NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Using *Flickr* to Identify and Connect Tourism Points of Interest

The case of Lisbon, Porto and Faro

Margarida do Carmo Duarte Pereira

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

USING *FLICKR* TO IDENTIFY AND CONNECT TOURISM POINTS OF INTEREST

by

Margarida do Carmo Duarte Pereira

Dissertation report presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Business Analytics

Co Supervisor: Prof Doutor Flávio Luís Portas Pinheiro

Co Supervisor*:* MSc Nuno Tiago Falcão Alpalhão

November 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading
to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 18th November 2022

ACKNOWLEDGEMENTS

First, I would like to thank my parents for always supporting my decisions, and for giving me all the conditions I needed to complete my master's degree. To all my friends and family for encouraging me and always being by my side, even if I was often absent.

To my master's colleagues, for not letting me give up and for being understanding about my professional situation; and to my companies, that allowing me to work and finish my studies without creating any adversity.

Last but not least, to Professor Flávio Furtado for sharing his knowledge and ideas that have enriched my thesis and, above all, for his patience and for pushing me along this time to always do more and better; and to Professor Nuno Alpalhão for his insights and for helping my work to achieve more interesting results that are closer to reality.

ABSTRACT

Understanding the movement of tourists helps not only the management of cities but also to enhance the most attractive places. The growth of people in social media allows us to have greater access to information about user preferences, reviews, and shared moments. Information can be used to study tourist activity.

Here, it is used geo-tagged photographs from the social media platform *Flickr*, to identify the locations of tourists' Points of Interest in Lisbon, Porto and Faro and quantify their relationship from the user's co-occurrence in the identified points.

The results show that, using standard clustering methods, it is possible to identify likely candidate Points of Interest. The association of the Points of Interest from users' social media activity (i.e., posting of photos) results in a non-trivial network that breaks geographical proximity. It was found that, in all the cities under study, historical places (such as churches and cathedrals), viewpoints and beaches are captured.

KEYWORDS

Tourist; Points of Interest; Geo-tagged photos; Clustering; Network Analysis

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

API	Application Programming Interface
DBSCAN	Density-Based Spatial Clustering of Applications with Noise algorithm
CRS	Coordinate Reference System
Eps	*Epsilon*
MinPts	Minimum number of data points in a cluster to the *DBSCAN* algorithm
NMI	Normalized Mutual Information
PCR	Projected Coordinate System
PoI	Point of Interest

1. INTRODUCTION

We have all seen or been tourists at some point in our lives. It is part of human beings to go out and discover new places. The movement of people around the world fuels the creation of new businesses, new jobs, and the increase in the production of goods and services. Tourism is considered one of the main economic activities worldwide (Dogru & Bukut, 2017; Pender & Sharpley, 2005). However, it is necessary to ensure that tourism is managed and that the conditions needed for it to continue to grow are created.

Identifying the most likely paths of tourists within cities can help the planning and management of the cities. Be it through the creation of new transport routes or the optimization of existing ones, improving access and communication about places to visit, or the development of target marketing campaigns, a smart data-driven approach to tourism has the potential to foster the growth of the industry (Habeeb & Weli, 2020). However, developing a smart and data-driven approach to tourism is challenged by the need to collect large volumes of data that can accurately represent tourists' activities in different locations.

Nowadays, social media is present in everyone's life. People share photographs from where they are and allow their followers to visualize them. These traces of user activities can reveal our identity and preferences, which allows users to communicate where, when, and with whom they have been, which activities they have done, what they have seen, and their likes and dislikes. Indeed, social media is changing the tourism culture, as well as how travelers decide on a new destination (Zeng & Gerritsen, 2014). This trove of data, which is available through users' social media activity public amounts, provides us with the information to reconstruct their behavior and trajectories.

Portugal is one of the European destinations and is also an example of exponential tourism growth. Like most European countries, it is a historic location that combines multiple viewpoints, museums, football stadiums, monuments, hotels, restaurants, and much more. These different Points of Interest are distributed unevenly across the country's geography, and tourists choose them according to their preferences, which might not be related to geographical proximity. Here the dissertation is focused on three particular cities - Lisbon, Faro and Porto – to try to answer the following research questions:

Q1 – Can we identify Points of Interest for tourists in Lisbon, Faro and Porto using the geo-location of photos and standard clustering methods?

Q2 – From a network perspective, what characterizes the paths taken by tourists through those Points of Interest, and what differentiates the cities under study?

The dissertation is divided into 7 chapters: the first one, presents some context and the main goals of the study; the second, where a tourism context is deepened and some related studies are reviewed; the third focus on the data extraction and treatment; the fourth and fifth where the techniques used are explained and both clustering and network analysis are explored; the sixth, presents the results obtained; and to finish, the seventh and eighth chapters list the main conclusions, limitations, and recommendations for future work.

2. RELATED WORK

2.1. SMART TOURISM

In many countries, Tourism is one of the main economic activities (K. Walton, 2012).

In Portugal, tourist activity revenues reached €18.4 billion in 2019, weighing 15,3% of the GDP. This ratio makes Portugal the 5th country with the largest contribution of tourism to GDP (INE, Estatísticas do Turismo - 2019, 2020; INE, Estatísticas do Turismo - 2020, 2021).

To remain competitive in the tourism market, countries, regions, and cities need to embrace new technologies and processes to continually innovate and grow. In that sense, Smart Tourism presents an important solution of optimal information flows – tourist's expenses, their participation in society, their movement patterns, etc – and supports efficient and effective policy and governance decision-making. Smart Tourism is a tourism that "integrates information with traditional and new forms of information dissemination and highlights the role of accurate and personalized information designed to meet the demands of tourists in the era of fast-growing wants for information and communication" (Li, Hu, Huang, & Duan, 2016).

Smart Tourism involves a set of best practices using the application of information and technology that we can see daily. In terms of accessibility, we can find cities with city routes, accessible beaches, and infrastructure, not to mention accessibility to information through city guides, tours, and tourist information offices. Sustainable transportations, alternative means of transportation and Natural preservations are also signs of Smart Tourism, since guarantee the sustainability of the cities. Furthermore, activities such as city walks, wine tastings and festivals promote cultural knowledge, being essential points in the quality of a trip and in the improvement of tourist satisfaction (Directorate-General for Internal Market, 2020). These aspects could be managed if the cities know the preferred destinations and their association.

Many aspects influence a tourist's choice of destination: "the size and expenditure of tourist time budgets; personal motivations, interests, travel group composition; and tourist knowledge of the destination" (Lew & McKercher, 2006). Hence, before any marketing campaign to make yourself known, to recommend new places, or to offer unique promotions to attract tourists, it is necessary to understand the reasons that bring tourists to a particular destination, and how destinations are related.

As social networks are increasingly used, and as they have a big impact on tourism, they are also a great source of information about tourists' behavior patterns. Social media as *Twitter*, *Instagram*, *Facebook*, etc, has enabled tourists to share their experiences, playing a significant role to help tourists plan their travels (Zeng & Gerritsen, 2014).

Using the photos tourists share on their social media accounts also helps to identify the most visited places, when are visited, and detect tourists' movement, as will be seen in the next section. Then, that information can be used to construct and optimize Smart Tourism.

2.2. TOURIST MOVEMENT PATTERNS

Understanding tourists' movement patterns through their social media activity is an efficient approach to understanding which locations in a city are more attractive and planning tourist flows. Indeed, many past studies have investigated the movement patterns of tourists in recent years. In particular, these studies have used Geo-tagged photos (Figure 2.1).

Figure 2.1 – World Map highlighting places studied in previous works. Highlighted in the map are the cities of United States of America (New York, San Francisco, Washington, Chicago, San Diego, Los Angeles and Grand Canyon), Portugal (Lisbon, Porto and Faro), Spain (Barcelona, Boí Valley, Madrid and Toledo), United Kingdom (London), France (Paris), Italy (Rome), Germany (Berlin), Danube River, The Netherlands (Amsterdam), Taiwan (Nantou), and Hong Kong.

It was started by founding studies based on the 3 cities that were the focus of the dissertation: Lisbon, Faro, and Porto.

In Lisbon, several studies about tourism and tourist behavior have already been carried. In 2015 was elaborated a quantitative evaluation of the tourist experience in Lisbon (Sarra, Zio, & Cappucci, 2015), aimed at measuring the satisfaction of tourists in the city. The study surveyed a group of 300 foreign and used an Item Response Theory[1] approach to the data. As a result of the study of various satisfaction hypotheses (if it depended on the characteristics of the place, the behavior of the tourist, and its location, among others), it was concluded that the satisfaction and the preferred places depend on the type of travel (number of nights, the goal of the travel, tourists' age, etc) – The tourists' satisfaction usually increase with the number of days spent in Lisbon, being the ideal number of days to visit the city between 4/5; and while people between the ages of 35-50 prefer places like Baixa and Chiado, which are also likely to be re-visited, young people visit Bairro Alto and Belém.

[1] IRT – based – uses several statistics models to obtain results through based on the relationship between individuals' performances

Moreover, in 2020 a similar study looked to identify the key determinants to a tourism destination choice (Pestana, Parreira, & Moutinho, 2019). Using a sample of 460 seniors that visited Lisbon (Portugal), it was concluded the different motivations and behavioral intentions of visiting Lisbon. For example, escape from routine, know new places and different cultures and lifestyles, stimulation of emotions and sensations, and cultural attractions are more valuable reasons to people between 55-60 years, while accessibility and weather are the motivations for older people.

In Faro, Tourism and its environmental implications were studied analyzing the observation of tourist use (Oliveira & Costa, 2020). Oliveira and Costa applied surveys to tourists and locals to determine the distances traveled, as well as ages and perceptions of the environment. It was concluded about the profile of tourism in Faro, where 80% of the tourists travel to Faro for tourism, while 18% for relaxation. Besides that, tourists travel on average for close to 8 nights. The city of Faro is the main attraction, with the beach and the Ria Formosa being destinations where most of the time is spent. From the locals, it was presumed that tourism is not a potential threat to the local economy since it is seen as a safeguard of promising local development, but there is a concern with the loss of environmental quality in the Faro region.

In 2015 a dissertation about the growth in tourist activity was elaborated using document analysis, participant observation, analysis of tourist surveys, and expert panel interviews (Bexiga, 2015). Surveys of tourists were able to identify mostly their profile and what they are looking for in the city, concluding that Faro is mainly visited by foreign tourists – especially European ones, being England the main source. Furthermore, almost half of tourists are over 60 years old, and the Historical Center, Churches, and Commerce are the preferred places to visit.

In Porto, the Sequence and Network Mining of Touristic Routes were studied based on *Flickr* Geotagged Photos (Silva, Campos, & Ferreira, 2019). It was used a database with around $8k$ photos to perform a Sequence Mining, a Social Network to measure relationships and flows between people, groups, and organizations; and a clustering algorithm to fulfil market segmentation. It was found that the most frequent attractions are Ponte D.Luís I and Douro River, being the individual's preferences concentrated around the river, and that the tourists visit around four points during their trips. Regarding the segmentation, it was not found segments different enough to apply specific marketing strategies; however, the data shows the high-level of education of the tourists, as well as an average age of above 25 years old.

However, survey-based studies, such as the most ones mentioned above, present several trade-offs. On one hand, surveys are efficient and allow to get a substantial amount of information per participant allowing to distill many desirable details. On the other hand, participants are dishonest, recruiting can be difficult (small sample size), they are time-consuming and costly to organize, and they offer an answer to a very narrow time window, meaning that the conclusions of a survey-base study have limited temporal validity in quickly changing environments. As such, researchers look for data that can by narrower in detail, but that can offer good proxies to answer questions at study, but with a large volume of observations (hundreds of thousands to millions). With this in mind, it was also examined studies using data extracted from other sources.

In 2015, the authors of "Urban magnetism through the lens of geo-tagged photography" (Paldino, Bojic, Sobolevsky, Ratti, & González, 2015) used information from photographs shared on *Flickr* about a group of different cities to conclude the relationship between world-known cities and points

within the cities themselves. They not only studied the global attractiveness (which cities are most visited) and local (which places within cities are most attractive) but also estimated the flows of visitors and identified the top activity hotspots of each city for both domestic and foreign tourists. To perform it, some distribution functions were used, as well as some particular algorithms – starting by defining who is a resident or tourist through pre-defined criteria; proceeding with a Ribbons visualization[2], to demonstrate the outgoing' fluxes between cities; measuring the ratio between the movement of each tourist's hometowns and the total activity of the respective city (with the total number of photographs taken by residents or not in the city), in order to understand how residents move (relative attractiveness); and calculating density values, using log-normal distributions[3] to represent the spatial distribution within each city. After the analysis, it was concluded that the attractiveness differs depending on the type of tourist - American cities are preferred by domestic's tourists, while European cities are preferred by foreign tourists, as well as that the flow of tourists traveling from Europe to the United States is greater than the opposite. Besides that, the favourite spots in New York City are the Downtown, New York City Hall, and the Metropolitan Museum of Art.

Still focusing on attractiveness, also 6 cities in Italy were investigated using *Flickr* photos to identify tourist behavior and cities attraction (Giglioa, Bertacchini, Bilotta, & Pantano, 2019). Here, the researchers started using an artificial neural network to identify the object of the photos, and then applied a clustering algorithm to find different groups of photos with similar objects and classify each photo into a specific category for each city. Then, behavior and attractiveness maps were created, helping to conclude about the annual trend of photography and which places are more interesting for both Italian and tourists.

As the previous studies, it also used Geo-tagged photos to modeling human mobility patterns in the United Kingdom (Barchiesi, Preis, Bishop, & Moat, 2015). Given the information about the geographic coordinates and the day and time photos were photographed, it was concluded the individual trajectory of each *Flickr* user and, using *DBSCAN* (a clustering algorithm), the researchers obtained different groups of locals in the country according to their geographic location. Then, a map with the different trajectories was created, to easy visualize how each person moves between the pre-defined clusters. After that, the Aggregated mobility was analyzed, and was computed a model that gives the probability of a user making a transition between two specific cities in the United Kingdom.

Since the dissertation aims to study a specific city, it was also explored other successful studies in other similar large cities.

In 2020, the tourist trajectories were estimated in Toledo (Domènech, Mohino, Inmaculada, & Moya-Gómez, 2020). To estimate that, spatiotemporal trajectories were defined for each user giving the number of photographs (more or equal to 2) and the time between these consecutive photographs (less than 2 hours) - and places visited after that time were considered in another sequence. After the trajectories being defined, the different hotspots in the city were evaluated according to the proportion between the number of spatiotemporal trajectories per street and the total number of spatiotemporal reconstructed. Lastly, it was applied two different statistical methods: a rank correlation test (more specifically, Spearman correlation - that studies the statistical dependence

[2] Ribbons Visualization – visualization that demonstrates path flows (Verma & Pang, 2004)

[3] Log-Normal Distribution – Continuous probability distribution

between two variables) and two linear regression models (Ordinary Least Squares—OLS[4], in relative and absolute terms). With the results, it was possible to produce maps with the percentage of visitor of the spatiotemporal trajectories per street segment and location and classify each place according to the tourists' preferences (primary, secondary, complementary, or off the beaten track attractions). Besides that, the correlation test and the OLS regressions allowed to conclude, for example, that retail and commerce are very associated with the trajectories, while residential uses are negatively correlated with that.

Besides Toledo, also New York City was studied. A related research was developed using Geo-Tagged Photos from *Flickr* to quantify tourist behavior patterns (Yang, Wu, Liu, & Kang, 2017). As in the previous case, they defined travel trajectories, but now according to the most popular places and considered that each point presents the geographic and time information, defining tourist trajectory in terms of a sequence of places ordered by the time that they are visited by the tourist. Using the graphs theory, it was identified individual mobility patterns, forming a network. Then, they found 13 motifs (sub-networks) able to describe 87.4% of all dataset. After the analysis of the motifs, they processed a Motif-Based Clustering, measuring the similarity between two users' Euclidean distance to calculate each pair of vectors and the average-link clustering criteria as a function of the pairwise distances of vectors in two clusters. The results were matrices that allow us to identify different clusters, in terms of discrete and consecutive sequence-based motifs, denoting groups of tourists with common travel motifs.

Regarding clustering, other studies also used clustering techniques to achieve their results. Now, two cases focused on the cities of Boí Valley, Spain, and Amsterdam, The Netherlands are present.

Both cities were studied using data from *Flickr* to find different groups of tourists that share the same characteristics in their photographs, as well as define their behavior in these locals. In the Boí Valley investigation (Donaire, Camprubí, & Galí, 2014) they started by defining different categories according to the places' characteristics (Nature, Heritage, Culture, and Tourist Services) and the type of tourist (resident or tourist). Following the study, it was analyzed the photo itself – the zoom, the presence of humans, if it was photographed in interior/exterior, etc).

Then, the researchers were able to conclude about the most photographic categories, the degree of human presence, how the tourist took the photograph (depending on the focus/zooming used), the preferred captured spaces (interior or exterior), and using Ward's method (clustering method) defined 4 different groups of tourists that present similar favorites places (the Global, Scenic, Detail hunters and Monument lovers).

In the Amsterdam case, it was started with a Data Collection phase, where data was extracted, cleaned, and divided among locals and tourists (Karayazi, Dane, & Vries, 2021). The cluster analysis used 2 different methods: *DBSCAN*, to define the tourist clusters and classify the hotspots, where was concluded that the tourists usually visit more museums, churches, and the Amsterdam Central Station; Finally, the OLS (dependent (attractive heritages) and independent (heritage attributes and supporting products)) and GWR (regression analysis across the space) analysis were performed to study the correlation between the places to explain heritage attractiveness.

[4] OLS - method for estimating the unknown parameters in a linear regression model

The above studies are mostly focused on finding the points most visited by tourists. However, there is also possible to find the correlation between these points and construct a Network of Tourist Points of Interest.

In 2018, *Flickr* photos from Beijing were studied to build a spatially embedded tourism hotspot network and study network characteristics (Wu, Huang, Peng, Chen, & Liu, 2018). The researchers started by finding the Points of Interest using Clustering by Fast Search and Finding of Density. Then, a network was constructed generating trajectories per user considering the chronological order in which they visited; after adding up all the points, they assigned tourist frequency as edge weights at the Point of Interest (*PoI*). Then, based on the extraction and topology of the links between the hotspots, the network was built. They concluded that the tourism hotspot network in Beijing is scale-free and small-world. Interconnected triplets tend to be formed by the edges with greater weight values, and a high-weighted edge is often connected by two high-degree vertices.

To create insights regarding travel behaviors in Hong Kong (Vu, Li, Law, & Ye, 2014), also a clustering algorithm was performed – in this case, using the *P-DBSCAN*[5] method – and then, applying the Markov Chain (a stochastic model) and the conditional probabilities properties, create travel patterns and routes done by tourists around the Points Of Interest resulted from clustering.

A similar study was performed on Danube River (Gede & Kádár, 2019) to find its tourist movement. However, in this case, the river was divided into long segments to detect movements between cities – it was formed a time series of data for each user and created a "travel graph" using the river segments as nodes and the edges if a user took photos in both nodes within two days. The result of the graph allows them, using modularity analysis[6], to identify different clusters of destinations.

Despite these studies present useful results, they still present some limitations. First of all, the Lisbon and Faro studies, besides being based on surveys, these studies are built with international tourists' information, so they do not consider Portuguese tourists, who end up having a great weight in tourism in those cities. Secondly, concluding what makes tourists choose them and what makes them more satisfied, gives the information on which places to improve, but does not allow us to conclude how cities can be improved as a whole or how the places can be connected.

Regarding the remaining studies, despite applying different and advanced methods, they do not present a comparative basis nor how to measure that the results are in line with reality. Furthermore, the methodologies are very specific to the cities studied, and needing to be changed if they are to be applied to different areas.

[5] P-*DBSCAN* – an *DBSCAN* derivation to geo-tagged photos

[6] Modularity fAnalysis – measures the strength of a network divided into modules

3. DATA EXTRACTION AND TREATMENT

Flickr (*www.flickr.com*) is one of the largest online photo repositories. It was created in 2004 in Canada, and it helps people to publish and share their photos in an easy and organized way. *Flickr* repository has over 60 million monthly active users and over 100 million photographers. There are users from at least 63 nations that publish and share photographs of all types from all places. The country with more active users is the USA (which constitutes more than 30% of users in the platform), while in Europe, only France (4%) and Germany (5%) stand out (Figure 3.1) with the percentage of users in the platform per country.

When a photo is published, it can be accompanied by a description, a direction for an album or another profile, and a tag. *Flickr* uses these tags to organize and classify its photos, and with that it allows its users to search for specific photos in an easily and simply manner.

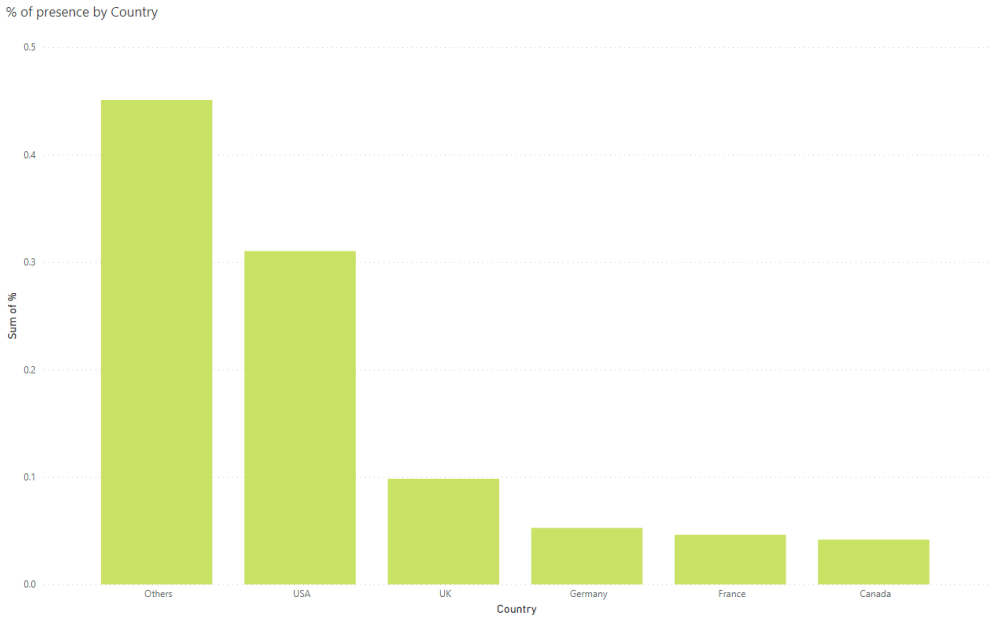


Figure 3.1 – Percentage of user's presence of *Flickr* by country

Flickr allows users to explore and use its data differently. It offers an Application Programming Interface (*API*), for users to access millions of photos and associated information from the photos (e.g., timestamps, geolocation, profiles, tags, comments, etc).

The *API* presents different and varied methods to query *Flickr* and retrieve information. For instance, the activity of the user, the type of cameras used to take the photos, locations, among others (Flickr, Flickr, 2020).

Naturally, the *API* presents some limitations. In order to use it, users need to create a personal key. Calls to the *API* are limited to 3600 searches, a threshold used to lock keys for abusive use of *API* services.

3.1. DATA

Although it is expected that the photos tagged with "Lisbon", “Porto” and “Faro” are taken within the cities, some of them were not.

It was used the *Flickr.photos.search* method from the *Flickr API* to retrieve the *ID* of the photos taken in each of the three cities. To avoid cases where the photo does not have the geolocation information it was set the argument *has_geo* to value 1. Then, to ensure that the data provided were taken in the regions that covered the cities, it was passed values to three more arguments: *lat, lon* and *radius.* These arguments make it possible to make a query that extracts photos geolocated in a radius over a lat-long coordinate. Since the maximum *radius* is limited to $32km$ by the *API*, it was necessary to run 3 different calls for Lisbon; 2 for Porto; and 1 for Faro. Below is listed the lat-long coordinated used as the center of each of the areas:

Lisbon

- Latitude: 38.92, Longitude: -9.2
- Latitude: 38.61, Longitude: -8.94
- Latitude: 38.89, Longitude: -8.80

Porto

- Latitude: 41.01, Longitude: -8.49
- Latitude: 41.23, Longitude: -8.57

Faro

- Latitude: 37.02, Longitude: -7.93

Finally, it was also limited search photos whose *timestamps* was between January 2010 and December 2021.

After that, for each *ID*, the *Flickr.photos.getInfo* was used to get the information about the location where the photo was taken (longitude and latitude), the timestamp when the photo was taken and the *ID* of the user who published it. To deepen the analysis and to study other variants, it was also possible to get information about the home country of the user, using the *Flickr.people.getInfo* method and the user ID. However, this last field presents around 75% missing values.

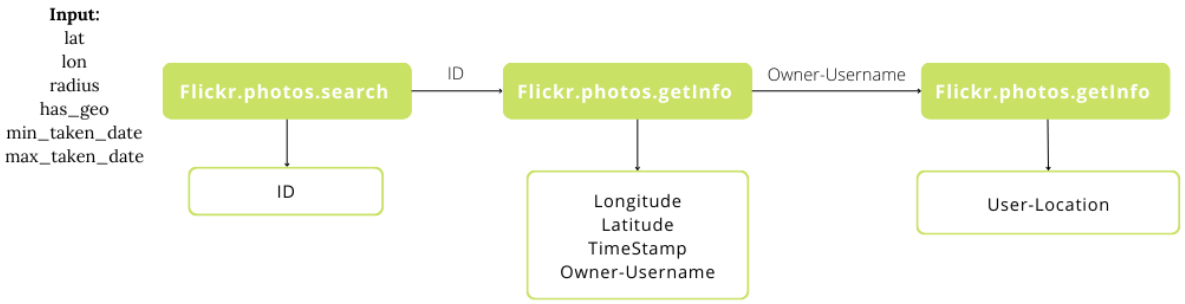


Figure 3.2 – Data download workflow

Hence, completed the data sourcing, each record in the dataset corresponds to the metadata associated with a photograph and includes the following fields:

- *ID*: unique ID of the uploaded photo
- *Owner-Username* - Unique *Username* of the person who uploaded the photo
- *Owner-Location* – Location from the person who uploaded the photo
- *TimeStamp* – The actual time stamp at which a photograph was taken
- *Latitude* – The latitude of where the photo was taken
- *Longitude* – The longitude of where the photo was taken

3.2. DATA CLEANNING

Many photographs were shared by companies or organizations - for example, "Iscte - Instituto Universitário de Lisboa", "Rede Munincipal de Bibliotecas de Almada" or "rtppt" - that are not in the interest of the study. Hence, the data cleaning started by discarding all records that could associate with organizations and companies, which given their high number of posted photographs were candidates for removal through outlier detection methods, such as the Interquartile Range (IQR) method - that was used. In any case, it was validated through *Owner-Username* that the suggested outliers did indeed fit the criteria of an organization or company.

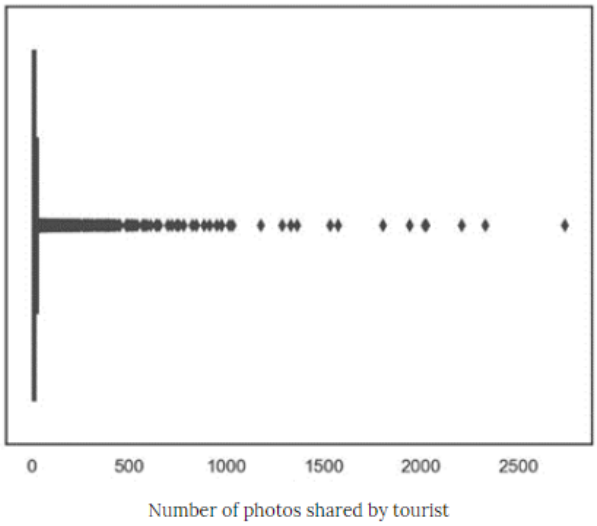


Figure 3.3 – Box Plot with number of photos per *ID* distribution using the *IQR* method

In addition, it was also observed that some users were not tourists but locals. Using the *Owner-Location,* it was possible to identify them using some keywords, such as the names of the cities, municipalities, and specific areas of the studied areas. These users, which represented 8% of the dataset, were also discarded.

Finally, photos taken in some clearly non-touristic spots in the cities of the study were also removed. These include locations of Airports and in Lisbon the *25 de Abril* and *Vasco da Gama* bridges.

The final dataset counted 154*k* photos shared by around 8k users between January 2010 and December 2021. Figure 3.4 shows the number of shared photos shared and the number of different users per year for the 3 cities. The number of tourists remains constant between 2010 and 2019. On the other hand, the data presents a clear decrease in both the number of tourists and photos taken after 2019.

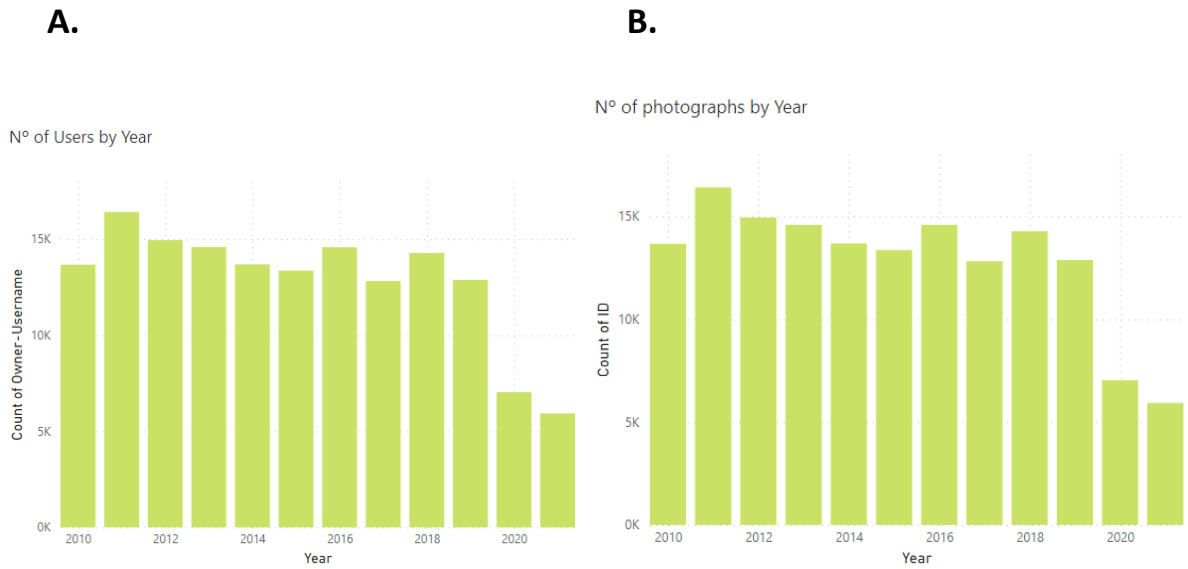
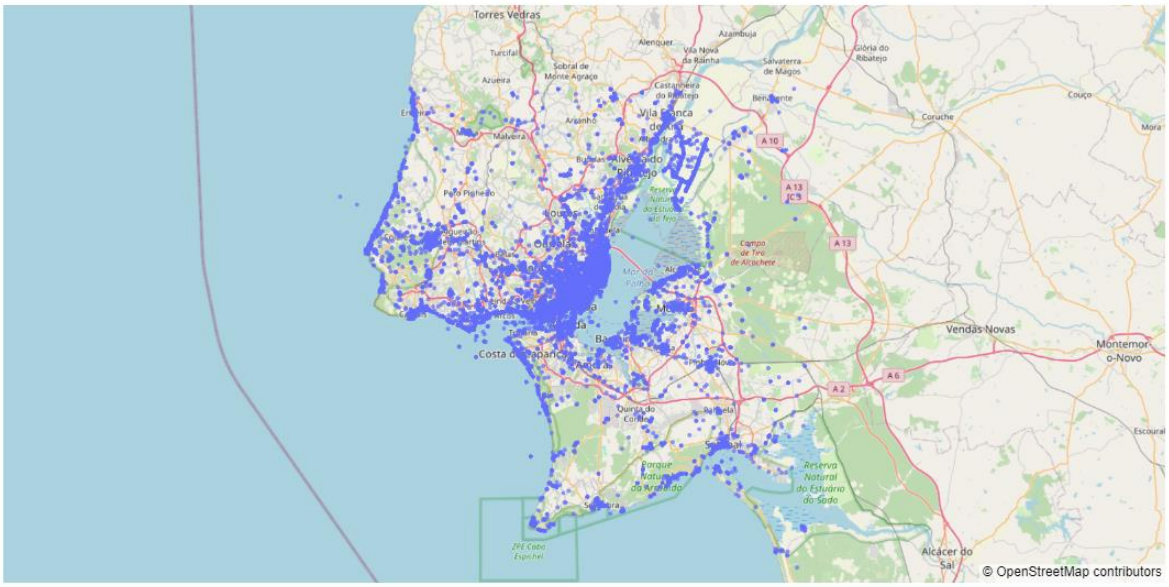


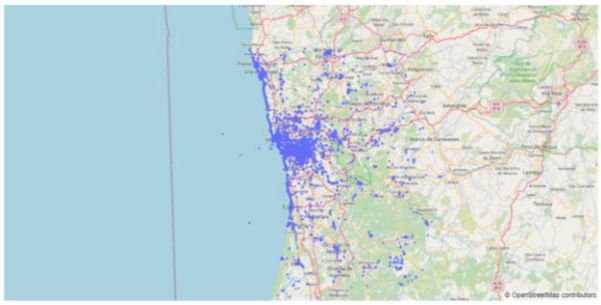
Figure 3.4 – Total Number of tourists that shared on *Flickr* photographs taken in the cities studied (A.) and total number of photographs of the cities studied shared on *Flickr* by year (B.)

Figure 3.5 shows the distribution of observations in Lisbon, Porto and Faro. The observations in Lisbon and Porto do not have the same concentration throughout the areas, presenting different densities between the city centers and the peripheries. On the other side, in Faro, we can observe an identical number of photos throughout the city.

A.



B.



C.

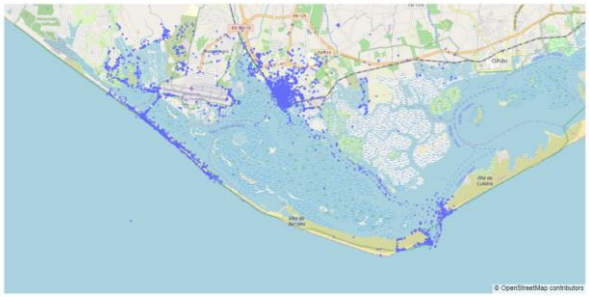


Figure 3.5 – Points distribution in the Metropolitan Area of Lisbon (A.), Metropolitan Area of Porto (B.) and Faro (C.)

4. IDENTIFYING POINTS OF INTEREST

One of the goals of this dissertation is to assess the potential to identify tourist Points of Interest using standard clustering methods from the geolocated photos shared by *Flickr* users.

Clustering is an unsupervised learning method that involves a set of techniques to identify natural groupings within multidimensional data – dividing the population into different groups of data points according to the similarity between these points. It focuses on minimizing the intra-cluster distance (the distance between the Points associated within the same cluster) and maximizing the inter-cluster distance (the distance between clusters).

This type of algorithm is usually used in Data Mining and pattern recognition. It is used to define Marketing Segmentations, Diseases classification, among others (Gorunescu, 2013). Clustering can also be used to identify, using geographic coordinates, locations with a remarkable number of observations.

There are several metrics to be used in a clustering algorithm: the Euclidean distance, Manhattan, Cosine, among others, that are chosen according to the type of data it will be dealing with. When talking about geographic coordinates, extra care is needed since calculating the distance between two points using latitude and longitude is not 100% correct – the Coordinates Reference System (*CRS*) is usually in a projected space where the Euclidean distance does not incorporate the curvature of the earth. To deal with this, the data was reprojected from the *CRS* to Projected Coordinate System (*PCR*). In *PCR* the data is represented in meters, designated by the distance from each point to two perpendicular axes to the flat map. These axes are chosen according to the UTM Zone (Universal Transverse Mercator) (Parr Snyder, 1994) and may differ from point to point. After the projection, Euclidean distances are freely used since it tells the data how to draw on a flat surface.

In order to find the best results, 3 different Clustering Algorithms were applied: the *DBSCAN*, the *K-Means,* and the *Agglomerative* (one type of Hierarchical Clustering Algorithm (Murtagh & Contreras, 2017)).

To apply the *DBSCAN*, two different parameters need to be defined *a priori*: the *Epsilon (eps)*, which defines the maximum distance to consider two points as neighbors; and the *MinPts*, defining the minimum number of data points in a cluster. For the *Epsilon,* an *Eps* visualization will be performed for each scale, while *MinPts* was considered 10 for all models. This value was chosen so as not to devalue any *PoI*, since the number of points per cluster will be evaluated later.

On the other hand, both *K-Means* and *Agglomerative* require previously the number of clusters in advance. To define the optimal number, *Elbow* plots (Shi, Wei, Wei, Wang, & Liu, 2021) will be analyzed. The *Agglomerative* needs one more element: the linkage criteria. The “Complete” one was chosen since this linkage considers the distance between two clusters as the distance between the two farthest points in two clusters, which seemed to be the most correct to use since the algorithm is aggregating points to identify areas of interest, and that by joining clusters, it considering that these two distant points belong to the same area.

4.1. CLUSTERING EVALUATION

- **Intrinsic Measures**

Intrinsic Measures are metrics that use internal information of the clustering process to evaluate the goodness of a clustering (Han, Kamber, & Pei, 2012), without requiring knowledge about the truth labels. Some criteria were taken into account, such as the *R-Squared*, when the solutions had the same number of clusters.

The *R-Squared* (or Coefficient of Determination) indicates the degree to which predictor variables explain the variation of predicted variables. In this study, it will indicate the variation in data explained by the relationship between the points and the clusters predicted.

The *R-squared* (R^2) is computed using the correlation coefficient (Miles, 2005):

$$R^2 = \left(\frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)} \times \sqrt{(n\sum y^2 - (\sum y)^2)}}\right)^2, \qquad (4.1)$$

where n is the total number of observations, x the independent variables (the input) and y the dependent variable (the predicted result).

The value resulted lies within the range 0 and 1, which means if the value is 0, the independent variable does not explain the changes in the dependent variable. However, a value of 1 reveal that the independent variable explains the variation in the dependent variable perfectly well.

In addition to the *R-Squared,* also the distribution of the metric features through the clusters mainly by their discrimination ability was considered, based on the visualizations and the distances between the centroids. Besides that, the number of clusters has impact in the model choice, since many clusters would make it harder to develop general marketing approaches, and only few clusters would not allow to clearly identify the *PoI*s.

- **Extrinsic Measures**

On the other hand, Extrinsic Measures are supervised metrics, where the ground truth is available. Since Lisbon is known by tourism and has been studied over the years, it is easy to find different pre-defined points of Interest in sites as *TripAdvisor*[7] and tourism blogs. Using those sources, a list formed by 142 points (100 inside Lisbon, and 42 outside Lisbon) was collected by searching about Lisbon, its activities, parks, emblematic constructions, etc; and the geographic coordinates were found using *Google Maps*, which returns the latitude and longitude of each point center. The list served as a baseline and point of comparison for the results, contributing to the discovery of a

[7] *TripAdvisor* – American online travel company

methodology that reduces the error of the models so that it can be applied in other cities where reality is not known, as it can be seen later.

Making use of the list, it first associated with a Point of Interest to each data entry - following the rule that each photograph was taken at the nearest pre-defined *PoI*.

After applying each model, the dataset ended up with 2 clusters identifications – the "*labels_true*", as it was explained above, and the "*labels_pred*", with the predictions from the model.

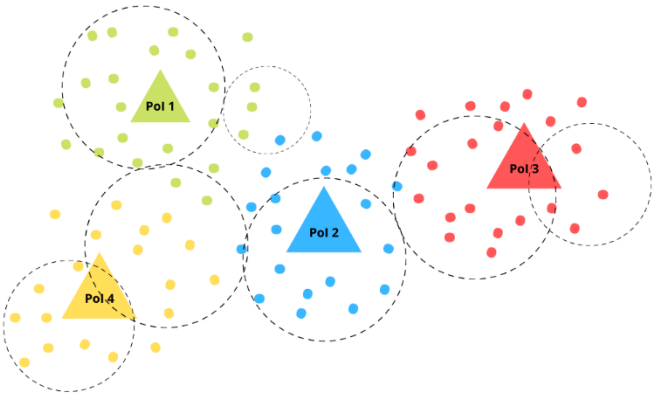


Figure 4.1 – Illustration of the points associated to each *PoI* and the centroids identified by the model (dashed lines)

One possible Extrinsic measure to use in this case is the Normalized Mutual Information (*NMI*). The *NMI* is a Mutual Information score that scales the results between 0 (no mutual information) and 1 (perfect correlation), and it is usually used to compare two partitions even when they come up with different number of clusters. This score can be interpretated as an average of other two measures: homogeneity and completeness.

The NMI is computed as (Kvålseth, 2017):

$$NMI\ (T, C) = \frac{2\ \times I\ (T; C)}{[H(T) + H(C)]} \qquad (4.2)$$

where T represents the class of true labels and C the class of prediction labels. The functions $H(.)$ and $I(T; C)$ correspond to the *Entropy* and *Mutual Information* respectively, and are defined as:

$$H(T) = \sum_{t}^{N} -P(T = t) \times \log(P(T = t)) \qquad (4.3)$$

$$I(T; C) = H(T) - H(T|C) \qquad (4.4)$$

where N is the total number of labels and $P(T = t)$ correspond to the probability of the label be equal to the value t, and $H(T|C)$ is the entropy if the class labels T within each cluster. Being M the total number of cluster labels, it can be computed as:

$$H(T|C) = \sum_{c}^{M} -P(C = c) \times \sum_{t \in T} P(T = t\ |C = c) \times \log(P(T = t|\ C = c)) \quad (4.5)$$

However, this method could be limitative, since it may be misleading and the value could be relatively high on solutions with a high number of clusters (Amelio & Pizzuti, 2015). Furthermore, it is not prejudicing the case that a *PoI* is not associated with any entry, and it is difficult to interpret since it doesn't measure how far/close are the solution to reality.

The comparison between the *PoI* and centroids resulted from the model would be simple if the number of clusters resulting from the models is equal to the number of Points of Interest identified – then, computing the distance between the centroid and the associated *PoI* results in the total distance between them, and thus be able to conclude how far the results are from reality. However, could it be so direct?

When applying the model to different cities, it is not expected to know which are the *PoI a priori*. So, the number of clusters is defined considering the methods presented above – which will most likely be different from what was found as real *PoI*.

This resulted in dealing with one of the following situations:

1. The number of centroids is equal to the number of Points of Interest. Is the model identifying the right *PoI*s? how could clusters be assigned to PoI? Was it just the closest? If yes, could it happen that 2 clusters are associated with the same PoI? (Figure 4.2 (C.))
2. The number of centroids is fewer than the number of Points of Interest (Figure 4.2 (A.)). Is the model identifying less locals than the real tourist interests? How the distance between the centroids and the *PoI*s should be measured if there is no one centroid for each *PoI*?
3. The number of Points of Interest is less than the number of Clusters (Figure 4.2 (B.)). In that case, could it be that all *PoI*s have not been identified, or is the model overestimating some places that are not considered tourism points? Should some centroids be discarded according to some criteria?

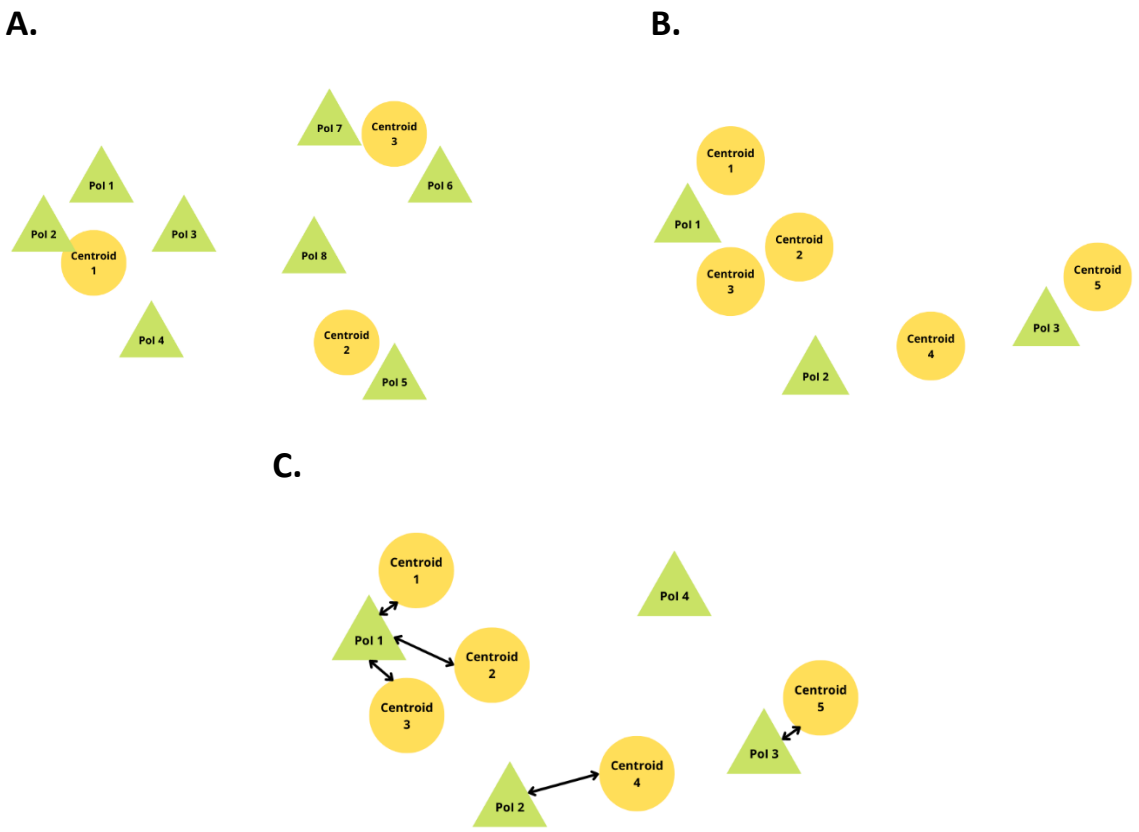


Figure 4.2 – Illustration of the situation where the number of centroids resulted by the model is less than the number of *PoI* identified *a priori* (A.)*;* Illustration of the situation where the number of centroids resulted by the model is more than the number of *PoI* identified *a priori* (B.); and Illustration of the situation where the number of centroids resulted by the model is more than the number of *PoI* identified *a priori* and how it is not possible to associate each *PoI* to a centroid by the nearest rule (C.)

So, this section's challenge is to find a measure to evaluate the methodology applied until this moment, being able to penalize the situations above and easily interpret the results.

It was started by defining D as a matrix of size $C \times T$, where C is the number of clusters identified by the model, and T is the number of *PoI* defined *a priori*. Each matrix entry (D_{ij}) corresponds to the distance between the centroid C_i and the *PoI* T_j.

Then, the method differs depending on the different values of centroids found:

- **Case 1: $C = T$**

  In this case, the method should associate each centroid to the nearest *PoI*, starting with the centroid with the smallest distance and blocking the *PoI* associated. Then, repeat the process for the remaining centroids, using the *PoI*s still available and blocking them every iteration.

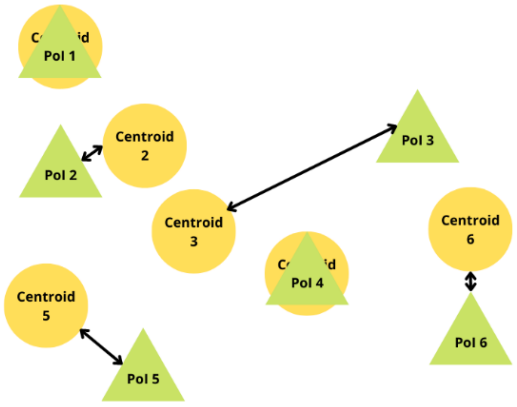


Figure 4.3 – Illustration of how the *PoI*s are associated to a centroid with the presented methodology

- **Case 2: $C < T$**

In this case, the main goal is not only to guarantee that all the centroids are connected with a *PoI*, but also that all *PoI* has at least one centroid. Then, the method starts by associating each centroid to a *PoI*, using an identical process as the previous one, but by centroid instead of *PoI.*

When repeating the process until all centroids present a *PoI*, it is enough to finish associating the *PoI*s with the closest centroid.

- **Case 3: $C > T$**

Finally, since the city presents fewer *PoI*s than the centroids identified by the model, it is needed to ensure that all the centroids are associated with the *PoI*s and that each *PoI* has at least one centroid connected. In addition, the excess number of centroids must be penalized.

Considering this, it is possible to start the method as it was done in Case 1 ($T = C$), associating each *PoI* to the nearest centroid, and minimizing the distance between each *PoI* and the centroid. Then, all that remains is to correct and associate the extra centroids that have been identified. To do that, the method associates the remaining centroids with the closest *PoI*.

Finally, the mean error distance of the centroids will be computed as the sum of the distances computed previously divided by the total number of centroids.

In addition to the average error distance, it is important to assess whether the number of centroids meets the number of *PoI*. With this, the frequency of the number of centroids per *PoI* should be better closer to 1.

5. NETWORK ANALYSIS

With the Points of Interests identified in the previous section, it is possible to study the correlation between each other. These correlations are defined by identifying links that are statistically significant according to the user's activity and using network analysis methods – and then create a network composed of both the points and them association.

To define if a relationship between 2 hotspots is statistically significant, the *p-value* of each link must be computed.

The method starts by defining M as a square matrix of size $N \times N$, where N is the number of *PoI*s. Each entry of M, say M_{ij}, indicates the number of users that uploaded in sequence two photographs in Points of interest i and j – that means that M is a co-occurrence matrix that informs co-visits of user to pairs of Points of interest. On this Matrix, the diagonal elements of M are zero, since it was ignored multiple photographs from the same user in the same Points of interest.

Then, the correlation ϕ-correlation matrix was computed, based on works like (Candia, Encarnação, & Pinheiro, 2019; Lyra, Curado, Damásio, Bação, & Pinheiro, 2021; Ronen, et al., 2014; Hidalgo, 2009). For each pair *(i,j),* if the correlation value is positive, the pair co-occur more often than expected based on their representation in the dataset alone, and if is negative otherwise.

$$\phi_{ij} = \frac{M_{ij}Z - M_i M_j}{\sqrt{M_i M_j (Z - M_i)(N - M_j)}} \quad (5.1)$$

Where M_i is the number of co-occurrences that were identified in the hotspot *i* with all the other hotspots. It is defined as $M_i = \sum_j M_{ij}$, Z is the total number of observations and N the total number of co-occurrences divided by 2, to represent the total number of occurrences.

$$N = \frac{\sum_i M_i}{2} \quad (5.2)$$

Then, using standard statistical inference methods (Gotelli, 2000), it is estimated the *p-value* associated with ϕ_{ij} by calculating the upper tail probability of obtaining a value equal or greater than ϕ_{ij}:

$$t_{ij} = \frac{\phi_{ij}\sqrt{D-2}}{\sqrt{1-{\phi_{ij}}^2}}, \quad (5.3)$$

Where D is defined as $D = max(M_i, M_j)$.

The final network is obtained by identifying the pairs of Points of interest with $\phi_{ij} > 0$ and a significance level of 0.05 (p −value), meaning that there is less than 5% chance they would be observed from pure chance.

6. RESULTS

6.1. CLUSTERING - LISBON

As defined in Section 4, the identification of centroids is performed through clustering algorithms.

Since the density of observations is skewed, being concentrated in Lisbon proper rather than in the surrounding areas, it was combined the results from independently performing the clustering at two different scales: The first includes the entire Lisbon Metropolitan area excluding observations within the city of Lisbon; The second includes only observations within the city of Lisbon.

It was started by applying the *DBSCAN*. To compute this model, the *MinPts* and the *Eps* needed to be found, and in Figure 6.1 it is possible to find the *Eps* visualization for photos from Lisbon and from outside Lisbon.

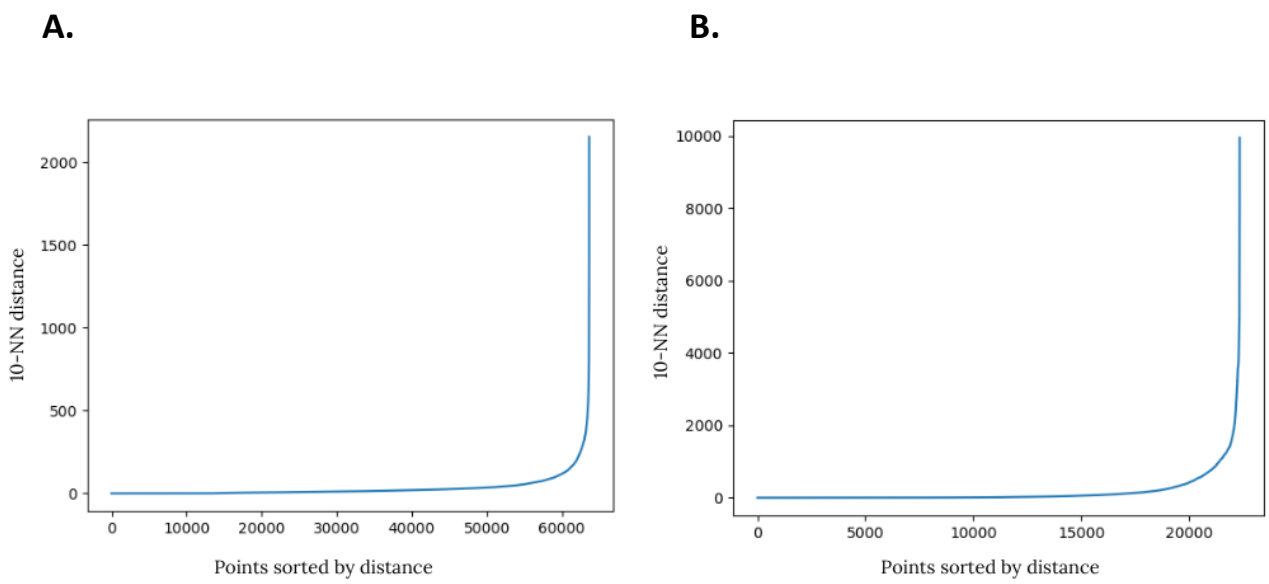


Figure 6.1 – *Eps* visualization for the data from Lisbon (A.) and from the outside Lisbon area (B.)

It was chosen a value of $10m$ for Lisbon, and $1000m$ for outside Lisbon, corresponding to the elbows to the figures.

To apply the remaining models (*K-Means* and *Agglomerative*), it was necessary to find the optimal number of clusters. For that, the *Elbow* plots were printed (Figure 6.2).

A. **B.**

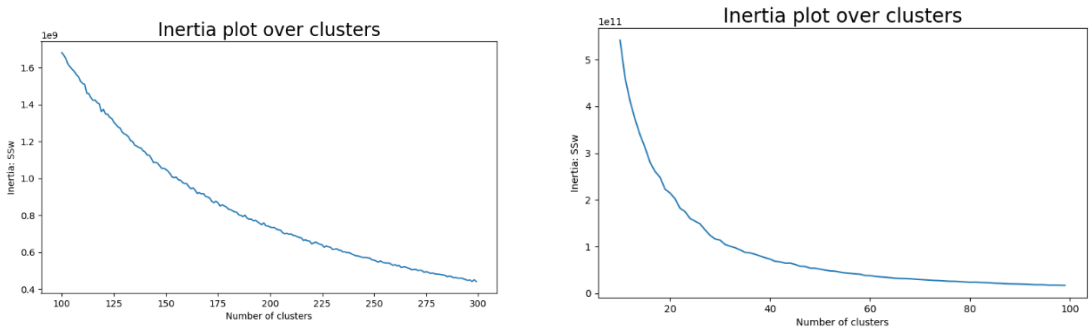


Figure 6.2 – *Elbow* visualization for the data from Lisbon (A.) and from the outside Lisbon area (B.) using *k-means* algorithm

After applying all the clustering algorithms, the results were merged in one dataset per algorithm, to apply the metrics explained in the section 4.1 to all the results and choose only one algorithm. The Table 1, that summarizes all the measures and parameters considered.

Table 1 – Clustering Evaluation metrics for 3 models – *DBSCAN*, *Agglomerative* and *K-means*

Model	***R-Squared***	***NMI***	***Nº Centroids***
DBSCAN	0.9820	0.6574	248
Agglomerative	0.9666	0.7645	190
K-means	0.9898	0.7889	190

Analyzing all the results presented in the table and considering the centroids distribution and their discrimination (Annexes 1), the Model chosen was the *K-means*.

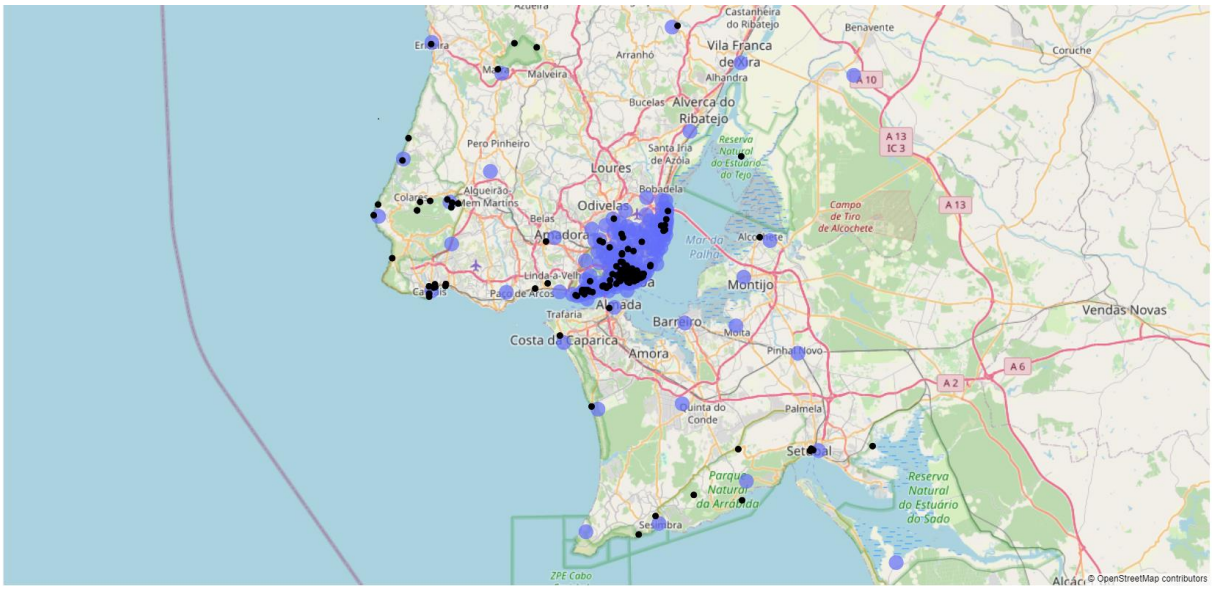


Figure 6.3 – Map of the Metropolitan Area of Lisbon with the centroids identified by *K-means* (blu points) and the pre-defined *PoI*s (black points)

The model chosen presents 190 centroids – 160 inside the city of Lisbon, and 30 outside Lisbon. Generally, the model overestimates the *PoI*s, since it finds more centroids than those that exist. Looking at these results, it is seen that they fell in two cases presented in section 4.1.: within Lisbon, there are 160 centroids > 100 *PoI*s, and outside Lisbon, there are 30 centroids < 42 *PoI*s.

Computing the error associated with each centroid, it was about $560m$ per centroid for the Lisbon model, and $6.65km$ for the model outside Lisbon. At first glance they seem high values, however, $560m$ within a city turn out to be little, considering that many clusters have more than that length (Figure 6.4) and that the *PoI* may have been identified at one end, and the centroid at the other. Outside Lisbon, we have examples like Setúbal and Mata de Sintra, large places that were only identified as one centroid. At the same time, it is a much larger area with a smaller number of centroids, so a higher error value would be expected.

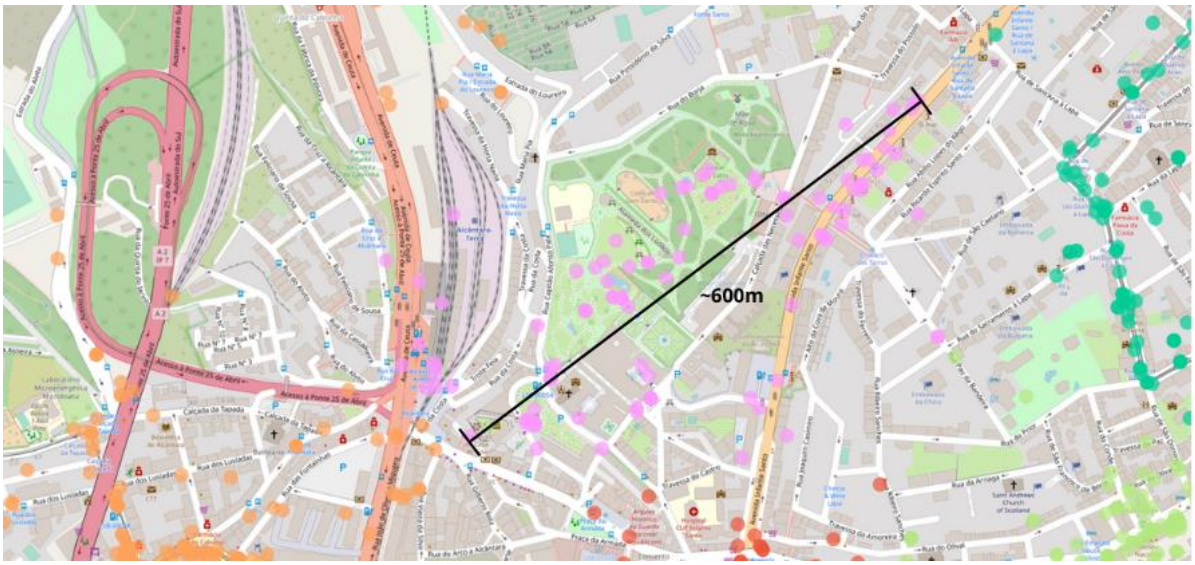


Figure 6.4 – Example of a cluster with a length greater than 560 meters, where the colours represent to the cluster to which each point is associated

Although the centroids are already defined, the number of observations per cluster is quite varied - there are both centroids with 7000 observations and with only 30.

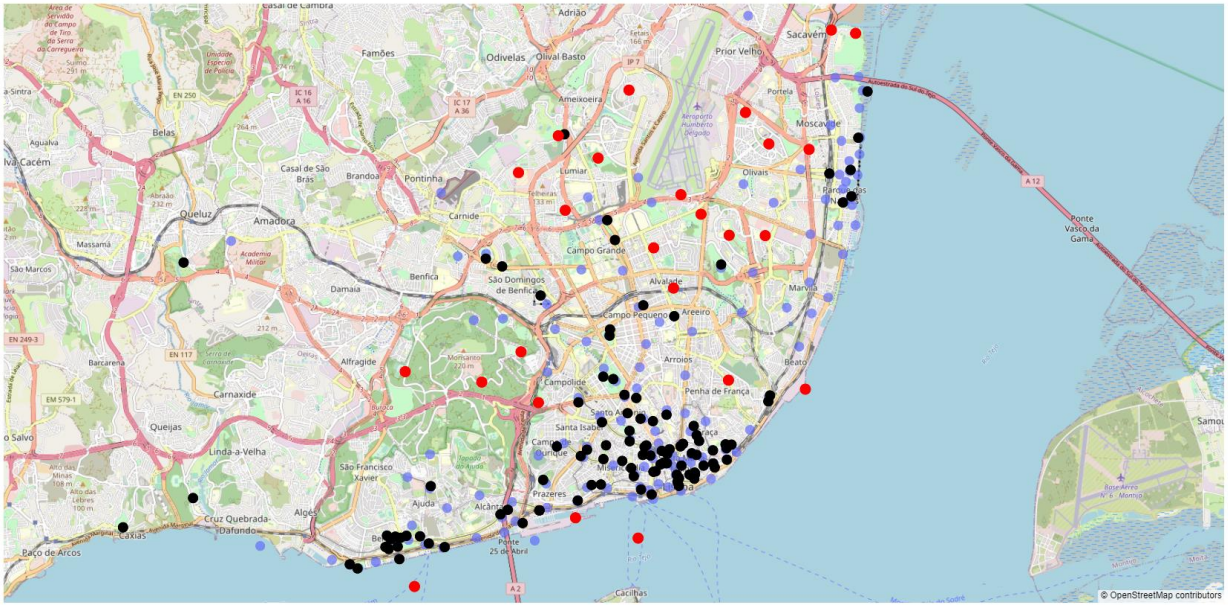


Figure 6.5 – Map of Lisbon with the centroids identified by the *K-means* with more than 50 observations (blue points), the centroids identified by the *K-means* with less than 50 observations (red points), and the pre-defined *PoI*s (black points)

Analyzing Figure 6.5, it is possible to see that most of the red centroids are not related to any *PoI*, presenting only a set of photographs taken on the spot, but not representing a particular tourist spot - perhaps because there was an event in those places or just something very specific that only

appealed to a minority of *Flickr* users. Removing these centroids, the Metropolitan of Lisbon ends with 164 centroids, and the error inside Lisbon decreases to $425m$.

Still, regarding the density of photographs taken, it is possible to conclude that there are 3 areas in Lisbon more popular than others – Belém, Baixa-Chiado, and Parque das Nações. These zones could be considered Hotspots since they represent a set of *PoI*s with a higher density of tourists than the rest. Outside Lisbon, Sintra also stands out from the other cities, as is possibly seen in Figure 6.6.

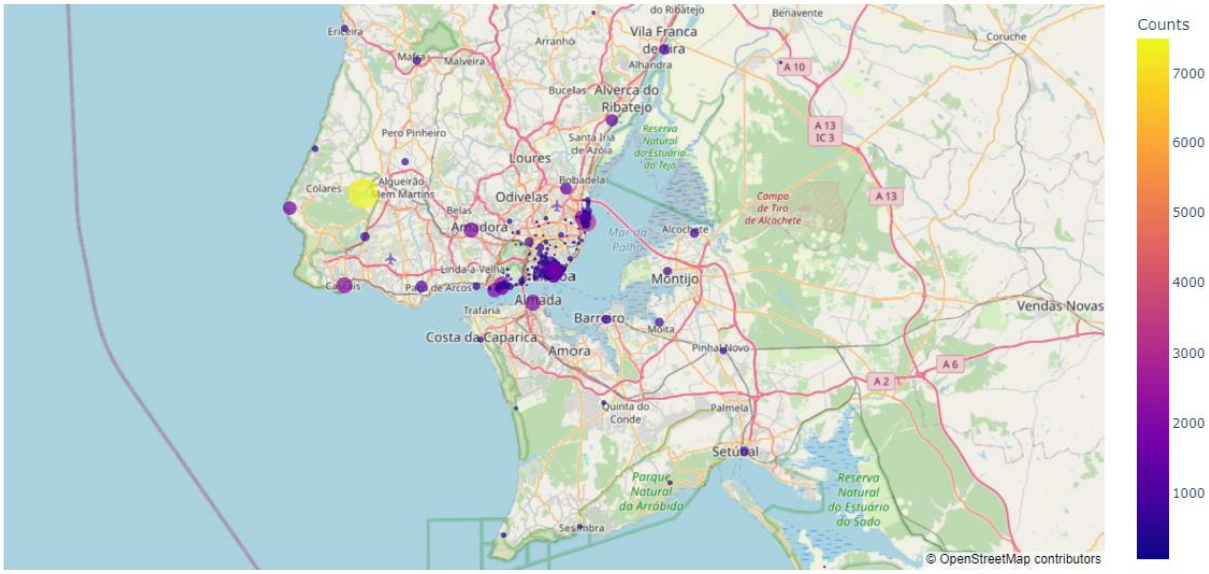


Figure 6.6 – Map with the centroids with size and colour according to the number of photographs taken

3 of the 4 hotspots have some points in common - Sintra, Belém, and Baixa-Chiado are historic areas, with old buildings and constructions. On the other hand, Parque das Nações is a much more recent area, with large modern buildings, but with activities such as the Oceanarium, Pavilion of Knowledge, Casino, etc.

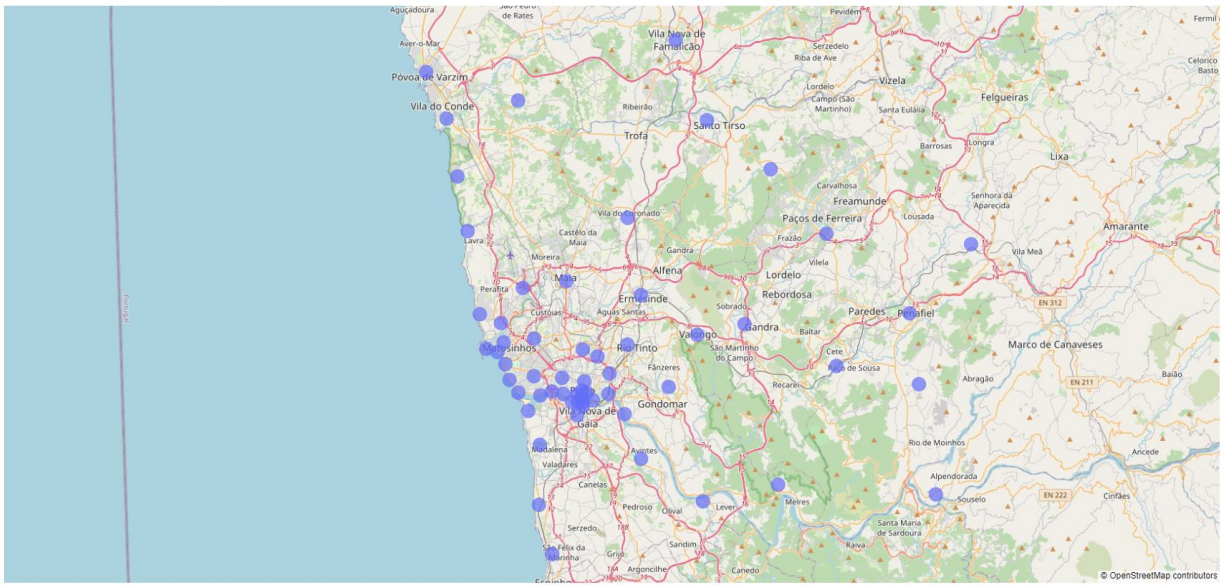
6.2. CLUSTERING – FARO AND PORTO

One of the goals which this study proposes is also to find *PoI*s in Faro and Porto. After finding the methodology that comes closest to the reality of Lisbon, it can be applied to these cities.

As in Lisbon, given the irregularity of the data density, the centroids of Porto were also calculated in 2 parts: one with data within the city of Porto, and another with data outside the city of Porto. In Faro, as it is only applied to the city and the distribution seems uniform, the model was applied only once, covering all Faro's data.

Applying the *Elbow* method to find the number of *PoIs*, there were identified 60 centroids for the all the Metropolitan Area of Porto and 30 centroids for Faro. After applying the *K-means* and removing the centroids with less than 50 observations, the cities ended up with 59 and 28 centroids in Porto and Faro respectively. Figure 6.7 represents these centroids.

A.



B.

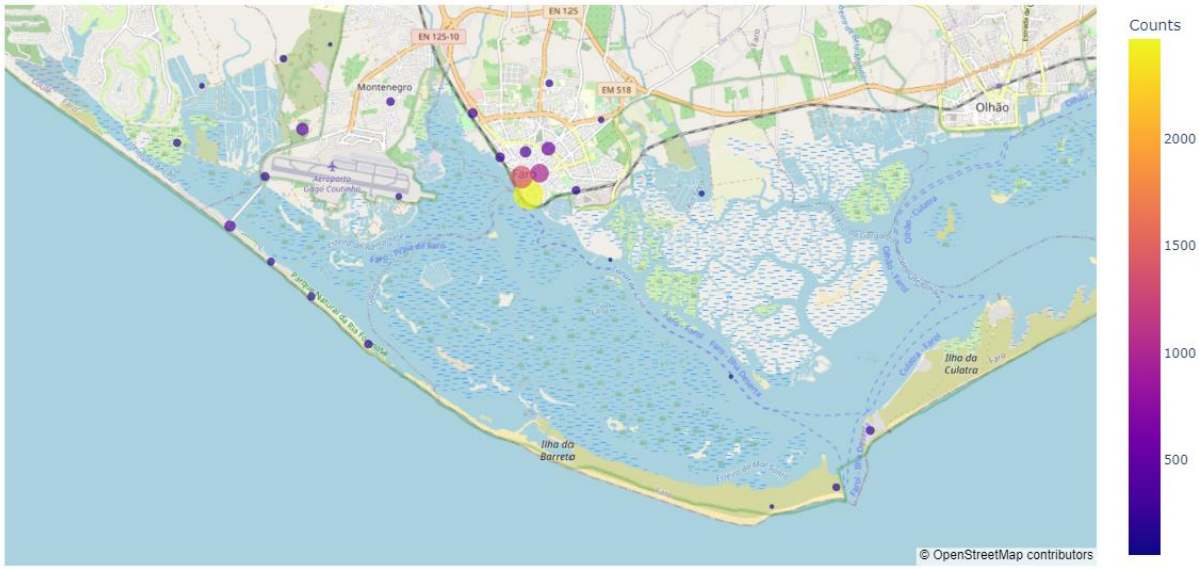
Figure 6.7 – Map with Metropolitan Area of Porto's centroids (A.) and with Faro's centroids (B.)

In Figure 6.7, the centroids seem equally spread throughout the Metropolitan Area of Porto, apart from a greater concentration in the Porto city area (Annex 4). At these points, there are historical buildings such as São Bento Station and the Aliados, as well as many viewpoints, squares, and beaches – also typical for receiving tourists. In addition, several points in the historic center and ends of bridges, where tourists are usually found to appreciate the streets and views.

In Faro, the largest concentration of centroids is in the historic center of Faro – like churches, gardens, and the Marina. In addition to these attractions, it was also possible to identify beaches along the coast of Faro.

In terms of tourist density (Figure 6.8), in Faro the preferences are found in the historic center of Faro – it's observable a hotspot consisting of 3 *POI*s: the Igreja da Sé tower and the Municipal Museum of Faro, the Marina of Faro and the downtown area of Faro.

A.

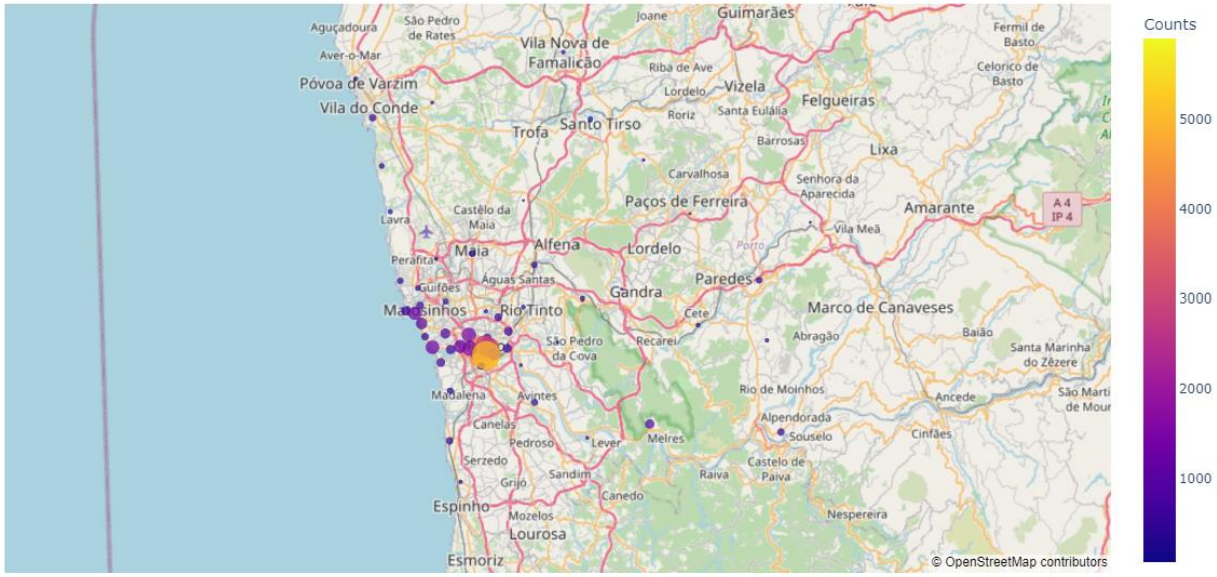


B.

Figure 6.8 – Map of Faro (A.) and zoom in center of Faro (B.) with the centroids with size and colour according to the number of photographs taken

In Porto the behavior is similar to what was found in Faro - a greater concentration in the center of Porto, decreasing as the distance to the city increases (contrary to what was seen in Lisbon). In the center it is still possible to identify 5 points with greater density than the others: the two banks of the D. Luís II bridge; the Infante Dom Henrique and Gomes Teixeira squares; and São Bento station.

A.



B.

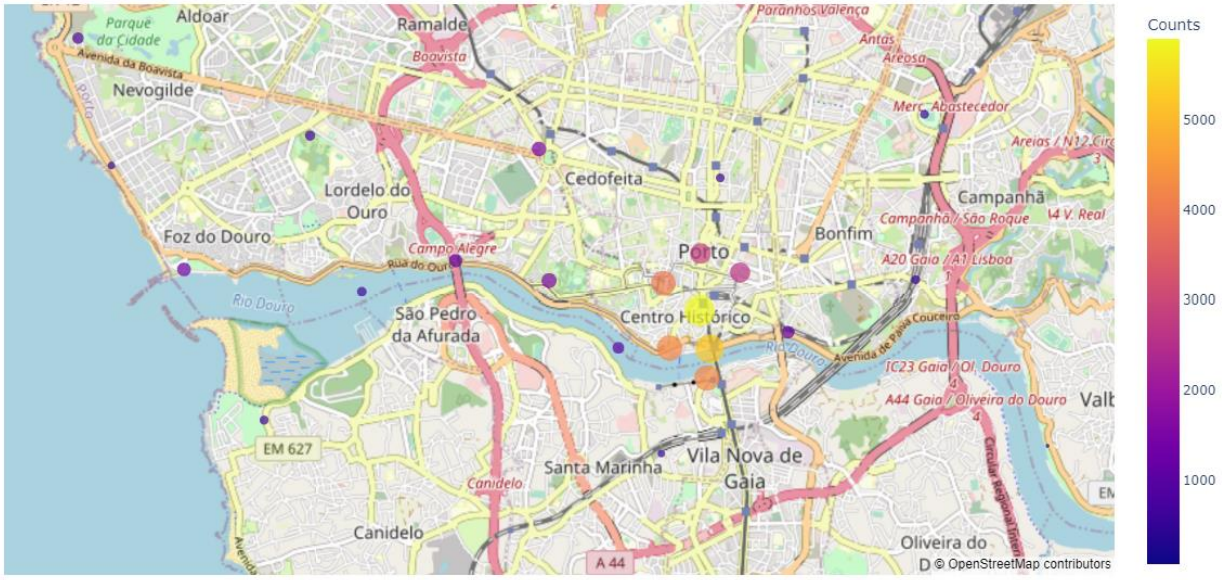


Figure 6.9 – Map of Metropolitan Area of Porto (A.) and zoom in Porto (B.) with the centroids with size and colour according to the number of photographs taken

6.3. NETWORK ANALYSIS

With the final centroids defined and the density analyzed, the study proceeded with the Network Analysis. The analysis was applied to all the centroids (inside and outside the cities). In Figure 6.10, we can find the network for Lisbon.

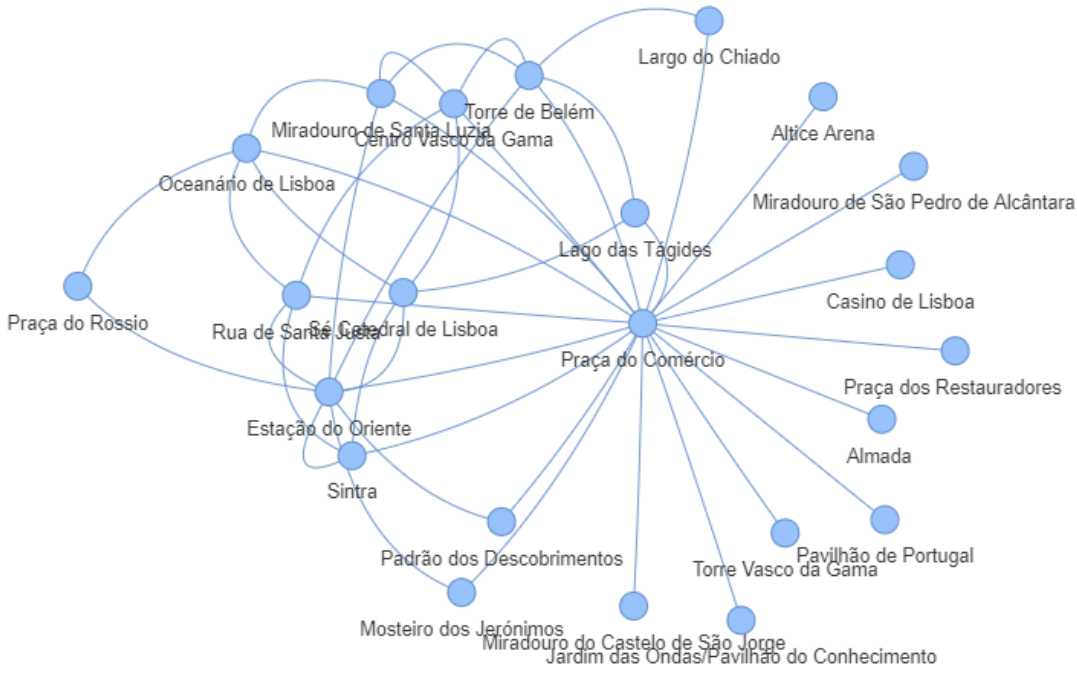
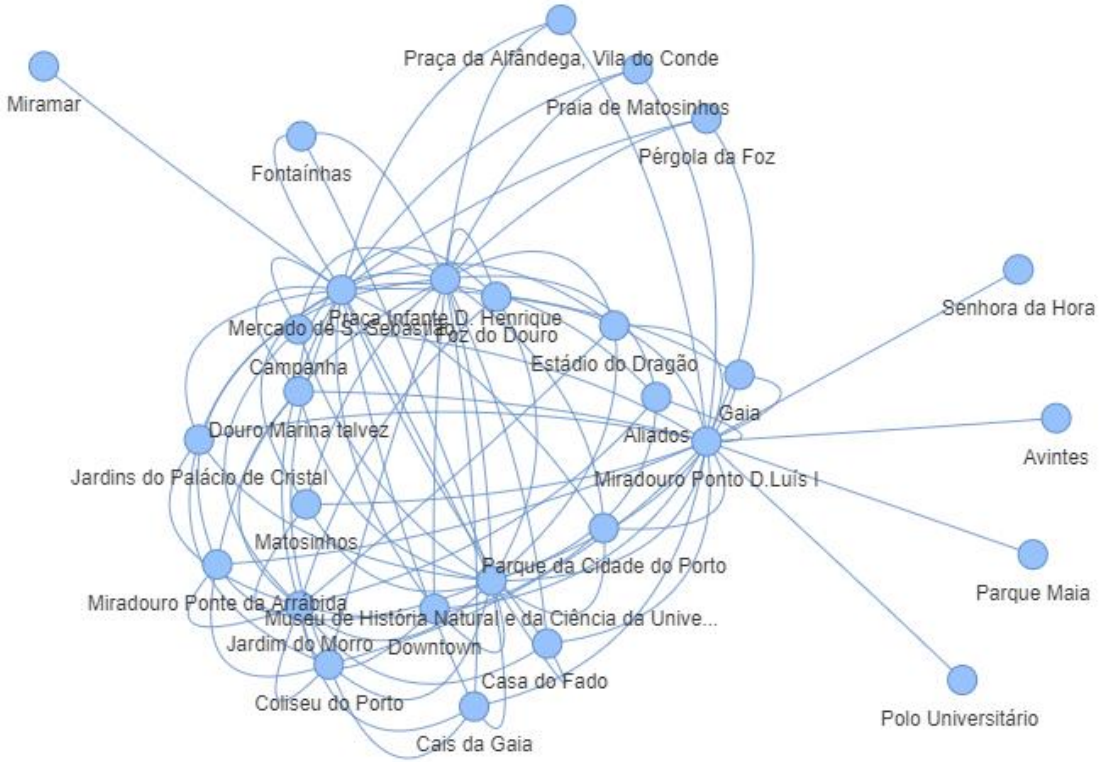


Figure 6.10 – Network connecting the most correlated centroids in the Metropolitan Area of Lisbon

Analyzing the zones identified in the Network, we see that it focuses on places that are part of the 4 Hotspots previously identified: Belém, Baixa-Chiado, Parque das Nações and Sintra – which in practice makes sense, being the points that tourists photograph the most, are also the ones that are most correlated and visit them together. There is possible to conclude that Praça do Comércio is the center of the Network – connected with a big part of the remaining points. Besides that, it is possible to see points not directly related to Praça do Comércio, such as Praça do Rossio and Oceanário de Lisboa, and Rua de Santa Justa, but that relate to each other and with places like Sintra, Catedral de Lisboa, Estação do Oriente, etc.

We can also take away from this analysis that not even the closest points are the most related, and that popularity counts much more in the tourist's decision when choosing which places to visit on their trip.

A.



B.

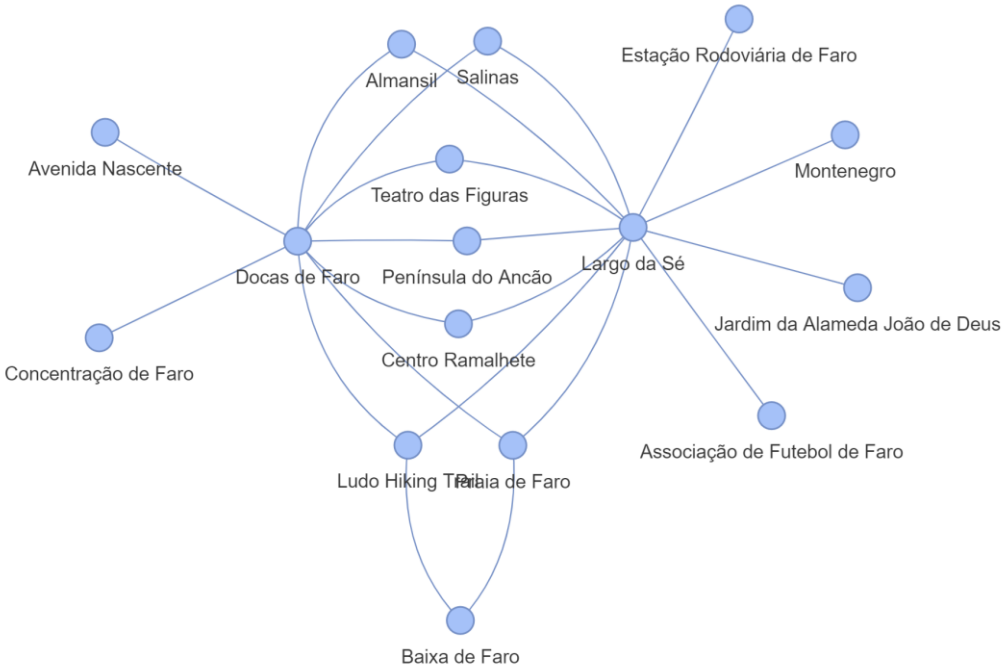


Figure 6.11 – Network connecting the most correlated centroids in the Metropolitan Area of Porto (A.) and Network connecting the most correlated centroids in Faro (B.)

Both Porto and Faro have more relative connections than Lisbon – perhaps because they are smaller cities, and it is easier for tourists not only to get around but also to plan, as there are fewer *PoI*s.

In Porto, there are many connections within the city itself, but also between the city center and some cities in the periphery, such as Maia, Matosinhos, and Vila Nova de Gaia. We can also identify points with a particular behavior, such as Mercado de S. Sebastião, Praça Infante D. Henrique and Miradouro Ponte D.Luís I, which despite showing a correlation between points within the city, take tourists to other places outside that center of more correlated points.

In Faro it is possible to find practically the same behavior: correlations between the historic center; adding only that tourists choose the center and then just one beach (otherwise, correlations were seen between the different beach areas). We can highlight Largo da Sé and Docas de Faro since they are points that are not connected but that serve as bridges to other Points of Interest.

7. CONCLUSION

It was started by getting the data from *Flickr* and cleaning it.

Then, the dissertation explored different clustering techniques to identify Points of Interest (*PoI*s) in a city where the reality is known (Lisbon). It was very useful to have this experience of *PoI*s defined, not only as a point of comparison for the different models but also to distinguish how and what types of clusters should be considered good candidates for *PoI*s. The proposed approach measures the average error of each cluster centroid and helps to guarantee a good balance between the location of *PoI*s and the natural dispersion of photos. The final error threshold for each centroid was set to $425m$ inside the city and $6.65km$ outside, which can be considered insignificant due to the area's size. Applying the method in the cities of Porto and Faro, it was possible to identify centroids that made sense - mostly historic centers, iconic and old buildings, and beaches/viewpoints.

The *DBSCAN*, despite being the preferred method for a huge part of the Literature Review, seems to show only a few results in Lisbon. In terms of metrics, it is the least close to reality. Analyzing the centroids on the map (Annexes 1), it is possible to see that the algorithm detects many centroids that represent only one *PoI* - perhaps because this *PoI* can be considered a Hotspot and not just a *PoI*. In other words, *DBSCAN* identifies areas with higher densities, discarding areas with fewer photographs but which are also considered *PoI*.

The Network Analysis allowed us to conclude that not all cities present the same behavior: while in Lisbon the correlated points are mostly located within the city (despite being distant), in Porto and Faro, in addition to several correlations within the city, they also present some connections with points from the peripheries - being them more distant villages around Porto or beaches in Faro. With this information, cities can optimize their resources, for example in transport, creating routes that pass through these points, or even improve access to them and keep prizing them.

8. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Although interesting conclusions and results were obtained, this study ends up having some limitations, especially regarding the data obtained. First, it is restricting the investigation to *Flickr* users and not taking into consideration all the tourists. Besides that, the *Location* field is a big part of a missing value, so it can be considered residents in the analysis and not just tourists as would be the main objective. In addition, the data does not present information about the tourists, which could be interesting to study the different characteristics of each one of them and how they are related to the *PoI*.

For future work, it can be recommended to explore the temporal dimension: for example, define the tourist routes using the characteristics of network analysis and according to the *PoI*s that tourists visited in chronological order, make use of the *TimeStamp* field. Furthermore, there is a possibility to study the difference between the *PoI*s over the years, and what has been the trend of tourists in terms of tastes and preferences.

9. REFERENCES

Amelio, A., & Pizzuti, C. (2015). Is Normalized Mutual Information a Fair Measure for Comparing Community Detection Methods. *2015 IEEE/ACM International Conference onAdvances in Social Network Analysis and Mining*, (pp. 1584-1585). Paris.

Barchiesi, D., Preis, T., Bishop, S., & Moat, H. S. (2015). *Modelling human mobility patterns using photographic data shared online.* Royal Society Open Source.

Bexiga, N. (2015). *Crescimento da atividade turística em Faro .* Universidade do Algarve.

Candia, C., Encarnação, S., & Pinheiro, F. L. (2019). The higher education space: connecting degree programs for individuais' choices. *EPJ Data Science*, 39.

Directorate-General for Internal Market, I. E. (2020). *European Capitals of Smart Tourism*. Retrieved from European Commission: https://smart-tourism-capital.ec.europa.eu/leading-examples-smart-tourism-practices-europe_en

Dogru, T., & Bukut, U. (2017). *Is tourism an engine for economic recovery? Theory and empirical evidence.* ELSEVIER.

Domènech, A., Mohino, Inmaculada, & Moya-Gómez, B. (2020). *Using Flickr Geotagged Photos to Estimate Visitor Trajectories in World Heritage Cities.* International Journal of Goe-Information.

Donaire, J. A., Camprubí, R., & Galí, N. (2014). *Tourist clusters from Flickr travel photography.* ELSEVIER.

Flickr, I. (2020, November 23). Retrieved from Flickr: https://www.flickr.com/

Flickr, I. (2020, April 30). *Flickr*. Retrieved from https://www.flickr.com/services/api/

Gede, M., & Kádár, B. (2019). Analysing tourism movements along the Danube river based on geotagged Flickr photography. *29th International Cartographic Conference (ICC 2019)* (p. 5). Tókio: International Cartographic Conference.

Giglioa, S., Bertacchini, F., Bilotta, E., & Pantano, P. (2019). Using social media to identify tourism attractiveness in six Italian cities. In *Tourist Management* (pp. 306-312). Elsevier.

Gorunescu, F. (2013). *Data Mining: Concepts, Models and Techniques.* 2011 Springer .

Gotelli, N. J. (2000). Null model analysis of species co-occurrence patterns. *Ecology*, 2606--2621.

Habeeb, N. J., & Weli, S. T. (2020). *Relationship of Smart Cities and Smart Tourism: An Overview.* HighTech and Innovation Journal.

Han, J., Kamber, M., & Pei, J. (2012). 10 - Cluster Analysis: Basic Concepts and Methods. *Data Mining (Third Edition)*, 443-495.

Hidalgo, C. A.-L. (2009). A dynamic network approach for the study of human phenotypes. *PLoS computational biology*.

INE. (2020). *Estatísticas do Turismo - 2019.* Lisbon: Instituto Nacional de Estatítica.

INE. (2021). *Estatísticas do Turismo - 2020.* Lisbon: Instituto Nacional de Estatística.

K. Walton, J. (2012, January 27). Retrieved from Britannica: https://www.britannica.com/topic/tourism

Karayazi, S. S., Dane, G., & Vries, B. (2021). *Utilizing Urban Geospatial Data to Understand Heritage.* International Journal of Geo-Information.

Kvålseth, T. O. (2017). On Normalized Mutual Information: Measure Derivations and Properties. *Entropy*, 15.

Lew, A., & McKercher, B. (2006). *MODELING TOURIST MOVEMENTS - A Local Destination Analysis.* ELSEVIER.

Li, Y., Hu, C., Huang, C., & Duan, L. (2016). *The concept of smart tourism in the context of tourism information.* ELSEVIER.

Liu, B., Huang, S., & Fu, H. (2016). *An application of network analysis on tourist attractions: The case of Xinjiang, China.* ELSEVIER.

Lyra, M. S., Curado, A., Damásio, B., Bação, F., & Pinheiro, F. L. (2021). Characterization of the firm–firm public procurement co-bidding network from the State of Ceará (Brazil) municipalities. *Applied Network Science*, 1-10.

Miles, J. (2005). R-squared, Adjusted R-squared. In *Encyclopedia of Statistics in Behavioral Science* (pp. 1655-1657). John Wiley & Sons, Ltd, Chichester.

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *Wiley Periodicals, Inc.*, 1-16.

Oliveira, D., & Costa, G. (2020, April-June). Tourism and its environmental implications: analysis from the observation of tourist use - the case of Faro, Portugal. *Pasos*, pp. 279-291.

Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C., & González, M. C. (2015). *Urban magnetism through the lens of geo-tagged photography.* EPJ Data Science.

Parr Snyder, J. (1994). *Map Projections: A Working Manual.*

Pender, L., & Sharpley, R. (2005). *The Management of Tourism.* Chennai, India: SAGE Publications Ltd.

Pestana, M. H., Parreira, A., & Moutinho, L. (2019). *Motivations, emotions and satisfaction: The keys to a tourism destination choice.* ELSEVIER.

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, E5616-E5622.

Sarra, A., Zio, S. D., & Cappucci, M. (2015). *A quantitative valuation of tourist experience.* ELSEVIER.

Shi, C., Wei, B., Wei, S., Wang, W., & Liu, H. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Wireless Communications and Networking*, 16.

Silva, A., Campos, P., & Ferreira, C. (2019). *Sequence and Network Mining of Touristic.* Springer Nature Switzerland.

Simmons, S. (2018). *1.09 - Metadata and Spatial Data Infrastructure.* ELSEVIER.

Su, K., Li, J., & Fu, H. (2011). *Smart City and the Applications.* Wuhan, Hubei, China: School of Computer Wuhan University.

UNESCO. (n.d.). *World Heritage List*. Retrieved from UNESCO World Heritage Centre 1992-2022: https://whc.unesco.org/en/list/

Verma, V., & Pang, A. (2004). *Comparative Flow Visualization.*

Vu, H. Q., Li, G., Law, R., & Ye, B. H. (2014). Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. In *Tourist Management* (pp. 222-232). Elsevier.

Wu, X., Huang, Z., Peng, X., Chen, Y., & Liu, Y. (2018). Building a Spatially-Embedded Network of Tourism Hotspots From Geotagged Social Media Data. *IRRRAccess*.

Yang, L., Wu, L., Liu, Y., & Kang, C. (2017). *Quantifying Tourist Behavior Patterns by Travel Motifs and Geo-Tagged Photos from Flickr.* International Journal of Geo-Information.

Zeng, B., & Gerritsen, R. (2014). What do we know about social media in tourism? A review. In *Tourism Managment Perspectives* (pp. 27-36). Australia: Elsevier.

Zeng, B., & Gerritsen, R. (2014). What do we know about social media in tourism? A review. *Elsevier*, 27-36.

ANNEXES

ANNEX 1 – FINAL CLUSTERING RESULTS (K-MEANS) FOR LISBON

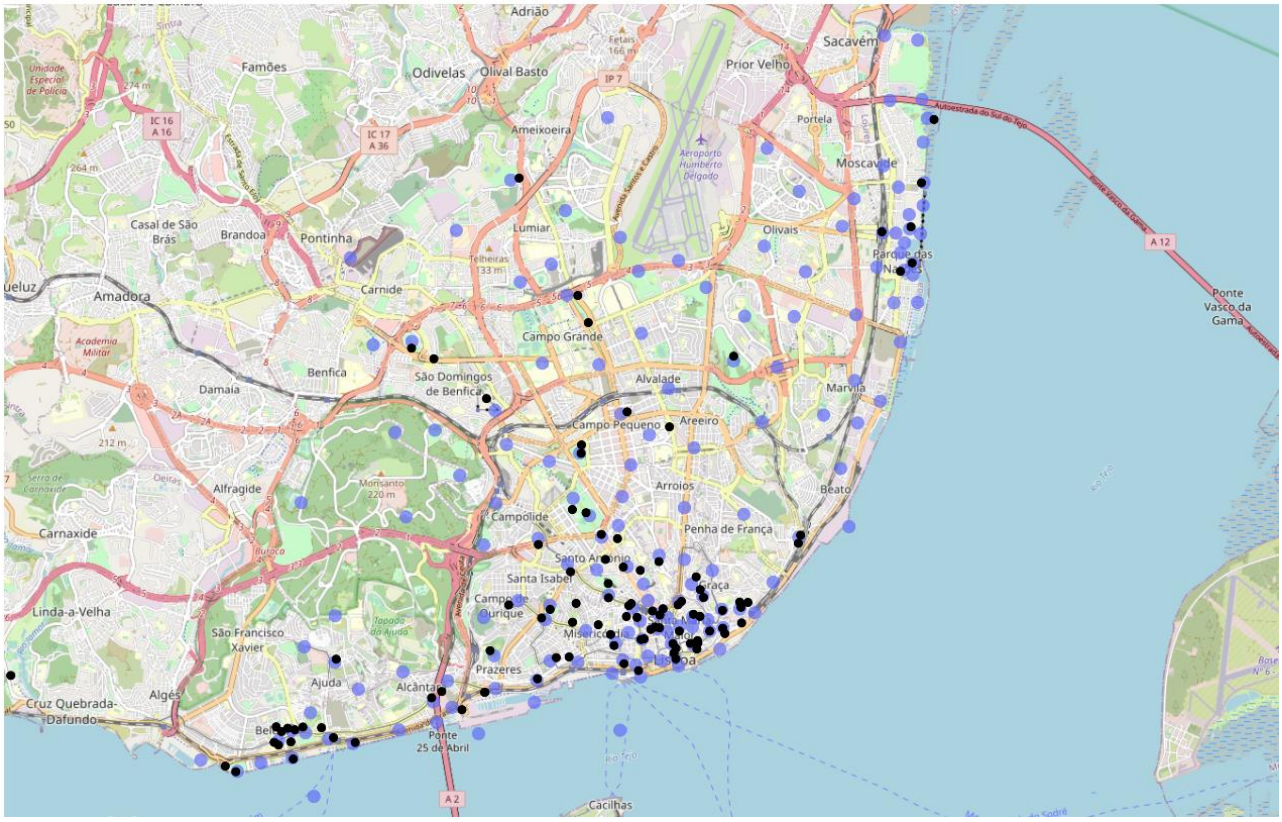


Figure A 1 – Map of Lisbon with the centroids identified by the *K-means* (blue points) and the pre-defined *PoI*s (black points) – *Zoom* in Lisbon

ANNEX 2 – AGGLOMERATIVE RESULTS FOR METROPOLITAN AREA OF LISBON

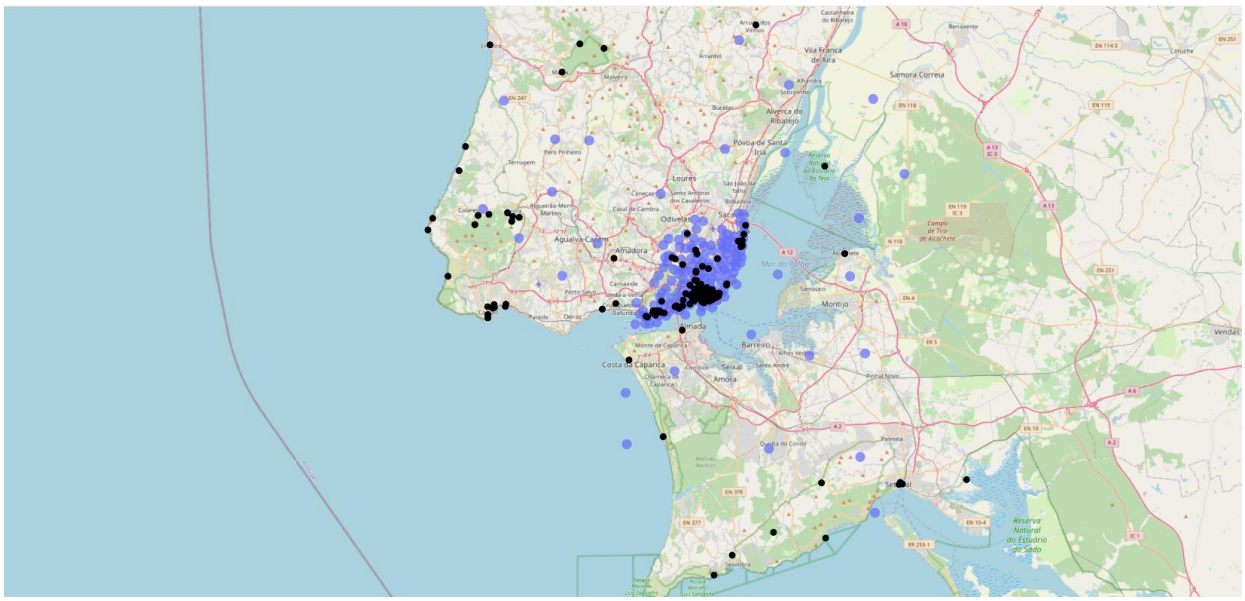


Figure A 2 – Agglomerative results (blue points) and PoIs (black points) for the city of Lisbon and outside Lisbon

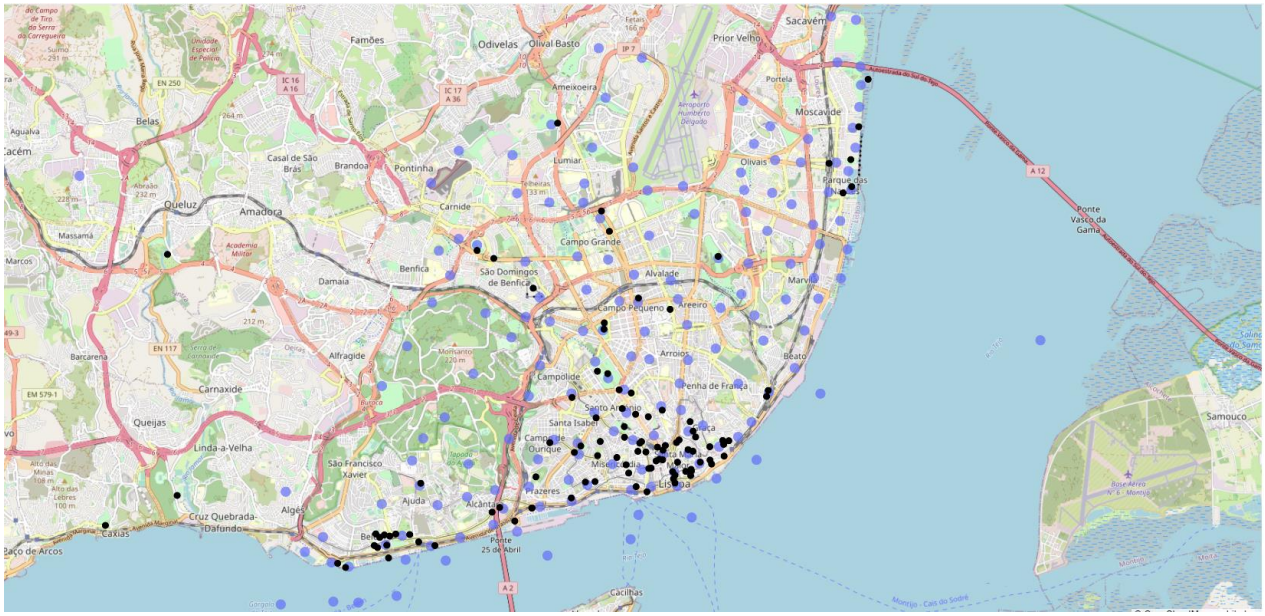


Figure A 3 – Agglomerative results (blue points) and *PoI*s (black points) for the city of Lisbon

ANNEX 3 – *DBSCAN* RESULTS FOR METROPOLITAN AREA OF LISBON

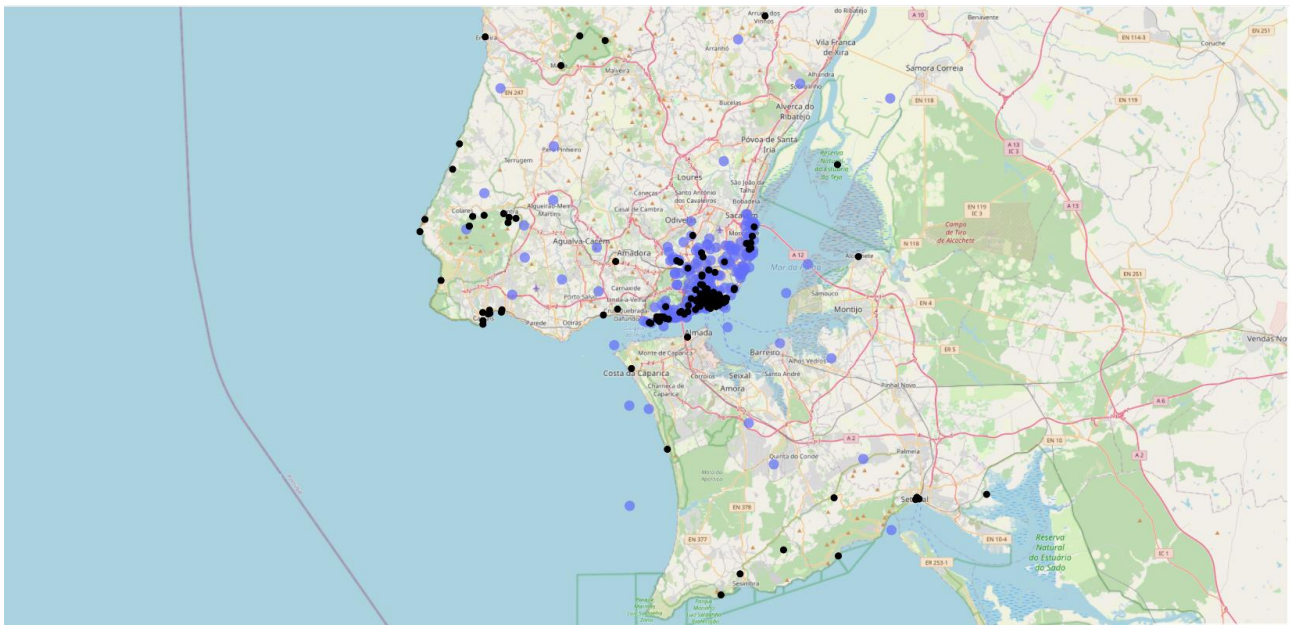


Figure A 4 – *DBSCAN* results (blue points) and *POI*s (black points) for the city of Lisbon and outside Lisbon

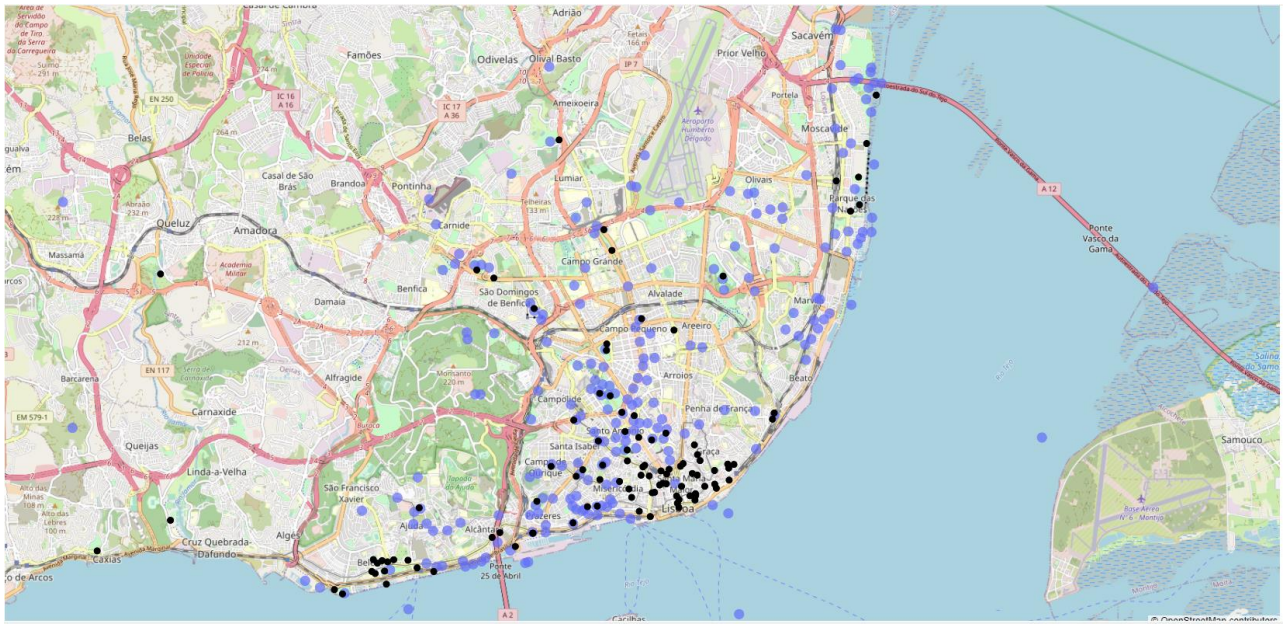


Figure A 5 – *DBSCAN* results (blue points) and *POI*s (black points) for the city of Lisbon

ANNEX 4 – *ZOOM IN* THE PORTO CENTROIDS

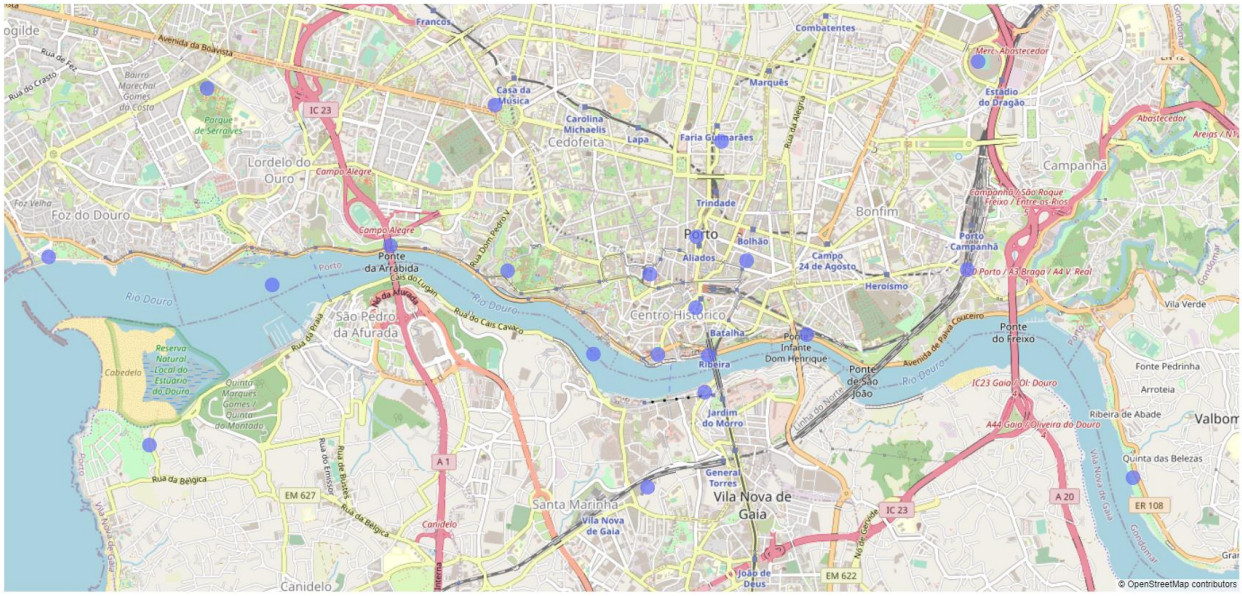


Figure A 6 – *Zoom in* the Porto centroids

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

Universidade Nova de Lisboa