GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

COVID-19: ANALYSIS OF THE PANDEMIC FACTORS IN THE COMMUNITY OF MADRID AND SPAIN

Autor: Javier Olmedo Trejo
Directores: Raquel Caro Carretero
Antonio Rua Vieites

Madrid
Junio de 2020

Agradecimientos

A mis padres, a mi hermana y a toda mi familia, gracias a quienes me han permitido llegar a este momento importante en mi vida y hacia quienes sólo puedo expresar mi sincero agradecimiento por apoyarme durante la etapa académica que hoy culmina.

A los profesores Raquel Caro Carretero y Antonio Rua Vietes por haber hecho de guía y estímulo durante la redacción de mi TFG, mostrándome su apoyo, comprensión y dedicación durante todos estos meses.

Especialmente, quiero agradecer a Raquel que me enseñase la importancia de las herramientas estadísticas en su asignatura Estadística I como forma de entender y explicar el mundo.

1. Introducción

Actualmente, el mundo se encuentra combatiendo una de las mayores amenazas para la sociedad globalizada en la que vivimos como son las pandemias. El nuevo Covid-19 (causado por el virus SARS-CoV-2) está expandiéndose a un ritmo que pocos esperaban, dejando infectados en casi todos los países e interrumpiendo nuestro normal funcionamiento de vida. El virus causa diversas manifestaciones clínicas [1*] que incluyen cuadros respiratorios que varían desde el resfriado común hasta cuadros de neumonía grave con síndrome de distrés respiratorio, shock séptico y fallo multiorgánico.

El 31 de diciembre de 2019, la Comisión Municipal de Salud y Sanidad de Wuhan [2*] (provincia de Hubei, China) informó a la Organización Mundial de la Salud sobre un grupo de 27 casos de neumonía de etiología desconocida, con una exposición común a un mercado mayorista de marisco, pescado y animales vivos en la ciudad de Wuhan, incluyendo siete casos graves. Tras la detección del primer paciente infectado, los nuevos casos diagnosticados se han incrementado a un ritmo exponencial creando verdaderos retos para la salud pública y tensando los límites de los sistemas sanitarios de todo el mundo.

Aunque la mayoría de los gobiernos han extendido distintas medidas de contención como confinamientos obligatorios o cierre de fronteras, se desconoce el alcance de la infección global y la duración de la emergencia actual. El desarrollo de vacunas y ensayos clínicos de posibles tratamientos se encuentran en marcha en estos momentos con una gran esperanza en que con estos se consigan controlar la expansión de la pandemia.

2. Estado de la cuestión

Por ahora, la investigación de esta nueva pandemia se está realizando en tiempo real con los avances y datos que nos deja día a día ésta. El mayor número de investigaciones hasta la fecha se han conseguido llevar a cabo en China por el mayor tiempo disponible desde el origen de la enfermedad en su territorio.

Sin embargo, numerosas instituciones tanto a nivel nacional como global están poniendo a disposición de la ciudadanía información y bases de datos para que se realicen nuevos estudios. Asimismo, se están anunciando iniciativas para desarrollar modelos predictivos y de identificación de factores de riesgo.

3. Motivación

La motivación principal de este proyecto nace, por un lado, de la necesidad de entender mejor la enfermedad que está causando graves problemas a nivel global. Por otro lado, se persigue conocer e identificar los factores de riesgo para un mayor control de la enfermedad con medidas más efectivas.

4. Objetivos del proyecto

El proyecto tiene como objetivos analizar la importancia de los distintos factores de riesgo, comprendiendo cómo afectan estos al actual desarrollo de la pandemia en los distintos niveles geográficos y agrupaciones regionales.

5. Metodología de trabajo

Para el trabajo se van a utilizar principalmente como metodologías de trabajo la revisión bibliográfica orientada hacia las distintas partes de los problemas descritos anteriormente y el análisis de bases de datos. Los pasos fundamentales que se seguirán serán, por un lado, el análisis exploratorio inicial de los datos recabados en relación con las variables COVID y, por otro, un análisis multivariante descriptivo por cada nivel geográfico y variable COVID diferente que comprenda herramientas estadísticas como la regresión lineal, ANOVA, clustering o el Análisis de Componentes Principales.

6. Recursos a emplear

Dentro de los recursos que se van a emplear se destacan investigaciones y estudios previos publicados en revistas de divulgación científica como los citados en el siguiente apartado, bases de datos abiertas como la del INE y programas informáticos de simulación y análisis como Matlab y Excel.

1. Introduction

Currently the world is fighting one of the greatest threats to the globalized society in which we live, namely pandemics. The new Covid-19 (caused by the SARS-CoV-2 virus) is spreading at a rate that few expected, leaving people infected in almost every country and disrupting our normal functioning of life. The virus causes a variety of clinical manifestations [1*] including respiratory conditions ranging from the common cold to severe pneumonia with respiratory distress syndrome, septic shock and multi-organ failure.

On 31 December 2019, the Wuhan Municipal Health and Sanitation Commission [2*] (Hubei Province, China) reported to the World Health Organization a group of 27 cases of pneumonia of unknown aetiology, with common exposure to a wholesale market of seafood, fish and live animals in Wuhan City, including seven severe cases. Following the detection of the first infected patient, new cases diagnosed have increased at an exponential rate creating real challenges for public health and straining the boundaries of health systems worldwide.

Although most governments have extended various containment measures such as mandatory confinement or border closures, the extent of the global infection and the duration of the current emergency is not known. Vaccine development and clinical trials of potential treatments are now underway with great hope that these will succeed in controlling the spread of the pandemic.

2. State of play

For now, research into this new pandemic is being carried out in real time with the advances and data that it leaves us every day. The greatest number of investigations to date have been carried out in China because of the longest time available since the origin of the disease in its territory.

However, many institutions both at national and global level are making information and databases available to the public for further studies. Likewise, initiatives are being announced to develop predictive models and to identify risk factors.

3. Motivation

The main motivation for this project arises, on the one hand, from the need to better understand the disease that is causing serious problems globally. On the other hand, it aims to know and identify the risk factors for a better control of the disease with more effective measures.

4. Project objectives

The project aims to analyse the importance of different risk factors, understanding how these affect the current development of the pandemic at different geographical levels and regional groups.

5. Working methology

For the work it will mainly be used as working methodologies the bibliographic review oriented to the different parts of the problems described above and the analysis of databases. The fundamental steps to be followed will be, on the one hand, the initial exploratory analysis of the data collected in relation to the COVID variables and, on the other hand, a descriptive multivariate analysis for each different geographical level and COVID variable that includes statistical tools such as linear regression, ANOVA, clustering or Principal Component Analysis.

6. Resources to be employed

Among the resources to be used, it can highlighted previous research and studies published in popular science journals such as those mentioned in the following section, open databases such as INE and simulation and analysis software such as Matlab and Excel.

Table of Contents

1. THEORETICAL FRAMEWORK 13

a. Pandemic profile 13

i. Definition of pandemic 13

ii. Difference between pandemic, epidemic and endemic 14

b. Risk factors 15

c. Influenza pandemic history 15

i. Influenza Pandemic of 1918 15

ii. 1930s research 16

iii. 1957-1958 Pandemic 16

iv. 1968 Pandemic 16

v. H1N1 Pandemic (2009) 17

d. Contextualization of the Covid-19 18

i. Chronologic spread 18

ii. Origin 19

iii. Symptomatology 20

iv. Measures taken 20

v. Impact on ethnical groups 21

2. STATE OF THE ART 22

a. SARS and MERS research 22

i. Transmission characteristics of MERS and SARS in the healthcare setting: a comparative study 22

ii. MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease 23

iii. Factors Influencing the Response to Infectious Diseases: Focusing on the Case of SARS and MERS in South Korea 23

b. COVID-19 worldwide research 24

i. Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study 24

ii. Modeling the Control of COVID-19: Impact of Policy Interventions and Meteorological Factors 24

iii. Temperature, humidity, and latitude analysis to predict potential spread and seasonality for COVID19 24

c. COVID-19 national research 25

i. National Epidemiological Research Network, Spain (Red Nacional de Investigación Epidemiológica, España) 25

ii. The vulnerability of Spanish provinces to covid-19 according to their age and co-residence structure: implications for (de)confinement 25

3. METHODOLOGY 26

a. Problem description 26

b. Data Collection 26

c. Analysis Method & Hypothesis 28

4. CASE STUDY 34

a. Exploratory analysis 35

i. Community of Madrid 35

ii. Spain 43

b. Analysis I – CAM 54

c. Analysis II – SPAIN 65

5. CONCLUSIONS 82

6. LIMITATIONS AND FUTURE LINES OF RESEARCH 85

SUSTAINALBLE DEVELOPMENT GOALS (SDGs) **87**
APPENDIX **89**
REFERENCES **92**

- Figures -

Figure 1 Dendogram, Matlab ____ *31*
Figure 2 Graphical representation of the P-value ____ *32*
Figure 3 Distribution of the total Incidence in the Community of Madrid ____ *35*
Figure 4 Distribution of the total Incidence by municipality per region ____ *36*
Figure 5 Total cases by region in the Community of Madrid ____ *36*
Figure 6 Distribution of the cumulative incidence reported in May ____ *37*
Figure 7 Distribution of the cumulative incidence reported in April ____ *37*
Figure 8 Distribution of the population rate by age group and total incidence of infection ____ *38*
Figure 9 Distribution of the population rate over 65-year-old per total incidence of infection ____ *39*
Figure 10 Distribution of the total incidence per density on the town ____ *39*
Figure 11 Distribution of air quality level characteristics per total incidence ____ *40*
Figure 12 Distribution total incidence per economic indicators ____ *41*
Figure 13 Distribution total incidence per death cause ____ *42*
Figure 14 Distribution of the total Incidence in Spain ____ *43*
Figure 15 Total incidence per autonomous community ____ *44*
Figure 16 Distribution of the 14-day cumulative incidence ____ *45*
Figure 17 Distribution of the 14-day hospitalizations cumulative incidence ____ *46*
Figure 18 Total hospitalization incidence per autonomous community ____ *46*
Figure 19 Distribution of the 14-day ICU cumulative incidence ____ *47*
Figure 20 Total ICU incidence per autonomous community ____ *47*
Figure 21 Distribution of the 14-day death cumulative incidence ____ *48*
Figure 22 Total death incidence per autonomous community ____ *48*
Figure 23 Distribution of the population rate by age group and total incidence in Spain ____ *49*
Figure 24 Distribution of the tourism and total incidence in Spain ____ *50*
Figure 25 National Health Survey ____ *51*
Figure 26 Health expenditure per capita by total incidence ____ *52*
Figure 27 Google trends historic search by total incidence ____ *53*
Figure 28 Average of concern per Autonomous Community ____ *53*
Figure 29 Scree plot PCA Analysis I ____ *55*
Figure 30 Clustering generated from the principal components in Analysis I ____ *58*
Figure 31 ANOVA means representation Analysis I ____ *59*
Figure 32 Air quality areas ____ *61*
Figure 33 Residual Analysis Model 1 Analysis I ____ *63*
Figure 34 Scree plot PCA Analysis II ____ *66*
Figure 35 Clustering generated from the principal components in Analysis II ____ *69*
Figure 36 ANOVA means representation Analysis II ____ *70*
Figure 37 Autonomous Communities of Spain: names (top) and colored cluster areas (bottom) ____ *72*
Figure 38 Residual Analysis Model 1 Analysis II ____ *75*

- Tables -

Table 1 Used variables in PCA Analysis I *54*
Table 2 Eigenvalues of PCA Analysis I *55*
Table 3 Loadings of principal components Varimax rotation Analysis I *56*
Table 4 Principal component names and parameters Analysis I *57*
Table 5 ANOVA in Analysis I *59*
Table 6 Mean contrast between the different clusters *59*
Table 7 Means of characteristics variables per cluster in Analysis I *60*
Table 8 Clusters characteristics Analysis I *62*
Table 9 Characteristics Regression Analysis I Model 1 ..... *62*
Table 10 Characteristics Regression Analysis I Model 2 (Cluster 1) *63*
Table 11 Characteristics Regression Analysis I Model 3 (Cluster 2) *63*
Table 12 Characteristics Regression Analysis I Model 4 (Cluster 3) *64*
Table 13 Significant factors Analysis I *64*
Table 14 Used variables in PCA Analysis II *65*
Table 15 Eigenvalues Analysis II *66*
Table 16 Loadings of principal components Varimax rotation Analysis II *67*
Table 17 Principal component names and parameters Analysis II *68*
Table 18 ANOVA in Analysis II *69*
Table 19 Mean contrast between the different clusters *70*
Table 20 Means of characteristics variables per cluster in Analysis I *71*
Table 21 Clusters characteristics Analysis II *74*
Table 22 Characteristics Regression Analysis II Model 1 ..... *75*
Table 23 Characteristics Regression Analysis II Model 2 (Cluster 1) *76*
Table 24 Characteristics Regression Analysis II Model 3 (Cluster 1) *76*
Table 25 Significant factors Analysis II Model 1 and 3 ..... *77*
Table 26 Characteristics Regression Analysis II Model 4 ..... *77*
Table 27 Characteristics Regression Analysis II Model 5 ..... *78*
Table 28 Significant factors Analysis II Model 4 and 5 ..... *78*
Table 29 Characteristics Regression Analysis II Model 6 ..... *79*
Table 30 Characteristics Regression Analysis II Model 7 ..... *79*
Table 31 Significant factors Analysis II Model 6 and 7 ..... *80*
Table 32 Characteristics Regression Analysis II Model 8 ..... *80*
Table 33 Characteristics Regression Analysis II Model 9 ..... *80*
Table 34 Significant factors Analysis II Model 8 and 9 ..... *81*
Table 35 Results Analysis II *81*
Table 36 SDGs conclusions *88*
Table 37 Variables names and meaning Community of Madrid *89*
Table 38 Variables names and meaning Spain *91*

1. THEORETICAL FRAMEWORK

a. Pandemic profile

i. Definition of pandemic

At the moment, according to the vast majority of public health experts as WHO and Nations' Ministries of Health, the world is in the midst of a global pandemic caused by the novel SARS-CoV-2 virus. During many years, the experts, politicians and media communication did not reach an agreement [1] about what was exactly a pandemic. Different generations of people had used the term for referring to cholera, AIDS, influenza and other diseases with a widespread, but it never reached a general consensus about the phenomenon.

Not being previously formally defined, the criteria [2] for the declaration of a pandemic was derived from the definitions of "pandemic alert phase" and not from a particular definition of a pandemic. After the H1N1 pandemic, the WHO [3] explained some characteristics for distinguishing seasonal influenza from the influenza pandemics. However, today the criteria for considering a disease as a pandemic is still diffuse.

Even if there is no single accepted definition of the term pandemic, there are common circumstances and parameters that can be identified and mostly accepted. The key features [4] that are considered to be part of a pandemic are:

- Wide geographic extension: it is spread over a large area, several countries or continents, and usually affects a considerable part of the population.
- Disease movement: the expansion can be traced from place to place.

- High attack rates and explosiveness: multiple cases appear within a short period of time.
- Minimal population immunity: there is no existence of vaccine and the human body is not able to face it by itself.
- Novelty: the term is used for targeting new diseases.
- Infectiousness: it has to be transmittable and not behavioural.
- Contagiousness: with a high contagious rate and transmission is usually produced from person to person.
- Severity: it generates serious complications and is a considerable risk for the health.

ii. Difference between pandemic, epidemic and endemic

Despite the fact that pandemic is not a term with an agreed definition, other types of diseases often follow it and are used alongside it. The terms [5] pandemic, epidemic and endemic are linked, but describe situations of varying severity and health alarm.

On the one hand, epidemic [6] is used to describe an increase (often sharp) in a population or in a particular region of cases of a disease that clearly exceeds what is expected under normal conditions during a given period. Some examples of epidemic are the Ebola, Zika, Syphilis or Cholera. On the other hand, endemic [7][8] is a term used to refer to a constantly present disease in a geographical area or in a population group and that which spread can be expected. As endemic examples [9], it is important to mention Chagas disease, Malaria, African Sleeping Sickness and Chicken Pox.

b. Risk factors

According to WHO, a risk factor [10] is defined as a characteristic, condition, or behaviour that increases the likelihood of getting a disease or injury. Even these factors are mostly presented individually, often they relate to each other. Risk factors can be grouped depending on their origin [11] as:

- Behavioral: based on the individual actions and can be modified by each one through lifestyle. i.e. Smoking, unprotected sex and physical inactivity.
- Physiological: those relating to an individual's body or biology. i.e. High blood pressure and high blood sugar.
- Demographic: those that arise from comparing people with each other affecting the overall population. i.e. gender and age.
- Environmental: those related to the context in which we live (from economic and politic factors to physical and biological factors). i.e. Air pollution and access to clean water and sanitation.
- Genetic: based on an individual's genes.

c. Influenza pandemic history

i. Influenza Pandemic of 1918

Considered the greatest recorded pandemic [12]. Named as "Spanish" [13] because during the World War I, Spain (one of the few major European countries which remained neutral) was the country which most informed about the advance of the disease [14]. It was caused by H1N1 and it had presumably an avian origin. The virus infected as much as 40 percent of the global population in 18 months [14]. Having an important mortality rate, the Spanish Influenza [15] killed more people than the Great

War (World War I). The high mortality in healthy people affected vast majority of the society, including from people younger than 5 years old to 20-40 years old, and 65 years and older [16]. This feature made the pandemic unprecedently severe.

ii. 1930s research

Smith, Andrewes, and Laidlaw isolated [17] influenza A virus in ferrets in 1933, and Francis isolated influenza B virus in 1936. In 1936, Burnet discovered that influenza virus could be grown in embryonated hens' eggs. All these events resulted [17] in the study of the characteristics of the virus and the development of inactivated vaccine during the next decade (1942 bivalent vaccine).

iii. 1957-1958 Pandemic

The 1957 outbreak was caused by the virus influenza A subtype H2N2. This virus did not derive from any previously identified subtype, but from a regrouping [18] of genes between human influenza viruses and those from different species of birds. It is considered that the process of gene reassortment was probably biologically supported by pigs. First detected in the Asiatic southwest [19], it spread rapidly all over the world. Worldwide it is estimated that between one million to two million people died [20]. Not having a very high mortality rate, due to its rapid expansion, it had a very serious impact on the economy. This pandemic is also known as *Asian flu* due to its origin.

iv. 1968 Pandemic

The 1968 pandemic was caused by an influenza A (H3N2). This virus comprised of two genes from an avian influenza A virus [21]. First detected in Hong Kong, it is estimated that one million people died around the world. The mortality rate became higher in people 65 years and older [22]. The following year of the pandemic high death

rates were seen. This pandemic is also known as *Hong Kong Influenza*. Nowadays being less severe due to the creation of vaccines, this subtype is found on approximately the 50% of the patients with seasonal flu [23].

v. H1N1 Pandemic (2009)

In June 2009, the WHO declared the H1N1 Pandemic, caused by a novel influenza A named (H1N1)pdm09 virus. The H1N1 was significantly different from the seasonal flu H1N1 viruses. It was composed of a combination of influenza genes not previously identified in people or animals [24]. The pandemic spread mainly through countries with a temperate climate in the form of two waves. Mortality was concentrated in people aged 20 to 50-year-old [25]. It is estimated that estimated to be responsible for between 151,700 and 575,400 deaths worldwide during the first year it circulated [26]. As the H3N2 virus, it is still found on approximately half of the patients diagnosed with seasonal flu [27]. This is the only pandemic flu identified in the 21st century previous to COVID19 Pandemic.

d. Contextualization of the Covid-19

i. Chronologic spread

During the month of December 2019, several cases of pneumonia of unknown origin were detected. All of them shared the common factor of having previously been in a live animal market in Wuhan City [28]. On 31 December 2019, the Wuhan Municipal Health and Sanitation Commission (Hubei Province, China) reported the phenomenon to the authorities, closing the market on 1 January 2020. On January 5, Chinese officials ruled out the possibility that this was a recurrence of the severe acute respiratory syndrome (SARS) virus - an illness that originated in China and killed more than 770 people worldwide in 2002-2003 [29]. On 7 January 2020, a novel type of virus from the *Coronaviridae* family, called SARS-CoV-2, was identified as the causative agent of the outbreak [30]. On January 11, China announced its first death from the virus, a 61-year-old man who had purchased goods from the seafood market and died of heart failure on January 9 [29].

On January 13, WHO reported in Thailand the first case detected outside China, being one woman who had arrived from Wuhan. On January 14, WHO admitted that it may have been limited human-to-human transmission of the coronavirus and that there was a risk of a possible wider outbreak [31], raising fears of a major outbreak as millions travelled for the Lunar New Year holiday. The cities of Wuhan, Xiantao and Chibi in Hubei province were placed under effective quarantine on January 23, suspending air and rail departures and affecting a total of 56 million people [29].

The WHO declared the novel coronavirus outbreak (2019-nCoV) a Public Health Emergency of International Concern (PHEIC) on January 30. In Spain, on January 31, the first case of COVID-19 was confirmed in La Gomera (Canary Islands)

in a German citizen, who had had close contact with another case confirmed [32]. During the month of February, the virus spread rapidly, having confirmed cases on numerous countries as India, Philippines, Russia, Spain, Sweden and the United Kingdom, Australia, Canada, Germany, Japan, Singapore, the US, the UAE and Vietnam.

On March 11, the World Health Organization declared that the public health emergency caused by COVID-19 was an international pandemic, concerned both by the alarming levels of spread and severity, and by the alarming levels of inaction [29]. In that moment, 118,000 cases of infection and 4,291 deaths had been reported in 114 countries [33].

During the next month the spread continued, and, on June 7th 2020, it has been detected 6,962,392 identified cases and 401,251 deaths for COVID-19 worldwide [34].

ii. Origin

To analyze the origin is necessary to study previous coronavirus that produced also respiratory complications. Of the 7 coronaviruses identified, only 3 can produce serious infection in humans [35]. The others are limited to producing mild symptoms. Both severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) were caused by coronaviruses, but they had a significant minor impact leaving 8,000 and 2,519 infections; and, 774 and 866 deaths respectively [36]. According to experts, three theories are mainly discussed as the origin of the COVID-19 pandemic.

Firstly, natural selection in animals before transmission to humans [35]. This origin could be possible due to the fact that similar coronaviruses subtypes have been found in animals on the same region. The possible animals that are suspected to have produced this transformation are the Malayan pangolins and the bats.

Secondly, natural selection in humans [35]. It could be possible that the SARS-CoV-2 predecessor infected humans and modified its genome during unrecognized human-to-human transmission.

Thirdly, selection during passage [35]. Researchers have been studying different coronaviruses in biosafety level 2 laboratories during many years. Also, escapes of SARS-CoV have been documented in the past. For all these reasons, the possibility that the coronavirus was accidentally released by a laboratory must be considered.

iii. Symptomatology

The majority of the patients develop gradually mild symptoms. The most common symptoms [37] are fever, tiredness, and dry cough. Some patients may have aches and pains, nasal congestion, runny nose, sore throat or diarrhea. These patients usually recover without specific treatment. Approximately 1 out of every 6 people who is infected of COVID-19 becomes seriously ill [36] and develops difficulty breathing. People over 60 years old are more likely to develop serious illness and death.

iv. Measures taken

Around the world, important measures have been applied for limiting the spread of the pandemic and preventing the health care system saturation. The most common measures are [38]:

- Lockdown. i.e. Most countries in Europe
- Mandatory quarantine. i.e. Israel
- Closure of schools and educational institutions. i.e. Japan
- Deployment of military troops. i.e. USA
- Sending text alerts, with information on infected people. i.e. South Korea

- Ban on mass meetings. i.e. Austria (100 people max.)
- Tight control including calls, real-time location sharing and in-person visits. i.e. Singapur

v. Impact on ethnical groups

Having this pandemic been originated in China, this fact has produced that Chinese communities situated around the world have suffered xenophobia attacks because of the coronavirus [39]. This cause has been amplified due to the use of "Chinese virus" expression by some politicians in reference to the current pandemic [40]. According to the sociologist Russell Jeung, the number of attacks against oriental communities has increased in USA over 50% in the four first weeks of the outbreak.

2. STATE OF THE ART

In recent months, research related to pandemic diseases has become particularly important. Previously, SARS or MERS crises were studied worldwide, being the number of cases and population infected much lower than at present. At this moment, it is possible to see how international scientific circles are updating the research carried out to date as the new COVID-19 pandemic is studied. It is found abundant information about its theoretical basis and the detail of the number of cases, but there is little information about its relationship with external factors such as environmental conditions, population mobility or per capita GDP. The following is some background information on the research carried out. These studies have been divided into:

a. SARS and MERS research

b. COVID-19 worldwide research

c. COVID-19 national research

a. SARS and MERS research

i. Transmission characteristics of MERS and SARS in the healthcare setting: a comparative study

Comparison of exposure and transmission patterns of large hospital clusters of MERS and SARS. In the research, it is analyzed different models of how both viruses spread comparing different regions as South Korea, Saudi Arabia and Toronto. The work shown the benefits of rapid case detection and strict adherence to infection control measures [41], which can rapidly reduce the risk of super-spreading events and therefore the size of the outbreaks. It also emphasizes the importance of individual patient data [41] and transmission tree information to dissect the progression of subcritical outbreaks of key interest.

ii. MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease

In this research, it is analyzed the importance of Super-Spreaders in the expansion of an infectious disease as the MERS, SARS and Ebola. It is explained that the environment and behavior of individuals can change radically how the virus infects others. Included in this behavior are the Super-Spreading events [42] as hospitals, enclosed housing complexes, and mass transportation. Not only through study of host and virus, but also, environmental dynamics [42] are important to delineate the relative contribution of each factor to the phenomenon of super-spreading.

iii. Factors Influencing the Response to Infectious Diseases: Focusing on the Case of SARS and MERS in South Korea

The study conducted an analysis to understand the correlation of factors influencing the response against infectious diseases from a Korean context. It was highlighted the importance of infectious disease response-related legislation [43], qualitative and quantitative supplementation of disaster response-related human resources, legal grounds for the authority of response personnel and preparation of protective measures. Lastly, it is recommended to establish a system for information [43] sharing and disclosure, as well as a cooperation system involving the central government, the local government, health centers and medical institutions.

b. COVID-19 worldwide research

i. Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study

The purpose of this study is to analyze the relationship between the pollution and the COVID-19 mortality rate. In the analysis, data from more than 3,000 counties in USA was used. Being fine particulate matter ($PM_{2.5}$) related to higher risk in diseases as SARS, seasonal influenza or pneumonia, the study concluded [44] that it also increased COVID-19 mortality rates with a long-term exposure of it.

ii. Modeling the Control of COVID-19: Impact of Policy Interventions and Meteorological Factors

In this study, it is shown that the COVID-19 spread trends in China are similar after an isolation period between 30 to 60 days. These results allow to conclude that a strict isolation strategy [45], as the Chinese, is effective and it has already had a great success in controlling the pandemic. Also, the study warns that the period in the province of Hubei that is necessary to maintain the isolation [45] must be from January to April. This means that for being completely useful, the isolation should be around 3 months. Finally, it highlights the similarities [45] between COVID-19 and SARS, and, therefore, links the changes in the weather as a fundamental factor in the control of the pandemic, as happened in the past with SARS.

iii. Temperature, humidity, and latitude analysis to predict potential spread and seasonality for COVID19

The study identifies a greater spread of the COVID-19 over regions and cities with similar weather indicators [46] as absolute humidity (4-7 g/m^3) and temperature (5-11ºC). In addition, it is stated that historically climatic characteristics such as

humidity or temperature [46] have influenced other human coronaviruses and pandemics such as MERS, SARS or seasonal influenza. Finally, it is argued that a better understanding of how coronaviruses react to changes in these parameters [46] would allow greater control over future outbreaks.

c. COVID-19 national research

i. National Epidemiological Research Network, Spain (Red Nacional de Investigación Epidemiológica, España)

Preparation of epidemiological reports on the situation of the coronavirus pandemic on national territory [47]. The first report published by this institution was on 28th February 2020 and, up to 21st April 2020, a total of 23 reports have been published [47] as more data on the infected population becomes available. The reports include analysis of cases reported in Spain, their demographic, clinical and epidemiological characteristics. The conclusions include an analysis of the infected person, the intensive care unit (ICU) patient and the deceased.

ii. The vulnerability of Spanish provinces to covid-19 according to their age and co-residence structure: implications for (de)confinement

In this study, the spread of COVID-19 is analyzed from the point of view of the demography and mobility inside Spain. It explains that [48] deconfinement strategies cannot be based exclusively on age criteria since the relatively young population, and with less risk of complications, can infect older people within households. Beyond age, deconfinement strategies and epidemiological models should incorporate more demographic and household data to prevent transmission of the disease to older persons in particular. It is recommended that aspects [48] such as density outside and inside the home should also be studied.

3. METHODOLOGY

a. Problem description

The object of this research is contributing to detect the various factors affecting the spread of the new Covid-19 on the Community of Madrid, in addition to identifying the impact of that analysis on the infected, tested and deceased in Spain.

Despite the fact that many investigations have analyzed the figures of infection, effects of the disease, symptomatology and profile of the infected person, not many have studied the factors that make this rapid transmission possible.

b. Data Collection

The data has been obtained from generally accessible public databases, so that the study can be replicated and extended in the future, when more data on the progress of the pandemic is available.

Firstly, the dataset of those infected by COVID in the Community of Madrid has been taken from the open data set: *Covid 19 -TIA by Municipalities and Districts of Madrid* [49]. The rest of the data of the variables have been obtained separately, on the one hand those of the city of Madrid by districts and, on the other, those of the rest of the municipalities.

The data for the Community of Madrid include:

i. Surface and population density from the dataset *Municipalities of the Community of Madrid* [50]
ii. Total population and its division by age groups from *Residents in the Community of Madrid by Age Range* [51]

iii. Air Quality including SO2, CO, NOX… from *Air Quality Network. Data from the current month* [52]

iv. Per capita income by municipality from *Estimate of the Municipal Gross Domestic Product* [53]

v. Social security members in general regime and totals from *Workers affiliated to the Social Security who work in the Community of Madrid*. [54]

The data from the city of Madrid include:

i. Surface by District from *Madrid Districts* [55]

ii. Population density, total population and its division by age groups from *Demographic Indicators by District* [56]

iii. Air Quality from *Air quality. Hourly data years 2001 to 2020* [57]

iv. Per capita income by district from *Bankinter* [58]

v. Social security members in general regime and totals by district from *Annual data city of Madrid: affiliations* [59]

Secondly, the data of those infected by COVID in Spain has been taken from the open data set: *Evolution of coronavirus disease (COVID-19)* [60]. The rest of the data of the variables have been obtained separately including:

i. Total population and its division by its age groups from *Population by Autonomous Community and Autonomous City* [61]

ii. Surface from *Surface extension of the Autonomous Communities and Provinces* [62]

iii. Tourism from *Travellers and overnight stays by Autonomous Community and province* [63]

iv. Health assessment survey from *National Health Survey* [64]

v. Health budgets from *Autonomous Community Budgets: Health* [65]

vi. Search trends in coronavirus, coronavirus test, coronavirus deaths… from *Google Trends* [66]

Among the variables used, the use of cumulative frequencies of 14 days to measure the incidence of the pandemic is due to the virus average latency period [67]. In addition, these 14-day periods start on Wednesday to avoid the weekend effect [68]. This phenomenon is due to data that have not been completed until Monday and are usually posted on Tuesdays.

Quantitative analysis of the data base is conducted by using the software program Matlab. Before starting the analysis, the data must be loaded from the tables “ESP.mat”, “ESP_N.mat”, “CAM.mat”, “CAM_N.mat”. The first two files contain the information related to variables and evolution of the pandemic, explained in previous sections, at the national level of Spain. The other two contain the information at the Community of Madrid (including the City of Madrid). Similarly, files ending in "_N" include standardized variables along with other indicators resulting from the analysis to be developed.

c. Analysis Method & Hypothesis

The hypothesis, based on the information gathered in the previous sections as a theoretical framework and state of the art, is centred on considering all those values that have been shown to have a direct impact on previous pandemics or preliminary studies of the current one and verifying that they have a good relationship with the expansion of outbreaks at the level of the Community of Madrid and Spain. To perform the analysis, three main questions will be answered:

- Does the incidence of the pandemic (infections, deaths...) vary from one population to another depending on each factor and in what way?
- Is it necessary to work with all the variables? Which ones can better explain the dependent variable analyzed?
- Are there patterns between the incidence of the pandemic in different areas or factors?

These steps are intended to make a model in such a way that the dependent variable is explained, as well as the relationship between the factors and different clusters.

d. Statistical concepts and analysis

This part defines and explains the different concepts and types of statistical analysis used in the development of the project. These are:

- Dependent and independent variable.

Dependent variable represents [69] the variable generated as an output after entering some input parameters, which would be the independent variables.

- PCA

PCA is a mathematical tool [70] that is used to reduce a large set of correlated variables into a smaller set of uncorrelated variables (principal components) that explain a good fraction of the variance. Each new principal component created attempts to represent the majority of the remain total variance. The components are structured as eigenvectors whose ponderation parameters are the eigenvalues (loadings).

- PC Scores

PC scores are [71] those created for each observation on each factor extracted that weights the observation on that principal component. It allows to place each variable in a plane of multivariate variability. This variable is standardized.

- Standardized variable

It is defined [72] as the number of standard deviations a given value takes from the mean of its sample or population. Standardized variables are used in statistics to compare data from different samples or populations.

- Clustering

It is a mathematical tool [73] that looks to find homogeneous subgroups (clusters) among the observations in a data set. There are two types of clustering:

  - Clustering of observations: allow data reduction identifying homogeneous groups of similar observations.
  - Clustering of variables: allow dimensionality reduction identifying similarities among variables.

- Hierarchical clustering

Clustering technique based on an agglomerative approach [73], identifying the two most similar clusters and fuse them. The representation of this technique is usually done in a dendrogram. The next figure [74] illustrates the graphical representation of the dendogram:

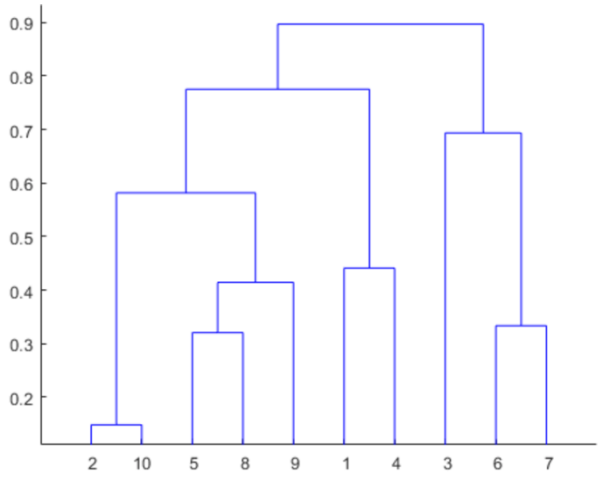

Figure 1 Dendogram, Matlab

- K-means clustering

It is based on a simple approach [73] for partitioning a data set into K distinct, non-overlapping clusters minimizing the quadratic quantization error. The K-means algorithm assigns each observation to exactly one of the K clusters (previously specified).

- Variance

Variance (σ^2) is a measurement [75] of the spread between numbers in a data set. That is, it measures how far each number in the set is from the mean and therefore from every other number in the set. It is calculated by taking the differences between each number in the data set and the mean, then squaring the differences to make them positive, and finally dividing the sum of the squares by the number of values in the data set.

- ANOVA

It is a mathematical method that allows to study [73] the variation of the expectation of the output (dependent variable) conditioned by the groups of the factors (inputs, independent variables). These factors must take discrete values

(even if they were continuous originally). ANOVA is the acronym of Analysis of Variance.

- Regression analysis

It is a mathematical analysis that looks for finding trends [76] of an output, given some input variables, obtaining an equation as a result that can be used for making predictions or finding relations between causes and consequences.

- Model fitting

It is a technique used for finding an optimal regression analysis that reports the p-value and coefficient of determination.

- P-value

It is the probability [77] of finding the observed, or more extreme, results when the null hypothesis (H0) of a study question is true. The term significance level (alpha) is used to refer to a pre-chosen probability and the term "P value" is used to indicate a probability that you calculate after a given study. The choice of significance level at which you reject H0 is arbitrary. Conventionally the 5% (less than 1 in 20 chance of being wrong), 1% and 0.1% ($P < 0.05$, 0.01 and 0.001) levels have been used. The next figure [78] illustrates the graphical interpretation in a normal distribution:

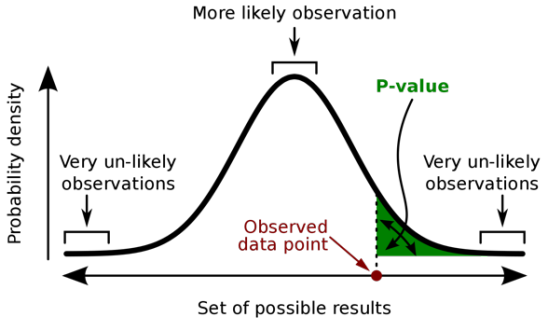

Figure 2 Graphical representation of the P-value

- Coefficient of determination (r^2)

It measures the dispersion degree of the point cloud around the line in a normalized way allowing to compare fits of different models and problem. The coefficient bounds between 0 (large dispersion) and 1 (very correlated).

4. CASE STUDY

The analyses made in this section will be carried out for the community of Madrid and Spain separately. They will be divided into two parts. First, an exploratory analysis of the data obtained will be developed with the help of various graphs and figures. Then, these data will be processed to generate models following different paths but with 4 fundamental processes:

- PCA to reduce the number of variables and a varimax rotation to improve the relation between the principal components and original variables
- Clustering to indicate different homogeneous groups within the data set
- ANOVA allowing to verify the independence between the clusters
- Adjusted regression model that explains the interaction between COVID variables and characteristics of the region analyzed

The conclusions obtained from the section of the case study will be explained at the end of the analysis of each path and summarized in the next section of this document.

a. Exploratory analysis

An exploratory analysis of the variables in the database created is carried out in order to discover and observe different patterns in the values, detect anomalies or errors when manipulating them during the process of creating this database, test initial hypotheses and check ideas based on the help of statistics and graphics. This is the first step to take when you start using a database to make models, not only because of its usefulness as described above, but also to get a general idea of the database.

i. Community of Madrid

- COVID incidence

COVID incidence or total incidence variables refer to the ones that give information about the total positive cases of coronavirus relatively to the population at the day they are reported, in the same way cumulative incidence refers to incidence in the 14 days period previous to the report. The following figures represent the different groups of municipalities according to the historical incidence of the pandemic since the beginning per number of municipalities and per area in the Community of Madrid.

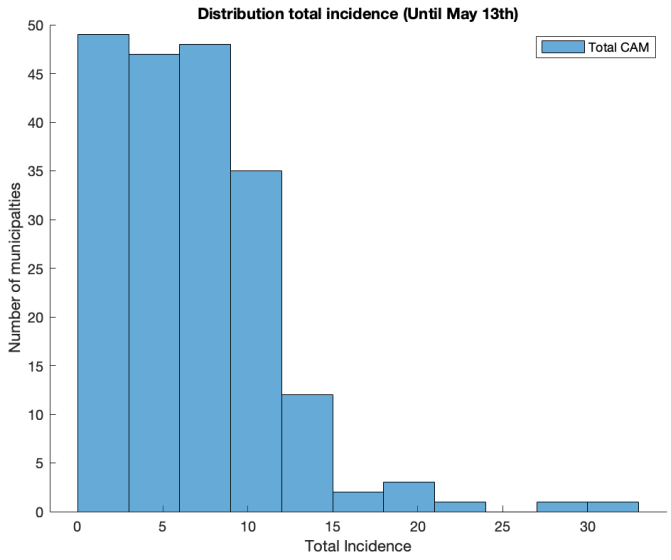

Figure 3 Distribution of the total Incidence in the Community of Madrid

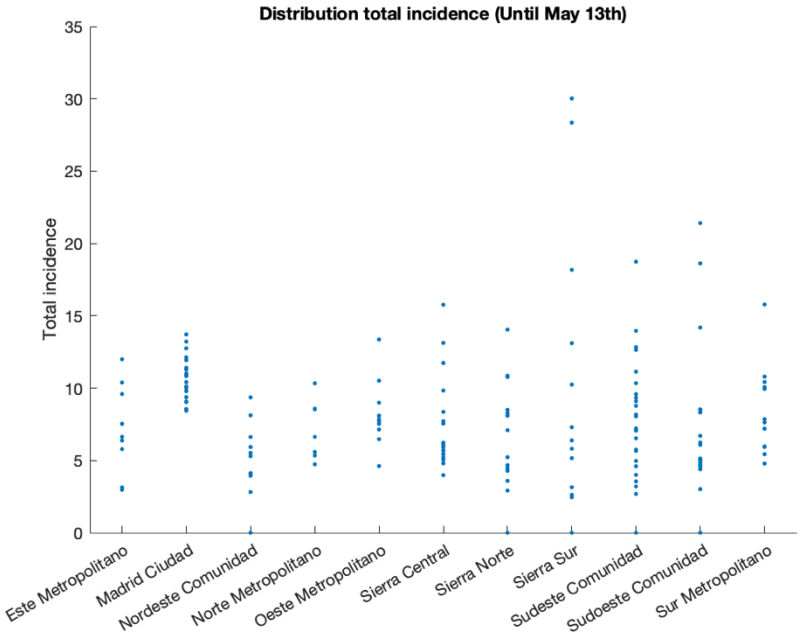

Figure 4 Distribution of the total Incidence by municipality per region

It is observed in the Figure 3 that the vast majority of the 199 municipalities has under 15 cases detected per 1,000 inhabitants. Only 8 exceed that level. In the Figure 4, it can be observed that those outliers are found in non-metropolitan areas. The following graph represents the number of cases distributed in the different administrative regions of the Community of Madrid.

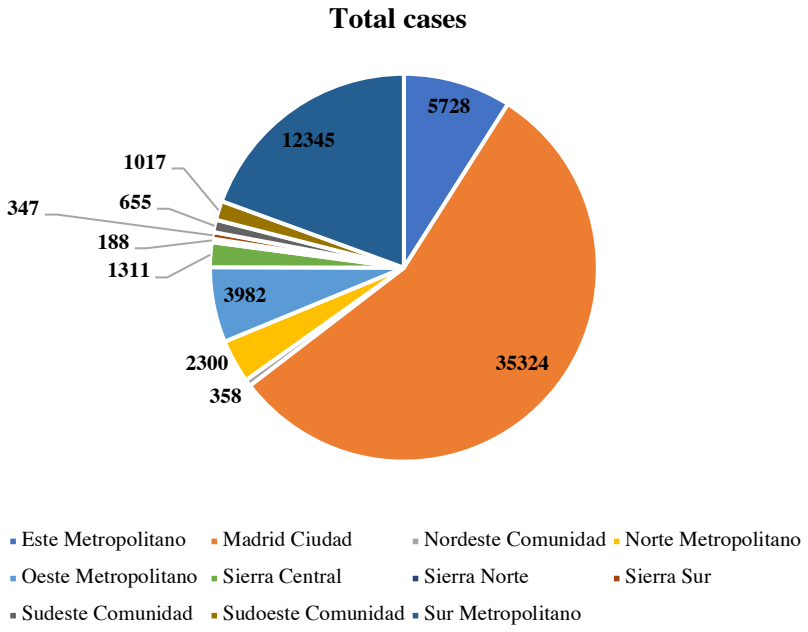

Figure 5 Total cases by region in the Community of Madrid

Analyzing the Figure 5, it can be claimed that the majority of cases occur in the city of Madrid, having a 5% infection rate of the total population.

Although the first case of coronavirus in the Community of Madrid occurred at the end of February, the autonomous administration began to report detailed daily reports by municipality as of April 8th. In the next figures, the evolutions of the cases reported during the month of April and May respectively are represented.

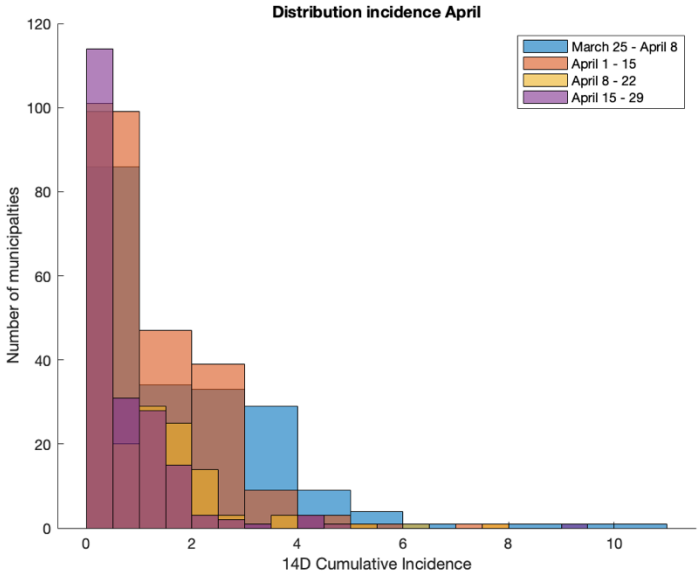

Figure 7 Distribution of the cumulative incidence reported in April

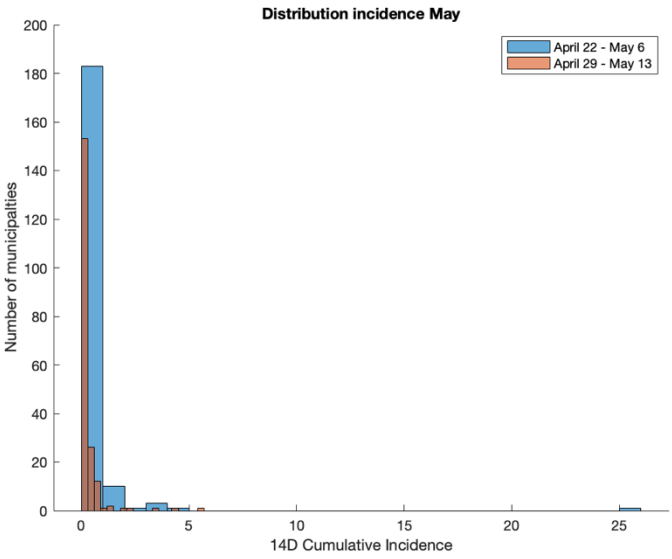

Figure 6 Distribution of the cumulative incidence reported in May

During both periods, a clear decrease on the number of cases detected in a 14-day period is shown over Figures 5 and 6. Also, even the trend is to reduce the pandemic impact, some outliers are presented in a minority way, verifying a characteristic of this pandemic such as the appearance of outbreaks.

- Features of the regions

Various research studies and experts have pointed out the relationship between the COVID pandemic and certain social groups characterized by age, such as high mortality in the elderly (+65 year old). The following figures show the relationship of the population rate in certain age groups to the incidence of coronavirus infection.

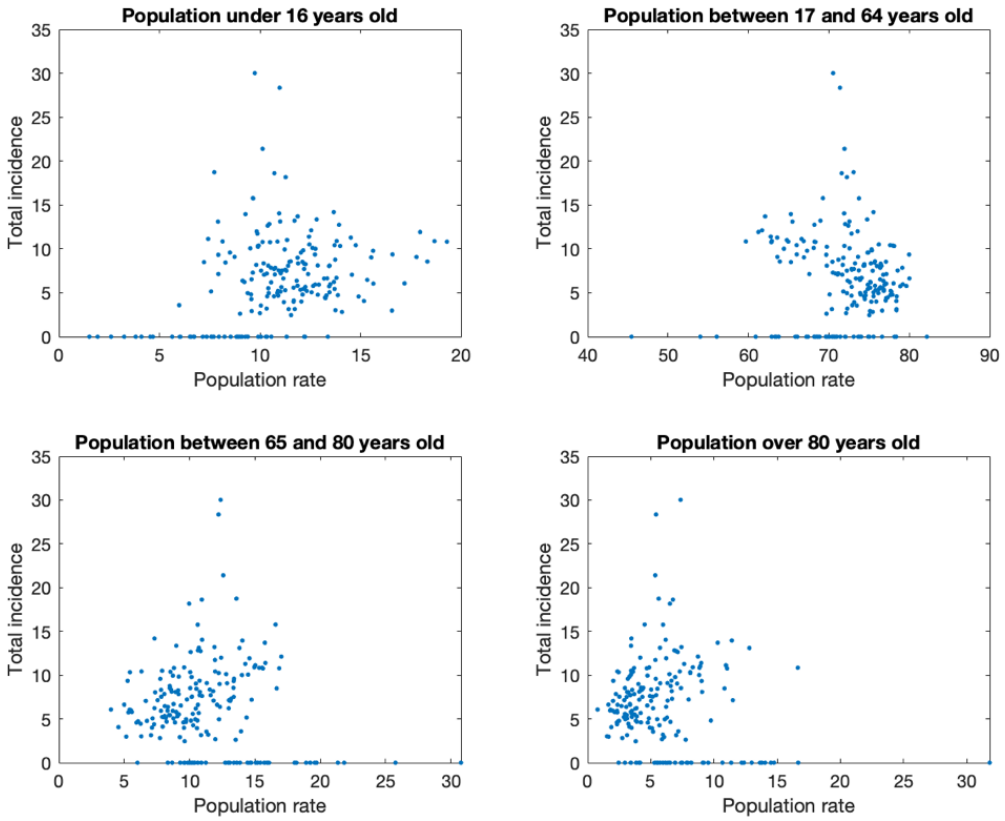

Figure 8 Distribution of the population rate by age group and total incidence of infection

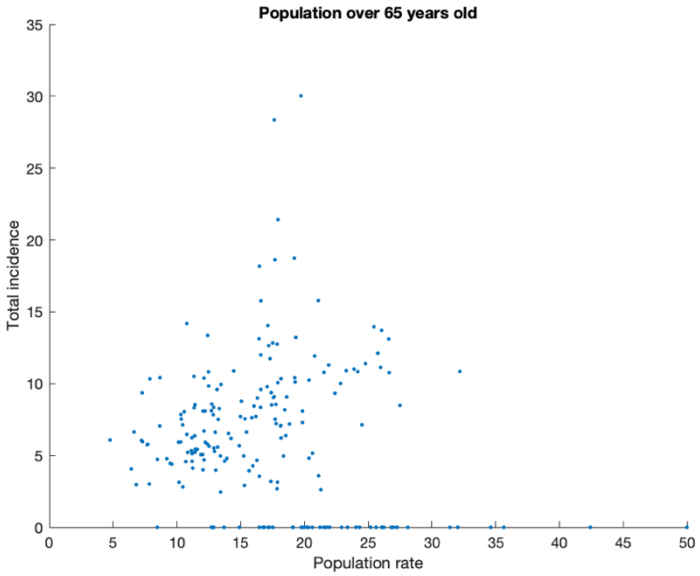

Figure 9 Distribution of the population rate over 65-year-old per total incidence of infection

Only with the exploration carried out in the Figures 8 and 9, it is not possible to determine a great relationship between any particular age group and the total incidence.

During previous pandemics, it has been shown that high population densities lead to a higher probability of infection and therefore generate a greater spread of infection. The following figure represents the population density in relation to the infection rate.

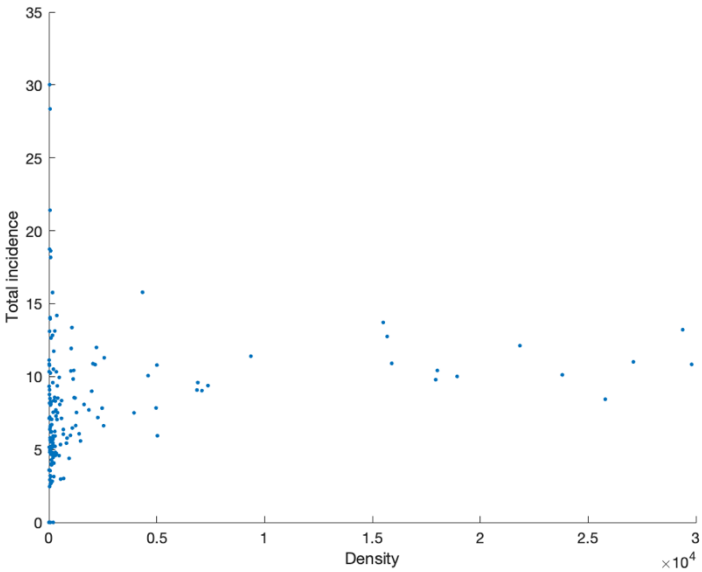

Figure 10 Distribution of the total incidence per density on the town

Zero incidence rates are found only in municipalities with very small densities. In higher infection rates, two trends are observed, one towards a level of infection that increases in proportion to the density and the other towards a very high level of infection at low densities.

Because the particular virus affects the respiratory system, it is interesting to analyze the state of the air quality with the levels of contaminants found in the air. The municipal data have been extracted by the areas of the air quality program of the community of Madrid. The following figure represents the different levels of contaminants with respect to the total incidence in the population by town.

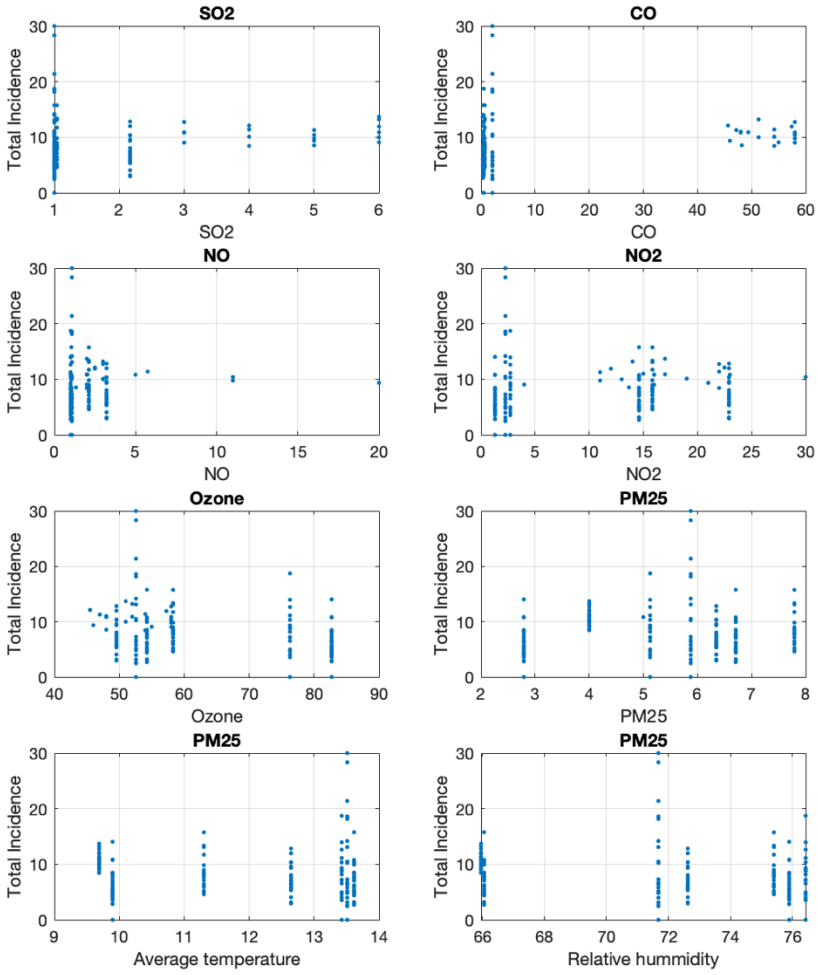

Figure 11 Distribution of air quality level characteristics per total incidence

In general, looking into the Figure 11 and the different indicators, it seems to exist a certain relation between the level of the pollutants and the incidence in some values like PM25 or SO2. Extreme values and dispersion are due to the city of Madrid.

Quality of life, life expectancy and access to healthcare, although in Spain it is a right for all citizens, these indicators tend to improve due to the economic situation in which the population finds itself. This is the reason why the relationship between economic indicators and total incidence is analyzed in the following graphs.

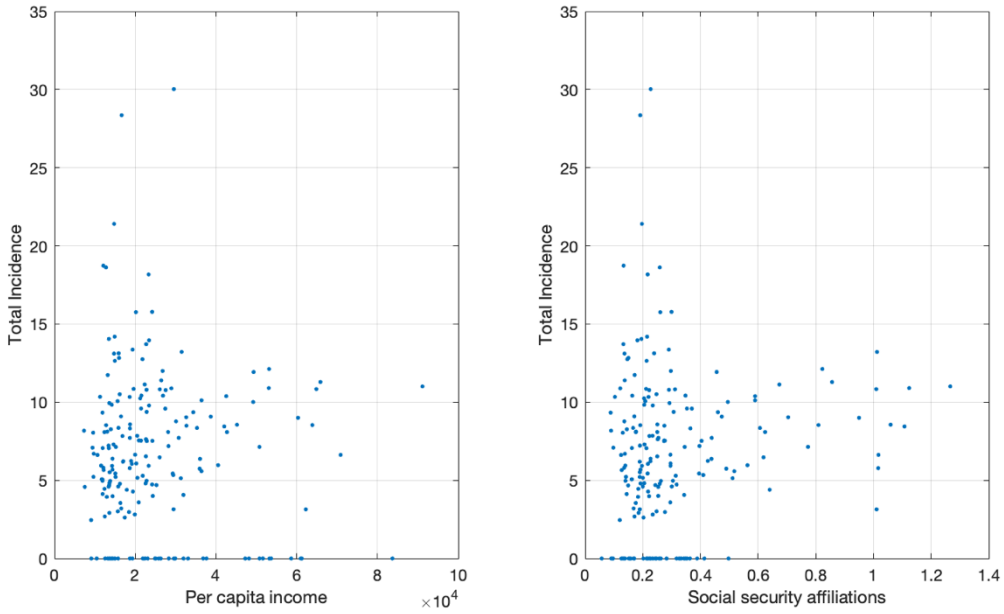

Figure 12 Distribution total incidence per economic indicators

In this initial representation on Figure 12, there is a slight trend towards increased detection of positive cases with a better economic situation. This increase could also be due, in addition to pay testing, to increased economic activity in that area.

One of the risk factors present in many diseases is the genetic one, not only because it can be the cause, but also because it could create a greater

vulnerability to infection. The causes of respiratory and infectious death could be strongly related to this factor, as well as to the set of social circumstances surrounding that population. The following figure shows different causes of death in relation to the total incidence in each municipality.

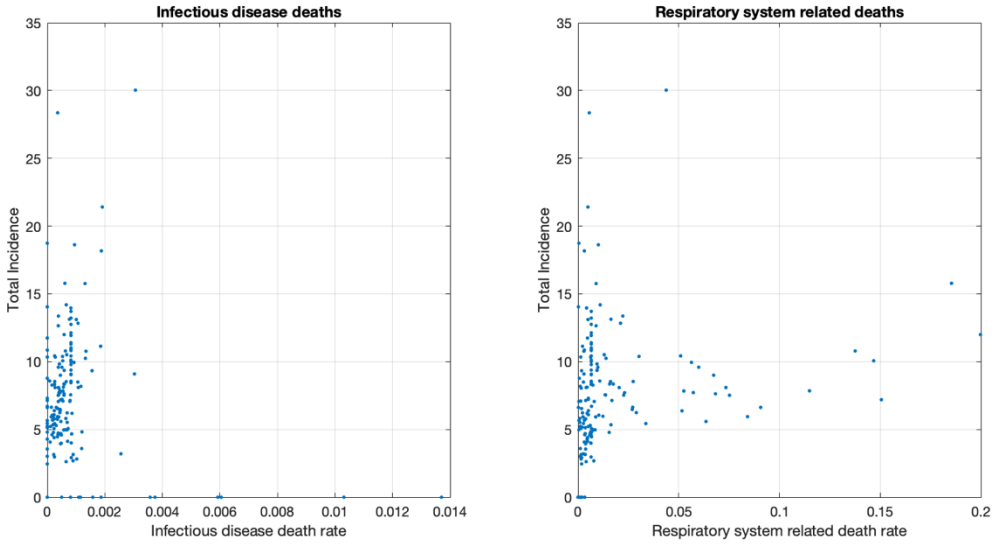

Figure 13 Distribution total incidence per death cause

In the Figure 13, a clear trend seems to be defined on the Infectious disease's deaths and the increase in total incidence. For respiratory system, 2 trends are developed but with also a direct relation between both variables.

ii. Spain

- COVID incidence

As in the community of Madrid, the analysis begin with the following figures representing the different groups of regions according to the historical incidence of the pandemic from the beginning per number of autonomous communities.

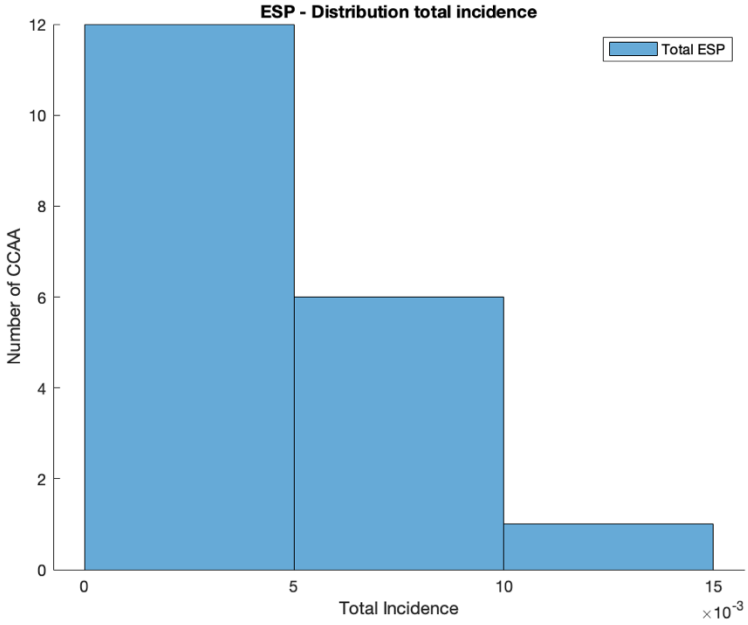

Figure 14 Distribution of the total Incidence in Spain

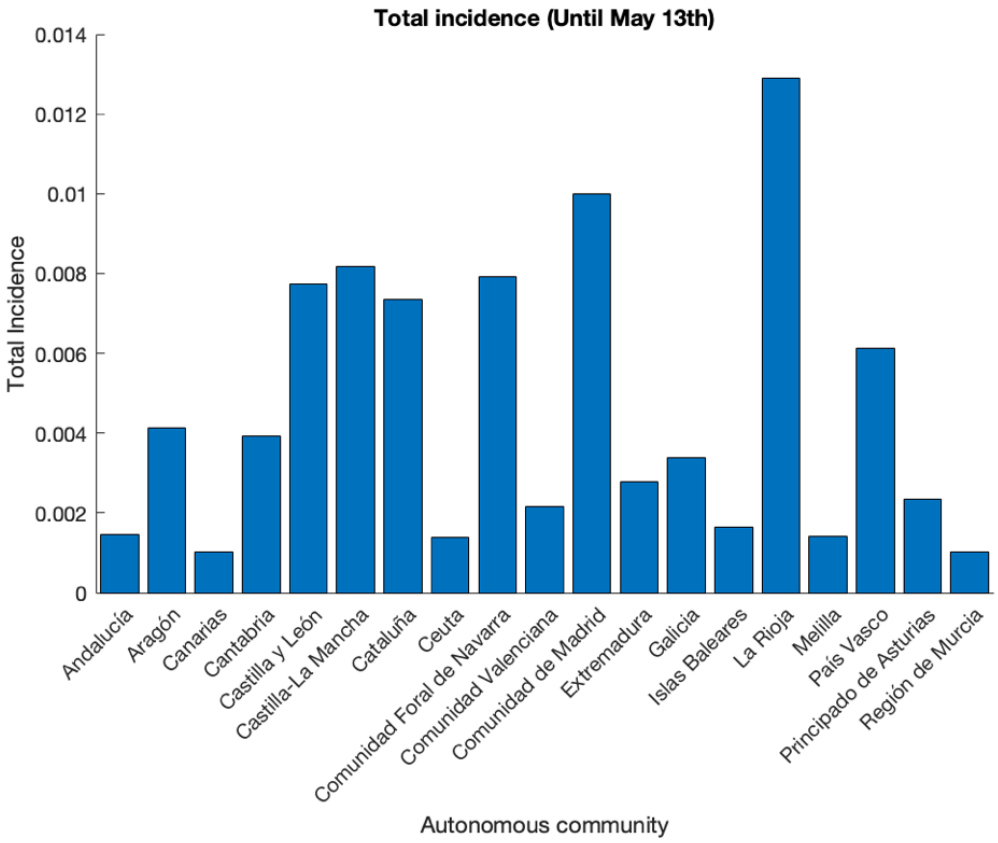

Figure 15 Total incidence per autonomous community

In this case, it can be seen in the Figure 14 that by extending the incidence to the whole of the country, the incidence is reduced, with most Autonomous Communities having a value of less than 10 cases detected per 1000 inhabitants. In the Figure 15 it can be observed that the only region that has a higher incidence than the Community of Madrid is La Rioja with a value of 12.9.

The distribution in the national territory due to the outbreaks has not been homogeneous. The following figure represents the variation in the accumulated cases reported every 14 days by Autonomous Community.

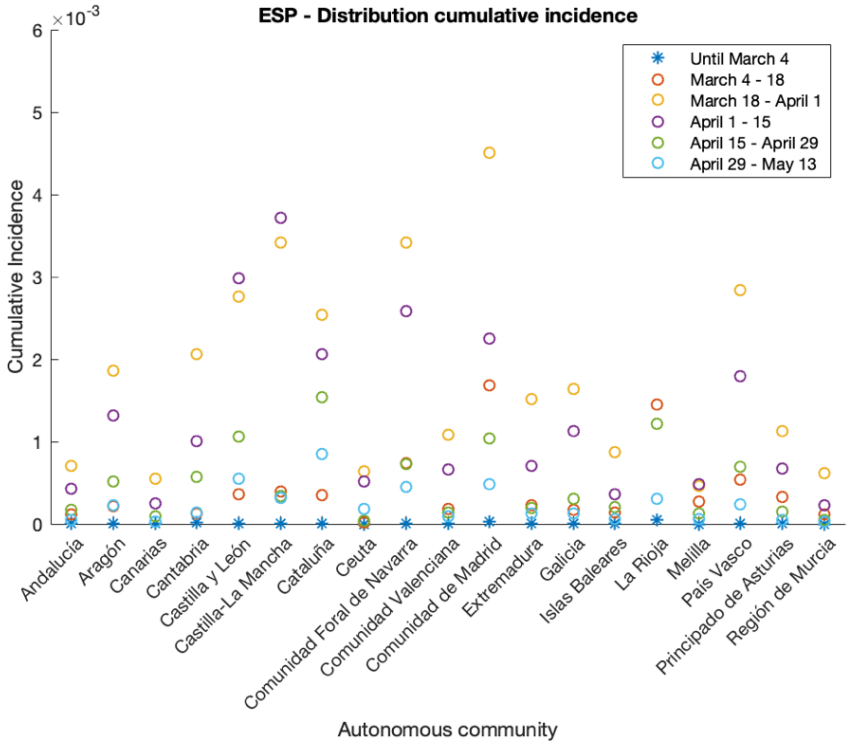

Figure 16 Distribution of the 14-day cumulative incidence

In the graph represented on Figure 16, it is seen how in the first month the incidence grows uncontrollably until reaching the peak of incidence between March 18th and April 15th in all regions and decreasing in the subsequent series. This behavior is very similar to the development of the cases in the Community of Madrid.

In addition to the positive cases detected, and unlike in the Community of Madrid, there is reported national data on the rate of deaths, hospitalizations and admissions to ICUs. This data is represented on the next figures.

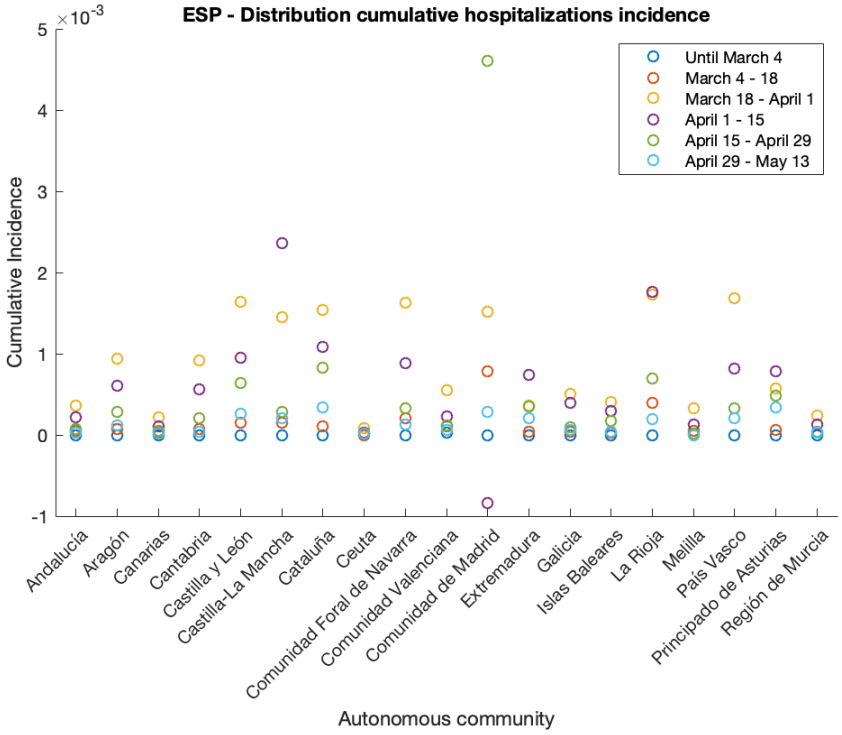

Figure 17 Distribution of the 14-day hospitalizations cumulative incidence

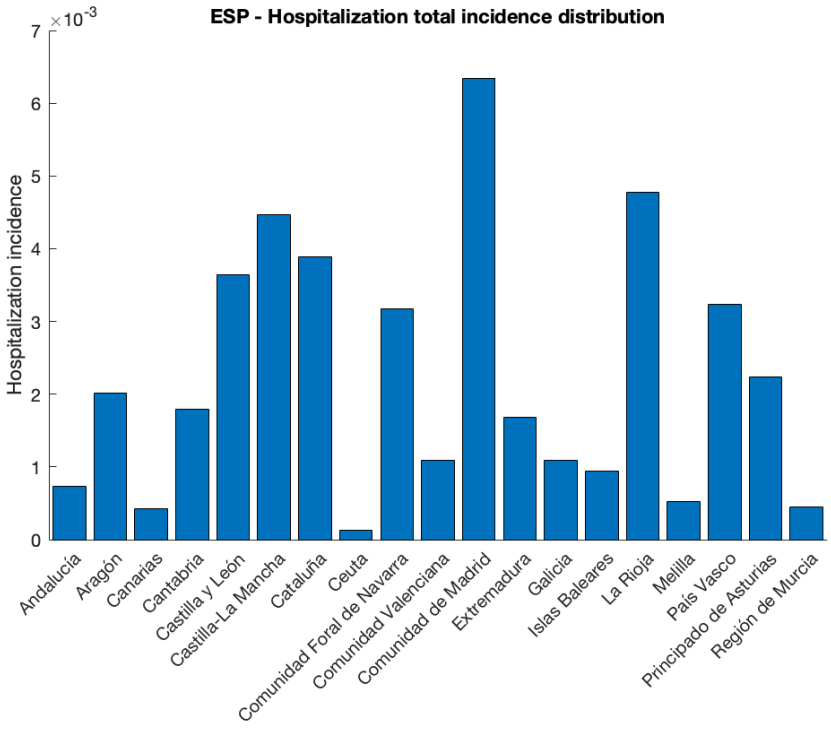

Figure 18 Total hospitalization incidence per autonomous community

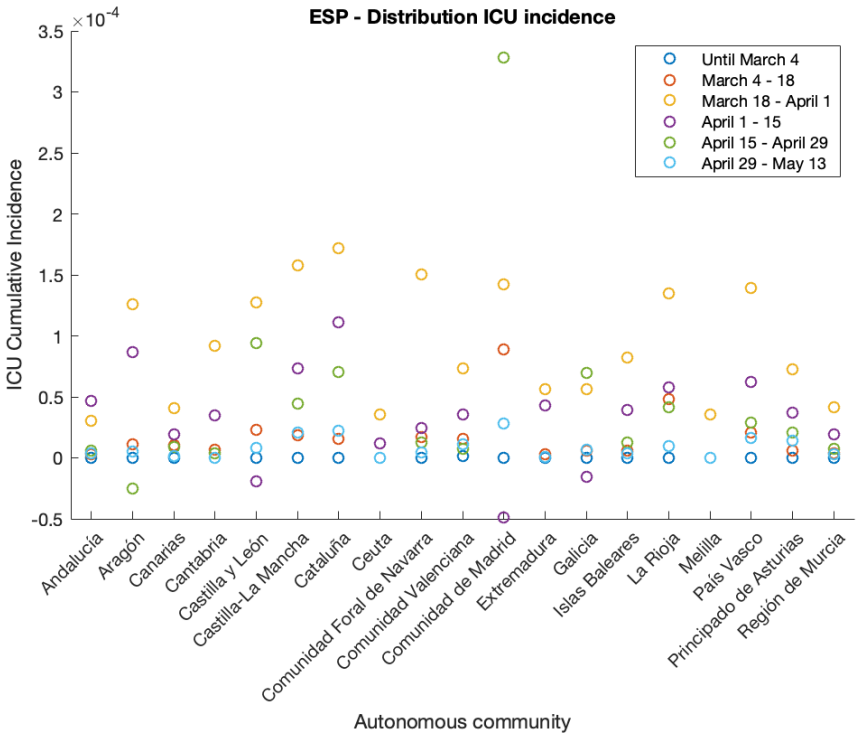

Figure 19 Distribution of the 14-day ICU cumulative incidence

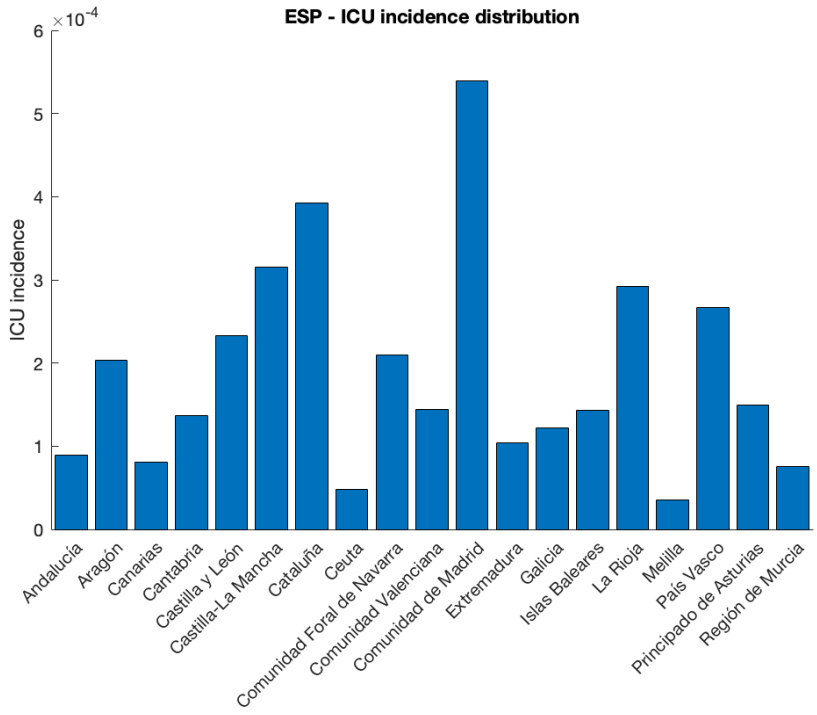

Figure 20 Total ICU incidence per autonomous community

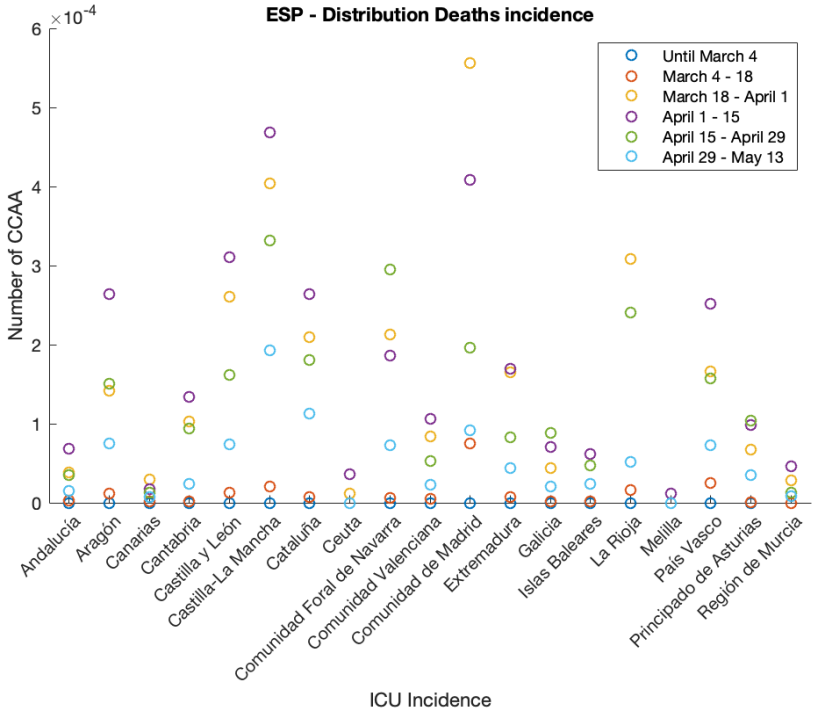

Figure 21 Distribution of the 14-day death cumulative incidence

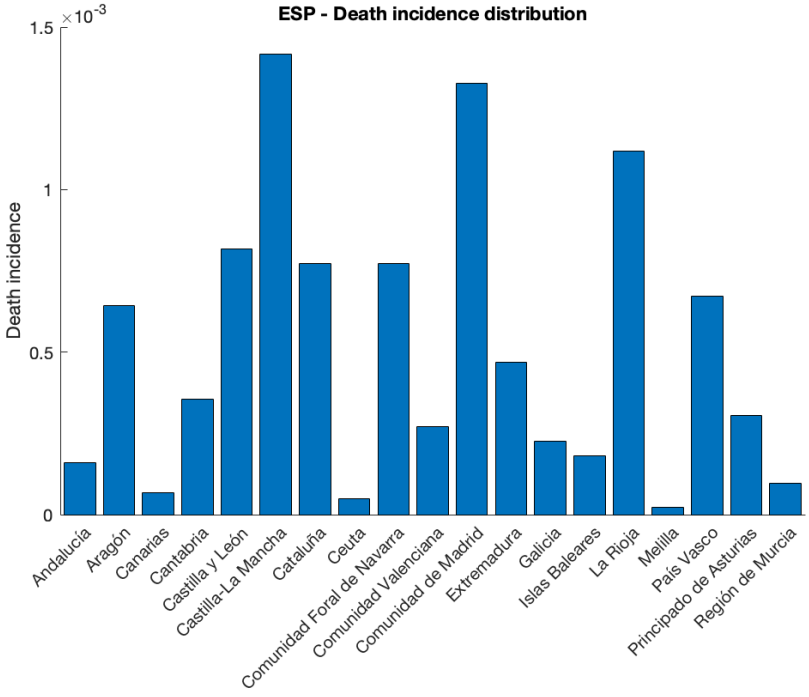

Figure 22 Total death incidence per autonomous community

In all the three variables represented on Figures 17, 19 and 21 it is observed a similar pattern in the distribution over the time of the incidence,

having reached the peak between March 18th and April 8th, and having proportional values in total between them. It must be highlighted the high death rate in Castilla-La Mancha over the rest of Autonomous communities despite having a smaller ICU and hospitalizations rate (Figure 22 vs 18 and 20).

- Features of the regions

Due to the reasoning explained above for the Community of Madrid, age is considered an important factor and the following graph shows the relationship of the population rate in different age groups with respect to the total incidence rate.

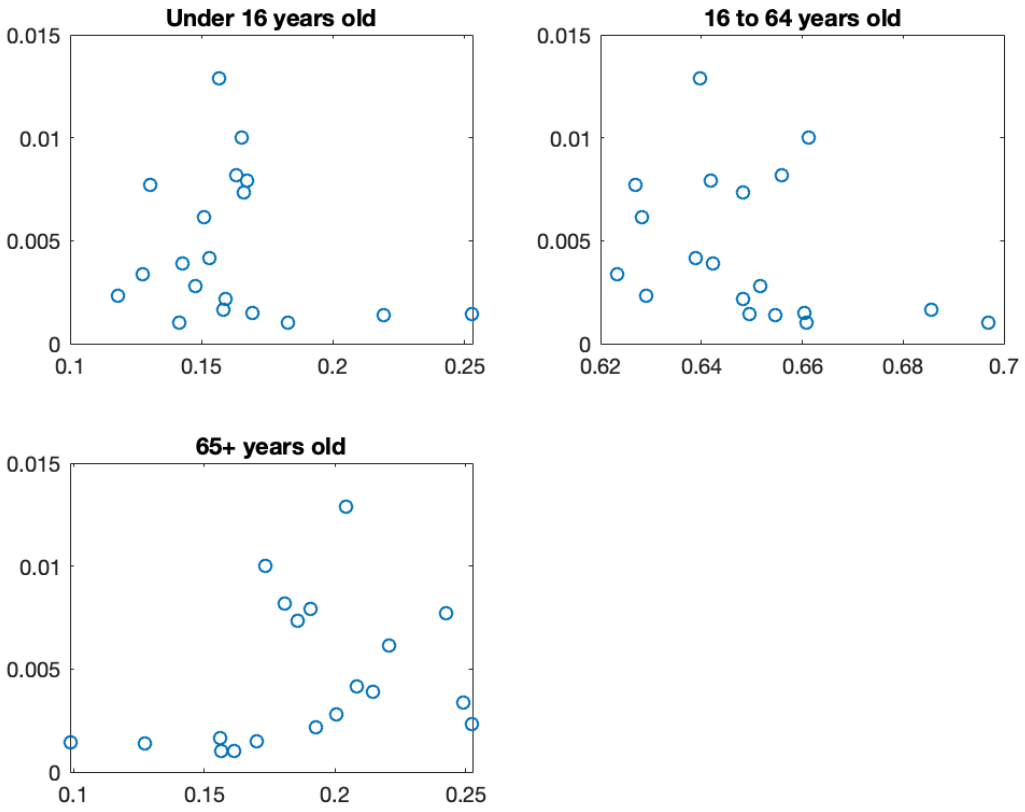

Figure 23 Distribution of the population rate by age group and total incidence in Spain

Mobility becomes a key factor in the spread of an infectious disease. In the case of a global pandemic, tourists can become vectors of transmission. Following this idea, we proceed to analyze the relationship between tourists and incidence rate.

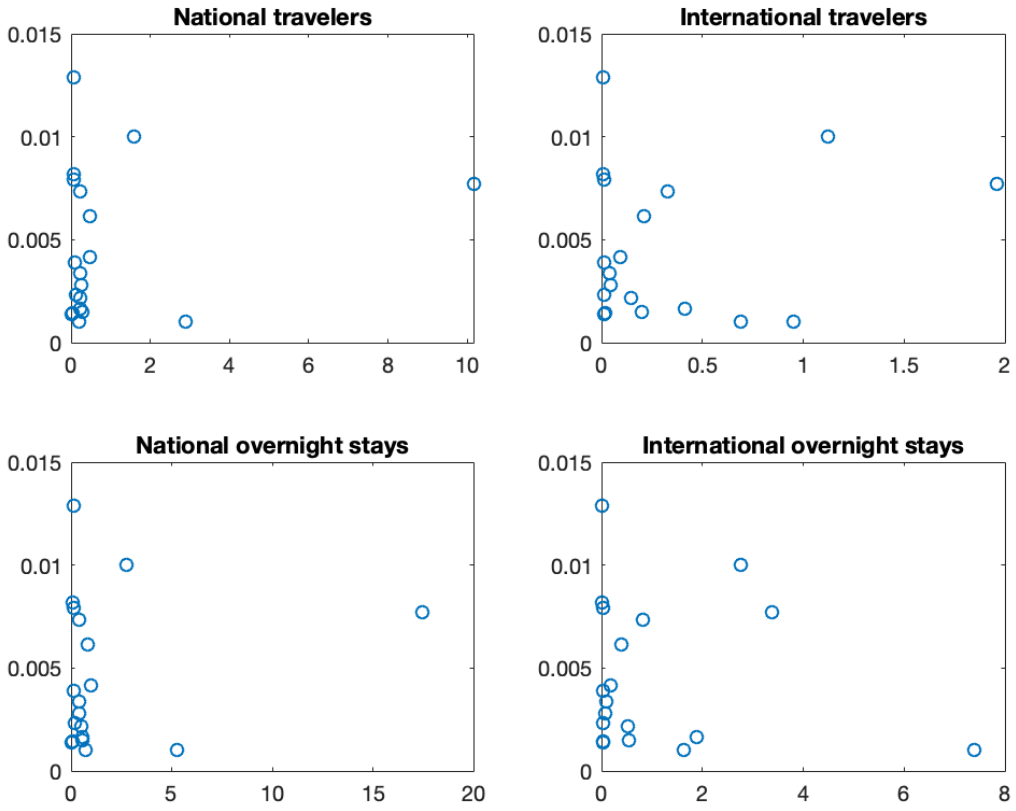

Figure 24 Distribution of the tourism and total incidence in Spain

The graph of the Figure 24 shows a slight trend in the number of international tourists in relation to the incidence.

Every year, the Spanish National Institute of Statistics carries out a quality survey on the public health service by Autonomous Communities and, being an indicator on the perception of the services, it is interesting to compare it with the incidence. The following figure shows the results of the mentioned survey.

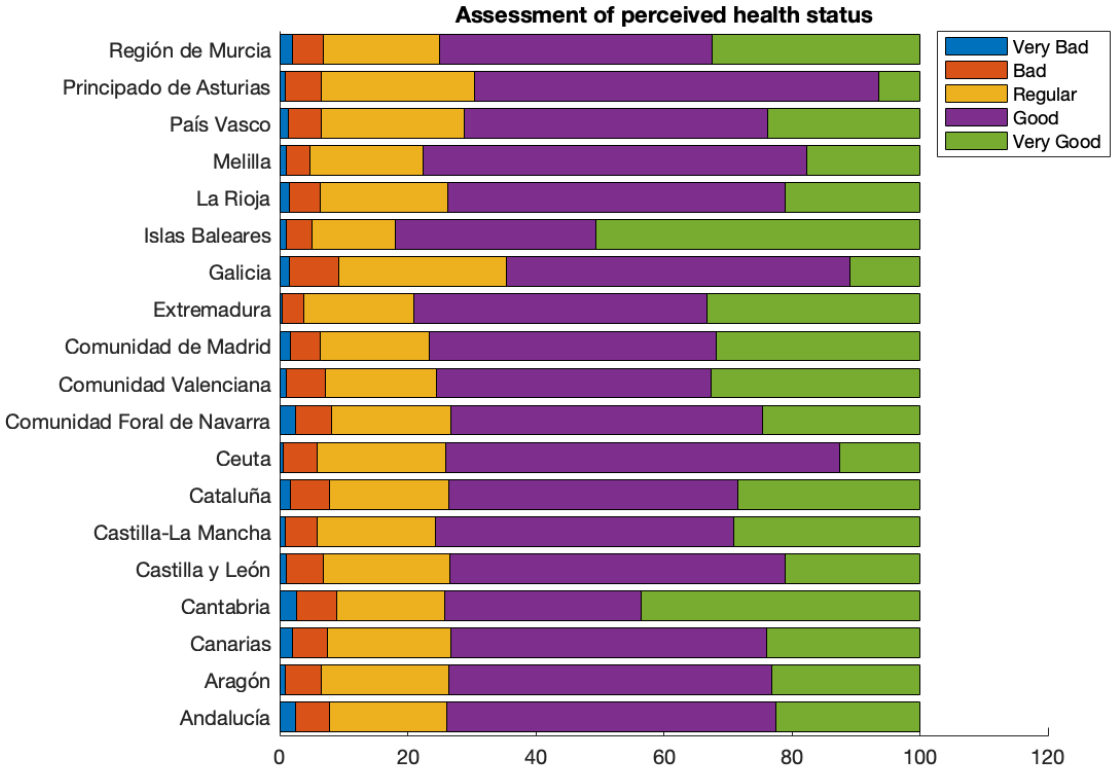

Figure 25 National Health Survey

From the survey results on Figure 25, it can be concluded that the community which perceives a worse health state is Galicia and the one which perceives the best one is the Balearic Islands. In the following analysis the relation with the COVID will be discussed.

Another important variable that could be related to the impact of how the health care system work is the resources and budget that this one has. In the next figure, the health expenditure is shown in comparison with the total incidence.

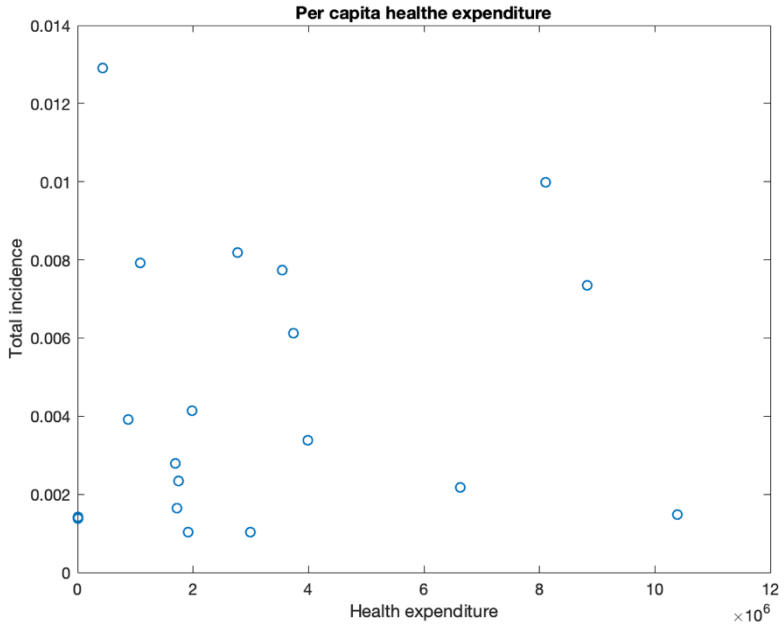

Figure 26 Health expenditure per capita by total incidence

The initial analysis on Figure 26 does not show a strong relationship between health expenditure and the total incidence of the COVID 19 pandemic directly in the graph.

More and more, the internet and connectivity has allowed that all the information can be registered. Google provides data about how the different searches on the internet vary its popularity during the time, giving important details in real time. In the next graph, it is shown the search of 4 topics related with the coronavirus and its impact with the incidence. Additionally, a variable has been created with the average of the 4 others called average of concern.

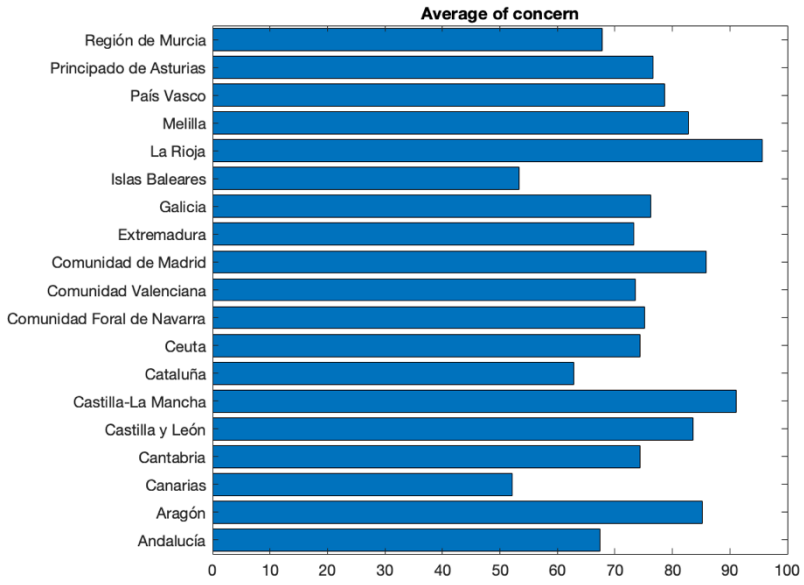

Figure 28 Average of concern per Autonomous Community

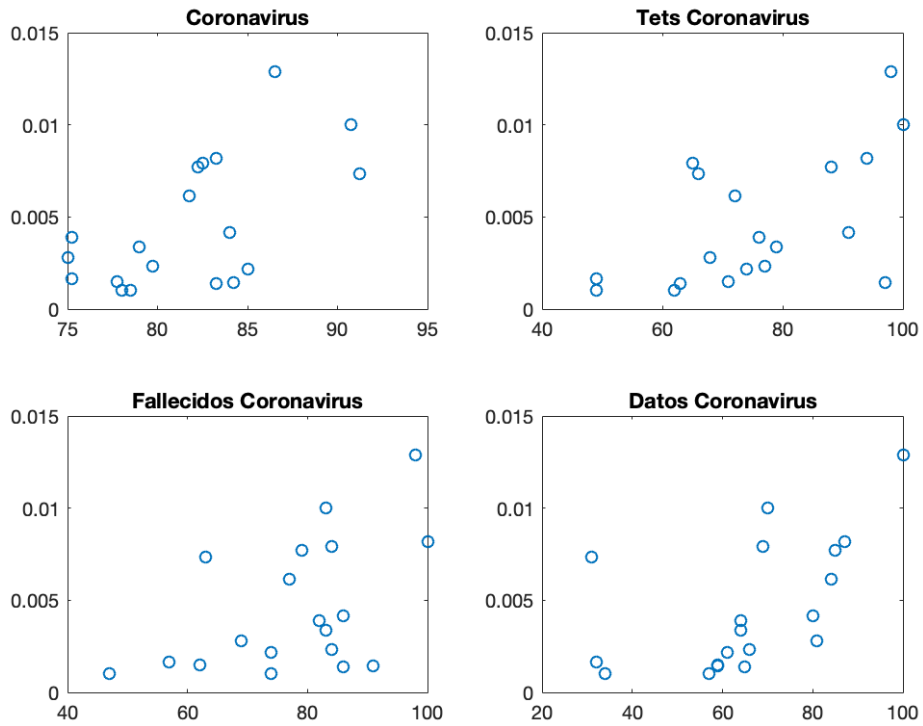

Figure 27 Google trends historic search by total incidence

In general, some trends can be derived from the initial analysis represented on Figure 27 and it could be a good relation between the Google Trends variables and the expansion of coronavirus in Spain.

b. Analysis I – CAM

In this part the analysis will be performed from the characteristics and features variables from each municipality in the Community of Madrid, shown in the previous section, generating new variables that can explain the existence of different clusters and its relationship with the pandemic.

As explained, the first step made is the creation of a PCA over different groups of variables to try to reduce the number of them. For this analysis, some variables have to be filtered in order to get full independent variables ready for the varimax rotation. The variables used are 16 from 18 originally collected, removing 2. This way, none of the loadings are zero. In the next table, the variables used are represented.

'Pob16aos'	'Pob80'	'Pob65'	'DecEI'
'SO2'	'CO'	'NO'	'Ozono'
'MSistResp'	'Densidad'	'PM25'	'NO2'
'TempMedia'	'HumRel'	'Rentapercapita'	'Totalafiliados'

Table 1 Used variables in PCA Analysis I

The results of PCA Analysis report can be interpreted in two different ways. The first criterion is based on the explained variance (Figure 29), from which all the variables that allow more than a certain minimum percentage of explanation are considered. The second criterion focuses on the eigenvalues that represent each of the main components (Table 2), so that from a minimum value that component is considered valid or not. Both methods are represented below.

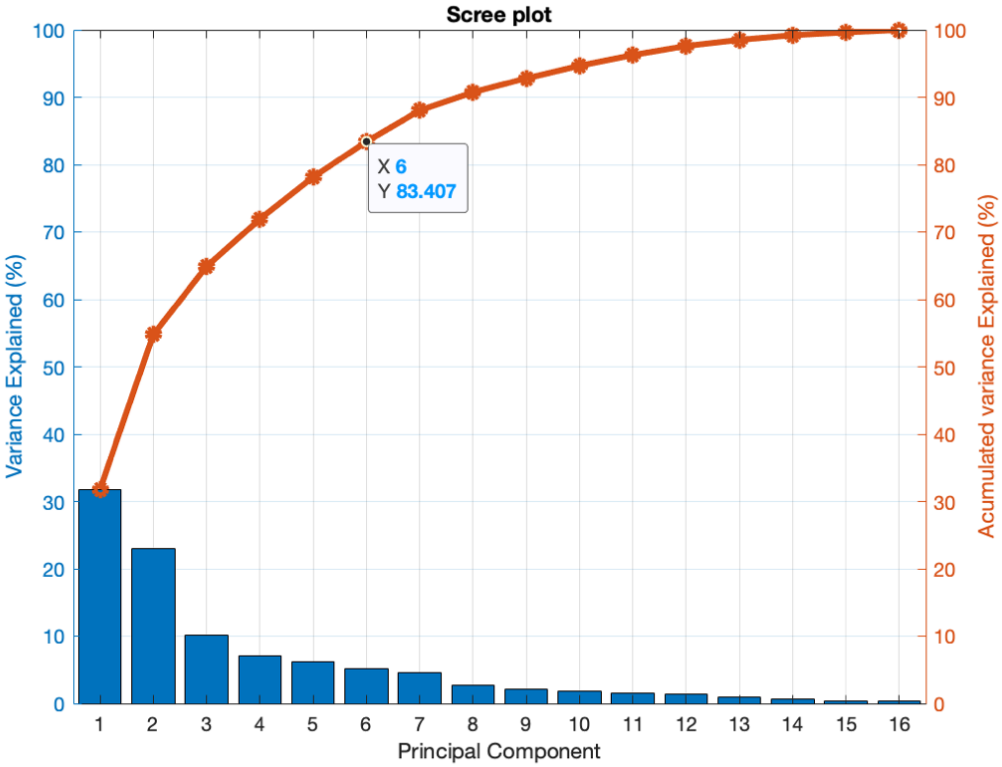

Figure 29 Scree plot PCA Analysis I

Principal Component	Eigenvalue
PC1	**5,08973929**
PC2	**3,67365841**
PC3	**1,62060487**
PC4	**1,12499969**
PC5	0,99877223
PC6	0,83826097
PC7	0,74329892
PC8	0,43134798
PC9	0,33175371
PC10	0,29767009
PC11	0,25463161
PC12	0,21487057
PC13	0,14509662
PC14	0,11165318
PC15	0,06687055
PC16	0,05786584

Table 2 Eigenvalues of PCA Analysis I

After executing the process represented in the Figure 27, it is extracted that 6 components (a third of the total) are necessary to explain more than 80% of the variance (83.41%).

Analyzing the eigenvalues in Table 2, it can be concluded that only 4 principal components have a value superior to 1, which is the average variance of the system. These 4 components are representative ones. The rest are below the mean and they are not considered.

Taking the most restrictive criterion, 4 is the number of principal components desired for developing the analysis. This way, the new components are built using the varimax rotation getting the 4 variables with their scores from the original observations. In the next table, the characteristics of the variables are represented.

	PC1	**PC2**	**PC3**	**PC4**
'Pob16aos'	0,5316	0,1546	-0,623	-0,106
'Pob80'	0,0084	-0,264	**0,8498**	0,0209
'Pob65'	0,0154	-0,245	**0,9602**	0,1123
'DecEI'	-0,002	-0,091	0,0664	-0,007
'MSistResp'	-0,011	0,4134	-0,027	0,1143
'Densidad'	0,6998	-0,048	0,0759	0,3498
'PM25'	-0,006	**0,8713**	-0,279	-0,058
'NO2'	0,4594	0,5597	-0,312	0,2207
'SO2'	**0,8431**	-0,06	-0,012	0,1921
'CO'	**0,8916**	-0,191	0,0464	0,2231
'NO'	0,553	0,0759	-0,103	0,0191
'Ozono'	-0,49	**-0,819**	0,1969	-0,073
'TempMedia'	-0,227	0,7378	-0,141	-0,252
'HumRel'	-0,662	-0,43	0,0583	-0,068
'Rentapercapita'	0,1673	-0,037	0,2455	**0,8006**
'Totalafiliados'	0,3187	0,1585	-0,111	**0,7656**

Table 3 Loadings of principal components Varimax rotation Analysis I

The result from the varimax rotation represented in the Table 3 is the creation of 4 new principal components that have a strong relationship with 8 original variables. The first component (PC1) represents principally pollution related information, having SO_2 and CO levels an important positive impact. The second one (PC2) is also related with the pollution information, but in this case are PM25 and Ozone levels the variables with a strong positive and negative impact respectively. The third one (PC3), has relevant information about the age distribution, with a positive impact on old people (Over 65 and 80 years-old) and a negative one on youngers (Under 16 years-old). The last component (PC4), has information mainly about the economic related aspects with a positive impact of Per capita income and total social security affiliations. With this information, the principal components have been named in relation with the variables that are more related to them in the following table:

Principal component (varimax rotation)	**Name**
PC1	Atmospheric Pollutants
PC2	Atmospheric Particulate Matter
PC3	Aging of the population
PC4	Economy

Table 4 Principal component names and parameters Analysis I

After obtaining the variables of study, the next steps aim to develop a clustering model based on the K-means method and validated by ANOVA. For doing this process, the 4 principal components obtained from the varimax

rotation of the original components are going to be used. In the next figure, it is shown the clusters behavior according to the principal components.

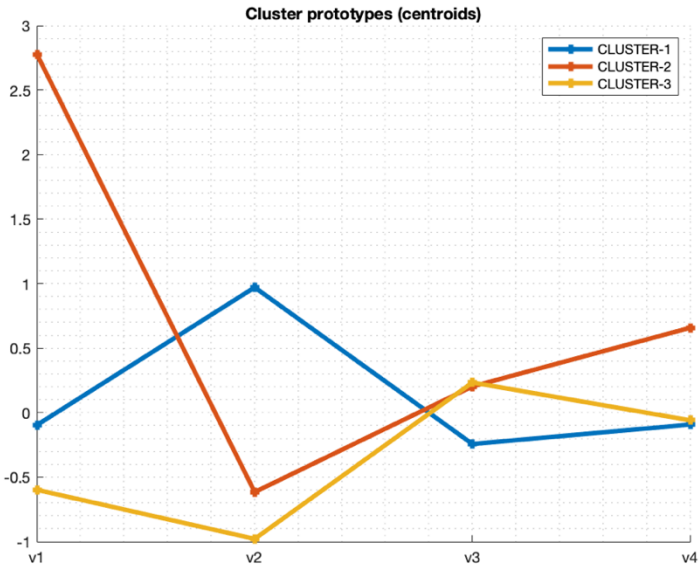

Figure 30 Clustering generated from the principal components in Analysis I

In principle, the clusters seem independent of each other and sufficiently different centroids in the Figure 30.

The elaboration of ANOVA between the different clusters has been developed in order, on the one hand, to verify that they are sufficiently different from each other, and on the other hand, to be able to understand how they are in general. The dependent variable used to do this analysis is the total incidence of the COVID 19 at May 13th (MIncidTot2). The following table represents the previously described analysis developed with different number of clusters.

Number of clusters (K)	Outliers	Prob>F	MS Error
2	6	3.55e-5	0.9213
3	**6**	**3.62e-12**	**0.7721**
4	6	5.37e-10	0.802
5	12	3.08e-11	0.769
6	12	7.95e-12	0.75

Table 5 ANOVA in Analysis I

The number of clusters finally developed has been chosen according to the Table 5. Being accepted by the ANOVA analysis different options, 3 clusters allow to have a minimum number of outliers and Prob>F (p-value) among a good MSE. Specifically, from this analysis can be extracted the following results.

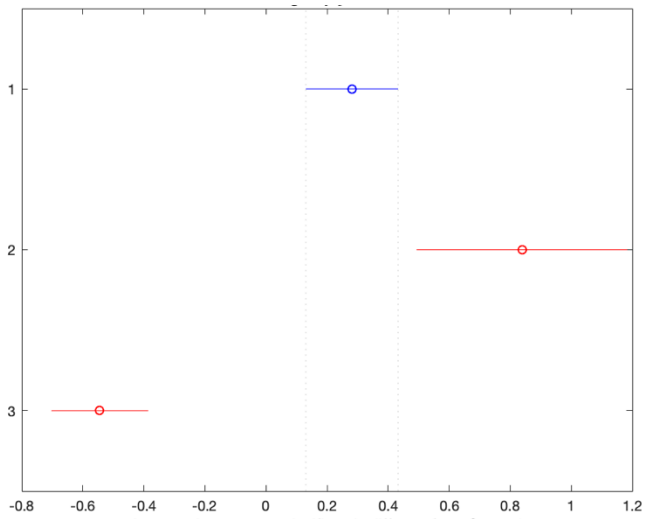

Figure 31 ANOVA means representation Analysis I

Cluster A	Cluster B	p-value
1	2	0.0232
1	3	2.16e-9
2	3	1.32e-9

Table 6 Mean contrast between the different clusters

After ANOVA analysis represented in the Figure 31 and Table 6, it can be concluded that all 3 clusters have means significantly different from each other with valid p-values in all the post-hoc mean contrast analysis. For a better understanding of how the clusters inside are, an analysis based on the characteristic variables is shown in the following table. Also, a figure has been included showing the different areas used for locating geographically the clusters.

Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**
Number of towns	96	21	82
Population	3,17 M	3,27 M	1,56 M
Density	843,95	**14254,96**	54,11
% Under 16 y.o.	11,65	**14,35**	9,08
% between 16 and 64 y.o.	74,55	65,62	70,99
% >65 y.o.	**13,80**	20,03	19,93
% >80 y.o.	**4,17**	7,14	7,54
Rel. humidity	71,12	65,96	76,04
Avg. temperature	12,83	9,69	10,89
NO2	14,04	17,01	**1,69**
SO2	1,29	**4,33**	1,00
CO	0,88	**50,40**	0,43
NO	1,82	**4,15**	1,00
Ozone	53,66	53,08	**80,91**
PM2.5	**6,67**	4,05	3,45
Per capita income	22267,03	**40590,98**	23989,95
% Employed	0,24	**0,54**	0,14
% SS Affiliations	0,32	**0,60**	0,22
% Infectious disease deaths	0,00058	0,00082	0,00084
% Resp. system deaths	**0,02659**	0,00660	0,00221

Table 7 Means of characteristics variables per cluster in Analysis I

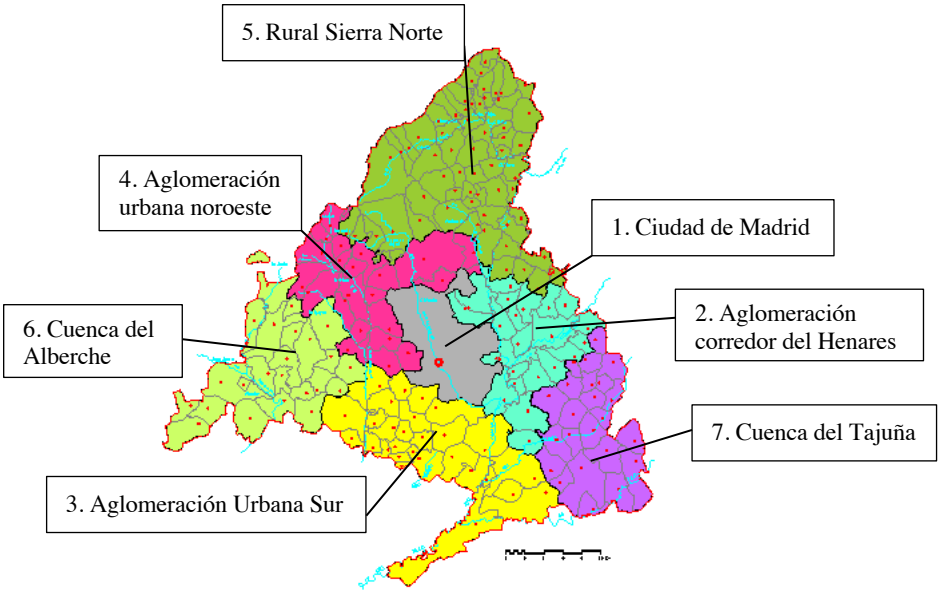

Figure 32 Air quality areas

From the previous Table 7, some conclusions can be extracted. The first cluster is the most numerous with 96 municipalities and districts. This cluster has less proportion of old people (population of more than 65-year-old is around 14% vs 20% than the other two). It has very high levels of fine particles which it could be related also with a very high rate of respiratory system related deaths. The cluster 1 is the second one in total incidence and it is located in the complete areas 2, 3, 4 and 6 of Figure 32, grouping all community urban agglomeration areas.

The cluster 2 groups the region of the city of Madrid, in the center of the autonomous community. This cluster has the highest total incidence rate of the 3 clusters. It has a younger population in proportion and very high levels of pollution in all the indicators, being especially extreme in SO2, CO and NO with respect to the other two groups. In the economic aspect, the number of people working in the city and the higher per capita income of its inhabitants stand out.

The cluster 3 consists of 82 populations located in zones 5 and 7 of Figure 32. As these are more rural areas, it has a much lower population and density than the other two clusters. In the same way, pollution levels are also very low. The economic activity is also smaller. This cluster has the lowest total incidence rate.

Cluster no.	No. of municipalities	Air quality areas	Characteristics
1	96	2, 3, 4, 6	- Urban area - Less old people - High pollution
2	1	1	- Urban area - Younger people - Very high pollution
3	82	5, 7	- Rural area - Low pollution

Table 8 Clusters characteristics Analysis I

Once the principal components to be used and the clusters generated have been analyzed, the explanatory models of the COVID19 pandemic are developed. This will be done using the linear regression tool. To begin with, the model is developed for the entire population of the Community of Madrid below.

$$MIncidTot1 \sim \beta_0 + \beta_1 * PC1 + \beta_2 * PC2 + \beta_3 * PC3 + \beta_4 * PC4 + \varepsilon$$

Equation 1 Analysis I Model 1

Coefficient	Estimate	p-value
(Intercept)	2.794e-16	1
$\widehat{\boldsymbol{\beta 1}}$	0.3632	**5.419e-09**
$\widehat{\boldsymbol{\beta 2}}$	0.3343	**8.331e-08**
$\widehat{\beta 3}$	-0.04587	0.4555
$\widehat{\beta 4}$	-0.04789	0.3891

Table 9 Characteristics Regression Analysis I Model 1

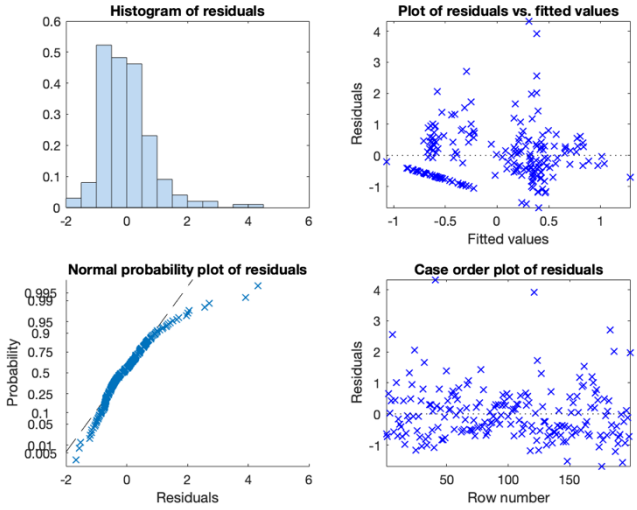

Figure 33 Residual Analysis Model 1 Analysis I

In this first analysis the clusters haven't been considered obtaining a model for the whole population of the Community of Madrid. For this case, $\widehat{\beta 1}$ and $\widehat{\beta 2}$ are presented as significant values obtaining a p-value inferior to 0.05. Both principal components have a positive value estimate meaning that their impact is directly proportional to the total incidence rate. In the next step, is going to be discussed a comparison between the whole model and the ones divided by clusters. For the cluster models, the equation is the same than the previously used Equation 1.

Coefficient	Estimate	p-value
(Intercept)	1.174	0.3030
$\widehat{\beta 1}$	0.2716	0.7301
$\widehat{\beta 2}$	-0.7625	0.4539
$\boldsymbol{\widehat{\beta 3}}$	0.4730	**0.1442**
$\widehat{\beta 4}$	0.1136	0.6659

Table 10 Characteristics Regression Analysis I Model 2 (Cluster 1)

Coefficient	Estimate	p-value
(Intercept)	1.281	0.008613
$\widehat{\beta 1}$	-0.2174	0.2226
$\widehat{\beta 2}$	-0.2129	0.4393
$\boldsymbol{\widehat{\beta 3}}$	0.3226	**0.06603**
$\widehat{\beta 4}$	-0.05365	0.2969

Table 11 Characteristics Regression Analysis I Model 3 (Cluster 2)

Coefficient	Estimate	p-value
(Intercept)	2.502	0.03780
$\widehat{\boldsymbol{\beta 1}}$	2.913	**0.03319**
$\widehat{\boldsymbol{\beta 2}}$	1.198	**0.002204**
$\widehat{\boldsymbol{\beta 3}}$	-0.4248	**0.002731**
$\widehat{\beta 4}$	0.5269	0.1848

Table 12 Characteristics Regression Analysis I Model 4 (Cluster 3)

Inside the analysis of each cluster, the behavior of the models changes radically, having more importance the factor $\widehat{\beta 3}$ as estimator rather than $\widehat{\beta 1}$ and $\widehat{\beta 2}$. In the cluster 1 and 2, $\widehat{\beta 3}$ is the only significant estimator if it is considered individually. The sign of the estimate is positive for both. On the other hand, cluster 3 has a as significant coefficients $\widehat{\beta 1}$, $\widehat{\beta 2}$ and $\widehat{\beta 3}$ with a positive sign estimate. In the next table the results are summarized, explaining the significant factor with the sign of the estimate.

	Significant factors	r^2
Total	Atmospheric Pollutants (+) and Particulate Matter (+)	0.261
Cluster 1	Aging of the population (+)	0.057
Cluster 2	Aging of the population (+)	0.363
Cluster 3	Atmospheric Pollutants (+), Particulate Matter (+) and Aging of the population (-)	0.227

Table 13 Significant factors Analysis I

The table highlights the difference in factors between the analysis of the entire community and those divided by cluster. While pollution is the most important at a global level, age has greater relevance when similar municipalities are used. This shows that, although the levels of contaminants in the air are very important and are closely related to respiratory problems, population trends due to age distribution strongly impact the expansion and development of the Covid-19 pandemic. Finally, it is determined that none of the models finds the economic factor as significant.

c. Analysis II – SPAIN

In this part the analysis will be performed from the characteristics and features variables from each Autonomous Community in Spain, shown in section a., generating new variables that can explain the existence of different clusters and its relationship with the pandemic.

As explained in section b., the first step is the creation of a PCA over different groups of variables to try to reduce the number of them. The variables used are 18 from 27 originally collected, removing 9. In the next table, the variables used are represented.

'Densidad'	'Menoresde16'	'De16a19aos'
'ymsaos'	'Viajerosnacional'	'Viajerosextranjero'
'De20a24aos'	'Muybueno'	'Bueno'
'Muymalo'	'Malo'	'GastoSanitariopercapita'
'Fallecidoscoronavirus'	'Datoscoronavirus'	'Coronavirus'
'Testcoronavirus'	'Pernoctacionesextranjero'	'Pernoctacionesnacional'

Table 14 Used variables in PCA Analysis II

As mentioned in the Analysis I, the results of PCA Analysis report can be interpreted in two different ways. One criterion is based on the explained variance (Figure 34), and the other one is based on the eigenvalues (Table 15). Both methods are represented on the following figure and table respectively.

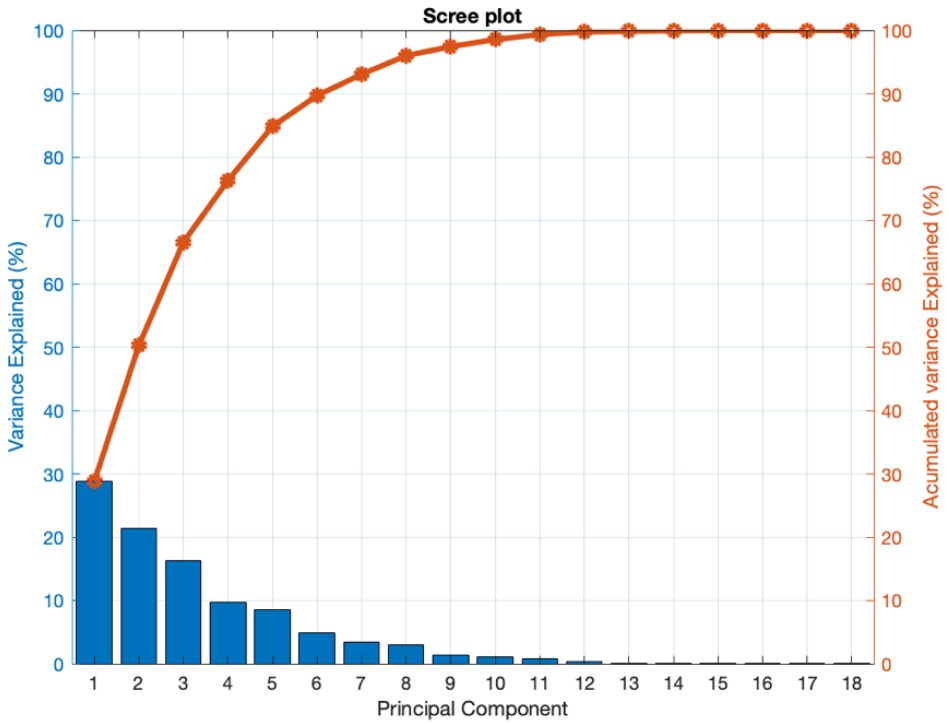

Figure 34 Scree plot PCA Analysis II

Principal Component	Eigenvalue
PC1	**5.1803**
PC2	**3.8608**
PC3	**2.9378**
PC4	**1.7535**
PC5	**1.5411**
PC6	0.8714
PC7	0.6056
PC8	0.5292
PC9	0.2580
PC10	0.2019
PC11	0.1414
PC12	0.0721
PC13	0.0235
PC14	0.0131
PC15	0.0082
PC16	0.0019

Table 15 Eigenvalues Analysis II

After executing the process, it is extracted the from Figure 34 that 5 components are necessary to explain more than 80% of the variance (84.85%).

Analyzing the eigenvalues in the Table 15, it can be concluded that only 5 principal components have a value superior to 1, which is the average variance of the system. These 5 components are representative ones.

Having both criterions the same result, 5 is the number of principal components desired for developing the analysis. This way, the new components are built using the varimax rotation getting the 5 variables with their scores from the original observations. In the next table, the characteristics of the variables are represented.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
'Densidad'	0.749	-0.057	0.044	0.392	-0.207
'Menoresde16'	**0.945**	-0.099	0.116	0.093	-0.198
'De16a19aos'	**0.948**	-0.072	0.085	0.074	-0.108
'De20a24aos'	**0.936**	-0.051	-0.174	0.062	-0.037
'ymsaos'	**-0.971**	0.132	0.122	0.082	-0.110
'Viajerosnacional'	-0.133	**0.986**	0.089	0.025	-0.004
'Viajerosextranjero'	-0.038	**0.863**	-0.036	-0.050	0.486
'Pernoctacionesnacional'	-0.128	**0.987**	0.075	0.024	0.008
'Pernoctacionesextranjero'	0.052	0.405	-0.340	0.011	**0.786**
'Muybueno'	0.090	0.015	-0.187	**-0.971**	0.089
'Bueno'	0.127	0.013	0.216	**0.944**	-0.092
'Malo'	-0.551	0.000	-0.059	0.287	-0.025
'Muymalo'	-0.078	-0.061	-0.172	-0.201	0.250
'GastoSanitariopercapita'	-0.139	0.047	0.024	-0.065	0.406
'Coronavirus'	0.181	0.010	0.524	0.185	0.242
'Testcoronavirus'	-0.023	0.083	**0.988**	0.093	-0.042
'Fallecidoscoronavirus'	0.015	-0.079	0.766	0.177	-0.432
'Datoscoronavirus'	-0.261	0.127	0.649	0.112	-0.350

Table 16 Loadings of principal components Varimax rotation Analysis II

The result from the varimax rotation, represented in the Table 16, is the creation of 5 new principal components that have a strong relationship with 11 original variables. The first component (PC1) represents principally age-related information, with the younger age segments having a positive impact, while those over 65 have a negative impact on the variable. The second one (PC2) gives information about the mobility and tourism, with a positive impact of this one. The third one (PC3), has relevant information about the google trends analytics, especially with coronavirus test related searches, and a positive impact of these. The fourth component (PC4), has information mainly about the health care public system opinion, having a positive impact around average answers (good, bad) and a negative with extremes (very good, very bad). The last component (PC5), it is linked to the length of stay for international tourists. With this information, the principal components have been named in relation with the variables that are more related to them in the following table:

Principal component (varimax rotation)	**Name**
PC1	Population's Youth
PC2	Mobility and tourism
PC3	Concern (Google trends)
PC4	Health care opinion
PC5	Length of stay (international)

Table 17 Principal component names and parameters Analysis II

As in the Analysis I, after obtaining the variables of study, the next steps aim to develop a clustering model based on the K-means method and validated by ANOVA. In the next figure, it is shown the clusters behavior according to the principal components.

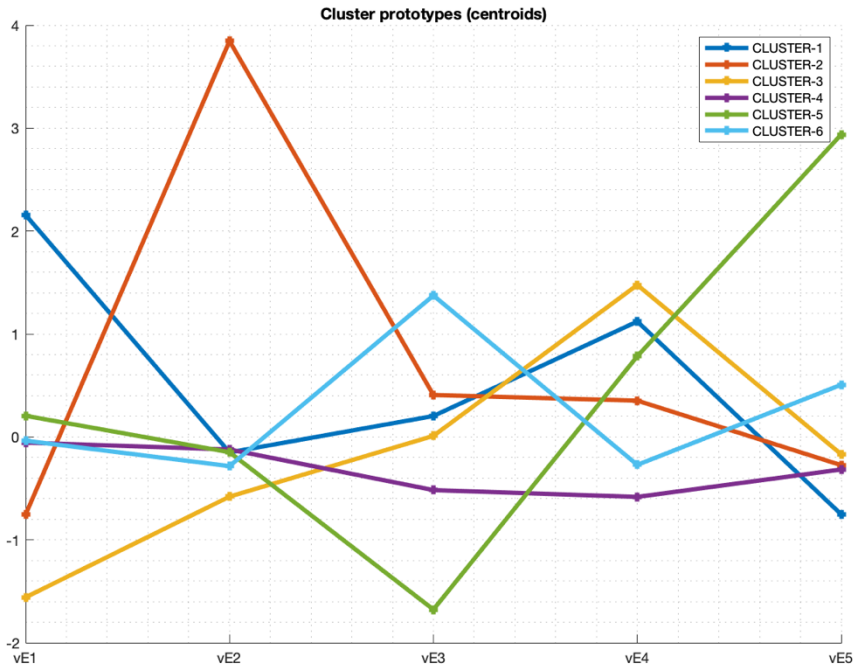

Figure 35 Clustering generated from the principal components in Analysis II

In principle, the clusters seem independent of each other and sufficiently different according to the Figure 35. However, several tests validated with ANOVA had to be processed before obtaining a satisfactory result. The dependent variable used to do this analysis is the total death rate of the COVID 19 at May 13th (MFallecidos2). The following figures represent the previously described analysis developed with different number of clusters.

Number of clusters (K)	Prob>F	MS Error
2	0.5391	1.035
3	0.6298	1.062
4	0.3706	0.980
5	0.2469	0.894
6	**0.0025**	**0.382**

Table 18 ANOVA in Analysis II

The number of clusters finally developed has been chosen according to the Table 18. Seeing how the p-value was reduced when the number of clusters was increased, the first scenario in which a result lower than 0.05 appeared was with 6 clusters.

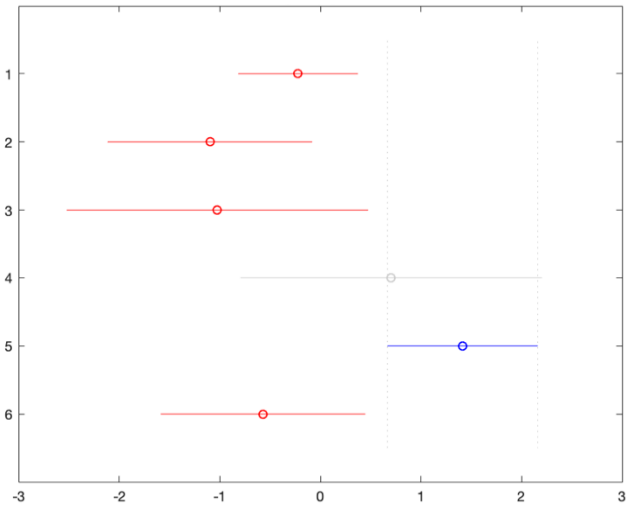

Figure 36 ANOVA means representation Analysis II

Cluster A	Cluster B	p-value
1	2	0.4883
1	3	0.8142
1	4	0.7149
1	5	0.0074
1	6	0.9756
2	3	1.0000
2	4	0.2308
2	5	0.0044
2	6	0.9506
3	4	0.4032
3	5	0.0345
3	6	0.9892
4	5	0.9006
4	6	0.5636
5	6	0.0251

Table 19 Mean contrast between the different clusters

After ANOVA analysis, the results of which are shown in the Figure 36 and Table 19, it can be concluded that all 3 clusters have means significant different from each other with good p-values in all the post-hoc mean contrast analysis. For a better understanding of how the clusters inside are, an analysis based on the characteristic variables is shown in the following table.

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
No. Communities	**9**	2	1	1	4	2
Population	**3.11 M**	**84.1 k**	2.21 M	2.35 M	2.56 M	1.84 M
Density	157	**7708**	297	**25**	234	**93**
Surface	24873	**11**	**7447**	**94224**	35064	20090
% <16 y.o.	0.161	**0.236**	0.142	0.131	0.160	0.123
% 16-19 y.o.	0.041	**0.056**	0.041	0.035	0.040	0.032
% 20-24 y.o.	0.050	**0.064**	0.053	0.043	0.049	0.039
% >65 y.o.	0.188	**0.113**	0.161	**0.243**	0.192	0.251
Tot. Travellers	0.755	0.041	1.145	**12.115**	0.858	0.195
Nac. Travellers	0.525	0.026	0.193	10.155	0.549	0.168
Intl. Travellers	0.230	0.015	**0.952**	1.960	0.309	0.028
Total stays	1.645	0.091	8.105	**20.847**	1.727	0.371
Nac. stays	0.983	0.066	0.721	17.473	0.983	0.311
Intl. stays	0.662	0.025	**7.384**	3.374	0.744	0.060
Very Good	**32**	15	24	21	26	9
Good	43	61	49	52	49	58
Regular	18	19	19	20	19	25
Bad	5	5	5	6	5	7
Very Bad	2	1	2	1	1	1
Per capita Health expenditure	1.447 k	**0.425 k**	1.358 k	1.506 k	1.379 k	**1.613 k**
Coronavirus	80	84	79	82	**86**	79
Test coronavirus	62	75	44	83	**86**	64
Death coronavirus	71	89	47	79	**92**	84
Data coronavirus	60	62	34	85	**84**	65

Table 20 Means of characteristics variables per cluster in Analysis I

Additionally, for identifying the geographical location of each cluster, a map of the Autonomous Communities has been represented in the following figure.

Figure 37 Autonomous Communities of Spain: names (top) and colored cluster areas (bottom)

From the previous Table 20, some conclusions can be extracted. The first cluster is the one that groups more Autonomous Communities (9), often on the coast, with a high population and a very good opinion about their health care system (32% of total). The second cluster includes the two autonomous cities of Ceuta and Melilla. Due to the special regime in these areas, the population and surface is much smaller than others, but their density is very high. Also, it is highlighted the youth of their population and low per capita expenditure on the health. The third cluster corresponds to the Canary Islands. It is important to highlight that this autonomous community has a very high level of visits of international visitors and overnight stays in relation to its population. In the fourth cluster it is found Castilla y León, the autonomous community with biggest surface and lowest density. The data shows that this community has older population than others and a very high rate of total visitors in relation to its population. The fifth cluster groups 4 autonomous communities in the interior of Spain as the Community of Madrid, Aragón, Castilla-La Mancha and La Rioja. It is highlighted the high rates of coronavirus related searches in Google Trends. The last cluster is formed by Galicia and Principado de Asturias, two communities that are next to each other in the north west of Spain. It has the second lowest population density and highest per capita expenditure. In the next table, the most important characteristics are highlighted.

Cluster no.	No. of Autonomous Communities	Characteristics
1	9	- Regions close to the coast - Good opinion about their healthcare
2	2	- Autonomous cities of Ceuta and Melilla - Low per capita expenditure on the health - Very youth population - Small surface with high density
3	1	- Canary Islands - Very high relative level of visits of international visitors and overnight stays
4	1	- Castilla y León - Biggest surface and lowest density - Very high rate of total visitors in relation to its population - Older population
5	4	- High rates of coronavirus related searches in Google - Communities in the interior of Spain
6	2	- Galicia and Principado de Asturias - Low population density - Highest per capita expenditure

Table 21 Clusters characteristics Analysis II

Once the principal components to be used and the clusters generated have been analyzed, the explanatory models of the COVID19 pandemic are developed. For doing this, the linear regression tool is going to be used. As the COVID information is more detailed at the country level, different phenomena are analyzed. In first place, the death rate is going to be studied.

$$MFallecidos2 \sim \beta_0 + \beta_1 * PC1 + \beta_2 * PC2 + \beta_3 * PC3 + \beta_4 * PC4 + \beta_5 * PC5 + \varepsilon$$

Equation 2 Analysis II Model 1

Coefficient	Estimate	p-value
(Intercept)	6.5739e-16	1
$\widehat{\beta 1}$	-0.25182	0.2048
$\widehat{\beta 2}$	0.085807	0.6570
$\boldsymbol{\widehat{\beta 3}}$	**0.62081**	**0.0058**
$\widehat{\beta 4}$	-0.2208	0.2627
$\widehat{\beta 5}$	0.16046	0.4017

Table 22 Characteristics Regression Analysis II Model 1

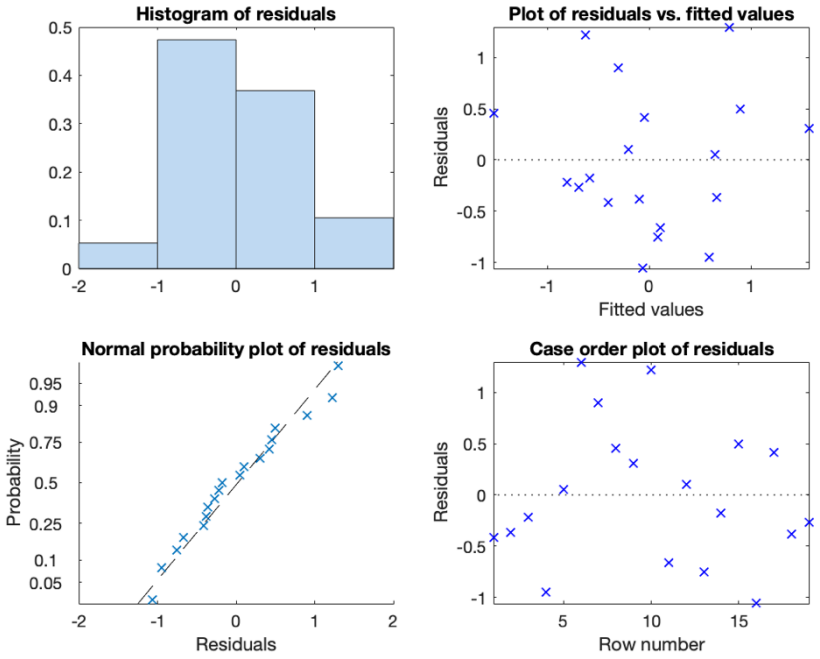

Figure 38 Residual Analysis Model 1 Analysis II

In this first analysis the clusters have not been considered, obtaining a model for the whole country. For this case, $\widehat{\beta 3}$ is presented as a significant estimator obtaining a p-value inferior to 0.05. The coefficient has a positive impact on the estimation of the death rate, meaning that the more value, the more death rate. In the next step, it is going to be discussed a comparison between the whole model and the ones divided by the clusters with more than the number of principal components, meaning that must have at least 5 Autonomous communities. For the cluster models, the equation is the same than the previously used Equation 2.

Coefficient	Estimate	p-value
(Intercept)	-0.51408	0.27255
$\widehat{\beta 1}$	-1.071	0.2103
$\widehat{\beta 2}$	-0.2048	0.77134
$\widehat{\beta 3}$	-0.82951	0.21762
$\widehat{\beta 4}$	0.4345	0.14613
$\widehat{\beta 5}$	-0.10377	0.84972

Table 23 Characteristics Regression Analysis II Model 2 (Cluster 1)

Having such a small sample, when performing a separate analysis of each cluster, the factors become less representative and it is not possible to draw clear conclusions from them considering all the principal components at the same time. For this reason, the equation must be changed deleting the 2 less representative principal components (PC2 and PC5). The final equation and its regression model are shown below.

$$MFallecidos2 \sim \beta_0 + \beta_1 * PC1 + \beta_3 * PC3 + \beta_4 * PC4 + \varepsilon$$

Equation 3 Analysis II Model 3

Coefficient	Estimate	p-value
(Intercept)	-0.46367	0.13124
$\widehat{\beta 1}$	**-1.2237**	**0.024006**
$\widehat{\beta 3}$	-0.83242	0.1028
$\widehat{\beta 4}$	0.43872	0.054002

Table 24 Characteristics Regression Analysis II Model 3 (Cluster 1)

From the previous Table 23, it can be concluded that analyzing the model per clusters, the first component, $\widehat{\beta 1}$, becomes more important being the only one with a p-value inferior to 0.05. The coefficient has a negative impact on the estimation of the death rate, meaning that the less value, the more death rate. In the following table, the results of the regressions related with mortality are shown.

	Significant factors
Total	Concern (Google trends) (+)
Per cluster	Population's Youth (-)

Table 25 Significant factors Analysis II Model 1 and 3

With this, we conclude from Table 25 that the only significant factor in the mortality rate at the national level is the concern in the population measured via Google Trends, having higher mortality rate the areas in which there is less concern (as expected, negative impact). When analyzed per cluster, using the cluster 1 as reference, aging of the population becomes more important, having higher mortality rate on the areas where the population is less youth.

Having completed the analysis of the first characteristic of coronavirus (death rate), we proceed to analyze other characteristics only at the national level as PCR positive incidence, hospitalized and admitted to ICU. Firstly, the PCR data is analyzed below.

$$MTotalPCR2 \sim \beta_0 + \beta_1 * PC1 + \beta_2 * PC2 + \beta_3 * PC3 + \beta_4 * PC4 + \beta_5 * PC5 + \varepsilon$$

Equation 4 Analysis II Model 4

Coefficient	**Estimate**	**p-value**
(Intercept)	2.6904e-16	1
$\widehat{\beta 1}$	-0.25882	0.2285
$\widehat{\beta 2}$	0.1121	0.59373
$\boldsymbol{\widehat{\beta 3}}$	**0.58526**	**0.013421**
$\widehat{\beta 4}$	-0.12227	0.5606
$\widehat{\beta 5}$	0.10949	0.59509

Table 26 Characteristics Regression Analysis II Model 4

When studied the table 24 and, as in the mortality rate, the most and only significant factor (p-value<0.05) is the Google Trends principal component parameter with a positive relation on the regression between its value and the

total incidence on positive PCR results, meaning that the more positive cases detected by PCR. In the same way that for the death rate, for the clusters analysis a reduced regression model is used, removing the 2 less representative factors (2 and 5) in the next analysis.

$$MTotalPCR2 \sim \beta_0 + \beta_1 * PC1 + \beta_3 * PC3 + \beta_4 * PC4 + \varepsilon$$

Equation 5 Analysis II Model 5

Coefficient	Estimate	p-value
(Intercept)	-0.4732	0.27039
$\widehat{\boldsymbol{\beta 1}}$	**-1.3512**	**0.063292**
$\widehat{\beta 3}$	-0.91756	0.19913
$\widehat{\beta 4}$	0.49602	0.11461

Table 27 Characteristics Regression Analysis II Model 5

From the previous Table 27, as in the death rate analysis, it can be concluded that studying the model per clusters, the first component, $\widehat{\beta 1}$, becomes the most relevant but not getting a p-value inferior to 0.05. The coefficient has a negative impact on the estimation of the death rate, meaning that the less value, the more death rate. In the following table, the results of the regressions related with mortality are shown.

	Significant factors
Total	Concern (Google trends) (+)
Per cluster	≈ Population's Youth (-)

Table 28 Significant factors Analysis II Model 4 and 5

With this, we can conclude from Table 28 that the only significant factor in the mortality rate at the national level is the concern in the population measured via Google Trends, having higher mortality rate the areas in which there is less concern (as expected, negative impact). When analyzed per cluster, using the cluster 1 as reference, even not reaching p-value < 0.05, aging of the

population becomes more important, having higher mortality rate on the areas where the population is less youth. Following the same structure, hospitalized models are analyzed below.

$$MHospitalizados2 \sim \beta_0 + \beta_1 * PC1 + \beta_2 * PC2 + \beta_3 * PC3 + \beta_4 * PC4 + \beta_5 * PC5 + \varepsilon$$

Equation 6 Analysis II Model 6

Coefficient	Estimate	p-value
(Intercept)	7.0938e-16	1
$\widehat{\beta 1}$	-0.29316	0.13497
$\widehat{\beta 2}$	0.11446	0.54484
$\widehat{\beta 3}$	**0.60268**	**0.0059928**
$\widehat{\beta 4}$	-0.17186	0.36694
$\widehat{\beta 5}$	0.24536	0.1971

Table 29 Characteristics Regression Analysis II Model 6

$$MHospitalizados2 \sim \beta_0 + \beta_1 * PC1 + \beta_3 * PC3 + \beta_4 * PC4 + \varepsilon$$

Equation 7 Analysis II Model 7

Coefficient	Estimate	p-value
(Intercept)	-0.4908	0.19215
$\widehat{\beta 1}$	**-1.3778**	**0.036055**
$\widehat{\beta 3}$	-0.9365	0.13679
$\widehat{\beta 4}$	0.46597	0.089426

Table 30 Characteristics Regression Analysis II Model 7

From the previous Table 29, it can be concluded that studying the model per clusters, the third component, $\widehat{\beta 3}$, becomes the only valid according to the p-values shown. The coefficient has a positive impact on the estimation of the hospitalizations, meaning that the more value, the more hospitalizations. In the following table, the results of the regressions related with mortality are shown. From the Table 30, the model shows that, when analyzed per clusters, the first component, $\widehat{\beta 1}$, becomes the most relevant getting a p-value inferior to 0.05 and having a negative impact in the hospitalization's regression. In the next table the results are summarized.

	Significant factors
Total	Concern (Google trends) (+)
Per cluster	Population's Youth (-)

Table 31 Significant factors Analysis II Model 6 and 7

In the table 31, a similar trend like in the death rate and positive PCR is seen. Now, following the same structure, admitted to ICU models are analyzed below.

$$MUCI2 \sim \beta_0 + \beta_1 * PC1 + \beta_2 * PC2 + \beta_3 * PC3 + \beta_4 * PC4 + \beta_5 * PC5 + \varepsilon$$

Equation 8 Analysis II Model 8

Coefficient	Estimate	p-value
(Intercept)	2.353e-16	1
$\widehat{\beta 1}$	-0.24543	0.22336
$\widehat{\beta 2}$	0.057364	0.76997
$\boldsymbol{\widehat{\beta 3}}$	**0.4942**	**0.02303**
$\widehat{\beta 4}$	-0.23809	0.23655
$\widehat{\beta 5}$	0.38071	0.064319

Table 32 Characteristics Regression Analysis II Model 8

$$MUCI2 \sim \beta_0 + \beta_1 * PC1 + \beta_3 * PC3 + \beta_4 * PC4 + \varepsilon$$

Equation 9 Analysis II Model 9

Coefficient	Estimate	p-value
(Intercept)	-0.42459	0.42837
$\boldsymbol{\widehat{\beta 1}}$	**-1.1791**	**0.16866**
$\widehat{\beta 3}$	-0.94668	0.29007
$\widehat{\beta 4}$	0.42179	0.26406

Table 33 Characteristics Regression Analysis II Model 9

These 2 additional analyses of the models 8 and 9 provide results that follow the same line that the previous ones, presenting the principal component 3 as the most significant factor at the national level and the principal component 1 as the most representative when analyzed per clusters. In the ICU

national level regression (Model 8) the p-value for principal component 5 is close to the acceptance limit.

	Significant factors
Total	Concern (Google trends) (+)
Per cluster	≈ Population's Youth (-)

Table 34 Significant factors Analysis II Model 8 and 9

On Table 34, it is represented in the analysis per clusters the first component ($\widehat{\beta 1}$) which, even did not reach a p-value smaller than 0.05, is still the variable that better performs according to the previous results (Table 33).

For summarizing the results from the Analysis II, the following table has been developed.

		Significant factors	**r^2**
Deaths	**Total**	Concern (Google trends) (+)	0.535
	Per cluster	Population's Youth (-)	0.728
PCR positive	**Total**	Concern (Google trends) (+)	0.452
	Per cluster	≈ Population's Youth (-)	0.60
Hospitalized	**Total**	Concern (Google trends) (+)	0.558
	Per cluster	Population's Youth (-)	0.67
Admitted to ICU	**Total**	Concern (Google trends) (+)	0.519
	Per cluster	≈ Population's Youth (-)	0.40

Table 35 Results Analysis II

5. CONCLUSIONS

In the Community of Madrid, the analysis began with 18 descriptive variables extracted from public databases that became 4 principal components after the varimax rotation: Atmospheric Pollutants, Atmospheric Particulate Matter, Aging of the population and Economy. With these components, 3 clusters were developed with a k-means model and validated with an ANOVA. Once the variables and the clusters were prepared, the regression analysis were run. The analysis for the Community of Madrid provided as relevant general factors for all the municipalities, the 2 principal components related to the pollution with a direct proportion. However, when looking into clusters with similar pollution data, aging of the population transforms also into a powerful estimator for measuring the impact of COVID-19 into the population. The results also showed that hypotheses based on economic factors divisions have been rejected and they are not valid.

For Spain, 27 variables were initially collected, developing 5 principal components with the varimax rotation: Population's Youth, Mobility and tourism, Concern (Google trends), Health care opinion and Length of stay (international). With these components, 6 clusters were designed in a similar way than in the Community of Madrid. From the regressions, a general statement can be extracted. In all the variables analyzed and models studied, the Google Trends factor has been demonstrated to be the most accurate and better fitted to estimate the development of the pandemic over the whole country, increasing the COVID variables proportionally to the value of the number of searches related with the pandemic. This fact provides a proof of how important have become the technologies and analytics to track changes in real time on our

daily life. When the analysis is generated into more similar areas/clusters the Population's Youth becomes a much more important factor in the estimation of all the pandemic characteristics, being indirectly proportional to the COVID variables. The rest of the principal components linked with Mobility and tourism or Health System Opinion has been rejected as valid estimators for predicting the expansion in COVID-19 at the national level.

With all this and analyzing other researches that have been carried out previously, some risk factors and potential variables that could influence the pandemic have been verified. On the one hand, the presence of super spreaders explained in the study of Wong G. *et al* [42]. In this research it is explained that the environment and behavior of individuals can change radically how the virus infects others [42]. Included in this behavior are the Super-Spreading events [42] as hospitals, enclosed housing complexes, and mass transportation. Already in the exploratory analysis it was possible to obtain in the COVID variables of the Community of Madrid indications of super spreaders and outbreaks due to the existence of outliers in some populations that were outside the general trend. On the other hand, the influence of meteorological factors in relation to the coronavirus was pointed out by the research of Jia J. *et al*. [45] and Sajadi M. M. *et al*. [46]. In these researches, it is highlighted the similarities [45] between COVID-19 and SARS, and, therefore, links the changes in the weather as a fundamental factor in the control of the pandemic, as happened in the past with SARS. Also, it is stated that historically climatic characteristics such as humidity or temperature [46] have influenced other human coronaviruses and pandemics such as MERS, SARS or seasonal influenza. Even there has not been a variable fully focused in meteorological factors, if the Table 3 is analysed,

those were included mostly in the components related to the pollution. Also, it has to be considered that the pollution levels change drastically with some factors as rainfalls. The pollution components have been the most relevant parameters in the models for the entire Autonomous Community proving that the hypothesis is valid in this study. Finally, age and demographics factors were pointed out as relevant in the study of Esteve A., Permanyer I. and Boertien D. [48]. The youth of the population has been placed as an important factor when the pollution factors are more similar in a specific area, so that this hypothesis has also been validated with the models carried out in this work.

6. LIMITATIONS AND FUTURE LINES OF RESEARCH

Once the project has been developed, additional lines of research have been discovered with which to improve and extend the depth of the problem studied.

In the first place, it would be interesting to carry out models related to deaths, hospitalisations and ICU admissions at a regional level, in the community of Madrid. Despite the existence of a total count of all these variables for the entire Autonomous Community, there are no public data at a municipal level, unlike those positive by PCR already used. This study would help to see how these variables behave in smaller sets and whether they are related or follow the same trend.

Secondly, an ambitious national municipal model could have been carried out in which the data produced from all of the Autonomous Communities were united in such a way that they managed to become much more diverse clusters based on more variables. In relation to the clusters in Spain, due to the smaller sample obtained, this one did not provide sufficiently precise information when developing the models for each one. However, if instead of using the autonomic division, municipalities were used, many patterns would be identified with greater precision.

Finally, as a limitation, it has been seen that public databases, such as pollution data, are based on the number of stations that exist in a given area. If tools with more detailed data could be accessed, some of the results obtained could be improved, again with greater precision.

Therefore, all the indications and recommendations extracted for future investigations follow the line of achieving a greater base of information to carry out the analyses to obtain a greater level of detail and a greater knowledge of the population and environment to be studied.

SUSTAINALBLE DEVELOPMENT GOALS (SDGs)

After executing a research and doing a deep reflection about how the SDGs could have an impact in the final thesis, it has been determined that the first and most important related goal is the SDG 3: Ensure healthy lives and promote well-being for all at all ages. During the last decades, Infectious Diseases as the Tuberculosis or HIV have continued spreading over the world. Different protection mechanisms like universal or affordable healthcare are not present in many parts of the world, and hospitals net or resources are often not good enough in developing countries, not being prepared for fighting against some outbreaks like the actual situation with the COVD-19. In the current pandemic, these infrastructures have demonstrated their importance in protecting the safety and health of citizens. This objective would be related to the society dimension and could be quantified measuring the mortality rate attributed to infectious diseases and air pollution. In this paper, those variables have been studied proving that their value is directly proportional to the mortality rate of the COVID-19. Secondarily, in relation to biosphere, it is highlighted the SDG 13: Take urgent action to combat climate change and its impacts. This goal looks to strengthen resilience and adaptive capacity to climate-related hazards and natural disasters in all countries and can be quantified measuring the number of deaths and persons affected by disaster per 100,000. In my thesis, the disaster is considered the pandemic, measuring infection, hospitalization and death rate in different regions and municipalities. Finally, also secondarily and in relation to the economy, it is considered the SDG 9: Develop resilient infrastructure, promote inclusive and sustainable industrialisation, and encourage innovation. This goal looks to develop infrastructure to support economic development and human well-being with equitable access for all and can be measured with the development expenditure as

a proportion of GDP. In the thesis, the variable used for measuring this effect has been the healthcare expenditure per capita.

The following table summarizes the conclusions obtained from this section:

SDG dimension	SDG identified	Role	Goal	Quantification
Biosphere	SDG13: Take urgent action to combat climate change and its impacts	Secondary	Strengthen resilience and adaptive capacity to climate-related hazards and natural disasters in all countries	Number of deaths and persons affected by disaster (COVID) per 100,000 people
Society	SDG3: Ensure healthy lives and promote well-being for all at all ages	Primary	By 2030, end the epidemics of AIDS, tuberculosis, malaria and neglected tropical diseases and combat hepatitis, water-borne diseases and other communicable diseases	Mortality rate attributed to household and ambient air pollution and infectious diseases (COVID)
Economy	SDG9: Develop resilient infrastructure, promote inclusive and sustainable industrialisation, and encourage innovation	Secondary	Develop quality, reliable, sustainable and resilient infrastructure, including regional and transborder infrastructure, to support economic development and human well-being, with a focus on affordable and equitable access for all	Research and health expenditure per capita or as a portion of GDP

Table 36 SDGs conclusions

APPENDIX

Variable name	Meaning
'AIncid14D'	Cumulative Incidence 14 days previous to April 8th
'AIncid14D1'	Cumulative Incidence 14 days previous to April 15th
'AIncid14D2'	Cumulative Incidence 14 days previous to April 22nd
'AIncid14D3'	Cumulative Incidence 14 days previous to April 29th
'MIncid14D'	Cumulative Incidence 14 days previous to May 6th
'MIncid14D1'	Cumulative Incidence 14 days previous to May 13th
'AIncidTot'	Total Incidence at April 8th
'AIncidTot1'	Total Incidence at April 15th
'AIncidTot2'	Total Incidence at April 22nd
'AIncidTot3'	Total Incidence at April 29th
'MIncidTot'	Total Incidence at May 6th
'MIncidTot1'	Total Incidence at May 13th
'Pob16aos'	Percentage of the population under 16 years old
'Pob1664'	Percentage of the population between 16 and 64 years old
'Pob80'	Percentage of the population over 80 years old
'Pob6580'	Percentage of the population between 65 and 80 years old
'Pob65'	Percentage of the population over 65 years old
'DecEI'	Incidence of deaths due to infectious diseases
'MSistResp'	Incidence of deaths due to respiratory system problems
'Densidad'	Density of the municipality / district
'PM25'	Particulate matter of less than 2.5 micrometres
'NO2_Nivel_Medio' / 'NO2'	NO_2 April average level
'SO2_Nivel_Medio_Abril' / 'SO2'	SO_2 April average level
'CO_Nivel_Medio_Abril' / 'CO'	CO April average level
'NO_Nivel_Medio' / 'NO'	NO April average level
'Ozono_Nivel_Medio' / 'Ozono'	Ozone April average level
'TempMedia'	Temperature April average level
'HumRel'	Humidity April average level
'Rentapercapita'	Per Capita Income
'Totalafiliados'	Percentage of population with social security affiliations

Table 37 Variables names and meaning Community of Madrid

Variable name	Meaning
'MPCRAcum14D'	Cumulative Incidence 14 days previous to March 4th
'MPCRAcum14D1'	Cumulative Incidence 14 days previous to March 18th
'APCRAcum14D'	Cumulative Incidence 14 days previous to April 1st
'APCRAcum14D1'	Cumulative Incidence 14 days previous to April 15th
'APCRAcum14D2'	Cumulative Incidence 14 days previous to April 29th
'MPCRAcum14D2'	Cumulative Incidence 14 days previous to May 13th
'MHospitalizados'	Total population Hospitalized at March 4th
'MHospitalizados1'	Total population Hospitalized at March 18th
'AHospitalizados'	Total population Hospitalized at April 1st
'AHospitalizados1'	Total population Hospitalized at April 15th
'AHospitalizados2'	Total population Hospitalized at April 29th
'MHospitalizados2'	Total population Hospitalized at May 13th
'MUCI'	Total population in ICU at March 4th
'MUCI1'	Total population in ICU at March 18th
'AUCI'	Total population in ICU at April 1st
'AUCI1'	Total population in ICU at April 15th
'AUCI2'	Total population in ICU at April 29th
'MUCI2'	Total population in ICU at May 13th
'MFallecidos'	Total deaths at March 4th
'MFallecidos1'	Total deaths at March 18th
'AFallecidos'	Total deaths at April 1st
'AFallecidos1'	Total deaths at April 15th
'AFallecidos2'	Total deaths at April 29th
'MFallecidos2'	Total deaths at May 13th
'Menoresde16'	Percentage of the population under 16 years old
'De16a19aos'	Percentage of the population between 16 and 19 years old

'De20a24aos'	Percentage of the population between 20 and 24 years old
'De25a34aos'	Percentage of the population between 25 and 34 years old
'De35a44aos'	Percentage of the population between 35 and 44 years old
'De45a54aos'	Percentage of the population between 45 and 54 years old
'De55a64aos'	Percentage of the population between 55 and 64 years old
'ymsaos'	Percentage of the population over 65 years old
'Viajerostotal'	Total travellers
'Viajerosnacional'	Total national travellers
'Viajerosextranjero'	Total foreign travellers
'Pernoctacionestotal'	Total overnight stays
'Pernoctacionesnacional'	Total national overnight stays
'Pernoctacionesextranjero'	Total foreign overnight stays
'Muybueno'	Opinion health care system: very good
'Bueno'	Opinion health care system: good
'Regular'	Opinion health care system: regular
'Malo'	Opinion health care system: bad
'Muymalo'	Opinion health care system: Very bad
'MaloMuymalo'	Opinion health care system: Bad + VeryBad
'GastoSanitariopercapita'	Per capita health expenditure
'Coronavirus'	Google trends: Coronavirus related search
'Testcoronavirus'	Google trends: Coronavirus test related search
'Fallecidoscoronavirus'	Google trends: Coronavirus deaths related search
'Datoscoronavirus'	Google trends: Coronavirus data related search
'Mediadepreocupacin'	Google trends: Average of all indicators

Table 38 Variables names and meaning Spain

REFERENCES

[1*] Centro de Coordinación y de Emergencias Sanitarias (Spanish Ministry of Health), Manejo en urgencias del COVID-19. Madrid, Spain, 2020.

[2*] Centro de Coordinación y de Emergencias Sanitarias (Spanish Ministry of Health), Enfermedad por coronavirus, COVID-19. Madrid, Spain, 2020.

[1] Altman L.K., Is this a pandemic define 'pandemic', *New York Times*, June 8th, 2009. Available at: http://www.nytimes.com/2009/06/09/health/09docs.html [Accessed March 31, 2020]

[2] Doshi P., The elusive definition of pandemic influenza, *Bulletin of the World Health Organization*, Volume 89, Number 7, July 2011, Pages 469-544. Available at: https://www.who.int/bulletin/volumes/89/7/11-086173-ab/es/ [Accessed March 31, 2020]

[3] WHO, Emergencies preparedness, response, *Pandemic (H1N1) 2009: Frequently Asked Questions*, 2009. Available at: https://www.who.int/csr/disease/swineflu/frequently_asked_questions/pandemic/en/ [Accessed March 31, 2020]

[4] Morens D., Folkers G.K., Fauci A.S., What Is a Pandemic?, *The Journal of Infectious Diseases*, Volume 200, Issue 7, October 1, 2009, Pages 1018–1021. Available at: https://doi.org/10.1086/644537 [Accessed March 31, 2020]

[5] Balzer, D., Pandemic vs. Endemic vs. Outbreak: Terms to Know, Mayo Clinic, March 10, 2016. Available at: https://newsnetwork.mayoclinic.org/discussion/pandemic-versus-endemic-versus-outbreak-terms-to-know/ [Accessed March 31, 2020]

[6] Idrovo A.J., Epidemics, endemics and conglomerates: basic concepts, Journal of the Faculty of Medicine of the University of Colombia, Volume 48, Number 3, 2020, Pages 175 – 180. Available at: https://revistas.unal.edu.co/index.php/revfacmed/article/view/19623/20690 [Accessed April 1, 2020]

[7] Torrado E. *et al*, Paracoccidioidomicocic: definition of the endemic areas of Colombia, Biomédica, Volume 20, Number 4, December, 2000, Pages 327 – 334. Available at: https://www.redalyc.org/pdf/843/84320408.pdf [Accessed April 1, 2020]

[8] Horcajada P., Padilla B., Endemia y epidemia. Investigación de un brote epidémico nosocomial, *Enfermedades infecciosas y microbiología clínica*, October, 2012, Pages 181 – 186. Available at: https://seimc.org/contenidos/documentoscientificos/eimc/seimc_eimc_v31n03p181a186.pdf [Accessed April 1, 2020]

[9] Endemic Diseases. Available at: https://byjus.com/biology/endemic-diseases/ [Accessed April 1, 2020]

[10] WHO, Risk Factors. Available at: https://www.who.int/topics/risk_factors/es/ [Accessed April 1, 2020]

[11] EUPATI, Risk factors in health and disease. Available at: https://www.eupati.eu/pharmacoepidemiology/risk-factors-health-disease/ [Accessed April 1, 2020]

[12] Moghadami M., A Narrative Review of Influenza: A Seasonal and Pandemic Disease, *Iranian Journal of Medial Sciences*, Volume 42(1), January, 2017, Pages 2 – 3. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5337761/ [Accessed 1 April 2020]

[13] Centers for Disease Control and Prevention, History of 1918 Flu Pandemic. Available at: https://www.cdc.gov/flu/pandemic-resources/1918-commemoration/1918-pandemic-history.htm [Accessed April 1, 2020]

[14] Andrews E., Why Was It Called the 'Spanish Flu?, *History Magazine*, January 12, 2016. Available at: https://www.history.com/news/why-was-it-called-the-spanish-flu [Accessed April 1, 2020]

[15] Billings M., The Influenza Pandemic of 1918, *Standford University*, June 1997. Available at: https://virus.stanford.edu/uda/ [Accessed: April 9, 2020]

[16] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), History of 1918 Flu Pandemic, March 21, 2018. Available at: https://www.cdc.gov/flu/pandemic-resources/1918-commemoration/1918-pandemic-history.htm [Accessed April 9, 2020]

[17] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), Influenza Historic Timeline, January 30, 2019. Available at: https://www.cdc.gov/flu/pandemic-resources/pandemic-timeline-1930-and-beyond.htm [Accessed April 9, 2020]

[18] Cervera C., La olvidada Gripe Asiática de 1957 que puso a prueba el sistema sanitario mundial con un millón de muertos, *ABC*, March 5, 2020. Available at: https://www.abc.es/historia/abci-olvidada-gripe-asiatica-1957-puso-prueba-sistema-sanitario-mundial-millon-muertos-202003050111_noticia.html [Accessed April 9, 2020]

[19] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), 1957-1958 Pandemic (H2N2 virus), January 2, 2020. Available at: https://www.cdc.gov/flu/pandemic-resources/1957-1958-pandemic.html [Accessed April 9, 2020]

[20] Rogers K, 1957 flu pandemic, *ENCYCLOPÆDIA BRITANNICA*, March 26, 2020. Available at: https://www.britannica.com/event/Asian-flu-of-1957 [Accessed April 9, 2020]

[21] Rogers K, 1968 flu pandemic, *ENCYCLOPÆDIA BRITANNICA*, March 25, 2020. Available at: https://www.britannica.com/event/Hong-Kong-flu-of-1968 [Accessed April 9, 2020]

[22] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), 1968 Pandemic (H3N2 virus), January 2, 2020. Available at: https://www.cdc.gov/flu/pandemic-resources/1968-pandemic.html [Accessed April 9, 2020]

[23] Instituto de Salud Carlos III, Informe semanal del Sistema de Vigilancia de la Gripe en España (SVGE), January 17, 2019. Available at: http://vgripe.isciii.es/PresentarHomeBoletin.do;jsessionid=2BF94ECCA6D8CA8620CB216DEB4EC2E5?boletin=1&bol=60637 [Accessed April 9, 2020]

[24] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), 2009 H1N1 Pandemic (H1N1pdm09 virus), June 11, 2019. Available at: https://www.cdc.gov/flu/pandemic-resources/2009-h1n1-pandemic.html [Accessed April 9, 2020]

[25] Vaqué J., Epidemiología de la gripe A (H1N1) en el mundo y en España, Archivos de Bronconeumología, *Elsevier Doyma*, 2010, number 46(Supl2), pp. 3-12. Available at: https://www.archbronconeumol.org/es-pdf-S0300289610700144 [Accessed April 9, 2020]

[26] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD), Ten Years of Gains: A Look Back at Progress Since the 2009 H1N1 Pandemic, July 31, 2019. Available at: https://www.cdc.gov/flu/spotlights/2018-2019/decade-since-h1n1-pandemic.html [Accessed April 9, 2020]

[27] Instituto de Salud Carlos III, Informe semanal del Sistema de Vigilancia de la Gripe en España (SVGE), January 17, 2019. Available at: http://vgripe.isciii.es/PresentarHomeBoletin.do;jsessionid=2BF94ECCA6D8CA8620CB216DEB4EC2E5?boletin=1&bol=60637 [Accessed April 9, 2020]

[28] ENFERMEDAD POR NUEVO CORONAVIRUS, COVID-19, Centro de Coordinación de Alertas y Emergencias Sanitarias, Dirección General de salud pública, calidad e innovación, 31 January 2020. Available at: https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/documentos/Informacion_inicial_alerta.pdf [Accessed April 16, 2020]

[29] Timeline: How the new coronavirus spread, *AL JAZEERA AND NEWS AGENCIES*, 15 April 2020. Available at: https://www.aljazeera.com/news/2020/01/timeline-china-coronavirus-spread-200126061554884.html [Accessed April 16, 2020]

[30] Informe sobre la situación de COVID-19 en España, Informe no 22. Situación de COVID-19 en España a 13 de abril de 2020. Equipo COVID-19. RENAVE. CNE. CNM (ISCIII), 13 April 2020. Available at: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/Informes%20COVID-19/Informe%20n°%2022.%20Situación%20de%20COVID-19%20en%20España%20a%2013%20de%20abril%20de%202020.pdf [Accessed April 16, 2020]

[31] WHO Timeline - COVID-19, WHO, 8 April 2020. Available at: https://www.who.int/news-room/detail/08-04-2020-who-timeline---covid-19 [Accessed April 16, 2020]

[32] Coronavirus COVID-19 outbreak in the EU, University Institute of Migration Studies, 24 March 2020. Available at: https://fra.europa.eu/sites/default/files/fra_uploads/spain-report-covid-19-april-2020_en.pdf [Accessed April 16, 2020]

[33] Coronavirus: qué significa que la OMS haya clasificado al covid-19 como pandemia, *BBC News Mundo*, 11 March 2020. Available at: https://www.bbc.com/mundo/noticias-internacional-51842708 [Accessed April 16, 2020]

[34] Johns Hopkins University, COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), Coronavirus Resource Center of Johns Hopkins University. Available at: https://coronavirus.jhu.edu/map.html [Accessed April 20, 2020]

[35] Andersen K. G. *et al*., The proximal origin of SARS-CoV-2, *Nature Medicine*, num. 26, pp. 450-452, 2020. Available at: https://doi.org/10.1038/s41591-020-0820-9 [Accessed April 18, 2020]

[36] COVID-19, MERS & SARS, National Institute of Allergy and Infectious Diseases, 6 April 2020. Available at: https://www.niaid.nih.gov/diseases-conditions/covid-19 [Accessed April 16, 2020]

[37] WHO, Coronavirus. Available at: https://www.who.int/health-topics/coronavirus#tab=tab_3 [Accessed April 20, 2020]

[38] BBC, Coronavirus: 6 medidas extremas adoptadas por las autoridades en la lucha contra el covid-19, *BBC News Mundo*, 11 March 2020. Available at: https://www.bbc.com/mundo/noticias-51832806 [Accessed April 20, 2020]

[39] Coronavirus COVID-19 outbreak in the EU, University Institute of Migration Studies, 24 March 2020. Available at: https://fra.europa.eu/sites/default/files/fra_uploads/spain-report-covid-19-april-2020_en.pdf [Accessed April 20, 2020]

[40] Guimón P., Temor en la comunidad asiática en Estados Unidos ante los ataques racistas por el coronavirus, *El País*, 24 March 2020. Available at: https://elpais.com/internacional/2020-03-23/temor-en-la-comunidad-asiatica-en-estados-unidos-ante-los-ataques-racistas-por-el-coronavirus.html [Accessed April 20, 2020]

[41] Chowell, G., Abdirizak, F., Lee, S. et al. Transmission characteristics of MERS and SARS in the healthcare setting: a comparative study. BMC Med 13, 210 (2015). Available at: https://doi.org/10.1186/s12916-015-0450-0 [Accessed April 22, 2020]

[42] Wong G., *et al*., MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease, Cell Host & Microbe, Volume 18, Issue 4, Pages 398-401, 14 October 2015. Available at: https://doi.org/10.1016/j.chom.2015.09.013 [Accessed April 22, 2020]

[43] Lee K., Jung K., Factors Influencing the Response to Infectious Diseases: Focusing on the Case of SARS and MERS in South Korea, *Int. J. Environ. Res. Public Health* 2019, 16(8), 1432. Available at: https://doi.org/10.3390/ijerph16081432 [Accessed April 22, 2020]

[44] Wu X., Nethery R., Sabath M. B., Braun D., and Dominici F., Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study, 2020, *Harvard University*. Available at: https://projects.iq.harvard.edu/files/covid-pm/files/pm_and_covid_mortality_med.pdf [Accessed May 2, 2020]

[45] Jia J., *et al*., Modeling the Control of COVID-19: Impact of Policy Interventions and Meteorological Factors, *Cornell University*, 6 March 2020. Available at: https://arxiv.org/abs/2003.02985 [Accessed April 22, 2020]

[46] Sajadi M. M., Habibzadeh P., Vintzileos A., Shokouhi S., Miralles-Wilhelm F., and Amoroso A., Temperature, humidity, and latitude analysis to predict potential spread and seasonality for COVID-19, *SSRN Electronic Journal*, January 2020. Available at: 10.2139/ssrn.3550308

[47] Instituto de Salud Carlos III, Informes COVID-19, Red Nacional de Investigación Epidemiológica, 16 Abril 2020. Available at: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Paginas/InformesCOVID-19.aspx [Accessed April 20, 2020]

[48] Esteve A., Permanyer I., Boertien D., La vulnerabilidad de las provincias españolas a la covid-19 según su estructura por edad y de co-residencia: implicaciones para el (des)confinamiento, *Centre d´Estudis Demogràfics*, Abril 2020. Available at: https://ddd.uab.cat/pub/worpap/2020/221367/perdem_a2020n_019_v2_iSPA.pdf [Accessed April 22, 2020]

[49] Covid 19 -TIA por Municipios y Distritos de Madrid, Portal de Datos Abiertos de la Comunidad de Madrid. Available at: https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos [Accessed May 6, 2020]

[50] Municipios de la Comunidad de Madrid, Portal de Datos Abiertos de la Comunidad de Madrid. Available at: http://datos.comunidad.madrid/catalogo/dataset/municipio_comunidad_madrid [Accessed April 22, 2020]

[51] Residentes en la Comunidad de Madrid por Rango de Edad, Portal de Datos Abiertos de la Comunidad de Madrid. Available at: https://datos.comunidad.madrid/catalogo/dataset/residentes_por_rango_edad [Accessed April 22, 2020]

[52] Red de Calidad del Aire. Datos del mes en curso, Portal de Datos Abiertos de la Comunidad de Madrid. Available at: http://datos.comunidad.madrid/catalogo/dataset/calidad_aire_datos_mes [Accessed April 22, 2020]

[53] Estimación del Producto Interior Bruto Municipal. Base 2015 Series homogéneas. 2015(p)-2018(a), Instituto de Estadística Comunidad de Madrid. Available at: https://www.madrid.org/iestadis/fijas/estructu/economicas/contabilidad/epibmb15tab.htm [Accessed April 22, 2020]

[54] Trabajadores afiliados a la Seguridad Social en alta que trabajan en la Comunidad de Madrid, Instituto de Estadística Comunidad de Madrid. Available at: https://www.madrid.org/iestadis/fijas/estructu/sociales/iss20.htm [Accessed April 22, 2020]

[55] Distritos de Madrid, Wikipedia. Available at: https://es.wikipedia.org/wiki/Anexo:Distritos_de_Madrid [Accessed April 22, 2020]

[56] Indicadores demográficos por Distrito, Banco de Datos del Ayto de Madrid. Available at: https://www.madrid.es/UnidadesDescentralizadas/UDCEstadistica/Nuevaweb/Demograf%C3%ADa%20y%20población/Indicadores%20Demográficos/C5000119.xls [Accessed April 22, 2020]

[57] Calidad del aire. Datos horarios años 2001 a 2020, Banco de Datos del Ayto de Madrid. Available at: https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=f3c0f7d512273410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default [Accessed April 24, 2020]

[58] Los barrios de Madrid con mayor renta media, Bankinter. Available at: https://blog.bankinter.com/economia/-/noticia/2019/10/22/barrios-madrid-mayor-renta [Accessed April 24, 2020]

[59] Ayuntamiento de Madrid. Afiliados que trabajan en la ciudad de Madrid según Régimen, Edad, Nacionalidad, Sector y Sección de actividad económica por Barrio y Sexo, para cada Distrito. Available at: https://www.madrid.es/UnidadesDescentralizadas/UDCEstadistica/Nuevaweb/Mercado%20de%20Trabajo/Afiliaciones%20SS/anuales/trabajan%20ciudad/I2220119.xls [Accessed April 24, 2020]

[60] Evolución de enfermedad por el coronavirus (COVID-19), Gobierno de España. Available at: https://datos.gob.es/es/catalogo/e05070101-evolucion-de-enfermedad-por-el-coronavirus-covid-19 [Accessed May 10, 2020]

[61] Población por comunidades y ciudades autónomas, Instituto Nacional de Estadística (España). Available at: https://www.ine.es/jaxiT3/Tabla.htm?t=2915 [Accessed May 10, 2020]

[62] Extensión superficial de las Comunidades Autónomas y Provincias, Instituto Nacional de Estadística (España). Available at: http://www.ine.es/inebaseweb/pdfDispacher.do?td=154090&L=0 [Accessed May 10, 2020]

[63] Viajeros y pernoctaciones por comunidades autónomas y provincias, Instituto Nacional de Estadística (España). Available at: https://www.ine.es/jaxiT3/Tabla.htm?t=2074 [Accessed May 10, 2020]

[64] INE, Encuesta Nacional de Salud. Available at: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176783&menu=resultados&secc=1254736195650&idp=1254735573175#!tabs-1254736195650 [Accessed May 10, 2020]

[65] Presupuestos de las Comunidades Autónomas: Sanidad, Expansión. Available at: https://datosmacro.expansion.com/estado/presupuestos/espana-comunidades-autonomas?sc=PR-G-F-31 [Accessed May 10, 2020]

[66] Google Trends, different searchs. Available at: https://trends.google.com/trends/ [Accessed May 10, 2020]

[67] Coronavirus: initial symptoms, *Redacción Médica*, 8 April 2020. Available at: https://www.redaccionmedica.com/secciones/sanidad-hoy/coronavirus-sintomas-iniciales-hasta-5-dias-despues-contagio-4776 [Accessed May 16, 2020]

[68] Coutelenq Á., "El "efecto fin de semana" hace repuntar los fallecidos y los contagiados del Covid-19 en Madrid", *COPE*, 7 April 2020. Available at: https://www.cope.es/actualidad/mas-madrid/noticias/efecto-fin-semana-hace-repuntar-los-fallecidos-los-contagiados-del-covid-19-madrid-20200407_672760 [Accessed May 16, 2020]

[69] Laerd Statistics. Available at: https://statistics.laerd.com/statistical-guides/types-of-variable.php [Accessed May 24, 2020]

[70] NC State University, Introduction to Principal Components and FactorAnalysis. Available at: ftp://statgen.ncsu.edu/pub/thorne/molevoclass/AtchleyOct19.pdf [Accessed May 24, 2020]

[71] NC State University, Introduction to Principal Components and FactorAnalysis. Available at: ftp://statgen.ncsu.edu/pub/thorne/molevoclass/AtchleyOct19.pdf [Accessed May 24, 2020]

[72] Economipedia, Puntuación estándar o tipificada. Available at: https://economipedia.com/definiciones/puntuacion-estandar-o-tipificada.html [Accessed May 24, 2020]

[73] Sánchez E., Clustering, Estadística II, April 2019.

[74] Matlab, Dendrogram. Available at: https://www.mathworks.com/help/stats/dendrogram.html [Accessed May 24, 2020]

[75] Hayes A., Variance, Investopedia, September 2019. Available at: https://www.investopedia.com/terms/v/variance.asp [Accessed May 24, 2020]

[76] Statistics How To, Regression Analysis. Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/ [Accessed May 24, 2020]

[77] Statsdirect, P Values. Available at: https://www.statsdirect.com/help/basics/p_values.htm [Accessed May 24, 2020]

[78] Liao C., p-value, *Wikipedia*, November 2014. Available at: https://commons.wikimedia.org/wiki/File:P-value_in_statistical_significance_testing.svg [Accessed May 25, 2020]