NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

GEOHEALTH. A GEOGRAPHIC INFORMATION SYSTEM FOR CLINICAL RESEARCH

Víctor González Gil

Work Project presented as partial requirement for obtaining the Master's degree in Advanced Analytics

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

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

GEOHEALTH. A GEOGRAPHIC INFORMATION SYSTEM FOR CLINICAL RESEARCH

by

Víctor González Gil

Work Project presented as partial requirement for obtaining the Master's degree in Advanced Analytics

Advisor: Fernando José Ferreira Lucas Bação (Int)

Co Advisor*:* Alberto Moreno Conde (Ext)

August 2021

ACKNOWLEDGEMENTS

Thank you to my supervisors, Fernando and Alberto, for providing guidance and feedback during the development of this project.

To my parents and the rest of my family for always believe in me and help me in all steps of my life, thank you very much. It would not be possible without their infinite support.

To all the people who, in one way or the other, were part of my life and experiences, thank you very much.

ABSTRACT

The objective of the project has been the development of a system for clinical, epidemiological, and translational research capable of associating environmental variables and geospatial information with clinical patient's data. The system aims to empower clinical researchers with a system that helps improve the generation of more meaningful and applicable results that directly benefit human health. The system has integrated pollution, pollen, and meteorological data as well as demographic information of the region of Andalusia, Spain. The heterogeneous data comes from different sources such as the Meteorological Spanish Agency (AEMET), the Spanish Ministry for the Ecological Transition, and the Demographic challenge or the National Institute of Statistics (INE). The project includes three main deliverables: the development of a system for clinical, epidemiological, and translational research, a geodemographic segmentation, and an air quality characterization.

The system, called "Geohealth", includes a module for the observation of all the data on a single map. This module allows researchers to observe and understand the geographical distribution of patients, the demographic information of the census sections, as well as the time series of the meteorological, pollution, and pollen data.

In addition, a second module called "Segmentations", includes the results of the first study developed for this project which aimed at discovering and understanding which are the characteristics of the different clusters of census sections among the province of Seville (Spain) through the development of two geodemographic segmentations.

A second study, based on the pollution times series data, has been developed and integrated into the system. The pollution time series have been processed and the different values, measured by the pollution stations, have been interpolated for the whole territory. Once each census section has assigned the values for each pollutant, they have been transformed into the Air Quality Health Index (AQHI) scale, developed by the Canadian government, which allows understanding the impact of the combination of the pollutant on health.

The system has been evaluated in collaboration with the Allergology department of the Virgen Macarena Hospital, with a dataset of asthmatic patients that have attended emergency care due to an asthma attack during the years between 2014 and 2019.

KEYWORDS

GIS; Asthma; Pollution; Clustering; Segmentation; Air Quality

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

SO2	Sulfur dioxide
NO2	Nitrogen dioxide
PM	Particles (Air pollutant)
O3	Atmospheric Ozone
PART	Particles
EHR	Electronic Health Record
INE	Spanish National Institute of Statistics
IECA	Andalusian Institute of Statistics and Cartography
ERD	Entity Relationship Diagram
AQI	Air Quality Index
AQHI	Air Quality Health Index

1. INTRODUCTION

The project begins with the objective of developing a geographic information system that facilitates researchers to study and understand the relationship between the patients and the demographic and environmental variables associated with them. Two studies are developed, one is a geodemographic segmentation, and the second one is an air quality characterization. The system has been evaluated with a dataset of asthmatic patients in collaboration with the allergology unit of Virgen Macarena Hospital (Seville, Spain). Moreover, other units have shown interest in using the system for their investigations.

1.1. CONTEXTUALIZATION

This project has been developed by the Innovation Area of Virgen Macarena Hospital in Seville (Spain) which is one of the regional hospitals of the Andalusian Public Health System. It has care, research, and teaching function and offers health care to a population of 481,296 people divided into three health districts (Hospital Virgen Macarena, 2021).

The system was born with the idea of relating each patient to the large amount of open data that exists in Andalusia (Spain). Many studies developed in other countries demonstrate the influence of patient's environmental variables on the prevalence of diseases. The different sections and utilities of the system have been evaluated with a dataset of patients who have been treated in the hospital emergency unit due to an asthma episode.

1.2. PROJECT'S BACKGROUND AND MOTIVATION

Asthma is a disease of high prevalence worldwide, with wide variation according to the country and according to the epidemiological criteria used for its definition. The prevalence range has been estimated between 2% and 12% of the population (Pearce et al., 2007; ECRHHS, 2002). In Spain, the prevalence of asthma in paediatric age has stabilized in the age group of 13-14 years old and increased in the age group of 6-7 years old, with a prevalence of diagnosis of asthma at some point in the life of 12.8% and 10.9% respectively (Garcia-Marcos et al., 2004). In adulthood, the prevalence range is between 1.1% and 4.7% of the population according to the geographical area studied (Martínez-Moratalla et al., 1999). Asthma is not only a major health problem but a significant economic cost to the public health system. The cost of asthma disease has been estimated at around 1% of total health spending in industrialized countries (Bateman et al., 2008). Furthermore, it is expected that the impact of asthma will increase in the coming years as a result of longer life expectancy, an increasing prevalence, and the emergence of new drugs and therapeutic modalities (Braman, 2006). It has been estimated that the annual cost of asthma in Spain is 1,480 million euros considering the prevalence based on the symptomatic diagnosis associated with bronchial hyperresponsiveness, and 3,022 million euros if only the symptomatic diagnosis of asthma are considered. According to this, 70% of this cost is attributed to the bad control of the disease (Martínez-Moragón et al., 2009).

A large number of epidemiological studies demonstrate the association between hospitalizations and concentrations of SO_2, NO_2, Ozone, and Diesel Exhaust Particles in asthmatic patients (Khreis et al., 2017). A study developed in Beijing, China demonstrated the direct relationship between the concentration of the pollutant PM and the health service use. Every 10 $\mu g/m^3$ increase in PM concentration on the same day was associated with a 0.67% increase in total hospital visits (Tian et al.

2007). Furthermore, new paediatric asthma cases could be attributable to NO_2 pollution concentration being urban centres the main affected areas (Achakulwisut et al, 2019). The percentage of new asthma cases attributable to NO2 pollution among main cities is ranged from 5-6% in Orlu, Nigeria, to 48% in Shanghai, China (Achakulwisut et al, 2019).

The advances made in the last years in data science and computation for data management, integration, data mining, visualizing, and others have opened a new world of solutions and approaches to study how to address many health problems and healthcare treatments (Shaban-Nejad et al., 2018). Since the 1970s when the United States began using the "Electronic Health Record" (EHR) (Atherton, 2011), the quantity of clinical data, that is available electronically, dramatically increased. Many advances have been made in clinical analytics such as techniques for analysing large quantities of data and getting new insights from that analysis (Bates et al., 2014). Many studies have already shown the impact of data science's techniques in the analysis and prediction of diseases. For instance, a study conducted in Taiwan has shown the potential of machine learning in the early detection of abnormalities in heart conditions, allowing to prevent of heart attacks and being the mortality rate drastically controlled (Mohan et al., 2019).

The IT infrastructure of the Andalusian Health Service integrates the information generated during patient care in the Electronic Medical Record (called DIRAYA) (Marín, J. M. M., & Cámara, S. B. 2008). This infrastructure includes clinical and demographic data about the 8 million patients it covers. Based on the data provided by the Population Database and the prevalence studies in Spain, it was estimated that this platform could analyse the evolution of more than 100,000 asthmatics in the region of Andalusia.

Due to the importance shown by other studies and the lack of tools in the Andalusian health system that integrate this large amount of heterogeneous data and that allow researchers to study this area simply and effectively, the hospital's innovation department has developed a system that is being tested in the hospital and has the intention of expanding to the entire Andalusian territory.

1.3. PROJECT'S OBJECTIVES

The main objectives of this project can be summarized in the following list:

1. Definition of the structure, technology, and tools used to develop the system.
2. Extraction and processing of the open data coming from different public institutions.
3. Development of the platform and integration of the heterogeneous data.
4. Integration of the patient's data into the system.
5. Development of a geodemographic segmentation and an air quality characterization.
6. Evaluation of the system with the asthma dataset.

1.4. SYSTEM'S STRUCTURE, TECHNOLOGY & TOOLS

The system is a web-based application developed using the open-source Python library "Dash". The project aims to develop a system made of open-source technologies and it will be available on the intranet of the Andalusian Health System. It has been integrated into the Innovation Area server using a Docker container for the integration of the application and a set of databases to store the data.

1.5. DEVELOPMENT OF THE STUDIES

Two studies have been developed and integrated into the system, providing valuable information for the researchers' studies.

1.5.1. Study I

This first study consists of carrying out a geodemographic segmentation that allows a better understanding of the different population groups existing in the province of Seville (Spain). This study will allow researchers to have a general idea of what the general characteristics of their patients are, according to where they live. This segmentation is integrated into the system and allows users to observe how the population of patients is distributed among the clusters and to understand what each cluster represents intuitively. The segmentations have shown the different homogeneous groups that exist in the province of Seville due to their characteristics.

1.5.2. Study II

The second study is focused on the processing and analysis of the pollution time-series, in order to make it useful for researchers, to study the relationship between diseases (such as the attendance to emergency care of patients due to an asthmatic event) and the environmental conditions. Using the system "Geohealth" for its representation and the processing carried out for the treatment of the pollution open data explained in the following chapters, the different regions of the province of Seville have been classified according to the levels of pollution registered by the measurement stations. The study has shown that the concentration of pollutants is directly related to the seasons and the environmental conditions due to pollution concentrations vary between the different areas of the region.

1.6. DOCUMENT STRUCTURE

This project's report begins with an introduction and a theoretical background research, to justify every decision made during the development of this project. Then, three chapters are exposed, the first one explaining the system's development and implementation, and then the methodology and results obtained for the two study cases. To conclude, in the final chapter the conclusions of this project have been discussed.

2. THEORETICAL FRAMEWORK

The advances made in the last years in data science and computation for data management, integration, data mining, visualizing, and others have opened a new world of solutions and approaches to study how to address many health problems and healthcare treatments. As Shaban-Nejad et al. (2018) say: "Theory, methods, and models from artificial intelligence (AI) are changing the health care landscape in clinical and community settings and have already shown promising results in multiple applications in healthcare including, integrated health information systems, patient education, geocoding health data, epidemic and syndromic surveillance, predictive modelling and clinical decision support".

This chapter presents an overview of the technology and methodology used during the development of this project.

2.1. SYSTEM'S TECHNOLOGY

2.1.1. Server, Docker & Anaconda

The system has been developed and integrated into the Innovation Area (Virgen Macarena Hospital) Linux server. It is connected to the intranet of the Andalusian Health System for data storage as well as the implementation of websites, apps, platforms, and databases for the development of projects.

The technology used for the integration and implementation of the project applications is Docker. Docker is based on the concept of containers, an encapsulation of an application with its dependencies. The use of containers provides many advantages, but the most remarkable is the possibility to create software in your computer, knowing that it will work identically in any other machine or server (Mouat, 2016, p.3).

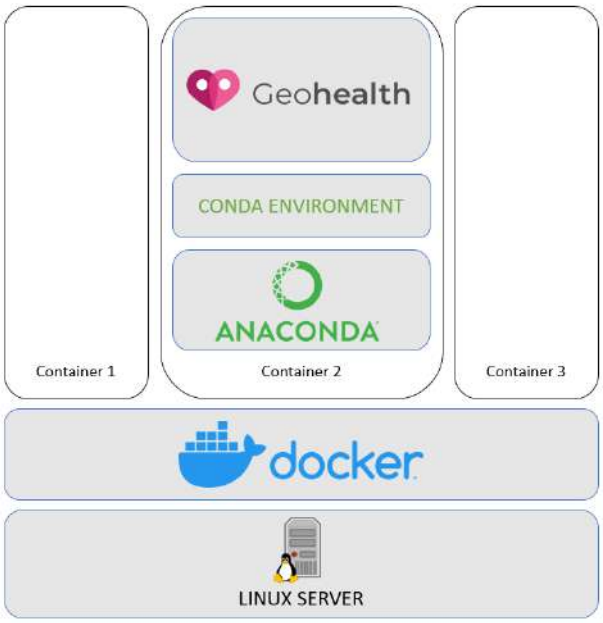


Figure 2.1 – Structure of the project's system

Source: Author

In order to run the system “Geohealth”, Anaconda has been integrated into the docker container. Anaconda is an open-source package management system able to run on the most famous OS systems (Windows, macOS, and Linux) (Conda, n.d.). Anaconda provides every dependency required by the Python application (libraries, tools ...etc.) creating virtual environments. A virtual environment is a tool that keeps all the dependencies needed by a specific project and allows the creation of an isolated space for it.

All this technology combined into a single system, following the structure observed in Figure 2.1, allows Geohealth to work in the Intranet of the Andalusian Health System, having access to the clinical datasets of the different projects.

2.1.2. Databases

PostgreSQL has been the chosen database management system because is being used already in other projects and following the priority of the Innovation Area of using, as much as possible, open-source software.

PostgreSQL is a free and open-source relational database management system with more than 30 years of history (PostgreSQL, n.d.). An advantage of PostgreSQL is the flexibility it offers in projects, being compatible and allowing the development of custom functions in some of the most famous languages such as Python and Java (PostgreSQL, n.d.).

Moreover, one of the reasons PostgreSQL has been chosen as the relational database administrator is due to the PostGIS extension. Geohealth is a geographic information system and the functions and advantages that this extension offers are very positive for the projects. PostGIS is open-source software that adds support for geographic objects to a PostgreSQL database. It provides many geospatial functions and contains all the spatial data types. Some of the functions are focused on table management such as “AddGeometryColumn” to add a column with the geometries, other focuses on geometry creations such as “ST_MakePolygon” to create the polygon from a list of points and others for geometry relationship such as “ST_Within” to know if a point or polygon is completely inside of another polygon (PostGIS, n.d.)

2.1.3. Python

Python has become one of the most important programming languages in the field of data science due to its high readability, flexibility, and its constantly evolving libraries (Nagpal et al., 2019). One of the main reasons for the popularity of the Python ecosystem is the huge set of libraries for scientific calculations, data visualization, data analytics, and others (Sutchenkov et al., 2020). Some of the most relevant libraries are aimed to improve data visualization such as Matplotlib, Seaborn, or Plotly. Pandas library provides high-performance, easy-to-use data structures, and data analysis tools. Numpy library allows users to work easily with multidimensional arrays of data and others such as Scikit-learn, Keras and TensorFlow allow implementing data mining or machine learning techniques.

2.1.4. Dashboards & Data Visualization

A dashboard is a visual display of the most important information with the aim of helping the understanding and analysis of the information and providing to the user a global overview. In order to

carry out its task quickly and efficiently, all the information is displayed consolidated, and organized on a single screen (Few, 2007).

Develop a dashboard should follow some fundamental basics and rules in order to improve efficiency. Some of the most important ones can be found in the following list (Karami et al., 2017):

- During the use of a dashboard, it must be possible to interact with the content by drilling down in the existing information, as well as the existence of filters, sliders, and others that allow customization and flexibility of the system.
- It should be simple to use and learn in order to not require a long training period.
- The use of different modules that keep the dashboard organized to analyse the information.
- Colours should enhance the information in a way that makes it easier to understand and analyse. Using the same colours to highlight groups and allow the information to be compared. Moreover, the use of a colour scale will allow to highlight information.
- A wide variety of graphs or very complex graphs could make observation and analysis a difficult task. Therefore, the correct selection of charts that allow a simple and legible dashboard is one of the most important parts.

Due to the high number of advantages offered to create an interactive dashboard written fully in Python, “Dash” has become one of the most popular frameworks used for data visualization in the field of Data Sciences. Dash is an open-source framework of Plotly that empowers users to create interactive full-stack web applications (Hossain et al., 2019) released in 2017. It is written on top of the most popular frameworks and libraries as Flask, React, and Plotly. Using Dash you can work purely in Python, the backend, and also the frontend can be written using this programming language, thanks to the independent library to create the web user interface. It is called “dash-html-components” and it provides wrappers for all HTML tags for Python. A second library called “dash-core-components” allows you to create user interface components such as input fields, buttons, sliders, and others (Sutchenkov et al., 2020). Moreover, the library “graph_objects” gives access to a wide range of different types of charts for visualization.

The importance of data visualization in the healthcare field is being clearly observed during the Covid-19 pandemic. Many data visualization platforms have been developed to inform media, authorities, and researchers about the situation of the pandemic around the world such as Johns Hopkings dashboard (Dong et al., 2020). A visualization system provides a representation of data specifically designed to allow people to carry out tasks more effectively. As Munzner (2014) said, “visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods”.

Previous developed studies and the lack of researches and systems in the region of Seville and Andalusia (Spain) have motivated the development of this project. An example of those studies was conducted at Oxford University showed the potential of data visualization computer-based systems in the development of research and helping in decision making. Visualization of the data in a specific system allowed a complete resolution of the outbreak and the associated isolation, demonstrating that multiple closely related but distinct strains were simultaneously present (Jolley et al., 2012). Data

visualization technology allowed to complete the whole process in less than 48 hours (Jolley et al., 2012).

2.2. DATA MINING & HEALTHCARE

Data Mining is a set of techniques that allow exploring large datasets in order to discover structures and patterns that explain the behaviour of the data (Hand et al., 2015). A data mining project consists of some fundamental steps that can be summarized in determining the objectives, processing the data, creating the model, analysing the results, and deploying the results (Ribas, 2018). Many techniques and algorithms exist in the data mining field, and they can be grouped into two categories according to the objective of the problem as can be observed in Figure 2.2.

One of the most famous techniques of Data Mining falls into the descriptive modelling category and its name is clustering. Clustering is the classification of patterns into groups, and it has been used in many contexts and disciplines. Cluster analysis encompasses different techniques and algorithms for grouping data that share a common behaviour into their respective category. Many algorithms exist and each one leads to a different result, but it does not exist an approach to select the best algorithm (Grekousis et al., 2012). It is necessary to understand how each algorithm works, the data that is going to be used, and the objective of the analysis in order to choose the best approach.

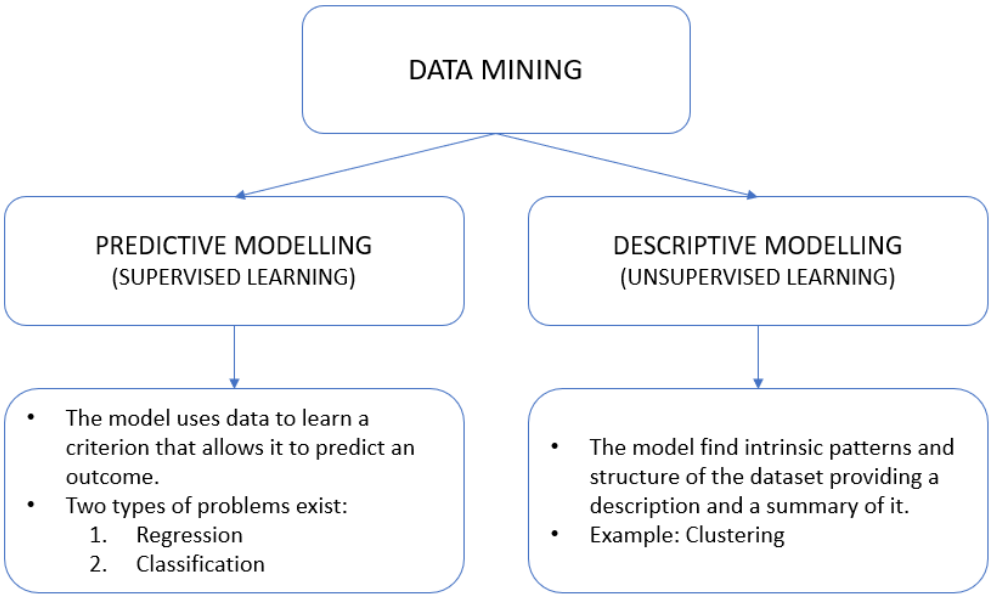


Figure 2.2 – Data Mining modelling categories

Source: Author

As an example of how useful data mining techniques can be in healthcare, a study developed by Requia et al. (2019) used a multivariate clustering approach to identify spatial patterns in pollution particles (PM2.5) in Eastern Massachusetts, United States. The segmentation was performed considering air pollution sources and geodemographic variables and it was discovered a substantial variation of clusters among PM2.5 components. Furthermore, variables such as land use, population density, and daily traffic were used to characterize clusters more effectively. The parameter used to estimate the effectiveness of each variable in characterizing clusters was R2 values. Larger R2 values indicate better

discrimination among the sites. For instance, it was discovered that population density had the highest discriminatory influence when the analysis was performed for some types of particles and land use for others. This study can help during the development of health campaigns and clinical decision-making using different strategies according to the region for some diseases such as asthma.

2.2.1. Geodemographic segmentation

Geodemography is a combination of three social sciences: Sociology, Geography, and Demography and it is applied in many fields such as urban planning, marketing analysis, public health, and others. Geodemographic segmentation, or geodemographic clustering, classifies a set of areas into different groups who share similar characteristics among multiple socioeconomic attributes (Singleton et al., 2013). This association can be focused on describing the generalities of the different regions or focused on a specific field such as health. Geodemographic segmentation is based on two principles (Sivadas, 2017):

- People living in the same area are likely to have similar characteristics.
- Areas can be categorized using the demographics of the households they contain.

Geodemographic segmentation has been widely used in marketing analysis to classify and characterize the different types of customers. It has demonstrated its potential to develop strategies according to the characteristics of the customer. Geodemographic segmentations started to be used due to their capability to describe the behaviour of the customers and, therefore, helping in market analysis (Longley & Clarke, 1995). Some systems have been developed by companies to help in this task such as Acorn or Mosaic but the use is not widely extended and exist a lack of understanding about this field and tools in the health sector (Abbas et al., 2009).

Some applications can be useful for the health sector such as (Abbas et al., 2009) :

- Population profiling: The profiles provide a general description of the population living in each area based on the most dominant attributes. The combination of this geodemographic information with health data can provide useful insights about the population.
- It can be used to study the probability of occurrence of an event due to the location of the patients or the relationship between the variables of a specific region with a particular health event. For instance, the probability of a patient needing emergency cares due to an asthma attack according to his location.

Many studies have been developed and have shown the potential of geodemographic segmentation such as the study presented by Petersen et al. (2011). The study showed the potential of it as a technique that can give valuable demographic understanding to many public sector applications. They presented how geodemographic segmentation offers a wealth of demographic information that can help define the best strategy, for example, for a health campaign.

Moreover, another study presented by Bright et al. (2020) found that geodemography can show some differences in the risk of emergency events in patients with cancer according to their location. They

propose that the use of geodemography could help to orient public health interventions for earlier detection, for instance, by screening programs or awareness campaigns.

2.2.2. Geospatial Data

Geospatial data is data about something in particular, an event or a phenomena that have a location on the surface of the earth. The location can be static such as the location of a pollution station or the location of a Hospital, or dynamic such as a moving vehicle. Geospatial data combines location information, attribute information, and sometimes temporal information (Stock & Guesgen, 2016). Location information is usually the coordinates of the location represented by latitude (angle from the equator in south-north direction) and longitude (angle from the prime meridian in west-east direction). The attribute information encompasses all the variables related to the object or event that it is being located and the temporal information exists when the information is also related to a date and time in particular.

2.2.3. Census Data

In Spain, a census section is the minimum territorial unit, being lower than the municipality. All municipalities are divided into one or more census sections and there is no point on their territory that is not included in a census section. Different sets of census regions within the same municipality are grouped in areas called census districts.

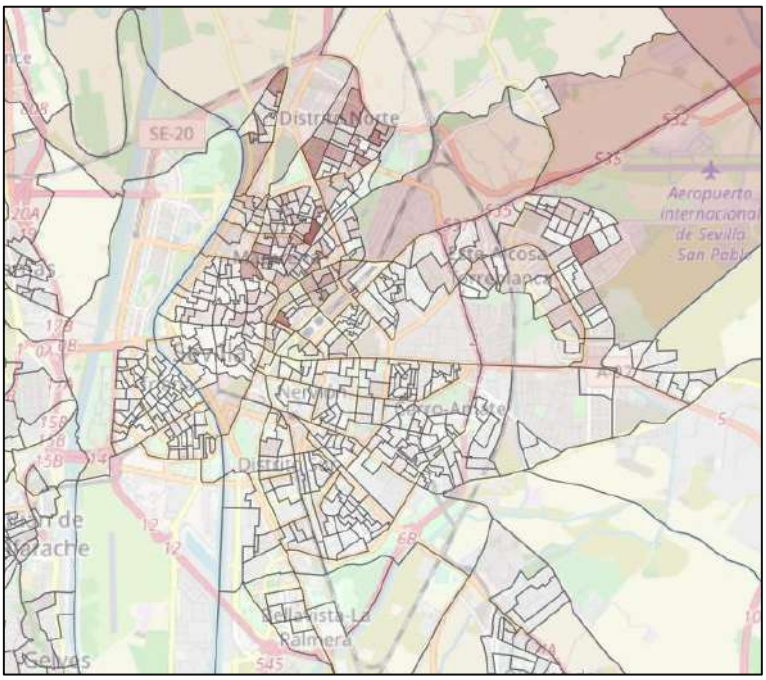


Figure 2.3 – Example of the census sections of the municipality of Seville.

Source: Geohealth.

The census survey and study has been carried out in Spain since the end of the 19th century and it is developed every ten years (INE, n.d.). Census sections are the units with the lowest level of statistical information about the Spanish population and are used, for example, to organize electoral processes.

All the information disseminated by the INE can be used free of charge by citizens, companies, and institutions and it is published on the website of the Spanish National Institute of Statistics (INE).

In the following figure (Figure 2.4) can be observed the different categories of information and some examples of the variables that are included in the census sections.

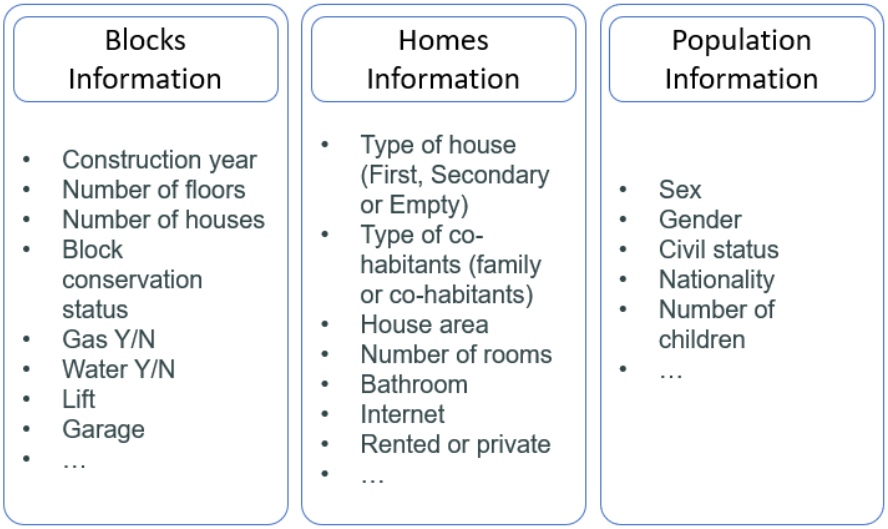


Figure 2.4 – Categories and variables of the census information

Source: Author. Adapted from the Spanish National Institute of Statistics information.

2.2.4. Environmental Data

Three environmental types of data have been chosen to be integrated into the system due to their, already shown by numerous studies, influence on human health. In the three following points, it has been described which are these variables and how they are measured.

2.2.4.1. Pollution

Andalusia (Spain) has a network of air quality measurement stations made up of more than 50 stations spread throughout the territory and with a special presence in urban areas. These stations capture the presence of some of the best-known pollutants due to their influence on air quality. Depending on the pollutant, they are measured in different units:

- **µg/m³**: sulfur dioxide (SO_2), nitrogen dioxide (NO_2), suspended particles smaller than 10 µm (PM_{10}), suspended particles smaller than 2.5 µm ($PM_{2.5}$), and ozone (O_3).
- **mg/m³**: carbon monoxide (CO).

Air Quality Health Index (AQHI)

The Air Quality Health Index or "AQHI" is a scale designed in Canada with the purpose of helping people understand better the air quality levels and the impact on their health. It has been designed to help people make decisions to protect their health by controlling the exposure to air pollution and adjusting the activity levels according to the quality index (Canada, 2019).

The index is calculated by combining three of the best-known pollutants, which are directly related to the impact of air quality on people's health: NO_2, O_3 and PART. O_3 and NO_2 are measured in parts per billion (ppb) while $PM_{2.5}$ is measured in micrograms per cubic meter (µg/m^3) (Stieb, D. M., 2008). The average values of each pollutant are calculated, and the next formula is applied to obtain the AQHI:

$$AQHI = \left(\frac{1000}{10.4}\right) x[(e^{0.000537\,x\,O_3} - 1) + (e^{0.000871\,x\,NO_2} - 1) + (e^{0.000297\,x\,PART} - 1)$$

The formula provides a number from 1 to 10 that will indicate the health risk according to the pollution level. As higher as the index is, higher will be the health risk as it can be observed in the following figure where are summarized the description of what each index means:

Health Risk	Air Quality Health Index	Health Messages	
		At Risk Population(i.e: Asthmatics)	General Population
Low	1 - 3	**Enjoy** your usual outdoor activities.	**Ideal** air quality for outdoor activities.
Moderate	4 - 6	**Consider reducing** or rescheduling strenuous activities outdoors if you are experiencing symptoms.	**No need to modify** your usual outdoor activities unless you experience symptoms such as coughing and throat irritation.
High	7 - 10	**Reduce** or reschedule strenuous activities outdoors. Children and the elderly should also take it easy.	**Consider reducing** or rescheduling strenuous activities outdoors if you experience symptoms such as coughing and throat irritation.
Very High	Above 10	**Avoid** strenuous activities outdoors. Children and the elderly should also avoid outdoor physical exertion.	**Reduce** or reschedule strenuous activities outdoors, especially if you experience symptoms such as coughing and throat irritation.

Figure 2.5 – AQHI categories and their descriptions

Source: Author. Adapted from Understanding Air Quality Health Index messages (Canada, 2015).

2.2.4.2. Pollen

Andalusia has a network of pollen level measurement stations made up of more than 60 stations spread throughout the territory. These stations capture the presence of some of the best-known types of pollen due to influence on people's health due to allergies and other health-related problems. The unit of measurement is the number of pollen grains divided by cubic meter (nº/m^3) and some of the best-known are Gramineas, Cupressaceous, Platanus, Olea, and others.

2.2.4.3. Meteorology

Andalusia has two large institutions dedicated to the measurement of meteorological variables:

- The State Meteorological Agency (AEMET): Spanish meteorological agency that has measurement stations distributed throughout the national territory, with a greater presence in urban centers.

- Andalusian Agroclimatic Information Network (RIA): This network has a greater presence in rural areas of Andalusia.

Both networks provide very valuable information covering both areas of the territory: rural and urban. Some of the variables measured are the minimum, maximum and average temperatures, humidity, wind, precipitation, pressure, and others.

2.2.5. Algorithms & Techniques

In this section it will be briefly describe the algorithms used in this project.

2.2.5.1. K-means

One of the most used clustering algorithms in data science and geodemographic segmentation is the k-means algorithm (Adnan et al., 2010. It falls into the group of partitioning methods and they organize the objects/observations into different groups or clusters (Han et al., 2011, p. 451). The partitioning algorithms organize the objects/observation in k partitions or clusters, being k smaller or equal to the number of objects/observations. The partitioning method used by the algorithms is based on the distance between each point, assuming that the points that are closer to one another share certain characteristics or similarities allowing the algorithm to divide the population of observations into different groups.

K-means is a centroid-based partitioning technique and that means that the algorithm uses the "centroid" (centre point) of each cluster as a representation of it. The difference between a particular observation (point) existing inside a cluster is measured by the Euclidean distance between the point and the centroid of the cluster it belongs. This measure allows us to calculate the quality of the cluster by the sum of squared error between all the population of observations. Using this function (Figure 2.6) as the objective function the algorithm tries to get the final cluster as compact as possible and as separate as possible between each other (Han et al., 2011, p. 452).

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c}_i)^2,$$

Figure 2.6 – Objective function of k-means algorithm

Source: Book "Data Mining. Concepts and Techniques". Page 451

The k-means algorithm starts by randomly choosing *k* points of the data, called seeds. Then, it allocates the rest of the data points to the nearest seed and the centroid is calculated for each group of points (cluster). Using this new calculated point (new seeds), the algorithm repeats the process of allocating each point to the nearest seed and creating a new version of the clusters. The algorithm repeats these steps until a convergence criterion is met (Adnan et al., 2010).

As it can be observed in Figure 2.7, in "(a)" three random seeds are chosen, and three clusters are created; in "(b)" clusters centroids are updated, and objects are reassigned according to the distance to the closest centroid. Finally, in "(c)", the algorithm finishes with the final clusters.

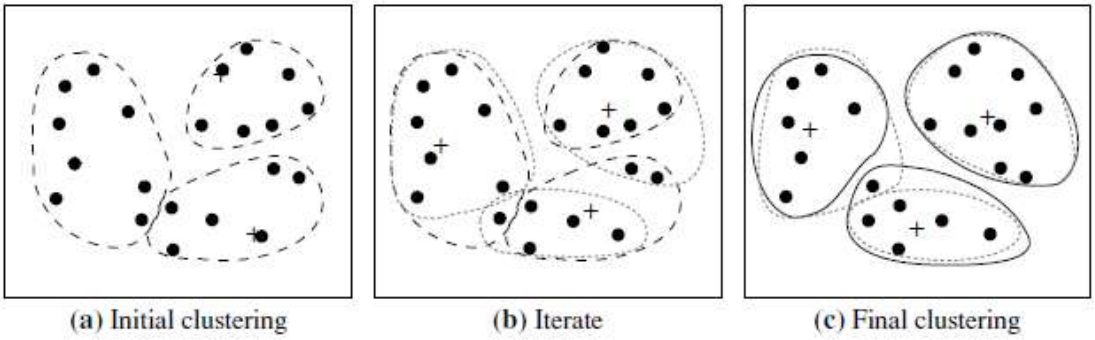


Figure 2.7 – Example of a clustering using the *k*-means method

Source: Book "Data Mining. Concepts and Techniques". Page 453

Moreover, Singleton & Longley (2009) illustrated how the final result of the algorithm is sensitive to the seeds that were randomly selected at the beginning, with consequences for the performance of the cluster model. They suggest that, in order to optimise the classification, the model should be run multiple times starting with different seeds to obtain a better result.

2.2.6. Elbow Diagram

K-means falls into the category of the unsupervised algorithm, but one parameter must be chosen to apply it: the number of clusters that are going to be created "k". There is not an exact method to choose the number of clusters and it depends on the problem that is being faced. One of the best-known methods to choose the number of clusters is the Elbow Diagram. This method consists of representing the sum of squared distances of samples to their closest cluster center, for a set of values of the variable k. The optimal number of clusters will be located at the "elbow" represented on the graph. This point corresponds to the number of clusters from which the amount of improvement in the sum of squared distances decreases considerably with each increase of k. This means that the complexity increases but the level of error from that point do not decrease substantially, to justify the introduction of greater complexity in the analysis. This can be observed, for a better understanding, in the following figure:

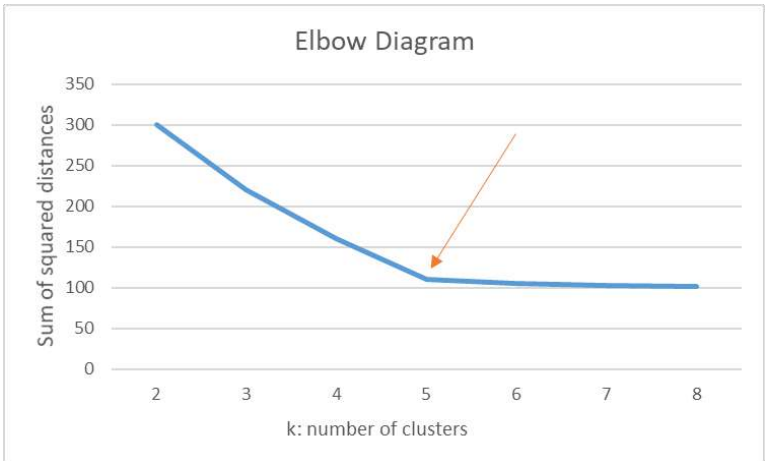


Figure 2.8 – Example of an Elbow Diagram. Optimal point represented by an orange arrow

Source: Author

2.2.7. Profiling Method

Once the number of clusters has been selected and the algorithms have calculated the different clusters, the profiling task must be performed to understand what each cluster represents. Some methods exist and the selected one is comparing the mean value of each cluster for each variable with the mean value of the whole dataset for each variable. The most significant differences with the mean will be considered to understand what each cluster represents. In the following figure can be observed an example of this technique:

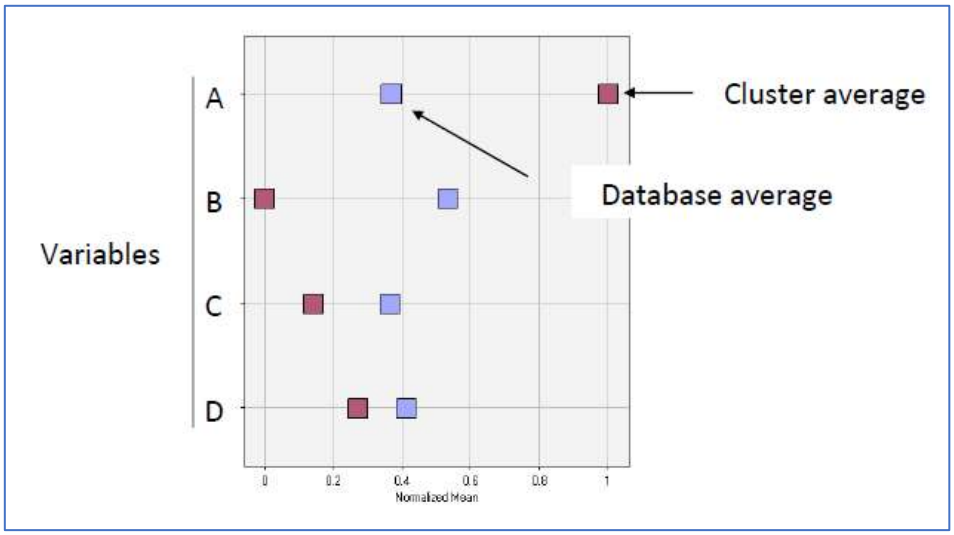


Figure 2.9 – Example of the profiling method.

Source: (Bação, F.L. 2019)

3. SYSTEM DEVELOPMENT & IMPLEMENTATION

In this chapter, it has been presented the methodology followed during the development of this project. Starting with the extraction and processing of the data, the assurance of the patient's data privacy, followed by an explanation of the whole system's structure.

3.1. OPEN DATA COLLECTION & PROCESSING

3.1.1. Geodemographic data

The geodemographic data has been obtained from two different sources:

- Spanish National Institute of Statistics (INE): Two datasets have been downloaded from the website of this institution. The first one contains the geometry of all the census sections of Spain and most of the data related to the population presented in this project, and a second one containing the average incomes per person.
    i. The first dataset [1] is made up of a set of ".csv" files containing the value of the variables per census section, another ".xls" with the metadata, and a ".shp" file containing the geometry of the census sections and some features related to them such as the identification codes, province, municipality, and others. Every variable presented in this dataset can be found in annex 10.1.
    ii. The second dataset[2] contains the average incomes per home and per person for every census section of Andalusia, Spain. For this study, the average net incomes per person have been chosen due to the non-dependence with the number of working members living at the same house.

- Andalusian Institute of Statistics and Cartography (IECA): Two datasets have been downloaded from this institution. The first one contains the unemployment rate for every census section and the second one contains a grid with some variables such as the mortality rate.
    i. The first dataset[3] is made up of a set of ".csv" files that correspond to the unemployment rates of every census section and for the years between 2016 and 2019. In order to simplify the dataset and get a single variable for the project, the average rates have been calculated for every section.
    ii. The second dataset[4] provides a grid of 250x250 meters with a 95% confidence interval of the mortality rate. Applying the function "overlay" of geopandas and the average value, is possible to obtain the mean value for each census section.

[1] Retrieved from www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm
[2] Retrieved from www.ine.es/dynt3/inebase/es/index.htm?padre=7132
[3] Retrieved from www.juntadeandalucia.es/institutodeestadisticaycartografia/poblacion_registros
[4] Retrieved from www.juntadeandalucia.es/institutodeestadisticaycartografia/

The variable “Population Density” has been created based on the “Total Population” variable and the area of each census section. Moreover, the variables that explain the distribution of gender, age, nationality, education, properties, and homes have been transformed into a percentage in order to standardize as much as possible the variables. It has been widely demonstrated how the scale in which the values of the variables are, notably affects during the application of algorithms.

The variables that have been considered important for the purpose of this project have been chosen and processed. All of them can be found in the following table and their distribution can be observed in annex 10.2:

Category	Name	Description	Type	Unit
POPULATION	POB1	Males	QUANTITATIVE	%
	POB2	Females	QUANTITATIVE	%
	POB3	People under 16	QUANTITATIVE	%
	POB4	People between 16-64	QUANTITATIVE	%
	POB5	People over 64	QUANTITATIVE	%
	POB6	People born in Spain	QUANTITATIVE	%
	POB7	People born outside of Spain	QUANTITATIVE	%
	POB8	Population density (nº/km2)	QUANTITATIVE	NUMBER
EDUCATION	EDU1	People with no studies	QUANTITATIVE	%
	EDU2	People with middle studies	QUANTITATIVE	%
	EDU3	People with complete studies	QUANTITATIVE	%
PROPERTIES	APA1	Principal Properties	QUANTITATIVE	%
	APA2	Secondary Properties	QUANTITATIVE	%
	APA3	Empty Properties	QUANTITATIVE	%
	APA4	Properties Density (nº/km2)	QUANTITATIVE	NUMBER
HOMES	HOM1	Homes of 1-2 people	QUANTITATIVE	%
	HOM2	Homes of 3 or more people	QUANTITATIVE	%
	HOM3	Homes Density (nº/km2)	QUANTITATIVE	NUMBER
LIVING CONDITIONS	LIF1	Average Net Incomes per person	QUANTITATIVE	€
	LIF2	Unemployment Rate	QUANTITATIVE	%
	LIF3	Mortality Rate	QUANTITATIVE	NUMBER

Table 3.1 – Selected and processed variables of the project

All these variables and the census section information (codes, names, and geometry) have been merged into a single dataframe.

Few missing values have been found in the dataset and they have been imputed with the average value of the census sections surround them. This decision has been taken due to the small number of missing values and assuming that two adjacent census sections have similar values with each another. In the following chapters, this methodology will be explained deeper.

3.1.2. Environmental data

Three types of environmental data have been integrated into the system, coming from three different sources:

- Ministry for the ecological transition and the demographic challenge: On the website of this institution is possible to find many datasets such as the historic pollution data measured by the pollution stations geographically distributed among the country[5]. The data have been analyzed and processed, assigning to each value ($\mu g/m^3$) the station and its coordinates in order to be represented on the map. An example of the data structure and its representation on the system can be found in the following figures:

	Name text	Province text	Municipality text	Date timestamp without time	SH2 double precision	SO2 double precision	NO2 double precision	CO double precision	PART double precision	O3 double precision	geometry geometry
1	AGUA AMARGA	ALMERIA	NIJAR	2010-01-01 00:00:00	[null]	3.1	9.3	[null]	13.3	55.6	0101000020E6100000154CBFD081F5FEB...
2	AGUA AMARGA	ALMERIA	NIJAR	2010-01-02 00:00:00	[null]	3	11.1	[null]	10.4	44.6	0101000020E6100000154CBFD081F5FEB...
3	AGUA AMARGA	ALMERIA	NIJAR	2010-01-03 00:00:00	[null]	3.2	11.5	[null]	11.2	41.9	0101000020E6100000154CBFD081F5FEB...
4	AGUA AMARGA	ALMERIA	NIJAR	2010-01-04 00:00:00	[null]	4.1	13.8	[null]	14.3	44.8	0101000020E6100000154CBFD081F5FEB...
5	AGUA AMARGA	ALMERIA	NIJAR	2010-01-05 00:00:00	[null]	4.2	10.2	[null]	8.5	62.1	0101000020E6100000154CBFD081F5FEB...
6	AGUA AMARGA	ALMERIA	NIJAR	2010-01-06 00:00:00	[null]	3	11.4	[null]	13.4	44.1	0101000020E6100000154CBFD081F5FEB...
7	AGUA AMARGA	ALMERIA	NIJAR	2010-01-07 00:00:00	[null]	3.2	13.4	[null]	7.6	45.5	0101000020E6100000154CBFD081F5FEB...
8	AGUA AMARGA	ALMERIA	NIJAR	2010-01-08 00:00:00	[null]	3.5	11	[null]	4.8	51	0101000020E6100000154CBFD081F5FEB...
9	AGUA AMARGA	ALMERIA	NIJAR	2010-01-09 00:00:00	[null]	2.9	10.7	[null]	6.4	50.4	0101000020E6100000154CBFD081F5FEB...

Figure 3.1 – Example of the pollution data structure

Source: PgAdmin. BBDD of Geohealth

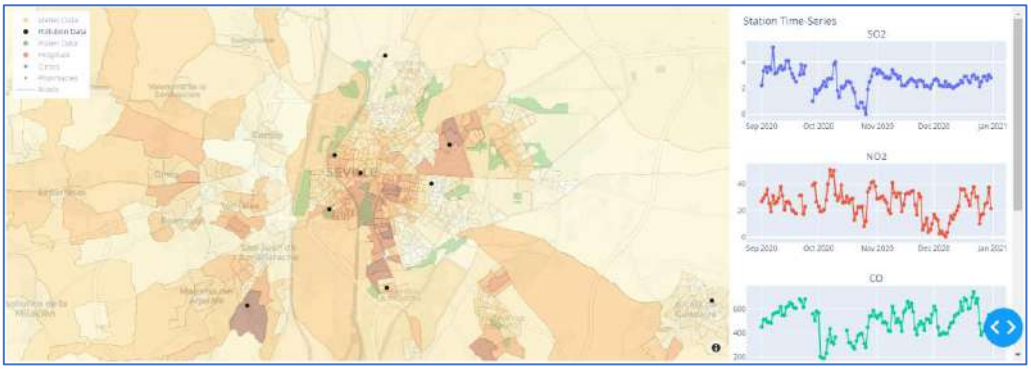
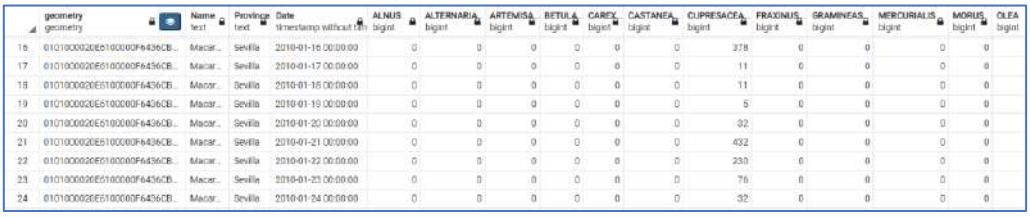


Figure 3.2 – Example of the pollution data representation. Black points represent the pollution stations

Source: Geohealth

[5] Retrieved from https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/

- Spanish Society of Allergology and Clinic Immunology (SEAIC): This institution provides data from different pollen stations spread among the country[6]. The data has been analyzed and processed, as pollution data has been done, assigning to each value (grain/m^3) the station and its coordinates. An example of the data structure can be found in the following figures and its representation on the system:



	geometry geometry	Name text	Province text	Date timestamp without time	ALNUS bigint	ALTERNARIA bigint	ARTEMISIA bigint	BETULA bigint	CAREX bigint	CASTANEA bigint	CUPRESACEA bigint	FRAXINUS bigint	GRAMINEAS bigint	MERCURIALIS bigint	MORUS bigint	OLEA bigint
16	0101000020E6100000F6436CB...	Macar...	Sevilla	2010-01-16 00:00:00	0	0	0	0	0	0	378	0	0	0	0	
17	0101000020E6100000F6436CB...	Macar...	Sevilla	2010-01-17 00:00:00	0	0	0	0	0	0	11	0	0	0	0	
18	0101000020E6100000F6436CB...	Macar...	Sevilla	2010-01-18 00:00:00	0	0	0	0	0	0	11	0	0	0	0	
19	0101000020E6100000F6436CB...	Macar...	Sevilla	2010-01-19 00:00:00	0	0	0	0	0	0	5	0	0	0	0	
20	0101000020E6100000F6436CB...	Macar...	Sevilla	2010-01-20 00:00:00	0	0	0	0	0	0	32	0	0	0	0	
21	0101000020E6100000F6436CB...	Macar...	Sevilla	2010-01-21 00:00:00	0	0	0	0	0	0	432	0	0	0	0	
22	0101000020E6100000F6436CB...	Macar...	Sevilla	2010-01-22 00:00:00	0	0	0	0	0	0	230	0	0	0	0	
23	0101000020E6100000F6436CB...	Macar...	Sevilla	2010-01-23 00:00:00	0	0	0	0	0	0	76	0	0	0	0	
24	0101000020E6100000F6436CB...	Macar...	Sevilla	2010-01-24 00:00:00	0	0	0	0	0	0	32	0	0	0	0	

Figure 3.3 – Example of the pollen data structure

Source: PgAdmin: BBDD of Geohealth

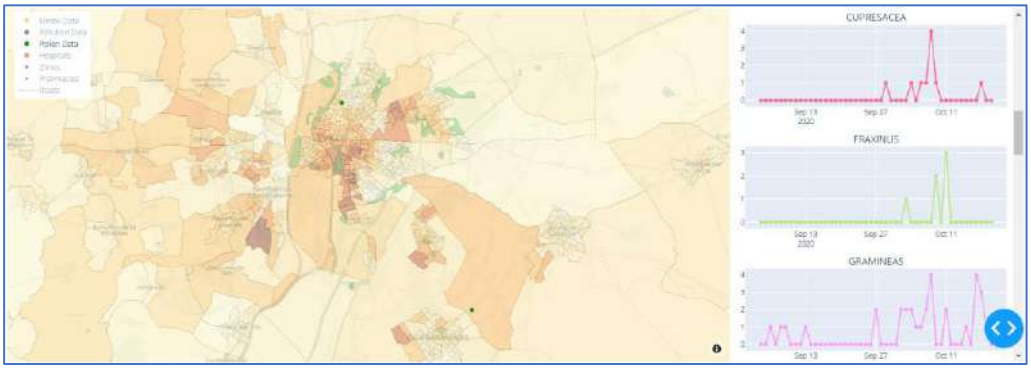


Figure 3.4 – Example of the pollen data representation. Green points represent the pollen stations

Source: Geohealth

- State Meteorological Agency of Spain (AEMET) and Agroclimatic Information Network of Andalusia (RIA) [7]: The two institutions provide the meteorological data integrated into the system. Both datasets have been analyzed, processed, and merged into a single dataset and the coordinates of the stations have been assigned to each one. An example of the data structure can be found in the following figure and its representation on the system:

[6] Retrieved from https://www.polenes.com/es

[7] Aemet retrieved from https://opendata.aemet.es/centrodedescargas/ and RIA retrieved from https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaws/swagger-ui.html

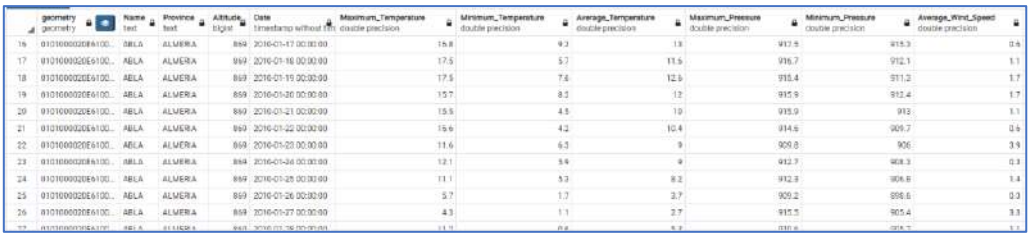


Figure 3.5 – Example of the meteorological data structure

Source: PgAdmin: BBDD of Geohealth

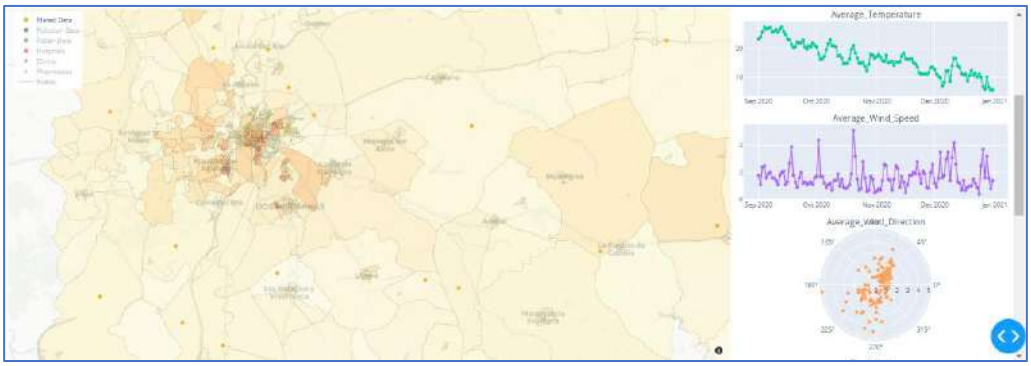


Figure 3.6 – Example of the meteorological data representation. Orange points represent the meteorological stations

Source: Geohealth

3.1.3. Andalusian Services Data

The Andalusian Institute of Statistics and Cartography (IECA) provides a set of ".shp" files[8] containing the information and location of some services of Andalusia (Spain) such as hospitals, clinics, pharmacies, and others. It has been created one dataset for each service in the system's database and in the following figures can be observed an example of the data structure and its representation on the system:

	Name text	Municipality text	Province text	Type text	Health_System text	geometry geometry
1	Hospital de la Cruz Roja de Almería	Almería	Almería	Hospital	SSPA	0101000020E61000002C3546A76F9E03C05EEB3...
2	Hospital de Especialidades Torrecárdenas	Almería	Almería	Hospital	SSPA	0101000020E6100000801111CC91E8703C0E603A...
3	Hospital el Toyo (Chare)	Almería	Almería	Hospital	SSPA	0101000020E61000008583AEB84D8002C026E11...
4	Hospital Sierra Norte de Sevilla (Chare)	Constantina	Sevilla	Hospital	SSPA	0101000020E61000002D603568177A16C06262E...
5	Hospital Comarcal Punta de Europa	Algeciras	Cádiz	Hospital	SSPA	0101000020E6100000021EF88D6D5C615C0D15B0...
6	Hospital de Especialidades Puerta del Mar	Cádiz	Cádiz	Hospital	SSPA	0101000020E61000002002D20AF01C19C099C79...

Figure 3.7 – Example of the hospital's data structure.

Source: PgAdmin: BBDD of Geohealth

[8] Retrieved from https://www.juntadeandalucia.es/institutodeestadisticaycartografia/DERA/g12.htm

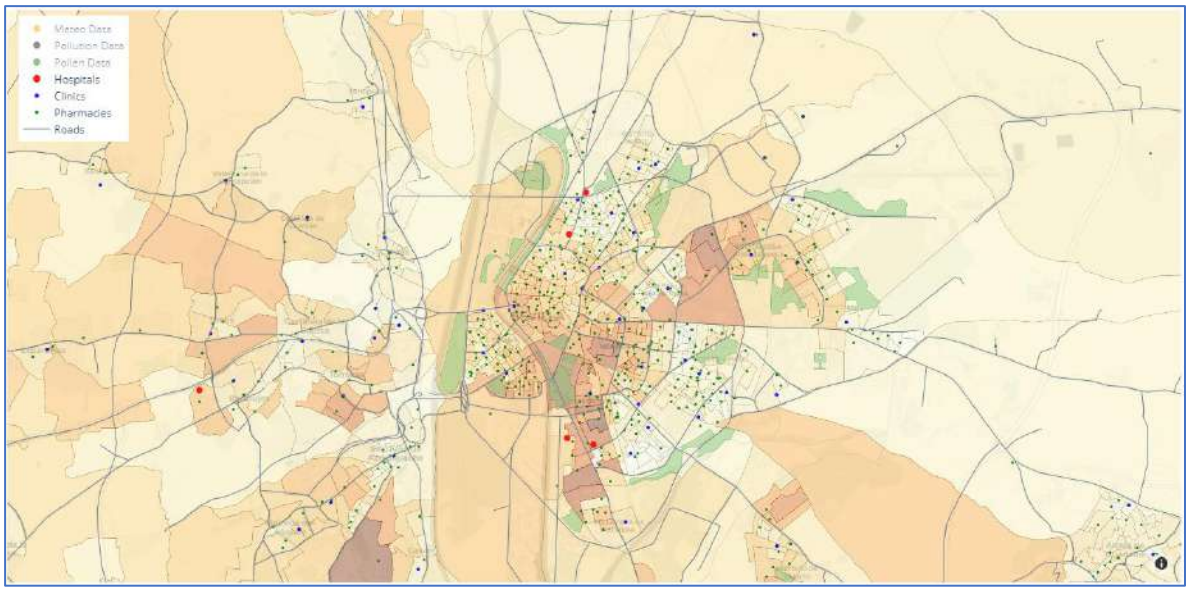


Figure 3.8 – Example of the services' data representation. Green points represent pharmacies, red points represent hospitals, blue points represent the clinics, and the lines represent the main roads.

Source: Geohealth

3.2. PATIENT'S INFORMATION PRIVACY

With the objective of not compromising at any time the privacy of the clinical and personal data of the patients, different strategies have been applied during the development and representation of the information into the system.

- Creation of a unique ID to identify each patient without showing their true health identification number.
- To carry out studies, the exact location of the patients' home has been replaced by the census section in which their home is located, assuming that the environment variables will be the same for all the area and because it is the minimum territorial unit for the statistical measurement of demographic information.
- Moreover, all the system's data is stored in a server of Virgen Macarena Hospital, with all the security measures guaranteed by the Andalusian health system.

3.3. STRUCTURE OF THE DATABASES AND SYSTEM

This chapter describes how is the databases structures and how is the communication between databases and the system. The project aims to expand the system to the different units of the hospital; thus, it needs to be adaptable according to the information that each study requires. Therefore, both the structure of the database and the system have been developed for that purpose.

3.3.1. Databases structure

A set of databases are necessary for the correct operation of the system. Each of them contains different sets of data as can be observed in the following figure and list:

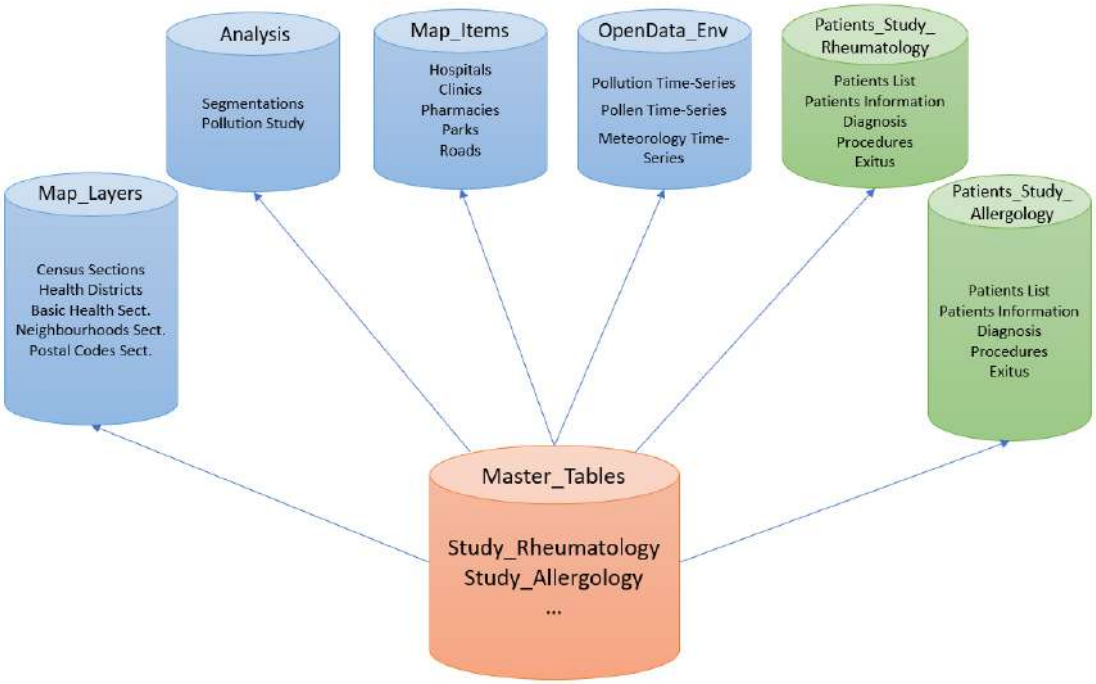


Figure 3.9 – Databases' structure

Source: Author

- "Map_Layers" Database: In this database are stored the different layers that are possible to be represented on the map. A set of layers such as census sections, postal code sections, health districts sections and others can be selected by the users according to their needs. Each one of them has incorporated the demographic information of the sections provided by INE and IECA.
- "Analysis" Database: This database stored the two studies developed in the project and it will contain the future studies and analysis that the Innovation Area of the hospital will develop.
- "Map_Items" Database: A set of services provided by IECA can be represented into the map such as hospitals, clinics, pharmacies, and others.
- "OpenData_Env" Database: This database stored the environmental data coming from the different institutions described in the previous chapter. Three tables make up this database storing the pollution, pollen, and meteorological time series and the location of the different stations.
- "Patients_Study_*" Databases: Each study has a database where is stored the data of the patients included on it. When a hospital's unit or researcher requests a new study, a new database is created especially for it and it is made up of different tables such as diagnosis, procedures, patients' information, and others, according to the needs of the study.

- "Master_Tables" Database: This database contains a set of tables that are key for communication with the dashboard. Each study has a master table that contains the list of layers, items, and data required by the study. As an example, in the following table is presented the master table used by the Rheumatology Unit of the hospital in one of their studies:

BBDD	Tables
Map_Layers	Census
Map_Layers	Basic Health sections
Map_Items	Hospitals
Map_Items	Pharmacies
Map_Items	Clinics
Analysis	Segmentations
Analysis	Pollution Study
OpenData_Env	Meteorology
Patients_Study_Rheumatology	** All **

Table 3.2 – Example of the Rheumatology study's master table

3.3.2. Communication BBDD-Dashboard

The dashboard knows what information needs to be loaded, for a particular study, by reading the corresponding master table. To access the study, the unit will use a personal URL provided by the team. This URL includes the name of the master table from which the dashboard extracts the data.

An example of this URL could be: "http://IP:Port/Study_Allergology".

The dashboard uses the component "Location" from the library "dash_core_components". Using a callback function, it is possible to extract the string from the URL and use it to perform the necessary queries to extract the data from the databases.

3.4. GRAPHIC MANUAL OF THE SYSTEM & STUDY CASES

In this chapter, some images of the system are exposed, as well as an explanation of the different functionalities. In addition, the two modules that contain the results of the two study cases developed in this project are presented. The introduction, methodology and analysis of the results of these two study cases can be found in the next two chapters. Moreover, a database provided by the allergology department will be used as an example to show the potential value and usability of this system for conducting clinical studies.

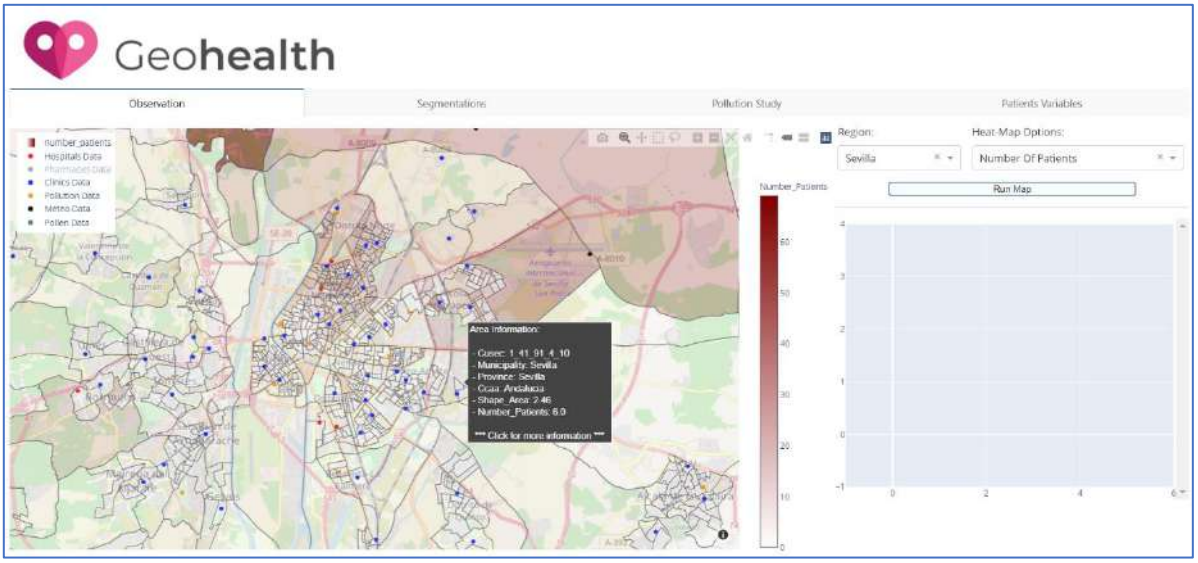
3.4.1. "Observation" Module

This module aims to work as an observation space where researchers can explore the information included in the study. All the information (such as demographic information, pollution data, meteorological data, pollen data, hospitals, and others) is integrated into a single map which allows users to interact with the data and perform selections for deeper observations and analysis of the information presented in the selected section. Two variables need to be selected to create the map: a specific region and a variable used to create a heat-map of the census sections.

Every item presented in the map is clickable and the information related to it will be shown on the right side of the map. According to the clicked item, different options are provided:

- Census section: The demographic information related to the clicked section or sections can be represented in two forms: In a table or graphs showing the value of each variable.
- Pollution, Pollen, or Meteorological station: The time series of the information captured by the station will be shown, allowing the researcher to choose a range of dates.
- Clinics, hospitals, and pharmacies: The information of the clicked item will be represented in a table.

In Figure 3.10 is presented the observation module for the region of Seville (Spain), being the variable selected for the heat-map, the number of patients per census section. Two images of the system are represented: the first one is the map before any item is clicked and the second one after the selection of the census sections of the neighborhood of the Virgen Macarena Hospital. Other examples of the possibilities provided by the system can be found in annex 9.6.



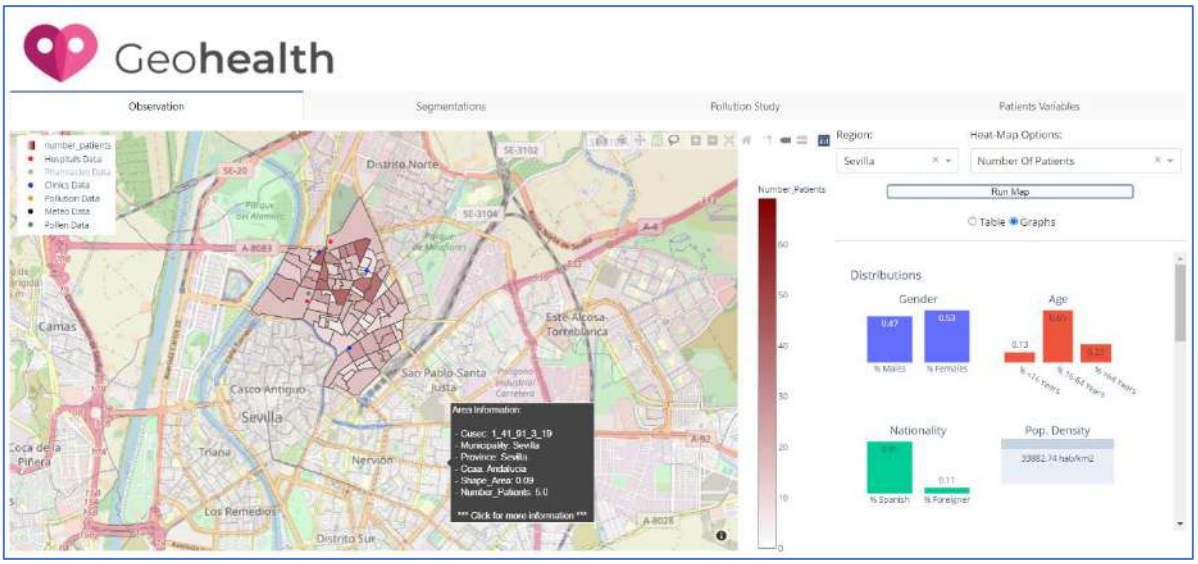


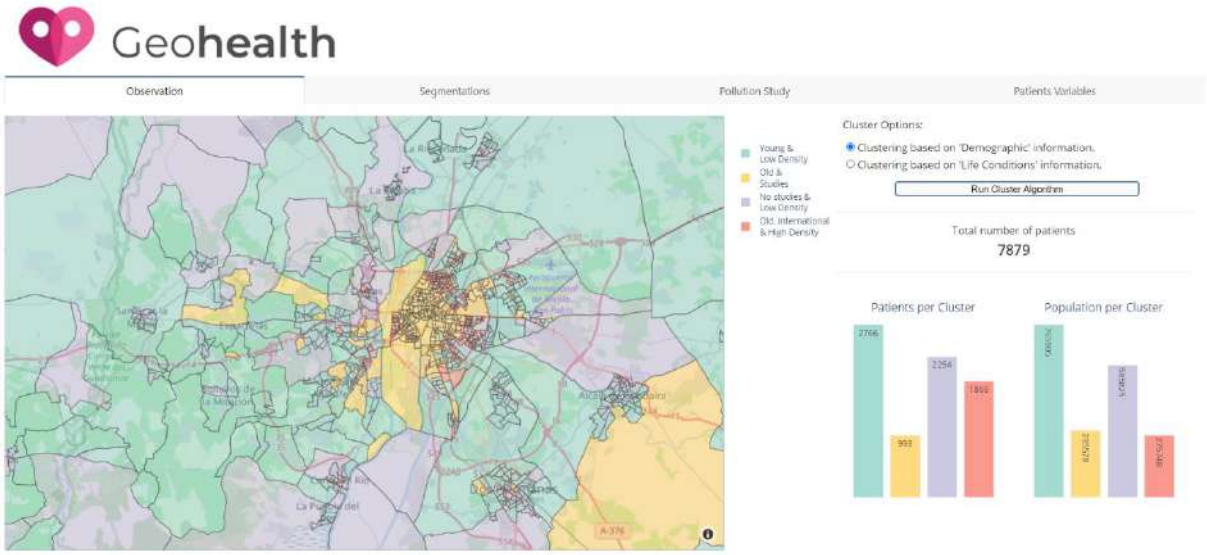
Figure 3.10 – "Observation" Module

Source: Geohealth

3.4.2. "Segmentations" Module

This module contains the two segmentations developed in this project and they are represented into a map. Furthermore, the distribution of the patients and the distribution of the population among the clusters, are shown into bar graphs to understand how the population of patients presented in the study is. At the bottom of the module, the clusters of the segmentation are explained using the mean value characterization method, as well as a table where a brief analysis has been written.

In Figure 3.11 is presented the "Segmentations" module for the region of Seville (Spain). The represented segmentation is the "demographic" information segmentation, and the distribution of the patients is shown in the bar graph. The images of the "Living Conditions" segmentation can be found in annex 9.7.



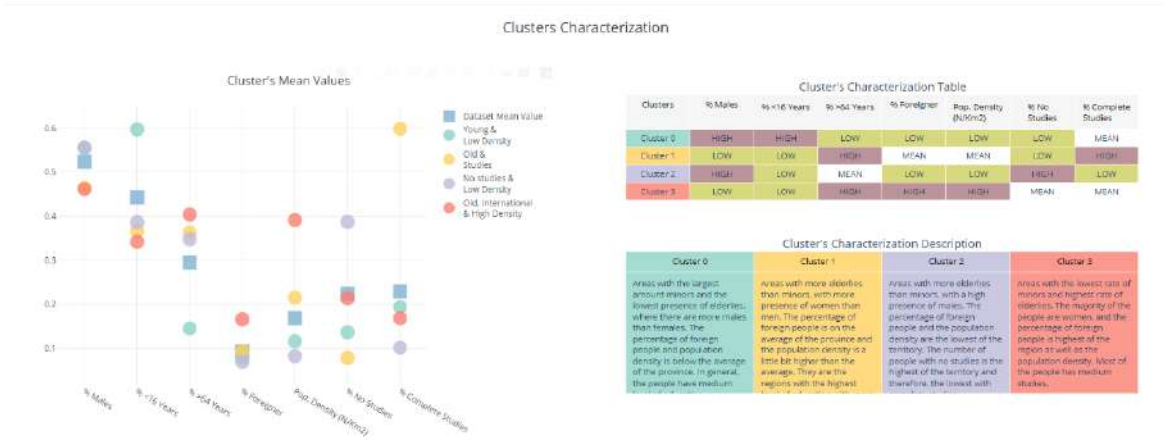


Figure 3.11 – "Segmentation" Module

Source: Geohealth

As it can be observed in figure 3.11, the patients included in the allergology study are distributed between the four clusters of the segmentation. Most of them fall into the "Young and Low Density", corresponding to the majority of the towns of the province of Seville, and some external areas of the city where the majority of the population are young people, the population density is low and in general, people have medium studies. Virgen Macarena Hospital attends the population living in the north of the city of Seville and the whole north of the province. Therefore, the large number of patients living in this cluster can be related to this fact. Moreover, it is the cluster with the largest population.

The second highest cluster is the "No Studies & Low Density" cluster. The most remarkable characteristic of this group is the large presence of elderlies and the low level of education presented in its population. The fact that it is the second-largest cluster of the segmentation and the characteristics and location (rural areas of the north of the province) of the census sections make it the second-largest cluster of patients.

The cluster with the lowest number of patients is the "Old & Studies" cluster. The population of this cluster is old and with high studies. The reason why it has a low number of patients may be because is one of the two clusters with the lowest population and most of the census sections fall into the area of the other big hospital of the city of Seville.

3.4.3. "Pollution Study" Module

The maps of the pollution studies have been integrated into this module with the aim of allowing researchers to understand the pollution conditions, of a specific date or range of dates, that suffer the patients of the study. Two options are provided:

- Create a new study: It is possible to represent the conditions of a specific date by the AQHI scale or the score calculated, for a range of dates, by the formula proposed in the methodology chapter. Moreover, the distribution of the patients among the different values of the conditions is shown in a bar graph. Finally, at the bottom of the module, the table with the information related to each condition is showed.

- Default studies: The monthly scores' studies developed in this project can be represented into the system and understand how the patients of the study are distributed among the different condition's categories.

In the following figures are presented two examples of studies developed by the "Pollution Study" module for the region of Seville (Spain). The first one is the map of the month of October from the second study developed in this project (Figure 3.12) and the second one is the map of the AQHI conditions of a single day (Figure 3.13). Other examples of the possibilities provided by the system can be found in annex 9.8.

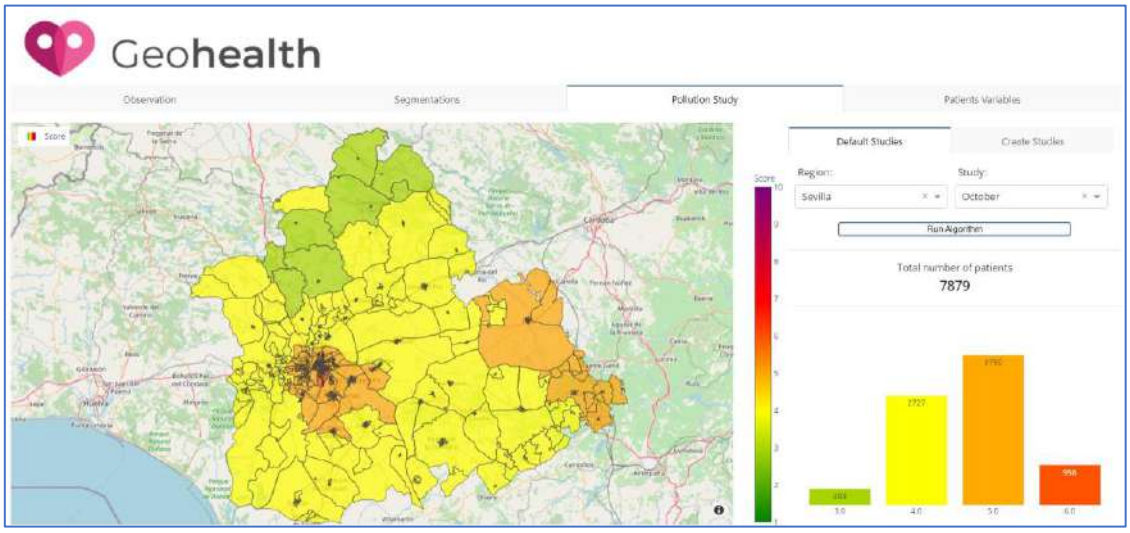
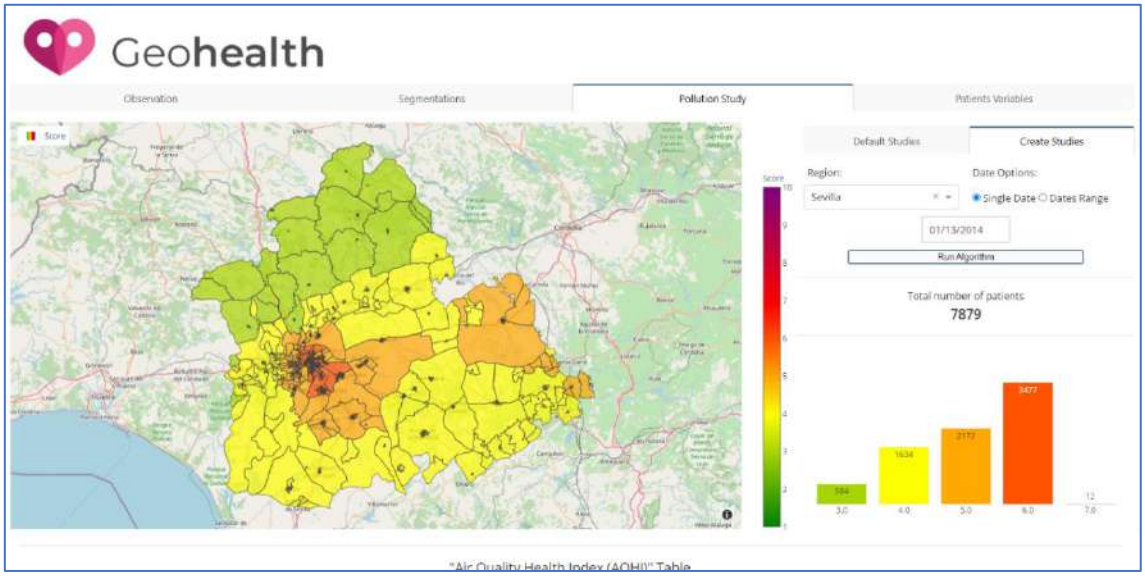
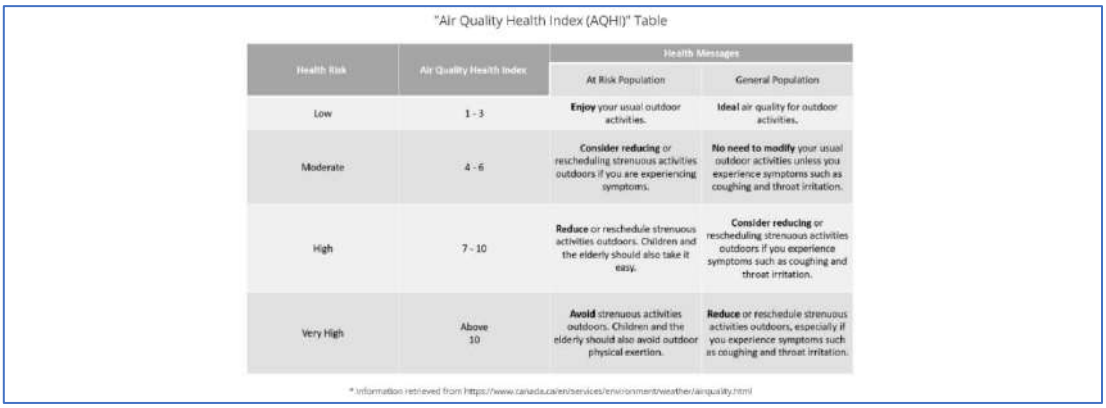


Figure 3.12 – "Pollution Study" Module. Study II October example

Source: Geohealth

In figure 3.13, the scores have been calculated using the formula proposed in the methodology chapter of this project and it is a scale between 1 to 10. It can be observed the distribution of patients among the different categories of air quality conditions for the month of October.





"Air Quality Health Index (AQHI)" Table

Health Risk	Air Quality Health Index	Health Messages	
		At Risk Population	General Population
Low	1 - 3	**Enjoy** your usual outdoor activities.	**Ideal** air quality for outdoor activities.
Moderate	4 - 6	**Consider reducing** or rescheduling strenuous activities outdoors if you are experiencing symptoms.	**No need to modify** your usual outdoor activities unless you experience symptoms such as coughing and throat irritation.
High	7 - 10	**Reduce** or reschedule strenuous activities outdoors. Children and the elderly should also take it easy.	**Consider reducing** or rescheduling strenuous activities outdoors if you experience symptoms such as coughing and throat irritation.
Very High	Above 10	**Avoid** strenuous activities outdoors. Children and the elderly should also avoid outdoor physical exertion.	**Reduce** or reschedule strenuous activities outdoors, especially if you experience symptoms such as coughing and throat irritation.

* Information retrieved from https://www.canada.ca/en/services/environment/weather/airquality.html

Figure 3.13 – "Pollution Study" Module. AQHI conditions on 13th of January of 2014.

Source: Geohealth

In figure 3.13, the AQHI has been calculated using the formula explained in the methodology chapter of this project and it is a scale between 1 to 10. It can be observed the distribution of patients among the different categories of air quality conditions on 13th of January of 2014 and the AQHI table explain what the recommendation is according to the index of each census section.

3.4.4. "Patients Variables" Module

This module provides a useful tool for researchers to represent and understand the distribution of the different patient variables (such as procedures, diagnosis, and others). A set of filters allow the selection of the variables and the values required by the user and a graph shows the distribution. Moreover, an XY plot can be created to observe if exist a relationship between a pair of variables.

Figure 3.14 – "Patients Variables" Module. Example of the variables and XY plot

Source: Geohealth

In Figure 3.14 are presented the distribution of two variables presented in the patient's information dataset included in the study case. The first variable represents the diagnosis of asthma between the years 2010 and 2015 and the second variable represents the spirometry procedures performed between 2010 and 2021. Finally, the XY plot combines the two variables allowing researchers to observe the patients (represented by the points) and the dates when they were diagnosed with asthma and they did the spirometry test.

4. STUDY I. GEODEMOGRAPHIC SEGMENTATION

This chapter contains the methodology, data and technology used for the development of the Study Case I, which correspond to a geodemographic segmentation. The results obtained by the application of the methodology proposed have been presented and analyzed, and the characteristics of the different clusters discovered by two segmentations have been discussed. In addition, a graphic analysis of the different groups is developed to help understanding their characteristics. This study has been developed and integrated into the system to provide researchers with a simple and intuitive tool with which they can carry out studies and apply techniques of data mining such as clustering.

4.1. OBJECTIVE

The objective of this study is the development of a geodemographic segmentation of the province of Seville (Spain) where Virgen Macarena Hospital is located and where its patients live. In order to complete this objective, cluster analysis techniques are applied, and the methodology followed can be summarized in problem definition, exploratory analysis and data processing, feature selection, application of the algorithm, and analysis of the produced segmentations.

4.2. PROBLEM DEFINITION

4.2.1. Identify the objective of the segmentation

The objective of a segmentation is the identification of homogeneous groups of census sections that have similar characteristics. Two types of segmentation have been developed and called: “Population” Segmentation and “Living Conditions” segmentation.

4.2.2. Data collection and features selection

The dataset used for the development of the segmentation is the same one used for the system observation module and the content of this dataset has already been exposed in the previous chapter. The set of variables have been grouped into five categories according to the information about the section that they provide: Population Information, Education Information, Properties Information, Homes Information, Living Conditions Information. From these five categories, the variables have been chosen and distributed into the two types of segmentation that have been developed, according to what can be observed in the following figure:

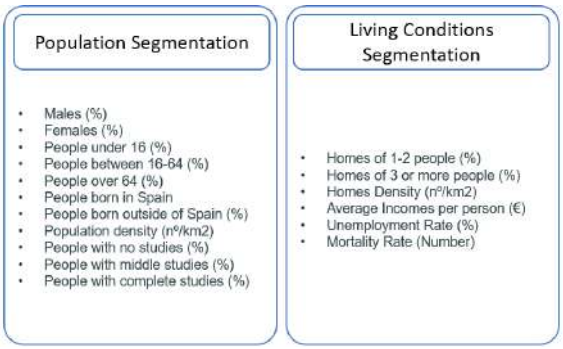


Figure 4.1 – Variables selected for each segmentation

Source: Author

4.3. EXPLORATORY ANALYSIS AND DATA PROCESSING

The segmentations have been carried out on the population of the province of Seville (Spain) using the information obtained from the census sections produced by INE. In the following list are presented some general parameters to understand better the analyzed region:

- Province: Seville
- Community: Andalusia
- Country: Spain
- Number of Municipalities: 105 municipalities
- Number of census sections: 1308
- Total population of the region: 1.923.620 habitants
- The area of the region is: 14.044 km^2

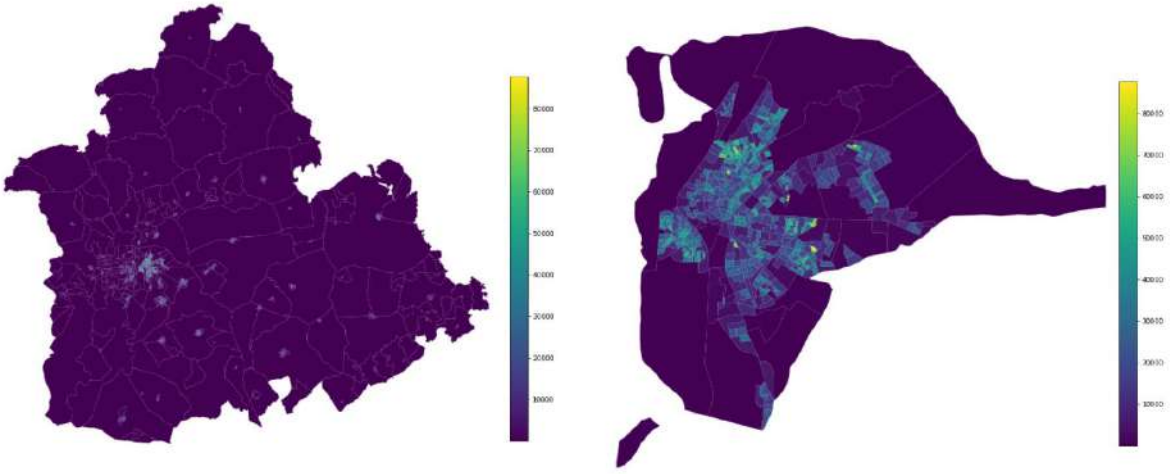


Figure 4.2 – (a) Province of Seville (Heat-map: Population Density); (b) Municipality of Seville (Heat-map: Population Density)

Source: Author

4.3.1. Creation of variables and imputation of missing values

As it was briefly explained in chapter 3.1.1, the variables had been selected from a dataset downloaded from the website of INE and IECA. During this process, some transformations were necessarily performed to group and simplify the data, standardize the data, and imputation some missing values. Some of these tasks are briefly explained in the following list:

- Creation of the variable "Population Density" per census section using the variables "Total Population" and the function "area" of Python for each census section's geometry. Using the formula

$$\frac{\text{“}Total\ Population\text{”}}{\text{”}Section\ Area\text{”}} \ \frac{(habitants)}{km^2},$$

the population density of each census section has been calculated.

- Some variables were summed and simplify into fewer variables such as the variables related to the educational level or the nationality. For example:
    - People with no studies = illiterates + Uneducated
    - People with middle studies = People with First-degree studies + Second degree
    - People with complete studies = People with tertiary studies
- Some missing values have been found and the strategy applied for their imputation has been the calculation of the average value of the adjacent sections to the one that has no value. It has been decided to apply this method due to the low number of missing values and under the assumption that two sections that are next to each other have a high probability of share similar values. In the following figure it can be observed the previous and after the imputation method for the variables EDU1 that represent the percentage of people with no studies:

Figure 4.3 – (a) Situation before the imputation (Part of Seville, Spain); (b) Situation after the imputation (Part of Seville, Spain); Missing values represented by red lines with grey background.

Source: Author

4.3.2. Univariate Analysis

The first step, to understand the dataset, study how the variables are distributed and know some characteristics of each variable such as mean, variance, minimum value, and others, is to develop a univariate analysis of each variable presented in the dataset. It has been developed in two steps: Descriptive statistics and visualization of histograms and boxplots to observe the presence of outliers.

1. In figure 4.4 it can be seen the descriptive statistics of the presented dataset (using the function “describe()” of pandas), from where it is possible to make a summary of the dataset by analyzing some of the variables. The dataset is made up of 1308 records corresponding to the census sections that form the province of Sevilla and the average population density of all the census sections is 14224 habitants/km2. The percentage of females is about 3% higher than the percentage of males and the percentage of minors is similar to the percentage of elderlies. Moreover, on average, 5% of the population living in a census section has born in another country and the percentage of people with complete studies is around 3% higher than the people with no studies. Finally, the average net incomes per person are around 9-10k € and the average unemployment rate is about 26% of the population.

	count	mean	std	min	25%	50%	75%	max
POB1	1308.0	0.488624	0.045545	0.30	0.4600	0.490	0.5200	0.66
POB2	1308.0	0.511422	0.045607	0.33	0.4800	0.510	0.5400	0.70
POB3	1308.0	0.172768	0.065728	0.00	0.1300	0.170	0.2125	0.39
POB4	1308.0	0.665894	0.072335	0.36	0.6200	0.670	0.7100	1.00
POB5	1308.0	0.162294	0.089488	0.00	0.0900	0.150	0.2200	0.55
POB6	1308.0	0.955015	0.062598	0.48	0.9400	0.980	1.0000	1.00
POB7	1308.0	0.057469	0.063264	0.01	0.0200	0.030	0.0700	0.52
POB8	1308.0	14762.023165	14224.394337	0.90	4326.0500	10557.550	21283.3250	87701.20
EDU1	1308.0	0.124182	0.084026	0.01	0.0500	0.110	0.1800	0.52
EDU2	1308.0	0.554213	0.102757	0.20	0.5000	0.560	0.6200	0.86
EDU3	1308.0	0.156430	0.132746	0.01	0.0600	0.110	0.2125	0.65
APA1	1308.0	0.810642	0.156953	0.06	0.7300	0.840	0.9400	1.00
APA2	1308.0	0.097271	0.089366	0.00	0.0400	0.070	0.1300	0.66
APA3	1308.0	0.168746	0.128074	0.00	0.0800	0.140	0.2200	0.83
APA4	1308.0	6993.912232	7002.730348	1.10	1908.2500	4794.900	10215.8250	36988.20
HOM1	1308.0	0.454855	0.135592	0.11	0.3600	0.450	0.5500	0.86
HOM2	1308.0	0.545130	0.135597	0.14	0.4500	0.550	0.6400	0.89
HOM3	1308.0	5645.739557	5621.087751	0.30	1520.1000	3843.500	8013.1500	30309.70
LIF1	1308.0	9379.173861	3210.050927	3326.00	7296.2500	8344.500	10686.2500	21525.00
LIF2	1308.0	26.648937	9.213846	8.83	20.3375	25.935	31.4400	62.69
LIF3	1308.0	1.046353	0.171531	0.57	0.9500	1.040	1.1100	2.75

Figure 4.4 – Descriptive statistics of the dataset.

Source: Author. Jupyter Notebook

2. In order to observe if exist values that could be considered outliers and be a problem for the development of this study, some visualization methods have been performed (the rest of the variables are represented in annex 10.3):

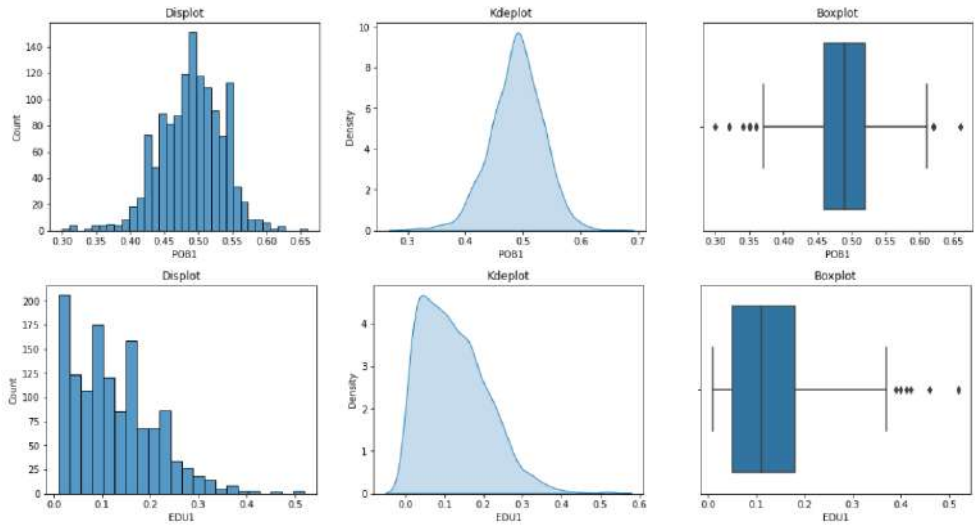


Figure 4.5 – Visualization of the distribution of the variable POB1 (% of males); Visualization of the distribution of the variable EDU1 (% of people with no studies).

Source: Author. Jupyter Notebook

This analysis concludes without finding any value far or very far from what is considered normal. Therefore, the application of outlier processing techniques is not necessary, and the values have been used for the segmentation.

4.4. Feature Selection

To reduce the dimensionality of the dataset and simplify the segmentation task, the correlation between the variables has been analyzed to eliminate the variables that are highly correlated with each other. It has been analyzed the correlation for both segmentations:

4.4.1. "Population" segmentation

The following figure presents the correlation matrix of the variables selected for the population segmentation. The type of correlation selected is Pearson because all the variable presented in this study are quantitative-continues:

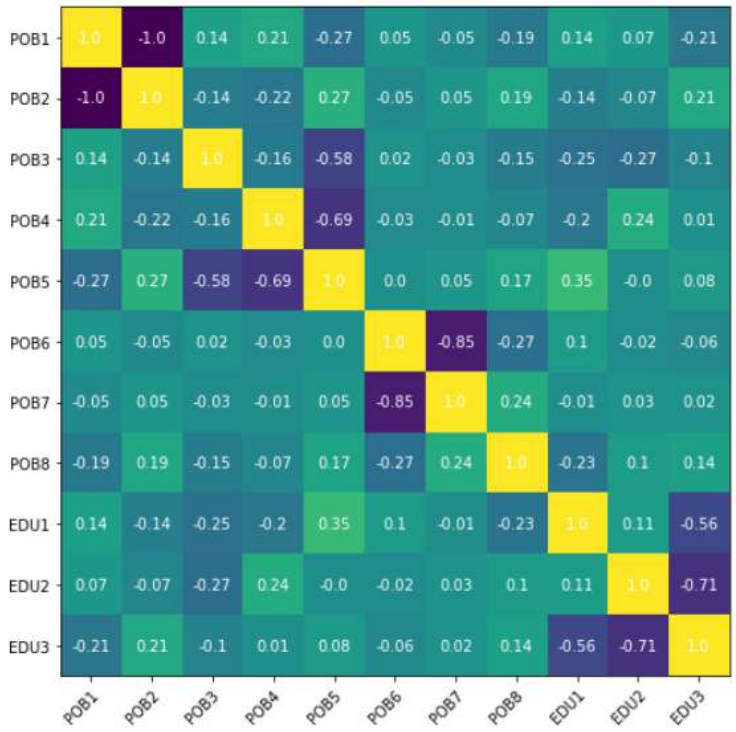


Figure 4.6 – Correlation Matrix of the variables selected for "Population" segmentation

Source: Author. Jupyter Notebook

Considering a high correlation between two variables more than 0.7 or -0.7, it can be observed that some of the variables are highly correlated with each other. In the following list are summarized the conclusions of this analysis:

- As it could be expected, variables POB1 and POB2 are completely correlated. This is logical due to the meaning of them, POB1 represents the percentage of males per census section and POB2 the percentage of females. One of these variables is selected for the study.
- Variables POB4 and POB5 are correlated with a value of -0.69. Then, variables POB3 and POB5 are chosen because they are considered more important for the study.
- POB6 and POB7 represent the percentage of people that were born in Spain or in another country. Then, the correlation between each other is high and one of them will be removed from the study.

- A similar situation happens with the education variables, in this case, EDU2 and EDU3 present a high correlation between each other. Then, EDU2 is removed, considering EDU1 and EDU3 more important for purpose of the study.

Finally, the selection of variables is presented in the following table:

Category	Name	Description	Type	Unit
POPULATION	POB1	Males	QUANTITATIVE	%
	POB3	People under 16	QUANTITATIVE	%
	POB5	People over 64	QUANTITATIVE	%
	POB7	People born outside of Spain	QUANTITATIVE	%
	POB8	Population density (nº/km2)	QUANTITATIVE	NUMBER
EDUCATION	EDU1	People with no studies	QUANTITATIVE	%
	EDU3	People with complete studies	QUANTITATIVE	%

Table 4.1 – Selected variables for the "Population" segmentation

4.4.2. "Living Conditions" segmentation

In the following figure is presented the correlation matrix of the variables selected for the "Living Conditions" segmentation. The type of correlation selected is Pearson because all the variable presented in this study are quantitative-continues:

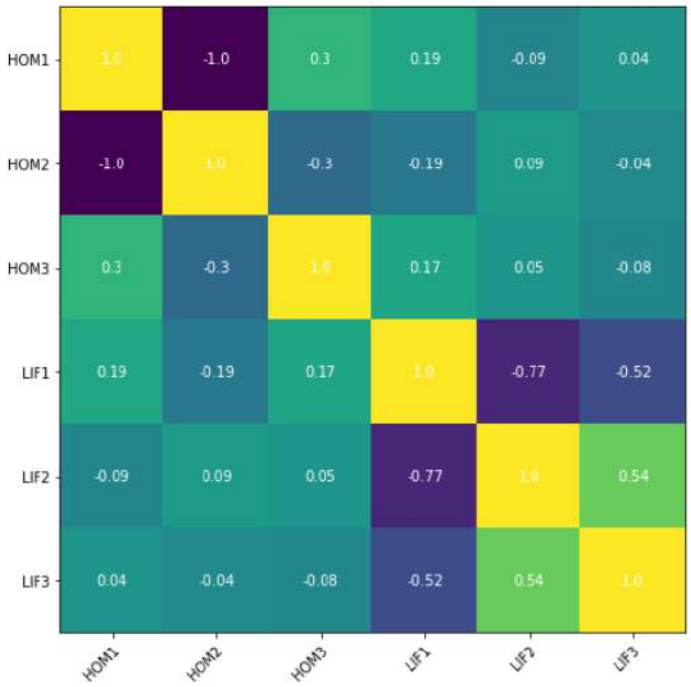


Figure 4.7 – Correlation Matrix of the variables selected for "Living Conditions" segmentation

Source: Author. Jupyter Notebook

As in the previous segmentation, considering a high correlation between two variables more than 0.7/-0.7, some of the variables that are highly correlated between each other will be removed from the study as it is summarized in the following list:

- Variables HOM1 and HOM2 represent the number of co-habitants per home divided into two groups, therefore, as it could be expected, the variables are completely correlated. Then, one of these variables is selected for the study.
- Variables LIF1 and LIF2 are correlated with a value of -0.77. LIF1 represents the average net income per person and LIF2 the unemployment rate per census section. Therefore, census sections with a high unemployment rate, are negatively related to the average incomes. Then, one of these variables is chosen for the study.

Finally, the selection of variables is presented in the following table:

Category	Name	Description	Type	Unit
HOMES	HOM1	Homes of 1-2 people	QUANTITATIVE	%
	HOM3	Homes Density (nº/km2)	QUANTITATIVE	NUMBER
LIVING CONDITIONS	LIF1	Average Net Incomes per person	QUANTITATIVE	€
	LIF3	Mortality Rate	QUANTITATIVE	NUMBER

Table 4.2 – Selected variables for the "Living Conditions" segmentation

4.5. APPLICATION OF THE ALGORITHM & ANALYSIS OF THE SEGMENTATIONS

Once the dataset has been processed and the features selected for each segmentation, a clustering algorithm has been applied to discover and understand the different homogeneous groups/clusters existing among the studied territory.

The chosen algorithm for the development of this study is one of the best-known algorithms for its great efficiency in exploring a dataset: K-means. As it has been explained in the theoretical framework chapter, the algorithm creates k groups from a set of observations/objects so that the members of each group are similar to each other.

4.5.1. "Population" segmentation

The variables selected for this segmentation have been standardized into the same scale (0-1) using the "MinMaxScaler()" function. The first step is to represent the elbow diagram corresponding to this dataset, in order to choose the appropriate number of clusters (Figure 4.8).

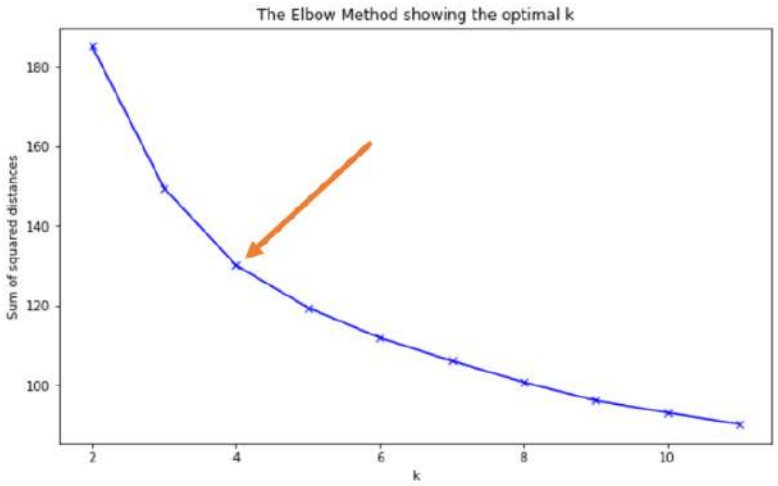


Figure 4.8 – Elbow Diagram of the "Population" Segmentation

Source: Author

Through the graph, it can be concluded that for this segmentation at least four distinct groups or clusters must be considered since it is verified at this point that the slope of the curve decreases for the following values of k. In order to not increase the complexity of the analysis and because the improvement would not be significantly better, the number of clusters chosen for this segmentation has been four.

The next step is the observation of the number of census sections that are contained in each cluster, being the objective of this phase to discover if exist a cluster with a low percentage of the sections falling into it. The analysis concludes without any modification due to the non-existence of a cluster with little representation of sections. A pie graph has been built in to help with this task as can be observed in Figure 4.9.

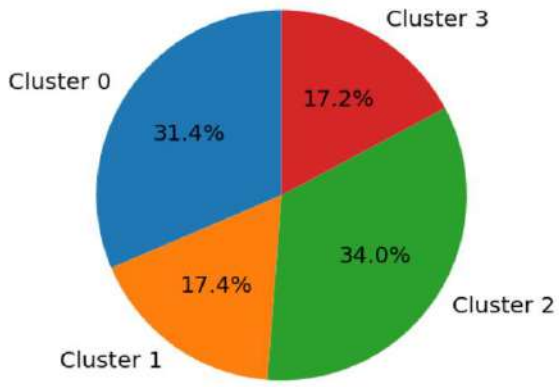


Figure 4.9 – Pie graph of the distribution of census sections among the clusters. "Population" Segmentation

Source: Author

Once the clusters are not going to be modified and they are considered definitive, it is time to move on to the next phase, called “Profiling”, where the main characteristics of each cluster will be analyzed to be useful for the researchers that use the system.

In the following figure are presented the values of the centroid of each cluster. These points are the mean value of each variable for each cluster and, therefore, its value will represent the entire group.

Clusters	POB1	POB3	POB5	POB7	POB8	EDU1	EDU3
Cluster 0	0.555236	0.385441	0.348088	0.070137	0.081928	0.389768	0.100754
Cluster 1	0.555621	0.595428	0.147813	0.079264	0.115788	0.138134	0.191148
Cluster 2	0.463876	0.339312	0.406523	0.164864	0.391616	0.212609	0.172786
Cluster 3	0.463618	0.370093	0.359372	0.092313	0.212790	0.076609	0.601148

Figure 4.10 – Mean values (centroids) of each cluster and each variable. “Population” Segmentation

Source: Author

The method selected for profiling the different clusters is to compare the mean of each variable and cluster with the mean of the entire dataset for each variable. This method is one of the most used and has helped to define the different clusters of this segmentation. The graph used for the development of this task is represented in the following figure:

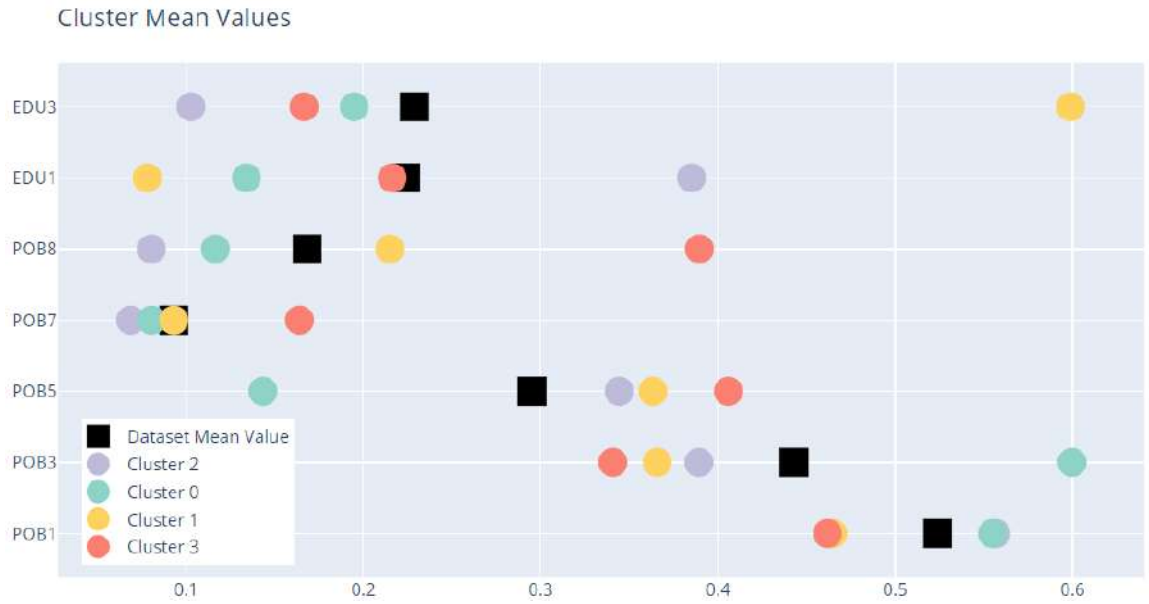


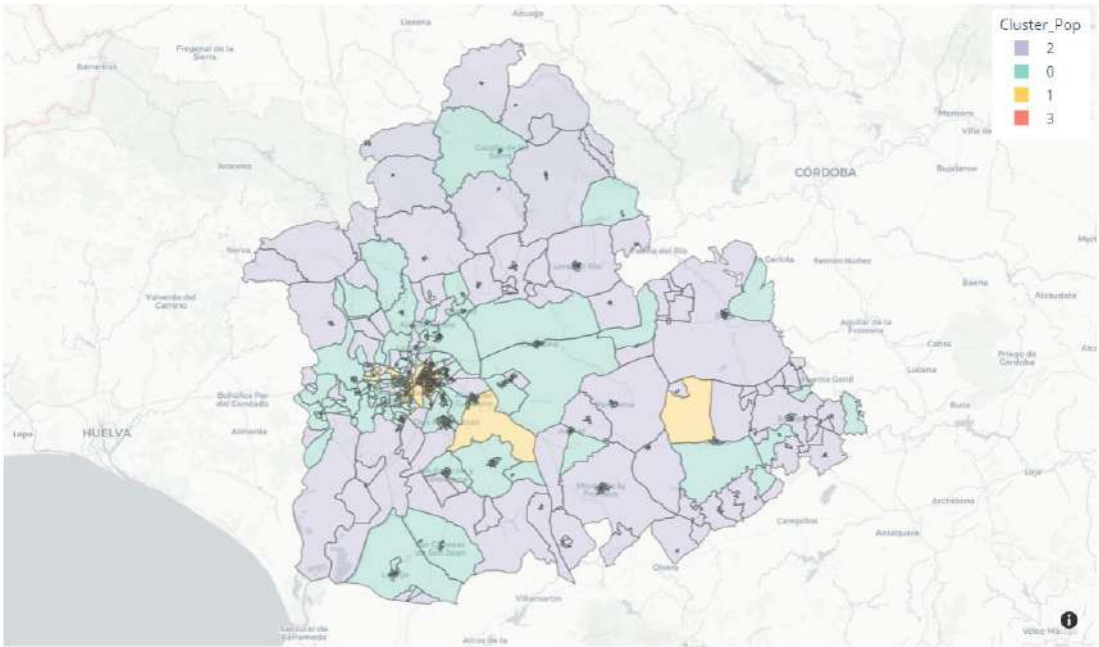
Figure 4.11 – “Mean value” profiling. “Population” Segmentation

Source: Author

Due to the analysis of the previous graph, it is possible to conclude the characterization of the different clusters of the segmentation, as shown in the following list:

- **Cluster 0: "Young & Low Population Density" Cluster.** Areas with the largest amount minors and the lowest presence of elderlies, where there are more males than females. The percentage of foreign people and population density is below the average of the province. In general, the people have a medium level of education.
- **Cluster 1: "Old & Studies" Cluster.** Areas with more elderlies than minors and with more presence of women than men. The percentage of foreign people is on the average of the province and the population density is a little bit higher than the mean. They are the regions with the highest level of education with more than half of the population with complete studies and where only a few of them have no studies.
- **Cluster 2: "No studies & Low Population Density" Cluster**. Areas with more elderlies than minors and with a high presence of males. The percentage of foreign people and the population density are the lowest of the territory. The number of people with no studies is the highest of the region and therefore, the lowest with complete studies.
- **Cluster 3: "Old, international & High Population Density" Cluster**. Areas with the lowest rate of minors and highest rate of elderlies. Most of the people are women, and the percentage of foreign people is the highest of the region as well as the population density. Moreover, most of the people have medium studies.

The following figure shows the distribution of the clusters over the territory of the province of Seville (Spain), as well as zoom in the capital of region:



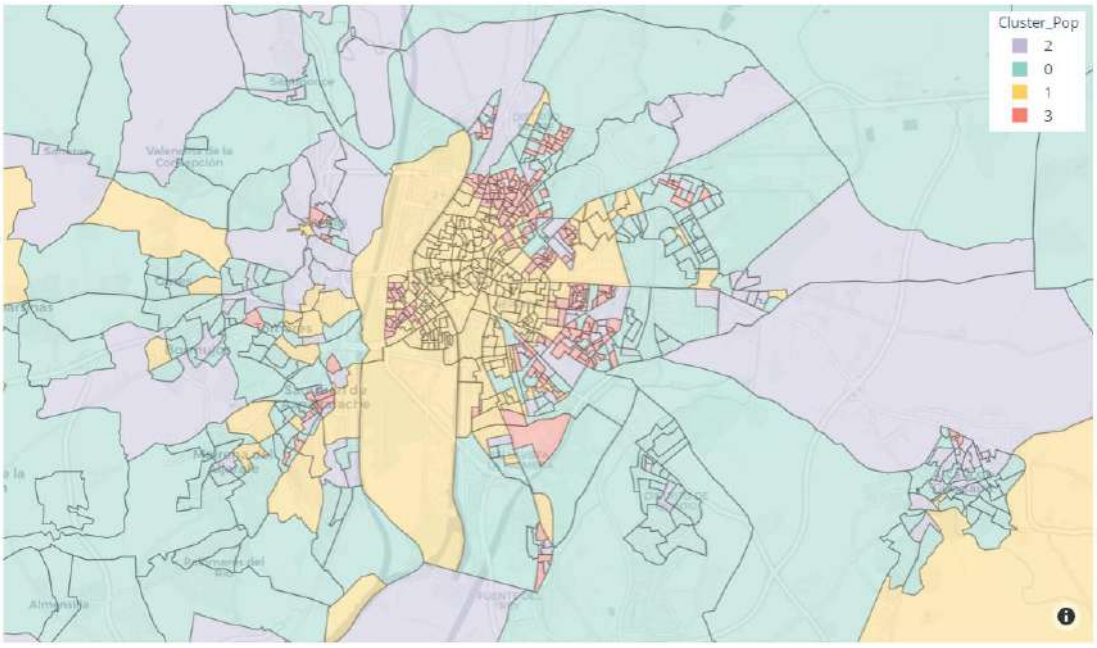


Figure 4.12 – Clusters among the territory. “Population” Segmentation

Source: Author

4.5.2. “Living Conditions” segmentation

The variables selected for this segmentation have been standardized into the same scale (0-1) using the “MinMaxScaler()” function. As in the previous segmentation, the first step is to represent the elbow diagram corresponding to this dataset to choose the appropriate number of clusters (Figure 4.13).

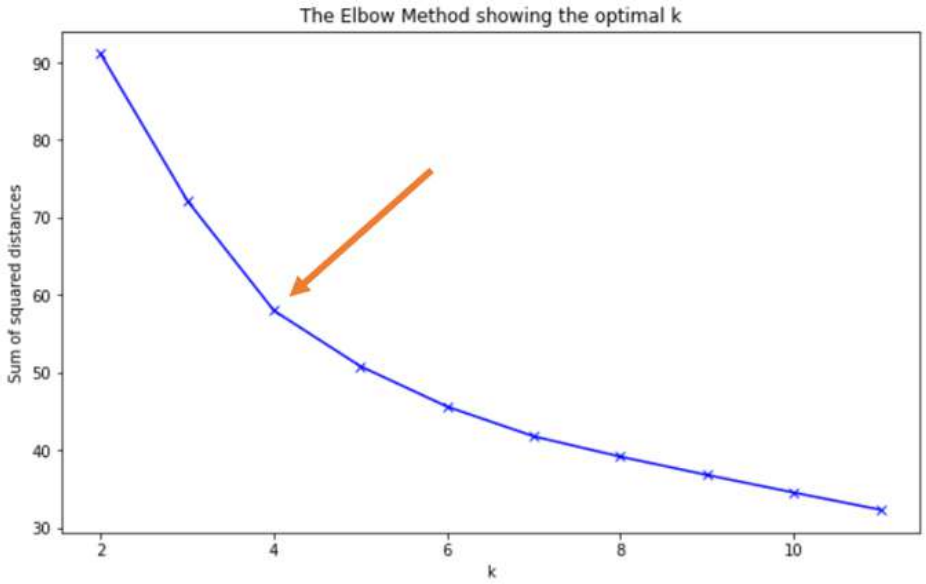


Figure 4.13 – Elbow Diagram for “Living Conditions” Segmentation

Source: Author

Through the graph, it can be concluded that for this segmentation at least four distinct clusters must be considered since it is verified at this point that the slope of the curve decreases for the following values of k. As in the previous one, the number of clusters chosen for this segmentation has been four.

The next step is the observation of the number of census sections that are contained in each cluster. The analysis concludes without any modification due to the non-existence of a cluster with a remarkable low representation of sections. A pie graph has been built in to help with this task as can be observed in Figure 4.14.

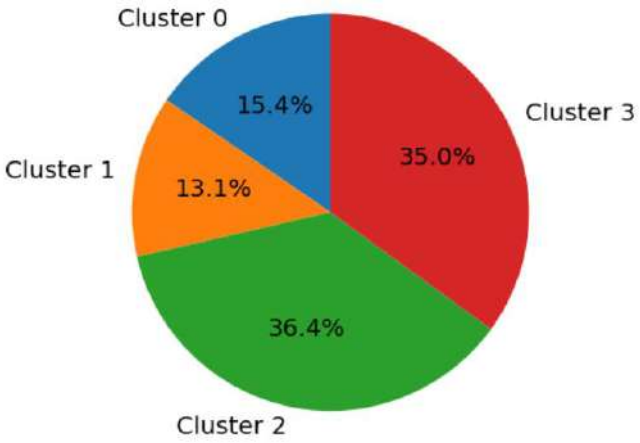


Figure 4.14 – Pie graph of the distribution of census sections among the clusters. “Living Conditions” Segmentation

Source: Author

Once the clusters are not going to be modified and they are considered definitive, it is time to the next phase where the main characteristics of each cluster have been analyzed to be useful for the researchers that use the system.

In the following figure are presented the values of the centroid of each cluster:

Clusters	HOM1	HOM3	LIF1	LIF3
Cluster 0	0.584942	0.259916	0.662029	0.160642
Cluster 1	0.536632	0.101947	0.252964	0.244006
Cluster 2	0.576466	0.546820	0.290025	0.231734
Cluster 3	0.282696	0.112123	0.287319	0.212365

Figure 4.15 – Mean values (centroids) of each cluster and each variable. “Living Conditions” Segmentation

Source: Author

The method selected for profiling the different clusters is the same as the previous segmentation. The results are exposed in the graph presented in the following figure:

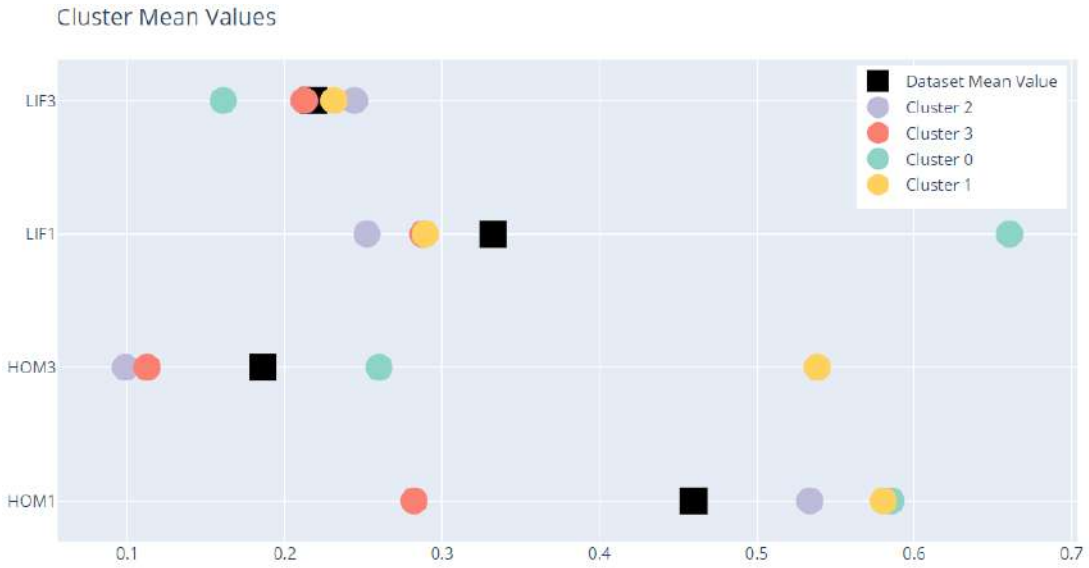


Figure 4.16 – "Mean value" profiling. "Living conditions" Segmentation

Source: Author

Due to the analysis of the previous graph, it is possible to conclude the characterization of the different clusters of the segmentation, as shown in the following list:

- **Cluster 0: "High" Cluster.** Areas where in most of the homes live one or two people. The home density is higher than the average of the region. The incomes are the highest of the region and the mortality rate is the lowest. These conditions make this cluster the group of sections where the living conditions could be considered the "best".
- **Cluster 1: "Medium & High Home Density" Cluster.** Areas where most of the population live in homes of one or two people. The home density is the highest in the region. The incomes are lower than the average and the mortality rate is on the average. These sections correspond to most populated areas of the capital, where the home density is high, and the people do not have "high living conditions".
- **Cluster 2: "Low" Cluster.** Areas where in most of the homes one or two people are living. The home density is the lowest of the region, as well as the incomes. The mortality rate is the highest of the region. Two type of areas are found in this cluster, the one corresponding to the "worst" census sections to live in term of living conditions inside the capital, and some rural and countryside regions where the conditions in terms of economy are low.
- **Cluster 3: "Medium & Low Home Density" Cluster.** Areas where most of the people live in homes of more than two co-habitants. The homes density and the incomes are lower than the average of the region and the mortality rate is in the mean. This cluster mainly represents the census sections of little towns where the home density is low, and the economy is under the average of the region.

The following figure shows the distribution of the clusters over the territory of the province of Seville (Spain), as well as zoom in the capital of the region:

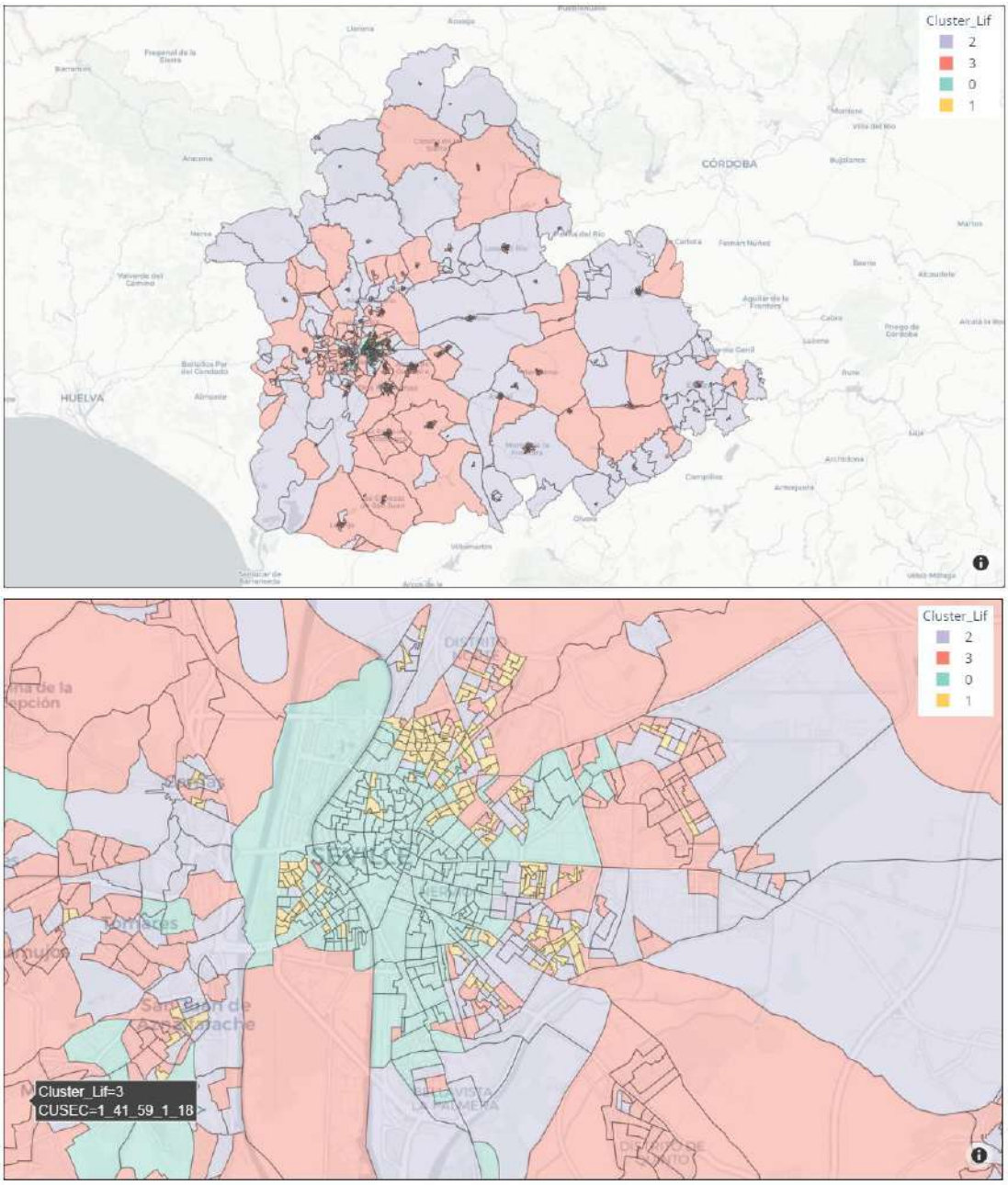


Figure 4.17 – Clusters among the territory. “Living Conditions” Segmentation

Source: Author

These two segmentations have been integrated into the system and allow clinical researchers to know in which cluster the patients fall into, as well as understand what each cluster represents. This provides valuable information, simply and effectively, during the development of studies, understanding what the demographic characteristics and “living conditions” of the patients and the people around them are.

4.5.3. Graphic analysis of the clusters

Some pictures of the city of Seville and surrounding towns have been chosen to characterize them for a better understanding of what each cluster represents. As it can be observed by analyzing the two distributions of the census sections among the territory, both segmentations divide the city of Seville and the surrounding towns into two similar homogeneous groups. This demonstrates the high correlation between the demographic variables and the "living conditions" variables used for both segmentations. Otherwise, just two of the clusters are presented for the areas located outside of the capital of the province. Therefore, four groups of images representing the clusters of both segmentations for the capital of Seville are presented in the following figures:

"High" & "Old & Studies" Clusters (Figure 4.18). These two clusters represent the city center and neighborhoods around it and some areas where the population has high incomes and the rate of people with complete studies is high. The city center and surrounding neighborhoods represent the most important part of the city where are located most of the business and the entertainment places. These sections represent the area with the greatest interest and therefore, their population has the greatest purchasing power.

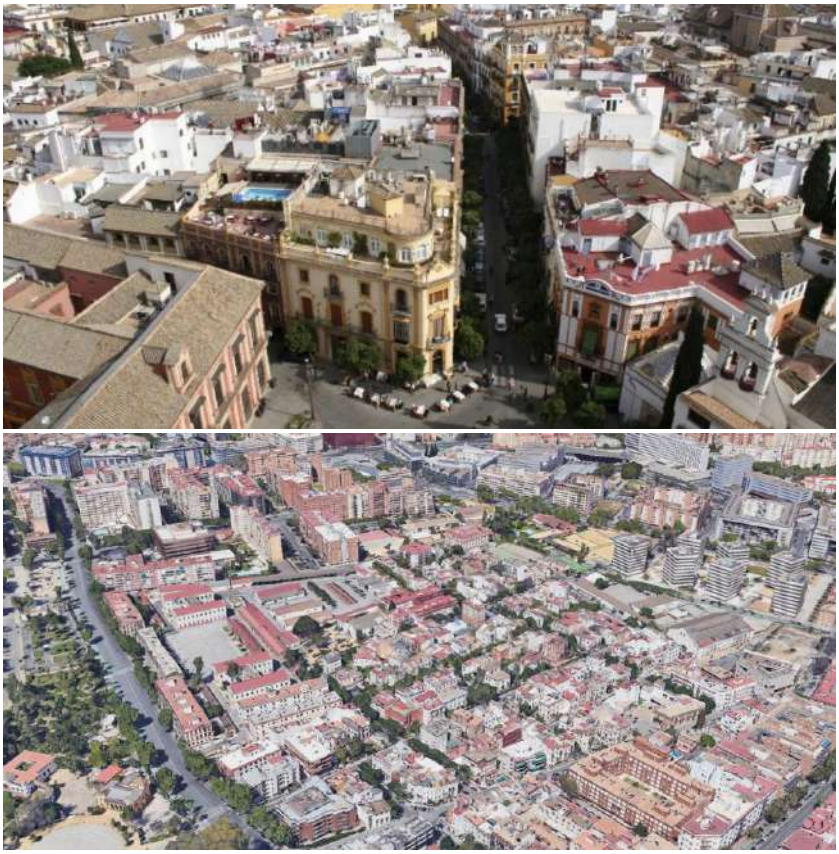


Figure 4.18 – "High" & "Old & Studies" Clusters descriptive images. (a) City center of Seville; (b) Neighborhood around the city center

Source: Retrieved from: (a) https://www.sevillacitycentre.com/sevilla-centro/; (b) Google Earth

"Medium & High Homes Density" & "Old, international & High Population Density" Clusters (Figure 4.19). These two clusters represent the external neighborhoods of the city and some areas in other towns. In this group of sections, the population and home density are the highest of the territory and their population has a medium-low purchase power with high rates of unemployment, lower studies, and lower incomes. Moreover, they are the regions where the rate of foreign people is higher.

Figure 4.19 – "Medium & High Homes Density" & "Old, international & High Population Density" Clusters descriptive images. (a) Neighborhood on the north of the city; (b) Neighborhood on the east of the city

Source: Retrieved from: (a) Diario Sevilla[9]; (b) Google Earth

[9] Retrieved from https://www.diariodesevilla.es/sevilla/linea-3-Tussam-Pino-Montano-Bermejales-rapida-expres-plan-movilidad-Norte-Macarena_0_1510049342.html

"Low" & "Young & Low Population Density" Clusters (Figure 4.20): These two clusters represent the areas where the population has the lowest salaries, highest mortality rate and in general, a low education level. The areas that fall into these clusters are the little towns, countryside areas and some of the neighborhoods with the lowest conditions in terms of security, incomes, and expectancy of life of the capital:

Figure 4.20 – "Low" & "Young & Low Population Density" Clusters descriptive images. South neighborhood of the city

Source: (a) El Mira[10]

"Medium & Low Homes Density" & "No studies & Low Population Density" Clusters (Figure 4.21): These two clusters represent the rural areas of the province and some medium-low purchase power areas of the city. In this group of sections, the population and home density are the lowest and in general, the habitants do not have studies.

Figure 4.21 – "Medium & Low Homes Density" & "No studies & Low Population Density" Clusters descriptive image. Example of a village in the north of the province

Source: VerPueblos[11]

[10] Retrieved from https://www.elmira.es/articulo/sevilla/

[11] Retrieved from https://www.verpueblos.com/andalucia/sevilla/el+pedroso/foto/369489/

5. STUDY II. AIR QUALITY CHARACTERIZATION

This study aims the processing of the pollution data integrated into the system in order to understand which are the pollution conditions that the patients of the different studies are living with. This chapter contains the methodology followed during the development of this study case and the results obtained for every month after processing the data and the application of the formula to transform the AQHI value of each day and census section, into the monthly score (scale 1-10), are exposed and discussed.

5.1. PROBLEM DEFINITION

5.1.1. Identify the objective of the study

The objective of the study is the extension of the data, measured by the different pollution stations, in order to have a pollutant concentration value for each census section presented in the territory. Moreover, a ranking of the worst census sections (or group of census sections) in terms of pollution and the risk for health, has been developed assigning a score to each zone. Once this step is done and knowing the census sections where the patients live, it has been possible to understand the pollution conditions they are exposed to.

5.1.2. Data collection and feature selection

The dataset used for the development of the segmentation is the same used for the system's observation module. It provides the time series of the pollution levels captured by the pollution stations every hour and from 2014 to 2019 (range of dates of this study). Five types of pollutants are measured by the stations (O_3, SH_2, NO_2, PART, and CO) and three of them have been selected to be processed in this study due to their demonstrated correlation with asthma exacerbations: NO_2, PART, and O_3.

5.2. EXPLORATORY ANALYSIS AND DATA PROCESSING

With the aim of expanding the pollution data to the entire territory of the province of Seville (Spain), data coming from stations of the contiguous provinces to the one of this study, have also been used.

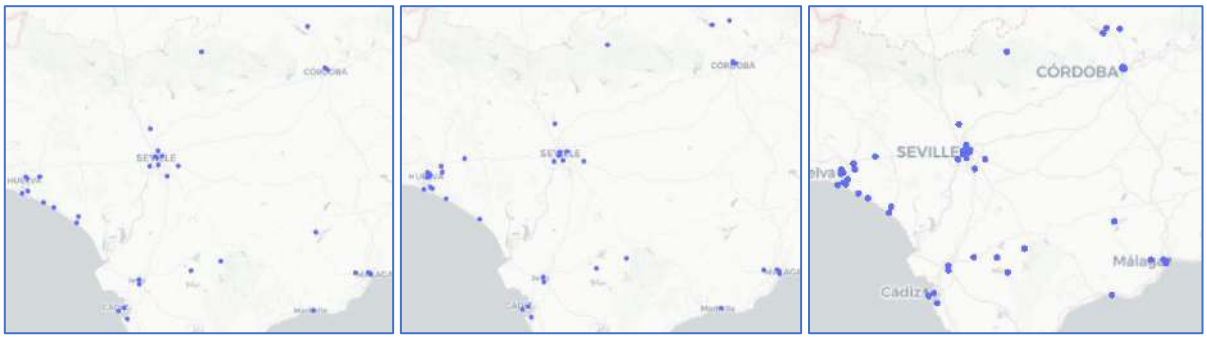


Figure 5.1 – (a) O3 stations; (b) PART stations; (c) NO2 stations

Source: Author

5.2.1. Processing the pollution data

The methodology followed to expand the data to each census section of the province of Seville, has been the creation of a grid of 250m x 250m areas over the whole territory as can be observed in Figure 5.2. This grid allows the use of the interpolation method and expands the pollution measurement captured by the stations over the whole grid and, therefore, the territory. Once they are calculated, the interpolated values are assigned to the centroid of each square of the grid. This method is performed for each day from 1st of January of 2014 to 31st of December of 2019.

Figure 5.2 – (a) Grid over Seville province; (b) Zoom-in Seville capital; (c) Zoom-in capital city center

Source: Author

Once the data has been interpolated among the grid, the next step is to calculate the average value of all the centroids that fall into every census section of the region of this study. To perform this step, the two layers (the census sections and the grid) are overlapped using the GeoPandas function “overlay”. The result of this step can be observed in the following figure where, as an example, it is represented the O_3 pollutant distribution for the municipality of Seville of the 10th of May of 2018 and the NO_2 pollutant distribution among the census sections for the 1st of January of 2016:

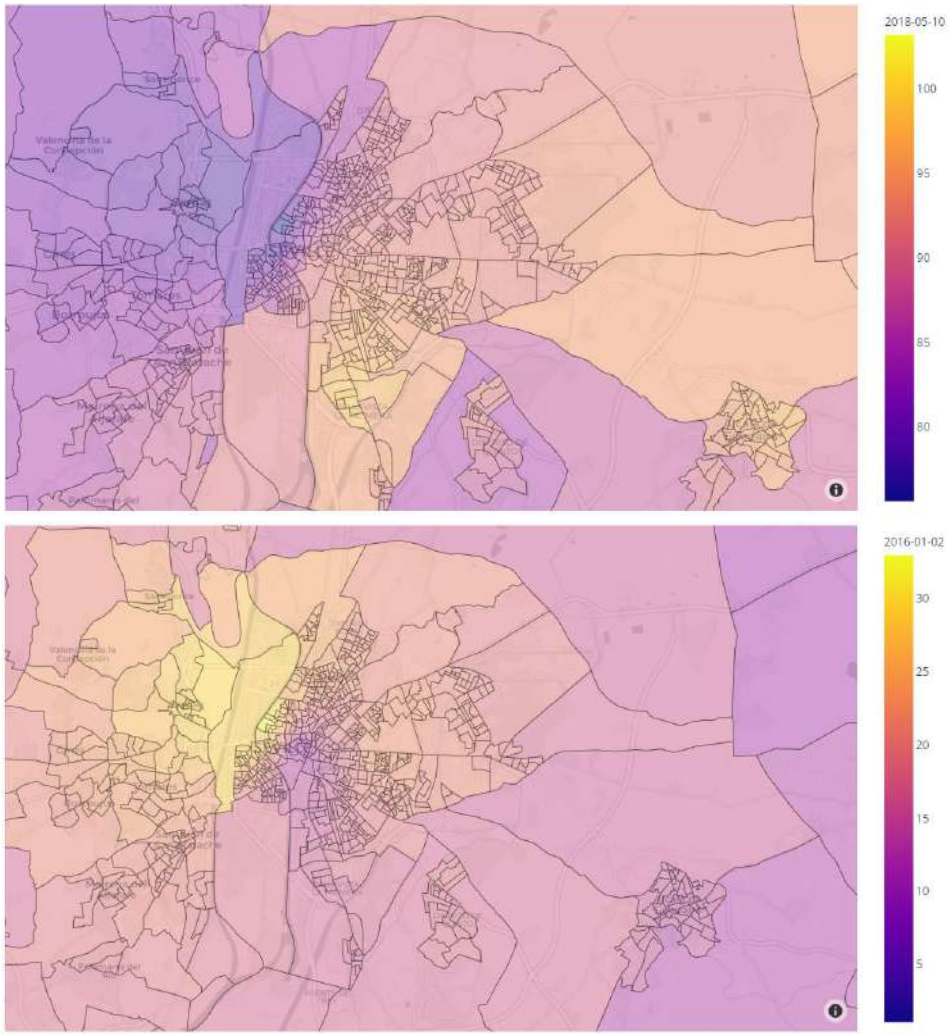


Figure 5.3 – (a) Overlapped O_3 grid example for the municipality of Seville for 10th of May of 2018; (b) Overlapped NO_2 grid example for the municipality of Seville for 1st of January of 2016

Source: Author

5.2.2. Air Quality Health Index (AQHI)

Being the objective of this study to create a ranking of the worst census sections in terms of pollution and health, it has been necessary to use a formula that assigns a score to each census section based on the concentration levels and the impact of air quality on health.

The Air Quality Health Index formula explained in the theoretical framework chapter is applied for each census section and the everyday average value for the three pollutants considered in the formula. The NO2 and O3 pollutants need to be transformed into ppb (parts per billion) to use the values in the formula. In normal conditions, 1 ppb of NO2 is equal to 1.88 µg/m^3 and 1 ppb of O3 is equal to 2.00 µg/m^3.

In the following figure are represented, as examples, the AQHI values for the territory of the study of the two dates selected in the previous point. For 10th of May of 2018, where three indexes have been

found corresponding the highest one to the city of Seville, and for 2nd of January of 2016, where four indexes have been found corresponding the worst one to the city center of Seville and the best ones to the north of the region:

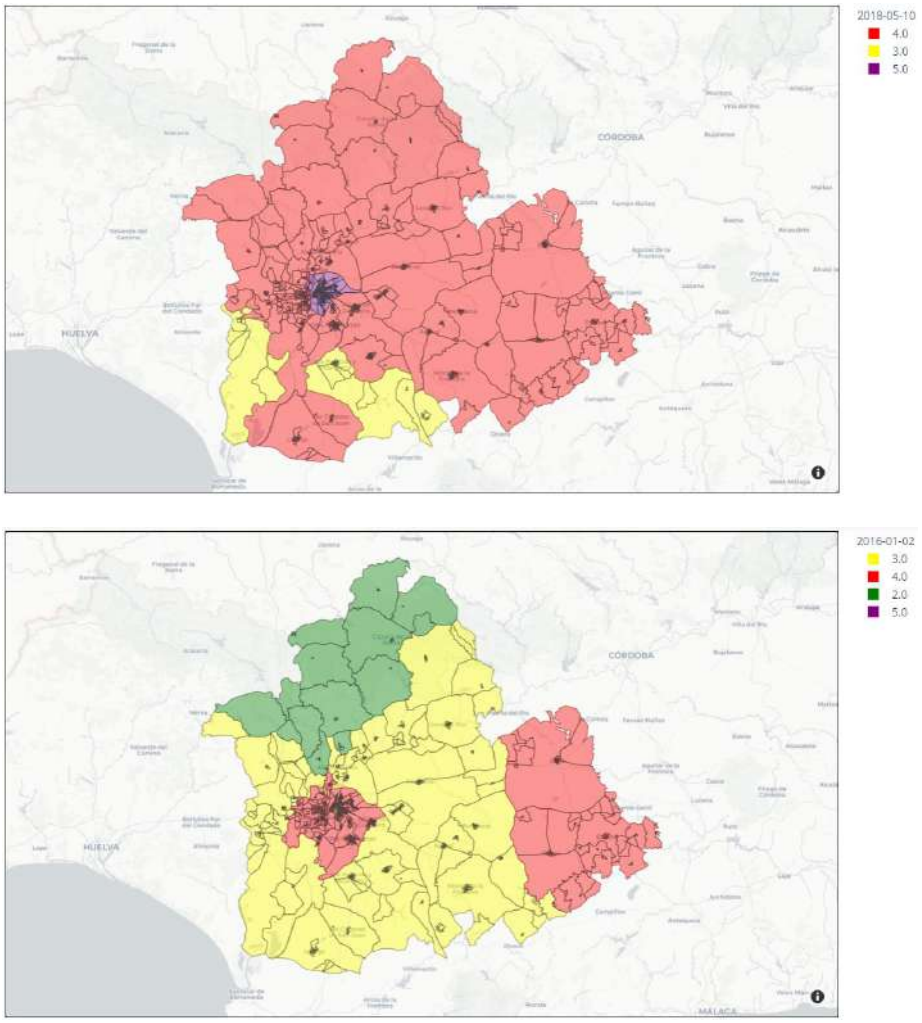


Figure 5.4 – (a) AQHI for 10th of May of 2018; (b) AQHI for 1st of January of 2016

Source: Author

5.2.3. Monthly average score

For this study, it has been determined the worst areas for each month of the year. This decision has been taken due to the relationship that the pollution concentration has with the seasons and therefore, the meteorological conditions. Then, assuming a similar environmental condition for the same months of each year, the ranking of each month has been calculated considering each day's pollution conditions for each month between 2014 to 2019.

The formula has been created assuming that the worse the conditions, the higher the score obtained by the census section and, therefore, the worse in terms of health. The percentage of days with each category of the AQHI has been calculated and used to calculate the scores of each section. The scores have a minimum possible value of one if the everyday conditions are corresponding to the index “1”,

and 10, if everyday conditions are corresponding to the “10”. In the following line is presented the formula applied for all the months of January between 2014 and 2019:

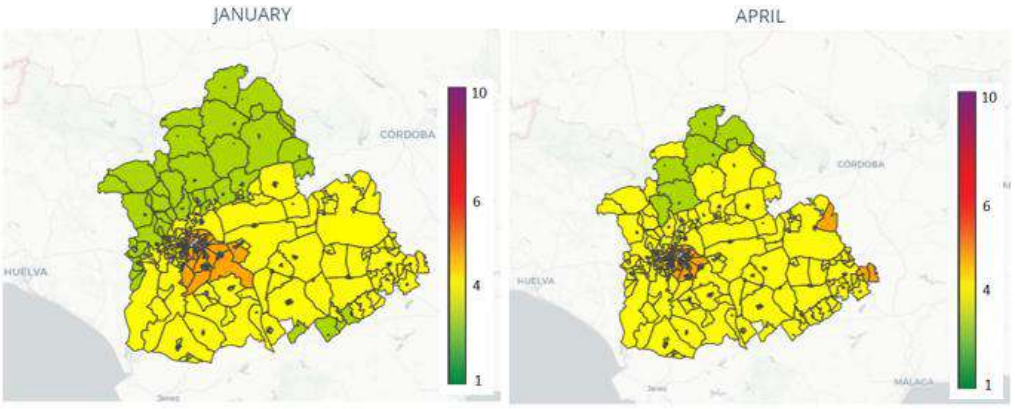
$$
\begin{aligned}
&Census\ Section\ Score \\
&= \left(\%\ of\ \text{"1"}\ days * \frac{10}{10.39}\right) + \left(\%\ of\ \text{"2"}\ days * \frac{20}{10.39}\right) + \left(\%\ of\ \text{"3"}\ days * \frac{30}{10.39}\right) \\
&+ \left(\%\ of\ \text{"4"}\ days * \frac{40}{10.39}\right) + \left(\%\ of\ \text{"5"}\ days * \frac{50}{10.39}\right) + \left(\%\ of\ \text{"6"}\ days * \frac{60}{10.39}\right) \\
&+ \left(\%\ of\ \text{"7"}\ days * \frac{70}{10.39}\right) + \left(\%\ of\ \text{"8"}\ days * \left(\frac{70}{10.39} + 1.0875\right)\right) \\
&+ \left(\%\ of\ \text{"9"}\ days * \left(\frac{70}{10.39} + 2.175\right)\right) + \left(\%\ of\ \text{"10"}\ days * \left(\frac{70}{10.39} + 3.2625\right)\right)
\end{aligned}
$$

The weights of each category in the formula have been chosen based on the study developed by Prof. Wong Tze Wai, et al. (2012). The study determined the excess of risk (% ER) of hospital admission due to respiratory diseases for each AQHI category. For each category between “1” and “7”, the %ER follows a constant ascending slope, but from category “8” the slope increases around 13%, and therefore, the weights are increased according to the slope. The values of the weights have been chosen in order to obtain a scale between 1 and 10.

5.3. RESULTS

5.3.1. Monthly scores

In the following figures are presented the most remarkable months in terms of the monthly score due to pollutant concentration and risk for health, calculated by the formula. The overall situation among the whole territory corresponds to a good-medium score (values vary between 3 and 5), as it can be observed in figure 5.5 where are represented four months corresponding to the seasons.



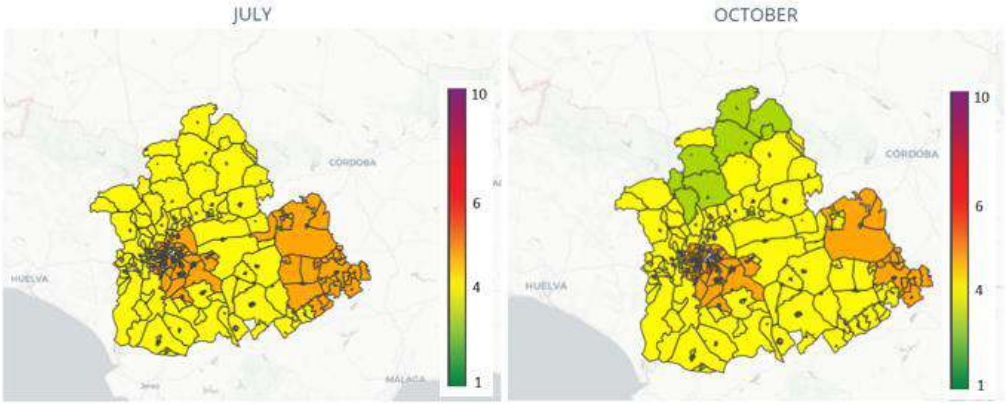


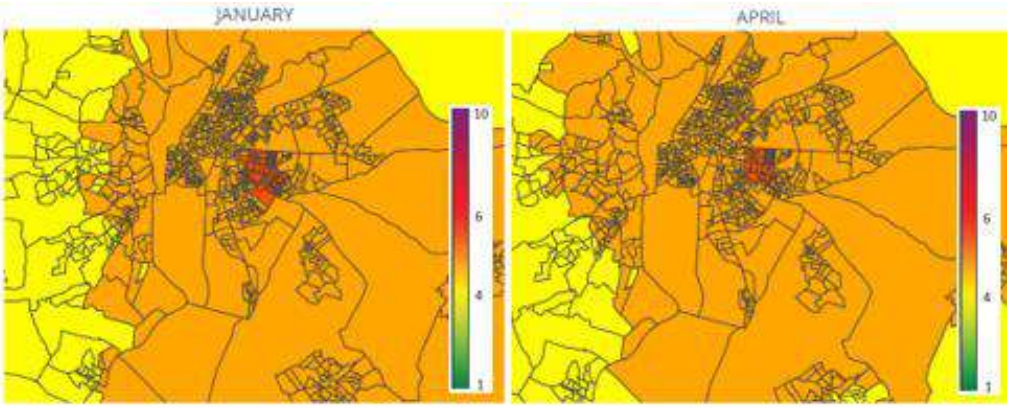
Figure 5.5 – Monthly scores for each season for Sevilla province

Source: Author

As it was expected, analyzing the images, it has been concluded that the conditions in the north of the province of Seville are more favorable in terms of pollution and health risk. This territory is the rural and mountainous part of the region and in none of the months has the result been higher than category four. Moreover, this pattern has been found in the southern part of the province but in a less clear way. This area shares certain characteristics with the northern region, but the populations are larger and the degree of industrialization in the area is slightly higher. These characteristics could explain the small differences that exist in terms of average pollution.

In the eastern region of the province, a clear pattern of worse conditions can be observed, especially during the spring and summer months, where these conditions could be influenced by the meteorological condition of those seasons.

Finally, and as could be expected, the worst conditions in the province happen in the central-west region where the city and capital of the region is located. In this region, the monthly average conditions do not fall below category 5. For a better analysis of this area, in the following figure can be observed four months, corresponding to the seasons, for the municipality of Seville and the towns around it.



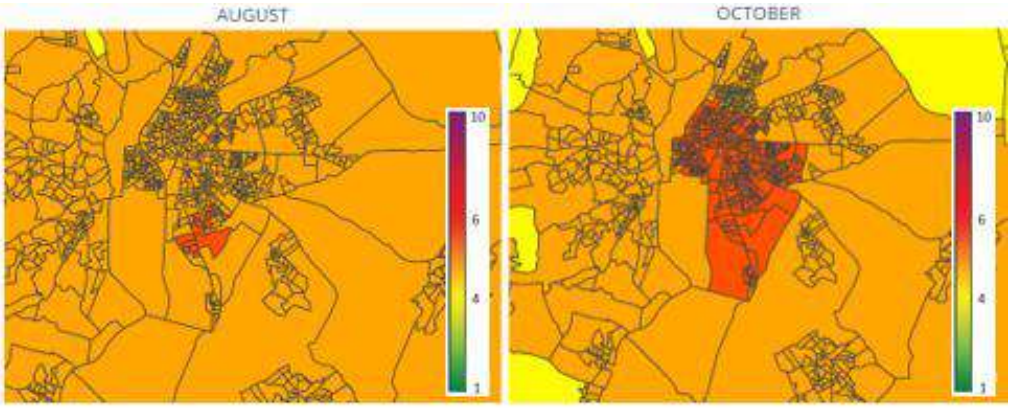


Figure 5.6 – Monthly scores for each season for Seville's capital and towns

Source: Author

The city of Seville generally presents a category 5 for the monthly averages. It should be noted, within the city of Seville, that it has been observed different areas that during some months of the year present results above the average, corresponding to category 6. These conditions could be related to the presence of one of the most important avenues of the capital and the presence of some industrial zones around those neighborhoods.

This study ends with the following conclusions:

- The best areas of the province to live in terms of pollution and health risk, are mainly the northern part of the region and, to a lesser extent, the southern area.
- The season of the year plays a fundamental role in the concentration of pollutants for all areas of the region. In the spring and summer seasons, the conditions are, in general, worse than in the colder months.
- For the capital of Seville and surrounding towns, the month of the year and its environmental conditions play an important role but with less impact on the variation of conditions.
- Some areas of the city stand out where the presence of worse conditions is more remarkable during the year. These areas are close to industrial zones and that could be directly related to these conditions.

The rest of the monthly scores for the region of Seville and for the capital the city can be found in annex 9.4 and 9.5

These 12 studies (one per month) and each day AQHI distribution, have been integrated into the system and will allow clinical researchers to know in which areas the patients live and observe if there is a relationship between the conditions and the event of the study. In the next point of this chapter, different images and information are exposed for a better understanding of the use of these maps in the system Geohealth.

6. CONCLUSIONS

This system, created by the author of this Work Project, within the “Geohealth” project of the Innovation Area of Virgen Macarena Hospital in Seville (Spain), was born with the motivation of providing valuable information to clinical researchers that until now could not obtain simply and intuitively. In particular, this project was born with the objective of providing the researchers, of the hospital's allergology unit, a system where the patient can be related to their environmental variables such as demographic or pollution variables.

During the development of the project, it has been extracted, processed, analyzed, and integrated different data from various sources in a system, providing clinical researchers current analysis techniques, such as the use of the k-means algorithm for spatial clustering. The project can be divided into three main blocks that have provided different results and utilities for the development of future clinical investigations.

The first block is an observation module, which provides simple and intuitive access to all the data integrated into the system by being represented on the same interactive map. By clicking on each object, it is possible to observe the demographic characteristics of a census section, the measurements of the pollution, meteorology, and pollen stations, as well as a heat map of the number of patients by census section or other demographic variables such as income, studies, sex, and others.

The second block is a module where the two geo-demographic segmentation made in the “Study I” of this project have been integrated. This study has allowed us to understand how it is and what are the characteristics of the population of the province of Seville (Spain) and therefore, understand what the characteristics of the patients that belong to the study are. Some of these characteristics are provided by demographics variables, studies, incomes, mortality rate, and others. Two segmentations have been carried out, one with demographic variables such as sex, studies, population density, nationality, and others, and a second segmentation, which has been called "Living Conditions Segmentation", with variables such as mortality rate, income, co-habitants per home.

Finally, the third module is focused on the processing and analysis of the pollution levels measured by the stations geographically distributed among the province of Seville. After the application of the methodology exposed in the chapter three, the results have been of great value to understand what the pollution conditions of each month of the year are, using the data between 2014 and 2019 and for each census section of the province. In addition, the module allows the observation of conditions for a particular date or for a range of dates using the Air Quality Health Index scale created by the Government of Canada, allowing researchers, for example, to explore the conditions that a specific patient can suffer due to its location.

All this makes “Geohealth” a relevant system which is already being used by some units of the hospital. Moreover, other units have already shown their interest in using it. The system created in this project has managed to improve the lack of tools and knowledge of the latest analysis techniques existing in Virgen Macarena hospital and in general in the entire Andalusian Health System.

Furthermore, the “Geohealth” system has already been approved in another study at an Andalusian level that aims to study all asthmatic patients in Andalusia, to analyze the conditions that may produce a greater risk for exacerbation of asthma in both minors and adults.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Some recommendations for future works have been summarized in the following list:

- Develop the segmentations with the variables from previous and future census sections dataset, provided by the Statistics National Institute, to understand the evolution of the region and know the characteristics of the patients according to the time they were diagnosed, had an asthma attack or another aspect.
- Develop the pollution study from previous years to understand the evolution of the conditions with the time.
- Develop similar studies with the rest of the variables integrated into the system.
- Develop studies that relate the pollution data and the meteorology information.

8. BIBLIOGRAPHY

Hospital Virgen Macarena. Retrieved from https://www.hospitalmacarena.es/entrada-blog/poblacion-de-referencia/

Pearce, N., Ait-Khaled, N., Beasley, R., Mallol, J., Keil, U., … Mitchell, E. (2007). *Worldwide trends in the prevalence of asthma symptoms: phase III of the International Study of Asthma and Allergies in Childhood (ISAAC). Thorax, 62(9)*, 758–766.

ECRHHS. (2002). *The European Community Respiratory Health Survey II. European Respiratory Journal, 20(5),* 1071–1079.

Garcia-Marcos, L., Quiros, A. B., Hernandez, G. G., Guillen-Grima, F., Diaz, C. G., Urena, I. C., … Garrido, J. B. (2004). *Stabilization of asthma prevalence among adolescents and increase among schoolchildren (ISAAC phases I and III) in Spain*. Allergy, 59(12),* 1301–1307.

Martínez-Moratalla J, Almar E, Sunyer J, et al. *Estudio Europeo del Asma. Identificación y tratamiento de individuos con criterios epidemiológicos de asma en adultos jóvenes de cinco áreas españolas. Grupo Español del Estudio Europeo del Asma. Arch Bronconeumol 1999; 35(5):* 223-228.

Bateman, E. D., Hurd, S. S., Barnes, P. J., Bousquet, J., Drazen, J. M., FitzGerald, M., … Zar, H. J. (2008). *Global strategy for asthma management and prevention: GINA executive summary. European Respiratory Journal, 31(1),* 143–178.

Braman, S. S. (2006). *The Global Burden of Asthma. Chest, 130(1),* 4S–12S.

Martínez-Moragón, E., Serra-Batllés, J., De Diego, A., Palop, M., Casan, P., Rubio-Terrés, C., & Pellicer, C. (2009). *Coste económico del paciente asmático en España (estudio AsmaCost). Archivos de Bronconeumología, 45(10),* 481–486.

Khreis, H., Kelly, C., Tate, J., Parslow, R., Lucas, K., & Nieuwenhuijsen, M. (2017). *Exposure to traffic-related air pollution and risk of development of childhood asthma: A systematic review and meta-analysis. Environment International, 100*, 1–31.

Tian, Y., Xiang, X., Juan, J., Sun, K., Song, J., Cao, Y., & Hu, Y. (2017). *Fine particulate air pollution and hospital visits for asthma in Beijing, China. Environmental Pollution, 230,* 227–233.

Achakulwisut, P., Brauer, M., Hystad, P., & Anenberg, S. C. (2019). *Global, national, and urban burdens of paediatric asthma incidence attributable to ambient NO2 pollution: estimates from global datasets. The Lancet Planetary Health.*

Shaban-Nejad, A., Michalowski, M., & Buckeridge, D. L. (2018). *Health intelligence: how artificial intelligence transforms population and personalized health. Npj Digital Medicine, 1(1).*

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). *Effective Heart Disease Prediction using Hybrid Machine Learning Techniques. IEEE Access,* 1–1.

Marín, J. M. M., & Cámara, S. B. (2008). Historia Digital de Salud. Tácticas a seguir a partir del caso Diraya en Andalucía. *RevistaeSalud. com*, 4(13).

Atherton, J. (2011). *Development of the Electronic Health Record. AMA Journal of Ethics, 13(3),* 186–189.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). *Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. Health Affairs, 33(7),* 1123–1131.

Few, S. (2007). *Perceptual Edge. Visual Business Intelligence Newsletter*. Retrieved from http://perceptualedge.com/articles/visual_business_intelligence/dboard_confusion_revisited.pdf.

Mouat, A. (2016). Using Docker: Developing and Deploying Software with Containers. (First Edition). O'Reilly Media, Inc.

Conda. (n.d.). *Conda User Guide. Retrieved March 13, 2021, from* https://conda.io/projects/conda/en/latest/

PostgreSQL. (n.d.). *About PostgreSQL. Retrieved March 11, 2021, from https://www.postgresql.org/about/*

PostGIS. (n.d.). *PostGIS References. Retrieved March 11, 2021 from https://postgis.net/docs/reference.html*

Nagpal, A., & Gabrani, G. (2019). *Python for Data Analytics, Scientific and Technical Applications. 2019 Amity International Conference on Artificial Intelligence (AICAI).*

Sutchenkov, A. A., & Tikhonov, A. I. (2020). Embedding Interactive Python Web Applications into Electronic Textbooks. 2020 V International Conference on Information Technologies in Engineering Education (Inforino).

Hossain, S., Calloway, C., Lippa, D., Niederhut, D., & Shupe, D. (2019). Visualization of Bioinformatics Data with Dash Bio. In *Proceedings of the 18th Python in Science Conference* (pp. 126-133).

Karami, M., Langarizadeh, M., & Fatehi, M. (2017). *Evaluation of Effective Dashboards: Key Concepts and Criteria. The Open Medical Informatics Journal, 11(1)*, 52–57.

Dong, E., Du, H., & Gardner, L. (2020). *An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases*

Munzner, T. (2014). *Visualization analysis and design*. CRC press.

Jolley, K. A., Hill, D. M. C., Bratcher, H. B., Harrison, O. B., Feavers, I. M., Parkhill, J., & Maiden, M. C. J. (2012). *Resolution of a Meningococcal Disease Outbreak from Whole-Genome Sequence Data with Rapid Web-Based Analysis Methods. Journal of Clinical Microbiology, 50(9),* 3046–3053.

Petersen, J., Gibin, M., Longley, P., Mateos, P., Atkinson, P., & Ashby, D. (2011). Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *Journal of Geographical Systems*, *13*(2), 173-192.

Bright, C. J., Gildea, C., Lai, J., Elliss-Brookes, L., & Lyratzopoulos, G. (2020). *Does geodemographic segmentation explain differences in route of cancer diagnosis above and beyond person-level sociodemographic variables?. Journal of Public Health.*

Requia, W. J., Coull, B. A., & Koutrakis, P. (2019). *Multivariate spatial patterns of ambient PM2.5 elemental concentrations in Eastern Massachusetts. Environmental Pollution.*

Stock, K., & Guesgen, H. (2016). *Geospatial Reasoning With Open Data. Automating Open Source Intelligence,* 171–204.

Instituto Nacional de Estadística (INE). (n.d.). *Cuestiones básicas sobre el Censo. Retrieved March 15, 2021 from https://www.ine.es/censos2011/censos2011_faq.htm*

Singleton, A. D., & Spielman, S. E. (2013). *The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom. The Professional Geographer, 66(4),* 558–567.

Sivadas, E. (1997). *A preliminary examination of the continuing significance of social class to marketing: a geodemographic replication. Journal of Consumer Marketing, 14(6)*, 463–479.

Longley, P., & Clarke, G. (1995). GIS for business and service planning.

Abbas, J., Ojo, A., & Orange, S. (2009). *Geodemographics – a tool for health intelligence? Public Health, 123(1)*, e35–e39.

Canada. (2019). *About the Air Quality Health Index. Retrieved April 18, 2021, from* https://www.canada.ca/en/environment-climate-change/services/air-quality-health-index/about.html

Canada. (2015). *Understanding Air Quality Health Index messages. Retrieved April 18, 2021, from* https://www.canada.ca/en/environment-climate-change/services/air-quality-health-index/understanding-messages.html

Stieb, D. M., Burnett, R. T., Smith-Doiron, M., Brion, O., Shin, H. H., & Economou, V. (2008). *A New Multipollutant, No-Threshold Air Quality Health Index Based on Short-Term Associations Observed in Daily Time-Series Analyses. Journal of the Air & Waste Management Association,* 58(3), 435–450.

Hand, D. J., & Adams, N. M. (2015). *Data Mining. Wiley StatsRef: Statistics Reference Online,* 1–7.

Ribas, E. (2018). *¿Qué es el Data Mining o minería de datos?*. Retrieved from https://www.iebschool.com/blog/data-mining-mineria-datos-big-data.

Grekousis, G., & Thomas, H. (2012). *Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods. Applied Geography, 34*, 125–136.

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, *5*(4), 83-124.

Adnan, M., Longley, P. A., Singleton, A. D., & Brunsdon, C. (2010). *Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases. Transactions in GIS, 14(3),* 283–297.

Singleton, A. D., & Longley, P. A. (2009). *Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. Papers in Regional Science, 88(3),* 643–666.

Bação, F. L. (2019) *Data Mining slides. Nova IMS.*

Wong TW, Tam WWS, Lau AKH, Ng SKW, Yu ITS, Wong AHS, Yeung D. (2012). *A Study of the Air Pollution Index Reporting System. The Chinese University of Hong Kong.*

9. ANNEXES

9.1. VARIABLES OF THE CENSUS SECTION'S DATASET

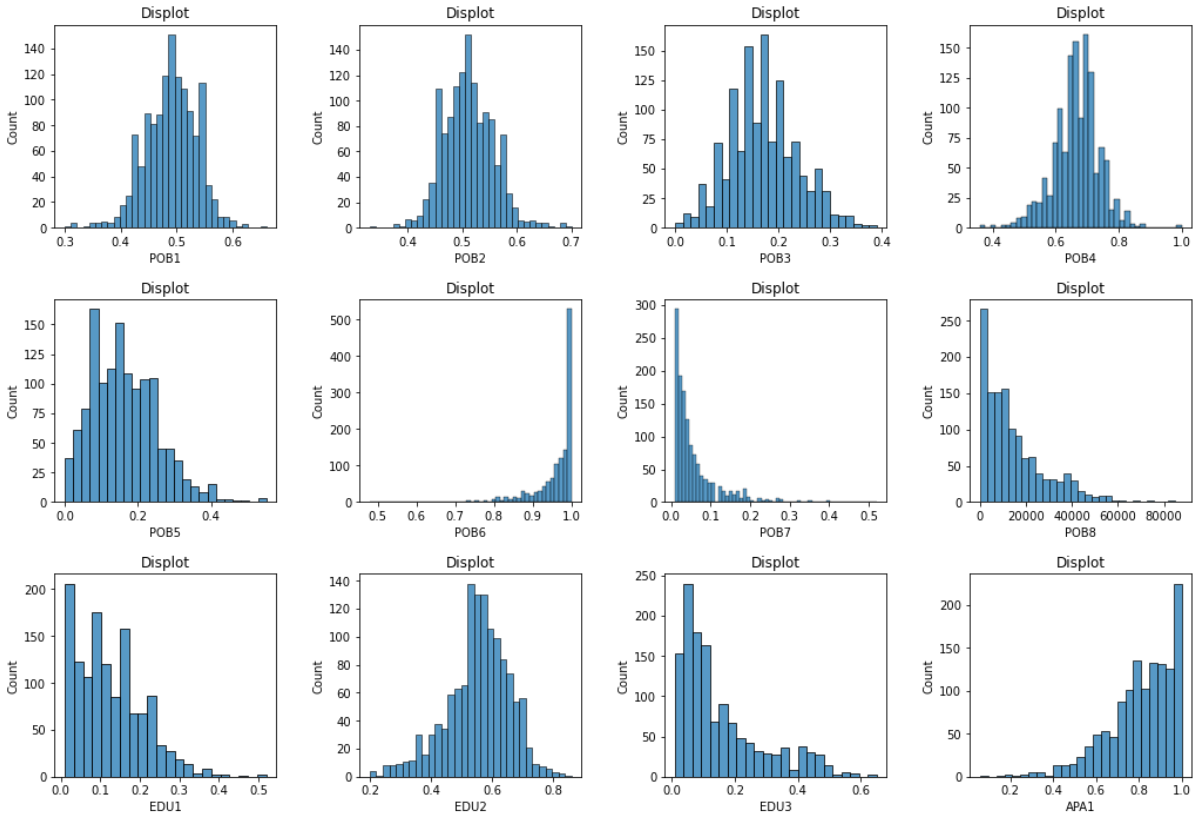
VARIABLES FROM INE				
CATEGORIA	NOMBRE	DESCRIPCION	TIPO	UNIDADES
POPULATION	POB	Total population	QUANTATIVE	NUMBER
	POB1	Men	QUANTATIVE	NUMBER
	POB2	Women	QUANTATIVE	NUMBER
	POB3	Persons under 16 years	QUANTATIVE	NUMBER
	POB4	People between 16 (including) and 64	QUANTATIVE	NUMBER
	POB5	People over 64 years	QUANTATIVE	NUMBER
	POB6	People who were born in Spain	QUANTATIVE	NUMBER
	POB7	People who were born in another EU	QUANTATIVE	NUMBER
	POB8	People who are born in a European country	QUANTATIVE	NUMBER
	POB9	People who are born in Africa	QUANTATIVE	NUMBER
	POB10	People who were born in Central America, in	QUANTATIVE	NUMBER
	POB11	People who were born in North America	QUANTATIVE	NUMBER
	POB12	People who were born in Asia	QUANTATIVE	NUMBER
	POB13	People born in Oceania	QUANTATIVE	NUMBER
	POB14	Spanish nationals who were born in Spain	QUANTATIVE	NUMBER
	POB15	Foreign nationals who were born in Spain	QUANTATIVE	NUMBER
	POB16	Spanish nationals who were born in another	QUANTATIVE	NUMBER
	POB17	Foreign nationals who were born in another	QUANTATIVE	NUMBER
	POB18	Spanish nationals who were born in a	QUANTATIVE	NUMBER
	POB19	Foreign nationals who were born in a	QUANTATIVE	NUMBER
	POB20	Spanish nationals who were born in Africa	QUANTATIVE	NUMBER
	POB21	Foreign nationals who were born in Africa	QUANTATIVE	NUMBER
	POB22	Spanish nationals who were born in Central	QUANTATIVE	NUMBER
	POB23	Foreign nationals who were born in Central	QUANTATIVE	NUMBER
	POB24	Spanish nationals who were born in North	QUANTATIVE	NUMBER
	POB25	Foreign nationals who were born in North	QUANTATIVE	NUMBER
	POB26	Spanish nationals who were born in Asia	QUANTATIVE	NUMBER
	POB27	Foreign nationals who were born in Asia	QUANTATIVE	NUMBER
	POB28	Spanish nationals who were born in Oceania	QUANTATIVE	NUMBER
	POB29	Foreign nationals who were born in Oceania	QUANTATIVE	NUMBER
	POB30	Spanish nationals	QUANTATIVE	NUMBER
	POB31	Foreign nationals	QUANTATIVE	NUMBER
	POB32	Men under 16	QUANTATIVE	NUMBER
	POB33	Men between 16 (inclusive) and 64	QUANTATIVE	NUMBER
	POB34	Men over 64 years	QUANTATIVE	NUMBER
	POB35	Women under 16	QUANTATIVE	NUMBER

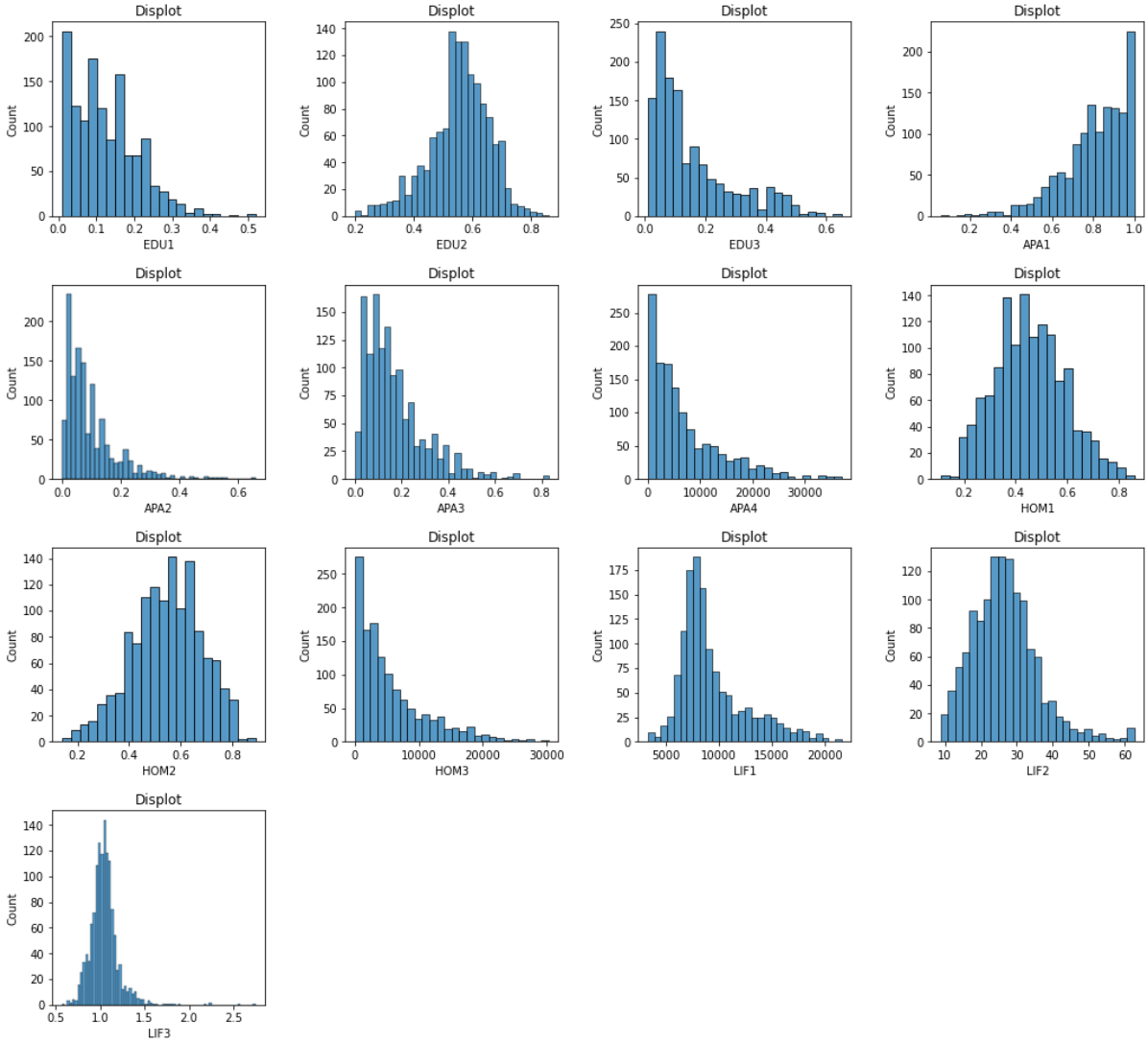
	POB36	Women between 16 (inclusive) and 64	QUANTATIVE	NUMBER
	POB37	Women over 64 years	QUANTATIVE	NUMBER
	POB38	Men with Spanish nationality	QUANTATIVE	NUMBER
	POB39	Men with foreign nationality	QUANTATIVE	NUMBER
	POB40	Women with Spanish nationality	QUANTATIVE	NUMBER
	POB41	Women with foreign nationality	QUANTATIVE	NUMBER
	POB42	Spanish nationals under 16 years	QUANTATIVE	NUMBER
	POB43	Persons of foreign nationality less than 16	QUANTATIVE	NUMBER
	POB44	Spanish nationals between 16 (including)	QUANTATIVE	NUMBER
	POB45	People of foreign nationality from 16	QUANTATIVE	NUMBER
	POB46	Spanish nationals over 64 years	QUANTATIVE	NUMBER
	POB47	Persons of foreign nationality more than 64	QUANTATIVE	NUMBER
	POB48	People with marital status Single	QUANTATIVE	NUMBER
	POB49	Married people	QUANTATIVE	NUMBER
	POB50	People with separate marital status	QUANTATIVE	NUMBER
	POB51	Divorced people	QUANTATIVE	NUMBER
	POB52	People with marital status widowed	QUANTATIVE	NUMBER
	POB53	Marital status Single people under 16 years	QUANTATIVE	NUMBER
	POB54	Marital status people between 16 (inclusive)	QUANTATIVE	NUMBER
	POB55	Marital status Single people with more than	QUANTATIVE	NUMBER
	POB56	Married people under 16 years	QUANTATIVE	NUMBER
	POB57	Married people between 16 (inclusive) and	QUANTATIVE	NUMBER
	POB58	People with marital status married over 64	QUANTATIVE	NUMBER
	POB59	People with marital status separate from	QUANTATIVE	NUMBER
	POB60	People with marital status separate from 16	QUANTATIVE	NUMBER
	POB61	People with marital status separately over	QUANTATIVE	NUMBER
	POB62	Divorced people less than 16 years	QUANTATIVE	NUMBER
	POB63	Divorced people between 16 (inclusive) and	QUANTATIVE	NUMBER
	POB64	Divorced people with more than 64 years	QUANTATIVE	NUMBER
	POB65	Marital status widowed people less than 16	QUANTATIVE	NUMBER
	POB66	Marital status widowed people between 16	QUANTATIVE	NUMBER
	POB67	Marital status widowed people over 64 years	QUANTATIVE	NUMBER
EDUCATION	EDU1	illiterates	QUANTATIVE	NUMBER
	EDU2	Uneducated	QUANTATIVE	NUMBER
	EDU3	People with first-degree studies	QUANTATIVE	NUMBER
	EDU4	People with second degree	QUANTATIVE	NUMBER
	EDU5	People with tertiary studies	QUANTATIVE	NUMBER
	EDU6	People without information on their level of	QUANTATIVE	NUMBER
	EDU7	illiterate men	QUANTATIVE	NUMBER
	EDU8	illiterate women	QUANTATIVE	NUMBER
	EDU9	Men with no education	QUANTATIVE	NUMBER
	EDU10	Women with no education	QUANTATIVE	NUMBER
	EDU11	Men with first-degree studies	QUANTATIVE	NUMBER
	EDU12	Women with first-degree studies	QUANTATIVE	NUMBER

	EDU13	Men with second degree	QUANTATIVE	NUMBER
	EDU14	Women with second degree	QUANTATIVE	NUMBER
	EDU15	Men with tertiary studies	QUANTATIVE	NUMBER
	EDU16	Women with tertiary studies	QUANTATIVE	NUMBER
	EDU17	Men without information on their level of	QUANTATIVE	NUMBER
	EDU18	Women without information on their level of	QUANTATIVE	NUMBER
	EDU19	Spanish nationals and illiterate	QUANTATIVE	NUMBER
	EDU20	Persons of foreign nationality and illiterate	QUANTATIVE	NUMBER
	EDU21	Spanish nationals uneducated	QUANTATIVE	NUMBER
	EDU22	Persons of foreign nationality without	QUANTATIVE	NUMBER
	EDU23	Spanish nationals with first-degree studies	QUANTATIVE	NUMBER
	EDU24	Persons of foreign nationality with first-	QUANTATIVE	NUMBER
	EDU25	Spanish nationals with second degree	QUANTATIVE	NUMBER
	EDU26	Persons of foreign nationality with second	QUANTATIVE	NUMBER
	EDU27	Spanish nationals with tertiary studies	QUANTATIVE	NUMBER
	EDU28	Persons of foreign nationality in tertiary	QUANTATIVE	NUMBER
	EDU29	Spanish nationals without information on	QUANTATIVE	NUMBER
	EDU30	Foreign nationals without information on	QUANTATIVE	NUMBER
	EDU31	People between 16 (including) and 64	QUANTATIVE	NUMBER
	EDU32	People over 64 years old and illiterate	QUANTATIVE	NUMBER
	EDU33	People between 16 (including) and 64	QUANTATIVE	NUMBER
	EDU34	People over 64 years without studies	QUANTATIVE	NUMBER
	EDU35	People between 16 (including) and 64	QUANTATIVE	NUMBER
	EDU36	People over 64 with first degree studies	QUANTATIVE	NUMBER
	EDU37	People between 16 (including) and 64	QUANTATIVE	NUMBER
	EDU38	People with more than 64 years with second	QUANTATIVE	NUMBER
	EDU39	People between 16 (including) and 64	QUANTATIVE	NUMBER
	EDU40	People over 64 years tertiary studies	QUANTATIVE	NUMBER
	EDU41	Persons under 16 years (no information is	QUANTATIVE	NUMBER
APARTMENTS	APA	Total Housing	QUANTATIVE	NUMBER
	APA1	housing Main	QUANTATIVE	NUMBER
	APA2	Secondary housing	QUANTATIVE	NUMBER
	APA3	Vacant housing	QUANTATIVE	NUMBER
	APA4	Home ownership by purchase fully paid	QUANTATIVE	NUMBER
	APA5	Home ownership by purchase with	QUANTATIVE	NUMBER
	APA6	Home ownership, inheritance or donation	QUANTATIVE	NUMBER
	APA7	Homes for rent	QUANTATIVE	NUMBER
	APA8	Housing ceded free or at low prices	QUANTATIVE	NUMBER
	APA9	Homes in other tenure	QUANTATIVE	NUMBER
	APA10	Houses less than 30m2	QUANTATIVE	NUMBER
	APA11	Properties between 30-45 m2	QUANTATIVE	NUMBER
	APA12	Properties between 46-60 m2	QUANTATIVE	NUMBER
	APA13	Properties between 61-75 m2	QUANTATIVE	NUMBER
	APA14	Properties between 76-90 m2	QUANTATIVE	NUMBER

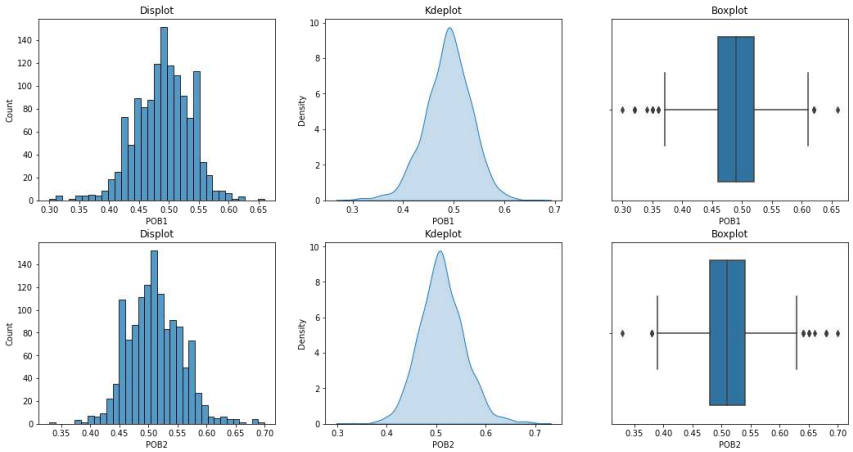
	APA15	Properties between 91-105 m2	QUANTATIVE	NUMBER
	APA16	Properties between 106-120 m2	QUANTATIVE	NUMBER
	APA17	Properties between 121-150 m2	QUANTATIVE	NUMBER
	APA18	Properties between 151-180 m2	QUANTATIVE	NUMBER
	APA19	Housing more than 180 m2	QUANTATIVE	NUMBER
	APA20	Housing with 1 bedroom	QUANTATIVE	NUMBER
	APA21	Homes with 2 bedrooms	QUANTATIVE	NUMBER
	APA22	Housing with 3 bedrooms	QUANTATIVE	NUMBER
	APA23	Houses with 4 bedrooms	QUANTATIVE	NUMBER
	APA24	Homes with 5 bedrooms	QUANTATIVE	NUMBER
	APA25	Houses with 6 bedrooms	QUANTATIVE	NUMBER
	APA26	Houses with 7 rooms	QUANTATIVE	NUMBER
	APA27	Homes with 8 rooms	QUANTATIVE	NUMBER
	APA28	Houses with 9 or more rooms	QUANTATIVE	NUMBER
HOME	HOM	Total Homes	QUANTATIVE	NUMBER
	HOM1	1 person Homes	QUANTATIVE	NUMBER
	HOM2	Homes of 2 people	QUANTATIVE	NUMBER
	HOM3	Homes of 3 people	QUANTATIVE	NUMBER
	HOM4	Homes of 4 people	QUANTATIVE	NUMBER
	HOM5	Homes 5 persons	QUANTATIVE	NUMBER
	HOM6	Homes of 6 or more people	QUANTATIVE	NUMBER

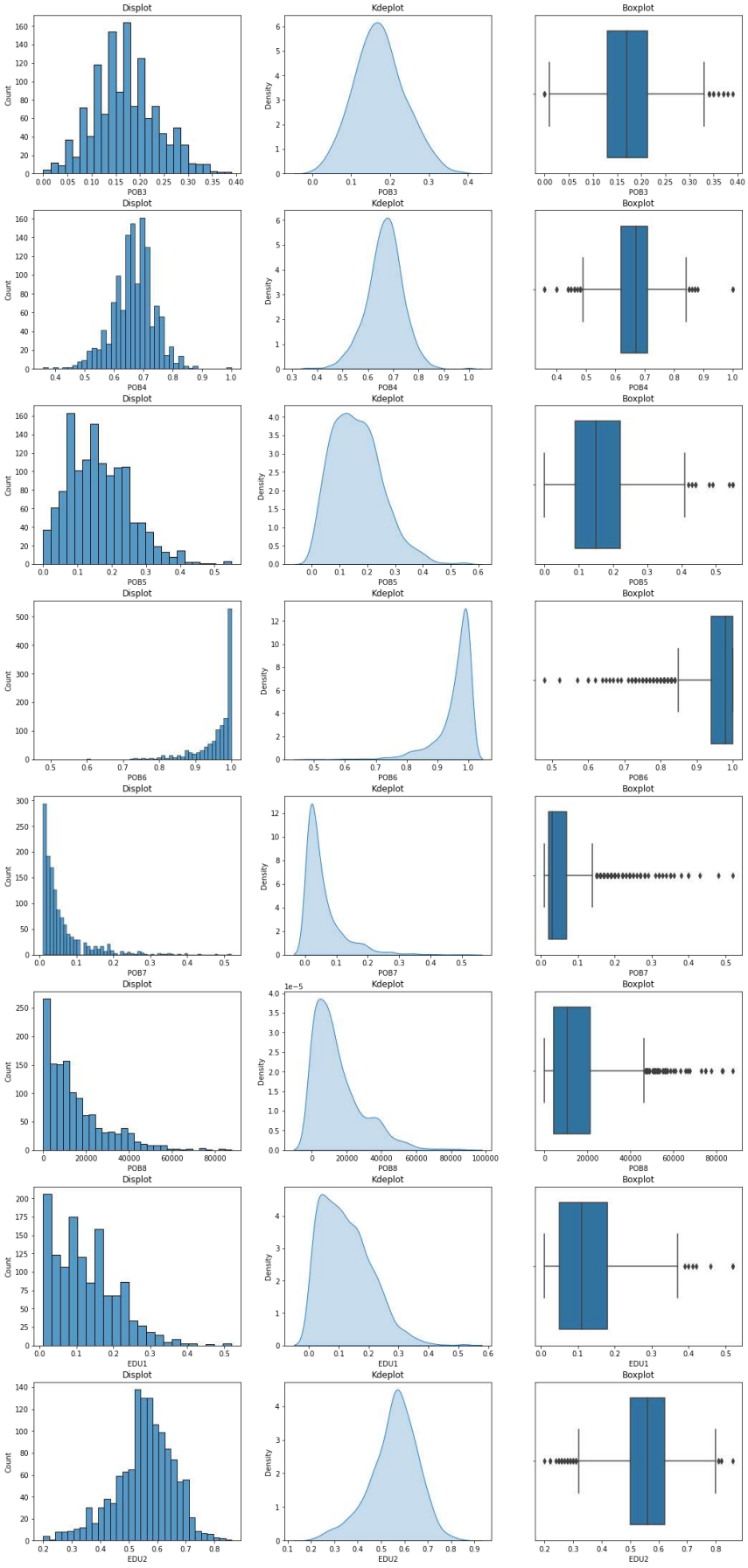
9.2. VARIABLES' DISTRIBUTION

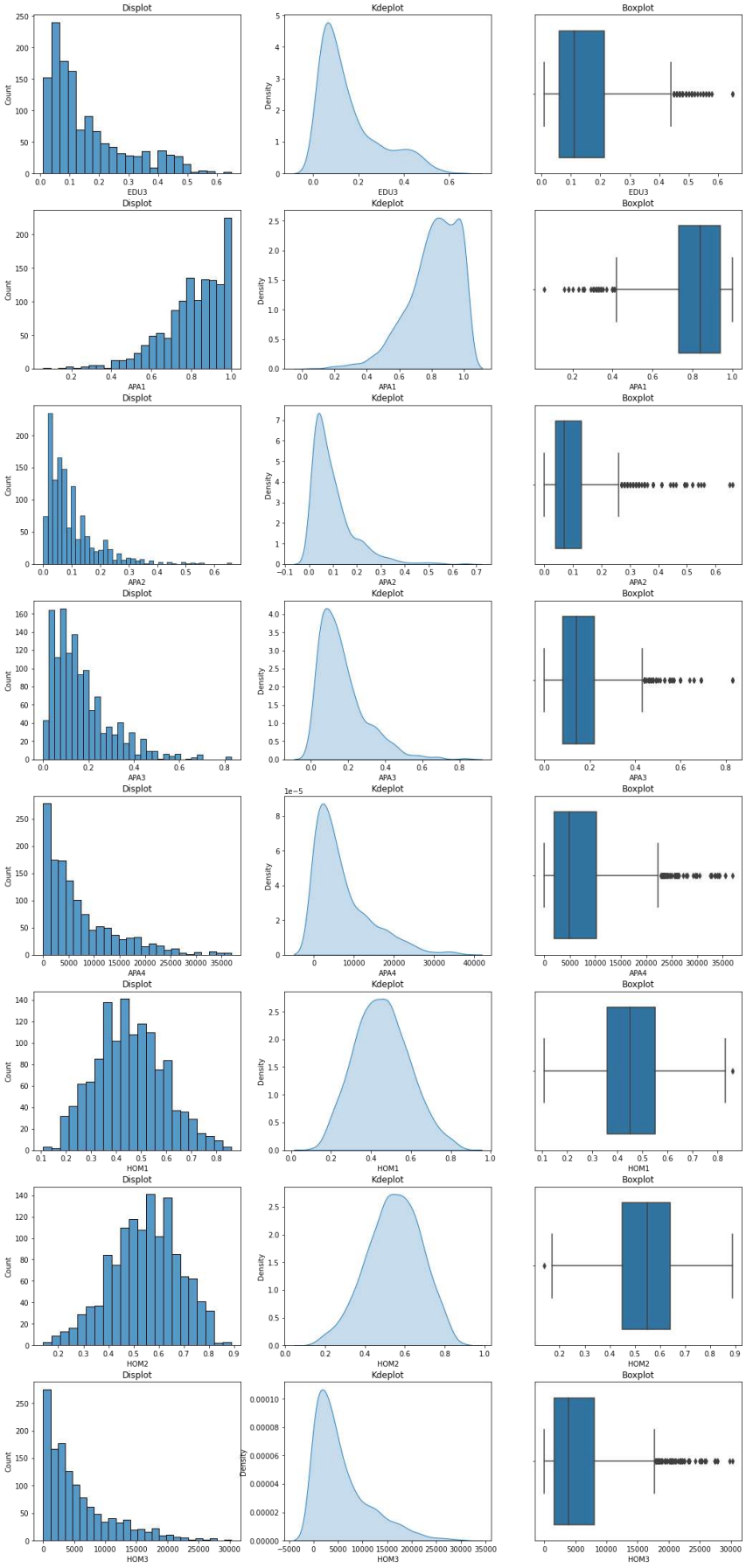


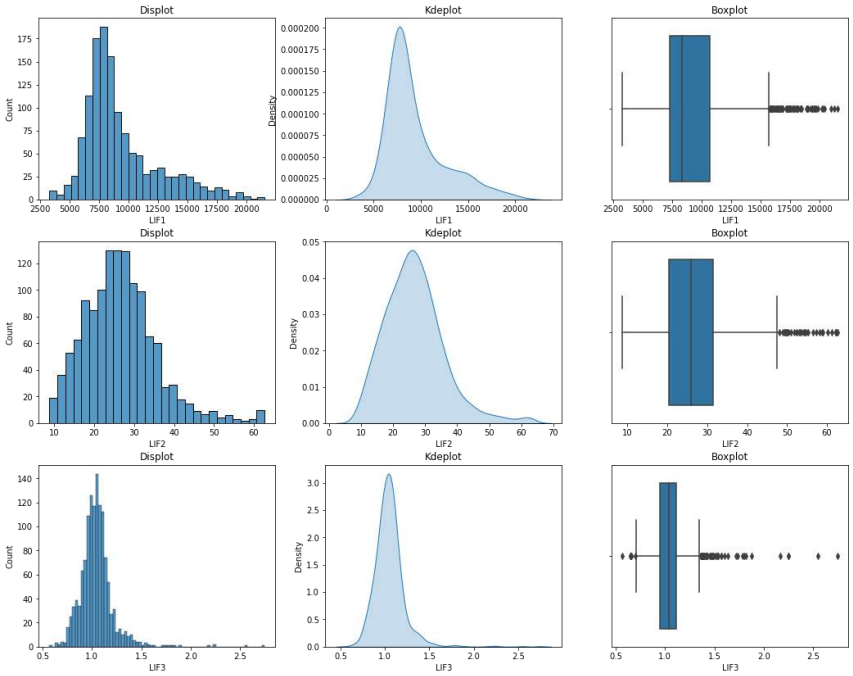


9.3. VARIABLES' DISTRIBUTION AND BOXPLOT

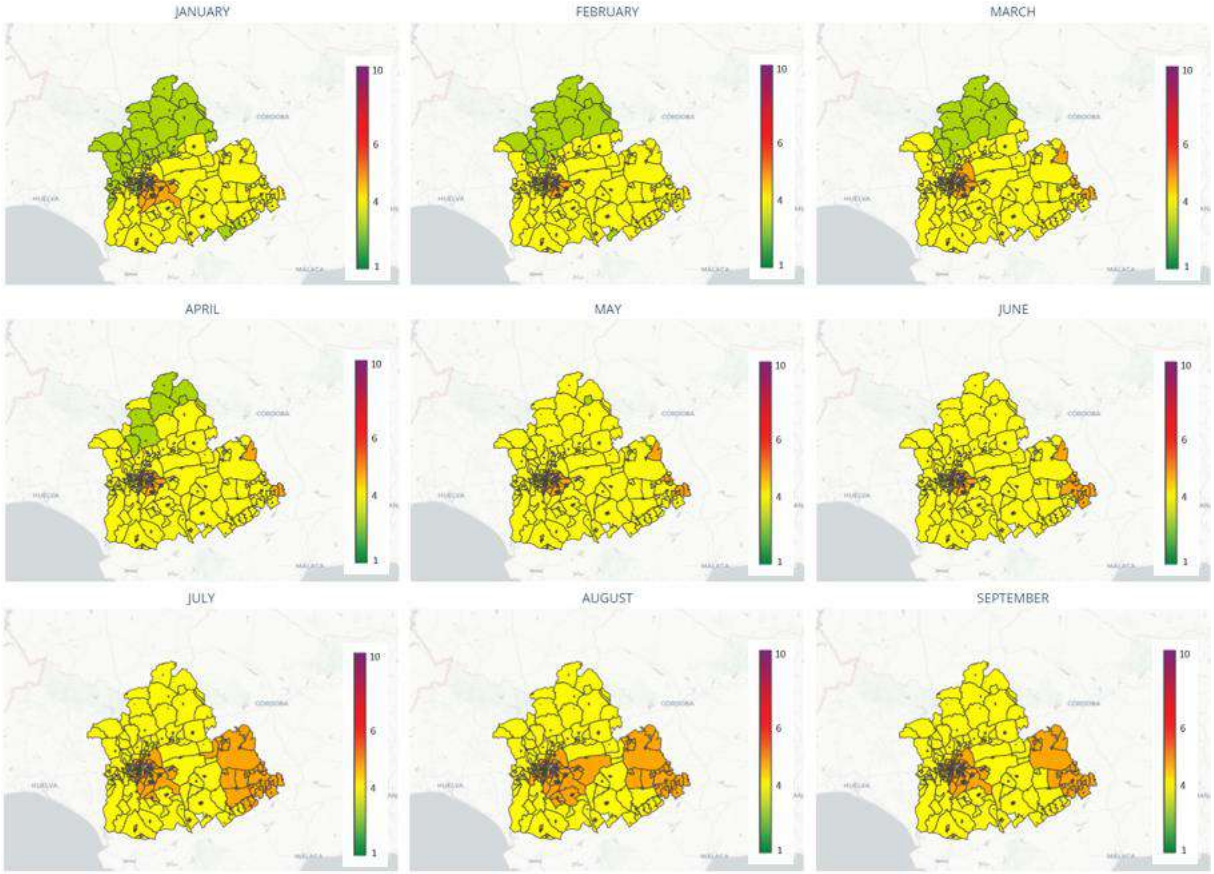


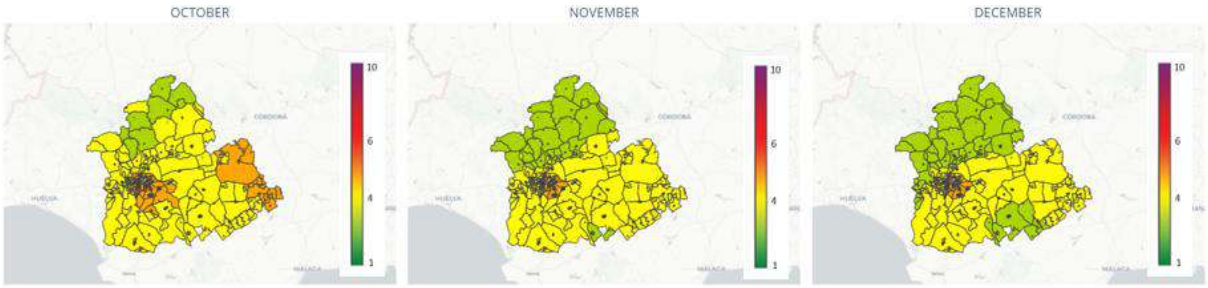

Displot
Kdeplot
Boxplot
Count
Density
POB3
POB4
POB5
POB6
POB7
POB8
EDU1
EDU2


Displot
Kdeplot
Boxplot
Count
Density
EDU3
APA1
APA2
APA3
APA4
HOM1
HOM2
HOM3

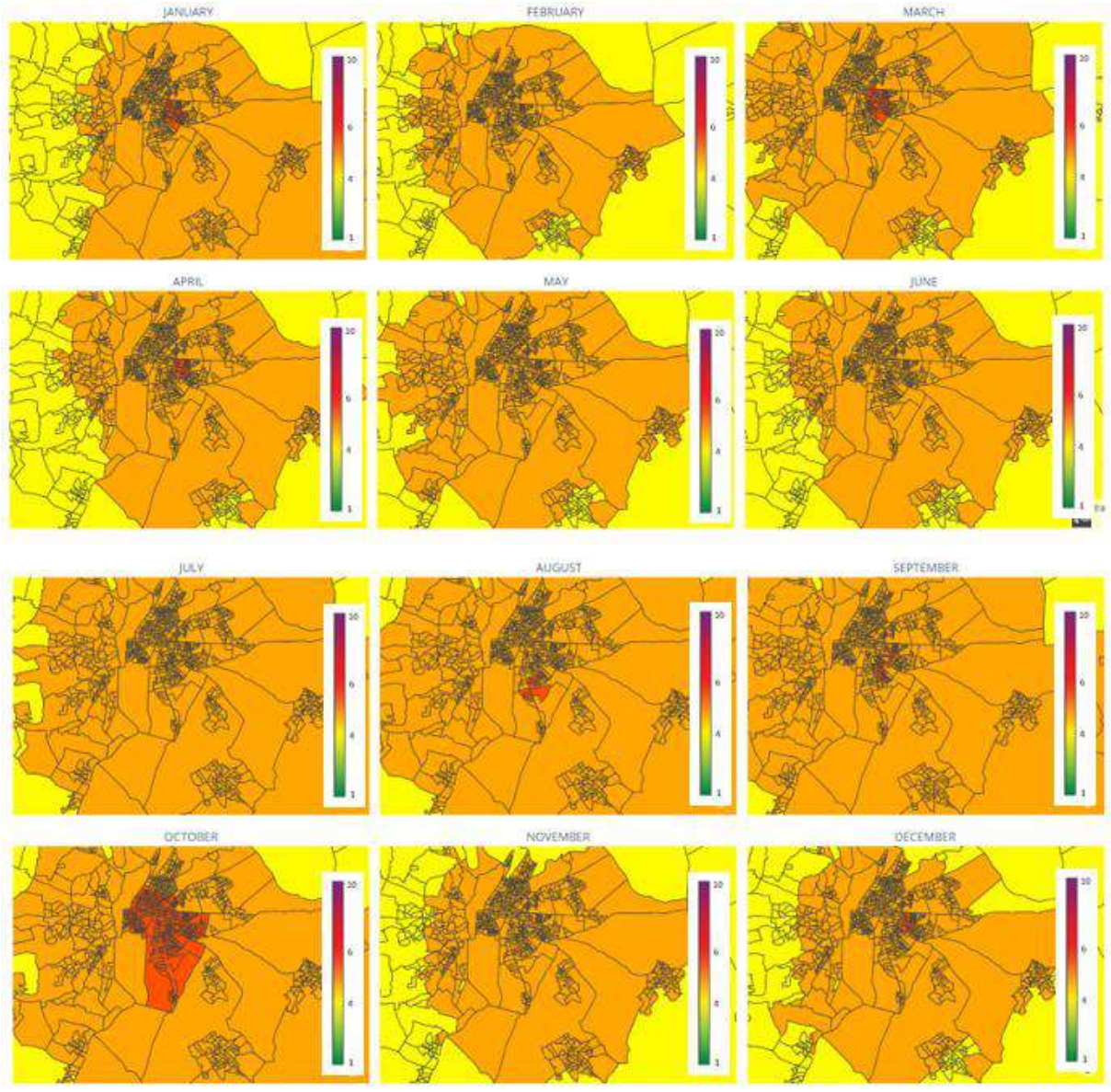


9.4. MONTHLY SCORES OF THE PROVINCE OF SEVILLE



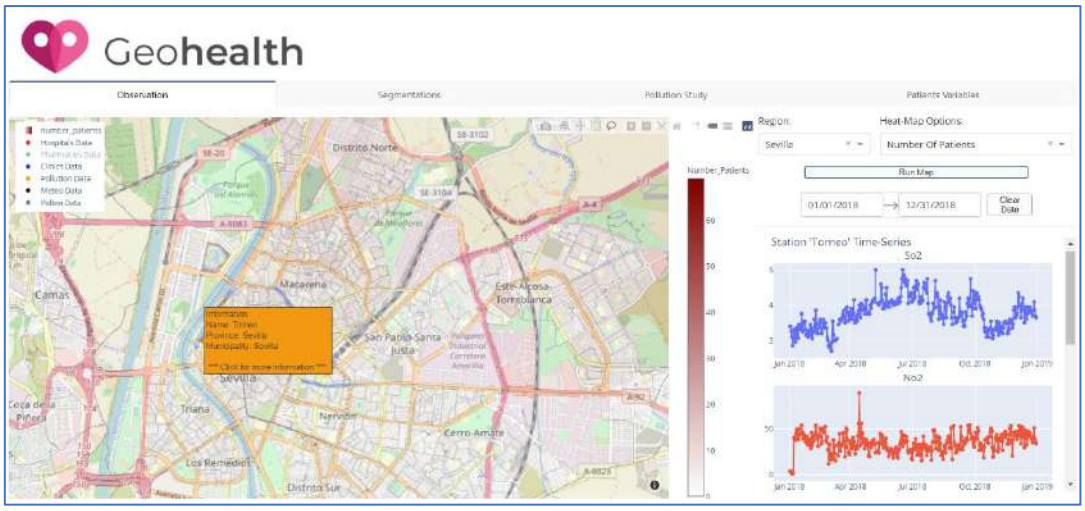


9.5. MONTHLY SCORES OF THE CAPITAL OF SEVILLE

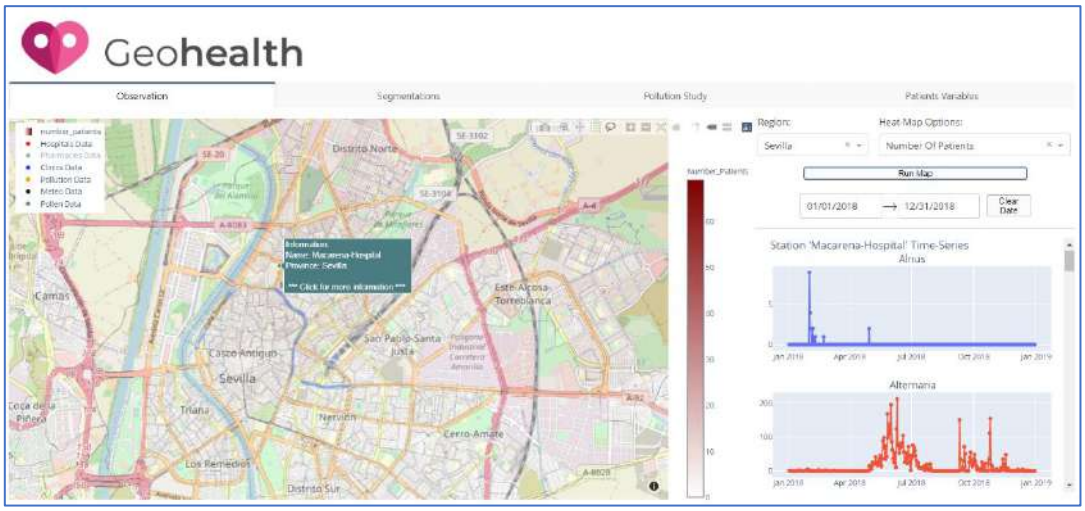


9.6. "OBSERVATION" MODULE

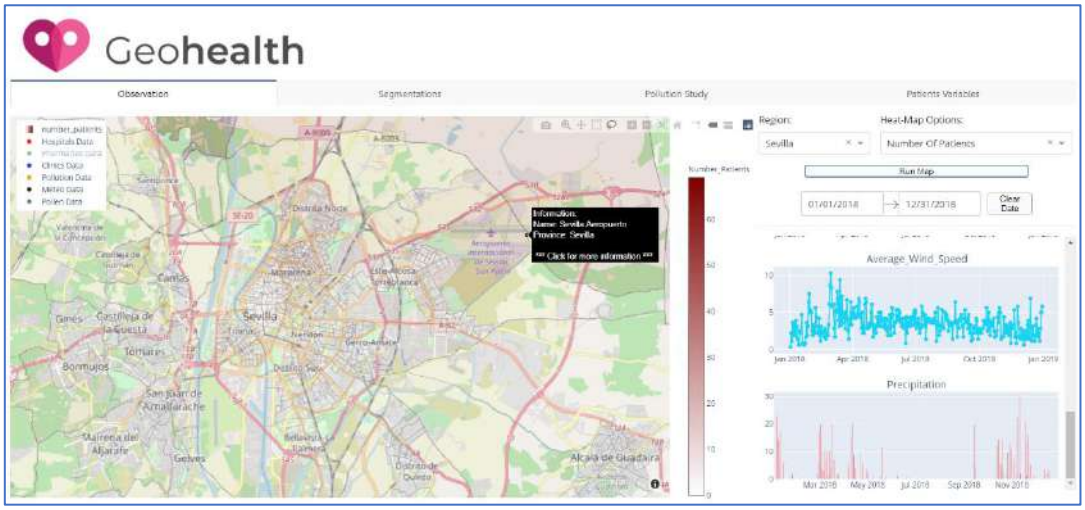
Pollution station clicked:



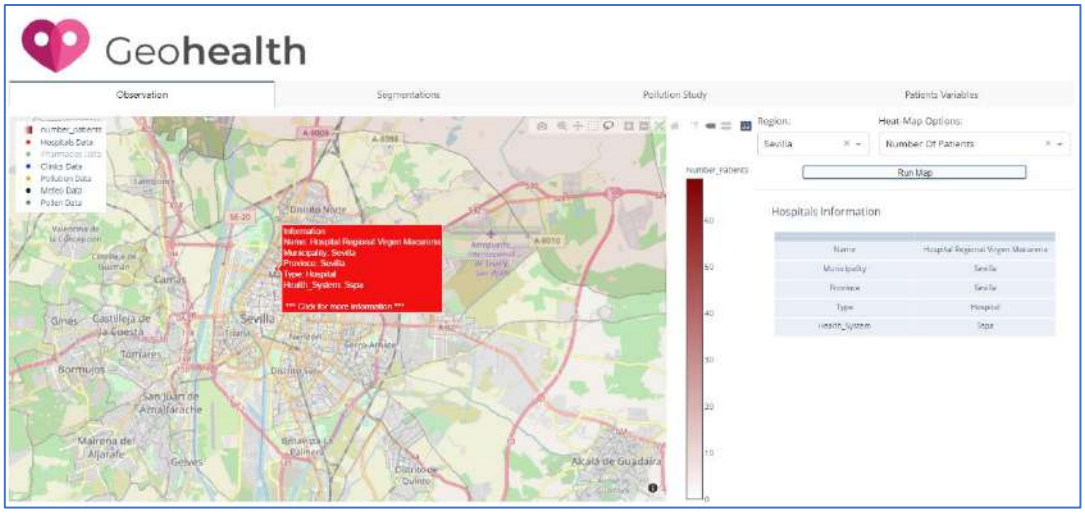
Pollen station clicked:



Meteorology station clicked:

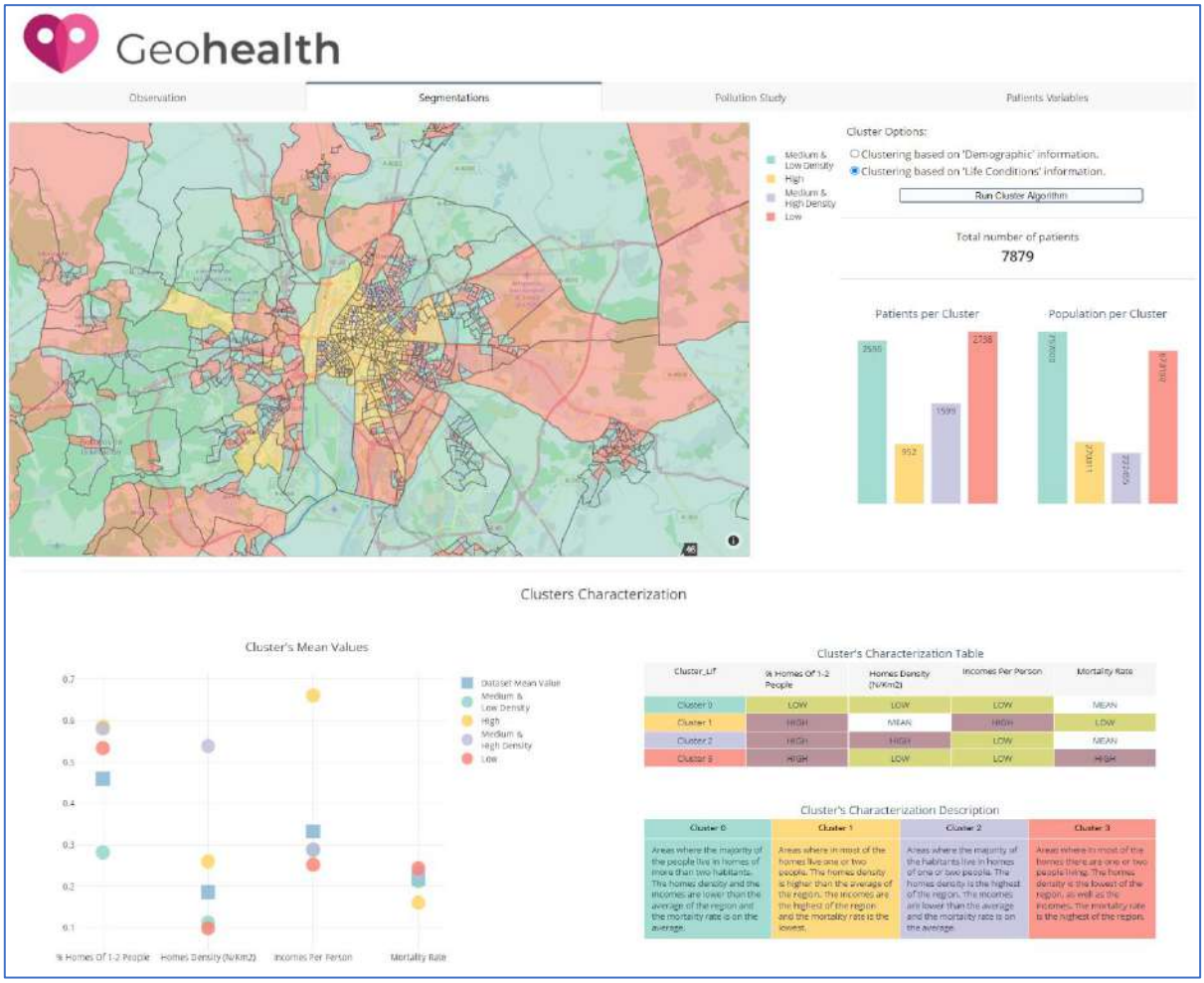


Hospital clicked:



9.7. "SEGMENTATION" MODULE

"Living Conditions" Segmentation:



9.8. "POLLUTION STUDY" MODULE

"Date Range" Study: