Masters Program in Geospatial Technologies

Site Selection Using Geo-Social Media

A Study For Eateries In Lisbon

Jaskaran Singh Puri

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

Site Selection Using Geo-Social Media
A Study For Eateries In Lisbon

Dissertation supervised by:

Prof. Dr. Marco Octávio Trindade Painho, PhD
NOVA Information Management School
Lisbon, Portugal

Co-supervised by:

Vicente De Azevedo Tang
NOVA Information Management School
Lisbon, Portugal

Prof. Dr. Sven Casteley, PhD
Universitat Jaume I Castelló, Spain

February, 2023

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, 20th February 2023
Jaskaran Singh Puri

[the signed original has been archived by the NOVA IMS services]

ACKNOWLEDGEMENTS

I feel fortunate enough to have been selected for this program of MSc in Geospatial Technologies. The program has offered me incredible opportunities for personal and academic growth. The program has provided me with fresh research perspectives, a high-quality education, and the chance to connect with amazing individuals who have enriched my experience.

I sincerely thank all the individuals and organizations who have contributed to my academic journey. Foremost, I am deeply grateful to Prof. Dr. Marco Painho for his invaluable advice and unwavering support throughout my thesis research and master's studies. I'd also like to express my gratitude to Prof. Dr. Sven Casteleyn and Prof. Vicente De Azevedo Tang for the amazing support, insights, and suggestions I received throughout my research journey.

I want to thank the academic staff at NOVA IMS and IFGI who have encouraged me to develop new skills, expand my knowledge, and achieve new milestones.

This thesis is one of the outputs of a funded project by the Fundacao para a Ciencia e a Tecnologia (FCT). The project is named as "CityMe" with the Project Reference (EXPL/GES-URB/1429/2021). I would like to express my gratitude to Prof. Dr. Marco Painho and Prof. Vicente Tang for not only granting me this invaluable scholarship opportunity, but also for their exceptional support throughout the entire duration of the course

SITE SELECTION USING GEO-SOCIAL MEDIA: A STUDY FOR EATERIES IN LISBON

ABSTRACT

The rise in the influx of multicultural societies, studentification, and overall population growth has positively impacted the local economy of eateries in Lisbon, Portugal. However, this has also increased retail competition, especially in tourism. The overall increase in multicultural societies has also led to an increase in multiple smaller hotspots of human-urban attraction, making the concept of just one downtown in the city a little vague. These transformations of urban cities pose a big challenge for upcoming retail and eateries store owners in finding the most optimal location to set up their shops. An optimal site selection strategy should recommend new locations that can maximize the revenues of a business. Unfortunately, with dynamically changing human-urban interactions, traditional methods like relying on census data or surveys to understand neighborhoods and their impact on businesses are no more reliable or scalable. This study aims to address this gap by using geo-social data extracted from social media platforms like Twitter, Flickr, Instagram, and Google Maps, which then acts as a proxy to the real population. Seven variables are engineered at a neighborhood level using this data: business interest, age, gender, spatial competition, spatial proximity to stores, homogeneous neighborhoods, and percentage of the native population. A Random Forest based binary classification method is then used to predict whether a Point of Interest (POI) can be a part of any neighborhood *n.* The results show that using only these 7 variables, an F1-Score of 83% can be achieved in classifying whether a neighborhood is good for an “eateries” POI. The methodology used in this research is made to work with open data and be generic and reproducible to any city worldwide.

KEYWORDS

Site Selection
Geographic Self-Organizing Maps
Geospatial Analysis
Retail
Social Media

ACRONYMS

POI	Point of Interest
GIS	Geographic Information System
GeoSOM	Geographic Self-Organizing Maps
AOI	Area of Interest
ARIMA	Autoregressive Integrated Moving Average
GDP	Gross Domestic Product
BERT	Bidirectional Encoder Representations from Transformers
NMF	Non-negative matrix factorization
LDA	Latent Dirichlet Allocation
MAUP	Modifiable Aerial Unit Problem
SCI	Spatial Competition Index
PPH	Posts Per Hexagon
UMAP	Uniform Manifold Approximation and Projection
HDBSCAN	Hierarchical Density Based Spatial Clustering
TM	Topic Modelling
BMU	Best Matching Unit
PCP	Principal Component Plane
RF	Random Forest

Table of Contents

Table of Figures

Index of Tables

1. INTRODUCTION

Portugal's capital, Lisbon, has seen significant economic growth in the last decade. The hilly coastal city is known for its multicultural integration and history, further complemented by some of the most beautiful landscapes in Europe. This recent surge in the internationalization of Lisbon has occurred mainly through studentification **[32]**, also referred to as a new class of transnational urban consumers, and the changes in fiscal policy and immigration rules **[42]** of Portugal, of which the Golder Visa is a prime example. This influx of multicultural populations has positively impacted the local economy. The food and drink industry, such as restaurants and cafes, has also been positively impacted while increasing retail competition for new owners.

At a granular scale, the diverseness has also given birth to changes in urban-consumer interactions **[7]**, which aims to study the relations between consumer demands and the urban retail market. In spatial terms, due to the emergence of different micro-segments of culture and ever-growing real-estate pricing, there has been a rise in retail fragmentation, weakening the concept of a common downtown in the city. This ambiguity of downtown has led to a huge problem for retail managers regarding the optimal location for their businesses in the city. For instance, a local retail store generally has its consumers as local residents, while a hotel does not **[12]**. This makes it crucial for hotel managers/owners to find the optimal spot in a neighborhood that serves mostly foreigners. A similar problem is faced by food/drink store owners as there is no clearly defined neighborhood in a city that can be targeted as an optimal location.

This research discusses a site selection strategy for eateries in Lisbon, specifically restaurants, cafes, and bars, as this industry now faces huge competition to serve the tourist sector. The restaurant's location has been consistently ranked as one of the most influential factors in its success **[43]**. The site selection for the food/drinks industry is rather more sophisticated because the food itself is analogous to cultural or ethnic differences. In some cases, having cultural knowledge about the food and the neighborhood can be a major competitive advantage. Accessibility, sociodemographics, and existing competition have also been repeatedly identified that can hinder eateries' success. As per **[27]**, an optimal location of a store should increase the footfall of consumers, thereby increasing profits.

It, therefore, becomes essential for the managers of restaurants/cafes/bars to consider their intended target audience, neighborhood, accessibility, and also proximity to tourist locations, especially in cities like Lisbon, before setting up a new place. Such ambiguities and challenges motivate a study about site selection for the city of Lisbon to understand the relationship between how different demographic groups in other regions interact with everyday businesses

and how this information can be fed back to retail managers for enhanced decision-making in future development. Several approaches, like machine learning **[13],** can address the complexity of factors influencing site selection and suggests using Convolutional Neural Networks as Point of Interest (POI) (location of a restaurant/bar/cafe outlet) prediction algorithm.

The geographic information systems (GIS) technology and its ability to integrate both spatial and non-spatial data have allowed several similar types of research like this to be implemented practically and at large scales, for instance, how GIS can help in understanding spatial phenomena using empirical research **[14]** methods. GIS has earned its reputation for analyzing and finding trends/patterns in spatial data. By combining demographic, economic, and geosocial data, it is possible to improve decision-making for site selection. Geosocial data, in particular, is about collecting data from social media platforms like Twitter, Instagram, and Flickr that come with geographic attributes. Such data can be essential for inferring the opinions and behaviors of users for a particular region. Collecting such insights can help business managers predict consumer behavior and make more informed decisions for site selection.

The research for site selection in Lisbon needs to be made available, while some studies have been conducted on cities in China **[20, 25, 27, 13]**. However, these studies do not consider the additional ever-changing factors of the city's socioeconomic and demographic landscape. Additionally, most existing studies focus on the quantitative analysis of spatial data. They do not consider the qualitative aspects of human-urban interactions and the impact of social media on consumer behavior.

This research aims to fill this gap by studying the use of geosocial data in the site selection of retail stores in Lisbon. Focus is given to using multidimensional data at a grid (hexagonal) level and understanding how these variables might influence the success of retail outlets in the city. Additionally, the availability of social media data will enable the examination of behavioral patterns in different sociodemographic groups. The significant advantage of using geosocial data is that it eliminates the need to rely entirely on census data to understand neighborhoods; in fact, it further enriches census data as it can be updated much more frequently. The highly generic data but the specific methodology for most of the previous research is also further improved in this paper as the previously mentioned methodologies did not focus on specific retail outlets like "cafes" or "bars." This approach can be improved by filtering the data sources for specific businesses while keeping the methodology generic, thus allowing a more personalized flow of spatial recommendations for each business.

Finally, by using sophisticated spatial clustering algorithms like Geographic Self-Organizing Maps (GeoSOM), it is possible to extract meaningful regions and use this information to enhance or develop new POI prediction algorithms.

1.2 Research Scope and Objectives

The research will focus on two specific objectives, thus, answering two research questions by the end of this work:

1. Can neighborhood features, such as extracting interests, POI reviews, competitions, and demographics information of a region using only social media as the primary data source, be used for predicting POI locations? Answering this would allow examining the relationship between these neighborhood features and the location of POIs in Lisbon.

2. Does a spatial clustering (GeoSOM) approach that can summarize homogeneous regions provide an advantage for predicting POI locations? The research will investigate the use of spatial clustering to group similar areas based on their neighborhood features and how important this information is.

Answering these two research questions will provide valuable insights for retail managers in making better decisions for site selection since the insights come from readily available datasets that analysts can take advantage of to enhance existing methods and practices. It would be possible to view the city from different perspectives by observing maps of different variables while allowing the flexibility to hand-pick variables that suit a specific business. This research has been constructed to be reproducible and generalized for any city in the world.

2. LITERATURE REVIEW

Emerging economies encourage the growth of modern lifestyles, which is highly influenced by having a wide variety of retail outlets across a particular region. Site selection of retail outlets is a spatial problem that can be answered using a wide variety of analytical pathways. Due to the complex ecosystem of society and economics, for example, cultural aspects, market potential, and consumer characteristics pointing out specific locations that can potentially maximize profits for business owners is not a straightforward task **[27]**. An effective site selection strategy can increase sales, reduce local competition, and comfort nearby residents. Moreover, it results in high profits **[27]** and promotes a virtuous circle of the economy. For instance, in previous studies for site-selection search, distance, population density, income levels, and traffic networks were some of the major factors influencing retail store success **[27]**. Another study in China **[20]** highlighted how regional population, road length, and the number of POis were significantly and positively correlated with consumption potentiality. Literature shows that it is still crucial to further study the factors in the context of geographical space and how they can influence people's behavior **[29]** or how sociodemographics concerning spatial location can influence the success of retail shops. This literature review brings forward previous research in the context of human-urban interactions, cities, social media, and geographical factors and observing how all of them can influence the market potential of different types of retail outlets.

Defining regions can help the governments personalize the administration services for certain regions and observe the socioeconomics of various regions. On the contrary, defining a region can also introduce ambiguity in terms of the significance of the region. For instance, a specific neighborhood can be a good viewing spot for a landscape with a huge inflow of people. However, it does not necessarily mean it becomes an ideal spot to put up a local retail outlet, as the nature of human-urban interaction for this particular region can be said to be touristic. A closed polygon can usually represent such an area on a map. The problem at this point is that regions in the real world are based on perceptions. A region can be known for tourism and as an educational hub, highlighting that human interactions in certain areas vary greatly **[10]**.

2.1 Geo-Social Data As A Proxy

The significance of understanding human-urban interactions is understood at this point; however, the spatial extent these interactions hold is still a gray area. By defining an Area of Interest (AOI) using a polygon, we would define a spatial phenomenon as discrete, which is always continuous in reality **[15]**. To study different regions in a city, we can use surveys or

street view images and understand POIs and their interaction with human activities up to some extent. However, they do not define the human interactions of a region or AOI **[39]**.

Over the last decade, social media has become a part of our daily routine. Digital platforms like Twitter, Flickr, and Instagram allow people to share their thoughts and opinions without filters and encourage free speech. Interestingly, when people post something on these platforms and tag it with a geo-location, they express the interactions they might have had with the specific place. Social media data, when geotagged, can be expressed as geosocial data. This kind of data has been used as a proxy to understand people's perceptions; for instance, **[16]** explains how social media contains information about people's behavior in geographic space and how AOIs keep changing over time and cannot be static. In site selection, such data can provide invaluable information on intent signals.

Check-in data can provide linguistic information with spatial attributes **[39]**; for example, as mentioned in the literature, “nice food” might depict a restaurant nearby to the geolocation of the post. Along the same lines, **[25]** discussed how “map-query” data could be used to estimate demand in a specific region and serve as a powerful feature for site selection. Similarly, by predicting the check-in numbers at given locations using a linear supervised learning model, the authors could recommend potential sites for retail stores **[40]**. Some authors used geosocial data to extract significantly more complex information on the demographic characteristics of the users. Demographic information in terms of spatial context can be described as geodemographics. Geodemographics is a classification of areas summarized by indicators of economic, social, and demographic characteristics **[18]**. To name some of the research work, **[19]** inferred the political orientation, gender, and ethnicity of Twitter users using a machine-learning approach. Similarly, the Latent attribute inference method was used to infer the age, gender, and political affiliation of Twitter users **[26]**.

2.2 Cities: A Social Media Perspective

City image, in general, refers to the perceptions, feelings, and opinions of individuals in and about different places in a city, which are extremely important for urban management, urban planning, urban cultural perceptions, and tourism resource development **[2]**. The literature has shown how social media is a great data source for investigating human interactions within the city, which is critical to infer user characteristics and support site selection. Geosocial data does provide high-resolution data, in the sense of granularity of insights; however, it is also important to understand how the city behaves as a whole to account for the fact that cities can be very different in terms of not just people but also geographical characteristics as well. One example was provided in **[23],** where it was proposed that

landmarks, edges, nodes, paths, and districts can be used to define a city. The "image" of the city can have hundreds of "dimensions," and all these dimensions can be projected onto geographic space. Geosocial media is a spatial proxy for several of these dimensions. For instance, thousands of geotagged photos were used to compare the differences in security, social hierarchy, and uniqueness between New York, Boston, Linz, and Salzburg **[1]**. When used in site selection, understanding social media's spatial, temporal, topical, demographic, and contextual features can help us infer how these interactions are distributed over space-time and find similar regions across the city **[11]**.

Research has been produced to determine how geographical variables influence retail sales, especially in recent years. For instance, when studying individual retail shops, traditional methods like Autoregressive Integrated Moving Average (ARIMA) were used to predict sales based on historical performance. These methods do not consider spatial unevenness and patterns in the context of consumer behavior **[22]**. On the other hand, to explore how economics plays a role in site selection, the authors **[20]** discussed how macro-scale and micro-scale determinants impact retail sales differently. Macro factors are generally extracted at a state or national level from static or not-so-frequently updated data sources like gross domestic product (GDP) or socioeconomic records. However, these factors do not provide meaningful information for retail businesses **[20]**. On the contrary, due to their higher resolution, micro-scale factors tend to be more insightful. This study also pointed out that retail sales are geographically correlated with nearby socioeconomic factors; the mechanism is, however, unclear. Other micro-scale factors like spatial proximity between retail shops influence the competition among each other and eventually affect the purchase behavior. Using Sina Weibo check-in data, **[27]** and **[13]** evaluate the competition of each shop by considering the distance and number of check-in records located in nearby regions.

2.3 Demographic & Topic Extraction from Social Media Data

The rise of social media usage over the last decade has enabled us to identify and characterize regions based on geographic coordinates attached to users' activities, also known as geotagging. In the past ten years, text mining of social media data has contributed to research on politics, marketing, public health, urban planning, transportation, and education **[35]**. This ultra-high resolution of data also brings along much noise as some regions can have various topics, making it harder to infer what a region might be primarily known for. In such scenarios, it makes sense to multiple aggregate posts together and extracts the most frequently occurring words. Although this is a complex task in itself, thanks to extensive research done in the field of natural language processing, it is now possible for algorithms to understand the semantic similarity of different sentences and assign common topics/keywords

to them **[31]**. To uncover common themes and text narratives, topic models have proven to be a powerful, unsupervised tool. They are the first step in understanding the public thoughts/opinion about various commercial stores.

One such algorithm is the Bidirectional Encoder Representations from Transformers (BERT) **[28],** which can understand the semantic context of words across multiple languages. In spatial awareness, a BERT-based Spatiotemporal neural network was developed for taxi demands within a city by observing historical trip data and further enriching the data with POI information **[33]**. Similarly, a BERT-based neural network model outperformed other natural language models without removing stopwords or similar text-cleaning processing for tweet classification **[30]**. Other traditional algorithms, the Latent Dirichlet allocation (LDA) and Non-negative matrix factorization (NMF), are also extensively used, but they disregard the semantic relationship among words **[21]**. For sparse and noisy data, as is a common case in social media posts, algorithms like LDA have been proven to be not so effective **[36]** in extracting meaningful semantic relations due to the lack of statistical features **[37]** and also because they give less weightage to co-occurring relations **[34]**. On the other hand, NMF is based on the linear algebra **[24]** approach to extracting topics.

Alongside people's opinions regarding a place, it is also important to realize who these people are and if there exists a relation between the demographics of people and their perceptions of a place. Demographic characteristics like political ideology, location, income, and gender **[35]** have been previously extracted from social media data. Additionally, the metadata of these social media users can help us infer the personal characteristics of people. There are several ways to infer users' demographics on social media, like computer vision and natural language processing of names/posts. For example, Analyzing first and last names with a character-based recurrent neural network, **[3]** achieved 73% accuracy in classifying 13 ethnicities. Profile pictures can help us infer the age/gender of users using machine learning and computer vision technologies. Deep convolutional neural networks have also achieved 95% accuracy in image-based gender classification **[4]**. Based on these previous findings, access to this information makes it all the more important to look at regions and their participants at a much finer resolution. The open accessibility of data and algorithms has made it possible to use this information and provide detailed insights for site-selection problems.

Although spatial phenomena are always continuous, in this case of point-based data, where users express their opinion regarding a particular place, such locations are always fixed. Since administrative regions are generally at a district/state level, the insights extracted from such regions are coarse in resolution and beyond the range of POI's scope of business. In spatial analysis, a smaller aerial unit for study increases the likelihood of getting the most precise outputs **[38]**. This concept is, however, also linked to the well-known Modifiable Aerial Unit

Problem (MAUP) **[41]**. As this research completely depends on people's social media usage in a particular city, the spatial units chosen can get biased by the distribution of such users; for example, certain spatial units might be very sparse while others would be highly dense. It is hence important to determine an ideal spatial unit for the research.

Aggregation of messages can lead to scale and zoning effects of MAUP. In this research, a hexagonal grid is used as the spatial unit. In the geometrical sense, squares, triangles, and hexagons are three shapes that can be superimposed over a space, covering it entirely without any breaks. In this context of spatial analysis, the ideal shape would be something with a low edge-to-zone proportion which a circle can achieve. Unfortunately, circles, when put together, would leave gaps while covering the space and cannot shape a nonstop matrix **[38]**. As a result, hexagons are the most rounded polygons available and tend to highlight any patterns that might exist along their curvature. Because of these properties, hexagonal fishnets are widely used nowadays, for instance, in vegetation mapping **[5]** and in understanding territorial control **[8]**.

2.4 Retail Optimization Using Site Selection

Naturally, site selection is a ranking problem that aims to find the most favorable locations for a retail outlet. The lack of availability of socioeconomic variables **[27]** on smaller scales with no reliable source of historical knowledge can lead to poor decisions for retail owners in the long term. Additionally, no set standards or organizations define rules while evaluating potential sites for commercial usage. Proposed in the 1900s, **[13]** the spatial interaction theory was introduced to highlight the importance of attractiveness and distance of consumers to a city's downtown. This theory was extended to observe people's movement patterns toward certain commercial districts. In a study by Reily **[6]**, it was concluded that the attraction of a city to consumers in its surrounding areas is positively correlated with the city's population size and negatively correlated with the spatial distance between customers and the city. In the past, when social media data was available in its early years, people used to rely on traditional methods like statistical or mathematical models **[17]** or conducting surveys and sometimes even elementary calculations or past experiences **[29]**. Since these methods were not scalable, they represented the opinions of a comparatively small sample of the population to what we can get from social media and were limited to a few variables; they sooner or later became ineffective.

To counter this, business consulting firms started to evolve in the markets where they would conduct large-scale surveys on online and offline platforms **[25]**. The data collected was further enriched by census and demographic data from the government. This paper addresses the gaps in such data collection since most traditional data sources cannot be updated

regularly or in an automated fashion. None of the abovementioned data collection methods discuss neighborhood profiling variables like sociodemographics or the most frequently discussed topics. In recent years, authors have been able to experiment with social media data with the possibility of machine learning to evaluate potential sites.

For instance, using K-means clustering on sales data to determine hot spots in the city and estimate demand locations **[22]**. A Neural-Network CNN approach was used to recommend retail sites using population density data, web check-ins, spatial competition, and road network (for retail distance). All the layers were supplied in raster form or images with target data in the form of sales **[13]**. A new spatial clustering algorithm GeoSOM **[11],** was used to detect spatiotemporal and semantic clusters of geo-tagged tweets to characterize areas of Greater London. The authors extended their work by using the clustering outputs as an extra feature to build a classification model to detect new POI locations.

3. METHODOLOGY

This section provides a detailed walkthrough of the various stages of implementing the methodology. Figure 1.1 shows an overview flowchart of how the different steps would come together to generate a map of the new POI locations for the study area of Lisbon, Portugal.

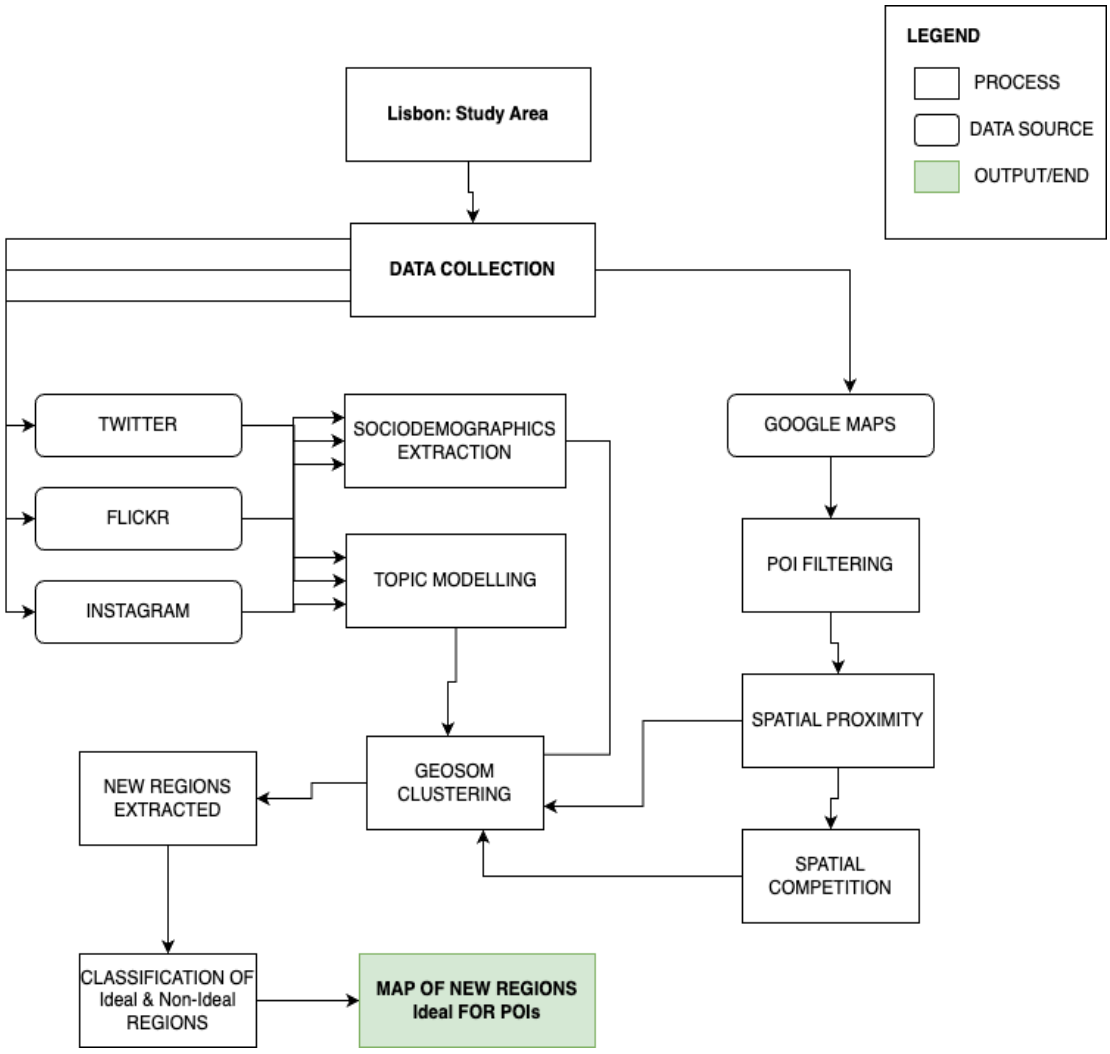


Figure 1.1 Flowchart for the overview of the complete methodology

3.1 Study Area

Situated in the westernmost region of Portugal, the capital city of Lisbon **[45]** is also the largest in Portugal. Like most cities, Lisbon is divided into smaller administrative units or parishes. There are currently 24 known parishes in Lisbon. Spread across 100 km^2 along the Tagus river; the city is an ideal destination for tourists worldwide and an important economic region for the country itself. Lisbon, being a capital city, also experiences a highly heterogeneous

group of cultures across various age groups and ethnicities. Figure 1.2 also shows the extent of the study area in the topmost image, while the bottom image shows the 24 parishes.

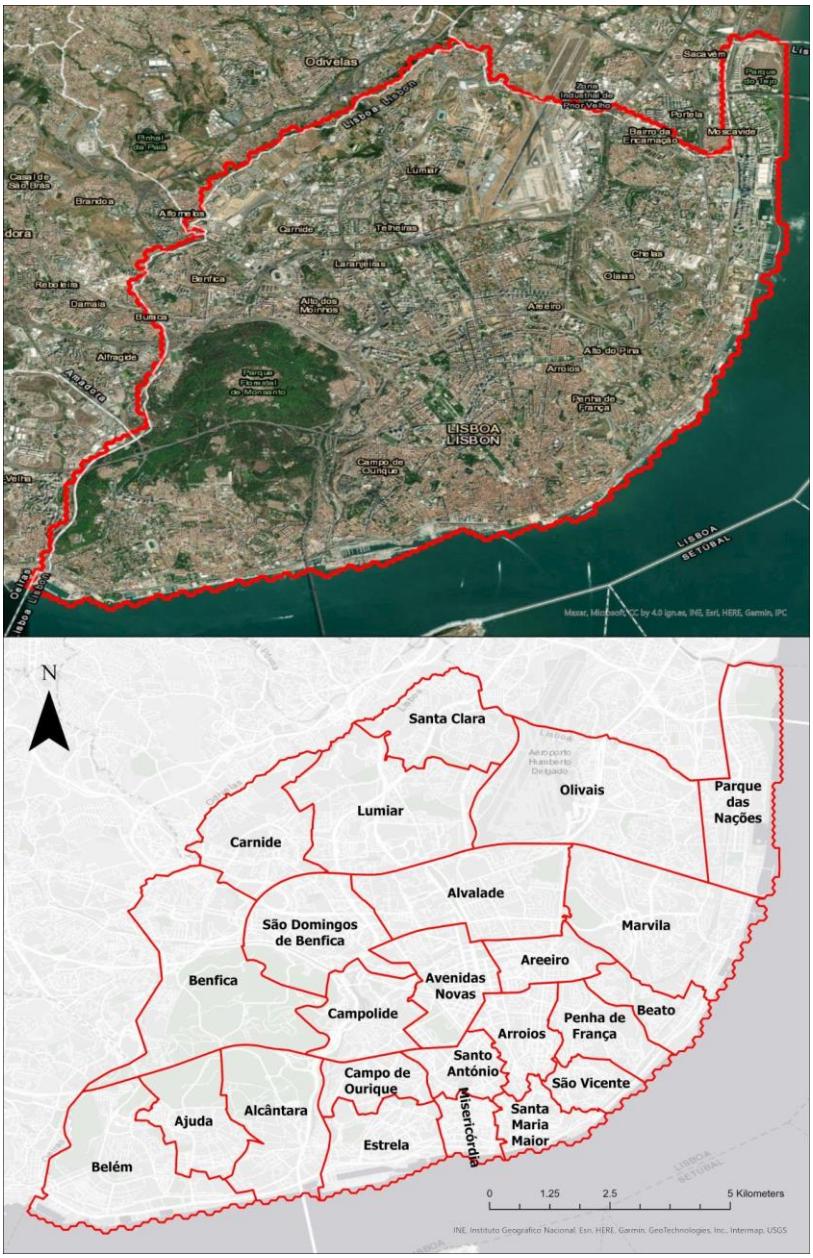


Figure 1.2 Study area extent of Lisbon. The top image shows the administrative boundary of Lisbon. The bottom image shows the 24 parishes.

Multilingual, highly populated, and high tourist inflow are some factors that make Lisbon a good study area for this research work. The city offers its people a good infrastructure for transport, education, and commercial services and hence becomes a hot spot for businesses across various industries. This research aims to observe the spatial patterns of a particular business category, namely “food & drink places” across various parishes/neighborhoods, and

to allow intelligent decision-making for retail managers when considering opening a similar business outlet in the city.

3.2 Data Collection & Processing

This research aims to not only extract data using tools published under the open source license but can also be used with our four sources of data points, ie. Twitter, Flickr, Google Maps, and Instagram. These tools have been developed previously by developers/researchers from the open-source community and are written using the Python programming language. It is also observed that the information of social media users used in this research work has been extracted only for the users who willfully kept their profile public (at the time of data collection). Furthermore, these tools provide highly flexible APIs to connect to social media websites, thus, allowing us to enable different kinds of filters, especially by extracting only those posts with geolocation properties attached to them. This section explains the data extraction methodology in detail for each of the four platforms.

3.2.1 Twitter

A famous microblogging platform that allows its users to share real-time updates and express opinions regarding everything and anything. It also allows users to share their geolocation to associate an event with a place, making it a platform for amplifying the voices of people and hence making it a primary data source for extracting people's thoughts and behavior regarding different retail shops.

The Twitter API requires separate API keys affiliated with a specific Twitter account. An open-source python library, **search-tweets-python,** was used to connect to the official Twitter website. The library provides an extensive set of filters that can be used to filter/extract specific archives of the Twitter database. Following is an example of a python command and different filters set up to extract the tweets.

python search-tweets-python/scripts/search_tweets.py --credential-file creds.yaml --max-pages 1 --max-tweets 5000000 --output-format a --results-per-call 100 ***--query "(bounding_box:[-9.237994149198911 38.68001599304764 -9.088636579438045 38.801883026428158] has:geo)"*** *--start-time "2017-01-01T00:00" --end-time "2022-04-06T22:59" --tweet-fields author_id,created_at,geo,id,source,text --place-fields country,geo,name,country_code,full_name --user-fields id,name,username,profile_image_url,location --expansions geo.place_id --filename-prefix thesis_twitter_data --no-print-stream --debug --results-per-file 2500*

The critical part of the query has been highlighted in bold which has two components in itself, *boundg_box* and *has:geo.* The *bounding_box* parameter expects a list of 4 coordinates in the order **west_longitude, south_latitude, east_longitude, and north_latitude.** The range of latitude and longitude should be ± 90 and ± 180, respectively. The *has:geo* component acts as a filter to exclude posts with no geolocation attributes attached. In this case, a bounding box over the Lisbon area was drawn in QGIS, and the resulting bounding box coordinates were passed in this previous query. The data was later clipped to the study area. Figure 1.3 shows the bounding box in blue used for data extraction and the actual study area (in black) for comparison.

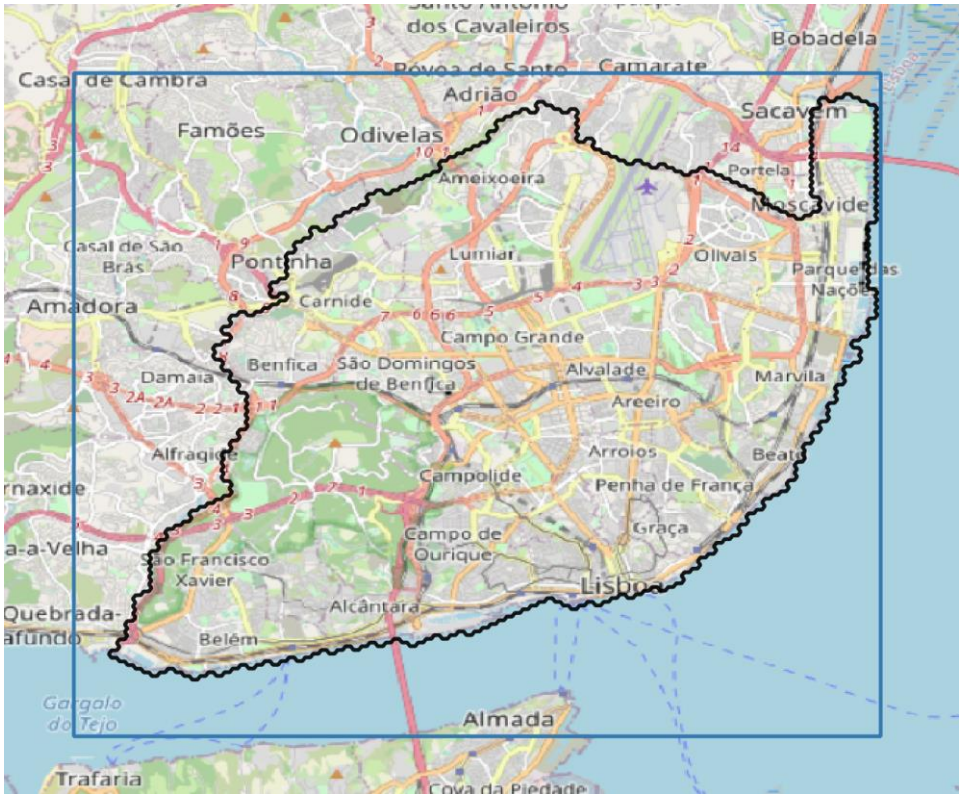


Figure 1.3. The bounding box for Twitter data extraction (in blue) and actual study area (in black)

Between May 2010 - February 2022, a total of 4.5 million (2) tweets were extracted and stored in a Postgres database. Total unique users numbered 255.200. The spatial distribution of Twitter posts can be observed in the bottom right corner of Figure 1.5, labeled as 4.

3.2.2 Flickr

Similar to the Twitter API, Flickr API also issues separate authentication API keys associated with one's account. The intention of using Flickr data points is similar to that of Instagram, although Flickr is used more professionally than Instagram's casual use. Nevertheless, Flickr

also allows the posting of images using geotags, indirectly allowing us to interpret what people think/perceive about a specific place based on the description they provide while posting.

FlickrAPI Python library was used to connect to Flickr and extract the data. The API, similar to Twitter, also allows geolocation-based search; however, in this case, the spatial extent is defined from a point's radius. In this case, a grid network of equally spaced points (in blue) 500 meters apart, as shown in Figure 1.4, was overlaid on Lisbon. The radius, in this case, for each was set to 250 meters, and the API was used to query Flickr for each of the 600 points.

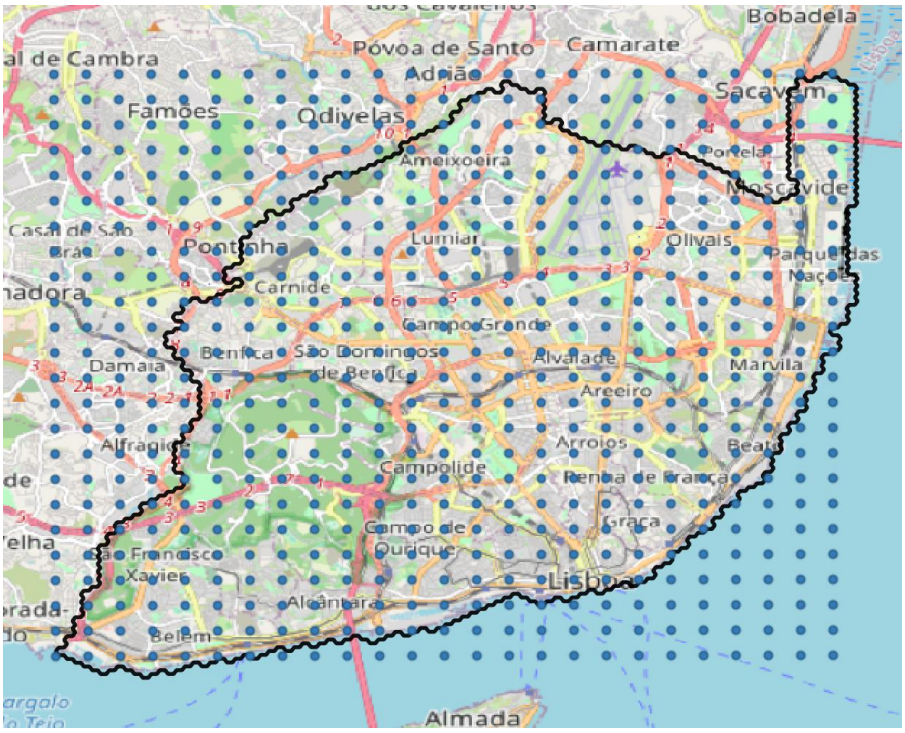


Figure 1.4 An equally spaced grid of points, 500 meters apart, is overlaid on the study area (in black).

From the early 2000s until 2022, 44.509 data points were retrieved with a unique user base of 3.063. The spatial distribution of Flickr posts can be observed in the top right corner of Figure 1.5, labeled as 2.

3.2.3 Instagram

The mobile-based application launched in the early 2010s crossed over a million users by 2012 and has now become the most common and convenient way to share your thoughts about a place/thing over the internet. From teenagers to young adults, the application has a diversified user base across age groups and genders. Like Flickr, anyone can publish geo-tagged images with a description on this app. A python-based library, **instagrapi** enabled us to extract data from public profiles. In this case, only a post's geolocation and descriptive part

are fetched alongside user information. The spatial extent for extraction was similar to that used for Flickr.

From 2021 to 2022, 35.657 posts were extracted with 5.912 unique users. The spatial distribution of Instagram posts can be observed in the top left corner of Figure 1.5, labeled as 1. Due to Instagram's privacy updates, it does not allow the extraction of data beyond one year from the data of data collection.

3.2.4 Google

The Google Maps API provides the essential data points for the research work, i.e., POIs. Using Google Maps also ensures that the latest data is used for analysis. Additionally, due to the high popularity of Google Maps, this data verifies being high-quality and regularly updated. Maps API is a vast platform to extract different map-based data features; however, only the POIs published on Google Maps are relevant for this research. Like the Flickr API, the Google Maps API also expects a point and a radius (in meters) as input. The overlaid Lisbon grid used in Flickr API was also used for POI extraction with a radius of 250 meters. A total of 19.660 POIs were extracted, and the attributes extracted were coordinates, names, ratings, reviews, and categories. The spatial distribution of Google POIs can be observed in the bottom left corner of Figure 1.5, labeled as 3.

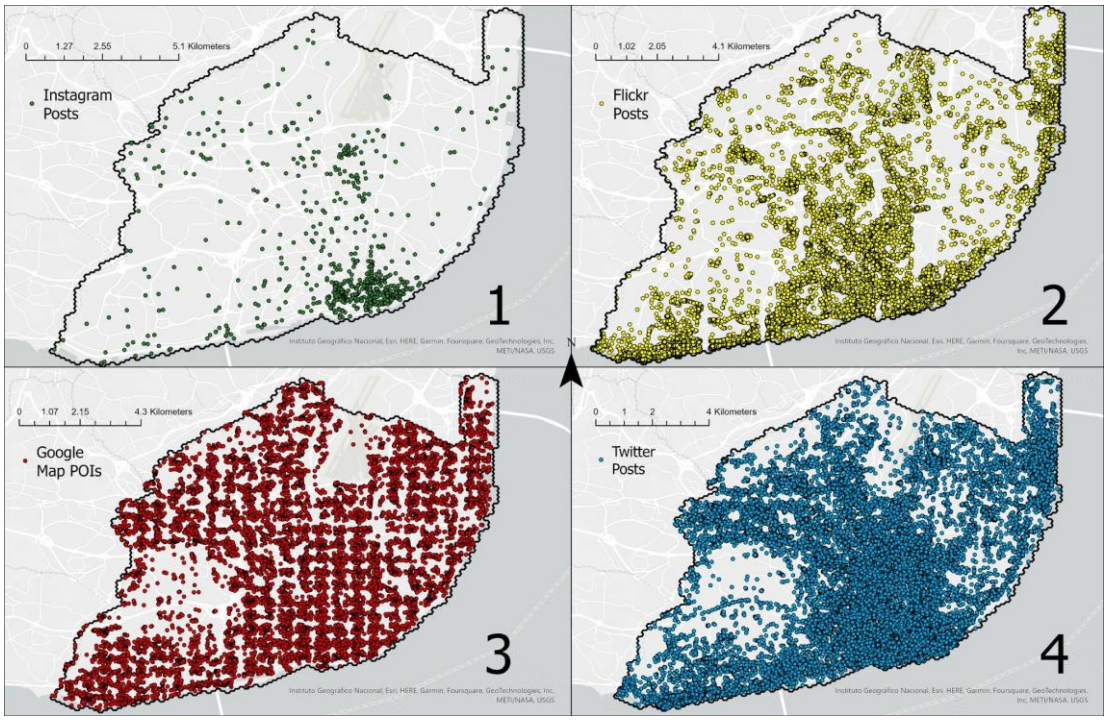


Figure 1.5 Spatial distribution of posts collected from each platform, in a clockwise direction, 1 Instagram, 2 Flickr, 3 Google POIs, and 4 Twitter

3.3 Data Variables

The variables were chosen to replicate the methodology in any city while relying only on social media platforms. This allows for continuous work on the latest version of data available for the specific town and can also increase the temporal dimension over time. The data extracted were present in two languages, English and Portuguese, which could be helpful while studying multicultural trends in a spatial context.

Since this methodology best works for a specific business type at a time, it is required to define this business **category/keyword** at the very beginning. These categories can be selected from the Google Maps POI category list, which would be related to the keywords of “food & drink” for our research scope.

The following list shows the variable names and their intended purpose for this study. How these variables were extracted and developed will be discussed in different sections:

1. Business Keyword Relevance: A value from the range 0 to 1 to depict how relevant is the specific “word” in a particular region. A high value shows the region has many social media posts related to this “word.” The variable name used for reference while performing clustering and POI prediction was **c_d_r_f_b.** This variable represents the mean relevancy of each of the five keywords, cafe, drinks, restaurants, food, and bar. However, for exploratory analysis, the five variables were also explored separately.

2. Average Age: An integer representing the average age of authors of social media posts.

3. Male Population: An integer between 0 to 100 represents the percentage of online male authors of social media posts.

4. Female Population: An integer between 0 to 100 represents the percentage of online female authors of social media posts.

5. Local/Foreign: A binary representation (True/False) to suggest whether the social media author is a local resident of Lisbon or a foreigner. This is inferred from the author's name

6. POI Proximity: A value scaled to 0-1 representing how many POIs are accessible within a buffer radius of 85 meters. The buffer is generated for each POI in the

discussion.

7. Spatial Competition Index (SCI): A number from within the range 0-1 representing the amount of competition a specific region faces to its connected neighbors.

8. Average POI Rating: A number in the range 0-5 shows the average rating relevant to the five business keyword POIs.

9. Cluster: This variable will be feature engineered using the spatial clustering algorithm, GeoSOM, to identify homogeneous regions based on the variables above.

Data sources were carefully chosen to maintain reproducibility and high temporal dimensions. The following Table 1.1 shows the variable names and their corresponding sources.

Platform	**Variables Extracted**
Twitter	Age | Gender | Business Keyword Relevance
Flickr	Age | Gender | Business Keyword Relevance
Instagram	Age | Gender | Business Keyword Relevance
Google Maps POIs	Age | Gender | Business Keyword Relevance | POI Rating | SCI | POI Proximity
Overall Temporal Range: **2010 - 2022**	

Table 1.1. Data sources and the variables extracted from each of them

3.4 Spatial Unit of Analysis

The accuracy of most mobile-based GPS has improved significantly over the last decade, enabling precise location of up to 1 meter. The spatial properties captured by various social media apps/websites rely mostly on GPS measurements while also using cell phone towers and WiFi for approximate triangulation of online users. The user may have turned off location sharing so that no geolocation information will be available in this case. We can safely assume

that the spatial resolution of the shared location will be very high, up to 30 meters in most cases **[44]**.

Considering such high-resolution data is available, it makes sense to use it for precise neighborhood profiling and understanding people's perception of places at a granular scale. Due to this property of geotagged data, it is beneficial to move beyond parish-level regionalization to even more large-scale regions by setting up a custom spatial grid over the study region.

For this purpose, a hexagonal grid of 0,02 SqKm area (~170m wide) is overlaid on Lisbon as seen in Figure 1.6. The area/width of each hexagon is large enough to account for any inaccuracies that might have occurred while approximating the user's geolocation by social media platforms. This allows us to infer that a point representing a social media post in a specific hexagon defines the hexagonal region in terms of perception of the place.

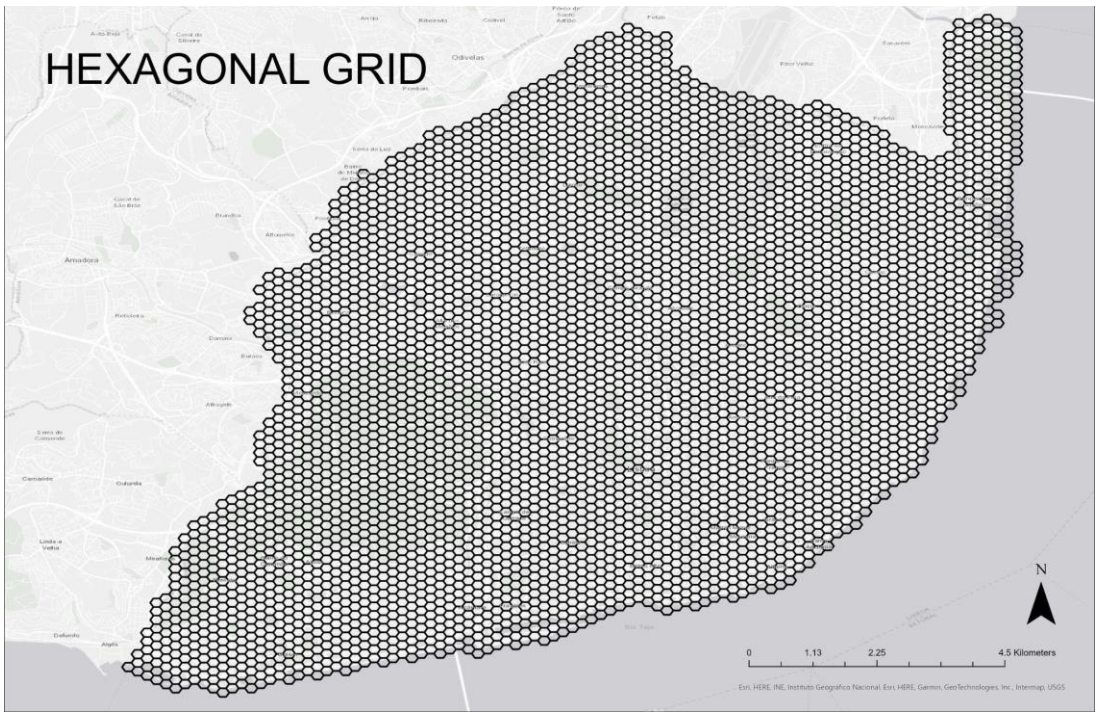


Figure 1.6. Spatial grid of 4.630 hexagons overlaid over Lisbon

The area of the hexagon is also ideal for evaluating it in terms of new POI locations as it is not too big to be vague, nor is it too small to be not able to drive/walk around. The total number of these hexagonal cells stands at 4.630, thus, allowing business professionals to perceive the city’s people and complexity in more than four thousand ways while also monitoring the city at granular scales for new possible POI locations.

The spatial grid was set up in QGIS using the “Grid Layer Tool” under “Vector > Research Tools.” The horizontal and vertical spacing was set to 170 meters each, and the CRS was inherited from the aforementioned study region. This spatial grid will be the base of analysis and visualization for the upcoming steps throughout this research.

3.5 Data Cleaning

At this point, only the raw form of the data was collected and stored in the postgres database. The reason for selecting postgres was to enable storing information along with spatial attributes and perform spatial queries on the same. Each data source was maintained in a separate table; for some data sources, the posts and author information were maintained separately. In total, there were ten tables for the four data sources. The following flowchart in Figure 1.7 depicts how the data was filtered down to a considerably smaller number by extensive cleaning processes and the final count of posts at each step.

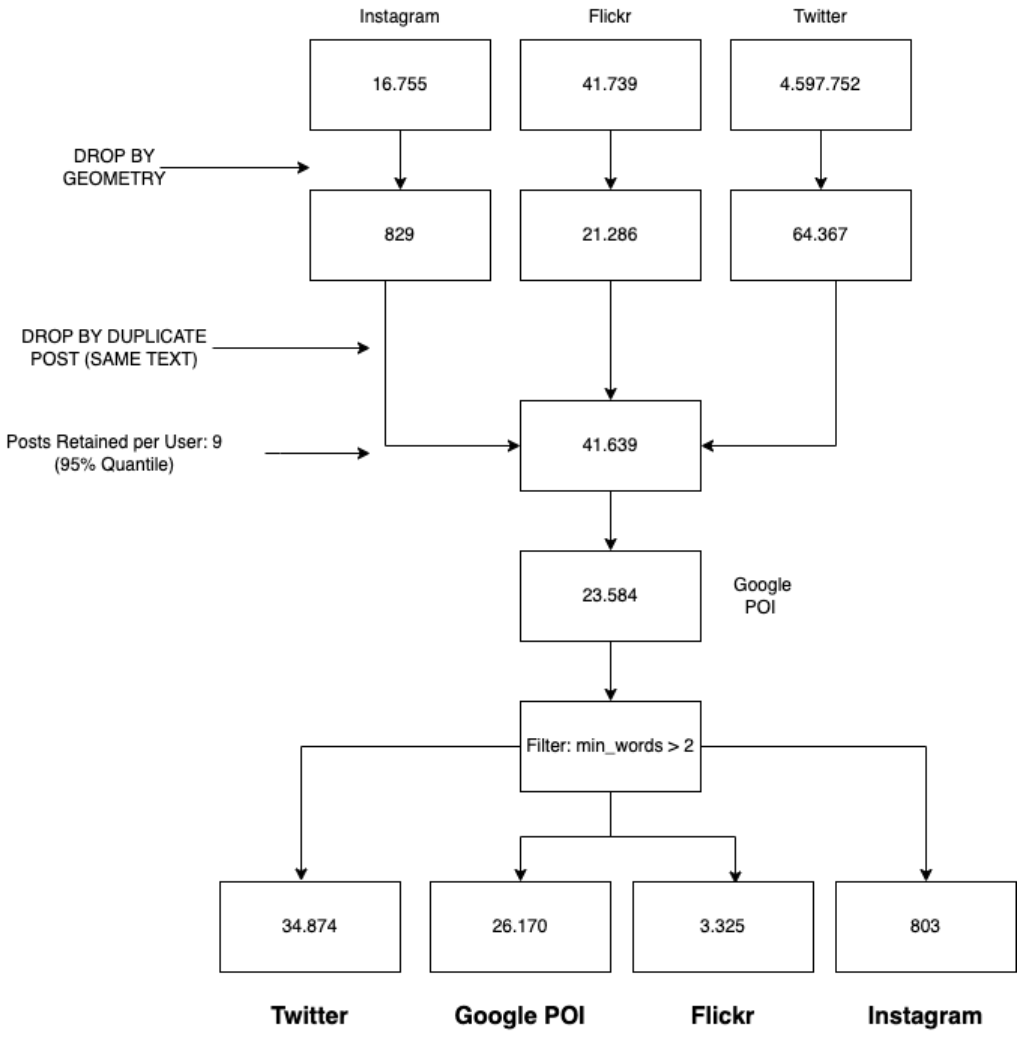


Figure 1.7. Flowchart showing the filtering process to clean the social media data

Table 1.2 shows the descriptive statistics of posts per hexagon (PPH). It can be seen that from the original 4630 hexagons, only 3532 hexagons had at least one post from all the platforms combined.

Total Count (Hexagons)	Mean No. of PPH	Min No. of PPH	Max No. of PPH
3532	22	1	1727

Table 1.2. Descriptive statistics of posts per hexagon

The data cleaning was divided into three significant steps:

1. Spatial Cleaning: As seen in the spatial distribution maps in Figure 1.3, the distribution is skewed towards some areas of the city. It was also observed that since all downloaded posts were geotagged, many were geotagged to the same lat/lon value. This happens because coordinates, due to privacy changes, are assigned to local POI or place coordinates registered in the database when user's precise location is not obtainable or enabled. After removing posts based on duplicate geometry, there were only **86.482** posts left across Twitter, Flickr, and Instagram.

   Furthermore, it is crucial to reproject the points to a common CRS to perform any kind of spatial operations across the data sources. A python library known as geopandas was used to reproject and change the CRS of all the data points and set them as **WGS84,** whose EPSG code is 4326.

   After all the corrections were made, the point data from the three social platforms were merged and exported into a single Shapefile.

2. Text Cleaning: The nature of social media posts is either textual or graphic, the graphical aspects of a post, like images/videos/audio, are out of the scope of this research, making it a text-dominated feature space. It is observed that the language used on social media for the area of Lisbon is generally multilingual, which is essential to note since the cleaning process should also respect the multilingual nature of the posts and perform cleaning appropriately. The text cleaning job is completed on all four data sources, including POI for its reviews. This process removes URLs, emojis, special characters, and even numbers. This leaves us with only alphabetic words in lowercase. The cleaning process is further refined by removing stopwords like articles and prepositions that do not provide any meaningful value about the context of the post. The stopwords removal process is done for Portuguese and English languages only, supported by the NLTK python library. Sentences with a word length of 2 or less

were also removed.

3. User-Based Cleaning: This step aims to reduce the skewness introduced by some highly active users on social media as it would bias the views of particular places to a single author. Additionally, if the posts are made very close by or in the exact location, it would again introduce skewness in the spatial distribution. To counter this, the number of posts per user was calculated, and an overall quantile range of the posts/user was observed. The number of posts up to 95% quantile per user was only 9, and as a result, the posts of users with more than nine posts were cut down to 9 by randomly selecting from a pool of their posts.

Additionally, the posts which had a word length of 1 were also discarded as they would not be beneficial in the further steps of topic modeling, where an n-gram range needs to be specified and needs to be greater than 1 for quality results. Figure 1.8 shows a sample dataframe of a cleaned subset of posts.

	hex_id	post_id	origin	tokens
65162	4629.0	57725	poi	filha experiências noventa princípios vim goog...
65163	4629.0	57726	poi	simpatia profissionais acima explicar processa...
65164	4629.0	57726	poi	ótimos resultados recomendo
65165	4629.0	57728	poi	always need translation translate time give gr...
65166	4629.0	57728	poi	awesome job great professionalism quality serv...
65167	4629.0	57728	poi	highly satisfied professionalism quality punct...
65168	4629.0	57728	poi	nice services time friendly guys course good p...
65169	4629.0	57728	poi	deliver good service work highly recommanded
65170	4630.0	41628	twitter	livro gratuito ibooks store aprendendo votar i...
65171	4630.0	41629	twitter	cedis universidade lisboa repercute encontro l...

Figure 1.8. A dataframe showing a clean sample of social media posts

3.6 Data Imputation

Due to the skewed spatial distribution of social media posts and other filtering processes, some hexagonal regions were left empty of any posts. Due to this, variables extracted from topic modeling and demographic extraction were left **NULL.** Leaving some variables as NULL can often be interpreted as 0 by some machine learning algorithms, leading to wrong results/interpretations. E.g., due to NULL age values for some hexagons, it can reduce the

average age in cluster/region. To counter this, multiple strategies are available to impute or fill in the missing values in a dataset. Common examples include imputing by mean and median.

More than simple statistical methods is needed to impute the spatially affected dataset's nature. A well-known sophisticated approach used in this research is Multivariate Imputation using Chained Equations **[46]**. The algorithm is available through the **sklearn** Python library under the module **sklearn.impute.IterativeImputer.**

The intuition of this algorithm is based on estimating the missing values by running a regression model where the missing value feature is the dependent variable. In contrast, other features act as independent variables. This is done round-robin until all missing values have been estimated. Estimating using regression provides a more robust and meaningful approximation of values than simple mean or median. The imputed data frame was then used as input for further data analysis and spatial clustering.

3.7 Business Category Definition

The objective of this work is to enable maximum reproducibility for POI region recommendation across business verticals and geographical regions; however, covering every business vertical is out of the scope of this work and, as a result of which, the focus is currently on a single business vertical of the widely known **Food & Drinks.** Defining the business context is of utmost importance as this would help guide and narrow down the analysis of how neighborhood profiling can be insightful for site selection in the Food & Drinks vertical when used with other business-specific variables.

Now that the business vertical is defined, the next step would be filtering down our POIs that fall in the vertical of **Food & Drink.** Since the POIs were retrieved from Google Maps API, it was possible to grab metadata like categories, reviews, and names of these points. There needs to be a clear distinction of whether a POI belongs to the vertical of **Food & Drink.** To counter this, a category variable with the POI data was manually vetted to find all keywords relevant to this vertical. Overall, there were 106 unique categories observed in the POI data. After manual selection, only five were found relevant to our vertical of interest. The POI filtration process for these five keywords is explained in the next section.

3.7.1 Data Filteration by Business

POI filtering was done in two separate steps. For certain POIs, there are no specific categories defined in Google Map API, and as a result, it was impossible to rely solely on this approach

of extracting all the “food and drink” POIs. To address this, the reviews of all POIs were also checked for any mention of the words from the following:

(coffee','cafe','galão','expresso','cappuccino','duplo','garoto','pingado','café','restaurant','restaurante','sande','sandwich','food','comida','petiscos','snacks','cafetaria','vega','búrguer','pizza','burger','ham','steak','meat','chicken','breakfast','lunch','dinner','brunch','presunto', 'bife', 'carne', 'frango', 'café da manhã', 'almoço', 'jantar')

The corpus includes words from Portuguese and English languages, the representation of the Portuguese words is the same as the ones in English for food items/products. For instance, galão is referred to as coffee with milk in Portuguese. Selecting reviews that mention any of these keywords gives us an additional 632 POIs. The POI was marked as relevant for this work if these words were found in the two metadata variables. The final set of 2.037 POIs can be seen in Figure 1.9

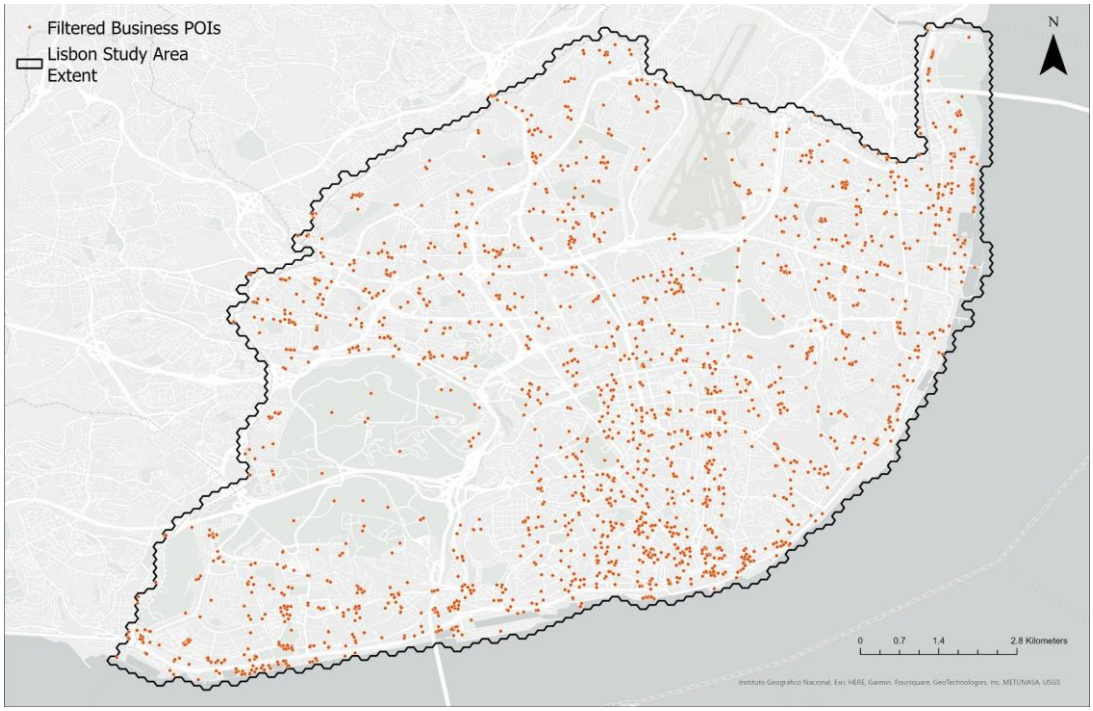


Figure 1.9 Spatial distribution of POIs after filtering up to five categories

3.8 Feature Engineering

The cleaned posts obtained from textual, spatial, and user cleaning was used in this step. The total number of posts included for topic modeling was **65.172** across all four platforms: Twitter, Instagram, Flickr, and Google POI reviews. However, since not all hexagons include the posts, only 3532 hexagons had at least one post.

BERTopic (Topic Model)

The basic idea of any topic model is to use a probabilistic method to find a group of words that belong to a certain topic (word). However, the underlying strength of the model comes from how the words of a corpus (e.g., a post) are represented. This process of representation of words is also termed feature extraction. Some common feature extraction approaches are bag-of-words, term frequency-Inverse document frequency (TF-IDF), and Word2Vec.

Conventional models like Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) use the bag-of-words feature extraction process; however, this representation does not hold any semantic relationships, for example, the words “king” and “queen” will not be treated as similar in the vector space (representation of words as numbers). This kind of representation only counts the number of occurrences of a word in a corpus. For example, the vector representation of the word **“geography”** in the sentence/corpus **“I love geography”** can be put as **[0 0 1]**. A statistical approach like TF-IDF was proposed to counter this limitation, giving higher weightage to rarely used words and less importance to frequently used words. Unfortunately, even this representation cannot hold semantic meanings; however, it is more representative of important words in a corpus.

More sophisticated models that enable the representation of words, like Word2Vec, define better semantic relationships as they project words in multi-dimensional space since they observe how often pairs/triplets of words appear together.

A recent topic modeling approach **BERTopic** was used for this research as it defines the process in three separate stages:

1. Document Embeddings: Instead of creating vector representation at the word level, the representation is developed at the document level. Bert, however, uses pre-trained (already developed models) to learn these document embeddings. The pre-trained model **paraphrase-multilingual-mpnet-base-v2** was used to generate document embeddings for this work.

   The reason for choosing this pre-trained model was that it was initially developed for semantic search. It is relevant for this work as it is required to know the other relevant topics in the input corpus for a specific business keyword. These results can only be inferred if the model understands semantic relationships, like “cafe” and “coffee” have the same context. This model also supports multiple languages, which is essential as different words in different languages hold the same meaning, and the nature of the input data is also multilingual. The model represents each sentence in 768-

dimensional vector space.

2. Dimensionality Reduction: Due to this dense representation of documents, it is required to re-project these vectors into smaller dimensions, as in high-dimensions, the concept of spatial distances differs less, which can affect the interpretation of similar/dis-similar topics. The Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm is used within Bert to reduce these dimensions and preserve local and global features

3. Dimension Clustering: The reduced clusters are then clustered using the Density-Based Clustering Based on the Hierarchical Density Estimates (HDBSCAN) algorithm, a hierarchical extension of the traditional DBSCAN that allows for clustering outliers separately.

Table 1.3 shows the parameters for UMAP, HDBSCAN, and BERTopic models while training.

MODEL	PARAMETERS
UMAP	n_neighbors=10 n_components=10 min_dist=0.0 metric='minkowski' random_state=10
HDBSCAN	min_cluster_size=10 metric='euclidean' cluster_selection_method='leaf' prediction_data=True
BERTOPIC	embedding_model="paraphrase-multilingual-mpnet-base-v2" language="multilingual" calculate_probabilities=True verbose=True top_n_words=5 min_topic_size=3 n_gram_range=(2, 4) nr_topics='auto'

Table 1.3 Model parameters used for UMAP, HDBSCAN, and BERTopic models

For topic extraction, a similar loop is run through all the hexagons, such that, the social media posts falling in a hexagon h are combined as one single corpus. BERT model is further trained in this loop for each hexagon separately. This allows for treating each geographical region as an independent entity and allows for more variation while interpreting clusters. The keyword similarity/match percentage of the five business keywords is then extracted for each hexagon separately by using the “find_topics()” function (also in a loop). However, this approach gives us the matching percentage for each matching word separately, as seen in Tables 3.4 - 3.6;

only the max of this percentage is treated as the final keyword relevancy for the hexagon. For each hexagon, a total of 5 keyword matching percentages are thus retained.

3.8.1 Topic Modeling

This section starts with a general overview of the interpretation of topic modeling results and the outputs extracted for this research, followed by a more detailed implementation.

The unsupervised process of extracting topics/keywords from data points by observing common patterns defines topic modeling (TM). This approach is usually enabled through clustering algorithms. In this case, extraction of topics can be helpful to interpret what online activity in different geographical regions represent/discuss/talk about. A topic extracted by a model is further divided into a set of words that describe the topic. Table 1.4 shows an example of the relationship between the topic “art” and “brunch” and the relevant words generated from the topic model, trained on a subset of social media posts. It can be observed that the model was able to identify relations across English and Portuguese effectively.

It is also possible to check how suitable a search term is to the specific input corpus and what topics are similar. Table 1.5 shows a list of topic IDs and their similarity score (0 - 100) for the input word and each topic. Finally, the words represented in one of the topic IDs are also shown in Table 1.6.

TOPIC	RELEVANT WORDS IN THE TOPIC
art	streetart art
	photography photooftheday
	artists art
	gallery museos
brunch	breakfast
	arroz picante
	food pasto
	food pasteisdebelem

Table 1.4 Relationship between a topic and its relevant words observed from the topic model

INPUT KEYWORD	OUTPUT TOPICS IDs	MATCH PERCENTAGE
food	16	94%
	34	75%
	104	58%
shopping	10	54%
	151	43%
	128	35%

Table 1.5 Relationship of an input topic, and its output of relevant topic IDs with matching percentage

INPUT TOPIC	OUTPUT TOPICS IDs	EXAMPLE WORDS
food	16	nikewaffle omela
		foodgram arroz torrado camarote
	104	cozinha curry
		lovecooking ricette kitchen
shopping	10	vintagestyle vintagefashion
		boutique vintage
	128	clothes roupa
		dress corset embroidery

Table 1.6 In reference to the above table, topic IDs with some of their respective words are shown

As we can see in the above tables, the words of a given topic usually represent the strength of the similarity for the topic. The similarity score also increases if the topic contains many relevant words. This metric can help infer the interest of a geographical region regarding a specific word. As in this case, the different business keywords were used to extract the

similarity in each hexagonal unit. Figure 1.10 shows a sample of the data frame given as input to train the topic model.

	hex_id	tokens	corpus
0	1.0	[avenida, igreja, ideia, felicidade, passar, a...	avenida igreja ideia felicidade passar alvalad...
1	3.0	[ohana, means, family, family, means, nobody, ...	ohana means family family means nobody gets le...
2	4.0	[auditório, pro, convencer, irmao, irmos, prai...	auditório pro convencer irmao irmos praia aman...
3	5.0	[sushi'clock, sakura, mariabritooo, passar, ma...	sushi'clock sakura mariabritooo passar mariabr...
4	6.0	[acabei, avenida, alvaro, pais, vídeo, set, co...	acabei avenida alvaro pais vídeo set construct...

Figure 1.10 A subset of the dataset provided as input for training a topic model.

3.8.2 Demographic Extraction

Socio-demographic extraction enables observing user-level distribution and how social media segments interact with their surroundings. By studying the user distribution, it is possible to infer if a specific segment interacts more/less with a particular type of business. In this work, three characteristics of a user are identified up to a certain accuracy, age, gender, and if a user is local/foreign. To identify if the user is local (in this context of the study, is the user Portuguese?) or a foreigner, it was possible to infer this by observing the first and last names of the authors. The first and last names were matched with an official list of 6.992 Portuguese names, out of which the names for the genders male and female were equally distributed.

Data Processing

Age and gender are user-level characteristics that are not openly available for privacy reasons. It is possible to infer these properties up to a certain level of accuracy by analyzing the profile pictures of the authors of the social media posts. The images processed for this purpose were already publicly accessible during data collection. As a result, as seen in Table 1.7, the number of such users extracted from the social media posts with publicly available user images was comparatively low for all the platforms except Google POI reviewers.

DATA SOURCE	TOTAL UNIQUE USERS	USERS WITH SOCIODEMOGRAPHICS
Twitter	18.069	7.412
Flickr	2.996	460
Instagram	10.360	5.912
Google POI Reviews	39.216	14.060

Table 1.7 User distribution across all platforms

Age/Gender Model

A pre-trained neural network model was used to predict age and gender for all the available profile pictures. This model was initially trained on 500.000 Twitter profile pictures. Currently, the model has a mean error of 3 years for predicting age; on the other hand, it has a 95% accuracy for predicting gender. The neural network used has the famously known architecture of WideResNet, which is commonly used for computer vision tasks.

The point vectors of final social media posts and POI review posts for which the demographics were identified are merged and intersected with the hexagonal spatial grid. Age attribute was averaged at the hexagonal level; hence, each hexagon had its average age-associated. Gender was divided into two separate variables, male and female, and represented as a percentage of the total. The Portuguese name check attribute was first summarized at the hexagonal level by taking a sum and finally represented as a total percentage.

Figure 1.11 shows two bi-variate maps for comparing the relevancy of the search term “bar” with age and Portuguese population, specifically how young-old (on the right) and local-foreign (on the left) groups interact with the business keyword **bar**. It can be seen that the parishes of Santa Maria Maior and Misericórdia seem to have a high density of young but foreign populations. At the same time, Benefica, for instance, is more attractive to older Portuguese people. The legend in the bi-variate maps reads as follows, HH (High-High), LH (Low-High), HL (High-Low), and LL (Low-Low).

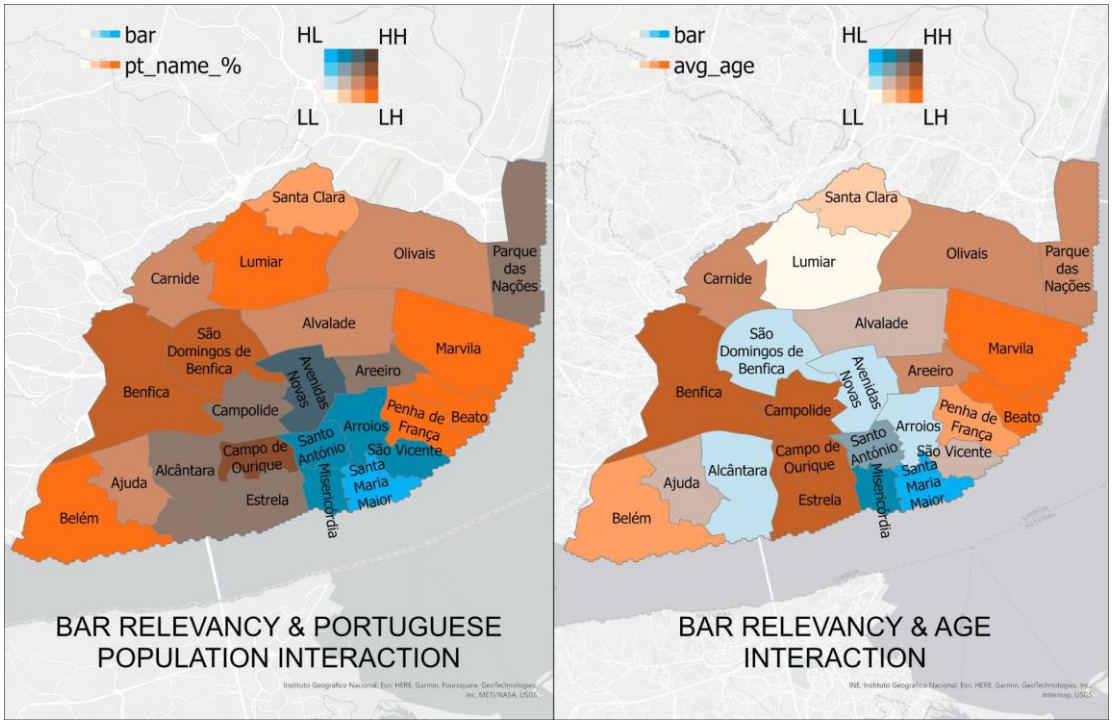


Figure 1.11 Bi-variate maps comparing two variables simultaneously. The variables “bar” relevancy and “Portuguese Population %” are on the left. On the right, “bar” relevancy and “Average Age” is shown.

3.8.3 Spatial Proximity

The proximity variable is developed from the perspective of the user/customer to calculate how many hexagons are in the vicinity of a POI’s radius of m meters. A buffer (circle) of 85 meters is drawn around each POI “p” (85 meters is chosen because the total diameter will be equal to the width of the spatial grid unit or the hexagon in this case). For each hexagon, the intersection is checked with such buffered regions. Figure 1.12 shows a representation of buffers drawn around the filtered business POIs, overlaid on the hexagonal grid.

Suppose there are no intersecting buffers with a hexagon. In that case, we can safely assume that the POIs are not easily accessible (by walking) from the center of a hexagon within a vicinity of approximately 85 meters on either side. The lack of other options can increase the importance of this hexagon “h.” On the other hand, many intersecting buffers would mean that there are a lot of accessible POIs from the hexagon “h.” Figure 1.13 shows the methodology used to obtain the resulting map of POI proximity for each hexagon.

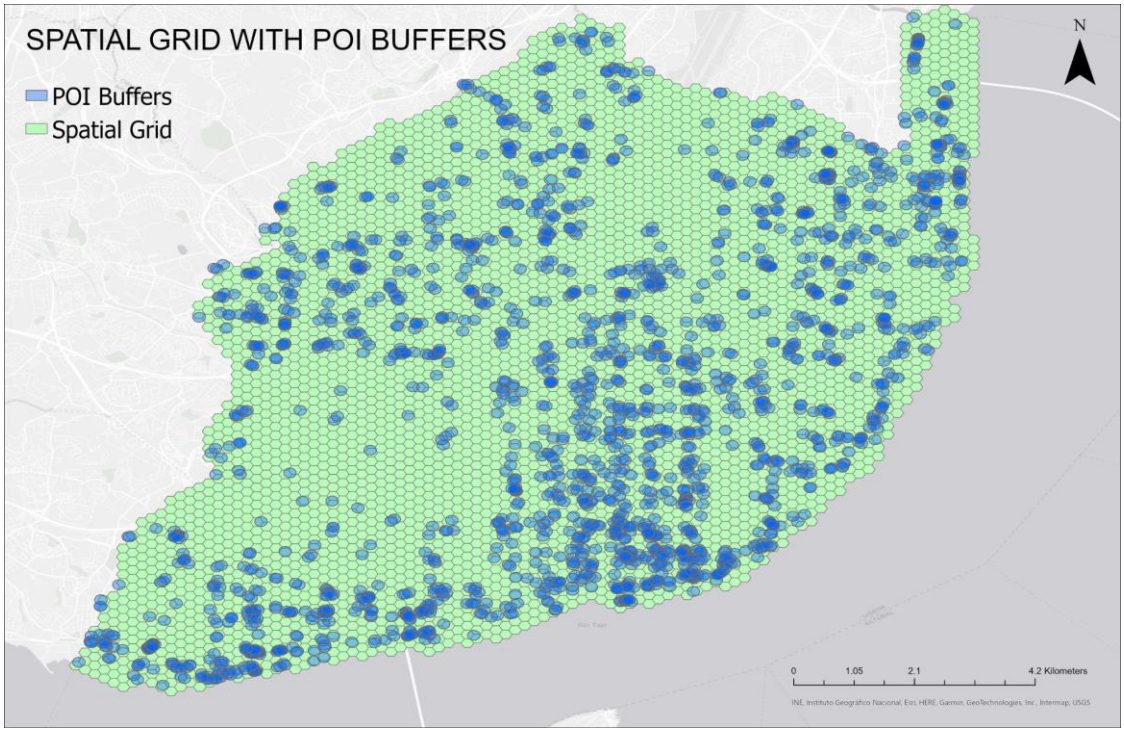
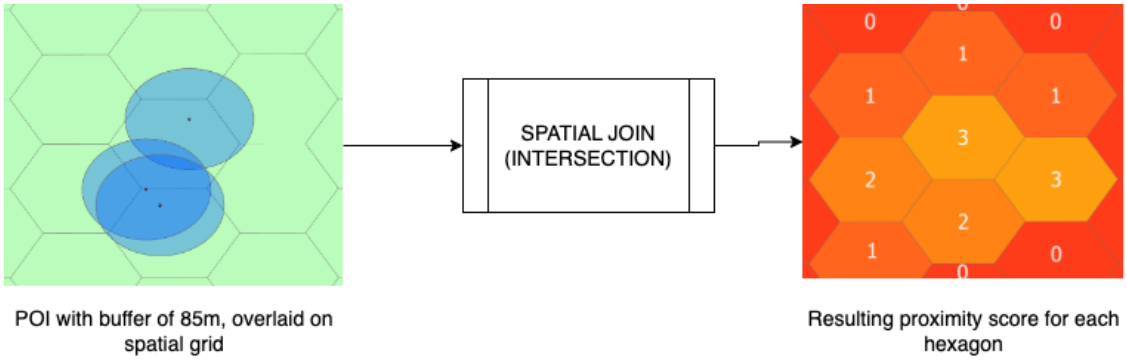


Figure 1.12 Map showing buffered regions drawn by taking POIs as the center points overlaid in the spatial grid



POI with buffer of 85m, overlaid on spatial grid

Resulting proximity score for each hexagon

Figure 1.13 POI buffers are intersected with the spatial grid, and the final intersection count is obtained.

The intersection count is summed at a hexagon level. A high count for a hexagon defines many intersecting buffers overlapping on a specific hexagon. Figure 1.14 shows the final resulting map; the count is normalized from 0 to 1.

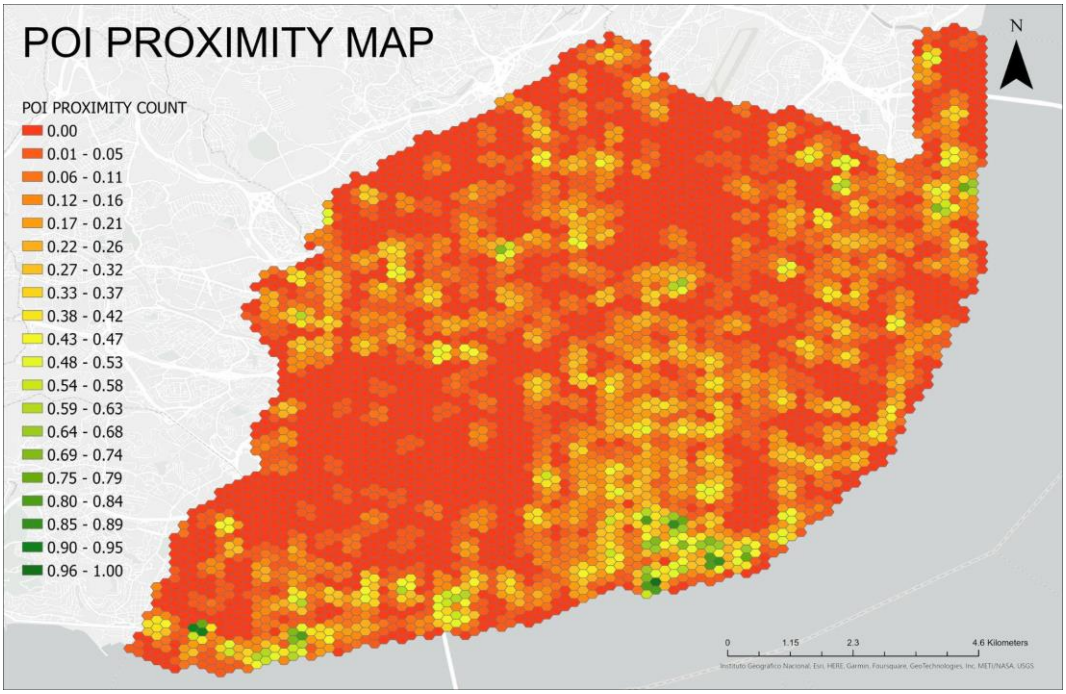


Figure 1.14 Final result of the intersection count, normalized to the range 0-1

3.8.4 Spatial Competition Index (SCI)

In any kind of retail site selection, revenue, and market share are some of the most critical factors that can alter the success of a store. The problem arises when these factors are unavailable or in a fuzzy state for newer stores. A simpler proxy of retail competition can address these two factors as the data is readily available and is mainly updated to the latest state. The competition index in this context is also known as the spatial competition index (SCI). The basic idea of an SCI is to evaluate the competitive relationship between a hexagon “h” and its neighboring six hexagons.

The previously constructed proximity variable provides information on the number of alternative stores available in a given vicinity; however, the SCI is a sophisticated statistical formula to evaluate the spatial distance within a defined but broader geographical region by also taking into account a business metric like sales. Due to the non-availability of sales data for every POI, the review of a store can be used as a metric to evaluate the success of a store in this case. The formula to calculate the SCI is shown in Figure 1.15

$$c_{ij} = \frac{{s_j}^{\lambda}}{1 + \ln\left(1 + D_{ij}\right)},$$

$$C_i = \sum_{j=0}^{n} c_{ij},$$

c = Hexagon Center
i = Hexagon Number
j = POI Number In Surrounding 6 Grids

D = Euclidean Distance
ln = Natural Log
s = Rating of the POI

C_i = Sum of all Competition Indexes for the Hexagon

Figure 1.15 Formula used for calculation of SCI. On the right is the description of variable names.

SCI is calculated for every hexagon in the study area. For every hexagon "h", the geometrical center of the hexagon is chosen as the source point, while the target points are the POIs in the hexagon "h" itself and the POIs in the neighboring six hexagons. A hexagon having zero POIs in itself and its neighbors will have an SCI value of 0, while the maximum value can be 1. For each pair of sources and the target points, the euclidean distance is measured, along with the rating of the target POI plugged into the formula. An SCI value is calculated for each center-to-POI pair w.r.t each hexagon and is finally averaged out for each hexagon. It should be noted that for the POI where no rating is available, they were not included in the calculation of SCI. Figure 1.16 shows the final output of the SCI map, while Figure 1.17 shows a zoomed-in version of the map with values of SCI and Avg. Rating of each hexagon.

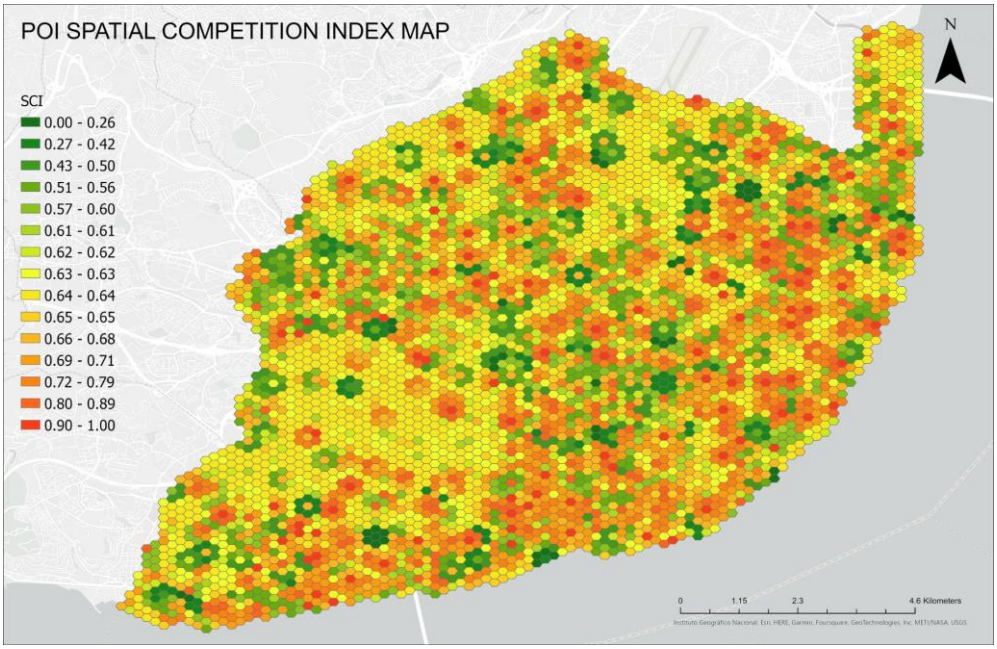


Figure 1.16 Final output of SCI calculation represented in a map

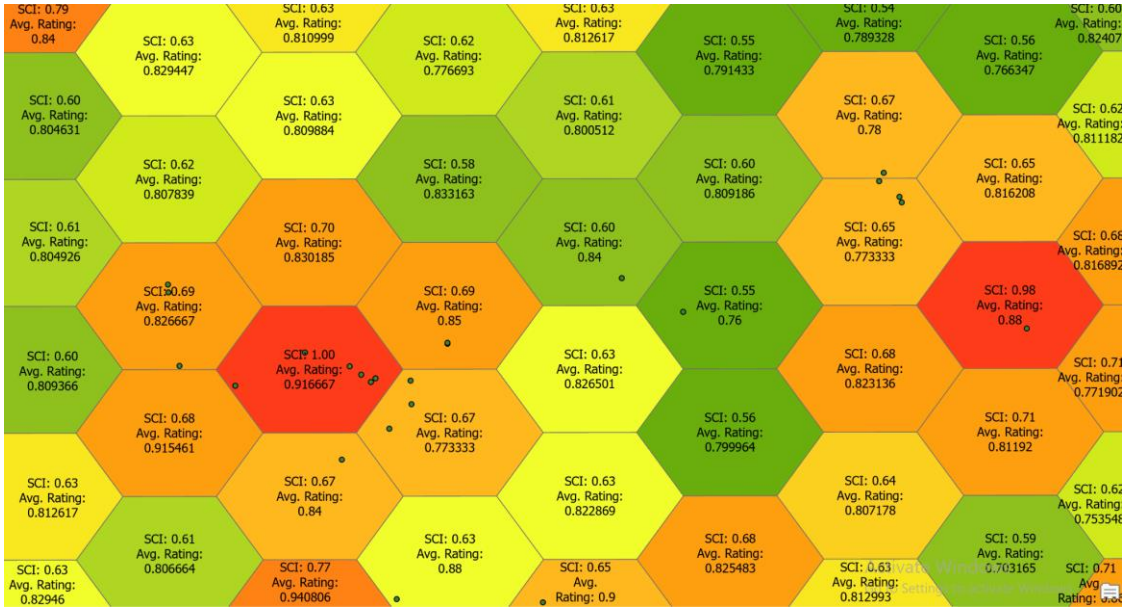


Figure 1.17 Granular representation of the map above with spatial unit, POIs, and the SCI, Avg. Rating for comparison

3.8.5 Spatial Clustering - GeoSOM

Geographical Self-Organizing Maps **[9]** are a spatial extension of the traditional SOM. SOM is a clustering algorithm; however, it can also be interpreted as a dimensionality reduction algorithm. The output of this algorithm is usually a two or three-dimensional feature map which is a mapping of the original high-dimensional data. GeoSOM solves two similar problems for this research. Firstly, due to the spatial nature of the problem at hand, i.e., POI site selection, a clustering algorithm should consider the spatial properties of the data and how the features can impact neighboring data points. Secondly, the algorithm should be able to perform efficiently in a high-dimensional space. In addition to these properties of GeoSOM, the algorithm is known to preserve topological patterns. This allowed patterns closer in space during the input and maintained while producing the output maps.

The objective of a GeoSOM algorithm is to find the Best Matching Unit (BMU), a node with the smallest euclidean distance between the input row and the output node. An input vector of available features is mapped to **n x n** (10 x 10 for this research) output nodes (output feature map or size of the map).

The GeoSOM suite tool was used for this research work to train a GeoSOM model. The software provides the outputs to be interpreted in six ways:

1. Geographical Map
2. U-Matrices or UMAT

3. Hit-map Plots
4. Parallel Coordinate Plots (PCP)
5. Boxplots and Histograms

UMAT diagram is the component that can be further analyzed to extract clusters. UMAT is the output feature map of size **n x n,** which would be defined before the training of GeoSOM. The map represents the distance of neighboring units in the input space. If visualized in gray-scale format, the higher distance units in input space (nodes in output space) will be darker than neighboring or closer units. As a result, the lighter nodes can be grouped (as they are already nearby) by using a drawing tool from the software and drawing a boundary around these nodes.

Figure 1.18 represents the UMAT on the left and the clusters drawn over it on the right. This methodology can then be used to infer multiple clusters from the UMAT. The tool also allows highlighting the most distant nodes (darkest color) by using a z-index filter. This filter, if enabled, colorizes (in yellow) the remote nodes for easier marking of boundaries.

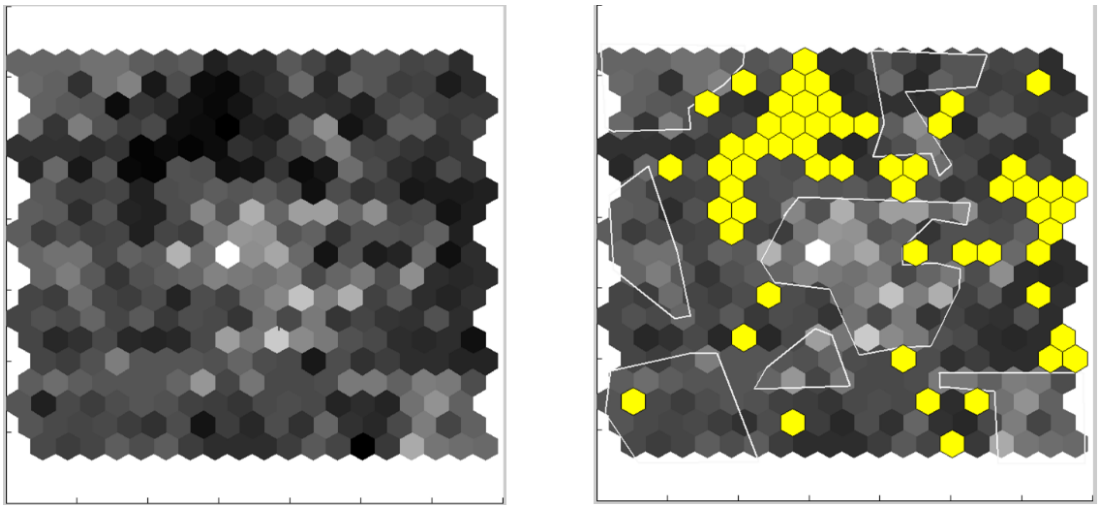


Figure 1.18 UMAP retrieved from GeoSOM (left). 7 Clusters drawn on the UMAP (right). The yellow hexagons highlight the nodes with the highest heterogeneity

As a result, the clusters formed are highly homogeneous, as seen in Fig. The clusters formed on UMAT can be visualized geographically in a map representation. Figure 1.19 shows the parameters in the GeoSOM suite tool for training the model. The component plans for each of the variables can be seen in Figure 1.20.

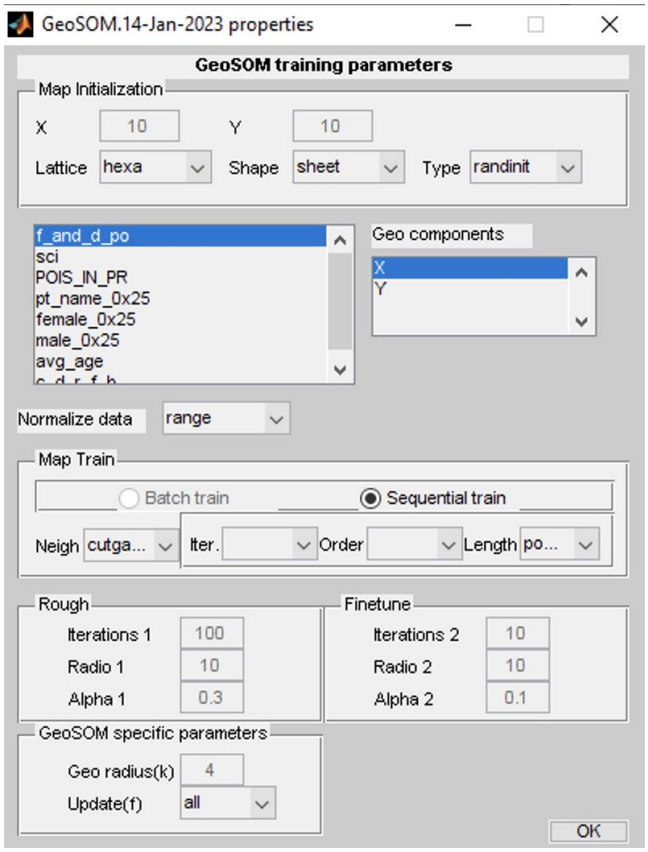


Figure 1.19 A window panel showing the final parameters set in GeoSOM suite

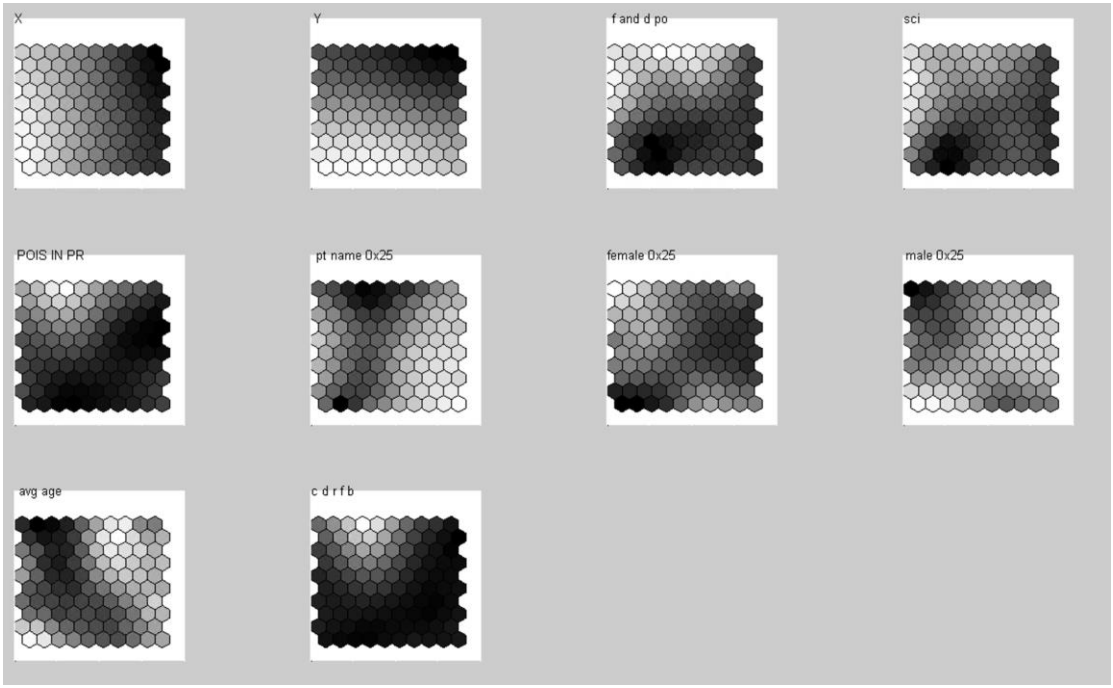


Figure 1.20 Component plans obtained after clustering in the GeoSOM suite for each variable.

3.8.6 Hexagon Classification / POI Location Prediction

In this section, the methodology to predict POI locations is discussed. The vector type of a POI is a Point, meaning, in an ideal scenario, the new predicted location for a POI should also be a Point. In this research work, however, a neighborhood or the hexagon is associated with each POI, and also because the basic spatial unit in this research work has been defined as a hexagonal grid which is why the prediction workflow has been designed in such a way that the predicted locations will be a polygon or the hexagon where the POI should be placed in. This defines the problem as predicting hexagons, if they are ideal for placing a POI or not, making it a classification problem, more specifically, a binary classification problem.

The classes for this problem can be defined as **Ideal or class label 1, Not-Ideal, or class label 0.** The variables chosen for GeoSOM clustering will be used as independent variables for the classification task. In addition, the clusters identified from the GeoSOM algorithm for each hexagon will also be used as an extra feature. The algorithm used for this purpose is the Random Forest Model. The algorithm is available through a Python library, **sklearn.ensemble.** Random Forest (referred to as RF now) is a decision tree-based algorithm. As mentioned, it is available under the “ensemble” module, which defines that RF combines multiple decision trees where each tree is trained on a different subset of samples from the original data (with replacement); this concept is also known as bagging. Additionally, each tree has a different set of features from the initial set of features present, which ensures no correlation between the trees and that the variance is reduced in the model. A voting mechanism of each tree hence realizes the importance of each feature. The model's output can be a class or a probability of belonging to each class.

In this case, the probabilities were extracted from the model, and the threshold used in this research was set to 0.5. Predicted values above or equal to 0.5 will be treated as they belong to class 1 (Ideal Hexagons), and values below 0.5 will be regarded as class 0 (Not-Ideal Hexagons).

4. RESULTS & DISCUSSIONS

This section explains the different intermediate and final results achieved from the different stages in methodology.

4.1 Neighborhood Profiles

It is known that the 24 Parishes identified by the government can also be considered neighborhoods on their own. The objective of disregarding these administrative units is to see a city from a different perspective, i.e., from the current dynamics of people in terms of their online behavior. Based on the spatial grid of 4.630 hexagons, it does not necessarily mean that there would be precisely the same number of neighborhoods formed. This is due to unavailability/significantly low activity on social media for some hexagonal regions. The following sections explain the observations made across each variable individually. It should be noted that the maps shown in this section were developed after imputing the missing values.

4.1.1 Exploring Topic Modeling Results

Figure 1.21 - 1.23 show the relevancy of the keywords cafe, food, and bar across the study region at a hexagonal and parish level.

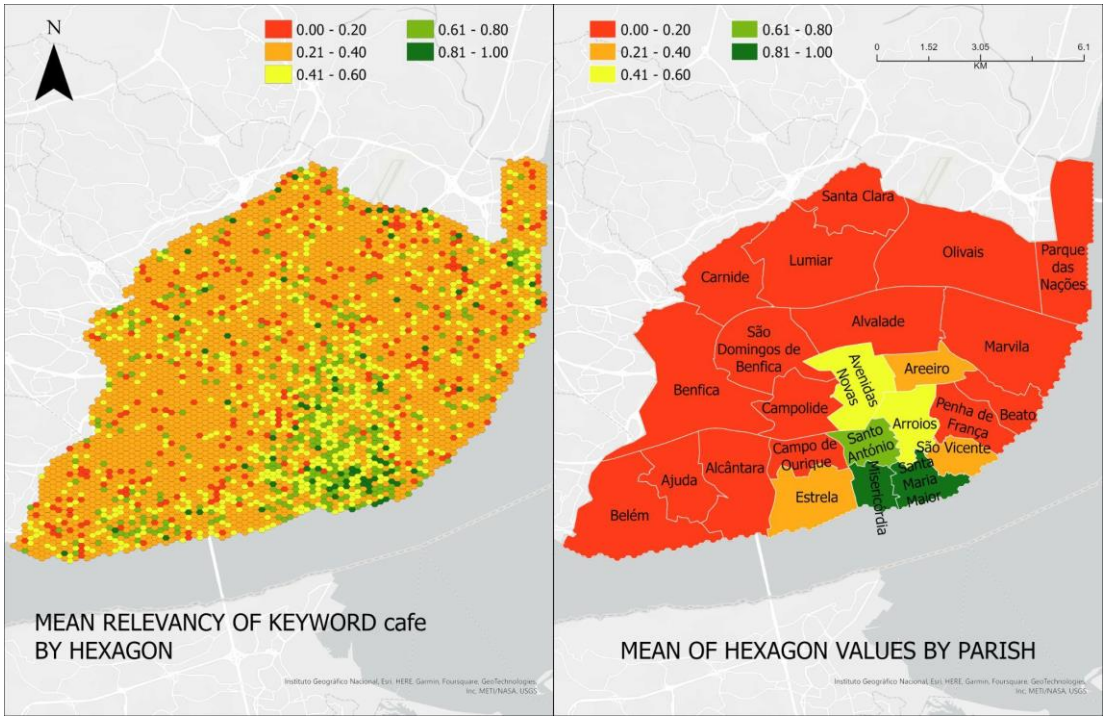


Figure 1.21. Maps for the keyword “cafe” relevancy at the hexagonal level (left) and the parish level (right)

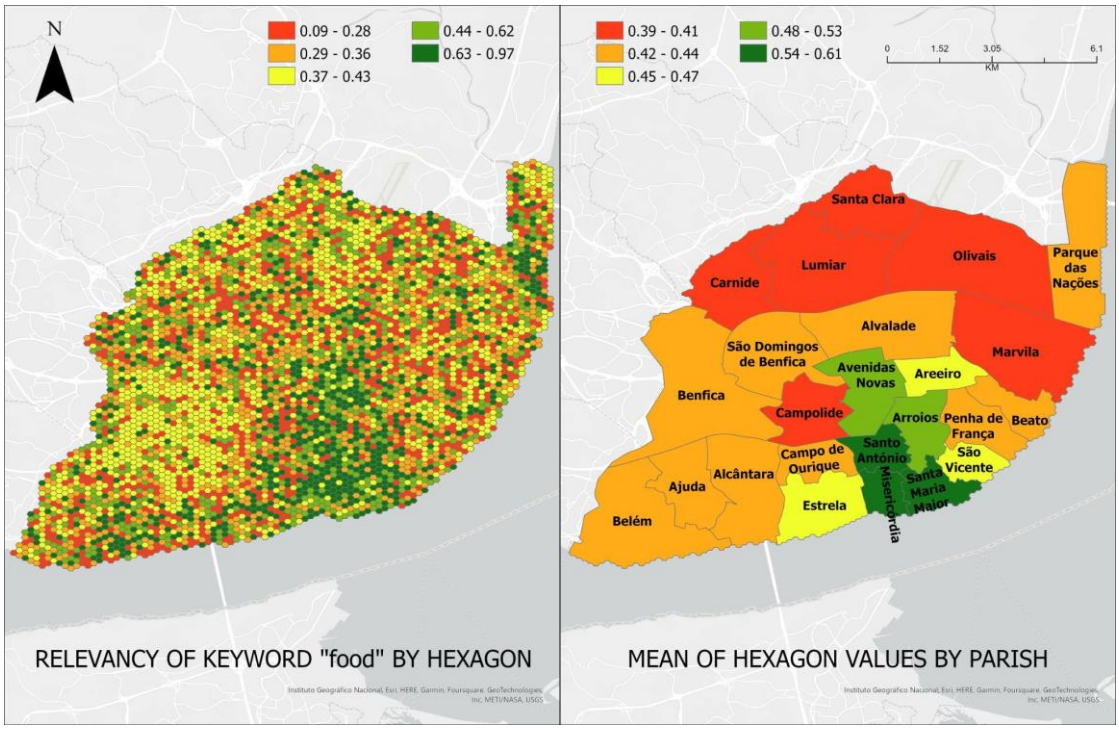


Figure 1.22. Maps for the keyword “food” relevancy at the hexagonal level (left) and the parish level (right)

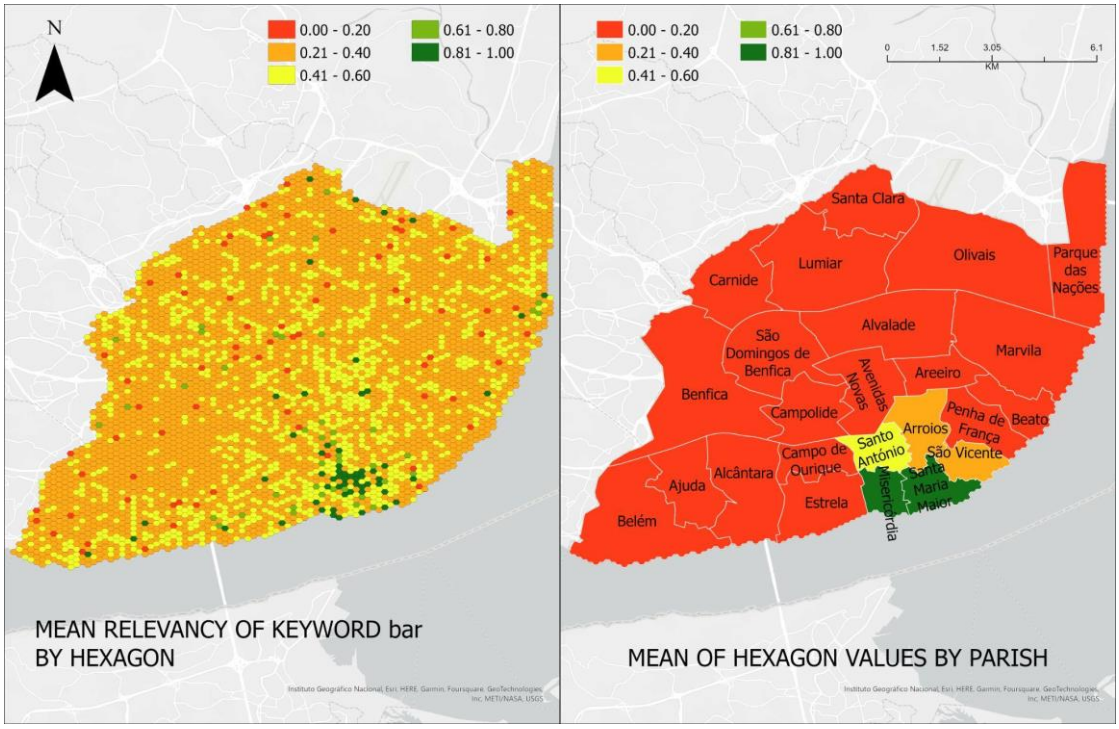


Figure 1.23. Maps for the keyword “bar” relevancy at the hexagonal level (left) and the parish level (right)

A straightforward interpretation that can be made from the above three figures is how the parishes Santa Maria Maior, Santo Antonio, and Misericordia are highly correlated across the three keywords. A similar trend can also be seen at a hexagon level. One explanation could be that these areas are more attractive to tourists and young people and famous for nightlife, which is why the three parishes are also known as the “downtown” of Lisbon.

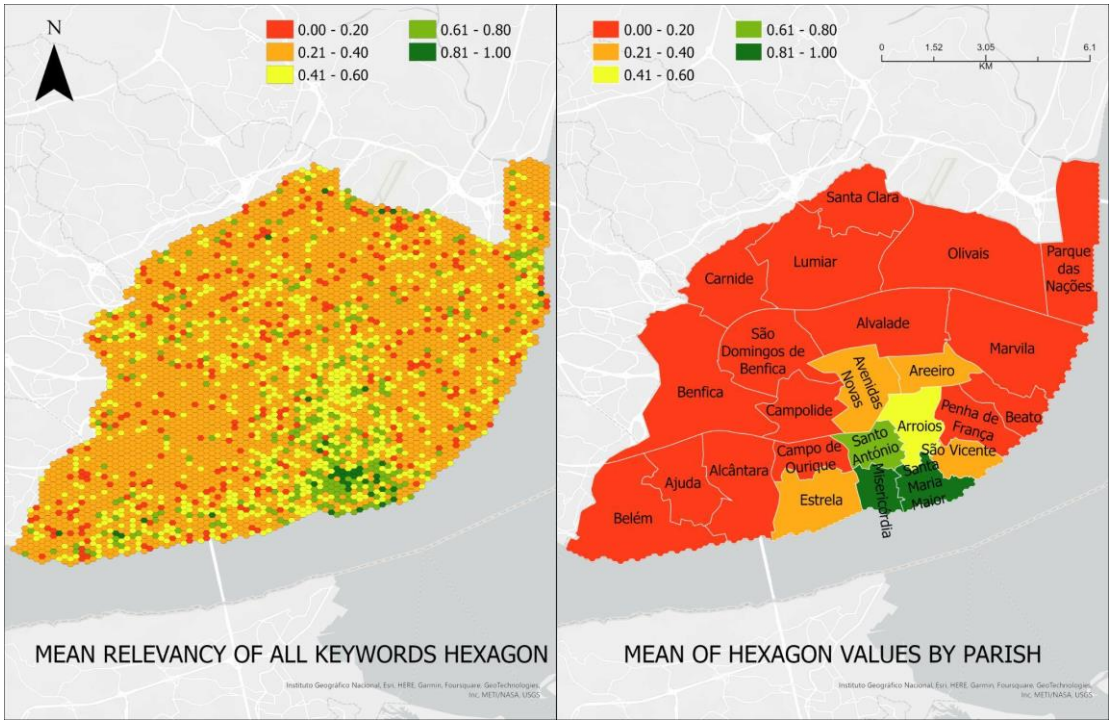


Figure 1.24. Maps for mean relevancy of all 5 keywords at the hexagonal level (left) and the parish level (right)

This allows a business manager to interpret that the people in specific parishes tend to talk and mention cafes/bars/food on social media more often. These maps also convey what regions are less popular for what keywords. For instance, any activities related to “bar” are not so talked about in the parishes of Belem/Benefica/Ajuda; however, for the other two keywords, the relevancy is higher. It can also be observed that the distribution of high/low relevancy is not random but is spread only through adjacent neighborhoods, taking the form of spatial correlation.

Figure 1.24 shows a better relevancy representation (variable: **c_d_r_f_b**) of the chosen business keywords. As we can see, the correlation for the known downtown region is also maintained in these pairs of maps. Parishes towards the coastal region have the highest relevancy while moving away from the coastal boundary significantly impacts the overall relevancy. Such representation of maps is helpful for decision-makers as it allows viewing the problem from different perspectives. E.g., in the parish Parque das Nacoes it can be observed that the parish is not so attractive for the chosen five keywords; however, after viewing at a

hexagonal level (finer resolution), some smaller regions can be called hotspots for the business and have high relevance as well. In contrast, the North, West, and East regions are less attractive in the same context.

To assess the accuracy of the topic modeling results, two hexagonal regions were chosen randomly from the spatial grid. Figure 1.25 shows an example of the different sets of topics generated for the specific geographical region in Figure 1.26.

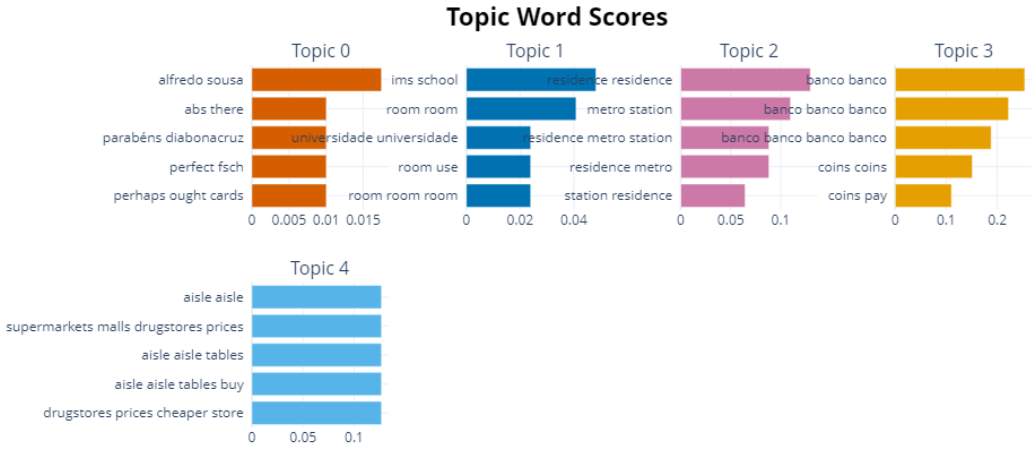


Figure 1.25. Topics generate for hexagon number 3648

These results are shown for hexagon number 3648, highlighted with red boundary in. The chosen hexagon is seen to be overlaid on the Universidade Nova de Lisboa campus. The topics in Figure 1.25, specifically from Topic 1, can be considered relevant to the actual geographical region in the discussion. E.g., the topic “ims school” refers to the NOVA IMS Institute, while the topics of “room” and Topic 0 refer to the Alfredo de Sousa Residence adjacent to the NOVA School of Law seen in the OpenStreetMap view.

Similarly, a comparison can be seen for the hexagon number 2782, highlighted in red in Figure 1.28, and the topics generated for this hexagon in Figure 1.27. The hexagon is overlaid on the famous park of “Jardim do Castelo de São Jorge” in Lisbon. The topics can also be seen as relevant to the place, specifically the words like castle, Castelo, and beautiful. Other topics like Fado refer to surrounding restaurants that play famous Fado music at these places.

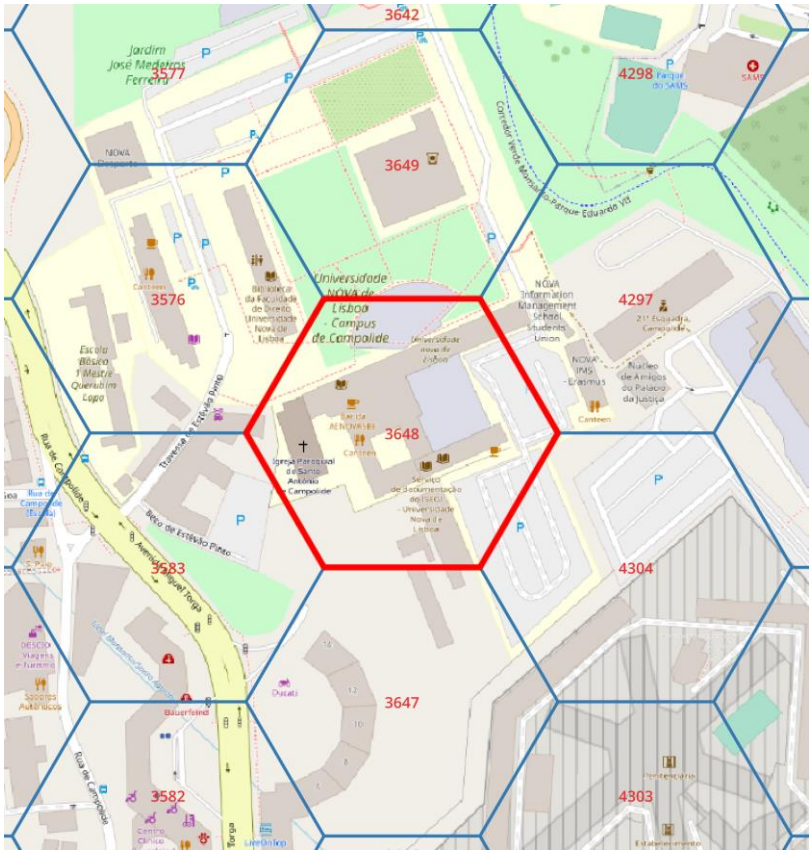


Figure 1.26. Hexagon number 3648, highlighted in red, overlaid on the Universidade NOVA de Lisboa campus

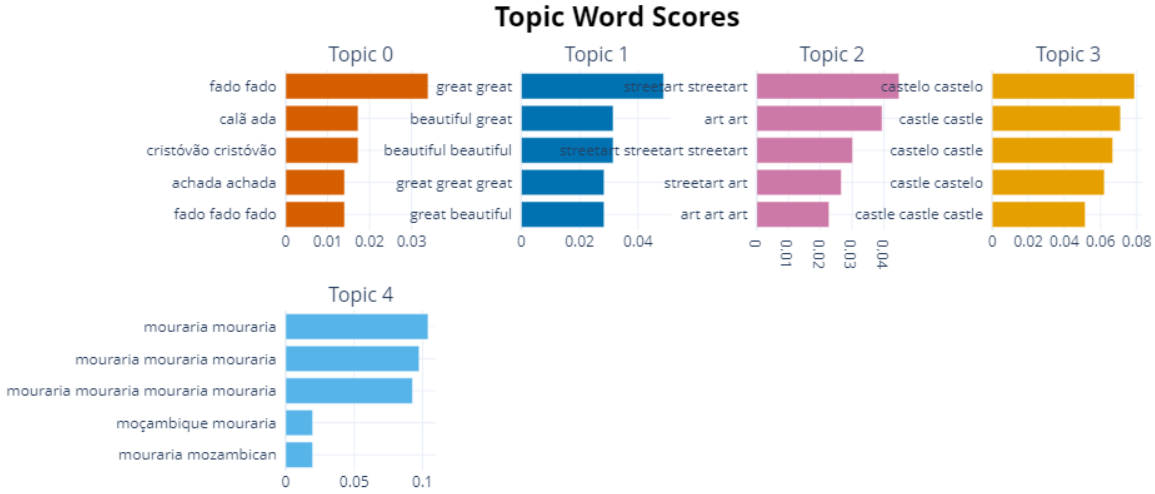


Figure 1.27. Topics generate for hexagon number 2782

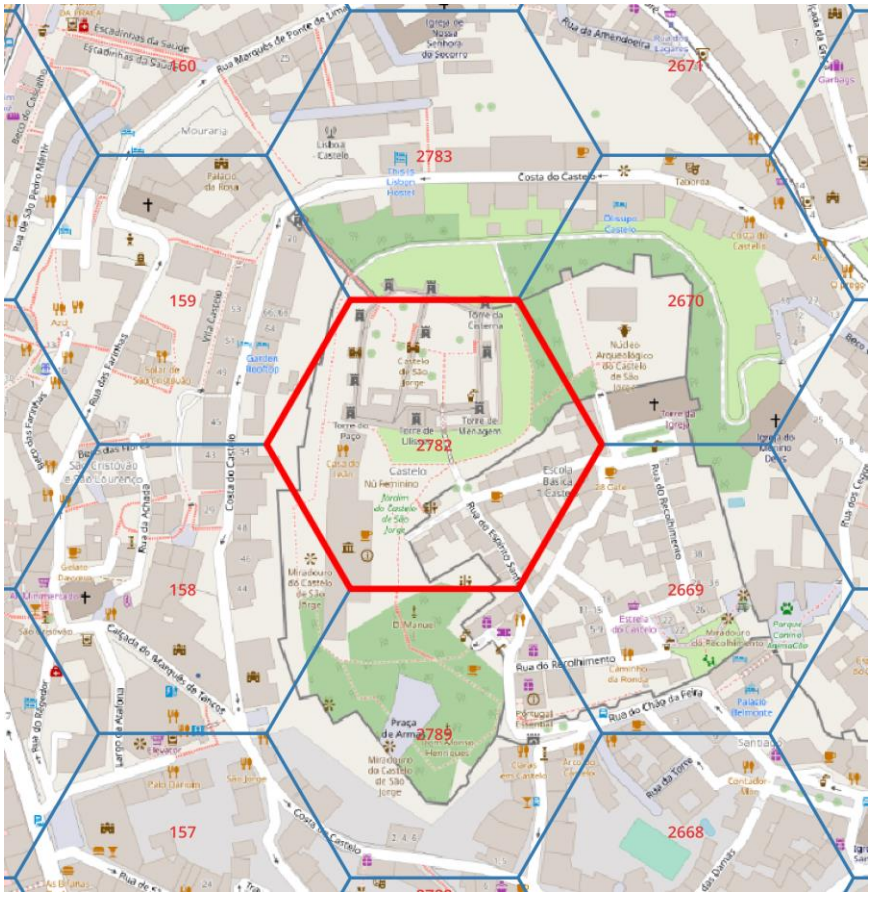


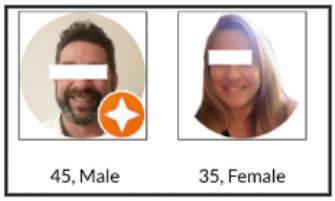
Figure 1.28. Hexagon number 2782, highlighted in red, overlaid on the Jardim do Castelo de São Jorge park

Overall, the topic modeling results are interpretable as they match the geographical characteristics of the hexagon and its surroundings. They also summarise the neighborhood into a few topics and give us a general idea of what a region is mainly known for.

4.1.2 Demographic Analysis

Deompgrahic attributes can further enrich the data by providing information on what segments of society are more or less inclined towards what kind of activities. To infer this kind of information, it is vital to identify the attributes of age, gender, and presence of the Portuguese population as accurately as possible.

Figure 1.29 shows a sample output table (right) with all three attributes identified for some users. The possible values for the variable “is_portuguese_name” can be 0 (False) or 1 (True), while for the variable “gender,” it can be either M (Male) or F (Female). On the left, gender and age attributes identified for two users can be seen.



name text	age integer	gender "char" (1)	is_portuguese_name integer
Yasmin	39	M	0
Gerardo Soto	37	M	0
Sandro Santos	33	M	1
Daniele Rizzo	26	M	0
Samuel Legrand	25	M	1
JustAnotherPO	39	F	0
Mauricio Ribeiro	33	M	1
Moitaz	31	M	0
Owen Leed	40	M	1
Jeff Bigham	38	M	0

Figure 1.29. On the left, age/gender is predicted for two users. On the right, is_portuguese_name is also identified from the names of users

Figure 1.30 and 1.31 shows the spatial distribution of males and females across the study region at the hexagonal and parish level. It can be seen that there is a high density of male online activity in the Benefica, Campo de Ourique, Santo António, and Parque das Nacoes and the lowest density is observed in Penha de França.

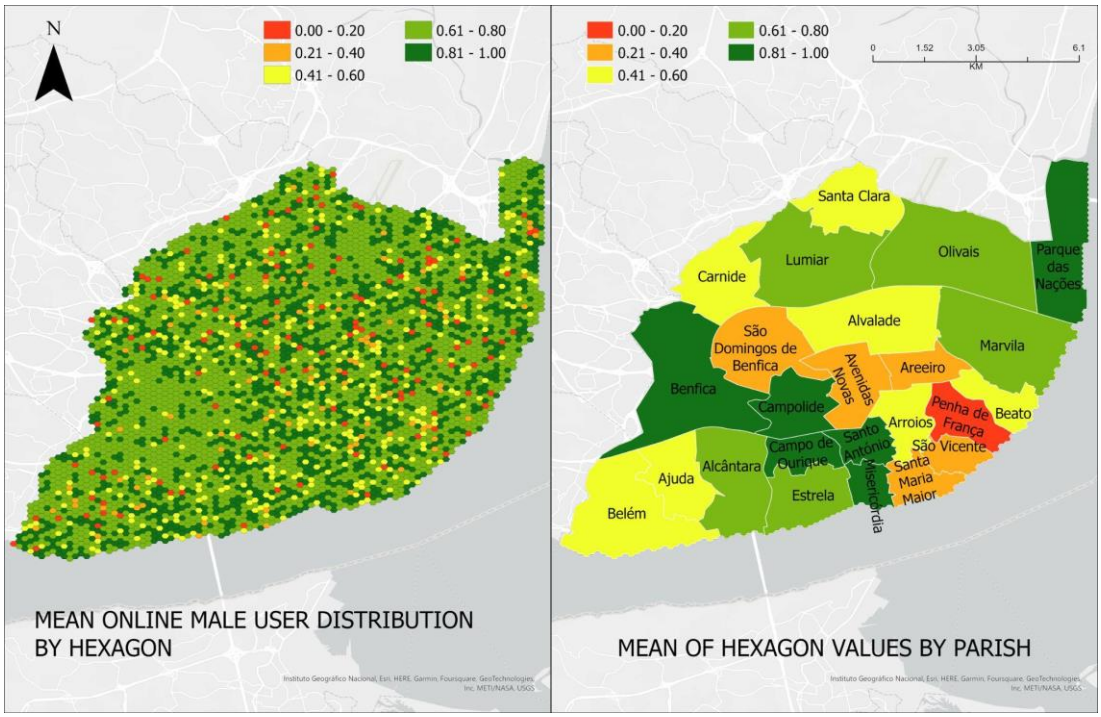


Figure 1.30. Online spatial distribution of the male population

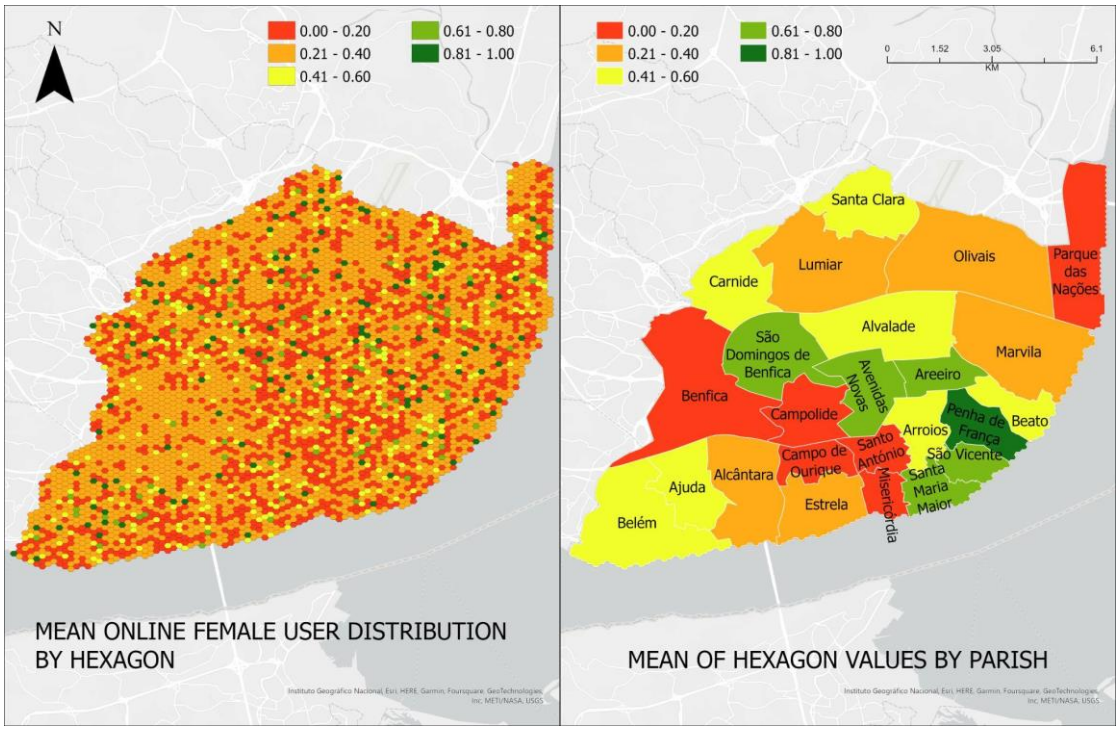


Figure 1.31 Online spatial distribution of the female population

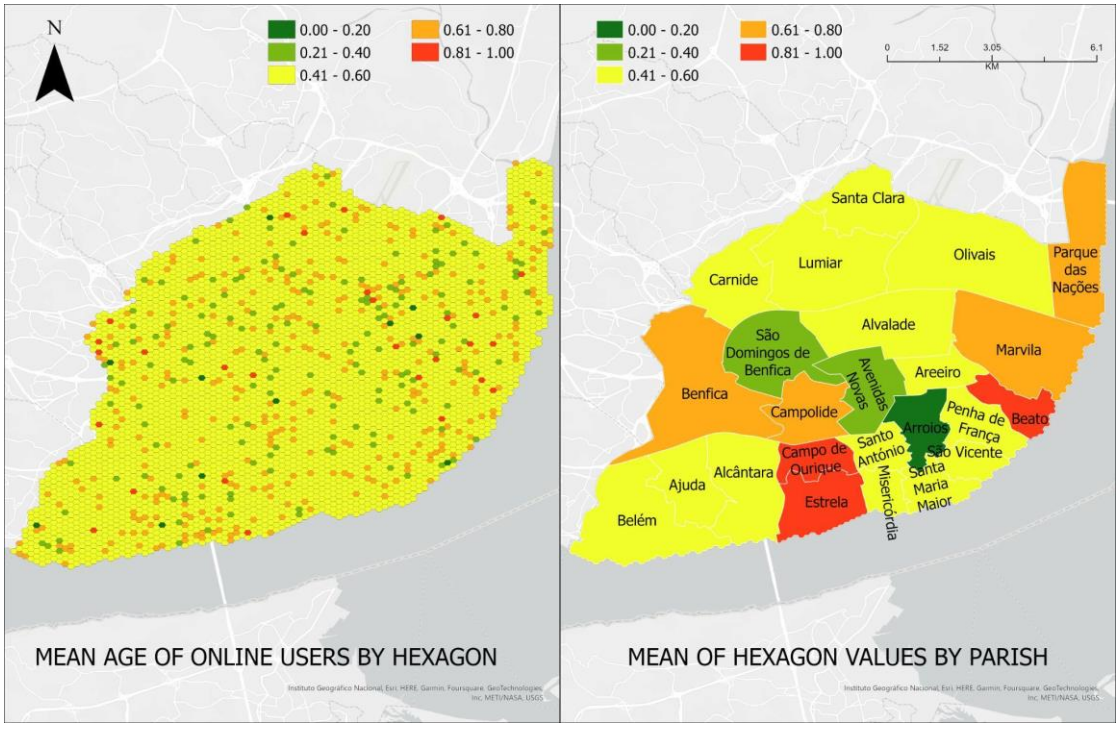


Figure 1.32 Spatial distribution of age inferred from online users

Figure 1.32 shows the average age distribution observed from the online user base. The figure shows the normalized version of age to compare the two maps on an equal scale; however, the average was observed from 32-34 years. Most parishes can be seen to have a similar age

distribution primarily; however, Arroios parish has the highest distribution of young people, whereas parishes like Estrela, Campo de Orique, and Beato have a high distribution of old people.

The map in Figure 1.33 shows the density of Portuguese online users. The parishes of Santa Maria Maior can be seen as having a higher concentration of tourists. At the same time, as we move towards the northern parts of Lisbon, the percentage tends to increase. To summarise, the online demographics of the city can be observed as primarily male-dominated, where only certain regions have a higher concentration of female users, most of whom are less than 30 years of age. A diverse distribution, however, can be observed for the population of online Portuguese users.

4.1.3 POI Spatial Maps: Reviews, Proximity, and Spatial Competition Index (SCI)

It is also essential to view a neighborhood's characteristics from the perspective of the POIs in the region. Inferring these variables can help us understand the accessibility, quality, and competition of the current POIs. For instance, a low SCI or low mean review in a hexagon combined with a high value of topic modeling relevancy variable can indicate a gap between demand and supply. Similarly, a high SCI is a good indicator for business managers to look for a location for their POI elsewhere. The hexagon maps are essential in this case as they can identify untouched sub-regions within the parish and can be a potential for the business.

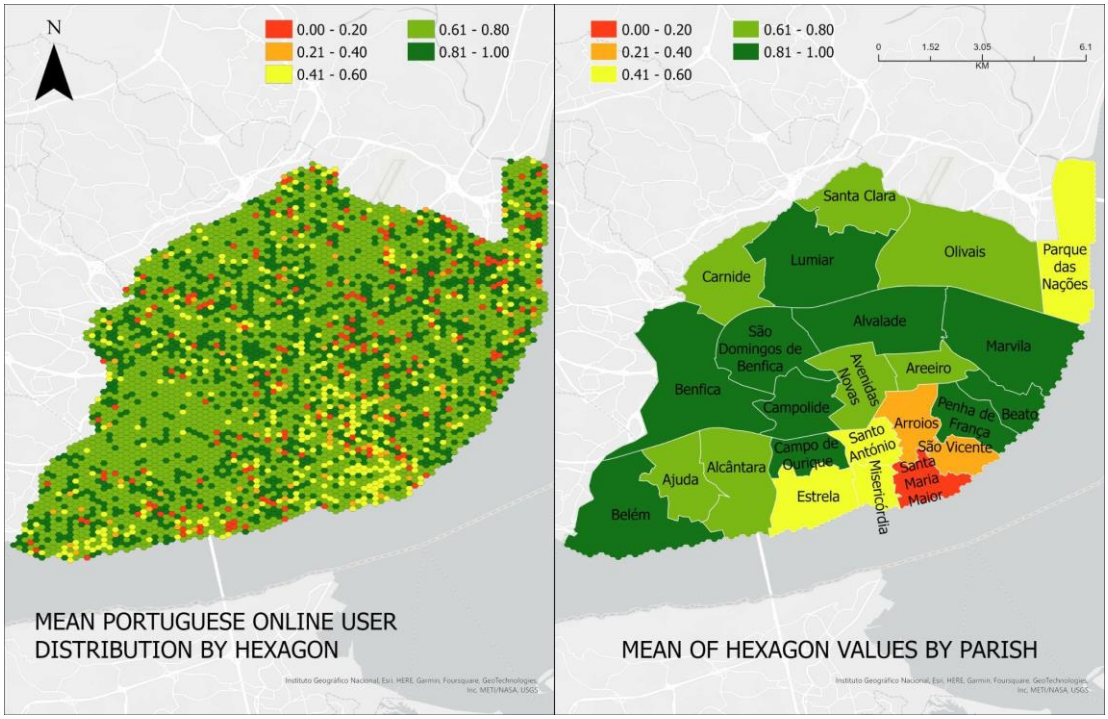


Figure 1.33. Spatial distribution of online Portuguese users

In Figure 1.34, the maps show the POI accessibility variable “POI_IN_PR.” It can be seen that most hexagons that have at least 10 POIs accessible from a hexagon “h” are mostly centered towards the southern-central parts of the city. The parishes of Benefica and Campolide have the least number of hexagons which have POIs in the vicinity of 85 meters from it. The parishes with the highest number of POIs accessible are also famously known for several touristic attractions.

In Figure 1.35, the SCI map is shown to compare which regions of the city have higher competition for the given POIs category. It should also be observed that variables of proximity and SCI are not correlated as they provide different kinds of information. However, both variables are based on the concept of spatial distance. For instance, the parish Arroios has a large number of accessible POIs, but the SCI value of this parish is lowest as compared to other parishes. This is because the SCI formula also considers the reviews variable, which signifies the quality of a POI. Hence, this allows business managers to realize that the parish may have many hexagons where the POI reviews, on average, are low. Interestingly, this can be confirmed by the map in Figure 1.36, which shows the reviews of the filtered POIs. For the parish Arroios, it can be confirmed that the region has a lower review score than other regions.

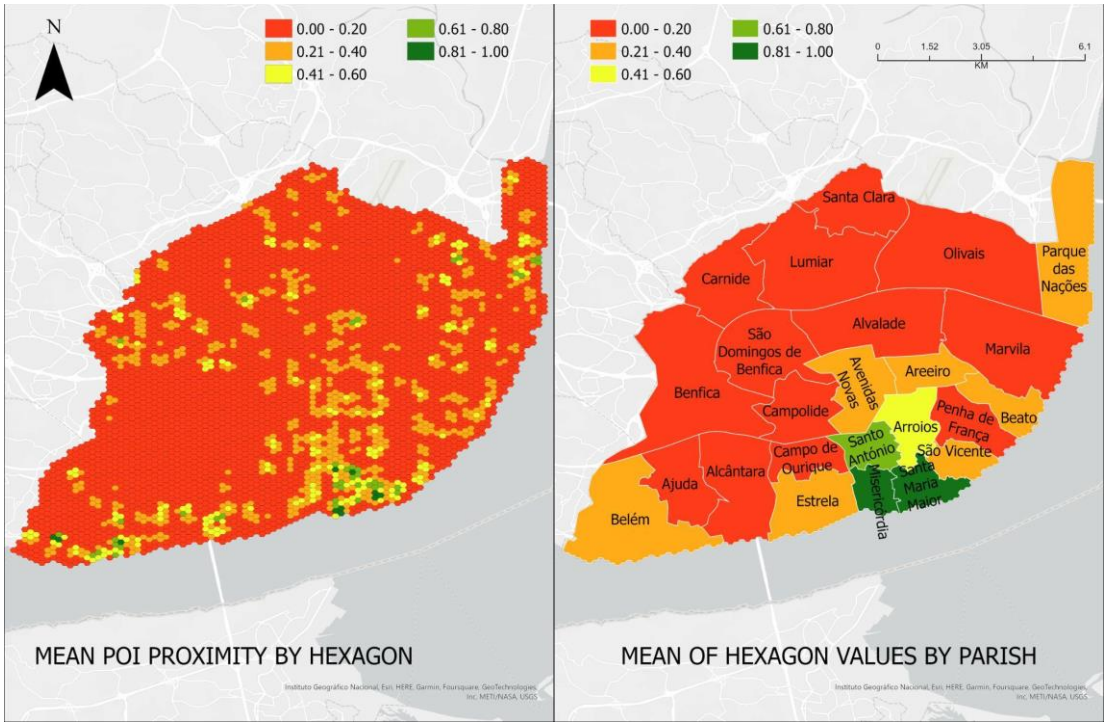


Figure 1.34 POI proximity map

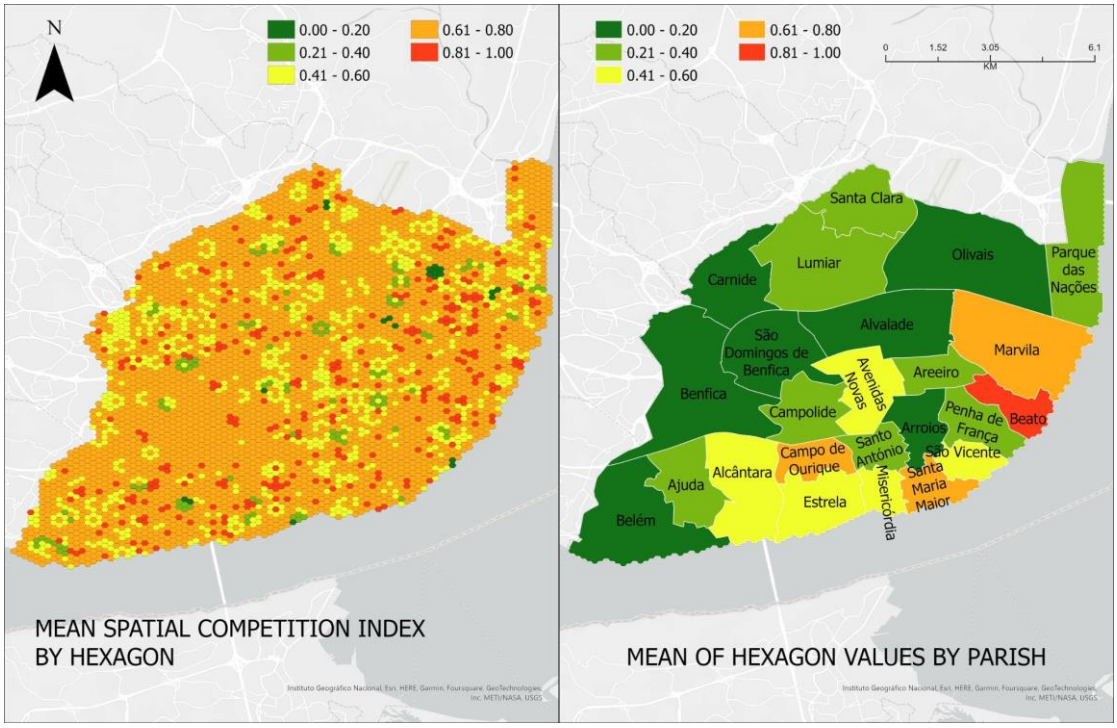


Figure 1.35 Spatial Competition Index (SCI) map

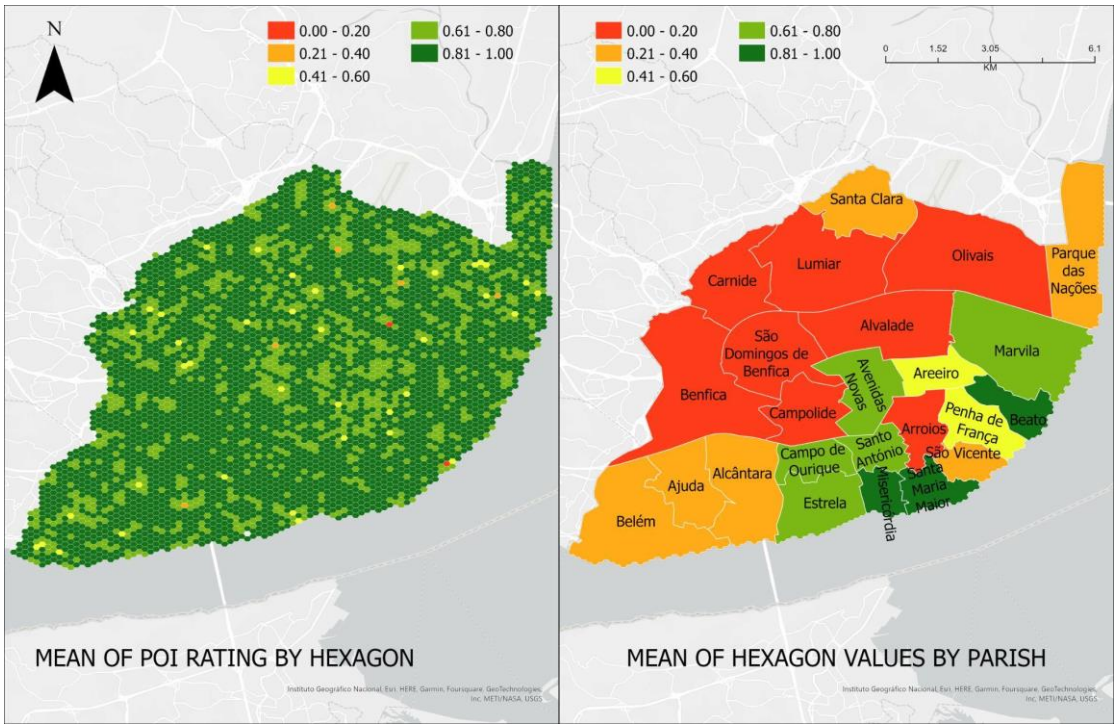


Figure 1.36 POI Rating map

4.2 Cluster Analysis

In this section, the results obtained from GeoSOM clustering are discussed. Figure 1.37 shows the obtained geographical maps of 7 clusters. On closer observation, it can be seen that the

clusters themselves take the shape of bigger parishes. For instance, we can see almost 6 clusters dividing the coastal line of Lisbon. The central and northern regions have also been separately identified as independent regions.

Figure 1.38 shows a better representation of parishes with clusters underneath them. It can be seen that the parishes of Carnide, Belem, and Parque das Nacoes are identified as independent homogeneous regions. Towards the southern/coastal parishes, Santa Maria and its four neighboring parishes also have a similar cluster of hexagons. Towards the central part, the parish of Alvalade is also seen to be primarily a homogeneous cluster of hexagons. The GeoSOM clustering has summarized the city well into seven regions with similar attributes across demographics, keyword similarity, and POI characteristics.

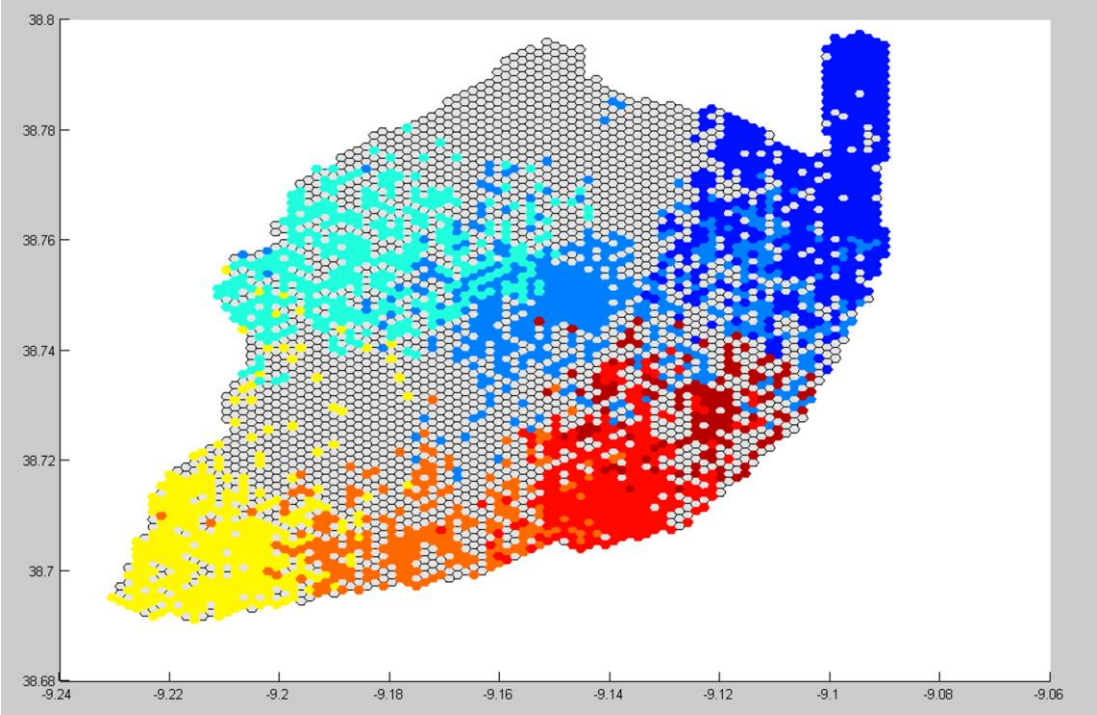


Figure 1.37. The 7 GeoSOM clusters were obtained from the GeoSOM suite tool

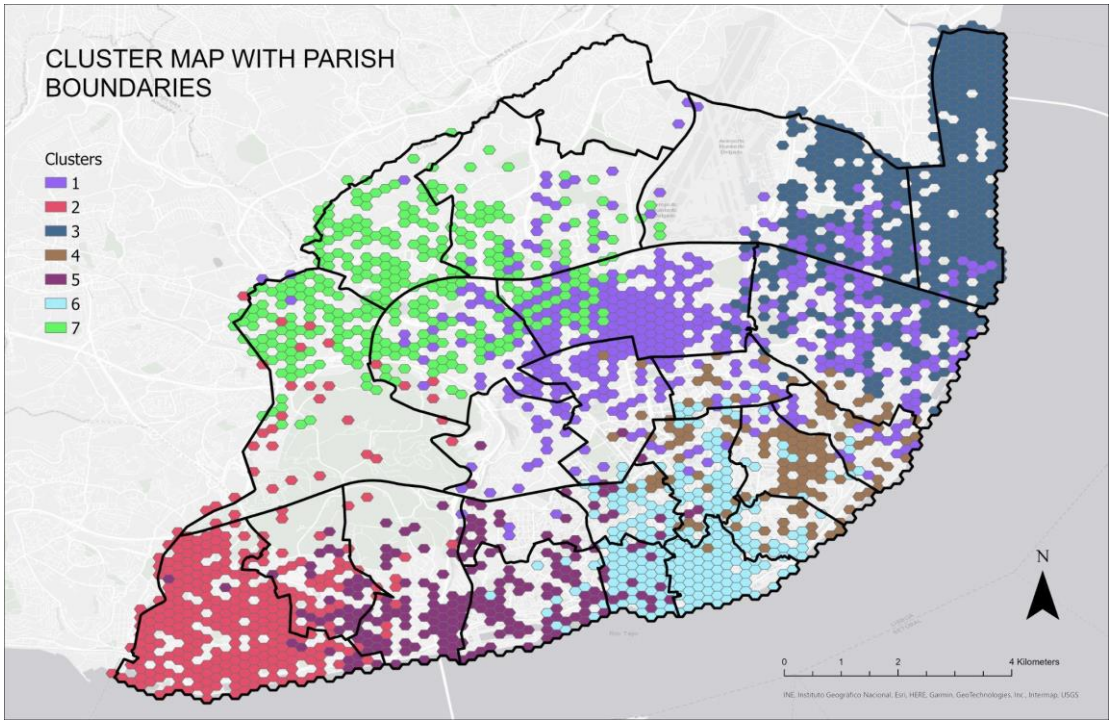


Figure 1.38. The 7 GeoSOM clusters with an overlaid Parish map

The identified regions can now be used as an extra feature for classifying hexagons if they are appropriate for placing a POI or not.

4.2.2 Region Comparisons Using GeoSOM Suite

This section provides a further granular comparison of the different regions identified by GeoSOM. Figure 1.39 and 1.40 highlight cluster six and compares **male-age-c_d_r_f_b** with **female-age-c_d_r_f_b** using the principal coordinate plots (PCP) visualization tool. PCP allows inference of high-dimensional data in a simplified view and highlights how the variables are distributed across different ranges. For instance, in the figure, for cluster six (as per Figure 1.40), it is seen that the relevancy of all five business keywords is high for both male and female groups, while the age group is also observed as similar. However, there is a higher concentration of social media activity by males than females in the given region. A similar comparison is made but for a different region, cluster 7 (as per Figure 1.40), in Figure 1.40. The given region includes some hexagons where the female population dominates the male population. Overall, the population (online) is much more widely distributed than in the previous region for cluster 7. On the contrary, the relevancy of all five business keywords as a mean can be observed as lower than cluster 6.

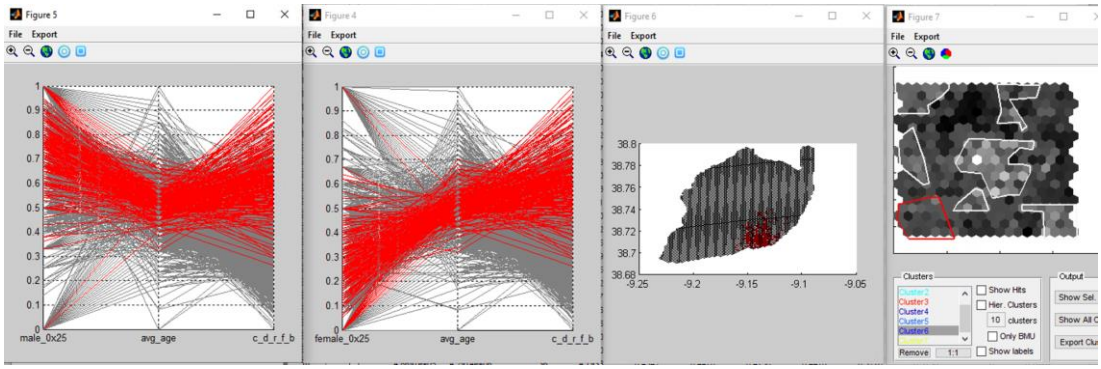


Figure 1.39 Principal Component Plot for **female-age-c_d_r_f_b** variables for cluster 6 (as per Figure 1.38)

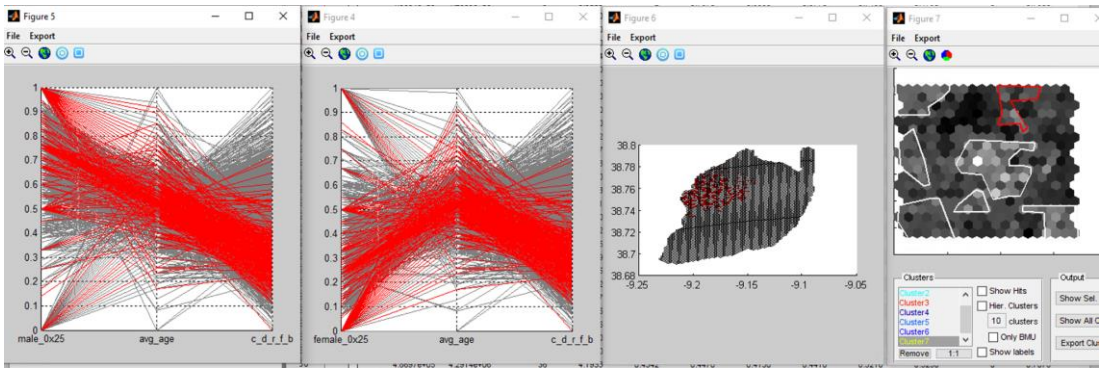


Figure 1.40 Principal Component Plot for **male-age-c_d_r_f_b** variables for cluster 7 (as per Figure 1.38)

Figure 1.41 and 1.42 compare the regions defined by clusters 6 and 3 (as per Figure 1.38) using PCPs. This comparison describes the influence of particular business keywords and their relation with age, i.e., **restaurant-bar-age.** In cluster 6, or the famously known downtown region of the city, a higher concentration of hexagons exists where restaurants and bars dominate the regions. There are no hexagons with 0 relevancy to the two business keywords. In Figure 1.42, cluster 3 (as per Figure 1.38) is mainly under the influence of parish Parque das Nacoes and has a massive drop in relevancy for bars and restaurants and with some influence of older age groups.

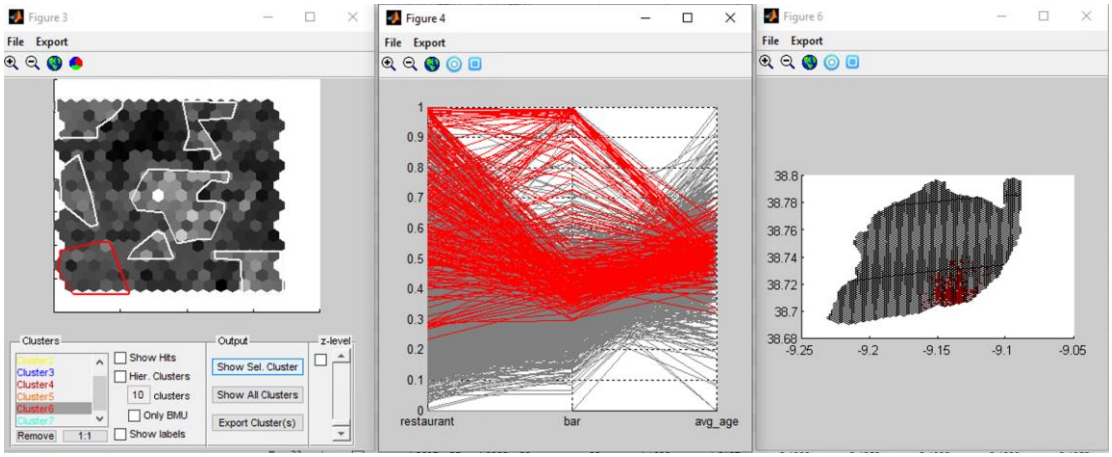


Figure 1.41 Principal Component Plot for **restaurant-bar-avg_age** variables for cluster 6 (as per Figure 1.38)

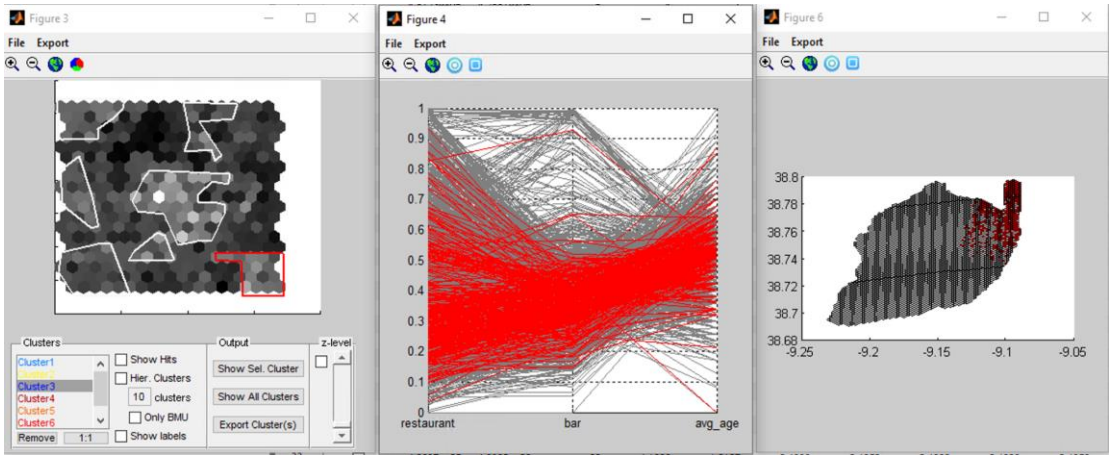


Figure 1.42 Principal Component Plot for **restaurant-bar-avg_age** variables for cluster 3 (as per Figure 1.38)

4.3 POI Prediction Using Random Forests

An 80/20 percentage split did the train/test split in a stratified way to ensure an even representation of all the classes in train and test subsets. Table 1.8 and Figure 1.43 show the classification performance on test data and the feature importance extracted from the model, respectively.

An overall average of 83% for the f1-score signifies a suitable classifier for the two classes. The model is 85% accurate in predicting irrelevant hexagons for the five business keywords and 80% in identifying the relevant hexagons. The support column in Table 1.8 shows how many samples of each class were present for testing. The model is helpful given the nature of dynamic data. For instance, this model can be reused every few months to observe how some hexagons become less/more relevant as their underlying variables change. A geographical map highlighting relevant hexagons allows government administrators to observe which

hexagons attract a specific business and what variables can be worked on to make other sets of hexagons attractive.

On the contrary, a POI manager can quickly monitor what potential locations/hexagons would be relevant for their business, study more critical variables, and make an informed decision accordingly. By not predicting specific points for a POI, this methodology of predicting hexagons/neighborhoods provides more flexibility and error tolerance while considering the ever-changing dynamics of geographical properties in a region.

Class / Metric	Precision	Recall	F1-Score	Support
0 (Non-Ideal)	86%	85%	85%	543
1 (Ideal)	79%	80%	80%	383
Overall Accuracy			**83%**	926

Table 1.8 Classification performance of the RF model on test data (test POIs)

In Figure 1.43, it is interesting to see spatial competition and POI reviews variables together account for more than 50% classification strength of the model. The “cluster” variable, an output of GeoSOM, provides about 5% of predictive power to the model. Due to the highly dominating male gender in the dataset, the gender features provide the minimum predictive power.

The final map of predictions can be seen in Figure 1.44. The map shows the hexagons that were predicted as Ideal in green. However, initially, they were marked as “Non-Ideal” since they did not have at least 1 POI inside them. It can be seen that most of the recommendations have moved towards the Norht-West-Center-East of the city except the South. This is because most POI regions (or hexagons) were predicted and already presented in most coastal parishes and were not discoveries. A cluster of markers is also seen towards the Airport of Lisbon.

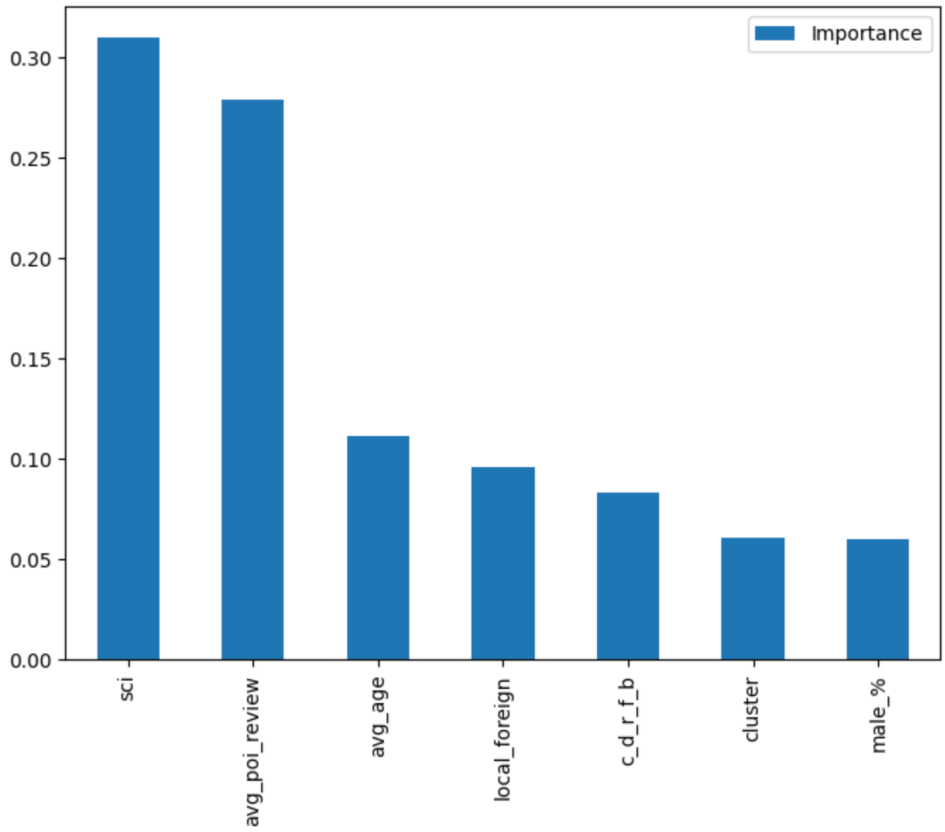


Figure 1.43 Feature importance retrieved from the RF model

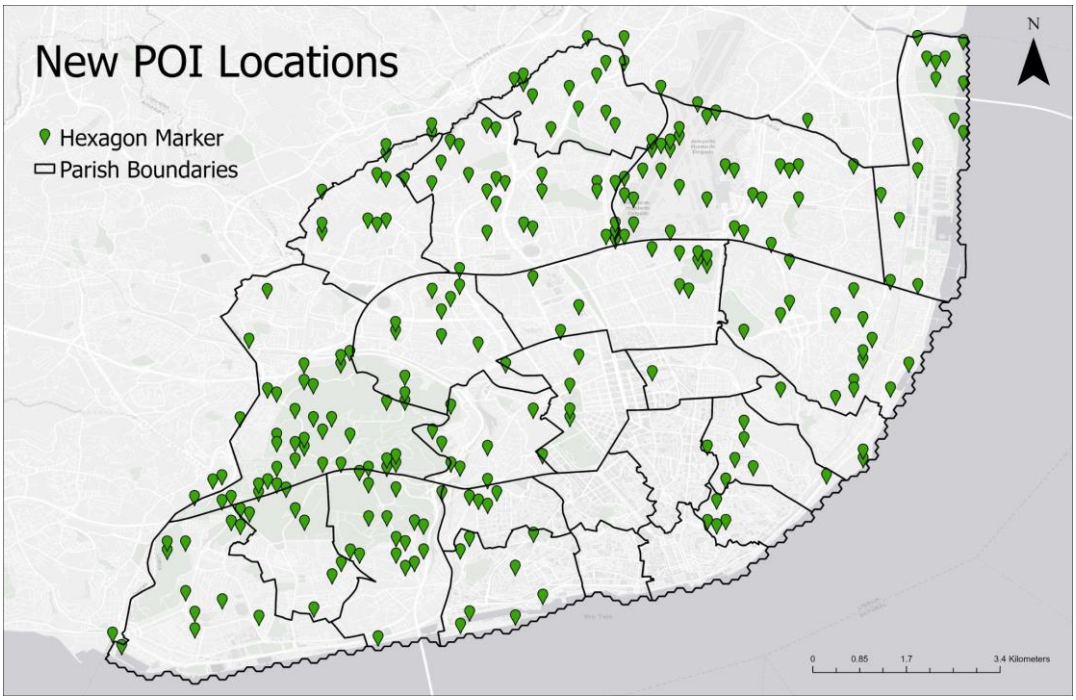


Figure 1.44 Newly predicted hexagonal regions are identified with green markers

5. FINAL DISCUSSION

5.1 Answering Research Questions

5.1.1 Can neighborhood features predict POI locations?

On observing the classification results, it is evident that the model could infer certain variables that were useful in distinguishing the hexagons with at least one POI compared to hexagons with 0 POIs. Site selection is a complex problem statement as the site (POI in this case) and its region is influenced bi-directionally. Many POIs in a specific neighborhood can influence the demographics of a region. On the contrary, a particular segment of demographics or other spatial constraints (like elevated region) can influence the success of POIs. The objective of this research work was to study the variables only in one direction, i.e., can certain spatial/non-spatial variables influence POIs site selection? The research methodology here enables business managers to adapt flexibly to new/old data, change cities, and control weights of variables as required.

The idea of using spatial characteristics like spatial competition and POI proximity and inferring different non-spatial variables at a hexagonal level was to mitigate the effect of demographic bias that might come in from using social media data. Similarly, extracting topics from social media posts has its challenges, like ambiguous meanings of words. For instance, the word “drink” can have different meanings for different users. Despite these challenges, the topic modeling results were made interpretable by observing the relevancy of the keyword in a region instead of using the words themselves. Additionally, since topic modeling algorithms can suffer from generating good topics from small sentences, hexagonal grids were used to address this issue by merging all sentences at a hexagonal level and then extracting topics.

Most of the new recommendations are clustered towards the northern parishes, the airport, and across parks of Benefica. As per the promoting factors of new POI locations **[43],** accessibility is crucial. A similar trend can be seen in Figure 1.44, where the POI locations are recommended in small clusters. Existing POIs are also a hindering factor for the birth of new POIs **[43],** which aligns with the results observed in Figure 1.44, since the new POIs' locations are now formed away from the parishes near the traditional downtown of Lisbon (at the coastal regions), for instance, Santa Maria Maior and Santo Antônio. Figure 1.45 shows the effect of dropping variables for classification in the order of highest to lowest importance. It can be seen that each of the neighborhood variables plays a vital role in predicting the POI locations.

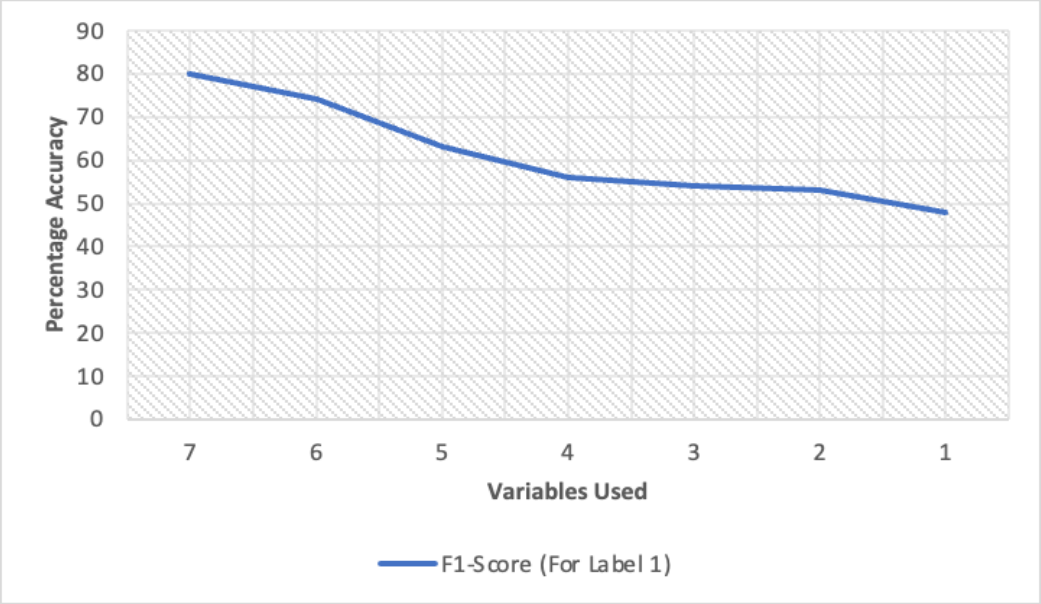


Figure 1.45 Line chart showing the drop in F1-Score for Label 1 on dropping variables.

The feature importance chart from Figure 1.43 shows the SCI variable largely influencing most of the decision-making for site selection. It is meaningful because a region with a high density of related POIs might not be the best place to create a new shop. By using multidimensional data, it is possible to address issues that individually exist with each variable. Furthermore, the formation of clusters from GeoSOM taking the form of a parish also signifies how varying the people's interest towards food & drinks POI over social media within a city. Overall, based on classification performance and the shape of clusters formed, it can be said that the variables chosen for this study were significantly helpful in deciding whether a certain hexagon provides ideal conditions for a POI or not.

5.1.2 Does GeoSOM clustering provide an advantage for predicting POI locations?

It is evident from the analysis presented in Figure that the variable of clusters contributes a relatively low proportion of predictive power, at approximately 5%. However, it is essential to note that the selection of the classification algorithm, in this case, Random Forest, may also impact determining the most influential features. Additionally, as the dataset size may vary, likely, the relative importance of variables may also fluctuate. Furthermore, the output of clusters is contingent upon the training parameters utilized in the GeoSOM algorithm. The utilization of cluster analysis in this context extends beyond mere commercial considerations.

The clustering of regions based on demographics and interests can provide valuable insights for government and local administration in formulating people-centric policies and programs. Additionally, it can aid in targeted marketing efforts for businesses and inform resource allocation in healthcare and urban planning areas. Furthermore, it can contribute to advancing

knowledge in the social sciences by illuminating regional patterns of behavior and cultural norms.

GeoSOM cluster analysis provides a significant advantage in traditional site selection. Identifying clusters with similar demographics and behaviors can help retail managers identify regions most likely attractive to their target audience. For example, cluster analysis can enable retailers to identify areas with high traction and purchasing power, allowing them to prioritize those areas for opening new stores.

5.3 Limitations

Since the objective of this study was to observe how spatial and non-spatial features of a neighborhood can contribute to the success of a POI, the implementation of this methodology can become complicated very quickly. To minimize this complexity and make the approach reproducible, there are possibilities that the aforementioned steps can suffer from a few limitations in certain situations. To begin with the spatial unit, for instance, the size of the hexagonal grid may not accurately capture the spatial dynamics of POI site selection. It is also known that any study is only as good as its dataset; hence, this study may only partially represent the entire social media behavior perfectly due to drawbacks of topic modeling and gender-age prediction algorithms. Similarly, the current dataset size is only a proxy to the real population and does not represent the variation in samples that a real population might have at a city level.

The study is also limited to the number of variables extracted basis the social media data. There might be many more variables that were also as important but should have been included as a part of this study, for instance, user demand and economic variables from census data. The study's use of social media data as a source of information about the user's interest within the hexagonal grid may not be accurate or reliable because the content of the posts may not have a direct connection to the location where they were posted. The use of social media data can also bring in bias, as in this work, for the skewed distribution of male and female users and not having any information regarding certain age groups.

5.4 Future Scope

It would be interesting to extend this work to other cities by using the exact methodology can compare how the importance of variables changes for predicting new POI locations. A more robust site selection algorithm is also required to include demand prediction variables. This can only be made possible if there is access to users' data of what they search on Google Maps, for instance.

For a fair comparison, other predictive algorithms can also be used for the final step of POI prediction. Another important variable can be the inclusion of temporal dimensions in the data. Observing how POIs change over time or how certain neighborhoods have evolved using older social media posts can enable analysts to see where is a certain neighborhood heading culturally, and these demands can change.

6. CONCLUSIONS

This research aimed to study the influence of spatial and non-spatial variables extracted from social media on POI site selection. Data sources for this work were only limited to online social media platforms, namely, Twitter, Flickr, Instagram, and Google POI Reviews. A total of 9 variables were extracted after data cleaning and feature engineering. POI proximity, SCI, c_d_r_f_b, and clustered regions were extracted by feature engineering. Topic modeling enabled the representation of hexagons using the percentage similarity of the five business keywords.

The results showed that certain variables helped distinguish hexagons with at least one POI from those with 0, with an F1-Score of 83%. The Spatial Competition Index (SCI) variable significantly influenced the decision-making for site selection. Additionally, the GeoSOM algorithm was used to form clusters, signifying the people's interest in food & drink POIs over social media. However, the significance of the regions extracted from GeoSOM did not seem huge as the variable importance of "clusters" is approximately 5%. Overall, it can be said that spatial variables play an important role in deciding an optimal location for site selection. The accessibility of certain POIs and the non-availability of alternative options in a neighborhood makes some categories, like restaurants, bars, or cafes, more likable than others. Sociodemographic factors like local/foreign and age can also be seen to have an impact on classification results, as seen in **[11].**

The research methodology used in this study enables business managers to adapt flexibly to new/old data, change cities, and control the weights of variables as required. The use of the GeoSOM algorithm provides a deeper understanding of people's interests and can be helpful in the future for similar studies. This research's use of available social media data sets an example of how this methodology can be used in other cities.

References

1. Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The Collaborative Image of The City: Mapping the Inequality of Urban Perception. *PLoS ONE*, *8*. doi:10.1371/journal.pone.0068400

2. Peng, X., Bao, Y., & Huang, Z. (2020). Perceiving Beijing's 'city Image' across different groups based on geotagged social media data. *IEEE Access*, *8*, 93868–93881. doi:10.1109/ACCESS.2020.2995066

3. Sood, G., & Laohaprapanon, S. (2018). *Predicting Race and Ethnicity From the Sequence of Characters in a Name*. Retrieved from http://arxiv.org/abs/1805.02109

4. Dhomne, A., Kumar, R., & Bhan, V. (2018). *Gender Recognition Through Face Using Deep Learning*. *132*, 2–10. doi:10.1016/j.procs.2018.05.053

5. Moln r, Z., & Bartha, S. (2007). Adrienne Ortmann-n6 Ajkai 12) & Szilvia R6v TM 1) Institute of Ecology and Botany of the Hungarian Academy of Sciences. *Folia Geobotanica*, Vol. 42, pp. 225–247.

6. Reilly, W.J. The Law of Retail Gravitation; University of California: Oakland, CA, USA, 1931.

7. Cachinho, H. (2014). Consumerscapes and the resilience assessment of urban retail systems. *Cities*, *36*, 131–144. doi:10.1016/j.cities.2012.10.005

8. Tao, R., Strandow, D., Findley, M., Thill, J. C., & Walsh, J. (2016). A hybrid approach to modeling territorial control in violent armed conflicts. *Transactions in GIS*, *20*, 413–425. doi:10.1111/tgis.12228

9. Henriques, R., Bacao, F., & Lobo, V. (2012). Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems*, *36*, 218–232. doi:10.1016/j.compenvurbsys.2011.11.003

10. Cohn, A. G., & Gotts, N. M. (1995). *ThèEgg-Yolk' Representation Of Regions with Indeterminate Boundaries*.

11. Psyllidis, A., Yang, J., & Bozzon, A. (2018). Regionalization of social interactions and points-of-interest location prediction with geosocial data. *IEEE Access*, *6*, 34334–34353. doi:10.1109/ACCESS.2018.2850062

12. Han, S., Jia, X., Chen, X., Gupta, S., Kumar, A., & Lin, Z. (2022). Search well and be wise: A machine learning approach to search for a profitable location. *Journal of Business Research*, *144*, 416–427. doi:10.1016/j.jbusres.2022.01.049

13. Ouyang, J., Fan, H., Wang, L., Yang, M., & Ma, Y. (2020). Site selection improvement of retailers based on spatial competition strategy and a double-channel convolutional neural network. *ISPRS International Journal of Geo-Information*, *9*. doi:10.3390/ijgi9060357

14. Kogure, K., & Takasaki, Y. (2019). GIS for empirical research design: An illustration with georeferenced point data. *PLoS ONE*, *14*. doi:10.1371/journal.pone.0212316

15. de Andrade, S. C., Restrepo-Estrada, C., Nunes, L. H., Rodriguez, C. A. M., Estrella, J. C., Delbem, A. C. B., & de Albuquerque, J. P. (2021). A multicriteria optimization framework for the definition of the spatial granularity of urban social media analytics. *International Journal of Geographical Information Science*, *35*(1), 43–62. doi:10.1080/13658816.2020.1755039

16. McKenzie, G., Janowicz, K., & Adams, B. (2014). A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography and Geographic Information Science*, *41*, 125–137. doi:10.1080/15230406.2014.880327

17. Owen, S. H., & Daskin, M. S. (1998). Strategic facility location: A review. *European Journal of Operational Research*, *111*(3), 423–447. Retrieved from https://EconPapers.repec.org/RePEc:eee:ejores:v:111:y:1998:i:3:p:423-447

18. Harris, R., Sleight, P., and Webber, R., 2005. Geodemographics, GIS and neighbourhood targeting. Chichester: Wiley

19. Pennacchiotti, M., & Popescu, A.-M. (2021). A Machine Learning Approach to Twitter User Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 281-288. https://doi.org/10.1609/icwsm.v5i1.14139

20. Wang, W., Wang, L., Wang, X., & Wang, Y. (2022). Geographical Determinants of Regional Retail Sales: Evidence from 12,500 Retail Shops in Qiannan County, China. *ISPRS International Journal of Geo-Information*, *11*. doi:10.3390/ijgi11050302

21. Liu, J., Meng, B., Wang, J., Chen, S., Tian, B., & Zhi, G. (2021). Exploring the spatiotemporal patterns of residents' daily activities using text-based social media data: A case study of beijing, china. *ISPRS International Journal of Geo-Information*, *10*. doi:10.3390/ijgi10060389

22. Wang, L., Fan, H., & Gong, T. (2018). The consumer demand estimating and purchasing strategies pptimizing of FMCG retailers based on geographic methods. *Sustainability (Switzerland)*, *10*. doi:10.3390/su10020466

23. L. Lynch, The Image ofthe City. Cambridge MA, USA: MIT Press, 1960

24. Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*. doi:10.3389/fsoc.2022.886498

25. Xu, M., Wang, T., Wu, Z., Zhou, J., Li, J., & Wu, H. (2016). *Demand driven store site selection via multiple spatial-temporal data*. doi:10.1145/2996913.2996996

26. Al Zamal, F., Liu, W., & Ruths, D. (2021). Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1), 387-390. https://doi.org/10.1609/icwsm.v6i1.14340

27. Wang, L., Fan, H., & Wang, Y. (2018). Site Selection of Retail Shops Based on Spatial Accessibility and Hybrid BP Neural Network. *ISPRS International Journal of Geo-Information*, *7*(6). doi:10.3390/ijgi7060202

28. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved from http://arxiv.org/abs/1810.04805

29. Cheng, E. W. L., Li, H., & Yu, L. (2007). A GIS approach to shopping mall location selection. *Building and Environment*, *42*, 884–892. doi:10.1016/j.buildenv.2005.10.010

30. Maulana, I., & Maharani, W. (2021). *Disaster Tweet Classification Based on Geospatial Data Using the BERT-MLP Method*. 76–81. doi:10.1109/ICoICT52021.2021.9527513

31. Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., … Kurzweil, R. (2019). *Multilingual Universal Sentence Encoder for Semantic Retrieval*. Retrieved from http://arxiv.org/abs/1907.04307

32. Caminero, L. M., & McGarrigle, J. (2022). Socio-spatial negotiations in Lisbon: Reflections of working-aged lifestyle migrants on place and privilege. *Population, Space and Place*. doi:10.1002/psp.2613

33. Cao, D., Zeng, K., Wang, J., Sharma, P. K., Ma, X., Liu, Y., & Zhou, S. (2021). BERT-Based Deep Spatial-Temporal Network for Taxi Demand Prediction. *IEEE Transactions on Intelligent Transportation Systems*. doi:10.1109/TITS.2021.3122114

34. Jaradat, S., & Matskin, M. (2019). On Dynamic Topic Models for Mining Social Media. In N. Agarwal, N. Dokoohaki, & S. Tokdemir (Eds.), *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining* (pp. 209–230). doi:10.1007/978-3-319-94105-9_8

35. Brandt, J., Buckingham, K., Buntain, C., Anderson, W., Ray, S., Pool, J. R., & Ferrari, N. (2020). Identifying social media user demographics and topic diversity with computational social science: a case study of a major international policy forum. *Journal of Computational Social Science*, *3*, 167–188. doi:10.1007/s42001-019-00061-9

36. Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, *163*, 1–13. doi:10.1016/j.knosys.2018.08.011

37. Cai, G., Sun, F., & Sha, Y. (2018). Interactive visualization for topic model curation. *CEUR Workshop Proceedings*, *2068*.

38. Mazhi, K. Z., Suryana, L. E., Davi, A., & Dewi, W. R. (2020). *Site selection of retail shop based on spatial analysis and machine learning*. 399–404. doi:10.1109/ICACSIS51025.2020.9263156

39. Sun, Y., Yin, H., Wen, J., & Sun, Z. (2020). *Urban Region Function Mining Service Based on Social Media Text Analysis*. *2020-August*, 170–177. doi:10.1109/ICSS50103.2020.00034

40. Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., & Mascolo, C. (2013). *Geo-Spotting: Mining Online Location-based Services for Optimal Retail Store Placement*. doi:10.1145/2487575.2487616

41. Kwan, M. P. (2012). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers*, *102*, 958–968. doi:10.1080/00045608.2012.687349

42. Montezuma, J., & McGarrigle, J. (2019). What motivates international homebuyers? Investor to lifestyle 'migrants' in a tourist city. *Tourism Geographies*, *21*, 214–234. doi:10.1080/14616688.2018.1470196

43. Wu, M., Pei, T., Wang, W., Guo, S., Song, C., Chen, J., & Zhou, C. (2021). Roles of locational factors in the rise and fall of restaurants: A case study of Beijing with POI data. *Cities*, *113*. doi:10.1016/j.cities.2021.103185

44. Aguilar, D. P., Barbeau, S. J., Labrador, M. A., Perez, A. J., Perez, R. A., & Winters, P. L. (2007). Quantifying position accuracy of multimodal data from global positioning system-enabled cell phones. *Transportation Research Record*, 54–60. doi:10.3141/1992-07

45. Cocola-Gant, A., & 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. doi:10.1177/0308518X19869012

46. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03