Master Thesis in Geographical Information Science nr 134

Mining geosocial data from Flickr to explore tourism patterns: The case study of Athens

Domna Kanari

2021

Department of

Physical Geography and Ecosystem Science

Centre for Geographical Information Systems

Lund University
Sölvegatan 12

Domna Kanari (2021). Mining geosocial data from Flickr to identify tourism patterns: The case study of Athens
Master degree thesis, 30/ credits in Master in Geographical Information Science
Department of Physical Geography and Ecosystem Science, Lund University

Mining geosocial data from Flickr to explore tourism patterns: The case study of Athens

---

Kanari Domna

Master Thesis, 30 credits, in Geographical Information Sciences

Supervisor Dr. Ali Mansourian

Department of Physical Geography and Ecosystem Science, Lund University

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Ali Mansourian, who supported my ideas and for being there not only as a supervisor but also as a mentor and motivator, as well as for his valuable comments that pushing me to think more.

A huge thank you to my mum, Sofia, and my dad, Giannis, for supporting my academic journey all these years without knowing exactly what I was studying ☺.

A thank you to my sister, Athina, and her partner, Christos for believing in me no matter what.

And last but not least, a thank you to my partner Nikolas, for his unlimited patience, for giving me strength, and for his belief that I can do anything in the world, the times I was ready to give up.

ABSTRACT

Social media are providing a new type of geo-tagged data that by processing them, new types of knowledge can be generated and used by decision-makers in different areas including tourism. It includes e.g., identifying touristic areas that are not prioritized or advertised by touristic authorities. This study aims to investigate a methodology to detect Areas of Interest (AOIs) and temporal distributions of visitors geotagged photos in Athens, by mining and analyzing geosocial data from the Flickr social media platform, from 2009 to 2019. The methodology of this research, divided into 5 stages of procedures: the geosocial data mining, the cleaning process of the data, the spatial clustering analysis, the construction of the database, and the visualization of the results through a Web-GIS application. The total amount of geosocial data harvested from the Flickr social media platform for this research was 157,314 and after the cleaning process was 77,659. To identify the most desired AOIs and the temporal distribution tendencies of the visitors, the HDBSCAN clustering algorithm was applied to the dataset. The algorithm produced 20 spatial clusters in popular areas of Athens and the results of the clustering analysis were stored in a database. To validate the results, 21 of the most famous Points of Interest (POIs) of Athens were gathered, mapped and the correlation between them and the produced AOIs was explored. Finally, the produced AOIs and the collected POIs were presented through a prototype Web-GIS platform among with temporal distribution statistics for each AOI. The results of this study showed that the HDBSCAN algorithm produced 8 new AOIs that were not suggested or advertised by tourism authorities. Also, the findings of this research demonstrate that the study area presents in general medium to high levels of seasonality with small exceptions and that the visitors are mostly from Europe, North America, and Asia. Despite some reliability issues geosocial data present, tourism agencies/authorities, urban planners, and policymakers must seriously consider exploiting such kind of data and understand the power of the tools they can create using location intelligence.

Keywords: Geography, Geographical Information Systems, GIS, Spatial Analysis, Spatial Clustering, Density-Based Clustering, HDBSCAN, Web-GIS, Tourism footprints, Social Media data, Geosocial data

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronym	Definition
AGI	Ambient Geographic Information
AOI	Area of Interest
API	Application Programming Interface
CPO	Chief Product Officer
CRUD	Create, Read, Update, Delete
CSS	Cascading Style Sheets
CSV	Comma-separated values
DBMS	Database Management System
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
ELSTAT	Greek Statistical Authority
EXIF	Exchangeable image file format
GDP	Gross Domestic Product
GIS	Geographical Information Systems
GML	Geography Markup Language
GUI	Graphical User Interface
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HTML	Hyper Text Markup Language
HTTP	HyperText Transfer Protocol
IDE	Integrated Development Environment
INSETE	Institute of the Association of Greek Tourist Enterprises
JSON	JavaScript Object Notation
KML	Keyhole Markup Language
LTS	Long-term Support
MVC	Model View Controller
OGC	Open Geospatial Consortium
OPTICS	Ordering points to identify the clustering structure
OS	Operating System
PHP	PHP: Hypertext Preprocessor
POI	Point of Interest
PPGIS	Public Participation GIS
REST	Representational State Transfer
SETE	Statistical data of Greek Tourism
SNFCC	Stavros Niarchos Foundation Cultural Centre
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
SSE	Sum of the Squared Error
UAOI	Urban Area of Interest
UNIX	Uniplexed Information and Computing System
URL	Uniform Resource Locator

VGI	Volunteered Geographic Information
WCS	Web Coverage Service
WFS	Web Feature Service
WMS	Web Map Service
XML-RPC	Extensible Markup Language remote procedure call
XML	Extensible Markup Language

List of Equations

List of Code-Snippets

1 Introduction

1.1 Background

1.1.1 Motivation

Tourism is one of the largest sectors in Greece and one of the greatest contributors to the Greek Economy, as the ¼ of the Gross Domestic Product (GDP) is generated by tourism highlighting both its catalytic importance for the national economy and employment (Pegkas 2020). This fact makes tourism extremely important and reveals opportunities for studying and exploring further ways, than the traditional, for its development.

The lack of research models based on identifying and analyzing the components that determine the country's tourism patterns using big spatial data from social media is one of the many problems Greece has to face. Such models can bring to light and underline weaknesses and strengths. Weaknesses can be limited down, and strengths can be used as growth points. Although there is literature that proves that geotagged social media data can be used as tools to analyze and visualize the preferences of tourists in urban areas (Salas-Olmedo et al. 2018; Katsoni and Segarra-Oña 2019), tourism authorities/companies, policymakers, and urban planners rarely take advantage of them to find solutions.

It must be mentioned that the main reason for this absence is the fact that the analysis of the tourism industry based on data and especially on big spatial data is a recent specialization that has not yet attracted the interest of many researchers in Greece. Thus, this research project, aims to contribute to this rising field of study in Greece and promote modern ways to study and analyze tourism patterns through geosocial data.

1.1.2 The importance of tourism in Greece 2009-2019

Tourism, in contrast with most activities in the primary and secondary sectors, is considered a horizontal activity. This means that tourism is an activity delimited by the product demand and services while the activities of the primary and secondary sectors are activities of production and product supply. Tourism affects many sectors of the economy such as transportation, accommodation, entertainment, shopping, and food & beverage services. Thus, tourism is an activity, that in any case, concerns many parts of the social and productive "web" of a country.

The economic crisis of 2011 in Greece, led many industries to limit their intentions for growth and to comply with the rigorous measures the European Central Bank and the International Monetary Fund imposed on the country (MarketLine Industry Profile: Travel & Tourism in Greece. 2020). However, tourism was the "fortification" against the recession and unemployment during the crisis years and then led the country to positive growth rates (Ikkos and Koutsos 2020). Studying and exploring statistical data about tourism in Greece for the decade 2009-2019 based on bulletins and studies published from the Greek Tourism Confederation (SETE) (SETE 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019), the Institute of the Association of Greek Tourist Enterprises (INSETE)(INSETE 2015; 2017b; 2017a; 2018b; 2018a; 2020) and the Greek Statistical Authority (ELSTAT), can be observed that from 2009 to 2012 the arrivals presented a gradual increase with a small decline in 2012 (Figure 1.1).

From 2013 until 2019 the arrivals raised continuously, with 2018 and 2019 holding the highest values, 30.1 and 31.3 million, respectively (Figure 1.1Table 1.1). Concerning the profits, between 2009 and 2012 there is a slight fluctuation and from 2012 to 2015 the receipts increased moderately. Although in 2016 they deteriorated, from 2017 to 2019, they improved significantly (Figure 1.2, Table 1.1).

Figure 1.1: International tourist arrivals 2009-2019
Author processing[1]

Figure 1.2: International tourism receipts 2009-2019
Author processing[2]

[1] *Data acquired from (SETE 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019), (INSETE 2015; 2017b; 2017a; 2018b; 2018a; 2020)*

[2] *Data acquired from (SETE 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019), (INSETE 2015; 2017b; 2017a; 2018b; 2018a; 2020)*

Table 1.1: Tourism Statistics in Greece 2009-2019

Tourism Statistics in Greece 2009-2019												
Year	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**Annual Growth rate (2009-2019)**
International tourist arrivals in millions	14.9	15	16.4	15.5	17.9	22	23.6	24.7	27.2	30.1	31.3	11%
International tourism receipts in billions €	10.4	9.6	10.5	10	11.7	13	13.6	12.7	14.2	15.6	17.7	7%
Contribution to GDP	15.9%	16.0%	15.8%	16.4%	16.3%	17.3%	18.5%	18.6%	19.7%	20.6%	20.8%	3%
Contribution to employment	17.7%	17.8%	17.6%	18.3%	18.0%	17.3%	23.1%	23.4%	24.8%	25.9%	21.7%	2%

Author processing[3]

Based on the bulletin "The contribution of Tourism in Greek Economy in 2019" (2020) and taking into consideration the information available in Table 1.1, tourism in 2019:

- presented an increase in receipts from international tourism by 13,1% or by 2.1 billion euros
- had a significant investment activity of 3.2 billion euros of which the 1.2 billion in domestic added value
- contributed to the country's GDP directly with 20.8% and from 27.5% to 33.1% indirectly
- contributed directly to employment with 21.7% and indirectly from 37.6% to 45.2%
- covered with the travel receipts 79.9% of the goods deficit. Those receipts are equal to 78.4% of the receipts from other country's exports.

From the above analysis, it is obvious on the one hand the importance of tourism in Greece and on the other hand its strength as a sector.

1.1.3 Tourism in Athens-The study area

Athens is the capital of Greece, it covers most of Attica County and it is surrounded by 5 mountains, Parnitha, Penteli, Ymittos, Mount Egaleo, and Mount Poikilo. It faces from the southwest the Saronic Gulf and the port of Piraeus, which is the largest port in the country and one of the largest in the Mediterranean[4]. On March 28, 2001, the modern Athens International Airport "Eleutherius Venizelos" became fully operational, replacing the old International Airport of Ellinikon and allowing the city to bloom as a tourist destination by attracting more foreign flights.

According to the Institute of the Association of Greek Tourist Enterprises (INSETE 2020) for the period, 2016-2019 Athens presented an annual growth rate of 10% in international arrivals, an annual growth rate of 16% in receipts, and an annual growth rate of 12% in overnights. The

[3] *Data acquired from (SETE 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019), (INSETE 2015; 2017b; 2017a; 2018b; 2018a; 2020)*

[4] *www.cityofathens.gr*

international tourist arrivals in 2019 increased by 1.4 million travelers compared with 2016 and the overnights exceeded 30 million (Table 1.2).

Table 1.2: Key Figures of International Tourism in Athens 2016-2019

Kye Figures of International Tourism in Athens 2016-2019					
Year	**2016**	**2017**	**2018**	**2019**	**Annual Growth Rate**
Arrivals in millions	4.5	5.1	5.7	5.9	10%
Receipts in billions €	1.7	2.1	2.3	2.6	16%
Overnights in millions	24.8	29.4	31.4	34.0	12%

Author processing[5]

Athens, apart from being an intermediate station for other Greek destinations, mainly islands, is also a pole of attraction for tourists who are looking, among other things, to explore its rich cultural heritage. Characteristics such as the ideal weather, the museums, the archaeological sites, and the great hospitality marked Athens out as a top destination and placed it between the most attractive European cities for tourism all year long(Ministry of Tourism 2014).

1.2 Social media & geographic information

Social media is the toolkit of online applications based on the ideological and technological foundations of Web2.0 as defined by (Kaplan and Haenlein 2010). In particular, they allow the creation and the exchange of content created by users. In the already developed social network applications the participants publicize daily several different types of data, which are part of different topics based on individuals' interests. Those data are a source of valuable information that can be used to analyze people's traces and footprints in urban areas.

Since 80% of the information shared in social media has a spatial reference (Hoefer et al. 1994), the participatory trend of the crowd who shared those data was expected to influence Geographic Information Science. In Geographic Information Science (GISc) the concept of public participation first appeared with the form of collective map creation and Goodchild (2007) introduced the term Volunteered Geographic Information (VGI) to describe this particular phenomenon. A successful example of a VGI and direct involvement is the OpenStreetMap[6] project, in which people are called to enrich the OpenStreetMap base map with earth surface elements as well as to add new entries to uncharted areas.

However, public involvement in various projects can be done also indirectly. In indirect evolvement, geospatial data that are freely distributed from the public are being exploited without being requested for any particular purpose. Applications such as Twitter, Flickr, Panoramio, in countries that support geographic location recordings, automatically save a geographic component in their metadata. Those data, which hold geographic information in their metadata, are referred to in the literature as geosocial data (Croitoru et al. 2014) and they are presented as a form of Ambient Geographic Information (AGI)(Stefanidis et al. 2013), the opposite of VGI. The indirect participatory geography and the AGI resulted from the use of

[5] *Data acquired from (INSETE 2020)*

[6] *https://www.openstreetmap.org/*

big volume multitude data with spatial reference from social media. Innovatively, combining heterogeneous data sources and formulating new hypotheses, several new products and research results have emerged. An example of AGI and indirect involvement that uses geosocial data is the project "Building Rome in a Day" (Agarwal et al. 2011). In this project photographs that were taken in Rome and uploaded to Flickr were used to create 3D models of the city's buildings.

This form of involvement is of particular importance because it provides new research perspectives in many fields and creates implicit opportunities both for researchers and stakeholders.

1.3 Aim & research questions

The abundance of tourist guides and tourist promotions created by tourism authorities or by municipalities includes mainly cultural, accommodation, and food and nightlife options. The disadvantage of those guides is the lack of information about which areas are the most attractive and famous according to the preferences as well as the spatial and temporal distribution of tourists. However, geosocial data gives the solution to this problem providing information about tourists that can be analyzed and contributed to the tourism decision-making process.

This research aims to disclose spatial and temporal patterns of tourists by mining spatial data from Flickr's social media platform.

To accomplish the overall aim, it can be divided into the following sub-objectives:

a) To study and implement methods and techniques to retrieve and analyze geosocial data
b) To identify important Areas of Interest (AOIs) in the city of Athens exploiting the traces tourists leave on social media
c) To examine the temporal and seasonal visiting trends of tourists on the AOI's
d) To determine the nationalities of tourists that have visited each AOI

Those objectives are related to the following research questions:

- ***RQ1(related to objective a):*** What are the existing methods and techniques to:
    - mining geosocial data?
    - reveal tourism spatial distributions from geosocial data?
- ***RQ2(related to objective a):*** What are the selected methods to be implemented in this study?
- ***RQ3(related to objective b):*** Does **(a)** the produced AOIs include corresponding POIs and did **(b)** the clustering analysis reveal new tourist spots?
- ***RQ4(related to objective c):*** What years did the AOIs present highest visitation rates?
- ***RQ5(related to objective c):*** Is the phenomenon of seasonality detected in the AOI's, or tourist concentration is distributed equally throughout the year?
- ***RQ6(related to objective d):*** Are tourists from different countries of origin prefer to visit specific AOI's more than others?

1.4 Thesis Structure

After this introductory chapter, Chapter **2** discusses previous studies about the importance of geosocial data and the methods and techniques applied to harvest them, as well as the spatial analysis methods used to disclose the spatial patterns from them. Also, it reviews the spatial context of famous social media platforms and provides a reference about the existing open-source Web-GIS technologies, that can be used in the visualization process of results. Through this chapter, an answer is given to **RQ 1**.

Chapter **3** describes the selected methodology to be implemented to mine, clean, store, process, and visualize the big geosocial dataset provided by Flickr. All the techniques and technologies used to produce the needed results are described, as well as code snippets and system architecture, are provided. Through this chapter, **RQ2** has been answered.

Chapter **4** presents the results produced by the selected methodology through tables, graphs, and a Web-GIS interface created as a supplement tool for this study, and a discussion section is provided. Through this chapter, **RQ 4,5, and 6** have been answered.

Finally, the last chapter is Chapter **5** which provides a Discussion on the results of this research, and a thorough answer to **RQ3** is given. After Chapter **5** a section about the conclusions of this work is provided with future recommendations.

2 Literature Review

2.1 Reviewing the spatial context of famous social media platforms

2.1.1 Flickr

Flickr was created by a Vancouver-based company the Ludicorp in 2004. Now after many different ownerships, Flickr belongs to SmugSmug and it is an active online photo-sharing social media platform containing more than 100 million active accounts and more than 10 billion photos and videos (Loizos 2019). It is probably the most popular web application in the world for organizing and sharing photos. Within Flickr users can upload photos, share them, and add metadata such as geolocation of the photograph, tags, license information, and personal information e.g., where they live, and many more.

The most interesting feature of Flickr which makes its data so powerful for spatiotemporal analysis is the existence of EXIF (Exchangeable image file format) data. Flickr's EXIF data is something like an ID for the photograph. When a photograph is captured by a digital camera, EXIF data are stored as metadata including information such as camera settings, date and time of capture, copyright ownership, and geolocation[7].

Virtually all the features of the various Flickr platforms -web, portable, desktop- are followed by long-term support (LTS)[8] Application Programming Interface (API). The Flickr API, which is going to be used in this study, allows external applications, websites, as well as individual developers to communicate with Flickr's database and exchange information. With over 10 billion photos (many of them with significant metadata) the Flickr community creates impressively rich data and great opportunities for researchers in different fields, and the Flickr API enables access to this data. Almost all the features offered by Flickr are made available through the API which is completely free to developers and anyone who wants to interact with it.

2.1.2 Facebook

Facebook is a social networking phenomenon with over 2.7 billion monthly users in 2020, founded in 2004 by 4 students of Harvard University (Mark Zuckerberg, Eduardo Saverin, Dustin Moskovitz, and Chris Hughes) (Hall 2020). Facebook allows users to create profiles and enrich them with information ranging from where they live and their favorite activities to their political and religious views. Each user can attach any file he/she desires (photo, video, image, text) in his/her personal profile which is called "post". Depending on the user's privacy settings this post can be visible to the public or specific users of his/her choice.

In spatial analysis terms, Facebook allows users to post their location to any Point of Interest (PoI) of the physical world, as well as the ability to add their own PoI. Those points, where users declare their location, are registered in the network database and contain the total number of the users who have visited them since their creation. This cumulative number can be beneficial for the planning process as it can quite easily shape the current state of an area by

[7] *https://help.flickr.com/en_us/understand-flickr-exif-data-r1ge02Xo1X*

[8] *Long-term Support is a complete and stable version of a product which is been supported by the company longer time (the company decides the range of the time, it could be months or years) than the standard edition of the same product. (Ashby 2018)*

recognizing the values of areas of interest and population concentration. The map of the PoIs of an area, after proper preparation, can be used for the extraction of data in real-time.

2.1.3 Twitter

Twitter was designed in 2006 by Evan Williams and Biz Stone and it is quite famous around the world with 340 million monthly active users and 500 million tweets sent per day (Britannica 2020). Twitter belongs to the category of micro-blogging social networks as it allows its users to compose short messages of up to 140 characters. Within the network, each user creates a communication channel to which she/he can add in addition to his/her messages and information flows from other channels (follow). This network of channels is visible to anyone with access to the service platform and is a popular way to stay up to date.

From a spatial point of view, Twitter allows a message to be geo-referenced in such a way that the physical space to which the message refers can be identified. Unfortunately, Twitter offers a sample of their data for free, only a 1% (unverified) sample. However, a 1% sample of Twitter is still in the order of a few million tweets a day. In each of those tweets, you can get a lot of information such as the text of the tweet, user profile information, and geolocation and you can get all of that information for retweets and quoted texts. In particular, the analysis of texts and messages can contain useful geo-information for accurately locating the reference position of a message.

A great example of exploiting the location of tweets and create a global map from big data is the huge project of Mapbox back in 2013 (Gundersen and Mapbox 2013). Analyzing geolocated messages on Twitter, the Mapbox platform creates a global thematic map that maps the areas where locals and visitors gather. It is found that these areas are not the same which can be translated as missed opportunities for visitors who do not know the beautiful places where locals gather. Through this visualization, questions can arise as to why this is happening, what the motivations are, and how this can change.

2.2 Related studies

2.2.1 The importance of geosocial data in tourism through related studies

Even nowadays where the evolution of technology and science provides a wide range of tools capable to automate the gathering and analyzing data on tourist behavior, old-fashioned data sources such as on-site observations, counting visitors, questionnaires, and surveys are still used to manage and plan tourism destinations. In the last decade, many researchers tried to prove that geosocial data can offer new approaches to study tourism patterns and tourist preferences without manual efforts. The below studies have been selected and reviewed as the most comprehensive, modern, and suitable to provide useful insights to the subject of this research and answer to the **RQ1** about what are the existing methods and techniques to reveal tourism spatial distributions from geosocial data.

Höpken et al. (2020) investigated if social media photo-sharing platforms such as Flickr can be reliable and useful to acquire information related to the spatial flow of tourists and their visitation tendency in Points of Interest(PoI's) and made a comparison between two of the most noteworthy clustering techniques that could determine PoIs in different situations. The results of their study, which has been conducted for the city of Munich, showed that the POI's

identified by the clustering algorithms could be related to already known POI's in the city and both of the techniques can be beneficial in different situations. This case-study approach determined, among other things, the appropriateness of Flickr for harvesting essential and valid data for identifying PoI's and tourists' footprints. Moreover, this study is particularly useful in studying which clustering methodology is the most efficient to identify PoI's from a tourist point of view, but it does not provide any clue about the temporal distribution of the visitors.

Unlike Höpken et al. (2020), in the study of Koutras et al. (2019) the spatiotemporal analysis of geotagged photos was the main purpose. After reviewing related works, they decided not to orient their study on comparing spatial clustering techniques, but to use the most popular one and give all their attention to the spatiotemporal analysis of geosocial data. Their goal was to expose, by applying a density-based clustering technique to geotagged photos from Flickr social platform, the most popular areas of interest (AOI's) in the city center of Athens and study the movement of the tourists in these areas. From their analysis emerged that the result areas merged with the already know famous landmarks which means that the analysis of geosocial data concerning tourism trends can produce reliable results. In addition, important statistical insights presented about the temporal preferences of the visitors in each area including what day, month and year present the highest visitation rate.

Focusing on the same context, as Höpken et al. (2020) and Koutras et al. (2019), of revealing areas of interest analyzing geosocial data from tourists, Halim et al.,(2018) explore the available opportunities in finding new tourism spots by the exploitation of big data in tourism in Sabah of Malaysia. The data used in this study were retrieved from the Application Programming Interface (API) of Twitter. The study consists of three stages, data extraction, data analysis, and the presentation of the data through maps. The spots resulted from their analysis were coincide with the majority of the already known attractions in many Sabah cities and their temporal analysis highlighted the fact that during holidays the visitation rate of some spots increased in comparison with a simple day. Although the majority of the spots found were related to the existing attractions of the cities, two new tourism spots were detected after applying an outlier analysis.

Other studies that support the idea that the analysis of geosocial data could be beneficial in tourism management, focused not only on the detection of AOI's or POI's but also on the investigation of the differences in spatial distribution between locals and tourists. In the study "Analyzing and visualizing the spatial interactions between tourists and locals: A Flickr study in ten US cities" (Li et al. 2018) the researchers used data from the Flickr social media platform to study the correlation between the preferences of the locals and the tourists in 10 US cities and if there are overlaps between the two datasets and they combined many spatial analysis methods to produce their results. After their analysis, they found that the range of intersection between tourists and locals varies across the cities. Some cities present highly mixed groups and other cities lower mixed groups. However, the study proves that in each city the locals prefer cultural and recreational destinations while tourists prefer more the city center with popular landmarks and historic sites. Although the first intention of this study was to compare the preferences of locals and tourists, its result could be efficiently used by planners and tourism managers to improve the detected areas. In addition to this scope, Garcia-Palomares et al.,

(2015) identified the most preferable and visited areas in 8 major European cities by analyzing geotagged photos from Panoramio taken by tourists and locals. They aimed to present the contribution of a photo-sharing social media platform such as Panoramio in the identification and analysis of popular tourist attractions. They extract the geosocial data from Panoramio, split the dataset based on if the photos were taken by tourists or locals, and analyzed their spatial distribution. Their results were similar to the results of the mentioned study of Li et al. (2018) while they discover that tourists show higher spatial concentration than locals and, in some cities, the concentration was denser than in others. Both studies provide a picture of how different traits locals and tourists can leave in a city and how they can be identified by exploiting geosocial data. However, they lack information on the temporal and seasonal preferences of visitors and residents.

Geosocial data can be used not only to identify tourist spots but also to reveal tourist routes. In their paper, Spyrou et al. (2015) exploit geotagged photos of Athens downtown derived from Flickr photo-sharing social media platform to inspect touristic user-generated routes, by employing an innovative two-level clustering pattern. The study's goal is to analyze those routes and choose the most characteristic for a specific area. With a focus on providing to the future visitor of Athens a desirable route based on geosocial data and the preferences of other visitors, this study processed big geosocial data, identified famous routes following the social traces of tourists, and used a clustering pattern to produce the proposed route. The researchers through this study highlighted that social media platforms with proper use can produce valuable information for tourism planning. Lastly, they tested the impact of their findings and how the proposed routes are appreciated using local people and concluded that their methodology has prospects and tourists will be benefited from such routes.

2.2.2 Clustering techniques for spatial analysis used in related studies

To identify AoI's from a large dataset of geotagged photos clustering techniques must be used. Clustering is a machine learning technique that is commonly used in grouping data points. In the clustering problem, a set of data is given, without the corresponding classes, and there is a need of an algorithm that will automatically group the data into clusters. For the clusters that are created, the data must be grouped correctly. This practically means that a cluster must be made up of objects, where each object is closer to the other object of the same cluster than another object of a different (Bhadane and Shah 2020). Using clustering analysis, important information can be acquired from the data points by investigating the different data point groups produced after the application of the clustering technique.

The majority of the researchers in the literature have used for their analysis a density-based clustering technique the Density-Based Spatial Clustering of Applications with Noise (DBSCAN). However, other interesting clustering techniques were found in the literature such as k-means, and a novel clustering approach the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). The aforementioned techniques will be further discussed after the presentation of the clustering techniques used in related studies.

As it is mentioned above many studies choose to cluster their data by using the DBSCAN methodology as it presented the least drawbacks. Muñoz et al. (2020) applied a DBSCAN

analysis to the spatial data they have gathered through Flickr and PPGIS(an online mapping platform) to explore differences in spatial distribution between the datasets of the aforementioned platforms. In the same context, Koutras et al. (2019) set 4 specific criteria to choose the best clustering technique found in literature and they decided that the DBSCAN meets all their rules for analyzing a big dataset from geotagged photos. Li et al.(2018) to avoid the fact that maybe a visitor post more than one photo of the same location, applied the DBSCAN clustering technique to prevent overlapping and to identify the number of unique destinations both tourists and residents visit. Hu et al. (2015) identified urban areas of interest and their change in ten years (2004-2014) for New York City (NYC), London, Paris, Shanghai, Mumbai, and Dubai by applying the DBSCAN technique to a point dataset retrieved from Flickr geotagged photos.

Other studies tried to make a comparison between clustering techniques to choose which one fits better to their needs. Höpken et al.(2020) compared two clustering techniques, k-means, and DBSCAN to identify PoIs from metadata of photos stored in Flickr. To achieve that they followed a specific methodology, they extracted and prepared the photo metadata (upload time, location, and user) from Flickr API, they clustered the data using both DBSCAN and k-means to identify the POI's and then they applied the association rule analysis (FP-growth algorithm) and sequential pattern mining (generalized sequential pattern algorithm) to reveal tourism patterns. Lee et al. (2014) referred to the advantages and drawbacks of two clustering techniques, k-means, and DBSCAN to understand which one fits their needs better. After the comparison, they decide that DBSCAN is more suitable for identifying spatial patterns in a dataset with noise. Kisilevich et al. (2010) aimed, among other tasks, to find attractive areas using data from two photo-sharing platforms, Flickr and Panoramio. To accomplish this specific task, they considered that two density-based clustering algorithms, OPTICS as well as DBSCAN would be an appropriate approach, and decided to use DBSCAN as the most efficient for this task.

An exception was the study by Chen et al. (2019). Unlike the previous studies, they provide a well-written reference to the advantages and the drawbacks the DBSCAN clustering techniques may have in extracting urban areas of interest from geosocial data and proposed a better and more advanced clustering technique of the DBSCAN the HDBSCAN which overcomes the possible disadvantages of its parent algorithm. They focused on identifying Urban Areas of Interest (UAOIs) using geotagged images and their metadata from Flickr. For this analysis, they proposed a unique methodology of spatial analysis techniques that combined the Hierarchical Density-Based Spatial Clustering with noise (HDBSCAN) and the 'a-shape' algorithm. In contrast to the traditional spatial techniques presented in the literature, this methodology produces higher accuracy.

All of the discussed studies provided valuable information for this research on various aspects. They study ways to harvest and analyze geosocial data, propose which is the best clustering technique based on their findings and support the general idea that geosocial data are a source that already should be used in tourism planning. Reviewing the literature, it has been find out that the majority of those studies applied DBSCAN to cluster their data, did not give much attention to the temporal and seasonal analysis of the data, and used graphs and static maps to

visualize their results. This study is going to contribute to this research area by using the HDBSCAN algorithm, a methodology that only Chen et al. (2019) applied, give attention to the temporal and seasonal analysis of the data, create an open-access API that anyone can use and request the data and build a modern web-GIS user interface to visualize the results so that a visitor or a tourism authority to easily interact with the map.

2.3 Clustering algorithms

2.3.1 K-means clustering algorithm

K-means algorithm (MacQueen 1967; Lloyd 1982) is one of the oldest and most widespread clustering algorithms. It is one of the most representative algorithms that use a quadratic error criterion since they define the sum of the squared error as an objective function as well as it is very easy to grasp because it uses a small number of computations (Jain et al. 1999).

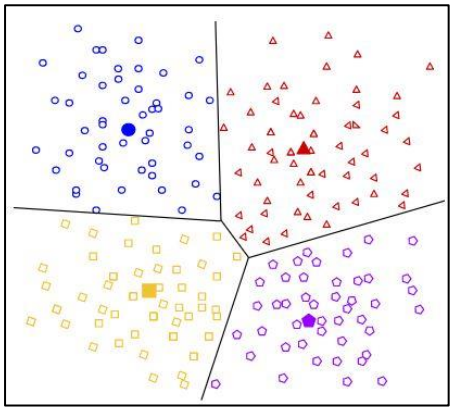


Figure 2.1: Example of k-means centroid clustering technique
Source: (Google Developers 2018)

In this section, the procedure of k-means is provided as described by Tan et al. (2006). In the begging, ks are selected as the initial centroids, which are also known as centers, with k being a constant that the user selects based on the desired number of groups he wants to achieve. Then each point is assigned to its nearest centroid (or center or middle) and k groups are formed. The centroid is then updated and recalculated based on the points belonging to the group to be placed in the center of the points as a mean. These steps of assignment and update are repeated until the points do not change group or respectively until the centroid remains the same.

The goal of grouping is expressed by an objective function that depends on the relationships of similarity or the distances between the points. In the case of partition algorithms and more specifically in the case of the k-means algorithm these relationships can be studied as the distances between their points and their centroids (Tan et al. 2006; Han et al. 2012; Google Developers 2018). Considering the Euclidean distance as a metric of distance, the sum of the square error (Sum of the Squared Error - SSE) can be defined as an objective function of the k-means algorithm, i.e., the sum of the Euclidean distance of each point from its nearest centripetal (center). If x is been defined as an object, C_i the i^{th} group, c_i the center of the i^{th} group, n the number of objects in the data set, n_i the number of objects in the i^{th} group, and k

the number of groups, then the sum of the square error SSE is defined presented in Equation 2.1(Tan et al. 2006):

Equation 2.1: SSE

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} |c_i - x|^2$$

Source: (Tan et al. 2006)

Using this mathematical formula as an objective function, k-means belongs to the category of squared error-based algorithms. The k-means algorithm aims to minimize this objective SSE function. Finally, based on the above objective function, it follows that the best grouping will be defined with the lowest SSE value.

2.3.2 DBSCAN clustering algorithm

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. 1996) algorithm is one of the most common density-based grouping algorithms. Unlike many new algorithms, it has a very well-defined grouping model called density-reachability, which connects points within certain distance limits. However, it only connects points that meet a specific density criterion defined as a minimum number of objects within a radius. Besides, a group consists only of points that are connected by density (density-connected) and can be taken in any arbitrary shape by the density of the data space (Schubert et al. 2017; Thompson 2019)

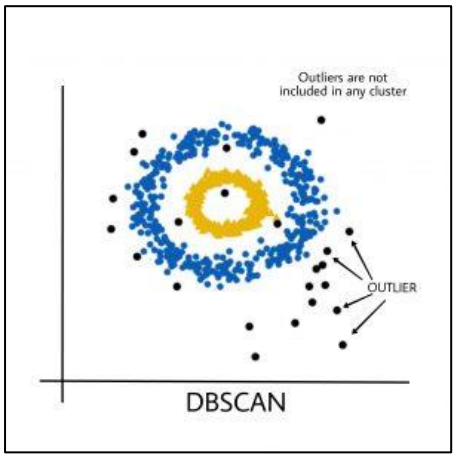


Figure 2.2: DBSCAN clustering algorithm
Source[9]

Here, some of the basic concepts of DBSCAN derived from Han et al. (2012) and Schubert et al. (2017) have been reviewed:

- The ε-neighborhood of an object (point) is the space centered on the object and the radius ε (radius ε, also known as eps and is given by the user).

[9] https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/ *(DBSCAN Clustering in ML | Density based clustering. 2019)*

- The density of a neighborhood can be calculated from the number of objects (points) in the neighborhood.
- Core objects are objects that have at least minPts of neighboring objects in their neighborhood (including themselves). And here the minPts number is provided by the user.
- For a core object q and an object p, it can be said that the object p is directly density-reachable from the object q, concerning ε and minPts, if p is inside the ε-neighborhood of q.
- An object p is accessible via density (reachable) from an object q, concerning ε and minPts, if there is a chain of objects $p_1,\ldots, p_n$ such that $p_1 = q$, $p_n = p$ and p_{i+1} is directly accessible through density from p_i
- Two objects p_1, p_2 are connected through density (density-connected), concerning ε and minPts, if there is an object q such that both objects p_1, p_2 are accessible through density from q.
- Objects that are not accessible by density from a kernel object are considered noise and do not belong to a group.

The operation of DBSCAN starts with a random point p and calculates its density, which is the number of points in the ε-neighborhood of p. If p is the core point, DBSCAN marks this point as a new cluster and then recovers all density reachable points from p and assigns them the same cluster label with p. Otherwise, point p is marked as a noisy point. DBSCAN then repeatedly collects the points that are density reachable from the core points. The process terminates when no new items can be added to any cluster. Meanwhile, if a point is not in the ε-neighborhood of a cluster, it is considered a noise point, which is a possible outlier. Thus, DBSCAN detects arbitrarily shaped clusters and is not affected by the data entry sequence. Each new item entered affects only one neighborhood (Ester et al. 1996).

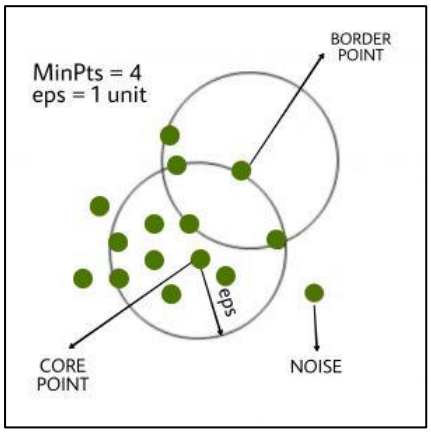


Figure 2.3: DBSCAN algorithm operation
Source[10]

[10] https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/ *(DBSCAN Clustering in ML | Density based clustering. 2019)*

2.3.3 HDBSCAN clustering algorithm

The HDBSCAN algorithm created by Campello et al. (2013) is an improved version of DBSCAN in which clustering is based on the distribution of data density in space. DBSCAN classifies in the same cluster the points that are close to each other, while at the same time it may leave points unclassified, especially in areas with low density. In this way, the algorithm succeeds in separating the main clusters from the noise created by the data. The extension to HDBSCAN is the transformation of DBSCAN into a hierarchical grouping algorithm and the extraction of clusters is based on their stability (stability of clusters). In HDBSCAN only the parameter ε is entered and the method applies the DBSCAN algorithm for a range of values of ε. The basic idea of the algorithm is that by gradually decreasing the parameter ε, that is, increasing the required density level of a cluster, then the cluster shrinks, but remains united, to a density level at which either the cluster splits into smaller ones or disappears completely. As a result, the HDBSCAN algorithm produces a clustering tree containing all the clusters discovered by the DBSCAN algorithm (Berba 2020). An example of data grouping using HDBSCAN is shown in Figure 2.4. It can be observed that the algorithm detects some major clusters, while several points remain unclassified and are considered noise.

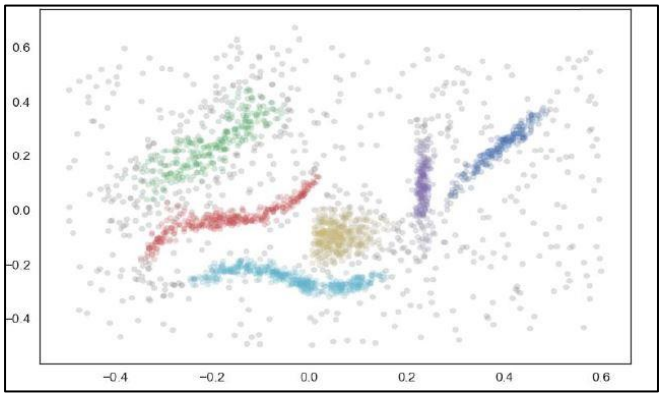


Figure 2.4: Clustering example with HDBSCAN
Source: (Berba 2020)

2.4 Review of open-source WEB-GIS technologies

2.4.1 Client Map libraries

2.4.1.1 OpenLayers

OpenLayers[11] is an open-source library, written in JavaScript, that allows the creation of dynamic maps on any web page, having the ability to present vector data, map levels, markers, and many other features, while it gives the developer the advantage of customizing its functions and features. It is freely available and offers developers a variety of tools for cartographic applications. Among other things, it provides ready-to-use functions for coordinate transformations and computations in cartographic projections, while supporting almost all known projection systems worldwide. It also offers many ready-made controls for maps such as zoom bars and memos for layers that are overlayed on cartographic backgrounds. It supports

[11] *https://openlayers.org/*

KML, GML, GeoJSON files, but also it can illustrate data that use OGC standards such as WMS and WFS. It can accept additional add-ons and is compatible with HTML and CSS3.

2.4.1.2 Leaflet JS

Leaflet[12] is a leading lightweight JavaScript open-source library for interactive maps. Leaflet was designed with simplicity, performance, and usability in mind, and it is very friendly for mobiles. Works very well on major desktop and laptop platforms, can be expanded with many add-ons, has a nice easy-to-use and well-documented API, and a simple, easy-to-read source code. Although it is a small JavaScript library, it is packed with almost every feature the user will need and if a feature is not available in the basic JavaScript library, it may be available as one of many add-ons available.

2.4.2 Map Server

2.4.2.1 GeoServer

GeoServer[13] is an open-source Web Server that allows its users to manage and share geospatial data. It is written in the Java programming language and since the Geoserver Project was launched in 2001 it has been constantly evolving by a wide community of users and developers around the world. It fully complies with the standards of the Open Geospatial Consortium (OGC) and through the web interface, it can accomplish many tasks along with the different data formats it serves. Some of the data formats are vector data, raster data, databases, cascade service data, application schemas. Regarding spatial data, GeoServer offers services such as Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), and others. It supports a huge number of reporting systems, both global and local (geocentric and geodetic) while allowing on-the-fly transformation from one system to another.

2.4.3 Databases

2.4.3.1 PostgreSQL

PostgreSQL is an open-source object-relational Database Management System (DBMS) with many and ever-increasing capabilities. It has been developing for over 20 years and is based on a proven good architecture that ensures reliability, data integrity, and proper operation. It runs on all major operating systems, such as Linux, UNIX (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64), and Windows and it is also ACID compatible (ACID compliant), which means that it complies with the following set of properties, Atomicity, Consistency, Isolation, and Durability (The PostgreSQL Global Development Group 2020). It is a complete relational system that supports multiple schemas per database, while its directory (information about tables, columns, views, etc.) is available through the Information Schema, as defined in the SQL standard. Regarding data integrity features, including primary keys, foreign keys with support for restricting and cascading updates/ deletes, check constraints, unique constraints, and not null constraints.

[12] https://leafletjs.com/

[13] http://geoserver.org/ (GeoServer. 2020)

2.4.3.2 MongoDB

MongoDB[14] is an open-source cross-platform database. It is described as a non-relational (NoSQL) database because it uses a structure consisting of dynamic documents in a form related to JSON instead of using a table-based relational structure. This adoption offers convenience and speed in integrating data into specific types of applications. In relational databases, data is usually stored in separate tables that are defined by the developer and a specific object can be shared between multiple tables. In the database documents, all the information of an object is stored as an entry within the database and each stored object may be different from all the others. At this time, MongoDB is the most widely used database for document-oriented storage[15].

[14] https://www.mongodb.com/what-is-mongodb *(What Is MongoDB? 2020)*

[15] https://db-engines.com/en/ranking *(DB-Engines Ranking. 2020)*

3 Methodology

The Methodology chapter describes the methods, techniques, procedures, and technologies used in this research. The study's methodology consists of 5 stages. The first stage contains the harvesting of the initial dataset from the Flickr API, mapping the POIs of Athens provided by popular tourism websites and a local tourism agency, and storing the acquired datasets in the database. The second stage is about cleaning and updating the data, while the third stage includes the mining of the restored data with a density-based clustering algorithm to detect spatial tourism clusters. In the fourth stage, the clustered data has been stored with their temporal characteristics in the database and make them accessible through the web by creating a backend system that serves as a RESTful API. In the final stage, the visualization of the spatial distribution of the tourism clusters and their temporal trends through a Web-GIS user interface has been made. Figure 3.1 presents the general methodology flowchart of this study and Figure 3.2 presents a thorough flowchart of the methodology which outlines carefully all the steps followed, from the source (initial data) to the knowledge (results).

In section 3.1 the methods that have been selected to achieve the aim and the sub-objectives of this study are presented and justified and an answer can be provided to **RQ2** while in section 3.1.1 an overview of the technologies used to create the Web-GIS application is provided. The next sections describe the implementation of the techniques and methods used in each of the five stages presented in Figure 3.2. Section 3.2 presents the requests made to the Flickr API to harvest the primary dataset, the sources used to collect the POIs of Athens, and the methods used to store the datasets in the database. Section 3.3 describes the functions and techniques applied to clean and preprocess the dataset and Section 3.4 explains the implementation of the HDBSCAN algorithm. Section 3.5 clarifies the procedures made to create the backend RESTful API of the Web-GIS application while section 3.6 explains the methods and technologies utilized to create the Web-GIS interface.

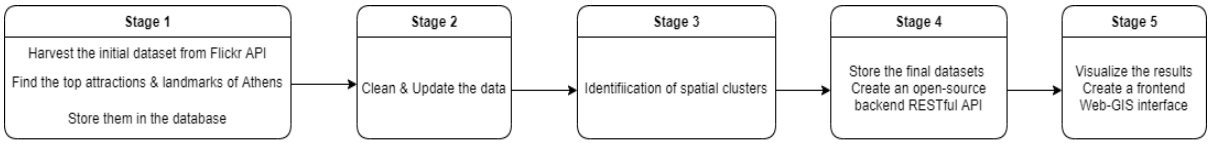


Figure 3.1: Methodology flowchart

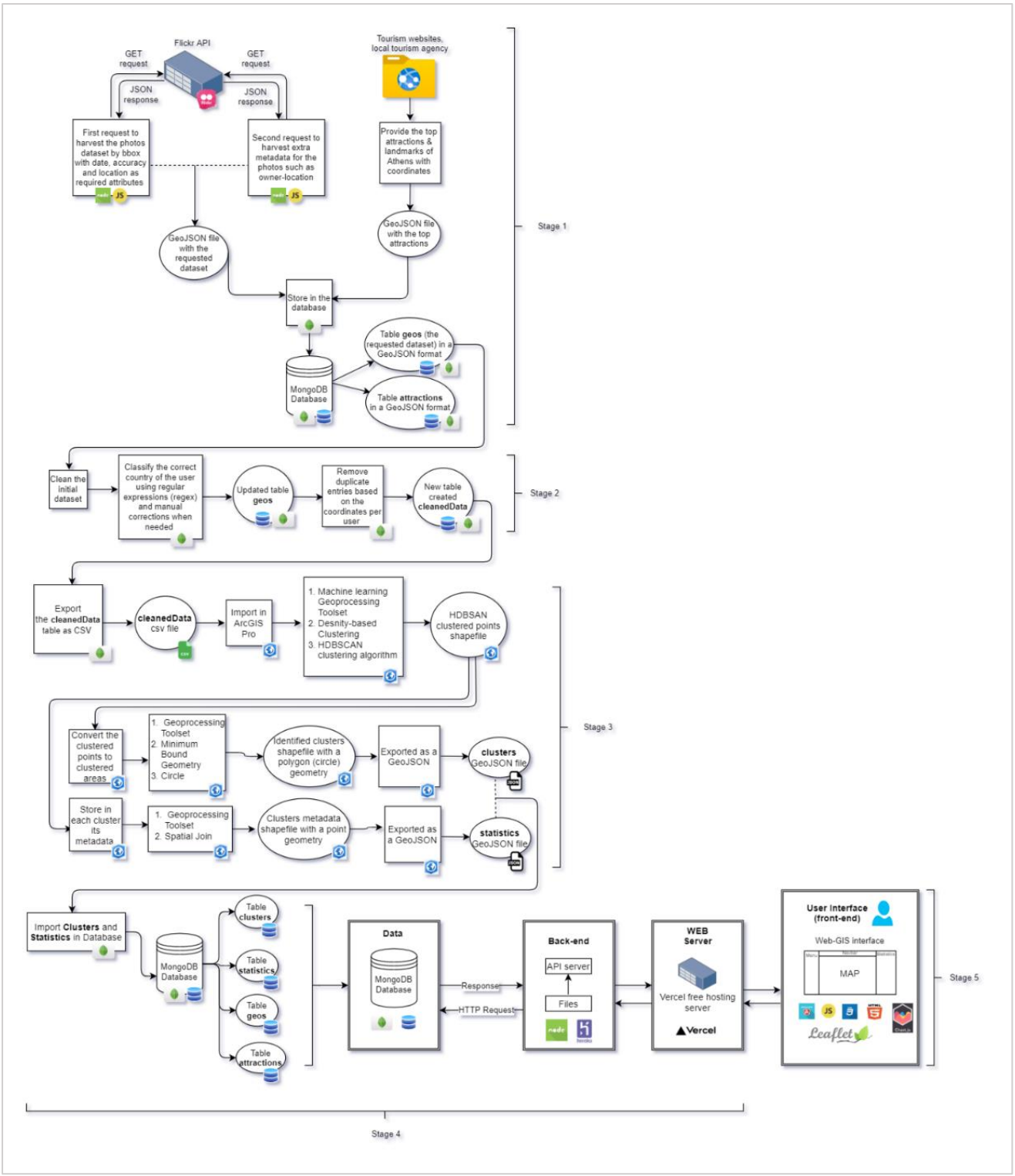


Figure 3.2: Detailed Methodology flowchart

3.1 Chosen Methods

In Chapter 2 related studies were explored and what methodologies and techniques they used to harvest, clean, and analyze geosocial data and their metadata to reveal spatial and temporal tourism trends. Inspired by those studies, this research, selected some of the methods they proposed and combined them with new methods to accomplish the overall aim of this research which is to identify tourism patterns by analyzing geosocial data.

To harvest the data from Flickr its API has been used, as the majority of the related studies did, and stored the received dataset in a non-relational database, the Mongo DB. MongoDB was chosen as the core database of this study since it uses a structure consisting of dynamic documents in a form related to JSON, the same as the form of the dataset. The greatest advantage of choosing MongoDB was that it is easier to set up and use for simple Geo data processes than PostgreSQL and in addition, it is very fast as document-object storage for handling large datasets containing geolocation (Makris et al. 2019). Moreover, MongoDB was selected firstly because this study did not need relations between the tables, secondly because MongoDB fitted better than PostgreSQL with the nature of the dataset, and last but not least, there was existing knowledge in MongoDB and JavaScript than with PostgreSQL and SQL.

To clean and preprocess the data two methods were applied to handle specific issues. The first issue was to classify correctly the users based on their country of origin. The majority of the related studies (Önder et al. 2014; García-Palomares et al. 2015; Koutras et al. 2019; Höpken et al. 2020) have used a certain technique to separate the locals from the tourists. The technique classifies as tourists, *the users that took photos in a short time of period,* and as locals *all the other users*. This technique is not very accurate and presents big disadvantages. One of its major disadvantages is that it could classify a tourist as a local wrongly if that person visits the study area many times and upload many photos and vice versa (Li et al. 2018). Moreover, with this technique the nationalities of the users cannot be found, which is of interest for this study, but only if they are categorized as locals or tourists based on the duration of the timestamps of the photos. For that reason, it has been decided to proceed by acquiring the hometown of the users requesting the Flickr API and separate the visitors according to their nationality. The second issue was to identify if duplicate or multiple same records exist in the dataset. That subject was resolved using MongoDB's specific tools and functions.

From the literature review, it has been discovered that the most suitable technique to group and analyze a big geosocial dataset with thousands of geotagged photos into spatial clusters is to use an unsupervised clustering technique, the density-based clustering technique. The reason an unsupervised clustering technique has been selected, was based on the nature of the data. This study searched for methods that could find clusters in the dataset without declaring what the clusters are. The algorithm of an unsupervised clustering technique figures out, in a data-driven way, what a cluster looks like based on the overall patterns within the data. The density-based clustering technique finds clusters based on feature locations and there are 3 different methods for density-based clustering: the DBSCAN, the HDBSCAN, and the OPTICS. The main characteristics of those methods can be found in Table 3.1.

Although the DBSCAN is the most popular and used method in the literature, this research experiments with the HDBSCAN method evaluating its advantages, presented by Chen et al.

(2019). One of its major advantages is that while it still does have that set of core distance, like the DBSCAN, the search distance can adjust depending on the data. It is a data-driven method, and it is self-adjusting. So, someone is going to find clusters of varying densities, and the only parameter that he/she is going to set is the minimum features required to be a cluster. Everything else will be data-driven. Moreover, by choosing this technique this study wants to contribute to the literature since not many related works using this technique has been found.

Table 3.1: Density-based clustering techniques main characteristics

DBSCAN	HDBSCAN	OPTICS
Uses fixed search distance	Uses a range of search distances	Uses neighbor distances to create reachability plot
Finds clusters of similar densities	Finds clusters of varying densities	Most flexibility for fine-tuning
It is fast	It is data-driven	Can be computationally intensive
	Requires less user input	

3.1.1 Technologies selected for the development of the Web-GIS application

3.1.1.1 WebStorm IDE

WebStorm is a powerful integrated development environment (IDE) by JetBrains. WebStorm provides full support for JavaScript, TypeScript, HTML, CSS as well as frameworks such as React, Angular, and Vue.js. It is also widely used to build mobile and desktop applications[16]. All the code pieces developed for the Web-GIS application were made exclusively with the use of this integrated development environment. WebStorm is considered a "smart" IDE with many features and automated tools that can make the process of writing code fast and easy.

3.1.1.2 Node.js

Node.js is an open-source cross-platform runtime environment, which is used to develop server-side web applications. It has an event-driven architecture that allows asynchronous communication and enables us to create web applications that run in real-time and interact perfectly with the user[17]. Node is based on an already widely used language, JavaScript and it was easier for us to use it for the server-side of the Web-GIS application in comparison with other server-side languages such as PHP, Java, C#, or Ruby.

3.1.1.3 Mongoose

Mongoose is a library that is used by Node.js and it can handle the communication with MongoDB database[18]. After defining the shape of each table, then the communication can easily be done with the database using the tools it provides. Essentially it provides a simple way to model the data of this application based on the format of the database tables. It also includes simple conditions to create queries to the database, check the validity of the queries, etc.

[16] https://www.jetbrains.com/webstorm/

[17] https://nodejs.org/en/about/

[18] https://mongoosejs.com/

3.1.1.4 Angular

Angular is an open-source framework based on the TypeScript programming language used to build dynamic web applications[19]. Agular's great characteristic is the software development method called MVC (Model View Controller). This method divides the application into 3 parts, the Model, which is responsible for managing the application's data, the View which is responsible for viewing the data with the help of HTML and CSS, and the Controller which is responsible for receiving input and perform functions and modifications to the Model[20]. For the frontend part of the Web-GIS application the Angular framework was chosen due to some important advantages:

1. In addition to being one of the most popular frameworks of 2020 according to google trends[21], it will officially have long-term support from Google.
2. Its component-based architecture has an easy structure and therefore it is easier for new developers to understand it.
3. It is a very strong ecosystem, as for years it has been enriched by packages, plugins, add-ons, and development tools, which means that in case of a problem there is a tool to solve it.
4. Familiarity with the Angular framework.

3.2 Harvesting and storing the initial data – Stage 1

Flickr provides a free API for non-commercial use with a wide range of services. The photos stored in Flickr include valuable metadata, such as where the photo was taken in latitude and longitude, the date that the photo was taken, where the user lives, and so on. The rich metadata of the photos and the fact that the Flickr API is free and accessible for the public made it the most efficient source to collect the primary dataset. Flickr API provides a variety of services to harvest data and each service has its purpose. To call Flickr's API and perform an action, a request must be send in REST, XML-RPC, or SOAP format to the API (Figure 3.2 – stage 1), declaring which method must be used, which arguments (metadata of the photo)must be received, and in what format response has to be (REST, XML-RPC, SOAP, JSON, PHP). For this study two services fulfilled the needs and were used to harvest geotagged photos and their metadata, the "flickr.photos.search" and the "flickr.photos.getInfo" services.

3.2.1 First Request to harvest the primary dataset

To collect the primary geosocial dataset for the study area the first request to Flickr's API is made (Figure 3.2 – stage 1). A request has been sent to the API endpoint using a REST format within the "request-promise" library of node.js a response has been received in a JSON format. The REST[22] Request format is the simplest request format that allows systems to communicate with each other efficiently. It consists of 4 HTTP request methods, the GET method which is used to retrieve data from a certain source, the POST method, which is used to create a new resource, the PUT method which is used to update a certain resource with or without an id, and

[19] https://angular.io/docs

[20] https://www.tutorialspoint.com/angularjs/angularjs_mvc_architecture.htm

[21] https://trends.google.com.br/trends/explore?date=today%205-y&q=Angular,React,Vue,Ember

[22] https://www.codecademy.com/articles/what-is-rest (What is REST? [no date])

the DELETE method which removes a specific resource by an id. To send the request, a script has been written (Code-Snippet 3.1) that used the "request-promise" library to retrieve the data from Flickr's API and the "flickr.photos.search" service with targeted arguments. This service is the fundamental tool to harvest the primary dataset because it returns a list of photos with standard arguments such as the photo id, the owner of the photo, and the title of the photo but also certain arguments a user can choose to receive provided by the Flickr API. The arguments used for the first request are explained below:

1. api_key (required): To use the services of the Flickr API and to be able to request and retrieve data, one has to apply for an API key. The process to get a personal API key is easy and straightforward and is implemented online answering some questions and explain why you want to use the API and what are your purposes.
2. min_taken_date (optional): This argument declares the minimum date that a photo has been taken. Photos with a taken date greater than or equal to this value will be returned. The date can be in the form of a MySQL datetime or Unix timestamp. This study, wanted to retrieve data from 2009 to 2019 so the min_taken_date started from 2009.
3. max_taken_date (optional): This argument declares the maximum date that a photo has been taken. Photos with a taken date less than or equal to this value will be returned. The date can be in the form of a MySQL datetime or Unix timestamp. As it is already mentioned the dataset must include photos from 2009 to 2019 so the max_taken_date given was 2019.
4. bbox (optional): The bbox – bounding box is a geographic rectangle area that can be used by a user to query photos inside the boundaries of this area. The bbox takes a comma-delimited list of 4 values as a parameter. These values are the minimum_longitude, minimum_latitude, maximum_longitude, and maximum_latitude and represent the bottom left corner and the right bottom corner of the bounding box. Since the study area of this research is the center of Athens, a bbox has been defined with longitude/latitude 23.5923,37.8058 as corner A and 23.9178,38.1479 as corner B.
5. accuracy (optional): This argument provides the accuracy level of the location information. It consists of 5 accuracy parameters and their corresponding values that a user can choose from. Those values are:
    a. World-level accuracy as value 1
    b. Country-level accuracy as value 3
    c. Region level accuracy as value 6
    d. City-level accuracy as value 11
    e. Street-level accuracy as value 16

    In this project, the parameter (e) with the value 16 is applied to the accuracy argument to receive the highest accuracy of the photo's location.

6. extras: The "extras" argument provides the user the possibility to extend his/her search and include in his/her request additional information about the photos. This study choose to use the "geo" and the "date-taken" parameters. The "geo" parameter provides the latitude and longitude of each record which is the information needed for the density-based cluster analysis and the "date-taken" parameter provides the time, day,

month, and the year a photo has been captured which is an essential element about the temporal analysis.

Although Flickr API is well documented and has many advantages, it also has some restrictions. It is very important to be aware of those restrictions, otherwise, the fetched data will contain many errors. The "flickr.search.photos" service that has been used for the first request produced only 4000 unique records in a single request, which means that if a request results in a 10000 records dataset the records above 4000 in the queue will be all the same. To solve this problem and to be sure that only unique records will be received, more requests were created to always receive under 4000 records. To achieve that a function was created (Code-Snippet 3.1) and called for every year, from 2009 to 2019. The response of the requests was returned as a JSON format and stored as objects in a JavaScript array.

Code-Snippet 3.1: First request code snippet

```
//const rp = require("request-promise"); a node.JS library for requests
//Request function to get the posts from flickr
const rp = require("request-promise");
const firstRequest = async (key, array, pages, year, monthA, dayA, monthB, dayB) => {
    for (let p = 1; p < pages + 1; p++) {
        const request = await rp({

            uri: 'https://www.flickr.com/services/rest/',
            qs: {
                method: "flickr.photos.search",
                api_key: key,
                min_taken_date: `${year}-${monthA}-${dayA}`,
                max_taken_date: `${year}-${monthB}-${dayB}`,
                bbox: "23.5923,37.8058,23.9178,38.1479",
                accuracy: 16,
                extras: "geo,date_taken",
                format: "json",
                nojsoncallback: 1,
                page: p
            }
        }).catch(error => {
            console.log('FLICKR API ERROR FIRST REQUEST');
        });
        //Store the data from the photos request to an array
        const postsData = JSON.parse(request);
        array.push(postsData);
    }
    return array
};
```

3.2.2 Second request to harvest additional information

One of the aims of this study is to explore the nationalities of the visitors. To obtain the information about the user's country of origin a second request must be made using a different service that returns additional metadata about the photos (Figure 3.1 – stage 1). The service "flickr.photos.getInfo" is the one used for the second request since it returns as a response extra metadata about the photos, such as the username of the user, the real name of the user, the country of the user, tags that might the user has added to the photo and the URL of the photo. The parameters required for this service are the API key and the id of the photo. To call the second request a loop through the ids of the photos must be made. (Code-Snippet 3.2).

The challenge of this request was to reduce the amount of time it requires. Unfortunately, Flickr API can handle only one request per second, and the "flickr.photos.getInfo" sends a request per photo id, which means that many hours are needed to obtain the extra data. To reduce the amount of time expected, the "chunk" method has been used of the "loadash" library of node.js and the JavaScript Promise.all() method. The "chunk" method breaks an array with many items into a new array with specified items and the Promise.all() method *"takes an iterable of*

promises as an input and returns a single Promise that resolves to an array of the results of the input promises" [23].

The "chunk" method was used to break down the array with the ids of the photos into smaller arrays of 100 items and the Promise.all() method was used to execute multiple requests at the same time (a request is a promise). Again, to send the second request to the Flickr API a script was written (Code-Snippet 3.2) that used the "request-promise" library and all the methods described above. The response returned in a JSON format and stored as objects in a JavaScript array exactly like the response from the first request. After storing the response, the array was updated with the dataset from the first request adding an extra key: value -the country of the owner- to each of the photo objects.

Code-Snippet 3.2: Second request code snippet

```
const _ = require("lodash");

async function requestChunk(arrayPhotoData, arrayPhotosIDs, key) {
    console.time("Get extra data for posts");
    const chunksPhotoIDs = _.chunk(arrayPhotosIDs, 100);
    for (const chunk of chunksPhotoIDs) {
        const secondRequests = [];
        for (const photoID of chunk) {

            const photo_url_request = rp({
                uri: 'https://www.flickr.com/services/rest/',
                qs: {
                    method: "flickr.photos.getInfo",
                    api_key: key,
                    photo_id: photoID,
                    format: "json",
                    nojsoncallback: 1
                }
            }).catch(error => {
                console.log(`FLICKR API ERROR SECOND REQUEST`)
            });
            secondRequests.push(photo_url_request);
        }
        await Promise.all(secondRequests).then((body) => {
            body.forEach((res) => {
                if (res) arrayPhotoData.push(JSON.parse(res));
            });
        }).catch(error => {
            console.log(error.message);
        });
        console.log(chunk.length, " extra data acquired");
    }
    console.timeEnd("Get extra data for posts");
}
```

3.2.3 Storing the received geosocial dataset

After harvesting the dataset with all the needed information, the process continued by storing it in the database (Figure 3.2 – stage 1). Since the visualization of the results was going to be through a Web-GIS interface using Leaflet, a JavaScript mapping API, the resulted dataset from the harvesting process and all the subsequent datasets that will be derived later, must be stored in the database as a GeoJSON format due to GeoJSON is a core technology in Leaflet.

To create the first table in the database a POST request method was used, and the table was given the name "geos". The "geos" table includes each photo with its metadata (id of the photo, the user of the photo, the year, month, day, and hour the photo was taken, and the country of the photo's user). In Code-Snippet 3.3 the schema of each photo object in the database is presented.

[23] https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global_Objects/Promise/all *(Promise.all() - JavaScript | MDN. [no date])*

Code-Snippet 3.3: The schema of the photos in the database

```
const geoSchema = mongoose.Schema({
    type: String,
    properties: {
        userID: String,
        ownerLocation: String,
        photo: {
            type: String,
            required: true,
            unique: true
        },
        yearTaken: String,
        monthTaken: String,
        dayTaken: String,
        hourTaken: String,
        id: mongoose.Types.ObjectId
    },
    geometry: {
        type: {
            type: String,
            enum: ['Point'],
            required: true
        },
        coordinates: {
            type: [Number],
            required: true
        }
    }
});
```

3.2.4 Map the existing top attractions & landmarks in Athens

To provide an accurate answer to RQ3 (*Does **(a)** the produced AOIs include corresponding POIs and did **(b)** the clustering analysis reveal new tourist spots?)* the top attractions and landmarks that already exist in Athens has to be found and stored in the database. To achieve this task and gather as much information as possible about the most visited attractions and landmarks of Athens, a search was made into two popular tourism websites the "Visit Greece"[24] and "Lonely Planet"[25] and an interview with the Chief Product Officer (CPO)[26] of a local tourism agency called "Travel Genius" [27]. The aforementioned data has been manually processed, with a coordinate pair and extra information. They stored in the database with the name "attractions" in a GeoJSON format and the schema of each attraction object can be seen in Code-Snippet 3.4

[24] https://www.visitgreece.gr/

[25] https://www.lonelyplanet.com/

[26] *Dimitris Papadopoulos, CPO of "Travel Genius"*

[27] https://travelgenius.gr/

Code-Snippet 3.4: The schema of attractions in the database

```
const attractionSchema = mongoose.Schema({
    type: String,
    properties: {
        name: String,
        category: String,
        img: String,
        source: String,
        sourceName: String
    },
    geometry: {
        type: {
            type: String,
            enum: ['Point'],
            required: true
        },
        coordinates: {
            type: [Number],
            required: true
        }
    }

});
```

3.3 Cleaning & updating the primary data – Stage 2

3.3.1 Classify the country of the user

As it has been already mentioned this study wanted to detect the nationalities of the visitors and for that reason, it used a technique that identifies the hometown of the visitor based on their profile on Flickr. However, some of the users do not provide the actual name of their country but instead, they add a nickname e.g., the Netherlands as Neverlands or some phrase. Also, as an alternative to their country's name, they add their city's name or different variations of the name of their country. To avoid those situations and to have a cleaner dataset this study needed to manage each condition differently (Figure 3.2 – stage 2).

The users that correctly declared their country were categorized easily. Regular expressions has been used and a list with country names to update the "geos" table with the actual name of the user's country. After that, the other steps were more complex. A second query made to the "geos" table to find which users provided a city instead of a country or provide an alternative country name. That step was both manual and automate because, first, the cities or the countries the users inserted in their profile must be written down in Table 3.2 and then they need to be updated. Finally, to deal with the users that did not give the actual name of their country, or they add as a country name the value null, or left the field empty, the "geos" table has been queried with certain conditions. If the field of the country of a user did not much with the country name inside the country list or it was null or empty, they categorized as "unknown country name".

Table 3.2

Input name	Updated with	Input name	Updated with	Input name	Updated with
Wien (Vienna), AT	Austria	United States	USA	Poey de Lescar	France
Ottawa	Canada	Berkeley, CA	USA	Bastia, Corsica	France
Copenhagen	Denmark	Las Vegas, U.S.A.	USA	Italia	Italy
Quarry Bay	Hong Kong	Concord, NC	USA	Alessandria	Italy
Seoul	Korea	Newton, Kansas	USA	Catania	Italy
México	Mexico	Port Wing, WI	USA	Venice	Italy
Tánger, Marruecos	Morocco	Somerville, MA	USA	Bavaria	Germany
Oslo, Norge	Norway	Daly City, CA, US	USA	Deutschland	Germany
Vers Chiang Mai, Thaïlande	Thailand	Sunnyvale, US	USA	Gütersloh	Germany
Tunis	Tunisia	chicago	USA	Berlin	Germany
Dubai, UAE	United Arab Emirates	Baltimore	USA	Leiden, Holland	Netherlands
Brasil	Brazil	Fremont, California	USA	Amsterdam, Nederland	Netherlands
Cheseaux, ШвейцарИя	Switzerland	US	USA	s-Hertogenbosch, Nederland	Netherlands
Rebstein, Rheintal	Switzerland	Los Angeles	USA	Rotterdam	Netherlands
Schweiz	Switzerland	St. Paul, MInnesota	USA	Piraeus, Grecia	Greece
Barcelona	Spain	New York City	USA	Thessaloniki	Greece
España	Spain	Denver, CO	USA	Athens	Greece
Tarifa	Spain	Brooklyn	USA	Ελλάδα	Greece
Weehawken, NJ	USA	Southampton, U.K.	United Kingdom	Athens, Hellas	Greece
Tarrytown, New York	USA	Birmingham	United Kingdom	Mons, Belgique	Belgium
NYC, Terra, Sol, Milky Way	USA	Southampton	United Kingdom	GENT, belgie	Belgium
Carmel, IN, US"	USA	London	United Kingdom	Brussels, Belgio	Belgium
Cincinnati, OH	USA	Croydon	United Kingdom	Brussels	Belgium
Washington, D.C.	USA	Wales	United Kingdom	Bruxelles	Belgium
Washington, DC	USA	Glasgow, Scotland	United Kingdom	Czech Republic	Czech Republic
Largo, FL, US	USA	Scotland	United Kingdom	prague, Czech republic	Czech Republic
San Francisco, CA, US	USA	England	United Kingdom	Paris	France
Oak Hill, VA, U.S.A.	USA	Grand Rapids, MI, US	USA	UK	United Kingdom
San Mateo, CA	USA	Lompoc	USA	Cambridge	United Kingdom

3.3.2 Remove the duplicate records

After the classification of the user's country name, the "geos" table was explored to see if duplicate and similar records exist (Figure 3.2 – stage 2). Many photos from the same user had the same coordinates at different times during the day and it is already known that it is physically impossible for a person to be to the same geolocation of a place more than one time after he/she departs.

To remove the duplicate records, the "aggregate" method was used with its functions to group the "geos" dataset more than once. The first query used the "$group" function to group the photos dataset by the user's id and then the second query grouped the already grouped users by the coordinates of the photos. With the "$addToSet" function only one record kept of each photo for the same user with the same coordinates. Moreover, reading the developer documentation of Flickr, it has been found out that Flickr has some big issues with time and sometimes it did not record the actual time a photo was captured. For that reason, the hour field was removed from the "geos" dataset using the same query. Finally, with the "$out" function a new table was exported with a clean dataset called "cleanedData" ready to be used for the

next step of the density-based clustering. The code used to remove the duplicates and the hour field from the "geos" dataset is given in Code-Snippet 3.5.

Code-Snippet 3.5: Duplicate records removal query

```
const duplicates = async (req, res) => {
    const results = await Geo.aggregate([
        {
            $group: {
                _id: "$properties.userID",
                country: {$first: "$properties.ownerLocation"},
                year: {$first: "$properties.yearTaken"},
                day: {$first: "$properties.dayTaken"},
                month: {$first: "$properties.monthTaken"},
                allLoc: {$addToSet: "$geometry.coordinates"}
            }
        },
        {$unwind: {path: "$allLoc", preserveNullAndEmptyArrays: true}},
        {
            $project: {
                "_id": 0,
                "lng": {$arrayElemAt: ["$allLoc", 0]},
                "lat": {$arrayElemAt: ["$allLoc", 1]},
                "userID": "$_id",
                "country": "$country",
                "year": "$year",
                "month": "$month",
                "day": "$day",
            }
        },
        {$out: "cleanedData"}
    ]);
```

3.4 Identify spatial clusters – Stage 3

3.4.1 The application of HDBSCAN

To find the AOI's that are formed in Athens by the geotagged photos of the visitors, the HDBSCAN method was applied to the cleaned dataset (Figure 3.2 – stage 3). To achieve that, the table "cleanedData" was exported as a CSV file from the database using MongoDB Compass[28] and imported into ArcGIS Pro. ArcGIS Pro provides its users with a Machine Learning Toolset that includes a density-based clustering tool. Within that tool, the intended dataset was selected and as the clustering method the Self-adjusting HDBSCAN method was applied. As it has been already mentioned HDBSCAN accepts only one parameter, the Minimum Features per Cluster. After many experiments, 450 features as minimum points per cluster were applied. From the experiments, was noticed that when the minimum points number is very high the number of the noisy points is increased while the number of clusters is decreased. On the other hand, while the minimum points number is low, the number of noisy points is decreased, and the number of clusters is increased. According to the amount of the data and their density, was decided that the 450 minimum points per cluster was the most reliable and accurate choice to produce clear clusters.

The HDBSCAN algorithm produced two types of points, the clustered points, and the noisy points. Although each cluster had a unique id and the number of its points stored in the attribute table, additional information such as the country of the user and the temporal attributes were missing. To be able to visualize better the produced AOI's and investigate the temporal distribution and the nationalities of the users for each area two more steps needed to be applied. The first step was to give the resulted clusters a shape that indicates their identity, and that

[28] *Mongo DB Compass is the graphical user interface (GUI) of MongDB which allows the user to interact visually with the data and import and export tables of data without coding. https://www.mongodb.com/products/compass*

shape is a circle. To convert the clustered points to circles the Minimum Bounding Geometry tool of ArcGIS Pro was used and the resulted shapefile was exported as a GeoJSON format with the name “clusters”. The second step was to give to each of the clustered points the missing information. To do that the spatial join tool was used which joined the attributes matching from the initial dataset to the clustered points. After that, the clustered points shapefile was exported with the extra information as a GeoJSON file with the name “statistics”.

3.5 Store the final datasets & create a public backend RESTful API - Stage 4

In the 4th stage, GeoJSON files, “clusters” and “statistics”, were imported in the database using the Compass GUI of MongoDB (Figure 3.2 – stage 4). Before importing them the schema for those datasets was designed and it is displayed in Code-Snippet 3.6 and Code-Snippet 3.7.

Code-Snippet 3.6: The schema of clusters in the database

```
const clusterSchema = mongoose.Schema({
    type: String,
    properties: {
        cluster_id: Number,
        frequency: Number,
        area: Number
    },
    geometry: {
        type: {
            type: String,
            enum: ['Point'],
            required: true
        },
        coordinates: {
            type: [Number],
            required: true
        }
    }

});
```

Code-Snippet 3.7: The schema of statistics in the database

```
const statisticSchema = mongoose.Schema({
    type: String,
    properties: {
        user_id: Number,
        cluster_id: Number,
        country: String,
        year: Number,
        month: Number,
    },
    geometry: {
        type: {
            type: String,
            enum: ['Point'],
            required: true
        },
        coordinates: {
            type: [Number],
            required: true
        }
    }
});
```

After the described procedures, the study`s database is complete and includes all the datasets needed to create the Web-GIS interface and visualize the results. The database consists of 5 tables: the “attractions” table, the “geos” table, which is the primary dataset, the “cleanedData”

table which contains the data after the cleaning process, the "clusters" table which is the dataset with the AOI's from the clustering procedure and the "statistics" table which contains additional information for each clustered point.

To make these datasets available and accessible through the web for the public and for the Web-GIS user interface a public backend RESTful API was created and deployed using Heroku[29]. As it has been already mentioned, REST is an architecture that is used broadly to design web applications. The fundamental criterion of this architecture is to use simple HTTP requests and not complicated mechanisms to achieve communication between two or more machines. The Internet itself is generally HTTP-based, so this can be thought of as a REST architecture for communicating with a client (say, a browser like a classic, everyday example). `*RESTful applications use HTTP requests to import - create - update (POST and PUT), read - access (GET), and delete data (DELETE). Therefore, REST uses all HTTP verbs for the basic CRUD (Create / Read / Update / Delete) procedures* (Elkstein 2008)`. Although the backend Restful API that created for this study is basically accepting only GET requests and responds in GeoJSON format, it can be further equipped with the rest of the HTTP methods (Post, Put, Delete).

The "attractions" and "clusters" datasets did not need any more process to be requested but the "statistics" dataset needed some queries to produce the desired information about each cluster. Each point inside the "statistics" dataset is grouped ($group function) by the cluster id and counted ($count function) to calculate the number of photographs each cluster contains. The second query grouped the dataset by cluster id and by year to investigate the number of photographs each cluster received each year from 2009 to 2019. The third query grouped the dataset again by cluster id and this time by month to detect the monthly distribution of the photos inside the clusters. Finally, the fourth query grouped the dataset by country to identify what kind of nationalities each cluster attracts. This procedure is long, and it is preferably not to provide a code snippet. However, the source code of the study is public and accessible through GitHub[30].

3.6 The design and structure of the Web-GIS user interface – Stage 5 (frontend)

Unlike other studies that use static maps, charts, and graphs to present their results, this research decided to create a single page Web-GIS application using modern technologies, to let the user interact with the results of this study (Figure 3.2 – stage 5). The user interface created with Angular 8 -a modern web development framework- HTML, CSS, Bootstrap4, a JavaScript library about charts called Chart.js, and the mapping library Leaflet.

As it is already mentioned in section 3.1.1 Angular helps the developer to create a complete and well-structured web application using an MVC pattern. The Angular MVC pattern as observed in Figure 3.3, consists of the View, Model, and Component (Controller). Models are common JavaScript objects, and they represent the part of the application which manages the retrieval and storage of the data. The View is what the user sees on the screen (HTML and

[29] *Heroku is a cloud platform that can be used as a server by developers. It is free for small scale applications, but it can be used by companies as it also provides an Enterprise version (https://www.heroku.com/home)*

[30] https://github.com/domikani/tourism-clusters-Athens-API

CSS). The Components (Controller) determine the interaction between Models and View. The 2-way data-binding means that there is a two-way communication between the Model and the View. Any data changes on the Model are reflected directly on the View and any changes on the View are shown automatically on the model. The Component is a fundamental element of Angular and an Angular project includes one or more components. A Component includes 3 files, the HTML (View) file, the CSS (View) file, and the ts (controller) file, and communicate with the View by inserting elements and tags in the Html code.

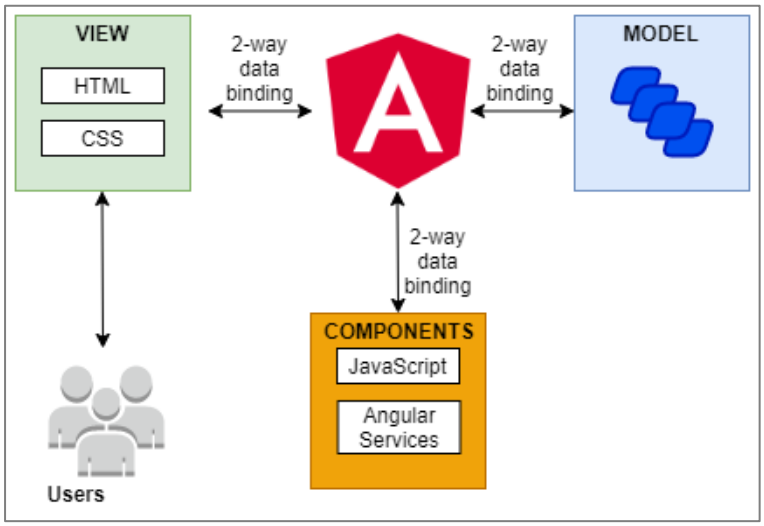


Figure 3.3: Angular MVC

The user interface consists of 4 parts, a navbar, a side menu on the left, a map on the center, and a sidebar on the right that includes the statistics section with graphs. For each of those parts, a component was created with its corresponding files (Html, CSS, ts) as well as three service files that help us to invoke functions stored in any of the components. The components and the services created were the following:

1. Map component
2. Menu component
3. Navbar component
4. Statistics component
5. Attractions service
6. Cluster service
7. Map service

The file structure of the aforementioned components and services are presented in Figure 3.4 and a sketch of the user interface is presented in Figure 3.5.

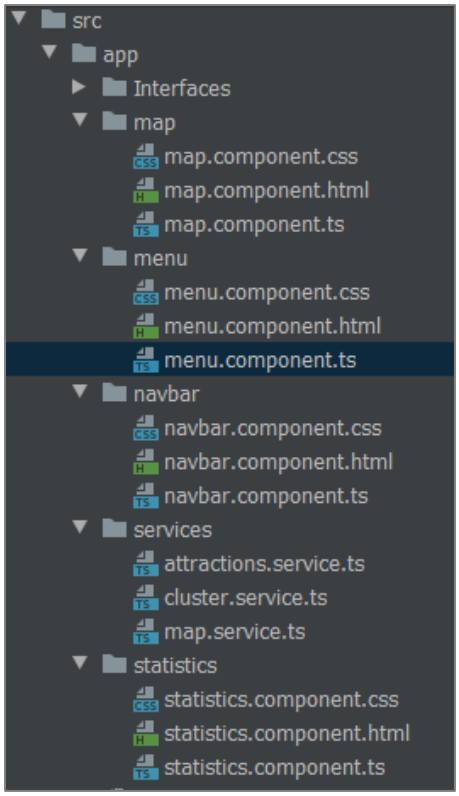


Figure 3.4: File structure of components and services of the Angular Web-GIS application

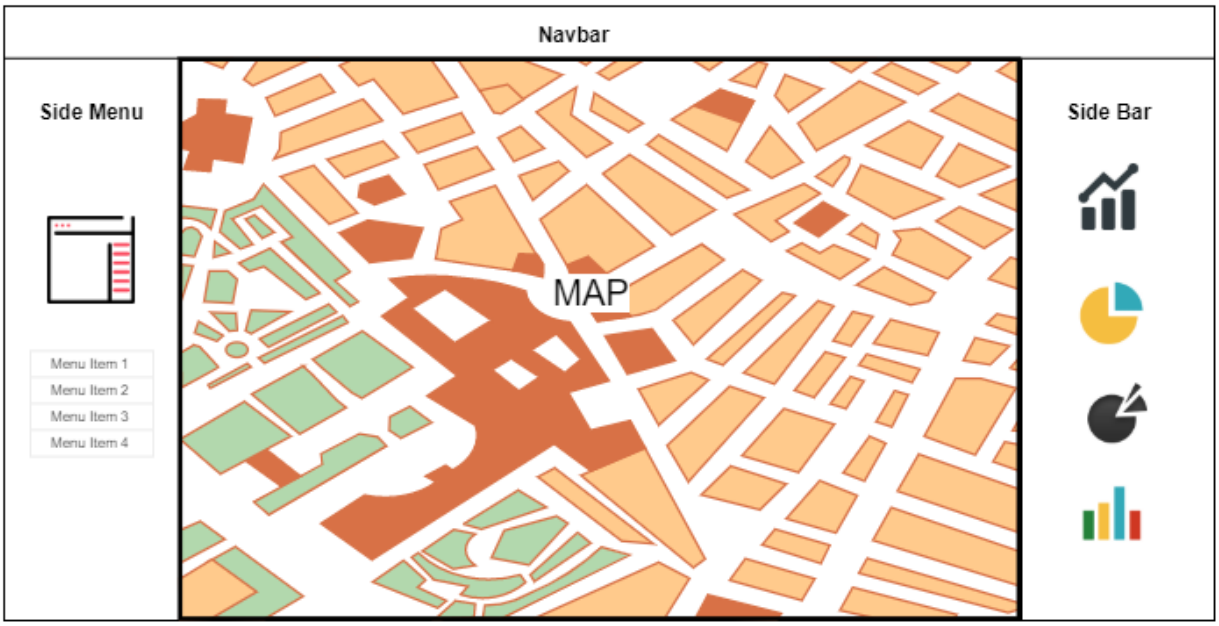


Figure 3.5: The Web-GIS user interface

The Web-GIS application is internally connected with the RESTfull API, and it has been deployed with Vercel[31]. The further explanation of the development of the Web-GIS application is not in the scope of this study, however, the source code of the user interface is public and accessible through GitHub[32]. Through the next chapter the results of the applied methodology are presented and the research questions 4,5 and 6 can be answered. Moreover, a presentation of the Web-GIS application is taking place, providing figures from the final product.

4 Results

4.1 Case study

This research aimed to reveal spatial and temporal patterns formed by visitors in Athens from 2009 to 2019, by harvesting and mining geotagged photos from Flickr social media platform. At this point, it is important to notice that visitors are considered both people from outside of Greece as well as from inside, and the results were constructed with geotagged photographs from both groups. Although this study identified the spatial clusters using both categories of foreigners and residents, it has been further investigated the nationalities of the visitors each AOI contains. Numerous techniques and methods were developed and applied to achieve the overall aim and the objectives of this study, and a novel approach was presented with a focus on the visualization of the results that promote the user interaction with the outcomes of this research.

4.2 Primary data collection

4.2.1 Geosocial data collection

The total number of geotagged photos (with their additional metadata), harvested from Flickr in eleven years period (2009-2019) for the area of Athens, was 157314. After collecting the primary dataset, the data was cleaned to remove several geotagged photos that have been captured and uploaded by the same user at the same coordinates at different times. The final dataset, after the cleaning process, contained 77,659 geotagged photos with their extra properties. A sample of a geotagged photo record stored inside the database is presented in Table 4.1, the sample has been chosen randomly.

Table 4.1: A sample of a geotagged photo record stored in the database after the cleaning process

Key	Value
_id (it is automatically created from MongoDB)	ObjectId(`6012a613b7c1445c0b52ef85`)
lng (longitude)	23.733299
lat (latitude)	37.983299
userID (the id of the user inside Flickr)	`71791926@N00`
country (the country name of the user after the classification)	`unknown country name`
year (the year the photo was taken)	`2009`
month (the month the photo was taken)	`05`
day (the day the photo was taken)	`01`

[31] *Vercel is a cloud platform for static sites and Serverless Functions. It is free and easy to deploy any front-end project (https://vercel.com/)*

[32] https://github.com/domikani/tourism-clusters-Athens-Client

4.2.2 Top attractions collection

In order to investigate if the resulted AOIs from the density-based clustering would merge with the most famous Points of Interests in Athens, the most popular attractions and landmarks in Athens were collected from two acclaimed tourism websites (Visit Greece and Lonely Planet) and a local tourism agency. Twenty-one POIs gathered and some of their characteristics were added manually. The collected POIs are displayed in Table 4.2 and Figure 4.1 (a, b).

Table 4.2: POIs

name	category	lng	lat	source	sourceName
Mount Lycabettus	Nature	23.74319	37.98194	https://travelgenius.gr/	Travel Genius
Asteras Vouliagmenis	Nature	23.77349	37.80939	https://travelgenius.gr/	Travel Genius
National Garden	Nature	23.73735	37.97369	https://travelgenius.gr/	Travel Genius
Kallimarmaro	Historic Stadium	23.74094	37.96859	https://travelgenius.gr/	Travel Genius
Museum of Illusions	Museum	23.72277	37.97687	https://travelgenius.gr/	Travel Genius
National Observatory of Athens	Observatory	23.71984	37.97332	https://travelgenius.gr/	Travel Genius
Eugenides Foundation-Planetarium	Cultural Centre	23.69658	37.94017	https://travelgenius.gr/	Travel Genius
Acropolis Museum	Museum	23.72896	37.96849	https://www.lonelyplanet.com/	Lonely Planet
Parthenon	Temple	23.72712	37.97151	https://www.lonelyplanet.com/	Lonely Planet
Acropolis	Historic Site	23.72638	37.97176	https://www.lonelyplanet.com/	Lonely Planet
Kerameikos	Historic Site	23.71823	37.97868	https://www.lonelyplanet.com/	Lonely Planet
Ancient Agora	Historic Site	23.72255	37.97525	https://www.lonelyplanet.com/	Lonely Planet
Byzantine & Christian Museum	Museum	23.74472	37.97493	https://www.lonelyplanet.com/	Lonely Planet
Benaki Museum of Greek Culture	Museum	23.74034	37.97609	https://www.lonelyplanet.com/	Lonely Planet
National Archaelogical Museum	Museum	23.73297	37.98906	https://www.lonelyplanet.com/	Lonely Planet
Temple of Olympian Zeus	Temple	23.73311	37.96948	https://www.lonelyplanet.com/	Lonely Planet
Stavros Niarchos Foundation Cultural Centre	Cultural Centre	23.6933	37.94127	https://www.lonelyplanet.com/	Lonely Planet
Syntagma Square	Square	23.73483	37.97559	http://www.visitgreece.gr/	Visit Greece
Philopappou Hill	Nature	23.72213	37.96762	http://www.visitgreece.gr/	Visit Greece
Odeon Of Herodes Atticus	Historic Site	23.72458	37.97089	http://www.visitgreece.gr/	Visit Greece
Naos Ifaistou	Temple	23.72144	37.97559	http://www.visitgreece.gr/	Visit Greece

Author processing[33]

[33] *The POIs provided by* https://travelgenius.gr/ , https://www.lonelyplanet.com/ , *and* http://www.visitgreece.gr/

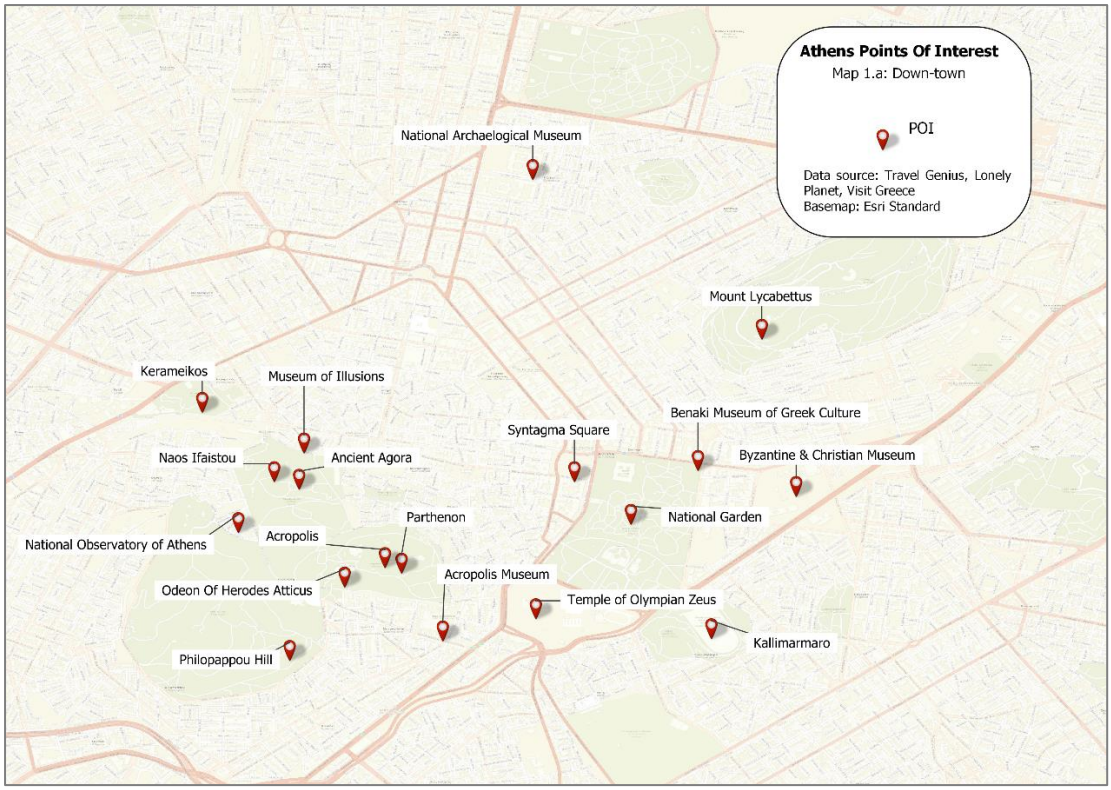
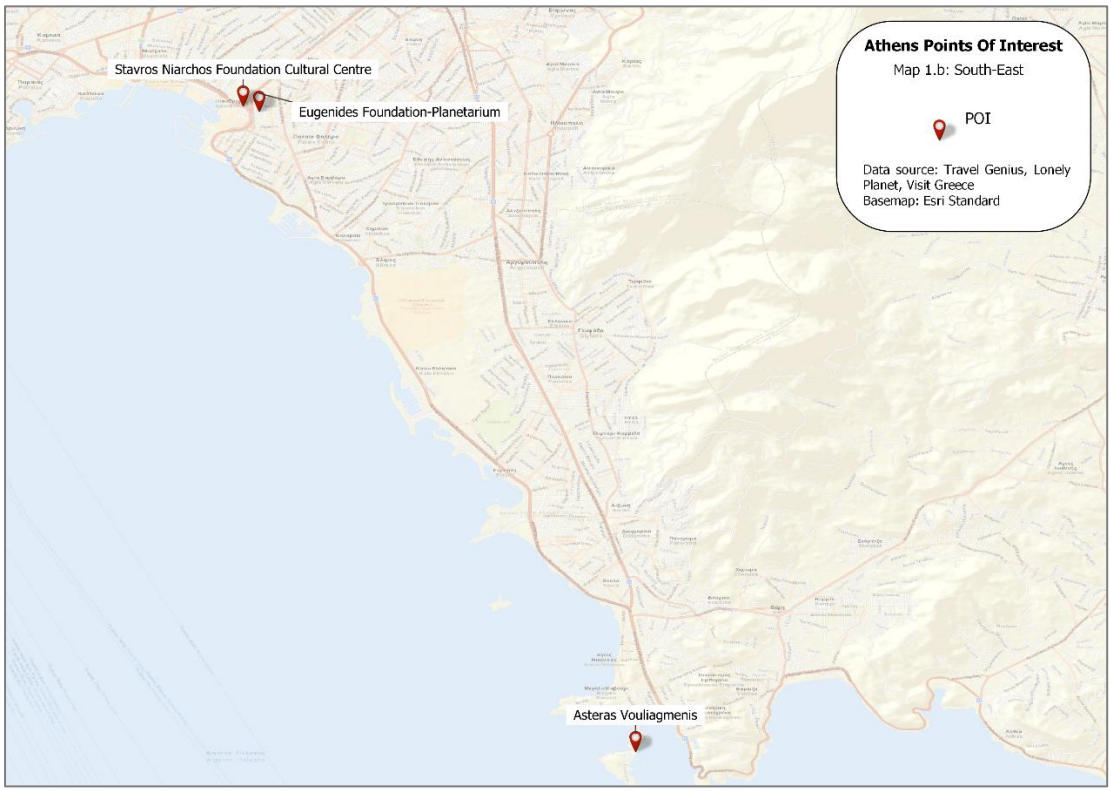


Figure 4.1: POIs map a) Athens - Down-town



b)Athens - South-East

4.3 Geosocial data mining

4.3.1 HDBSCAN application

To identify Areas of Interest by exploring big geosocial data and obtain information about the visitors that concentrate in these areas, an unsupervised clustering technique was used called density-based clustering. The density-based clustering technique provides 3 methods, as it has been already mentioned in section 3.1: the DBSCAN, the HDBSCAN, and the OPTICS method. Due to its efficiency, the HDBSCAN method was applied to the data and different values selected as minimum points per cluster (MinPtc) to explore which is the best value for the dataset. As it has been mentioned before, a high value of MinPtc will produce fewer clusters with higher density and a low value of MinPtc will produce more clusters with lower density. Based on the dataset which contained a small number of data (77,659 records) we experimented with different values declared as MinPtc. Each generated dataset from the HDBSCAN method includes an output feature class with a probability field which is the probability the feature belongs in its assigned group and an outlier field, designating the feature maybe an outlier within its own cluster (when the value is higher more the feature is more likely to be an outlier). After the experimentation, the dataset generated with the 450 MinPtc parameter was chosen because the majority of its points, in comparison with datasets produced with different values, presented very high probability to belong in their assigned cluster as it can be seen from Figure 4.2 and a very low probability to be outliers as it can be seen from Figure 4.3.

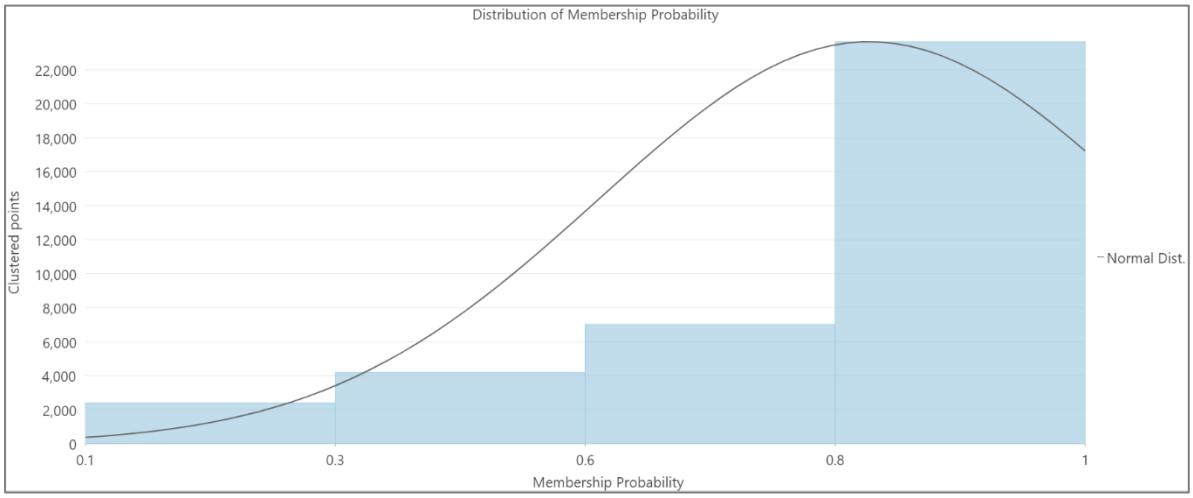


Figure 4.2: Distribution of Membership Probability

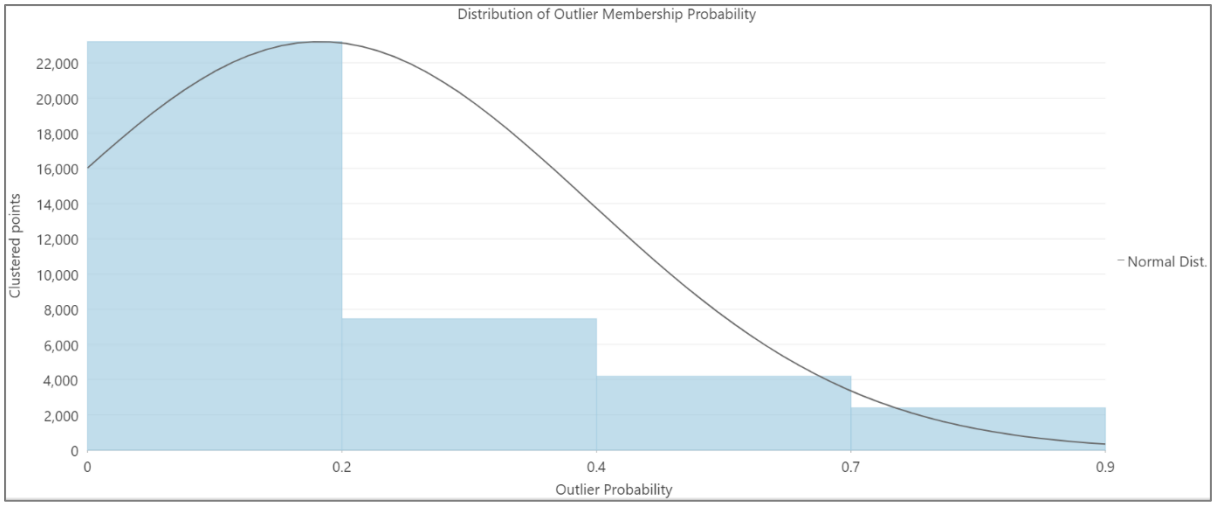


Figure 4.3: Distribution of outlier Membership Probability

4.3.2 Identified AOIs

The number of the extracted clusters was 20 and a name was given to them based on the area they were located (Table 4.3). The produced clusters were formed with points which is not the best way for a user to interpret them as Areas of Interest. To make them clearer to interpret through a Web-GIS interface, later on, the clustered points were converted into circles and symbolized using graduated colors based on the frequency of the points each cluster included. The final visualization of the AOIs will be presented in section 4.5, however, a primary sample of the clusters converted into AOIs is given created with ArcGIS pro (Figure 4.4, Figure 4.5).

Table 4.3: Identified AOIs

Cluster Id	Frequency of geotagged photos	AOI name
1	657	Vouliagmeni
2	856	Glyfada
3	2475	Chalandri-Marousi
4	1345	Flisvos-Stavros Niarchos
5	4006	Piraeus
6	972	Lykabettus
7	1912	Exarcheia
8	1056	Kallimarmaro
9	541	Gkazi-Kerameikos
10	851	Akadimia Athinon
11	3725	Syntagma
12	1663	Styloi Olympiou Dios
13	482	Mitropoli Athinon
14	992	Naos Ifaistou
15	1315	Acropolis Museum
16	10778	Acropolis
17	643	Areopagus Hill
18	624	Romaiki Agora
19	1661	Monastiraki
20	694	Archaia Agora

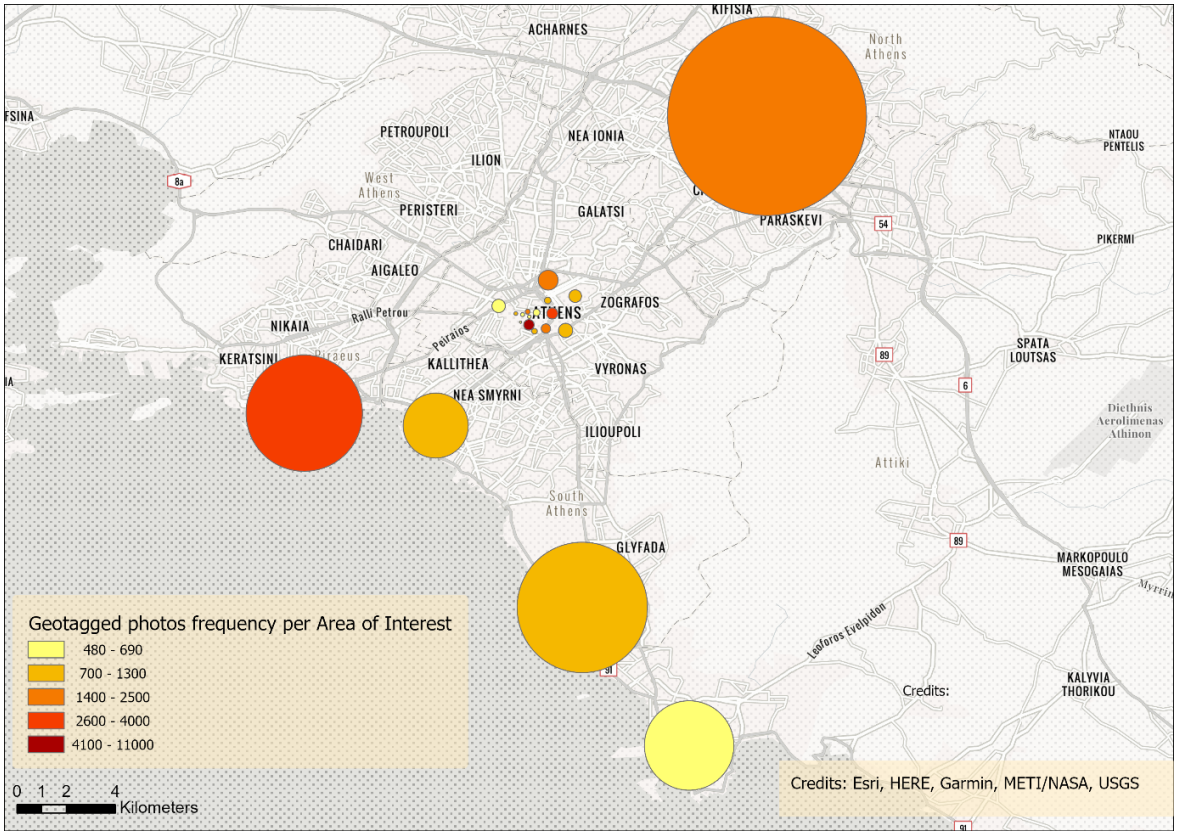


Figure 4.4: Map of identified AOIs

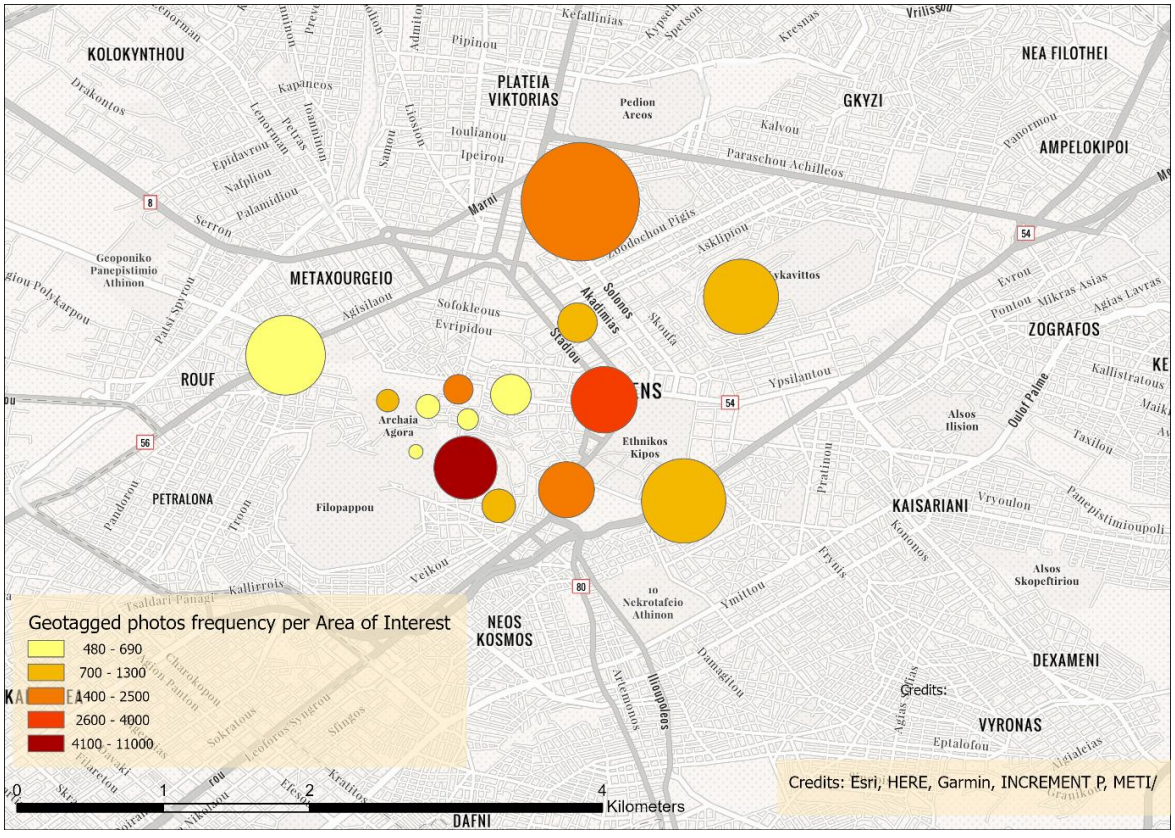


Figure 4.5: Map of Zoomed-in AOIs in the center of Athens

4.3.2.1 Correspondence assessment between the produced AOIs and the POIs

To validate the agreement of the produced AOIs from the HDBSCAN analysis with real-world data, this study explored if they meet with existing and popular landmarks of Athens that were already collected and referred to them as POIs (Table 4.2). By investigating the association between the resulted AOIs and POIS, it has been found out that 60% of the AOIs (12 out of 20 clusters) have corresponding POIs, 40%(8 out of 20 clusters) of the AOIs did not match with any of the existing POIs and only 28.5% (6 out of 21 POIs) of the landmarks were located outside of the clustered zones. The fact that the majority of Athens's most famous attractions matched with the produced clusters, means that the HDBSCAN algorithm successfully recognized the most visited areas in Athens, and in addition, it produced 8 new areas of interest. Table 4.4 shows the POIs that correspond to the identified AOIs, the cluster ID of the AOIs and the number of geotagged photos each AOI contains.

Table 4.4: Produced AOIs with corresponding real-world data (POIs)

Produced AOI	**Cluster ID**	**Number of geotagged photos**	**POIs detected inside AOIs**
Vouliagmeni	1	657	Asteras Vouliagmenis
Glyfada	2	856	None
Flisvos-Stavros Niarchos	4	1345	Stavros Niarchos Foundation Cultural Centre
			Eugenides Foundation-Planetarium
Piraeus	5	4006	None
Kallimarmaro	8	1056	Kallimarmaro
Styloi Olympiou Dios	12	1663	Temple of Olympian Zeus
Acropolis Museum	15	1315	Acropolis Museum
Acropolis	16	10778	Parthenon
			Acropolis
			Odeon of Herodes Atticus
Areopagus Hill	17	643	None
Romaiki Agora	18	624	None
Archaia Agora	20	694	Ancient Agora
Monastiraki	19	1661	None
Mitropoli Athinon	13	482	None
Naos Ifaistou	14	992	Naos Ifaistou
Gkazi-Kerameikos	9	541	Archaeological Site of Kerameikos
Syntagma	11	3725	Syntagma Square
Akadimia Athinon	10	851	None
Lykabettus	6	972	Mount Lykabettus
Exarcheia	7	1912	National Archaeological Museum
Chalandri-Marousi	3	2475	None

4.4 Distributions

4.4.1 Temporal distributions

To exploit the additional metadata of the geotagged photos concentrated in the AOIs, the month and the year the photo was taken were used to explore the temporal patterns of each area separately. Furthermore, since there were available data about the monthly frequency of the geotagged photos, the seasonality of the study area was measured using the Gini Coefficient of inequality index. Gini Coefficient index is a higher quality mean to measure the seasonality of an area. It can take on values between 0 and 1, with lower numbers representing a greater degree of equality and thus less seasonality (Koenig and Bischoff 2002). The results of the temporal analysis are presented in section 4.4.1.1 and section 4.4.1.2 among with the results of the Gini Coefficient index. Exploring further the monthly and yearly distribution results, the following observations can be drawn individually for the AOIs:

1. Vouliagmeni: Vouliagmeni is a luxury resort town, built very close to the sea, it is located 17 kilometers south of the center of Athens and it is mostly known for its clean beaches, for the expensive real-estate market, and the seafood restaurants. Monthly this area shows a significant concentration of visitors in January, March, and April and the least concentration in December. Yearly, there is a peak in visitation rate in 2017 and a sudden decrease in 2018 and 2019.
2. Glyfada: Glyfada is one of the largest and oldest seaside suburbs of Athens. It is located next to Vouliagmeni near Mount Ymittos. Glyfada presents the largest visitors' concentration in February, March, and April and the smallest in December while 2009 is the year with the maximum visitors.
3. Chalandri-Marousi: Chalandri and Marousi are two famous neighborhoods in Athens with a strong local market, many cafes, popular cocktail bars, and modern restaurants. This AOI presents a considerable concentration of visitors in October, and in a year time scale in 2013 and 2010.
4. Stavros Niarchos: The Stavros Niarchos Foundation Cultural Centre (SNFCC) is a *`public space, where everyone has free access and can participate in a multitude of cultural, educational, athletic, environmental, and recreational activities and events*[34]. In this AOI all months gather a satisfying number of visitors except July. Yearly, 2010 presents the greatest number of visitors, with 2009, 2013, and 2014 following with relatively same numbers.
5. Piraeus: Piraeus is the biggest port of Greece connecting Athens with the Aegean islands. It is also a very lively area famous for its architecture and the seafood taverns. Piraeus presents high levels of visitation in all months with a sharp increase in March. On a yearly basis, 2009 and 2014 are the most visited years.
6. Lykabettus: *`Lykabettus Hill is the highest point of Athens, which is known for a nice view of the Acropolis, the Temple of Olympian Zeus, Panathenaic Stadium, and the Ancient Agora`* (Katsoni and Segarra-Oña 2019, p.415). This

[34] https://www.snfcc.org/en/snfcc/meet-the-snfcc

AOI presents a considerably high number of visitors in May and July while in 2013 and 2017 there is a growth in visitation rate.

7. Exarcheia: Exarcheia is an artistic, politicized, and eccentric neighborhood of Athens with alternative cafes, stores, bars, and restaurants. It is very famous among artists and students. Its monthly visitation graph illustrates a rise in visitors' concentration in September and March and its yearly distribution graph, the visits fluctuate through the eleven years.
8. Kallimarmaro: `*Kallimarmaro is the old Olympic Stadium of Athens. It is the only stadium in the world built entirely of marble. The first Olympic Games in modern history were held there (1896)*` (Katsoni and Segarra-Oña 2019, p.415). Although Kallimarmaro shows the largest concentration of visitors in March, the rest of the months, except February and January, present also high numbers of visitors. The year 2018 presents an elevated rise of visits but all the years concentrate a pleasing number of visits.
9. Gkazi-Kerameikos: Gkazi is a very famous urban area of Athens. It is one of the most visited areas by night since it gathers a majority of nightclubs, bars, and restaurants. The monthly graph of this AOI displays a significant rise in visits in March and July and the yearly graph illustrates a great concentration of visitors in 2009.
10. Akadimia Athinon: The Academy of Athens is an intellectual institution that aims to cultivate and promote Sciences and Fine Arts, as well as scientific research and study. Visitors tend to gather in this AOI in July while 2010 is the year with the most photographs.
11. Syntagma: `*Syntagma is the central square of Athens. It is located in front of the Old Royal Palace which is housing the Greek Parliament*`(Katsoni and Segarra-Oña 2019, p.415). Syntagma square shows a large concentration of visitors in October and May and 2010 and 2009 are the years with the most visits.
12. Styloi Olympiou Dios: The `Styloi Olympiou Dios` was one of the largest historical sites `*in antiquity and close to Hadrian's Arch, which forms the symbolic entrance to the city*`(Katsoni and Segarra-Oña 2019, p.415). The analysis of the temporal data of this AOI showed that the majority of people visit this area in September and May while in the year time scale analysis there was a fluctuation in the number of visitors.
13. Mitropoli Athinon: The `Mitropoli Athinon` is the Orthodox Cathedral of Athens, and its construction began in 1842 and was completed in 1862. Monthly, July presents a significant number of visitors compared to the other months while 2013 gathered the highest visitation rate.
14. Naos Ifaistou: The `Naos Ifaistou` is one of the most well-preserved ancient temples in Greece and is located in the area of Thission. The temple of Hephaestus concentrates the greatest number of visitors in October and in July, while visits went up and down from 2009 to 2019.
15. Acropolis Museum: The Museum of Acropolis is an archaeological museum dedicated to the findings of the archaeological site of Acropolis. In this AOI

September and May concentrate the most visits while 2012 and 2011 present the highest numbers of photos, respectively.

16. Acropolis: *`Acropolis is the site of some of the most important masterpieces of worldwide architecture and art, the most renowned of which is the Parthenon temple`*(Katsoni and Segarra-Oña 2019, p.414). The site of Acropolis presents the highest number of visitors in October; however, September, August, July, June, and May show a considerable number of visitors too. In 2011 and 2013 Acropolis accepted the most visitors.
17. Areopagus Hill: Areopagus Hill is a historic site that served as the highest court in ancient Greece. The majority of the people prefer to visit this site in May and August while it presents a high visitation rate in 2017 and 2015.
18. Romaiki Agora: The Romaiki Agora is an archeological site located in the center of Athens. From the analysis of the geotagged photos of this AOI, the most visitors are distributed in August, May, and September. The years 2010, 2017, and 2018 present the most visits.
19. Monastiraki: Monastiraki is a famous area that is surrounded by ancient landmarks, lively cafes, taverns, and local stores. The most visitors on a monthly scale are presented in May and August and on a yearly scale in 2009.
20. Archaia Agora (Ancient Agora): The Ancient Agora is an open space located near the Acropolis. In ancient times it was mainly the economic center of the city. This AOI shows a significantly increased number of visitors in August while in 2009 visits reached a peak.

4.4.1.1 Yearly distribution results

From Table 4.5 it can be seen that the AOIs present a bigger concentration of geotagged photos in early years from 2009 to 2014 with small exceptions such as `Vouliagmeni`, with the highest concentration detected in 2017, `Exarcheia` and `Kallimarmaro` AOI which presented high visitation rates in 2018, and `Areopagus` and `Romaiki Agora` AOIs with the greatest number of geotagged photos identified in 2017. To examine better the yearly distribution of the geotagged photos in each AOI, trend analysis has been used to calculate the percentage change for the geotagged photos over a period of time. As it can be seen in Table 4.5 the concentration of the geotagged photos presented a decrease throughout the studied years in the majority of the AOIs and only the AOIs `Mitropoli Athinon` and `Archaia Agora` presented a slight increase of 5% and 26% respectively.

Table 4.5: Yearly distribution of geotagged photos

Yearly distribution of geotagged photos found in AOIs													
Cluster Id	**AOI name**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	Percentage change
1	Vouliagmeni	120	41	35	47	83	50	14	9	228	8	22	**-82%**
2	Glyfada	303	52	65	144	63	107	37	32	24	12	17	**-94%**
3	Chalandri-Marousi	273	611	145	142	890	253	65	33	28	29	6	**-98%**
4	Flisvos-Stavros Niarchos	209	231	60	111	191	213	67	80	76	46	61	**-71%**
5	Piraeus	589	263	368	456	418	607	259	243	381	247	175	**-70%**
6	Lykabettus	65	104	80	67	145	100	114	83	139	43	32	**-51%**
7	Exarcheia	193	287	123	84	299	124	235	96	136	248	87	**-55%**
8	Kallimarmaro	110	113	95	95	110	107	95	69	60	117	85	**-23%**
9	Gkazi-Kerameikos	130	64	44	54	90	63	18	20	16	12	30	**-77%**
10	Akadimia Athinon	126	134	82	76	71	75	91	43	58	57	38	**-70%**
11	Syntagma	588	655	451	410	293	278	307	186	224	200	133	**-77%**
12	Styloi Olympiou Dios	139	161	189	116	172	164	209	129	168	144	72	**-48%**
13	Mitropoli Athinon	43	43	26	38	82	34	21	45	50	55	45	**5%**
14	Naos Ifaistou	101	110	101	68	97	91	109	54	102	69	90	**-11%**
15	Acropolis Museum	143	122	143	157	134	109	92	107	87	119	102	**-29%**
16	Acropolis	1001	922	1619	941	1390	762	922	772	923	877	649	**-35%**
17	Areopagus Hill	46	64	47	58	55	26	95	51	100	72	29	**-37%**
18	Romaiki Agora	60	70	44	54	54	65	50	46	70	69	42	**-30%**
19	Monastiraki	253	162	114	182	151	161	131	81	148	181	97	**-62%**
20	Archaia Agora	68	99	65	54	52	59	68	30	74	39	86	**26%**

4.4.1.2 Monthly distribution results and Seasonality index

The overall trend that can be observed concerning the monthly distribution of geotagged photos, is that May and October have gathered the highest number of photos in the majority of the AOIs as well as March, September, July, August, June, and April presented big concentrations accordingly. Although the colder months such as January, February, November, and December appeared to be the months with the lowest numbers of geotagged photos in most of the AOIs, `Vouliagmeni` and `Glyfada` presented the biggest concentration in geotagged photos in January and February, respectively.

Further analysis has been made, using the monthly dataset of the AOIs to investigate the phenomenon of seasonality in Tourism. To explore the seasonality of the AOIs the Gini Coefficient Index was used. `*The Gini Coefficient (GC) is a numerical measure of the degree of inequality in the number of visitors across the months of the year. It is derived from the Lorenz curve which displays the cumulative frequency of the ranked observations starting with the lowest number. The Gini Coefficient is equal to the area between the Lorenz curve and the 45-degree line divided by the whole area below the line. GC can take on values between 0 and 1, with lower numbers representing a greater degree of equality and thus less seasonality*` (Koenig and Bischoff 2002, p.9) The equation of the Gini Coefficient can be displayed as follows:

Equation 4.1: GC equation

$$GC = 1 + \left(\frac{1}{n}\right) - \left(\frac{2}{(n^2 \cdot y_0)}\right) \cdot (y_1 + 2y_2 + 3y_3 + \ldots + ny_n)$$

Koenig and Bischoff 2002 (Koenig and Bischoff 2002, p.9)

`with:

n = number of observations (i.e., 12 in the case of monthly data)

$_0$*y = mean of observations (i.e., the average number of trips)*

$_n$*y y y y* $_{1, 2, 3,\ldots,}$ *= individual observations in decreasing order of magnitude`* (Koenig and Bischoff 2002, p.9)

Though Table 4.6 it can be observed that the AOIs `Vouliagmeni`, `Mitropoli Athinon`, `Glyfada`, `Gkazi-Kerameikos`, `Flisvos-Stavros Niarchos`, `Exarcheia`, and `Chalandri-Marousi` present medium to high seasonality while the rest of the AOIs present medium to low seasonality. In a general trend it can be pointed out that the study area present medium levels of seasonality with the AOIs that are mostly archaeological sites to present high seasonality values and equally concentration of tourists all the year.

Table 4.6: Geotagged photos GC seasonality index

Seasonality index of geotagged photos											
AOI	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Acropolis	0.3	0.2	0.5	0.4	0.5	0.4	0.3	0.3	0.3	0.4	0.3
Acropolis Museum	0.4	0.3	0.5	0.4	0.4	0.3	0.2	0.4	0.5	0.6	0.3
Akadimia Athinon	0.3	0.4	0.3	0.5	0.5	0.3	0.6	0.5	0.6	0.6	0.5
Archaia Agora	0.3	0.5	0.4	0.5	0.4	0.6	0.5	0.4	0.6	0.5	0.6
Areopagus Hill	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.5	0.4	0.5	0.7
Chalandri-Marousi	0.6	0.6	0.4	0.5	0.8	0.6	0.5	0.4	0.7	0.8	0.6
Exarcheia	0.5	0.5	0.3	0.4	0.7	0.6	0.5	0.5	0.7	0.6	0.4
Flisvos-Stavros Niarchos	0.6	0.7	0.4	0.4	0.4	0.6	0.5	0.6	0.5	0.5	0.7
Gkazi-Kerameikos	0.5	0.3	0.4	0.4	0.7	0.6	0.6	0.8	0.7	0.5	0.8
Glyfada	0.7	0.7	0.5	0.6	0.5	0.5	0.5	0.7	0.6	0.7	0.7
Kallimarmaro	0.5	0.4	0.2	0.4	0.5	0.5	0.4	0.3	0.6	0.5	0.5
Lykabettus	0.4	0.5	0.4	0.5	0.5	0.4	0.6	0.4	0.6	0.5	0.4
Mitropoli Athinon	0.5	0.4	0.5	0.5	0.7	0.6	0.6	0.4	0.6	0.6	0.7
Monastiraki	0.3	0.2	0.3	0.4	0.3	0.2	0.3	0.4	0.5	0.5	0.5
Naos Ifaistou	0.3	0.3	0.6	0.4	0.6	0.5	0.4	0.5	0.6	0.4	0.4
Piraeus	0.4	0.3	0.6	0.4	0.4	0.4	0.3	0.4	0.5	0.5	0.6
Romaiki Agora	0.3	0.4	0.3	0.5	0.3	0.5	0.5	0.4	0.6	0.5	0.6
Styloi Olympiou Dios	0.4	0.4	0.4	0.4	0.4	0.4	0.5	0.3	0.5	0.6	0.5
Syntagma	0.3	0.5	0.4	0.4	0.3	0.3	0.5	0.3	0.4	0.5	0.4
Vouliagmeni	0.7	0.6	0.6	0.4	0.7	0.4	0.5	0.8	0.8	0.8	0.7

4.4.2 Visitors' distribution

To understand better the distribution of the visitors in the study area the country of origin of the users was explored and the concentration of different nationalities was examined in the AOIs. Because there were too many countries to summarize in a graph, the countries of the users per continent were classified. From Figure 4.6 it can be observed that two groups stand out, those groups are the users classified in the cleaning process with `Unknown nationality` and users from Greece. However, comparing the rest of the groups we can see that the greater part of visitors is from Europe, North America, and Asia.

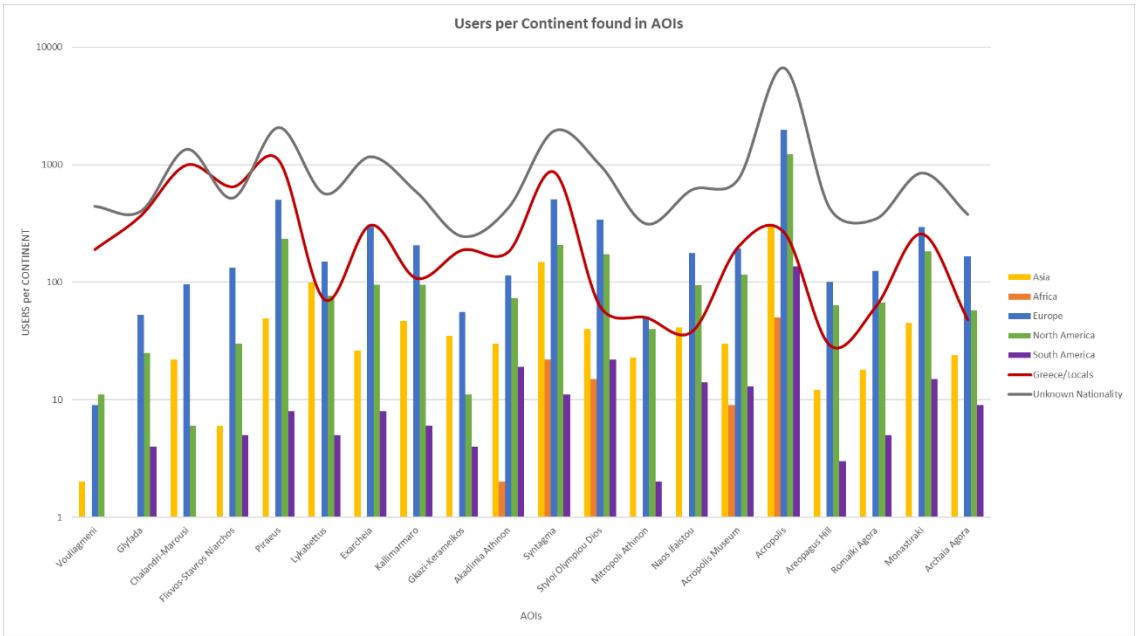


Figure 4.6: Visitors distribution per continent

4.5 Web-GIS interface

To visualize its findings, this study created a Web-GIS interface using modern web technologies and the Leaflet mapping library. The Web-GIS interface is hosted online, and it is available for anyone through the URL https://tourism-athens.vercel.app/. The interface of the Web-GIS application homepage is shown in Figure 4.7.

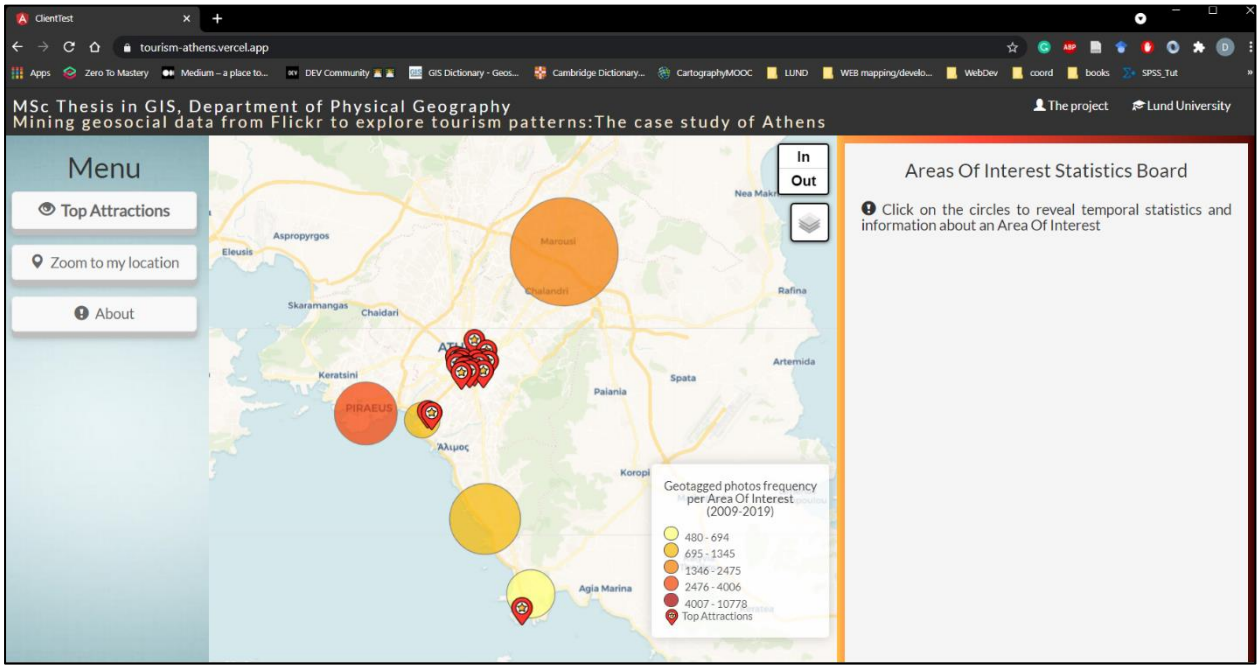


Figure 4.7: Web-GIS interface homepage

As it can be seen the interface consists of 4 parts:

1. The navbar (Figure 4.8) holds the title of this study and two buttons. The first button navigates the user to the source code of this project and the second button navigates the user to the webpage of the Department of Physical Geography of Lund University.

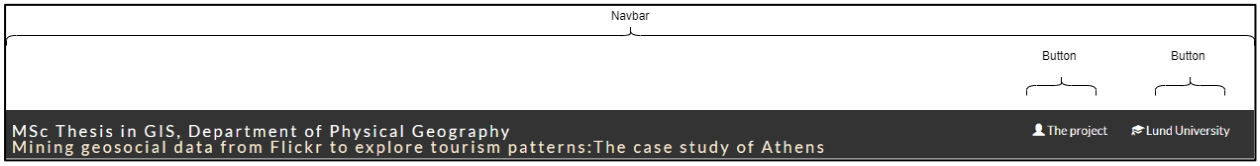


Figure 4.8: Navbar

2. The `Menu` (Figure 4.9) left sidebar includes the `Top Attractions button`, the `Zoom to my location` button, and the `About` button. When the user clicks the Top Attractions button a card-list with all the top attractions and landmarks of Athens is revealed (Figure 4.10). This card list includes photos of the attractions that were gathered with some additional information. Those cards are also clickable and navigate the user to a specific attraction on the map. The `Zoom to my location` button identifies the location of the user and the `About` button gives the user information about the project (Figure 4.11).

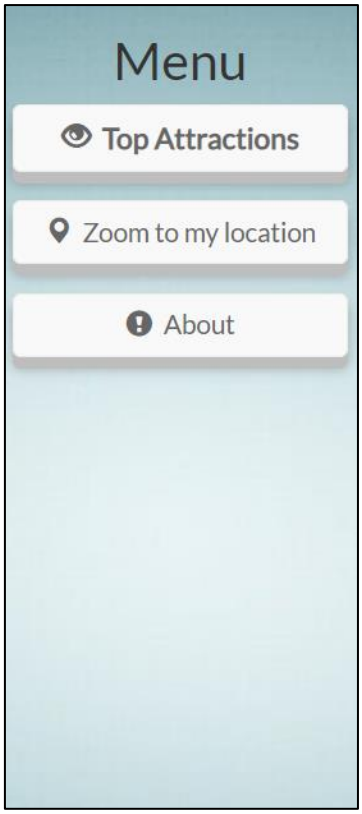


Figure 4.9: Menu sidebar

Figure 4.10: : `Top Attractions` button functionality

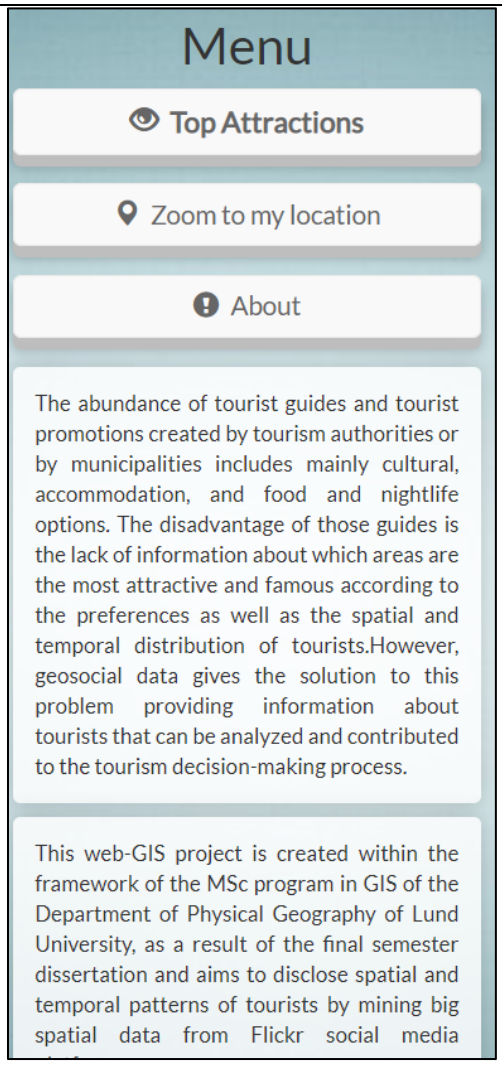


Figure 4.11: : `About` button functionality

3. At the middle part of the interface lays a map presenting the identified AOIs as circles and the top attractions as pins (Figure 4.12, Figure 4.13). Both the AOIs and the top attractions pins are clickable, and the user can interact with them and receive information. If a user clicks on an AOI, a pop-up is showed giving information about the AOI and provide a button that reveals temporal statistics about the AOI in the right-sidebar of the interface (Figure 4.14). The red pins represent the POIs of Athens that have gathered from tourism websites and a local tourism agency. The user can interact also with the pins hovering the mouse over them (Figure 4.15).
4. The last part of the interface is the right-sidebar, the `Statistics Board`, which contains statics about the AOI. When a user clicks on an AOI and clicks the button `Show Statistics`, 3 graphs are drawn at the right-sidebar (Figure 4.16, Figure 4.17). One graph about the yearly distribution of the visitors in the AOI, a second graph about the monthly distribution of the visitors in the AOI, and a third graph about the distribution of the nationalities of the visitors identified in the AOI.

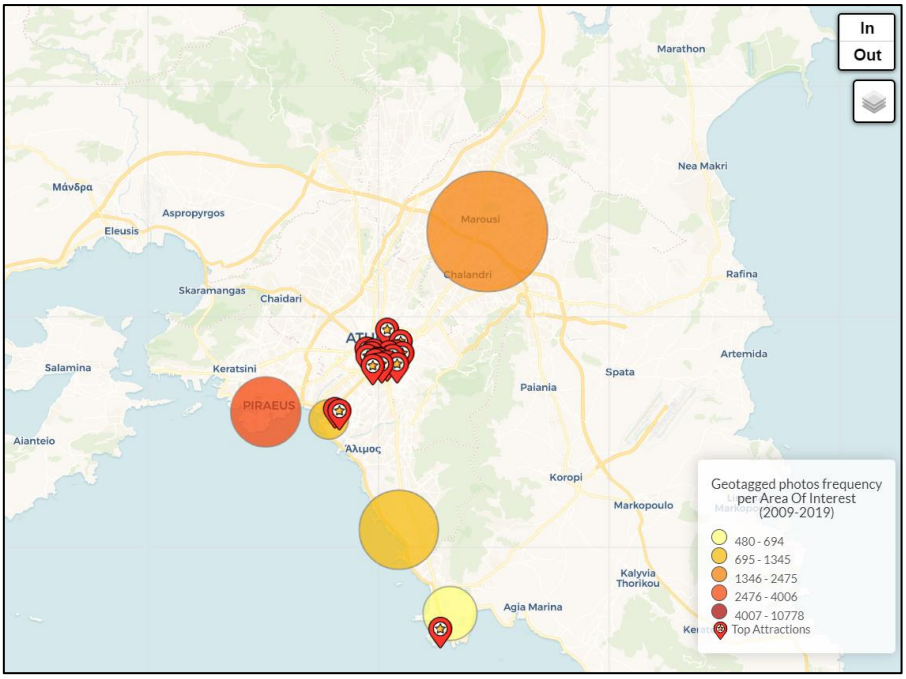


Figure 4.12: The Map

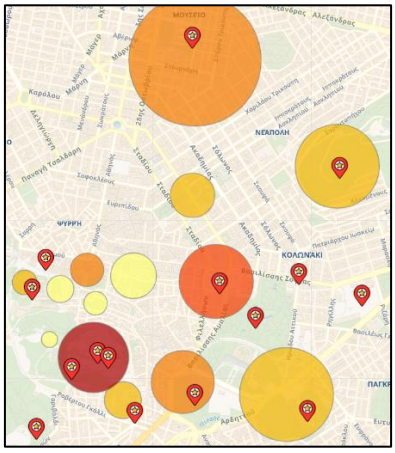


Figure 4.13: Zoom in AOIs & POIs

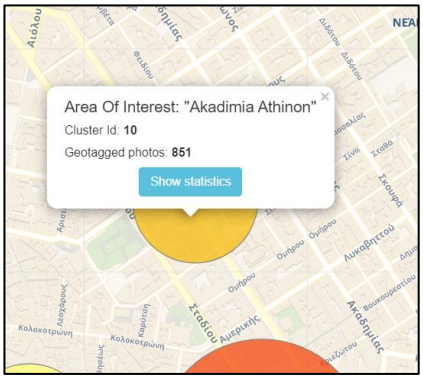


Figure 4.14: User clicking on an AOI

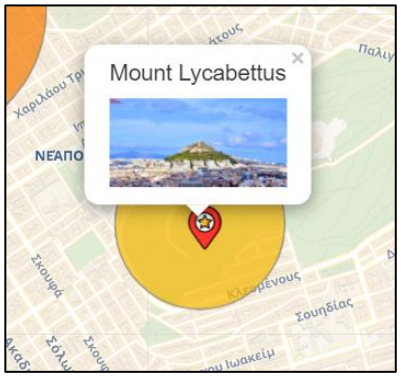


Figure 4.15: User hovering on a Top Attraction pin

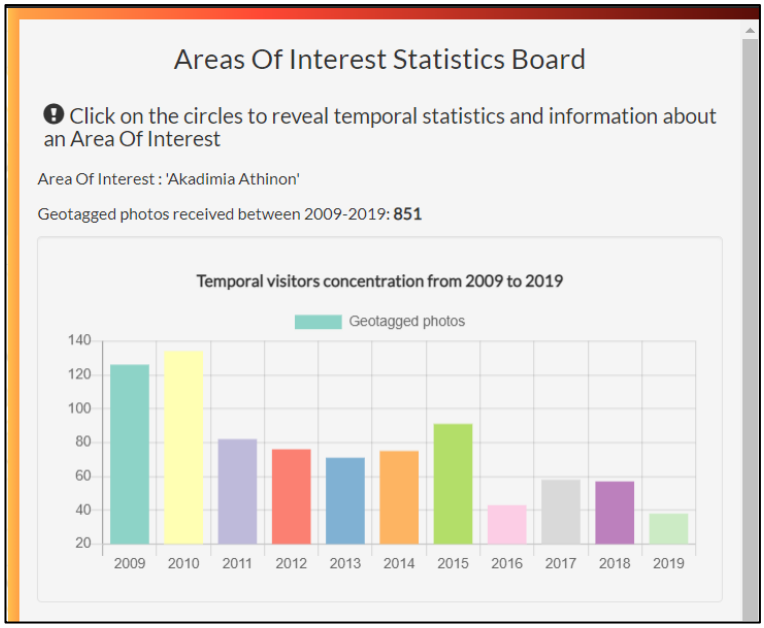


Figure 4.16: Statistics board

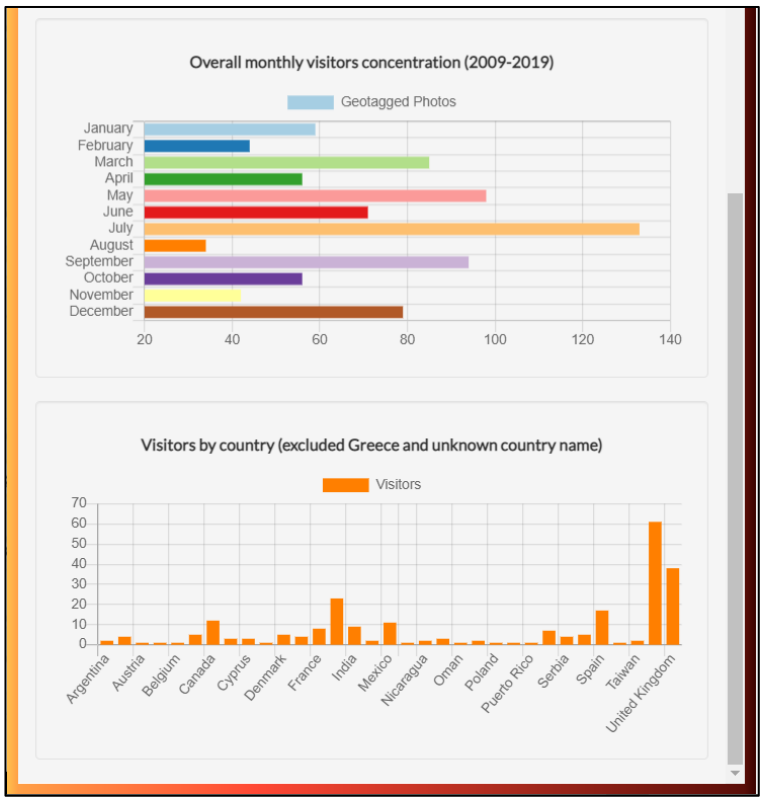


Figure 4.17: Statistics board

5 Discussion

The Flickr social media platform

To accomplish the overall aim and the sub-objectives of this study the Flickr social media platform was used to retrieve geotagged photos. Although there are additional photo-sharing social media platforms that provide geosocial data such as Twitter, Facebook, and Instagram, they have some limitations. Twitter as the most famous microblogging social media platform, is extensively used for online event detection (Farnaghi et al. 2020) and it is more suitable for detecting events that people tweeting about and not from a touristic perspective. Moreover, there are some restrictions on what someone can analyze using Twitter, which played a decisive role in not choosing it for this research. First, and most importantly, Twitter does not provide data from the past. This study, needed to retrieve data from 2009 until 2019 which was impossible to do using Twitter, and second, it provides around 1% of its data for free. Concerning Facebook and Instagram, they have changed their privacy rules and many permissions are needed to use their APIs. From June 29, 2020, Instagram disabled the `Basic Permission` to the Instagram Legacy API and third-party apps no longer have access to the legacy API[35]. Both Facebook[36] and Instagram require an App Review process to provide their data to anyone interested, and that process was not in the scope of this study. Nevertheless, those social media platforms share also one main drawback: When a user uploads a geotagged photo in those platforms, instead of the initial coordinates of where the photo was taken to be stored in the database, the coordinates from where the user uploaded the photo are preserved and stored (Leung et al. 2017). Having that in mind, this research is aware that geotagged photos from Instagram, Facebook, and Twitter might not hold the actual location of where the photo was taken. However, Flickr is widely used in the literature for harvesting geotagged photos to analyze tourism patterns and maintain the primary coordinates of where the photo was taken, and these are the main reasons this study has chosen Flickr to collect the dataset for this study.

The flaws of the collected dataset

Using Flickr, 157,314 geotagged photographs were collected between 2009 and 2019 for the center of Athens, which is a decent but not a great amount of data. Exploring further the dataset some errors were detected. Flickr sometimes uploads multiple photos from the same user with the same coordinates at different times which is not realistic. Some studies (García-Palomares et al. 2015; Salas-Olmedo et al. 2018; Koutras et al. 2019) identified that bug and removed the multiple photos uploaded from the same user at the same location but only when one photo is taken one minute later than the other. This study decided not to follow this method but instead to keep only one photo from the same user at a specific location. Although this technique tends to be more accurate, more data was lost in comparison with the aforementioned studies. Moreover, this study applied a technique proposed by Li et al. (2018) and modified it, to find the country of origin of each user. Compared to associated studies (Önder et al. 2014; García-Palomares et al. 2015; Koutras et al. 2019; Höpken et al. 2020) that categorized the users into

[35] https://www.instagram.com/developer/ (Instagram Developer Documentation. [no date])

[36] https://developers.facebook.com/docs/permissions/reference (Permissions Reference - Graph API - Documentation. [no date])

two groups, locals and tourists, using a technique based on the timestamps of the photos, and exclude the locals from their dataset, this study achieved to acquire extra information that those studies lack about the nationalities of the visitors based on information the users register on their profile and not to lose data excluding users that might be locals. Moreover, Flickr uses an automate way to add an accuracy values to the photos holding a geographic location. Flickr adds this value according to the zoom level of the map when the photo was geotagged. Those values range from (a)World-level accuracy as value 1, (b) Country-level accuracy as value 3, (c) Region level accuracy as value 6, (d) City-level accuracy as value 11 and (e) Street-level accuracy as value 16. Although this study used only photographs with a value equal to 16, it is considered that some of the photographs may lack high level of spatial accuracy.

Validation of spatial clusters – AOIs

The HDBSCAN algorithm successfully handles and clusters the geotagged photos that were retrieved from Flickr, into 20 clusters that represent the most visited areas in Athens. The areas that were identified from the algorithm are popular areas of Athens and by exploring the map through the Web-GIS application it can be seen that the majority of the AOIs include many of the top attractions and landmarks of Athens that has been collected through tourism authorities. However, some of the identified areas do not include any of the collected top attractions. This phenomenon can be justified by exploring further the nature and the location of those areas.

The AOIs `Areopagus Hill` and `Romaiki Agora` are very close to where Acropolis is located and thus, they do not be advertised as much as Acropolis by tourism authorities. This is happening because it is inevitable for someone to visit Acropolis and neglect to visit those areas. In some ways, the promotion of Acropolis is covering those areas too. One possible reason why the `Mitropoli Athinon` was identified as an AOI, without including any POIs, is that many religious events take place there since the Mitropoli is the most known cathedral in Greece, and many photographs taken there are concern weddings and the Easter and Christmas services. The `Akadimia Athinon` AOI consists of the majority of the universities of Athens and does not include any POI reasonably. It is probably identified as an area of interest due to the big number of photographs uploaded by students and not by visitors. This AOI is not important from a tourist perspective, but it can be from an education perspective. The rest of the AOIs found without merging with a POI such as `Monastiraki`, `Chalandri-Marousi`,` Piraeus` and `Glyfada` are famous areas in Athens among visitors, both locals, and foreigners, for shopping, nightlife, food, and recreation but they are lack attention from tourism consultants. Of course, except for all these reasons that some AOIs did not blend with POIs, one additional cause can be the fact that this study did not gather all the POIs that exist in Athens but the top 21 provided by tourism experts.

Apart from AOIs that did not contain any POIs there were found some POIs in areas where the HDBSCAN algorithm did not detect any cluster. Those POIs were the Museum of Illusions, the Byzantine and Christian Museum, the Benaki Museum of Greek Culture, the National Garden, the National Observatory of Athens, and Philopappou Hill. Three of the POIs found in areas without clusters are museums which is very logical since not many photos can be taken inside a museum. The other three POIs, although they are considered to be in the top 21

attractions of Athens, it is believed that did not gather a satisfactory number of photos to be recognized as areas of interest due to a bad marketing campaign.

HDBSCAN accomplished to identify almost 3 times more clusters with different densities, compared to the study of Koutras et al. (2019) that identified only 7 clusters using the DBSCAN algorithm in the same study area as ours, proving that the declarations of Chen et al. (2019) that HDBSCAN is more efficient than DBSCAN are worth to explore more. At this point, this study also want to remark that the fact of the number of POIs that was collected through the primary stage of data collection was almost being the same as the produced AOIs is a coincidence.

Statistics

Expanding the research of Koutras et al. (2019) in exploring the temporal statistics of the AOIs this research further analyzed the yearly distribution of geotagged photos, the seasonality of the AOIs and tried to identify the distribution of the nationalities of the visitors in the AOIs.

Concerning the yearly distribution of geotagged photos, by investigating Table 4.5, it can be said that in general the years from 2009 to 2013 presented the highest geotagged photos concentration between 2009 to 2019. Comparing those results with the results of the Institute of the Association of Greek Tourist Enterprises (INSETE 2020) (

Figure 1.1) it is obvious that they are not matching. The official statistical report shows a stable increase during the years 2016 to 2019 in tourist arrivals, and the results for the yearly distribution of the geotagged photos present the exact opposite assumption. It is believed that this happened because Flickr has lost its popularity from 2016 and onwards and many users have shown their preference on Instagram[37][38].

Regarding the monthly distribution of geotagged photos, this study noticed that the months with the most geotagged photos in the study area were mostly the months of Spring, Summer, and Autumn except November, December, and January. However, the months from November until February gathered a satisfactory number of geotagged photos. The AOIs that presented high seasonality have some characteristics that explain their high values of the Gini Coefficient index. To be more specific the `Vouliagmeni`, `Glyfada`, and `Flisvos-Stavros Niarchos` AOIs are located near to the sea and visitors tend to visit those areas mostly on summer months. The `Exarcheia`, `Chalandri-Marousi` and `Gkazi-Kerameikos` concentrate the majority of the most popular cocktail bars of Athens which justify their seasonality. The AOIs that presented medium to low levels of seasonality were areas of archaeological interest and monuments which means that visitors tends to visit such areas all year around.

Finally, concerning the nationalities distribution of the visitors, this study did not achieve to present the actual existing condition. Although it has been found that Europeans, North Americans, and Asians tend to visit more the identified AOIs the number of the users classified

[37] https://ferdychristant.com/the-rise-fall-and-resurrection-of-flickr-ca1850410ee1 (Christant 2020)

[38] https://www.investopedia.com/articles/markets/082015/why-instagram-winning-over-flickr.asp (Tarver 2020)

with `Unknown nationality` were too many that it is sure that this study have lost valuable information during the process. A more holistic approach should certainly be applied and combined with the study`s technique to provide more accurate and rich information on the preferences of the users based on their nationality. Acquiring knowledge like that, tourism consultants, authorities, and tourism agencies could provide an individual tourism plan to each of the visitors.

Future Recommendations

This research contributed to the literature by applying methods that were less discussed in related studies and should be considered for more investigation and provide a reusable source code that with some adjustments can be used by other researchers and developers for related studies. Although the advantages of this study are satisfactory, future work must be made to improve its findings. First of all, the method applied to classify the countries of the users should be further explored and modified to recognize the countries that were not declared correct and categorized as `Unknown country`. Secondly, more geosocial data should be harvested from other photo-sharing social media platforms -if it is possible due to privacy issues- to produce more accurate temporal statistics and their correspondence with the official statistics reports should be reviewed. Finally, the Web-GIS application can be improved to offer more interactions to the user such as the possibility to evaluate and rate each of the AOIs by creating a personal account and even to be able to add a new area that he/she thinks is of interest to other visitors, providing comments and photographs.

The methodology and the prototype Web-GIS created in this study can be used as a primary tool by tourism authorities and agencies for decision making, to explore which areas tend to concentrate a great number of visitors and regulate their flow, and to promote new areas of interest according to their popularity derived from analyzing real-time social media data. From visitors' perspective, the Web-GIS interface can be a useful tool to explore which areas are the most popular in Athens based on other visitors' preferences and plan their trip accordingly.

Conclusion

This research aimed to identify spatial and temporal patterns of tourists by mining geotagged photos from Flickr's social media platform. The aim was divided into 4 sub-objectives and 6 relevant research questions related to the sub-objectives applied to accomplish the overall aim.

Concerning ***RQ1*** the methods and techniques found in the literature review to mining geosocial data and detect spatial clusters is the application of clustering techniques such as DBSCAN, HDBSCAN and k-means. Regarding ***RQ2*** the methods selected to be implemented in this study were:

- the Flickr API to harvest the data
- the MongoDB database to clean and store the data
- the HDBSCAN algorithm to analyze the data and produce spatial clusters
- web-mapping techniques to visualize the results

The key findings resulted from ***RQ3*** is that HDBSCAN algorithm successfully identified 20 spatial clusters of different densities. The 60% of those clusters included corresponding POIs and the clustering analysis revealed 8 new AOIs in the study area.

The answers to the questions ***RQ4*** are that the phenomenon of seasonality is indeed detected in the AOIs, the yearly visitation rate presented a gradual decrease trend from 2009 to 2019 and the highest numbers of geotagged photos detected the years from 2009 to 2014. Furthermore, the main discoveries resulted from ***RQ6*** are that the majority of the visitors detected in the AOIs were Europeans, North Americans, and Asians.

However, the results on temporal analysis and the visitors` nationalities distribution that shape the answers to these questions present a lack of reliability based on two reasons: first, it is believed that the temporal data acquired from Flickr for the years 2016 to 2019 were not accurate because Flickr lost many users during the recent years with the appearance of Instagram, and second, concerning the nationalities distribution, the majority of the users did not declare their true country of origin and as a result many of those data disqualified.

A spatiotemporal analysis of geosocial data and a Web-GIS interface as a tool for visitors and tourism authorities are provided through this research. The geosocial data used for this study were harvested from Flickr social media platforms and the primary data acquired was 157,314. The final dataset used for the spatiotemporal analysis after the cleaning process was 77,659 geotagged photos concerning the study area of the city of Athens while an alternative method of those proposed by related studies was applied to classify the countries of the visitors. The HDBSCAN clustering algorithm effectively produced 20 clusters as AOIs since the majority of them corresponding to already popular areas of Athens including some of the most known top attractions and landmarks of Athens. In addition, the clustering analysis achieved to produce 8 new clusters that were not promoted by tourism authorities. Moreover, the clustering analysis revealed the temporal and the seasonal distributions of the geotagged photos in each AOI as well as which nationalities are more interested to visit Athens.

References

Agarwal, S. et al. 2011. Building Rome in a day. *Communications of the ACM* 54(10), pp. 105–112. doi: 10.1145/2001269.2001293.

Ashby, D. 2018. APIs For Dummies®, 3rd IBM Limited Edition., p. 51.

Berba, P. 2020. Understanding HDBSCAN and Density-Based Clustering. Available at: https://towardsdatascience.com/understanding-hdbscan-and-density-based-clustering-121dbee1320e [Accessed: 22 November 2020].

Bhadane, C. and Shah, K. 2020. Clustering Algorithms for Spatial Data Mining. In: *Proceedings of the 2020 3rd International Conference on Geoinformatics and Data Analysis*. ICGDA 2020. New York, NY, USA: Association for Computing Machinery, pp. 5–9. Available at: https://doi.org/10.1145/3397056.3397068 [Accessed: 24 February 2021].

Britannica 2020. Twitter | History, Description, & Uses. Available at: https://www.britannica.com/topic/Twitter [Accessed: 17 November 2020].

Campello, R.J.G.B. et al. 2013. Density-based clustering based on hierarchical density estimates. In: Pei, J. et al. eds. *Advances in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer, pp. 160–172. Available at: http://dx.doi.org/10.1007/978-3-642-37456-2_14 [Accessed: 22 November 2020].

Chen, M. et al. 2019. Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems* 21(1), pp. 89–109. doi: 10.1007/s10109-018-0284-3.

Christant, F. 2020. The rise, fall and resurrection of Flickr. Available at: https://ferdychristant.com/the-rise-fall-and-resurrection-of-flickr-ca1850410ee1 [Accessed: 24 March 2021].

Croitoru, A. et al. 2014. Geoinformatics and Social Media: New Big Data Challenge. In: *Big Data*. CRC Press, pp. 207–232. Available at: http://www.crcnetbase.com/doi/abs/10.1201/b16524-12 [Accessed: 22 October 2020].

DB-Engines Ranking. 2020. Available at: https://db-engines.com/en/ranking [Accessed: 23 November 2020].

DBSCAN Clustering in ML | Density based clustering. 2019. *GeeksforGeeks* 6 May. Available at: https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/ [Accessed: 22 November 2020].

Elkstein, D.M. 2008. Learn REST: A Tutorial. Available at: http://rest.elkstein.org/2008/02/what-is-rest.html [Accessed: 1 April 2021].

Ester, M. et al. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise., p. 6.

Farnaghi, M. et al. 2020. Dynamic Spatio-Temporal Tweet Mining for Event Detection: A Case Study of Hurricane Florence. *International Journal of Disaster Risk Science* 11(3), pp. 378–393. doi: 10.1007/s13753-020-00280-z.

García-Palomares, J.C. et al. 2015. Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography* 63, pp. 408–417. doi: 10.1016/j.apgeog.2015.08.002.

GeoServer. 2020. Available at: http://geoserver.org/ [Accessed: 23 November 2020].

Goodchild, M.F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4), pp. 211–221. doi: 10.1007/s10708-007-9111-y.

Google Developers 2018. Clustering in Machine Learning. Available at: https://developers.google.com/machine-learning/clustering [Accessed: 20 November 2020].

Gundersen, E. and Mapbox 2013. Visualizing 3 Billion Tweets. Available at: https://blog.mapbox.com/visualizing-3-billion-tweets-f6fc2aea03b0 [Accessed: 17 November 2020].

Halim, M.A. et al. 2018. Discovering New Tourist Attractions Through Social Media Data: A Case Study in Sabah Malaysia. In: *2018 IEEE 8th International Conference on System Engineering and Technology (ICSET)*. Bandung: IEEE, pp. 157–161. Available at: https://ieeexplore.ieee.org/document/8606373/ [Accessed: 29 May 2020].

Hall, M. 2020. Facebook | Overview, History, & Facts. Available at: https://www.britannica.com/topic/Facebook [Accessed: 17 November 2020].

Han, J. et al. 2012. *Data Mining: Concepts and Techniques*. Elsevier Inc. Available at: https://experts.illinois.edu/en/publications/data-mining-concepts-and-techniques-2 [Accessed: 22 November 2020].

Hoefer, R. et al. 1994. Geographic Information Systems and Human Services. *Journal of Community Practice* 1(3), pp. 113–128. doi: 10.1300/J125v01n03_08.

Höpken, W. et al. 2020. Flickr data for analysing tourists' spatial behaviour and movement patterns: A comparison of clustering techniques. *Journal of Hospitality and Tourism Technology* 11(1), pp. 69–82. doi: 10.1108/JHTT-08-2017-0059.

Hu, Y. et al. 2015. Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems* 54, pp. 240–254. doi: 10.1016/j.compenvurbsys.2015.09.001.

Ikkos, A. and Koutsos, S. 2020. INSETE, The contribution of Tourism in Greek Economy in 2019., p. 20.

INSETE 2015. The contribution of Tourism in Greek Economy in 2014., p. 17.

INSETE 2017a. The contribution of Tourism in Greek Economy in 2015., p. 19.

INSETE 2017b. The contribution of Tourism in Greek Economy in 2017., p. 20.

INSETE 2018a. The contribution of Tourism in Greek Economy in 2016., p. 19.

INSETE 2018b. The contribution of Tourism in Greek Economy in 2018., p. 20.

INSETE 2020. Regional Statistics | Attica Region. *Insete* April. Available at: https://insete.gr/statistika-stoixeia-perifereion/ [Accessed: 22 October 2020].

Instagram Developer Documentation. [no date]. Available at: https://www.instagram.com/developer/ [Accessed: 23 March 2021].

Jain, A.K. et al. 1999. Data clustering: a review. *ACM Computing Surveys* 31(3), pp. 264–323. doi: 10.1145/331499.331504.

Kaplan, A.M. and Haenlein, M. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons* 53(1), pp. 59–68. doi: 10.1016/j.bushor.2009.09.003.

Katsoni, V. and Segarra-Oña, M. eds. 2019. *Smart Tourism as a Driver for Culture and Sustainability: Fifth International Conference IACuDiT, Athens 2018*. Cham: Springer International Publishing. Available at: http://link.springer.com/10.1007/978-3-030-03910-3 [Accessed: 8 April 2020].

Kisilevich, S. et al. 2010. Event-Based Analysis of People's Activities and Behavior Using Flickr and Panoramio Geotagged Photo Collections. In: *2010 14th International Conference Information Visualisation*., pp. 289–296. doi: 10.1109/IV.2010.94.

Koenig, N. and Bischoff, E.E. 2002. Seasonality of Tourism in Wales - A Comparative Analysis., p. 33.

Koutras, A. et al. 2019. Towards Developing Smart Cities: Evidence from GIS Analysis on Tourists' Behavior Using Social Network Data in the City of Athens. In: Katsoni, V. and Segarra-Oña, M. eds. *Smart Tourism as a Driver for Culture and Sustainability*. Springer Proceedings in Business and Economics. Cham: Springer International Publishing, pp. 407–418. doi: 10.1007/978-3-030-03910-3_28.

Lee, I. et al. 2014. Exploration of geo-tagged photos through data mining approaches. *Expert Systems with Applications* 41(2), pp. 397–405. doi: 10.1016/j.eswa.2013.07.065.

Leung, R. et al. 2017. Understanding tourists' photo sharing and visit pattern at non-first tier attractions via geotagged photos. *Information Technology & Tourism* 17(1), pp. 55–74. doi: 10.1007/s40558-017-0078-3.

Li, D. et al. 2018. Analyzing and visualizing the spatial interactions between tourists and locals: A Flickr study in ten US cities. *Cities* 74, pp. 249–258. doi: 10.1016/j.cities.2017.12.012.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2), pp. 129–137. doi: 10.1109/TIT.1982.1056489.

Loizos, C. 2019. Flickr owner SmugMug emails subscribers with an urgent request: help us find more paying users. *TechCrunch* 20 December. Available at: https://social.techcrunch.com/2019/12/19/flickr-owner-smugmug-emails-subscribers-with-an-urgent-request-help-us-find-more-paying-users/ [Accessed: 16 November 2020].

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. The Regents of the University of California. Available at: https://projecteuclid.org/euclid.bsmsp/1200512992 [Accessed: 22 November 2020].

Makris, A. et al. 2019. *Performance Evaluation of MongoDB and PostgreSQL for spatio-temporal data*.

MarketLine Industry Profile: Travel & Tourism in Greece. 2020. *Travel & Tourism Industry Profile: Greece* , pp. 1–52.

Ministry of Tourism, G.N.T.O. 2014. *Athens | Attica guide*. 12th ed. Athens: Bibliosynergatiki S.A.

Muñoz, L. et al. 2020. Using crowdsourced spatial data from Flickr vs. PPGIS for understanding nature's contribution to people in Southern Norway. *People and Nature* n/a(n/a). Available at: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1002/pan3.10083 [Accessed: 27 May 2020].

Önder, I. et al. 2014. Tracing Tourists by Their Digital Footprints: The Case of Austria. *Journal of Travel Research* . Available at: https://journals-sagepub-com.ludwig.lub.lu.se/doi/10.1177/0047287514563985 [Accessed: 27 May 2020].

Pegkas, P. 2020. Interrelationships between tourism, energy, environment and economic growth in Greece. *Anatolia* 0(0), pp. 1–12. doi: 10.1080/13032917.2020.1795893.

Permissions Reference - Graph API - Documentation. [no date]. Available at: https://developers.facebook.com/docs/permissions/reference/ [Accessed: 23 March 2021].

Promise.all() - JavaScript | MDN. [no date]. Available at: https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global_Objects/Promise/all [Accessed: 10 March 2021].

Salas-Olmedo, M.H. et al. 2018. Tourists' digital footprint in cities: Comparing Big Data sources. *Tourism Management* 66, pp. 13–25. doi: 10.1016/j.tourman.2017.11.001.

Schubert, E. et al. 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems* 42(3), pp. 1–21. doi: 10.1145/3068335.

SETE 2009. Statistical data of Greek Tourism 2009. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2010. Statistical data of Greek Tourism 2010. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2011. Statistical data of Greek Tourism 2011. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2012. Statistical data of Greek Tourism 2012. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2013. Statistical data of Greek Tourism 2013. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2014. Statistical data of Greek Tourism 2014. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2015. Statistical data of Greek Tourism 2015. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2016. Statistical data of Greek Tourism 2016. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2017. Statistical data of Greek Tourism 2017. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2018. Statistical data of Greek Tourism 2018. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

SETE 2019. Statistical data of Greek Tourism 2019. Available at: https://sete.gr/el/stratigiki-gia-ton-tourismo/vasika-megethi-tou-ellinikoy-tourismoy/ [Accessed: 7 October 2020].

Spyrou, E. et al. 2015. Mining tourist routes from Flickr photos. In: *2015 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*., pp. 1–5. doi: 10.1109/SMAP.2015.7370093.

Stefanidis, A. et al. 2013. Harvesting ambient geospatial information from social media feeds. *GeoJournal* 78(2), pp. 319–338.

Tan, P.-N. et al. 2006. Introduction to Data Mining., p. 169.

Tarver, E. 2020. Why Instagram Is Winning Over Flickr. Available at: https://www.investopedia.com/articles/markets/082015/why-instagram-winning-over-flickr.asp [Accessed: 24 March 2021].

The PostgreSQL Global Development Group 2020. PostgreSQL 12.2 Documentation., p. 2698.

Thompson, J. 2019. Choosing the Right Clustering Algorithm for your Dataset. *KDnuggets* October. Available at: https://www.kdnuggets.com/choosing-the-right-clustering-algorithm-for-your-dataset.html/ [Accessed: 22 November 2020].

What Is MongoDB? 2020. Available at: https://www.mongodb.com/what-is-mongodb [Accessed: 23 November 2020].

What is REST? [no date]. Available at: https://www.codecademy.com/articles/what-is-rest [Accessed: 9 March 2021].

Series from Lund University

Department of Physical Geography and Ecosystem Science

Master Thesis in Geographical Information Science

1. *Anthony Lawther:* The application of GIS-based binary logistic regression for slope failure susceptibility mapping in the Western Grampian Mountains, Scotland (2008).
2. *Rickard Hansen:* Daily mobility in Grenoble Metropolitan Region, France. Applied GIS methods in time geographical research (2008).
3. *Emil Bayramov:* Environmental monitoring of bio-restoration activities using GIS and Remote Sensing (2009).
4. *Rafael Villarreal Pacheco:* Applications of Geographic Information Systems as an analytical and visualization tool for mass real estate valuation: a case study of Fontibon District, Bogota, Columbia (2009).
5. *Siri Oestreich Waage:* a case study of route solving for oversized transport: The use of GIS functionalities in transport of transformers, as part of maintaining a reliable power infrastructure (2010).
6. *Edgar Pimiento:* Shallow landslide susceptibility – Modelling and validation (2010).
7. *Martina Schäfer:* Near real-time mapping of floodwater mosquito breeding sites using aerial photographs (2010).
8. *August Pieter van Waarden-Nagel:* Land use evaluation to assess the outcome of the programme of rehabilitation measures for the river Rhine in the Netherlands (2010).
9. *Samira Muhammad:* Development and implementation of air quality data mart for Ontario, Canada: A case study of air quality in Ontario using OLAP tool. (2010).
10. *Fredros Oketch Okumu*: Using remotely sensed data to explore spatial and temporal relationships between photosynthetic productivity of vegetation and malaria transmission intensities in selected parts of Africa (2011).
11. *Svajunas Plunge:* Advanced decision support methods for solving diffuse water pollution problems (2011).
12. *Jonathan Higgins:* Monitoring urban growth in greater Lagos: A case study using GIS to monitor the urban growth of Lagos 1990 - 2008 and produce future growth prospects for the city (2011).
13. *Mårten Karlberg:* Mobile Map Client API: Design and Implementation for Android (2011).
14. *Jeanette McBride:* Mapping Chicago area urban tree canopy using color infrared imagery (2011).
15. *Andrew Farina:* Exploring the relationship between land surface temperature and vegetation abundance for urban heat island mitigation in Seville, Spain (2011).
16. *David Kanyari*: Nairobi City Journey Planner: An online and a Mobile Application (2011).

17. *Laura V. Drews:* Multi-criteria GIS analysis for siting of small wind power plants - A case study from Berlin (2012).
18. *Qaisar Nadeem:* Best living neighborhood in the city - A GIS based multi criteria evaluation of ArRiyadh City (2012).
19. *Ahmed Mohamed El Saeid Mustafa:* Development of a photo voltaic building rooftop integration analysis tool for GIS for Dokki District, Cairo, Egypt (2012).
20. *Daniel Patrick Taylor*: Eastern Oyster Aquaculture: Estuarine Remediation via Site Suitability and Spatially Explicit Carrying Capacity Modeling in Virginia's Chesapeake Bay (2013).
21. *Angeleta Oveta Wilson:* A Participatory GIS approach to *unearthing* Manchester's Cultural Heritage '*gold mine'* (2013).
22. *Ola Svensson:* Visibility and Tholos Tombs in the Messenian Landscape: A Comparative Case Study of the Pylian Hinterlands and the Soulima Valley (2013).
23. *Monika Ogden:* Land use impact on water quality in two river systems in South Africa (2013).
24. *Stefan Rova:* A GIS based approach assessing phosphorus load impact on Lake Flaten in Salem, Sweden (2013).
25. *Yann Buhot:* Analysis of the history of landscape changes over a period of 200 years. How can we predict past landscape pattern scenario and the impact on habitat diversity? (2013).
26. *Christina Fotiou:* Evaluating habitat suitability and spectral heterogeneity models to predict weed species presence (2014).
27. *Inese Linuza:* Accuracy Assessment in Glacier Change Analysis (2014).
28. *Agnieszka Griffin:* Domestic energy consumption and social living standards: a GIS analysis within the Greater London Authority area (2014).
29. *Brynja Guðmundsdóttir:* Detection of potential arable land with remote sensing and GIS - A Case Study for Kjósarhreppur (2014).
30. *Oleksandr Nekrasov:* Processing of MODIS Vegetation Indices for analysis of agricultural droughts in the southern Ukraine between the years 2000-2012 (2014).
31. *Sarah Tressel:* Recommendations for a polar Earth science portal in the context of Arctic Spatial Data Infrastructure (2014).
32. *Caroline Gevaert:* Combining Hyperspectral UAV and Multispectral Formosat-2 Imagery for Precision Agriculture Applications (2014).
33. *Salem Jamal-Uddeen:* Using GeoTools to implement the multi-criteria evaluation analysis - weighted linear combination model (2014).
34. *Samanah Seyedi-Shandiz:* Schematic representation of geographical railway network at the Swedish Transport Administration (2014).
35. *Kazi Masel Ullah:* Urban Land-use planning using Geographical Information System and analytical hierarchy process: case study Dhaka City (2014).
36. *Alexia Chang-Wailing Spitteler:* Development of a web application based on MCDA and GIS for the decision support of river and floodplain rehabilitation projects (2014).
37. *Alessandro De Martino:* Geographic accessibility analysis and evaluation of potential changes to the public transportation system in the City of Milan (2014).
38. *Alireza Mollasalehi:* GIS Based Modelling for Fuel Reduction Using Controlled Burn in Australia. Case Study: Logan City, QLD (2015).

39. *Negin A. Sanati:* Chronic Kidney Disease Mortality in Costa Rica; Geographical Distribution, Spatial Analysis and Non-traditional Risk Factors (2015).
40. *Karen McIntyre:* Benthic mapping of the Bluefields Bay fish sanctuary, Jamaica (2015).
41. *Kees van Duijvendijk:* Feasibility of a low-cost weather sensor network for agricultural purposes: A preliminary assessment (2015).
42. *Sebastian Andersson Hylander:* Evaluation of cultural ecosystem services using GIS (2015).
43. *Deborah Bowyer:* Measuring Urban Growth, Urban Form and Accessibility as Indicators of Urban Sprawl in Hamilton, New Zealand (2015).
44. *Stefan Arvidsson:* Relationship between tree species composition and phenology extracted from satellite data in Swedish forests (2015).
45. *Damián Giménez Cruz*: GIS-based optimal localisation of beekeeping in rural Kenya (2016).
46. *Alejandra Narváez Vallejo:* Can the introduction of the topographic indices in LPJ-GUESS improve the spatial representation of environmental variables? (2016).
47. *Anna Lundgren:* Development of a method for mapping the highest coastline in Sweden using breaklines extracted from high resolution digital elevation models (2016).
48. *Oluwatomi Esther Adejoro:* Does location also matter? A spatial analysis of social achievements of young South Australians (2016).
49. *Hristo Dobrev Tomov:* Automated temporal NDVI analysis over the Middle East for the period 1982 - 2010 (2016).
50. *Vincent Muller:* Impact of Security Context on Mobile Clinic Activities A GIS Multi Criteria Evaluation based on an MSF Humanitarian Mission in Cameroon (2016).
51. *Gezahagn Negash Seboka:* Spatial Assessment of NDVI as an Indicator of Desertification in Ethiopia using Remote Sensing and GIS (2016).
52. *Holly Buhler:* Evaluation of Interfacility Medical Transport Journey Times in Southeastern British Columbia. (2016).
53. *Lars Ole Grottenberg*: Assessing the ability to share spatial data between emergency management organisations in the High North (2016).
54. *Sean Grant:* The Right Tree in the Right Place: Using GIS to Maximize the Net Benefits from Urban Forests (2016).
55. *Irshad Jamal:* Multi-Criteria GIS Analysis for School Site Selection in Gorno-Badakhshan Autonomous Oblast, Tajikistan (2016).
56. *Fulgencio Sanmartín:* Wisdom-volkano: A novel tool based on open GIS and time-series visualization to analyse and share volcanic data (2016).
57. *Nezha Acil:* Remote sensing-based monitoring of snow cover dynamics and its influence on vegetation growth in the Middle Atlas Mountains (2016).
58. *Julia Hjalmarsson:* A Weighty Issue: Estimation of Fire Size with Geographically Weighted Logistic Regression (2016).
59. *Mathewos Tamiru Amato:* Using multi-criteria evaluation and GIS for chronic food and nutrition insecurity indicators analysis in Ethiopia (2016).
60. *Karim Alaa El Din Mohamed Soliman El Attar:* Bicycling Suitability in Downtown, Cairo, Egypt (2016).

61. *Gilbert Akol Echelai:* Asset Management: Integrating GIS as a Decision Support Tool in Meter Management in National Water and Sewerage Corporation (2016).
62. *Terje Slinning:* Analytic comparison of multibeam echo soundings (2016).
63. *Gréta Hlín Sveinsdóttir:* GIS-based MCDA for decision support: A framework for wind farm siting in Iceland (2017).
64. *Jonas Sjögren:* Consequences of a flood in Kristianstad, Sweden: A GIS-based analysis of impacts on important societal functions (2017).
65. *Nadine Raska:* 3D geologic subsurface modelling within the Mackenzie Plain, Northwest Territories, Canada (2017).
66. *Panagiotis Symeonidis*: Study of spatial and temporal variation of atmospheric optical parameters and their relation with PM 2.5 concentration over Europe using GIS technologies (2017).
67. *Michaela Bobeck:* A GIS-based Multi-Criteria Decision Analysis of Wind Farm Site Suitability in New South Wales, Australia, from a Sustainable Development Perspective (2017).
68. *Raghdaa Eissa*: Developing a GIS Model for the Assessment of Outdoor Recreational Facilities in New Cities Case Study: Tenth of Ramadan City, Egypt (2017).
69. *Zahra Khais Shahid*: Biofuel plantations and isoprene emissions in Svea and Götaland (2017).
70. *Mirza Amir Liaquat Baig*: Using geographical information systems in epidemiology: Mapping and analyzing occurrence of diarrhea in urban - residential area of Islamabad, Pakistan (2017).
71. *Joakim Jörwall*: Quantitative model of Present and Future well-being in the EU-28: A spatial Multi-Criteria Evaluation of socioeconomic and climatic comfort factors (2017).
72. *Elin Haettner*: Energy Poverty in the Dublin Region: Modelling Geographies of Risk (2017).
73. *Harry Eriksson*: Geochemistry of stream plants and its statistical relations to soil- and bedrock geology, slope directions and till geochemistry. A GIS-analysis of small catchments in northern Sweden (2017).
74. *Daniel Gardevärn:* PPGIS and Public meetings – An evaluation of public participation methods for urban planning (2017).
75. *Kim Friberg:* Sensitivity Analysis and Calibration of Multi Energy Balance Land Surface Model Parameters (2017).
76. *Viktor Svanerud:* Taking the bus to the park? A study of accessibility to green areas in Gothenburg through different modes of transport (2017).
77. *Lisa-Gaye Greene*: Deadly Designs: The Impact of Road Design on Road Crash Patterns along Jamaica's North Coast Highway (2017).
78. *Katarina Jemec Parker*: Spatial and temporal analysis of fecal indicator bacteria concentrations in beach water in San Diego, California (2017).
79. *Angela Kabiru*: An Exploratory Study of Middle Stone Age and Later Stone Age Site Locations in Kenya's Central Rift Valley Using Landscape Analysis: A GIS Approach (2017).
80. *Kristean Björkmann*: Subjective Well-Being and Environment: A GIS-Based Analysis (2018).
81. *Williams Erhunmonmen Ojo*: Measuring spatial accessibility to healthcare for people living with HIV-AIDS in southern Nigeria (2018).

82. *Daniel Assefa*: Developing Data Extraction and Dynamic Data Visualization (Styling) Modules for Web GIS Risk Assessment System (WGRAS). (2018).
83. *Adela Nistora*: Inundation scenarios in a changing climate: assessing potential impacts of sea-level rise on the coast of South-East England (2018).
84. *Marc Seliger*: Thirsty landscapes - Investigating growing irrigation water consumption and potential conservation measures within Utah's largest master-planned community: Daybreak (2018).
85. *Luka Jovičić*: Spatial Data Harmonisation in Regional Context in Accordance with INSPIRE Implementing Rules (2018).
86. *Christina Kourdounouli*: Analysis of Urban Ecosystem Condition Indicators for the Large Urban Zones and City Cores in EU (2018).
87. *Jeremy Azzopardi*: Effect of distance measures and feature representations on distance-based accessibility measures (2018).
88. *Patrick Kabatha*: An open source web GIS tool for analysis and visualization of elephant GPS telemetry data, alongside environmental and anthropogenic variables (2018).
89. *Richard Alphonce Giliba*: Effects of Climate Change on Potential Geographical Distribution of Prunus africana (African cherry) in the Eastern Arc Mountain Forests of Tanzania (2018).
90. *Eiður Kristinn Eiðsson*: Transformation and linking of authoritative multi-scale geodata for the Semantic Web: A case study of Swedish national building data sets (2018).
91. *Niamh Harty*: HOP!: a PGIS and citizen science approach to monitoring the condition of upland paths (2018).
92. *José Estuardo Jara Alvear*: Solar photovoltaic potential to complement hydropower in Ecuador: A GIS-based framework of analysis (2018).
93. *Brendan O'Neill*: Multicriteria Site Suitability for Algal Biofuel Production Facilities (2018).
94. *Roman Spataru*: Spatial-temporal GIS analysis in public health – a case study of polio disease (2018).
95. *Alicja Miodońska*: Assessing evolution of ice caps in Suðurland, Iceland, in years 1986 - 2014, using multispectral satellite imagery (2019).
96. *Dennis Lindell Schettini*: A Spatial Analysis of Homicide Crime's Distribution and Association with Deprivation in Stockholm Between 2010-2017 (2019).
97. *Damiano Vesentini*: The Po Delta Biosphere Reserve: Management challenges and priorities deriving from anthropogenic pressure and sea level rise (2019).
98. *Emilie Arnesten*: Impacts of future sea level rise and high water on roads, railways and environmental objects: a GIS analysis of the potential effects of increasing sea levels and highest projected high water in Scania, Sweden (2019).
99. *Syed Muhammad Amir Raza*: Comparison of geospatial support in RDF stores: Evaluation for ICOS Carbon Portal metadata (2019).
100. *Hemin Tofiq*: Investigating the accuracy of Digital Elevation Models from UAV images in areas with low contrast: A sandy beach as a case study (2019).
101. *Evangelos Vafeiadis*: Exploring the distribution of accessibility by public transport using spatial analysis. A case study for retail concentrations and public hospitals in Athens (2019).
102. *Milan Sekulic*: Multi-Criteria GIS modelling for optimal alignment of roadway by-passes in the Tlokweng Planning Area, Botswana (2019).

103. *Ingrid Piirisaar*: A multi-criteria GIS analysis for siting of utility-scale photovoltaic solar plants in county Kilkenny, Ireland (2019).
104. *Nigel Fox*: Plant phenology and climate change: possible effect on the onset of various wild plant species' first flowering day in the UK (2019).
105. *Gunnar Hesch*: Linking conflict events and cropland development in Afghanistan, 2001 to 2011, using MODIS land cover data and Uppsala Conflict Data Programme (2019).
106. *Elijah Njoku*: Analysis of spatial-temporal pattern of Land Surface Temperature (LST) due to NDVI and elevation in Ilorin, Nigeria (2019).
107. *Katalin Bunyevácz*: Development of a GIS methodology to evaluate informal urban green areas for inclusion in a community governance program (2019).
108. *Paul dos Santos*: Automating synthetic trip data generation for an agent-based simulation of urban mobility (2019).
109. *Robert O' Dwyer*: Land cover changes in Southern Sweden from the mid-Holocene to present day: Insights for ecosystem service assessments (2019).
110. *Daniel Klingmyr*: Global scale patterns and trends in tropospheric NO2 concentrations (2019).
111. *Marwa Farouk Elkabbany*: Sea Level Rise Vulnerability Assessment for Abu Dhabi, United Arab Emirates (2019).
112. *Jip Jan van Zoonen*: Aspects of Error Quantification and Evaluation in Digital Elevation Models for Glacier Surfaces (2020).
113. *Georgios Efthymiou*: The use of bicycles in a mid-sized city – benefits and obstacles identified using a questionnaire and GIS (2020).
114. *Haruna Olayiwola Jimoh*: Assessment of Urban Sprawl in MOWE/IBAFO Axis of Ogun State using GIS Capabilities (2020).
115. *Nikolaos Barmpas Zachariadis*: Development of an iOS, Augmented Reality for disaster management (2020).
116. *Ida Storm*: ICOS Atmospheric Stations: Spatial Characterization of CO2 Footprint Areas and Evaluating the Uncertainties of Modelled CO2 Concentrations (2020).
117. *Alon Zuta*: Evaluation of water stress mapping methods in vineyards using airborne thermal imaging (2020).
118. *Marcus Eriksson*: Evaluating structural landscape development in the municipality Upplands-Bro, using landscape metrics indices (2020).
119. *Ane Rahbek Vierø*: Connectivity for Cyclists? A Network Analysis of Copenhagen's Bike Lanes (2020).
120. *Cecilia Baggini*: Changes in habitat suitability for three declining Anatidae species in saltmarshes on the Mersey estuary, North-West England (2020).
121. *Bakrad Balabanian*: Transportation and Its Effect on Student Performance (2020).
122. *Ali Al Farid*: Knowledge and Data Driven Approaches for Hydrocarbon Microseepage Characterizations: An Application of Satellite Remote Sensing (2020).
123. *Bartlomiej Kolodziejczyk*: Distribution Modelling of Gene Drive-Modified Mosquitoes and Their Effects on Wild Populations (2020).
124. *Alexis Cazorla*: Decreasing organic nitrogen concentrations in European water bodies - links to organic carbon trends and land cover (2020).

125. *Kharid Mwakoba*: Remote sensing analysis of land cover/use conditions of community-based wildlife conservation areas in Tanzania (2021).
126. *Chinatsu Endo*: Remote Sensing Based Pre-Season Yellow Rust Early Warning in Oromia, Ethiopia (2021).
127. *Berit Mohr*: Using remote sensing and land abandonment as a proxy for long-term human out-migration. A Case Study: Al-Hassakeh Governorate, Syria (2021).
128. *Kanchana Nirmali Bandaranayake*: Considering future precipitation in delineation locations for water storage systems - Case study Sri Lanka (2021).
129. *Emma Bylund*: Dynamics of net primary production and food availability in the aftermath of the 2004 and 2007 desert locust outbreaks in Niger and Yemen (2021).
130. *Shawn Pace*: Urban infrastructure inundation risk from permanent sea-level rise scenarios in London (UK), Bangkok (Thailand) and Mumbai (India): A comparative analysis (2021).
131. *Oskar Evert Johansson*: The hydrodynamic impacts of Estuarine Oyster reefs, and the application of drone technology to this study (2021).
132. *Pritam Kumarsingh*: A Case Study to develop and test GIS/SDSS methods to assess the production capacity of a Cocoa Site in Trinidad and Tobago (2021).
133. *Muhammad Imran Khan*: Property Tax Mapping and Assessment using GIS (2021).
134. *Domna Kanari*: Mining geosocial data from Flickr to explore tourism patterns: The case study of Athens (2021).