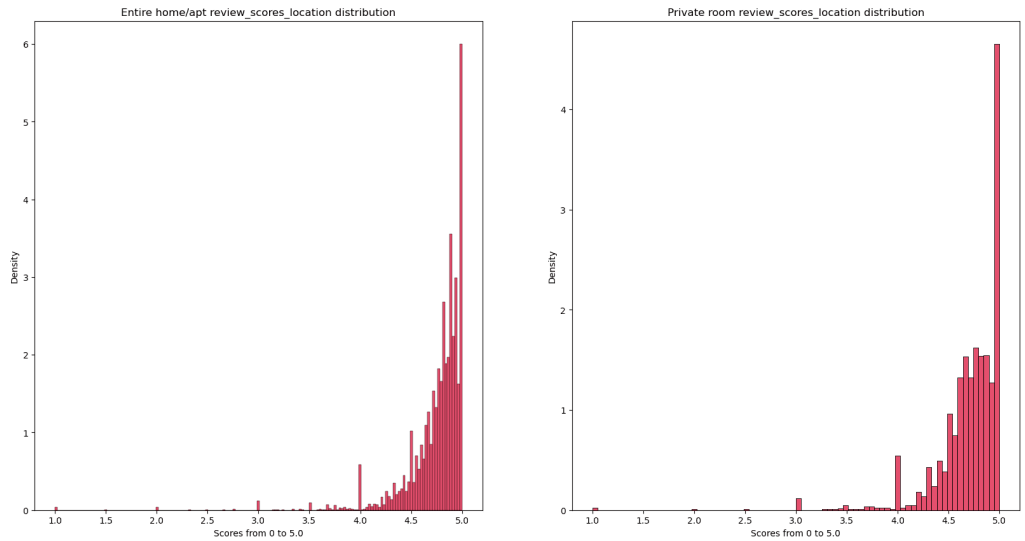


FIGURE 5: Location reviews distribution by house type

Previous results show how defining price behavior for Airbnb needs a more refined analysis that accounts for more variables than just scoring system to location. The results shown in the following section aim at providing richer insights that will better explain the spatial rules of Airbnb pricing in Lisbon.

Global Moran's I

In essence, the global statistic gives a quantifiable value to what is visible through the map represented above and will show how highly-priced offerings tend to cluster in the city center. The hypothesis for the Global statistic is:

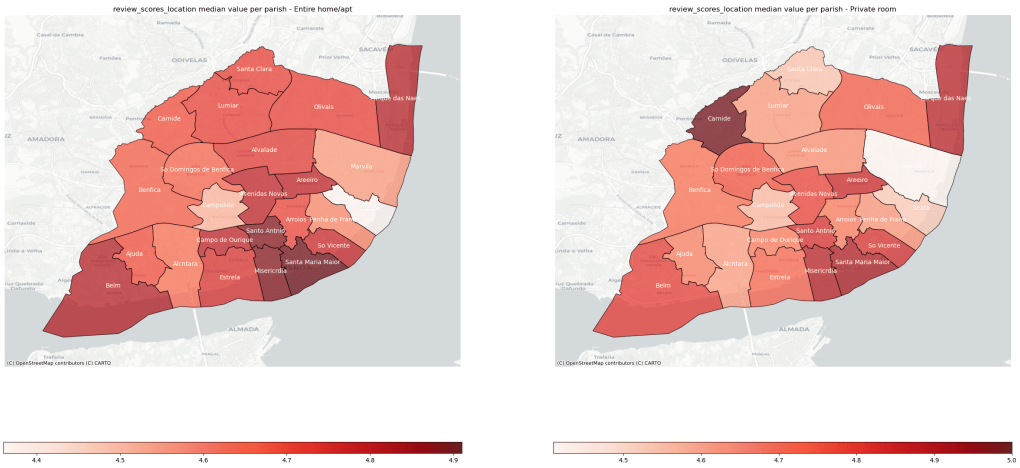


FIGURE 6: Location reviews scores by median value distribution by Parish

- H0: Price is randomly distributed among Airbnb offerings in Lisbon
- H1: Price follows a non-random distribution among Airbnb offerings in Lisbon

The value of the Global Moran's I statistic for the city of Lisbon is 0.10, which indicates a positive spatial autocorrelation, therefore a tendency of Airbnb offerings to cluster. The statistic is significant given a pseudo p_value equal to 0.001, which allows rejecting the null hypothesis with 99.9% confidence. It needs to be clarified that this p_value differs from the one in a, say, linear regression, but explaining how this value is calculated is beyond the scope of this work.

This statistic suggests that clusters are present in the study area within high values, but fails to provide more information about where these clusters are located, this is, therefore, the question intended to be answered in the next sections.

Local Moran's I

Local Moran's I statistic gives insights regarding where and which kind of clusters are present in the study area. Figure 7 shows the values of the local statistic for each observation. As shown in 2.3, Local Moran's I is a localized version of the parent statistic, the Global. Essentially it is possible to observe, in the city center and in the Parish 'Parque des Naçoes' a cluster of High-High values, meaning that there is a positive association of High values.

High-High (HH) values clusters are located in the city center, slightly skewed to the left, therefore in the area of Bairro Alto, a very touristic neighborhood, very prone to gentrification, as shown by Mendes (2021) in a specific analysis of the main characteristics of the area. An interesting hot spot is the one that can be noticed in Parque des Naçoes. The result is surprising given the distance from the center and from historical points of interest. To explain the result it's worth giving more information regarding the location.

The area of Parque des Naçoes has been modernized for the Expo World Fair in 1998, representing one of the most successful urban development of recent times in Lisbon (Ochoa, 2022). The area is in fact known as a wealthy and residential, with modern buildings and services for the people who decide to live there, which could explain the reason behind the agglomeration of high values together.

The other observable clusters are those where there exists a positive association of lower values, meaning areas where offerings with low prices tend to appear. Low-Low values cluster mainly in the parishes of Arroios, Areeiro and Penha da França. The results follow the topological features of these areas, which do not have relevant spots for mass tourism and present a less expensive housing market, positioning in the mid-lower end of the ranking of price per square meters, as shown in Table VIII. Although it is trivial to notice agglomeration of Low-Low values in the peripheral areas, it needs to be observed also a high concentration of HL values, which are considered outliers together with non-significant values in grey.

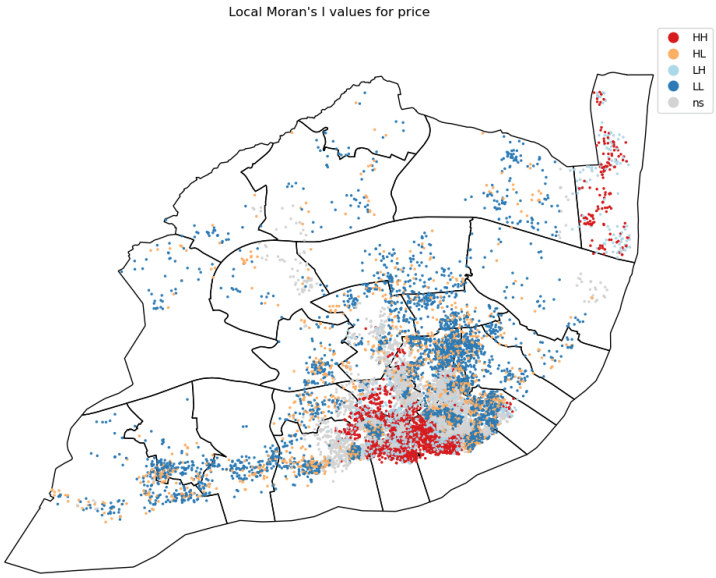


FIGURE 7: Local Moran's I based on price

4.1 Natural Language Processing Analysis

The following results section embeds an innovative approach and provides a topic-modeling text frequency analysis with the scope of understanding the factors promoting higher prices and, consequently, gentrification in Lisbon for the critical areas indicated through the spatial statistics.
The approach recalls the one implemented by Santos et al. (2022), where a topic-modelling analysis is carried out on the description of Lisbon Airbnb, but it takes distance from it for 2 reasons:

1. It looks at a more granular view implementing an analysis by parish.
2. Looking at the share of each topic provides information about which of them is stressed the most by the host, delivering a sentiment analysis of the description as well.

The NLP analysis is carried out on the description of each Airbnb listing. As a first step the 10 most common words per parish are retrieved, the purpose is to identify the terms mostly used by the hosts when publishing their property on the website in a specific neighborhood. A practical example is provided in Figure 8.

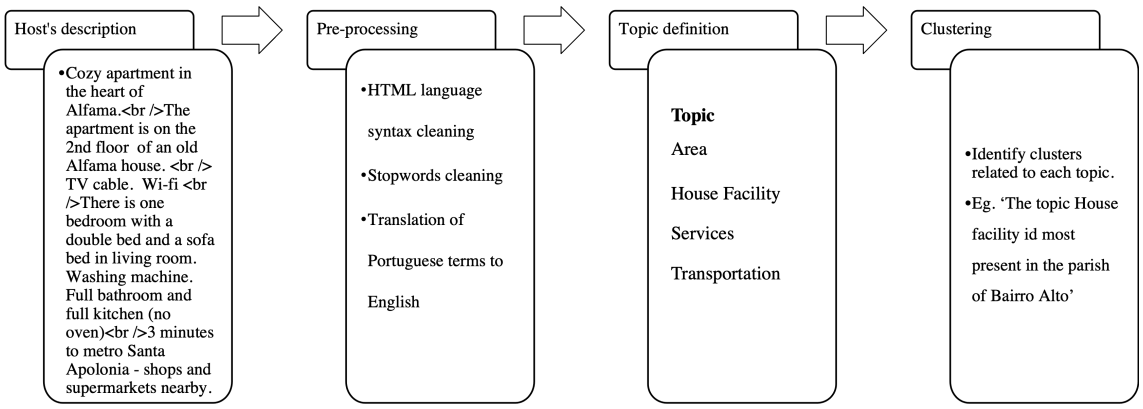


FIGURE 8: NLP process example

The findings presented in table VII allow for the categorization of the terms into 4 topics, based on the following criteria:

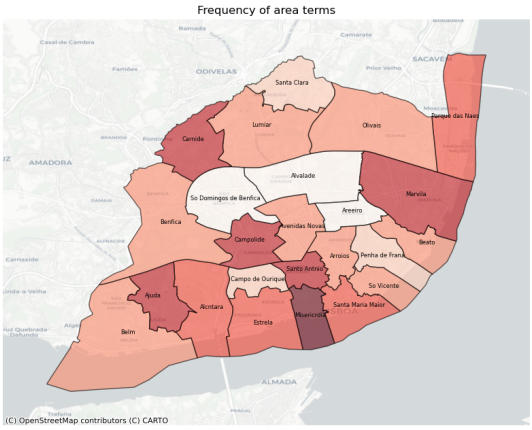
1. Area: those words strictly related to the points of interest in the city
2. House facility: those words related to the amenities.
3. Services: those words related to the services such as bars or shopping centers
4. Transports: those words related to transportation means

The common terms among the parishes divided by topic are then shown in Table IV.
Next, a high-level view of the terms distribution is presented.

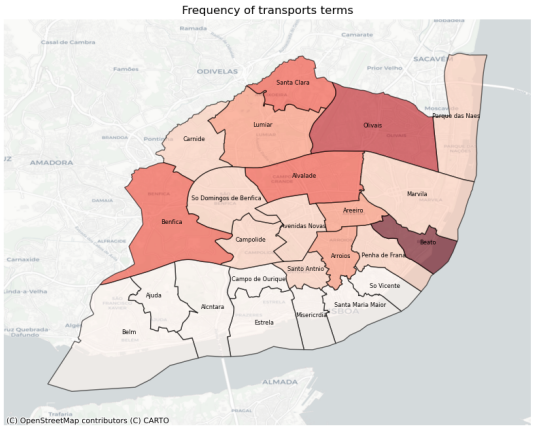
Topic	Most common words
Area	ajuda, avenida, bairro, beato, belem, belém, benfica, campolide, carnide, castle, center, central, centro, chiado, cidade, downtown, estrela, factory, graça, heart, liberdade, lx, nações, parque, pombal, river, saldanha, santos, telheiras
House Facilities	banho, bathroom, bathrooms, bedroom, comfortable, conforto, cozinha, garden, kitchen, machine, máquina, modern, quiet, renovated, wifi, tv
Services	amoreiras, café, colombo, hospital, restaurants, shopping, walk
Transports	aeroporto, airport, bus, metro, metros, oriente, station, stations, subway, train

TABLE IV: Main topics in Airbnb descriptions

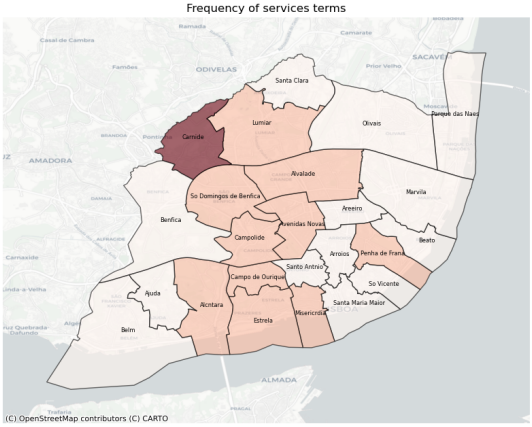
The idea on which this result is based is the following: the terms of a single topic may appear more than once in a set of most common words of a parish. Hence the occurrence, among the top 10 words, is explanatory of the importance of a topic in one neighborhood compared to the others. As an example, words related to the topic of transportation, like *'bus'* and *'metro'* appear 2 times in Alvalade neighborhood offerings, but 1 time in Benfica, explaining that hosts give higher importance overall to transportation in Alvalade. The results of this experiment are visible in Figure 9



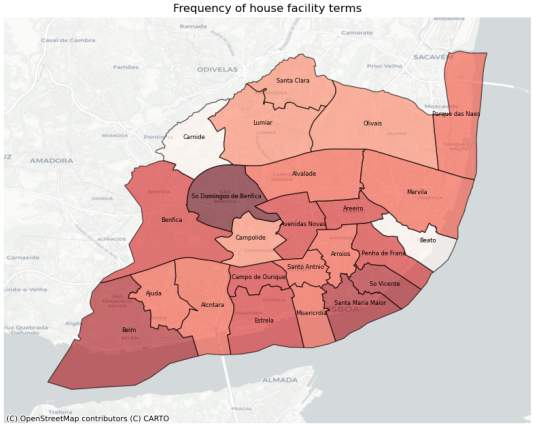
(a) Distribution of terms related to Area



(b) Distribution of terms related to Transports



(c) Distribution of terms related to Services



(d) Distribution of terms related to House facilities

FIGURE 9: Visual distribution of set of terms per parish

Results show how terms related to transportation are highly present in the parishes on the eastern side of Lisbon, specifically in the neighborhood of Beato, Olivais and Campo Grande, with the addition of Benfica, where 3 out 10 words are transport related. Central and Western areas do not have transportation terms among the top ten most common.

Terms related to the area, contrary to the transportation ones, are mostly present in the center of the city, specifically in the parishes of Misericordia and Santo Antonio, respectively containing 5 and 4 out of 10 terms related to the area.

Service-related terms have the lowest distribution among the Topics explored. With the exception of Carnide, in none of the other parishes the frequency is higher than one word out of the 10 most common. It is also notable how in many parishes, especially at the left and right extremities, these kinds of terms never appear.

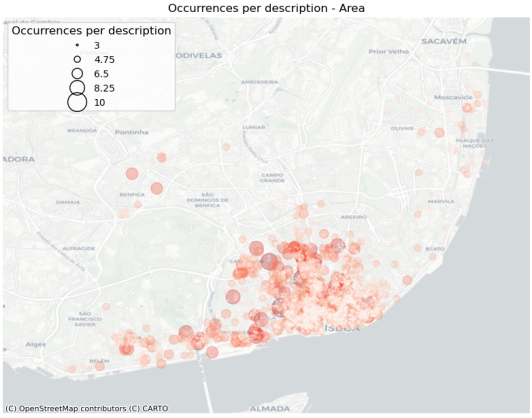
Lastly, House facility terms are very present among all parishes, the lowest frequency is 30%. Even if not clear, central parishes seem to have the highest frequency as well as Belém on the far left.

To explore the result presented in Figure 9, a more granular view is displayed.

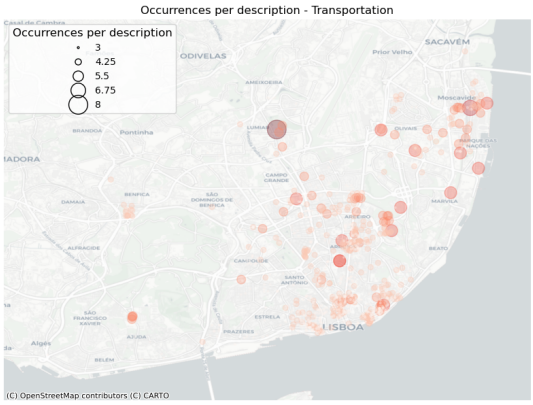
This time the occurrence of each term is calculated per single description. The purpose of this experiment relies on the idea that a host who values a certain characteristic of the Airbnb will stress it more times in the description to make it more visible.

As noted previously, terms related to transportation are highly present on the right side of Lisbon. In figure 10 it is possible to observe how the count of words related to means of transport per single description increases as moving out of the center of the city. With peaks near Areeiro, Parque des Nacoes and the airport, as well as in Lumiar. Another repeating pattern noted before is shown here with those terms related to the area, Airbnb hosts in Lisbon, in fact, particularly stress the importance of the area when the offering is situated in the center, as well as in the western side, especially near LX factory, a very touristic area of Lisbon. Service-related terms do not present a clear distribution, although it is possible to notice how clustering is present around the Colombo commercial center above the parish of Benfica. Lastly, hosts seem to enhance the house facilities mainly in the center of the city, but it's not possible to identify a clear cluster. Figure 10c shows that terms related to services are more stressed in the peripheral area of the city and mentioned in the offerings located in the center of the city.

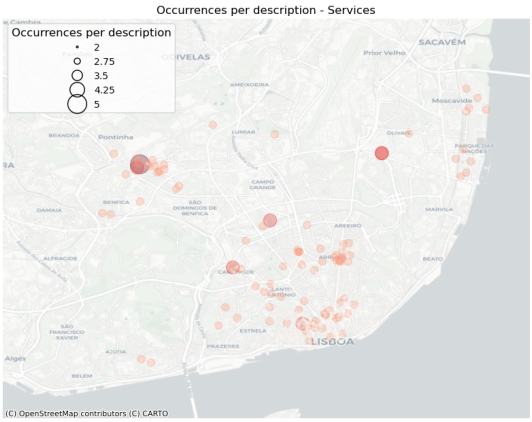
The frequency of terms related to the area topic in the descriptions is concentrated in the historical part of Lisbon, slightly shifted towards west with peaks in the area near the bridge, a very touristic spot. House facility frequency extends also further from the center of Lisbon and doesn't present peaks in any area, therefore making it difficult to interpret.



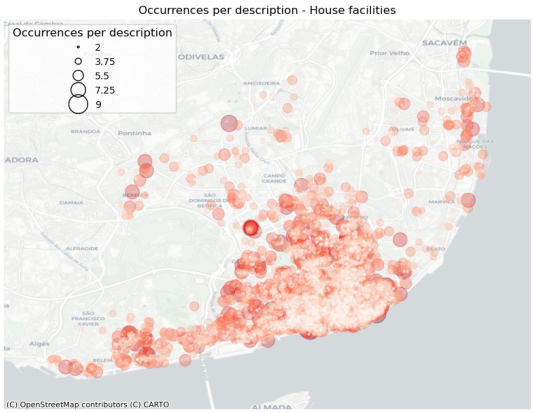
(a) Frequency of Area terms per offering description



(b) Frequency of Transport terms per offering description



(c) Frequency of Service terms per offering description



(d) Frequency of House terms per offering description

FIGURE 10: Visual distribution of terms per description

4.2 Airbnb and Subway network relationship

Following the interesting results presented in 4.1 regarding transportation, here is investigated the relationship between the main mean of transportation in Lisbon, the subway, and the offerings in Lisbon.

To better incorporate the subway system to give importance to the Metro stations, a network has been constructed.

Figure 11 shows that most of the Airbnb in Lisbon can be found within 1 km from each metro station, reaching a peak of 220 listings at a distance of 500 meters. The distribution then flattens showing just a small increase at 4.5 km from the closest metro station.

Lisbon counts for 4 metro lines: Amarelha, Azul, Verde, and Vermelha, corresponding, respectively, to the colors yellow, blue green, and red, therefore the result in Table V shows that the green line is the has the highest count of closest Airbnb, slightly more than 1800 offerings than the blue line. This finding is not associated with a higher median price per night, which is actually 2€ lower than the blue line. The red

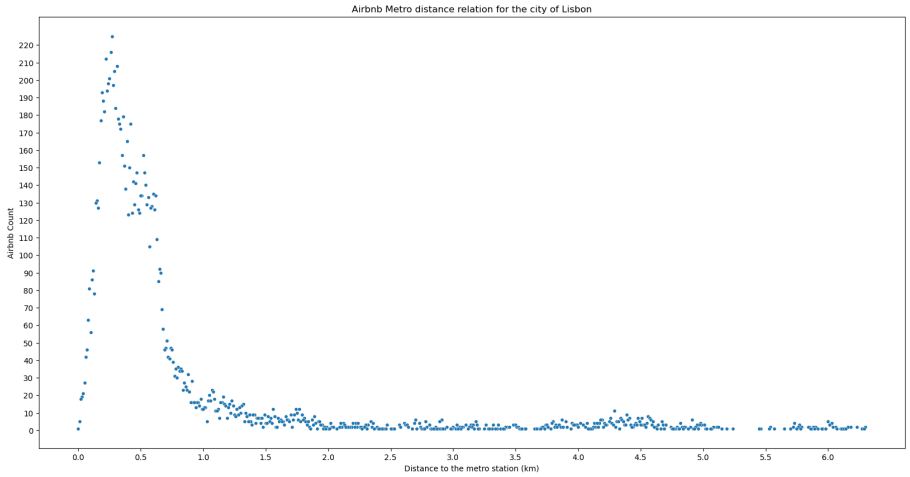


FIGURE 11: Relationship between Airbnb and the nearest metro station, count of listings

line presents the lowest count of offerings nearby, but not a relevant decrease in median price. The count of offerings near the yellow line is lower than the green and blue ones but shows a median price equal to that of the blue line, which counts for almost double the offerings. For the computation of the results above, metro stations where a crossing of two lines occurs were not counted for ease of interpretation. By recalling results in Figure 7 , the highest count of listings is located where the clustering of Low-Low values appears, specifically in the neighborhood of Arroios and Areeiro. This result suggests a possible relation between the count distance to each metro station, namely, accessibility, and the price of a listing. The answer to this is provided in the subsequent sections.

Nearest Metro station line	Airbnb count	Median price	Median distance
green	5226	99.0 €	330 meters
blue	3393	101.0	430 meters
yellow	1418	101.0 €	520 meters
red	579	90.0 €	590 meters

TABLE V: Figures per Metro Line

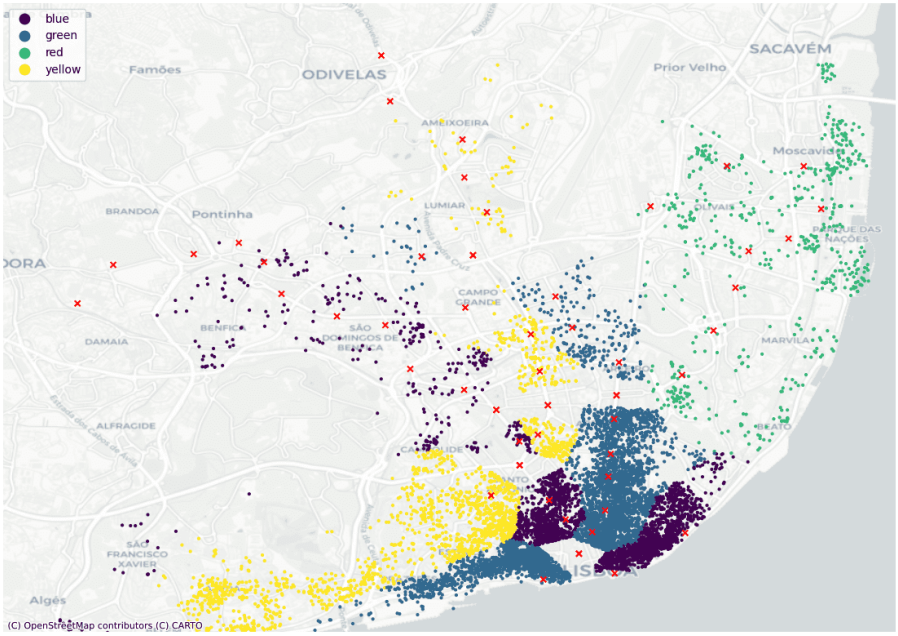


FIGURE 12: Visual Metro distribution

4.3 Bivariate Local Moran's I: Price and Transportation terms relationship

The further investigate and link the natural language processing analysis on the descriptions of Airbnb listings and the pricing, a bivariate local Moran's I is computed. The clustering provided through this statistic depicts the following cases:

- High-High values represent clusters of high prices along the distances to the stations
- Low-Low values represent clusters of low prices along the distances to the stations

Figure 13 shows how clusters of high values follow the metro lines, especially of the green and the blue ones, suggesting a positive relationship between the pricing and the proximity to a metro station. Moreover, high values in the area of Parque des Naçoes and Odivelas suggest that proximity to transportation hubs such as airport and bus terminals (Oriente) influences the pricing of short-term rentals. In line with the hypothesis that proximity to metro stations positively influences pricing, the values in the area of Estrela and Lapa show clustering of lower values, since no subway line passes through these neighborhoods.

The next section provides additional insights into the relationship between the subway system and the pricing of Airbnb in Lisbon by interpreting the metro system as a network and leveraging the analysis on its most important feature, betweenness centrality.

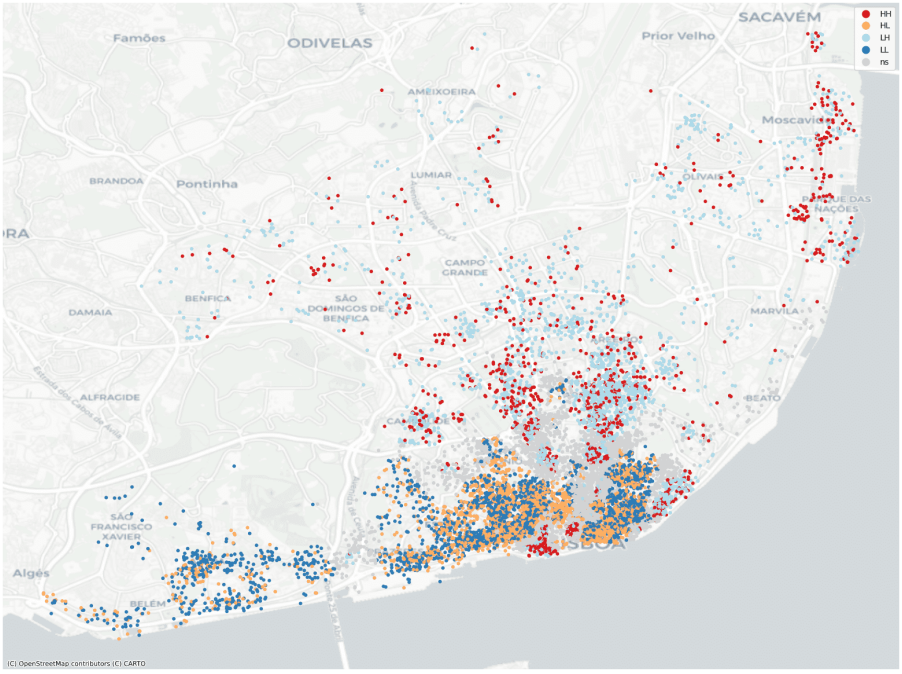


FIGURE 13: Bivariate Local Moran's I between Transportation terms frequency and pricing

4.4 Price and Transportation relation - results from topological network features

Metro network, as pointed out in section 3.2, embeds the topological features of an un-directed graph, where the nodes correspond to stations and edges to the lines connecting them, where distance can be considered the weight of each one. This section considers the most explanatory topological feature, betweenness centrality, and shows its relation with each Airbnb by taking into account the distance calculated in section 4.2. Table X contains the main centrality measures by station.

The results show an interesting phenomenon worth exploring in detail, the prices distribution seems to be concentrated where the distance from the metro stations is lowest and the betweenness centrality doesn't seem to have as much effect as stated in the literature, (Martínez and Viegas, 2009), (Dorantes et al., 2011), which affirms a strict positive of betweenness centrality and pricing.

5 CONCLUSIONS

5.1 Introduction

The main purpose of this work is to answer two strictly related research question emerging from the current Lisbon's short term rental situation, namely "How do the short-term rentals affect Lisbon's property industry" and "Which factors are promoting gentrification". In order to answer such questions the objectives posed are to 1) Identify clusters of short-term rentals in Lisbon and 2) Identify factors and topological features promoting higher prices.

The methodology of this work bases its first results on ground statistics largely adopted by the literature (i.e. Moran's I) and proceeds at exploring a new methodology, natural language processing NLP, and text analysis, to gain more insights. The hypothesis formulated by means of term frequency distribution analysis is then ultimately tested by means of the spatial statistic to understand the clustering of properties' pricing in light of the new insights.

OBJECTIVES	METHODOLOGY	RESULTS AND IMPLICATIONS
IDENTIFY CLUSTERS OF SHORT-TERM RENTALS OFFERINGS IN LISBON	GLOBAL MORAN'S I	- Confirmation of literature results: "closeness to the city center is directly linked to higher pricing." - Identify a special case: Parque das Nações, which challenges the previous assumption.
IDENTIFY FACTORS & TOPOLOGICAL FEATURES PROMOTING HIGHER PRICES	TOPIC MODELLING	- Identify a cluster of terminology related to public transportation in the eastern side of the city - Suggests that hosts highly value proximity to public transport - Result is in line with the robustness and resiliency of the transportation network Aparicio et al. (2022) - Results supported by the literature: "positive relations between real estate properties and proximity to the stations"
ADDITIONAL FINDINGS	BI-VARIATE MORANS I	- Links the results of the Local Moran's I and the Topic Modelling. - Confirms that hosts emphasizing the adjacency to public transportation tend to offer higher prices. - Confirms the presence of high prices near the transportation hubs

TABLE VI: Methods' summary

5.2 Discussion

Airbnb in Lisbon is a very complex peer-to-peer accommodation platform and the results shown in the previous section helped in defining the spatial relation behind it by means of statistics and visual representations. Figure 3 confirms literature results and assumptions that the majority of the listings can be found in the central urban areas of the city, therefore near the main points of interest, highlighting the unevenness of the distribution of pricing in the same area. To further gain insights around the specific clustering of

the values, Global Moran's statistics, with a 99.9% confidence, confirmed the presence of positive spatial auto-correlation between Airbnb offerings based on pricing, and, therefore, the presence of clusters in the study area. Local Moran's I consequently shows the presence of positive value clusters in the center of the city, specifically in the neighborhoods of Misericordia, Santa Maria Maior, and Santo Antonio. This result is expected, as shown by previous research, closeness to the city center is directly linked to higher pricing. The Local statistics presented a very surprising result, as it shows a cluster of high prices in Parque des Nações, located almost 5 km away from the city center. This finding is particularly interesting considering the results present in most of the literature, which supports the distance from the city center as having a negative influence on short-term rentals. As mentioned by Ochoa (2022), the area has been re-qualified and presents indeed high living standards.
This result is also important in the context of gentrification and it provides evidence that an increase in pricing is taking place, therefore providing regulatory agencies with insights on where to implement their monitoring rules to target sustainable goals 11.1 for a safe and affordable housing market.

The Natural Language Processing analysis shows very interesting results for what concerns means of transportation importance when a host is describing the offer. The results in Figure 9 show how terms related to transportation like 'metro', 'bus', 'station', or airport are mainly present in the eastern side of the city and, most importantly in the peripheral area, suggesting that hosts highly value the proximity of their offerings to means of transport in this side of the city. Transportation network has in fact been proven to be resilient and robust according to the findings presented by Aparicio et al. (2022). An implication on prices would be that the hosts, knowing the importance of the transportation lines for tourists to visit the city, are increasing the price based on the distance from the main means of commuting. The most emphasized transportation by the hosts is the subway. Metro system is in fact well developed in the city of Lisbon and represents one of the main means of transport used. Hosts' descriptions in the neighborhood of Areeiro and Arroios mostly stress the importance of subway transportation, aligning with the result presented in table V, restating the importance of the green line. Another area where it is observable a high frequency of Transport related terms is Olivais, the airport is very close as well as the Oriente bus station, both key hubs when arriving in and departing from Lisbon.
These findings are supported by Martínez and Viegas (2009) and Dorantes et al. (2011), in which the researchers confirm the positive relationship between the pricing of real estate properties with the accessibility to the subway system. Proximity to transportation hubs is a key insight for the regulatory agencies to take into account, as stated by Revington (2015) access to transportation infrastructure may bring about land use changes associated with gentrification and, as the study continues, proximity to transports is the sole influence on higher pricing of such houses. House amenities and other neighborhood characteristics can sometimes also promote unsafe housing conditions, a key aspect that the SDG target 11.1 aims at resolving.

To conclude the analysis, bi-variate local Moran's I is computed to statistically link the results from

Local Moran's I on prices (Figure 7) and the findings related to the Metro network with the goal of showing if there exists a correlation between the prices and the frequency of the transportation terms in each description. Figure 13 shows that relevant High values clusters are present near the transportation hubs, confirming that hosts emphasizing the adjacency to mean of commute tend to price their offerings higher. The results also show how the areas of Lapa and Estrela currently don't present high pricing values correlated with the proximity to the subway system but the Municipality of Lisbon has already started the extension of the current network to cover also these zones. In this sense and considering the results obtained in this research, a proposal for future research would be to analyze the effective change in price due to the changes in land use, promoting more accessibility to transportation.

5.3 Contributions

The review of the literature reported in section 2 emphasized the need of enriching the methods used in spatial analysis in the context of the accommodation industry. With this finding in mind, the research contributed to the current bag of methodologies by merging traditional approaches (spatial regression) with more modern techniques (NLP), demonstrating the possibility of analyzing the short term rental industry with new tools.

On top of renovating the techniques it also provides with a framework of analysis which can be scalable to other geographies where the short-term rental industry is developing, which can be useful for many parties involved both in the developement and the regulation of the phonomenon, as outlined in section 5.5.

5.4 Limitations

This work considers a limited amount of data for the analysis, which is here and in the literature explored considered enough to sample the phenomenon analyzed. Short term rentals platform are numerous and most of them have restrictions on accessing their data, therefore a portion of the market is here missing. Given that, Airbnb is still the most used short term rental platform in Lisbon[14] and occupies the majority of the current market. The choice of collecting data for a certain moment in time clearly limits the possibility of showing trends and evolution of the phenomenon, bit it allows for a better elaboration of the current situation, by focusing on the spatial dimension.

5.5 Considerations for future research

As pointed out in 5.4, this work considers a precise point in time for the analysis and it, therefore, doesn't include a time dimension in the methodology. Given the nature of the phenomenon, being it in

[14]https://www.phocuswright.com/Travel-Research/Research-Updates/2022/the-most-used-booking-platforms-for-short-term-rentals-and-what-makes-them-attractive

quick development, it would be interesting to implement these techniques on observations over time with the purpose of spotting the emergence or dissolution of clusters. Along with the expansion of the methodology to the time dimension, the approach could also be stretched to other cities with similar characteristics in Europe, so to not only spot similarities between similar phenomenon across the territory but also by raising awareness of gentrification phenomena and insights to regulate them at a higher level. That is to say not limiting intervention to single governments actions, but bringing relevance at the European Union level.

REFERENCES

Abrate, G., Sainaghi, R., and Mauri, A. G. (2022). Dynamic pricing in airbnb: Individual versus professional hosts. *Journal of Business Research*, 141:191–199.

Adamiak, C., Szyda, B., Dubownik, A., and García-Álvarez, D. (2019). Airbnb offer in spain—spatial analysis of the pattern and determinants of its distribution. *ISPRS International Journal of Geo-Information*, 8(3):155.

Anselin, L. (2019). A local indicator of multivariate spatial association: Extending geary's c. *Geographical Analysis*, 51(2):133–150.

Aparicio, J. T., Arsenio, E., Santos, F. C., and Henriques, R. (2022). Lines: multimodal transportation resilience analysis. *Sustainability*, 14(13):7891.

Basu, S. and Thibodeau, T. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, 17(1):61–85.

Celata, F., Capineri, C., and Romano, A. (2020). A room with a (re)view. short-term rentals, digital reputation and the uneven spatiality of platform-mediated tourism. *Geoforum*, 112:129–138.

Cocola-Gant, A. and Gago, A. (2021). Airbnb, buy-to-let investment and tourism-driven displacement: A case study in lisbon. *Environment and Planning A: Economy and Space*, 53(7):1671–1688.

Costa, C., Stevic, I., Veríssimo, M., and Ferreira, M. (2019). *Short-term accommodation and overtourism in Portuguese urban centres.*, pages 167–189. CABI.

Costa, C. J. and Aparicio, J. T. (2020). Post-ds: A methodology to boost data science. In *2020 15th iberian conference on information systems and technologies (CISTI)*, pages 1–6. IEEE.

Derrible, S. (2012). Network centrality of metro systems. *PLOS ONE*, 7(7):1–10.

Dorantes, L. M., Paez, A., and Vassallo, J. M. (2011). Analysis of house prices to assess economic impacts of new public transport infrastructure: Madrid metro line 12. *Transportation Research Record*, 2245(1):131–139.

Fischer, M. and Wang, J. (2011). *Spatial Data Analysis: Models, Methods and Techniques*. Springer.

Gant, A. C. (2016). Holiday rentals: The new gentrification battlefront. *Sociological Research Online*, 21(3):112–120.

Garrison, W. and Marble, D. (1962). The structure of transportation networks. *U.S. Army Transportation Command, Technical Report*, 62-II:100.

Gyódi, K. and Łukasz Nawaro (2021). Determinants of airbnb prices in european cities: A spatial econometrics approach. *Tourism Management*, 86:104319.

Hong, I. and Yoo, C. (2020). Analyzing spatial variance of airbnb pricing determinants using multiscale gwr approach. *Sustainability*, 12(11):4710.

Jover, J. and Cocola-Gant, A. (2022). The political economy of housing investment in the short-term rental market: Insights from urban portugal. *Antipode*, 55.

Lee, B. X., Kjaerulf, F., Turner, S., Cohen, L., Donnelly, P. D., Muggah, R., Davis, R., Realini, A., Kieselbach, B., MacGregor, L. S., et al. (2016). Transforming our world: implementing the 2030 agenda through sustainable development goal indicators. *Journal of public health policy*, 37:13–31.

Lestegás, I., Seixas, J., and Lois-González, R.-C. (2019). Commodifying Lisbon: A Study on the Spatial Concentration of Short-Term Rentals. *Social Sciences*, 8(2):1–15.

Martínez, L. M. and Viegas, J. M. (2009). Effects of transportation accessibility on residential property values: Hedonic price model in the lisbon, portugal, metropolitan area. *Transportation Research Record*, 2115(1):127–137.

Mendes, L. (2021). Transnational gentrification and the housing market during pandemic times, lisbon style. *Urban Geography*, 42(7):1003–1010.

Ochoa, R. (2022). The "expo" and the post-"expo": The role of public art in urban regeneration processes in the late 20th century. *Sustainability*, 14(2):985.

Oxoli, D. (2019). Exploratory approaches in spatial association analysis: Methods, complements, and open gis tools development. Ph.D. thesis.

Palm, R. (1978). Spatial segmentation of the urban housing market. *Economic Geography*, 54(3):210–221.

Peralta, S., Pereira dos Santos, J., and Gonçalves, D. (2020). Short-term rental bans and housing prices: Quasi-experimental evidence from lisbon. *SSRN Electronic Journal*.

Pereira, S. M. (2020). Regulation of short-term rentals in lisbon: strike a balance between tourism dependence and urban life. *Urban Research & Practice*, 15:477 – 504.

Revington, N. (2015). Gentrification, transit, and land use: Moving beyond neoclassical theory. *Geography Compass*, 9(3):152–163.

Santos, M., Ribeiro, R., Batista, F., and Correia, A. (2022). Examining airbnb guest satisfaction tendencies: A text mining approach. *Current Issues in Tourism*, 25(22):3607–3622.

Singh, N. and Botelho Azevedo, A. (2021). Geografia da airbnb em barcelona e lisboa:: Un estudo comparativo. *Finisterra*, 56(118):151–173.

Sun, S., Zhang, S., and Wang, X. (2021). Characteristics and influencing factors of airbnb spatial distribution in chinas rapid urbanization process: A case study of nanjing. *PLOS ONE*, 16:1–19.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240.

Wachsmuth, D. and Weisler, A. (2018). Airbnb and the rent gap: Gentrification through the sharing economy. *Environment and Planning A: Economy and Space*, 50(6):1147–1170.

Xu, F., Hu, M., la, L., Wang, J., and Huang, C. (2019). The influence of neighbourhood environment on airbnb: a geographically weighed regression analysis. *Tourism Geographies*, 22:1–18.

Zhang, Z. and Chen, R. J. (2019). Assessing airbnb logistics in cities: Geographic information system and convenience theory. *Sustainability*, 11(9):2462.

A APPENDIX

TABLE VII: Top 10 terms per neighborhood

Neighbourhood	**10 Most common words**
Belém	bathroom, bedroom, Belém, close, garden, home, house, kitchen, river, wifi
Santa Maria Maior	ama, bathroom, bedroom, center, comfortable, floor, heart, kitchen, renovated, wifi
Avenidas Novas	bathroom, bedroom, center, enjoy, kitchen, metro, modern, restaurants, rooms, wifi
Lumiar	apartamento, bathroom, bedroom, center, conchas, kitchen, metro, quinta, station, telheiras
Misericórdia	bairro, bathroom, bedroom, Chiado, comfortable, floor, heart, kitchen, note, restaurants
Estrela	bathroom, bedroom, comfortable, enjoy, Estrela, floor, home, kitchen, river, wifi
São Vicente	ama, bathroom, bedroom, enjoy, floor, Graça, kitchen, river, typical, view
Arroios	apartamento, bathroom, bedroom, center, floor, house, kitchen, metro, note, wifi
Campo de Ourique	bathroom, bedroom, Campo, Estrela, flat, home, kitchen, neighborhood, Ourique, restaurants
Santo António	avenida, bathroom, bedroom, floor, kitchen, Liberdade, metro, Pombal, real, wifi
Campolide	Amoreiras, bedroom, Campolide, center, house, kitchen, neighborhood, Pombal, shopping, well
São Domingos de Benfica	apartamento, cada, café, conforto, gratuito, kitchen, metro, modern, spacious, todo
Penha de França	apartamento, bathroom, bedroom, central, enjoy, floor, home, kitchen, metro, wifi
Parque das Nações	bathroom, bedroom, free, kitchen, modern, Nações, Oriente, Parque, river, wifi

TABLE VII: Top 10 terms per neighborhood (continued)

Neighbourhood	**10 Most common words**
Alcântara	apartamento, bathroom, bedroom, Belém, factory, home, kitchen, LX, restaurants, Alcântara
Alvalade	Alvalade, apartamento, bathroom, bedroom, bus, house, kitchen, metro, restaurants, train
Carnide	apartamento, Carnide, center, centro, Colombo, house, Luz, metro, restaurants, shopping
Areeiro	apartamento, bedroom, free, kitchen, machine, meda, metro, note, station, things
Ajuda	Ajuda, apartamento, bathroom, bedroom, Belém, flat, house, kitchen, quiet, river
Olivais	aeroporto, airport, bathroom, casa, kitchen, metro, minutos, Nações, Oriente, Parque
Beato	airport, apartamento, Beato, distance, downtown, ideal, Olaias, station, stations, subway
Marvila	apartamento, bedroom, center, cidade, comfortable, kitchen, metro, min, modern, river
Benfica	bathroom, bedroom, Benfica, bus, center, kitchen, km, quiet, station, train
Santa Clara	airport, away, beautiful, casa, cozinha, exempt, free, kitchen, km, place

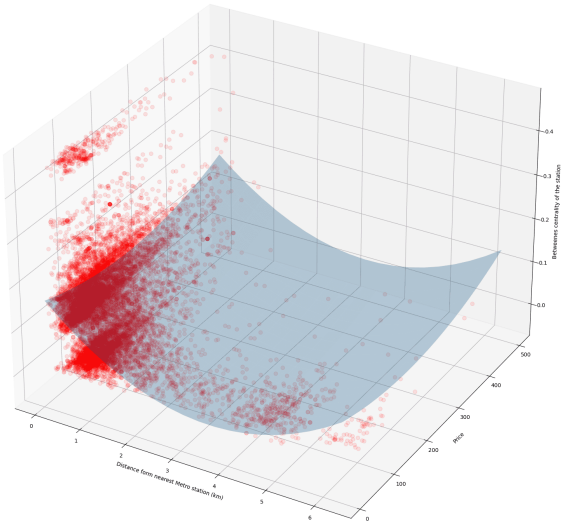


FIGURE 14: A 3-dimensional view of the relation between price, betweenness centrality, and distance to each station

	Q2 2020	Q3 2020	Q4 2020	Q1 2021	Q2 2021	Q3 2021	Q4 2021	Q12022	Q2 2022
STO.ANTONIO	5 550€	5 540€	5 544€	5 425€	5 359€	5 455€	5 435€	5 624€	€5547
MARVILA	2 863€	2 765€	2 742€	2 682€	2 924€	3 121€	3 335€	3 493€	5.297 €
STA. Ma MAIOR	5 111€	5 372€	5 914€	5 128€	4 207€	4 058€	4 157€	4 500€	€4536
PQ.NACOES	4 217€	4 076€	3 972€	4 000€	4 070€	4 335€	4 431€	4 390€	4.525 €
MISERICORDIA	4 839€	4 963€	4 754€	4 380€	4 401€	4 010€	4 375€	4 392€	4.412 €
C. OURIQUE	3 859€	3 687€	3 699€	3 553€	3 548€	3 716€	3 988€	4 128€	4.366 €
ESTRELA	4 200€	4 223€	4 223€	4 004€	3 835€	3 965€	4 061€	4 167€	4.254 €
AV. NOVAS	3 745€	3 645€	3 645€	3 621€	3 861€	4 167€	4 243€	4 182€	€4214
ALVALADE	3 634€	3 623€	3 590€	3 415€	3 704€	3 748€	3 808€	3 944€	4.103 €
BELEM	3 522€	3 577€	3 806€	3 653€	3 647€	3 743€	3 824€	3 904€	4.048 €
SAO VICENTE	3 375€	3 300€	3 470€	3 471€	3 429€	3 550€	3 409€	3 581€	3.730 €
ARROIOS	3 380€	3 538€	3 650€	3 541€	3 375€	3 315€	3 450€	3 634€	3.710 €
AREEIRO	3 321€	3 380€	3 368€	3 359€	3 395€	3 382€	3 500€	3 566€	3.655 €
CARNIDE	3 076€	3 165€	3 411€	3 416€	3 266€	3 344€	3 490€	3 583€	3.589 €
AJUDA	3 200€	3 027€	3 000€	3 129€	3 157€	3 163€	3 333€	3 485€	3.584 €
CAMPOLIDE	2 956€	3 096€	3 333€	3 110€	3 455€	3 577€	3 374€	3 421€	3.559 €
S.D. BENFICA	3 385€	3 333€	3 341€	3 307€	3 350€	3 385€	3 362€	3 469€	3.530 €
ALCANTARA	3 164€	3 128€	2 965€	2 924€	2 995€	3 092€	3 309€	3 459€	3.497 €
LUMIAR	3 040€	2 744€	2 740€	€2748	2 787€	3 136€	3 211€	3 297€	3.313 €
P. FRANCA	3 012€	2 967€	2 987€	2 849€	2 840€	3 037€	3 156€	3 136€	3.159 €
BENFICA	2 722€	2 753€	2 701€	2 668€	2 778€	2 826€	2 897€	2 993€	3.061 €
BEATO	2 461€	2 296€	2 518€	2 480€	2 490€	2 678€	2 762€	2 849€	2.878 €
OLIVAIS	2 500€	2 481€	2 527€	2 595€	2 730€	2 775€	2 805€	2 798€	€2785
STA.CLARA	2 261€	2 102€	2 102€	2 097€	2 350€	2 440€	2 501€	2 475€	2.589 €

TABLE VIII: Prices per square meters for each neighbourhood

TABLE X: Centrality measures

station	Betwenness	Eigenness	Centrality	Degree	
Reboleira	0.000	0.000	0.000	0.020	1
Amadora Este	0.041	0.000	0.000	0.041	2
Alfornelos	0.080	0.000	0.000	0.041	2
Pontinha	0.117	0.000	0.000	0.041	2
Carnide	0.153	0.000	0.000	0.041	2
Colégio Militar	0.187	0.000	0.001	0.041	2
Alto dos Moinhos	0.219	0.000	0.002	0.041	2
Laranjeiras	0.250	0.000	0.005	0.041	2
Jardim Zoológico	0.279	0.000	0.011	0.041	2
Praça de Espanha	0.306	0.000	0.021	0.041	2
São Sebastião	0.332	0.000	0.058	0.061	3
Parque	0.094	0.000	0.019	0.041	2
Marquês de Pombal	0.224	0.000	0.015	0.082	4
Avenida	0.142	0.000	0.005	0.041	2
Restauradores	0.128	0.000	0.001	0.041	2
Baixa Chiado	0.170	0.000	0.000	0.082	4
Terreiro do Paço	0.041	0.000	0.000	0.041	2
Santa Apolónia	0.000	0.000	0.000	0.020	1
Odivelas	0.000	0.000	0.009	0.020	1
Senhor Roubado	0.041	0.000	0.027	0.041	2
Ameixoeira	0.080	0.000	0.039	0.041	2
Lumiar	0.117	0.000	0.082	0.041	2
Quinta das Conchas	0.153	0.000	0.233	0.041	2
Campo Grande	0.227	0.000	0.627	0.082	4
Cidade Universitária	0.144	0.000	0.353	0.041	2
Entre Campos	0.169	0.000	0.210	0.041	2
Campo Pequeno	0.196	0.000	0.090	0.041	2
Saldanha	0.443	0.000	0.122	0.082	4
Picoas	0.137	0.000	0.032	0.041	2
Rato	0.000	0.000	0.005	0.020	1
Telheiras	0.000	0.000	0.238	0.020	1

station	Betwenness	Eigenness	Centrality	Degree	
Alvalade	0.103	0.000	0.472	0.041	2
Roma	0.129	0.000	0.168	0.041	2
Areeiro	0.155	0.000	0.105	0.041	2
Alameda	0.458	0.000	0.135	0.082	4
Arroios	0.156	0.000	0.025	0.041	2
Anjos	0.131	0.000	0.007	0.041	2
Intendente	0.106	0.000	0.001	0.041	2
Martim Moniz	0.093	0.000	0.000	0.041	2
Rossio	0.091	0.000	0.000	0.041	2
Cais do Sodré	0.000	0.000	0.000	0.020	1
Aeroporto	0.000	0.000	0.000	0.020	1
Encarnação	0.041	0.000	0.001	0.041	2
Moscavide	0.080	0.000	0.001	0.041	2
Oriente	0.117	0.000	0.001	0.041	2
Cabo Ruivo	0.153	0.000	0.002	0.041	2
Olivais	0.187	0.000	0.005	0.041	2
Chelas	0.219	0.000	0.015	0.041	2
Bela Vista	0.250	0.000	0.039	0.041	2
Olaias	0.279	0.000	0.084	0.041	2

TABLE IX: Correlation scores by parish of price and location review scores

Correlation score	**Parish**
-0.0	Areeiro
-0.006	Penha da França
-0.017	So Domingos de Benfica
-0.019	Campo de Ourique
-0.028	Belm
-0.039	Ajuda
-0.05	Arroios
-0.058	Beato
-0.09	Alvalade
-0.106	Canmpolide
-0.115	Lumiar
-0.649	Carnide
0.002	Santa Clara
0.016	Olivais
0.021	São Vicente
0.045	Parque das Naçoes
0.051	Santo Antonio
0.064	Alcantara
0.066	Misericordia
0.070	Avenidas Novas
0.088	Estrela
0.093	Marvila
0.145	Sanat Maria Maior
0.148	Benfica

Parishes	% Area terms	% services terms	% services terms	% house facility terms
Belm	0.0 %	20.0 %	0.0 %	60.0 %
SantaMariaMaior	0.0 %	30.0 %	0.0 %	60.0 %
AvenidasNovas	10.0 %	20.0 %	10.0 %	50.0 %
Lumiar	20.0 %	20.0 %	10.0 %	30.0 %
Misericrdia	0.0 %	50.0 %	10.0 %	40.0 %
Estrela	0.0 %	30.0 %	10.0 %	50.0 %
SoVicente	0.0 %	20.0 %	0.0 %	60.0 %
Arroios	20.0 %	20.0 %	0.0 %	40.0 %
CampodeOurique	0.0 %	10.0 %	10.0 %	50.0 %
SantoAntnio	10.0 %	40.0 %	0.0 %	40.0 %
Campolide	10.0 %	40.0 %	10.0 %	30.0 %
SoDomingosdeBenfica	10.0 %	0.0 %	10.0 %	70.0 %
PenhadeFrana	10.0 %	10.0 %	10.0 %	50.0 %
ParquedasNaes	10.0 %	30.0 %	0.0 %	40.0 %
Alcntara	0.0 %	30.0 %	10.0 %	40.0 %
Alvalade	30.0 %	0.0 %	10.0 %	40.0 %
Carnide	10.0 %	40.0 %	40.0 %	0.0 %
Areeiro	20.0 %	0.0 %	0.0 %	50.0 %
Ajuda	0.0 %	40.0 %	0.0 %	40.0 %
Olivais	40.0 %	20.0 %	0.0 %	30.0 %
Beato	50.0 %	20.0 %	0.0 %	0.0 %
Marvila	10.0 %	40.0 %	0.0 %	40.0 %
Benfica	30.0 %	20.0 %	0.0 %	50.0 %
SantaClara	30.0 %	10.0 %	0.0 %	30.0 %

TABLE XI: Percentage of terms for each topic