Technological Forecasting & Social Change

An International Journal

Usefulness of open data to determine the incidence of COVID-19 and its relationship with atmospheric variables in Spain during the 2020 lockdown

Jose Jacobo Zubcoff, Jorge Olcina, Javier Morales, Jose-Norberto Mazón, Asunción M. Mayoral

PII:	S0040-1625(22)00629-1
DOI:	https://doi.org/10.1016/j.techfore.2022.122108
Reference:	TFS 122108
To appear in:	*Technological Forecasting & Social Change*
Received date:	4 March 2022
Revised date:	2 September 2022
Accepted date:	16 October 2022

Please cite this article as: J.J. Zubcoff, J. Olcina, J. Morales, et al., Usefulness of open data to determine the incidence of COVID-19 and its relationship with atmospheric variables in Spain during the 2020 lockdown, *Technological Forecasting & Social Change* (2022), https://doi.org/10.1016/j.techfore.2022.122108

Usefulness of open data to determine the incidence of COVID-19 and its relationship with atmospheric variables in Spain during the 2020 lockdown

Jose Jacobo Zubcoff[2], Jorge Olcina[2], Javier Morales[1], Jose-Norberto Mazón[2], Asunción M. Mayoral[1]

[1]Miguel Hernandez University

[2]University of Alicante

Corresponding author:
Jose Jacobo Zubcoff
email: jose.zubcoff@ua.es
Universidad de Alicante
Carretera San Vicente del Raspeig s/n
03690 San Vicente del Raspeig – Alicante (Spain)

Acknowledgements

This work is funded by GVA-COVID19/2021/103 project from "Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana".

Abstract. The SARS-CoV-2 pandemic and the spread of the COVID-19 disease led to a lockdown being imposed in Spain to minimise contagion from 16 March 2020 to 1 May 2020. Over this period, measures were taken to reduce population mobility (a key factor in disease transmission). The scenario thus created enabled us to examine the impact of factors other than mobility (in this case, meteorological conditions) on the incidence of the disease, and thus to identify which environmental variables played the biggest role in the pandemic's evolution. Worthy of note, the data required to perform the study was entirely extracted from governmental open data sources. The present work therefore demonstrates the utility of such data to conduct scientific research of interest to society, leading to studies that are also fully reproducible. The results revealed a relationship between temperatures and the spread of COVID-19. The trend was that of a slightly lower disease incidence as the minimum temperature rises, i.e. the lower the minimum temperature, the greater the number of cases. Furthermore, a link was found between the incidence of the disease and other variables, such as altitude and proximity to the sea. There were no indications, however, in the study's data, of a relationship between incidence and precipitation or wind.

Keywords: COVID-19, incidence, atmospheric variables, lockdown, confinement, open data

1 Introduction

Efforts have been made, using available and accessible information, in almost all scientific disciplines to understand the incidence of the SARS-CoV-2 virus and the contagious COVID-19 disease, transmitted from person to person. In Spain, the comprehensive and uniform lockdown of the population between 16 March and 1 May 2020 that was decreed under the state of alert, allowed to minimise the effect of mobility, a key factor in the transmission of the disease (Courtemanche et al., 2020; Nolan, 2021; Sahoo & Sapra, 2020; Pan et al., 2020). In fact, mobility dropped sharply during the lockdown – by approximately 90% –, as suggested by the Apple report[1] (among others). Taking advantage of such an unusual situation, the objective of the present study was to examine the effect of geography and climatology on the incidence of the disease during this lockdown stage, in order to identify the key environmental factors in the pandemic's evolution under conditions of restricted population mobility. To this end, we used the open data provided by public administrations on the number of confirmed contagions across Spain's different autonomous communities (ACs), broken down by day and at the minimum geographic health districts or municipal level. This level of disaggregation allowed us to situate all the data, geographically and over time, and thus link the values of various sociological, geographic, climatic and meteorological variables (also retrieved from open data sources) to COVID-19 incidence, determining their impact on contagion. The

[1] https://www.apple.com/covid19/mobility

identification of the environmental factors that favour contagion is undoubtedly critical to define future policies directed towards the mitigation of SARS-CoV-2 transmission.

After having performed an intensive search across the open data web portals of Spain's different regional administrations, open and disaggregated information corresponding to the lockdown stage with a sufficient level of detail (i.e. daily and by municipality or by basic health district) was only found for 6 of Spain's 17 autonomous communities. In parallel, daily geographic and atmospheric data were obtained from each accessed area (i.e. from the most representative or nearest weather station), on the open data portal of the Spanish State Meteorology Agency (AEMET). Population sizes, needed to calculate the incidence rates, was extracted from the National Geographic Institute (IGN).

The data available and used in this work allowed us to analyse, in a situation of restricted population mobility and in the accessed Spanish geographic areas:

- The evolution over time of COVID-19 incidence rates (confirmed cases) per 1,000 inhabitants.
- The geolocated incidence per health district.
- The association between contagion and atmospheric as well as geographic factors.
- The prediction of contagion rates based on the most relevant atmospheric and geographic factors.

Finally, our work has been based entirely on open data sources available on government portals, which allows us to reflect on the usefulness of open government data in conducting scientific studies for the benefit of society. Therefore, the novelty of our work lies in forecasting whether open data could be used to make informed decisions in critical situations, such as the pandemic. Our paper aims to show the usefulness of available open government data to create models to anticipate the incidence of COVID-19 and contagion rates by taking advantage of the unusual lockdown situation in Spain. Our models and results suggest that open data would allow governments and practitioners to have insights on crisis scenarios, thus planning strategies to overcome them.

To sum up, contribution of our paper is twofold: (i) development of models to determine incidence and evolution of COVID-19 based on atmospheric and geographic factors, and (ii) using models to forecast whether open data would be useful for mitigating crisis scenarios, such as COVID-19 pandemic.

Our paper is structured as follows: Section 2 describe some relevant related work. Our methodology is stated in Section 3, while results are explained in Section 4 together with discussion of the proposed models. Conclusions and future works are sketched out in Section 5.

2 Related work

Containing the COVID-19 pandemic has become a challenge for global society and especially for science, in all domains.

Different models have been proposed by using a variety of data sources to determine incidence and evolution of COVID-19 and its influence on some social issues. For example, (Simionescu & Raišienė, 2021) use data from Google Trends to study unemployment rate during the pandemic and to enhance employment expectations. Different approach focuses on forecasting pandemic, such as (Miranda & Devezas, 2022) in which authors propose a model for explaining the different behaviour of the spreading of the COVID-19 disease among different countries and continents. Importantly, attempts have been made, in environmental, statistical and territorial disciplines, to decipher spread patterns of the SARS-CoV-2 coronavirus according to various atmospheric and environmental conditions. Scientific studies conducted throughout 2020 on the relationship between atmospheric and climatic variables and SARS-CoV-2 contagion are inconclusive. In most cases, results are preliminary (Gutiérrez-Hernández & García, 2020) and highlight the prominence of the human factor (mobility and contact) over environmental or atmospheric variables, which would be a complementary explanation of the pandemic's territorial spread.

In this regard, the Berkeley Earth report (Rohde, 2020) on the relationship between the COVID-19 pandemic and atmospheric conditions reduced the weight of the impact of meteorological and climatic factors on the pandemic's territorial spread, based on a range of studies published in affected countries. It also questioned the idea that summer weather conditions in the northern hemisphere contribute to lower transmission rates, as obtained in some countries that implemented strict control protocols and pandemic management (ISS, 2020; Olsen et al., 2020; Dezan, 2020). Xu et al (2020) point to similar conclusions on the weak link between climate factors and disease spread under conditions of population contact and mobility.

Morawska and Milton (2020) as well as Jones et al. (2020) point to two relevant issues regarding the relationship between the pandemic and atmospheric and environmental aspects: 1) the determination of climatic parameters affecting coronavirus transmission; and 2), the role of airborne transmission (aerosols). The detection of such relationships would allow the determination of the environments that are more and less favourable to the spread of the disease. These issues, however, continue to generate discussions within the World Health Organisation itself (WHO, 2020). Our work focuses on the first issue under discussion and seeks to identify the climatic parameters that present an association with COVID-19 transmission based on disease incidence data in Spain during a period of minimal population mobility.

As far as the effect of temperature is concerned, under controlled conditions, Chin et al. (2020) found that at 4^{o}C, the virus is very stable over an extended period. On the other hand, at 22^{o}C, there is a reduction in its stability following a 7-day incubation, and after 14 days, no infectious virus is detected at this temperature. At 37^{o}C, no infectious viruses were detected beyond the first day. Holtmann et al. (2020) found an association between room temperature and the time required to diagnose 100 new cases of COVID-19, based on the first 100 cases diagnosed in the different countries addressed in the study. They further observed that ambient and low temperatures appeared to be associated with a faster spread of COVID-19 during the early stage of the outbreak. Gupta (2020) concluded that for every 1^{o}C of temperature increase above 5^{o}C, there was a 10% decrease in the COVID-19 transmission rate.

Extending the analysis to other atmospheric conditions such as humidity or radiation, Araújo & Naimi (2020) identified that the optimal disease propagation conditions were temperatures around 5.81ºC (CI95%=(-3.44, 12.55)°C) and radiation values around 112.78 W/m² (CI95%=(61.07,170.96)W/m²). Bukhari & Jameel (2020) found a reduction in cases as from 17^{o}C on average and 9 g/m^{3} of absolute humidity. Ficetola and Rubolini (2020) detected greater propagation in areas with an average temperature of 5^{o}C and a specific humidity of 4-6 g/m$^{3.}$ Sajadi et al. (2020) point to high propagation when average temperatures range between 5-11 ^{o}C, specific humidity between 3 and 6 g/kg and absolute humidity between 4 and 7 g/m$^{3.}$ Bu et al. (2020) studied the weather conditions during the first weeks of the COVID-19 outbreak in Wuhan, and found that the environmental conditions favourable to the survival and spread of the virus were: a temperature between 13 and 24^{o}C; a relative humidity between 50 and 80%; and a monthly precipitation of less than 30 mm. In this latter study, temperatures above 24^{o}C seemed to slow the progress of the epidemic. Wang et al. (2020) analysed the effect of air temperature and humidity on COVID-19 transmission in 100 Chinese cities, using the daily values of the effective reproduction number (between 19 January and 10 February 2020). They concluded that high temperatures and humidity were significantly associated with lower virus transmission, suggesting—in their view—that the arrival of summer and the rainy season in the northern hemisphere could significantly reduce COVID-19 transmission, as in the case of the flu virus. Shi et al. (2020) found that the highest incidence of COVID-19 affected areas with a temperature around 10^{o}C, and an absolute humidity of 7 g/m $^{3.}$ Ma et al. (2020) analysed daily data on COVID-19 deaths, weather parameters and air pollution data, from 20 January to 29 February 2020 in Wuhan, China. As in the study of Shi et al. (2020), using a widespread additive model, they found that daily mortality from COVID-19 was positively associated with daily temperature variation and negatively associated with absolute humidity. In Brazil, Rosario et al. (2020) detected that over a 47-day time series in March and April 2020 in Rio de Janeiro, contagion incidence was lower under conditions of high temperatures (Daily Average T^{o} >23^{o}C), high relative humidity (RH >70%) and constant wind (>2m/s). Ramadhan et al. (2020) studied this relationship in Jakarta, obtaining high correlations with average temperature and air quality. Brassey et al. (2020) point out that, based on the evidence in the studies they reviewed, cold and dry conditions could have an impact on the spread of COV-2 SARS, affecting the stability of the virus and having an effect on the host. Ward et al. (2020), in their study of cases between February and May in New South Wales, highlight that for every 1% decrease in relative air humidity, the number of cases increased by between 7-8%. Other works have compared the climatic effects in several cities affected by COVID-19 against the climatic conditions of other less affected cities (Sajadi et al., 2020).

Mecenas et al. (2020), based on the analysis of 517 articles published throughout 2020 in impact journals included in databases and editorial repositories, conclude that warm and humid climates reduce the impact of COVID-19. They also specified, however, that temperature and humidity variables alone do not explain all the factors underlying the variability of disease transmission.

Table 1 summarises the conclusions mentioned above regarding the impact of different atmospheric parameters obtained from the studies of Araujo & Naimi (2020), Bukhari & Jameel (2020), Ficetola & Rubolini (2020), Sajadi et al. (2020), Wang et al. (2020) and Olcina, Biener and Martí (2020).

Table 1. Synthesis of atmospheric thresholds that favour the spread or reduction of SARS-CoV-2 infections, in the consulted bibliography

ATMOSPHERIC ELEMENT	FAVOURS CONTAGION	REDUCES CONTAGION
TEMPERATURE	Cold temperatures (4-12^{o}C)	Warm temperatures (>25^{o}C)
PRECIPITATION	Variable (<30 mm./month)	---
ABSOLUTE HUMIDITY	3-9 g/m^3	> 9g/m^3
RELATIVE HUMIDITY	<60%	>60%
SOLAR RADIATION	< 61 W/m^2 (average)	> 20 MJ/m^2

Elaborated by the authors based on the studies reviewed in this section: Araujo & Naimi (2020), Bukhari & Jameel (2020), Ficetola & Rubolini (2020), Sajadi et al. (2020), Wang et al. (2020), and Olcina, Biener and Martí (2020).

Based on studies on disease incidence in Spain, Oto-Peralías (2020) analysed the correlation, at the provincial level, between confirmed cases of COVID-19 and a range of geographic, meteorological and socioeconomic variables. He found a negative relationship between the average temperature of the months of February and March 2020 and the number of cases of COVID-19 accounted for on 1 April (R^2=0.62). The temperature and population density together would explain up to 66% of the variation of confirmed COVID-19 cases. In a joint work by the Carlos III Health Institute (ISCIII) and the State Agency for Meteorology (AEMET) (AEMET & ISCIII, 2020), the number of new daily contagions per 100,000 inhabitants, accumulated over 14 days in March (March 13-26), was associated in the different Spanish autonomous communities, with the average temperature recorded in each community during the same period. A negative exponential relationship was found between the two variables (the higher the temperature, the lower the contagion rate), with R^2=0.52 in March, and R^2=0.62, with added data for the first week of April (26 March to 5 April). Lastly, at the provincial level, Briz-Redón & Serrano-Aroca (2020) analysed the relationship between the incidence of COVID-19 and different environmental factors, using space-time modelling techniques and considering both fixed and random effects. They found no evidence whatsoever that higher average, minimum or maximum temperatures were related to a reduction in COVID-19 cases. They clarified, however, that their results should be interpreted with caution, given the uncertainty of the data themselves and the confusion factors. In addition, they warned against the risk of extrapolating conclusions to other temperature ranges and that other non-meteorological, spatial and temporal effects should be addressed.

Olcina, Biener and Martí (2020) also focused on the provincial level and highlighted that the synoptic situations during the month of February 2020 supported a lower development of contagion in areas with mild winter weather, such as Alicante, which encourages outdoor living and activities, in contrast to other regions of the interior or northern Spain with lower temperature records (T^{o}). They also noted the good air quality (low levels of NO_2 and of PM2.5 and 10 particles) and summer weather conditions (monthly average >25^{o}C; tropical nights T^{o} >20^{o}C, constant coastal sea breeze and high solar radiation values >20 MJ/m^{2}) as conditions that would possibly explain a lower incidence of contagions in Alicante, compared to the higher incidence in other regions of northern and central Spain.

Other works have studied the relationship between the UV index, hours of sunshine and prevalence of COVID-19. In Takagi et al. (2020), a meta-regression of the prevalence of COVID-19 was conducted in cities of over 500,000 inhabitants in the United States, and a higher number of hours of sunshine as well as a greater UV index were found to be associated with a lower prevalence of COVID-19. Another similar study, this time on the influenza virus (Ianevski et al., 2019), found a similar pattern in northern European areas, with data from 2010 to 2018. Other studies, however, have not found such an association with temperature or UV index (Yao et al., 2020), based on data from 224 cities in China.

Thus, based on the prior research, the relationship between warm temperatures, high relative humidity and air movement with a lower incidence of COVID-19 contagions -without taking other factors (social, mobility variables) into account- seems to present significant reliability (Dbouk, 2020). Hence the recommendations of health authorities in the affected countries and of the WHO itself on the use of outdoor spaces or constantly ventilated indoor spaces for activities and social life.

Nevertheless, only few studies considered the possible effects under conditions of restricted population mobility during lockdown periods, when the impact of mobility on transmission is attenuated and patterns relating to the atmospheric variables specific to each place may appear (Ma et al., 2020; Liu et al. 2021; Pincombe et al., 2021). Ma et al in 2020 studied with a GAM approach the possible relationships of environmental variables with COVID-19 mortality in the Chinese population, finding relationship with daily temperature range and negative relationship with humidity. Liu et al in 2021 used a regression discontinuity in time (RDiT) design to estimate the effectiveness of lockdown policy interventions. However, they did not include environmental effects in the study, only comparing discontinuities according to the restrictions introduced and their effect over time. Pincombe et al. in 2021 studied the effectiveness of national containment and closure policies at different income levels during the COVID-19 pandemic. and closure policies at the national level at different income levels during the COVID-19 pandemic. They used regression models to compare results across 113 countries, although they did not take into account environmental or climatic variables, focusing on aspects of constraints and income per country. Few studies have considered the pandemic's temporal evolution jointly with climatic conditions, and the possible relationship with the geographic diversity of the areas under study (Briz-Redón & Serrano-Aroca, 2020; Sun et al. 2020). Briz-Redón et al. in 2020 analysed the relationship between temperature and the evolution of the COVID-19 pandemic in Spain. They did not study the daily temperature range or any other climate variable. In addition, they used unofficial data that are not currently available on the website they linked to. They have used a mixed effects regression model fitted with an INLA approximation. They did not find any relationship with temperature at the level of provinces in Spain. Sun et al. in 2020 studied the relationship between special geographical information (latitude, longitude and latitude) and the daily cumulative number of people infected with COVID-19 in China. They also proposed a regression model but did not include climatic data (such as temperature, humidity, etc.). They found a negative correlation between latitude/altitude and the number of daily COVID-19 infected cases. In this regard, we present a novel approach in this work studying the possible effect of various atmospheric and geographic variables on the spread and evolution of COVID-19 across Spain during the strictest lockdown period, which was decreed for the entire population between 16 March and 1 May. The situation generated similar patterns of restricted mobility throughout Spain, allowing us to study virus transmission under conditions of restricted population mobility.

Finally, it is worth noting that there are works that focus on studying the usefulness of big data for dealing with social issues in certain regions, such as the proposal in (Dubey et al, 2019) that investigates the effects of big data on social and environmental performance in Indian manufacturing organisations. Under this big data umbrella, our work aims to study the usefulness of open government data (health and atmospheric data) to determine the incidence of COVID-19. Actually, our work was based entirely on available open data sources from government portals, thus highlighting the key role the FAIR principles (Findable, Accessible, Interoperable and Reusable data) in data management (Wilkinson et al., 2016).

3 Methodology

The following sections describe the data sources used in the present work, the process of integrating them, and the methodology used for data analysis.

3.1 Data used in the analysis

The data was retrieved from the web portals in which the ACs have been publishing data on the evolution of the COVID-19 disease in their territory, sometimes broken down by municipality or by basic health district (that is, the smallest geographic health area delimitation, used as a reference for the planning and organisation of primary care teams, a group of health and non-health professionals who, in essence, are responsible for attending to the population in the area, as defined by Spain's Ministry of Health[2]). The input data give the number of new cases of daily COVID-19 infections (confirmed by a positive PCR) per municipality or basic health district.

With regard to meteorological and geographic data, we retrieved the meteorological data from stations closest to the districts (through the AEMET open data portal), as well as the population size of each municipality within each basic health district (municipal information was provided by the National Centre for Geographic Information (CNIG)).

3.1.1 Contagion records in the ACs' Open Data Portals

The ACs that offered open data portals and complete information on new daily cases of COVID-19 contagion are listed below, together with their restrictions:

- **Castile and Leon**: provides rates of sick people per basic health district.[3]
- **Catalonia:** provides the number of COVID-19 cases disaggregated by sex and municipality.[4]
- **Valencian Community:** details the number of cases and rates per 100,000 inhabitants per basic health district.[5]
- **Basque Country:** lists the number of cases in the Basque Country by municipality and date.[6]
- **Madrid:** accumulated incidence by municipality and district in Madrid.[7]
- **Navarra:** provides the number of cases of COVID-19 by basic health district.[8]

[2] https://www.mscbs.gob.es/ciudadanos/prestaciones/centrosServiciosSNS/hospitales/introduccionCentro.htm

[3] https://datosabiertos.jcyl.es/web/jcyl/set/es/salud/tasa-coronavirus-zonas-basicas-salud/1284942912395

[4] https://dev.socrata.com/foundry/analisi.transparenciacatalunya.cat/jj6z-iyrp

[5] https://dadesobertes.gva.es/dataset/361a20e5-63ec-42ce-88ec-5b5f09fb8f46

[6] https://opendata.euskadi.eus/w79-dataset/es/contenidos/ds_informes_estudios/covid_19_2020/es_def/index_es.html

[7] https://datos.comunidad.madrid/catalogo/dataset/7da43feb-8d4d-47e0-abd5-3d022d29d09e

[8] https://gobiernoabierto.navarra.es/es/open-data/datos/datos-zonas-basicas-salud-covid-19

To make the analysis viable, and given the level of disaggregation we encountered, the study generically refers to the "district" as the minimal geographic agglomerate for which public disaggregated information exists, and which may correspond to a municipality or to a basic health district.

Table 2 illustrates the number of records available only for the ACs that provided disaggregated information, together with the start and end dates (consulted as of May 30 and which this study is based on) for the accessible data, which is also indicated using the week number in 2020.

Table 2. Information on the retrieved COVID-19 open data.

AC	No. records	From	To	Initial week	Final week
Castile and Leon	13554	2020-03-01	2020-05-29	09	22
Basque Country	12917	2020-03-21	2020-05-29	12	22
Catalonia	7470	2020-02-28	2020-05-29	09	22
Madrid	15681	2020-02-27	2020-05-29	09	22
Navarre	1624	2020-03-26	2020-05-29	13	22
Valencian C.	1012	2020-04-13	2020-05-26	16	22

Excluding the Valencian Community, the remaining five provide information from 21 March onwards. The available data from 9 March (week 9) to 31 May (week 22) were used essentially to relate the conclusions to the stage of total lockdown (between 14 March and 2 May), plus a pre- and post-period, covering the final de-escalation stage.

3.1.2 Weather data (atmospheric variables) and geographic data

The State Agency for Meteorology (AEMET) offers open and downloadable data on its open data portal[9] (AEMET), including minimum and maximum daily temperatures from Spain's main observatories. This open portal from AEMET overcomes some of the pitfalls exposed by (Gutierrez-Corea et al, 2013) regarding weather data, and although, main observatories of AEMET are located in provincial capitals, airports and mountain areas, they provided valid and sufficient information to

[9] https://opendata.aemet.es

conduct the analysis included in the study, since each basic health district was linked to its nearest main observatory of reference.

However, other atmospheric variables such as humidity, hours of sunshine, UV, etc. were not fully consulted because they were not openly available on the AEMET portal.

The atmospheric variables that were obtained from the AEMET open data portal and were eventually used were the following:[10]

- Minimum temperature.
- Average temperature.
- Maximum temperature.
- Temperature range (maximum - minimum).
- Precipitation.
- Average wind speed.
- Maximum gust.
- Maximum pressure.
- Minimum pressure.

Geographic data, referring to the geographic location (longitude, latitude) and altitude of each of the basic health districts, were obtained from the open data portal (called the "Download Centre") of the National Geographic Information Centre (CNIG)[11] in its geographic information section of reference (basic topographical data required to represent the territory).

3.2 Data integration processes

To fulfil the goals of this study by using open data entirely extracted from governmental sources, some issues remained to be solved. Actually, heterogeneous open data coming from three different sources (weather data, geographical data, and COVID-19 data) had to be harmonized and integrated to get the required quality. To do so, a brokering approach was followed, in the sense stated by (Gallagher et al. 2015), to design data integration processes that transform input data in the adequate form for consumption. Specifically, the following data integration processes were performed in order

[10] If any value for any of the atmospheric variables did not exist for any given weather date and station, the value was defined as NA (i.e. data not available).

[11] http://centrodedescargas.cnig.es

to collect, sort and standardise the input data, in order to obtain a usable dataset of quality for the analysis.

These data integration processes allowed the solving of various input data problems:

- Some ACs reported cases of people affected by COVID-19 on a given day, while others offered cumulative totals up to a specific date. In the latter situation, daily cases were obtained from those of the cumulative total by designing the corresponding query that used *"LAG and LEAD analytic functions"*.
- Contagion records were related to geographic location data (longitude, latitude and altitude of the National Geographic Information Centre) via the municipality name.
- The records were linked to the most representative weather station of each basic health district, based on geographic location.
- The number of inhabitants in each municipality was calculated using the data published by the National Geographic Information Centre of the National Geographic Institute.[12] To calculate that of each basic health district, we summed up the number of inhabitants of the municipalities included in the district.

Three datasets were obtained from these designed data integration processes:

1. Dataset 1 (DB1): All available data on the number of COVID-19 cases, and their incidence in terms of rates per 1,000 inhabitants (cases/population x 1000), along with climatic and geographic variables, disaggregated by district and per day. This database was used for preliminary descriptive applications and to create the aggregated databases DB2 and DB3 described below.
2. Dataset 2 (DB2): Given the daily variability and weekly seasonal behaviour detected, part of the analysis unfolded by aggregating the number of cases per week (and their rates), to obtain a clearer view of the pandemic's temporal evolution. In this case, we also calculated the logarithm of weekly incidence rates, such as the logarithm of the ratio between the number of weekly contagions, adding one unit to avoid non-determinations, and the district population size. In parallel, using the daily temperature records, we calculated the weekly minimum, maximum and average temperatures in addition to each week's maximum daily temperature variation.

[12] http://centrodedescargas.cnig.es/CentroDescargas/catalogo.do?Serie=CAANE

3. Dataset 3 (DB3): Leaving out weather information and retaining only geographic location, the overall aggregate contagion rate (per 1000 inhabitants) was calculated during the period under study based on the total number of contagions.

3.3 Statistical analysis methodology

The study was broken down into three steps based on the available information:

1. A preliminary descriptive study based on the temporal and territorial evolution of COVID-19.
2. A correlation study to identify the variables most related to incidence, conducted through a principal component analysis (PCA) of weekly aggregate rates (DB2) and aggregated rates for the whole period (DB3). This allowed us to detect association patterns between the atmospheric variables available and the spatial distribution of COVID-19 across both considered time spans: weekly and whole period. To study the weekly incidence rate data, we used the atmospheric data for the week during which the contagion occurred, i.e. two weeks before each case was counted and recorded in the database. A classification or clustering was then conducted, using the K-Mean (Hartigan & Wong, 1979) technique, to identify variables related to each other, and therefore using shared information to explain the incidence.
3. Both the descriptive analysis and the PCA guided the initial variable selection so as to propose reasonable models and to predict the pandemic's incidence and evolution. Normal linear models with nested effects were used, with the log response of the incidence rates by week (DB2). Model selection was resolved in terms of the AIC (Akaike, 1974). The model was validated in accordance with the usual procedures for testing the linear model hypothesis (found in any basic manual, for example Chaterjee et al., 2000).

3.4 Software

In this section, we describe the software that was used to process the input data, as well as the design of the developed data integration processes. We also describe the software and libraries employed in the data analysis.

3.4.1. Pentaho Data Integration

The Pentaho Data Integration[13] software provides tools to develop data integration processes, called ETL (Extraction/Transformation/Load), that is, processes that collect data from various sources, transform it by applying certain functions, and store it in a unique schema. We used the Spoon tool (Pentaho Data Integration GUI) to develop the necessary data integration processes.

In addition, the processes were subsequently introduced into a *data hub* (called EasyDataHub4COVID[14]) in which the updated data was automatically incorporated. In this way, the collection and integration of data will evolve as the various ACs publish more data.

The Pentaho Data Integration GIS plugin developed by ATOL[15], available on the marketplace[16] and in the GitHub repository[17] was also used. This plugin made it possible to spatially process the input data (weather station data and data from basic health districts)

Figure 1 shows an example of a developed ETL process, illustrating how the COVID-19 data of the Valencian Community was collected, processed and how data from weather stations was integrated. Once the Valencia Input Data on COVID-19 infections was collected, we standardised a range of data such as the date format ("Manage Date Format") and the number of daily cases ("Compute Daily Cases"). For its part, data from the nearest weather stations were associated with each basic health district. All these data sources were then gathered into a single output containing all the data needed for the later analysis ("Output Data").

[13] https://community.hitachivantara.com/s/article/data-integration-kettle

[14] The EasyDatahub platform allows you to easily publish and share open data. This platform is the outcome of the research project "Platform for the publication and consumption of open data for a smart city (Publi@City)" (Ref: TIN2016-78103-C2-2-R) funded by Spain’s Ministry for Science and Innovation. It has been adapted to the pandemic crisis and includes COVID-19 data as well as atmospheric data. The platform can be consulted at https://wake.dlsi.ua.es/datahub/

[15] https://www.atolcd.com/

[16] https://marketplace.hitachivantara.com/pentaho/

[17] https://github.com/atolcd/pentaho-gis-plugins/tree/master/pentaho-gis-plugins

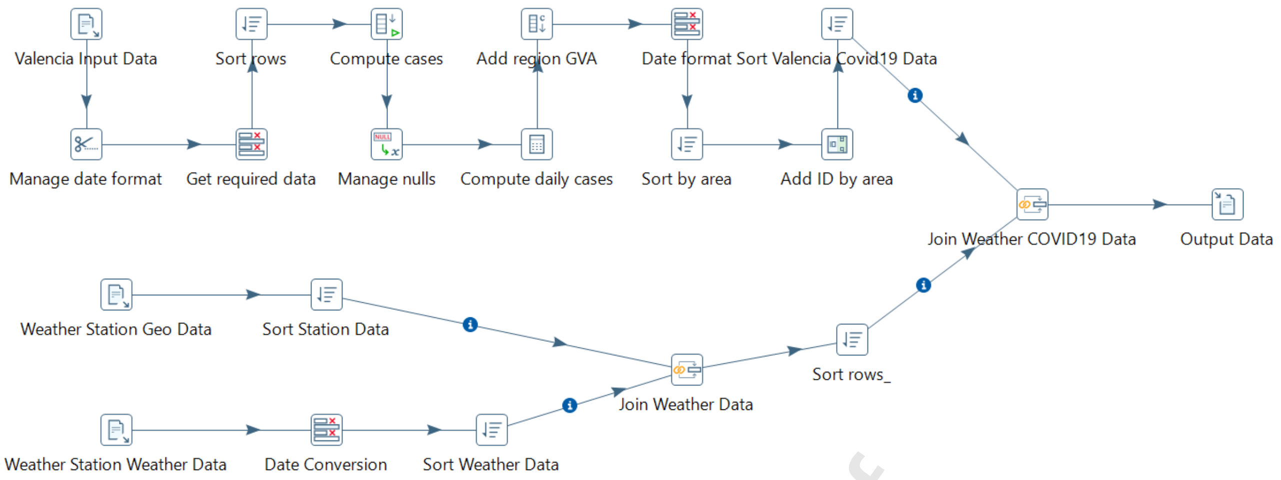

Figure 1. An excerpt of the ETL process designed in Pentaho Data Integration for processing input data.

3.4.2. R and RStudio

The descriptive and predictive analysis was resolved with R (R Core Team, 2020) and the RStudio interface (RStudio Team, 2020).

4 Results

4.1 Descriptive analysis

The temporal analysis of the incidence of COVID-19 over time (days), was obtained from DB1, for the aggregated rates both per AC and overall, in Spain (a representation by district was less informative due to the large amount of information: 1,055 different districts). All graphs by AC showed the same behaviour displayed in Figure 2, relating to overall rates in Spain: a periodic pattern of weekly incidence, with peaks between Tuesdays and Fridays, and a significant drop at weekends. This behaviour could not represent a real systematic infection pattern but could rather have been due to administrative gaps based on test delays, confirmation and case records. As a day effect is not of interest, weekly seasonality was avoided by aggregating weekly data. This aggregation had an impact on the loss of daily climatic information and required the use of weekly averages and/or extremes. In Figure 2, the white background indicates the period of total lockdown in Spain, clearly illustrating the effect of lockdown on incidence, once the peak was reached during week 13 at the end of March; grey backgrounds correspond to incomplete lockdown periods.

Conversely, the weekly rate, aggregated per AC and shown in Figure 3, captures the effect of time on the propagation/reduction of the pandemic in a much better way. In Figure 3, the notable differences in COVID-19 incidence across Spain's various regions are distinctly visible, with the Castile and Leon AC being the most affected by COVID-19 based on the data available. Also clear in the figure is the availability of data over time for all ACs considered in the study, the Valencian Community presenting the greatest deficit of data (data being provided only as from week 16).

Figure 4 shows the notable variations that can be found within districts and between them by AC, per week (the points represent weekly rates by district, the boxes delimit the quartiles within the AC, and the lines indicate the minimums and maximums). The high variability during the most critical weeks of the pandemic, in weeks 12 to 15, is key, as well as the high variability between districts in Castile and Leon and Catalunya. This variability will be examined later using the proposed statistical models, based on climatic and geographic variables.

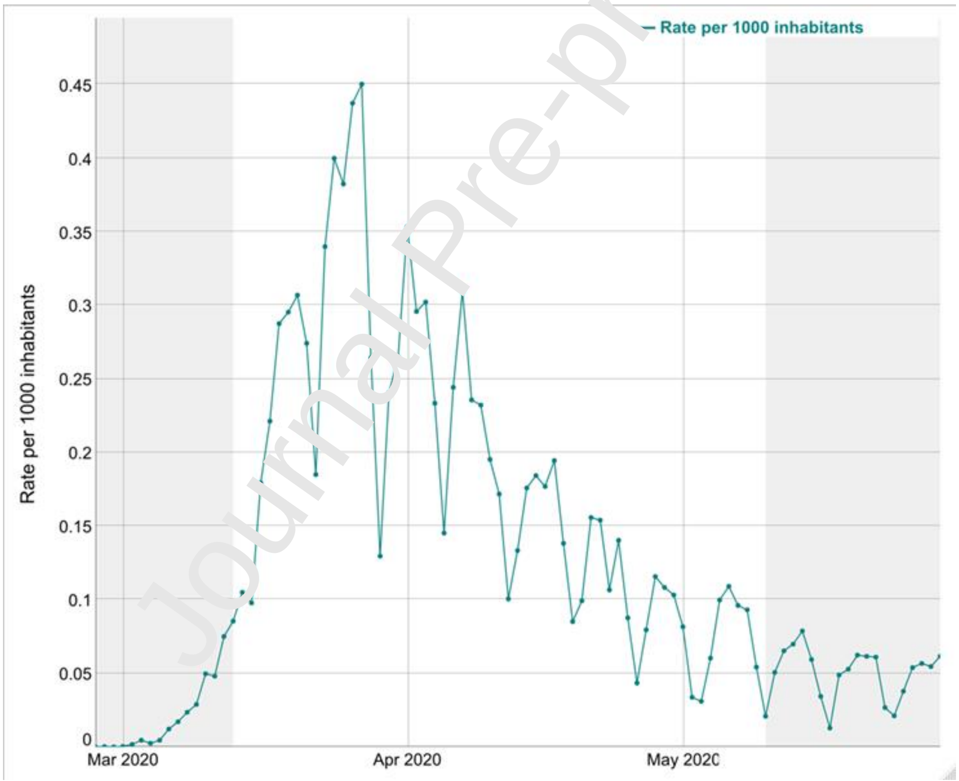

Figure 2. Evolution of the COVID-19 incidence rate (per 1,000 inhabitants) in Spain, between weeks 9 and 22 of 2020. The white background represents Spain's total lockdown period; the grey background represents incomplete lockdown.

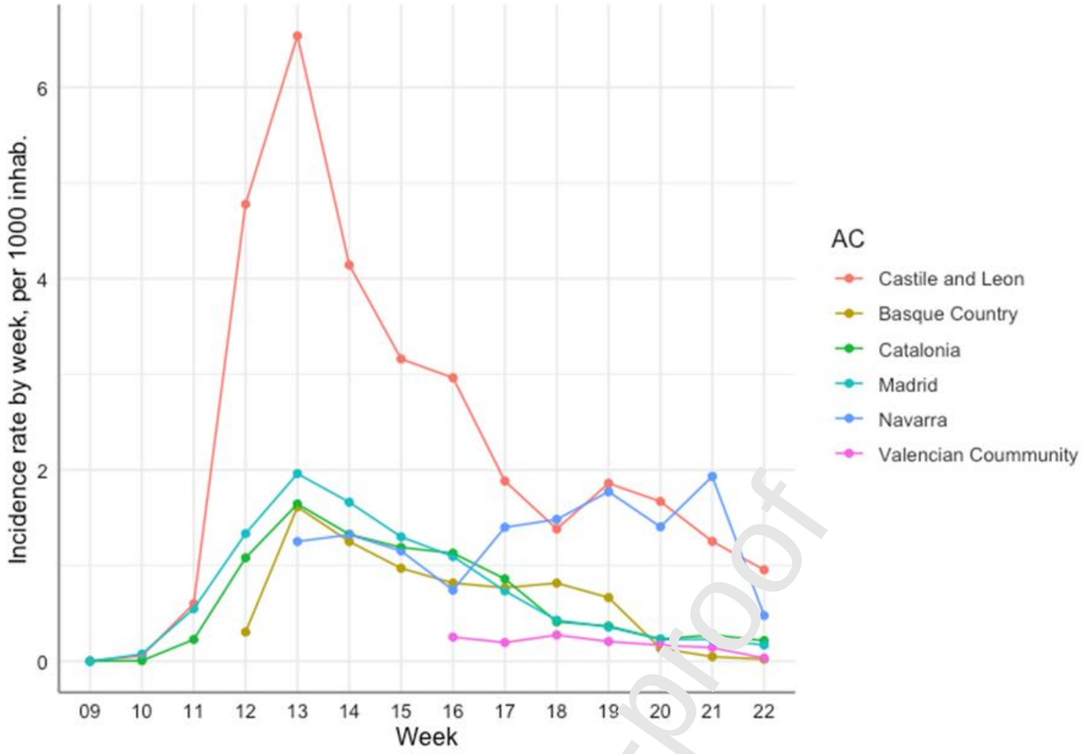

Figure 3. Evolution of COVID-19: weekly incidence rate (per 1000 inhabitants) in Spain, shown per AC, between weeks 9 and 22 of 2020.

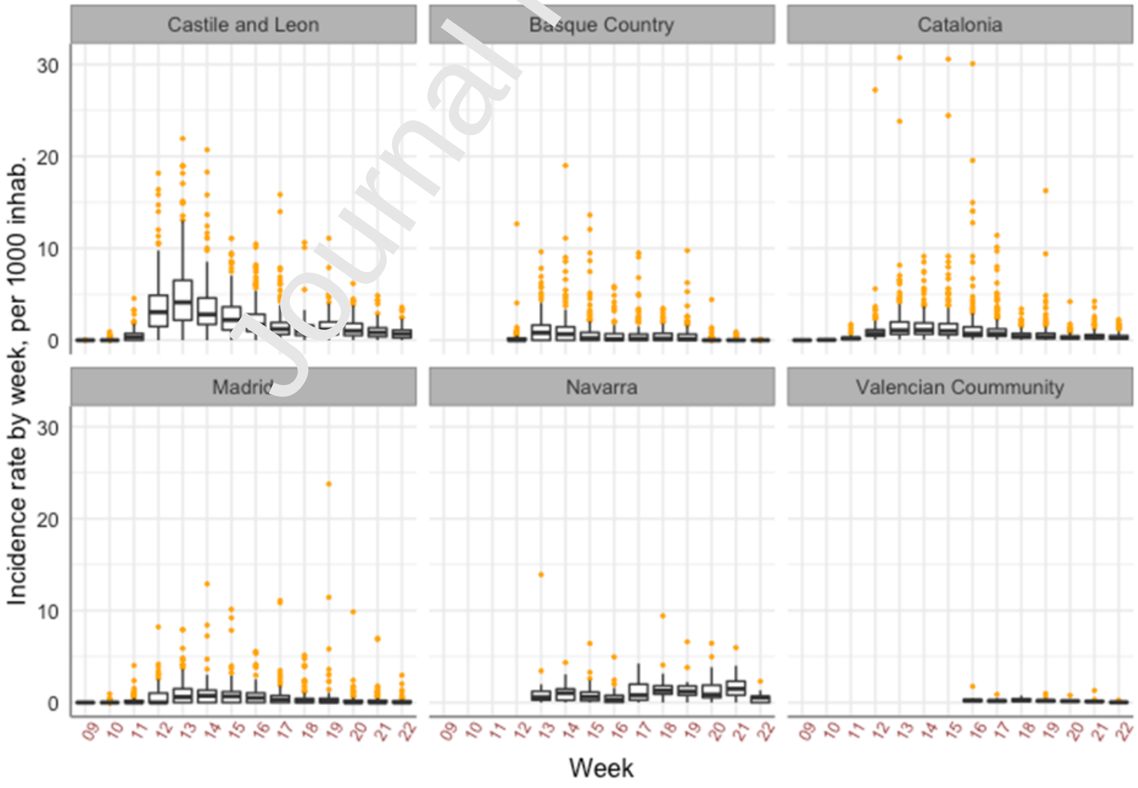

Figure 4. Evolution of COVID-19: weekly incidence rate (per 1000 inhabitants) in Spain, shown by district and AC between weeks 9 and 22 of 2020.

4.1.1 Geographic and weather descriptive data

To explore the effect of geographic variables on the incidence of COVID-19, rates for the whole period (per 1000 inhabitants) by district at the end of the study period (DB3) are displayed in Figure 5. More severe incidences are characterised by darker colours. In addition, Figure 5, which shows many blank territories, illustrates the great deficiencies of open data websites across Spain’s different autonomous community administrations.

The map in Figure 6 shows the incidence rate for the whole period by district (represented by the point radius) in relation to the minimum temperature recorded over the period of interest (differentiated by colour). In absolute terms, Castile and Leon (the largest region, centre-right on the map) was the most affected AC among those considered in the study, with incidence rates above 30% in many districts (identified with more intense colours). Navarra (above Castile and Leon, to the right), and Catalonia (in the upper right-hand corner of the map) show significant rates. The low incidence in the Valencian Community is due to a lack of data, which was only available as from week 16, once the pandemic’s severity had subsided.

This global analysis provided clues for the identification of patterns on the impact of the pandemic as Spain’s lockdown came to a close.

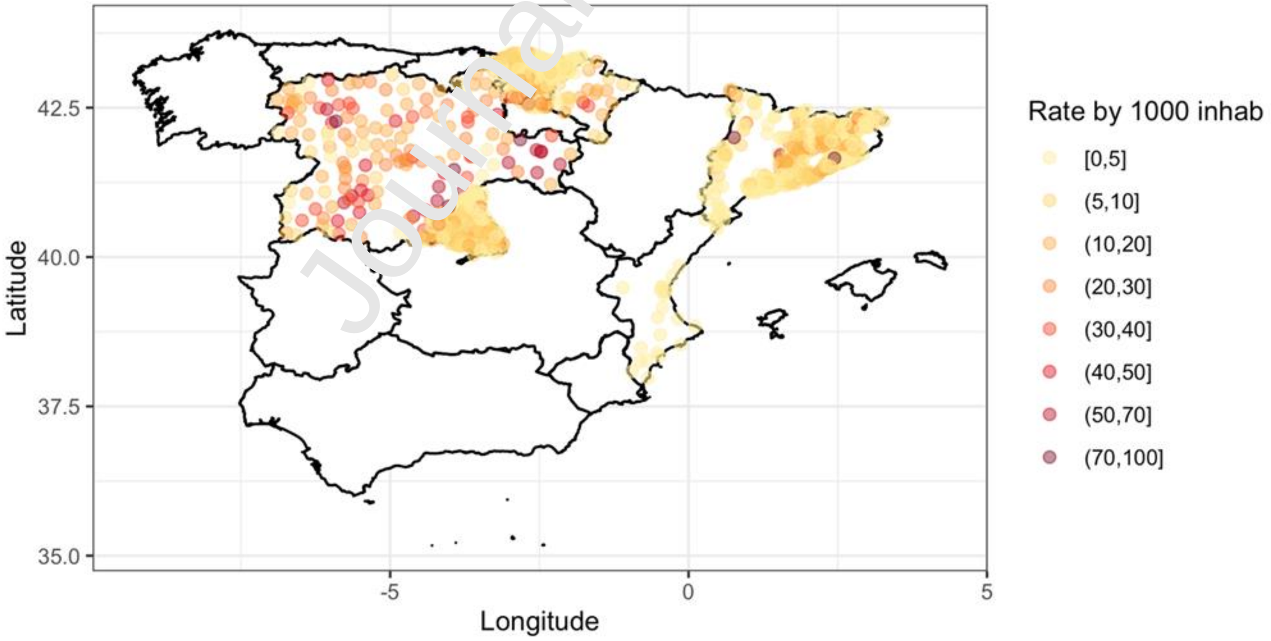

Figure 5. Spatial pattern in terms of rates per 1,000 inhabitants.

As one can observe, areas with higher minimum temperatures (colours turning towards orange and excluding the Valencian Community due to a deficit of available data as commented earlier) present a lower incidence of COVID-19. Conversely, areas with lower minimum temperatures (dots tending towards blue) present a higher incidence rate.

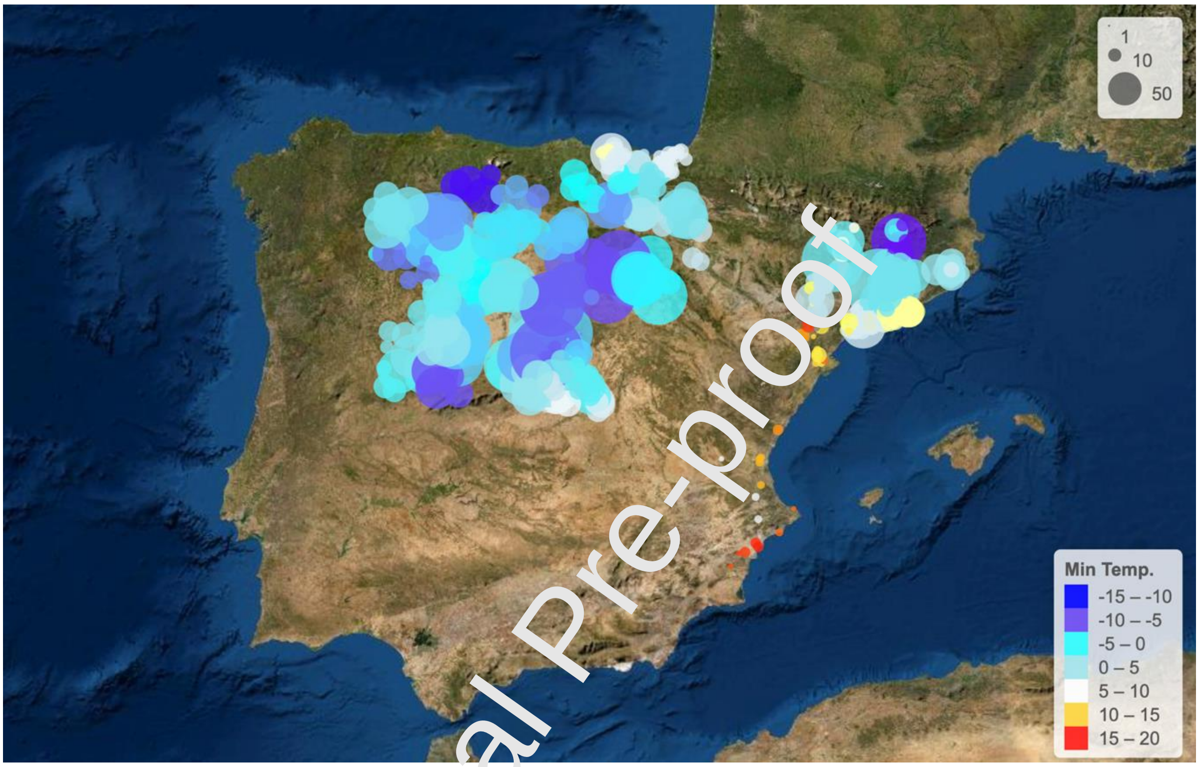

Figure 6. Incidence rate of COVID-19 per 1000 inhabitants and its relationship with minimum temperatures.

4.2 Main Component Analysis of climatic and geographic variables.

The Principal Component Analysis (PCA) technique was applied to identify possible relationships between atmospheric variables and the incidence of COVID-19 in the observed districts in the DB2-14 dataset. All the variables represented in Figure 7 are those that define the PCA plane. The first two dimensions of the PCA explain 62% of the total inertia. The variables with the greatest weight in the inertia are identified with the longest vectors. The shortest vector is the incidence rate, implying a slight association of these atmospheric variables with COVID-19 incidence. Since the COVID-19 incidence vector points to the left and downward, the variables in the third quadrant (with the incidence rate) will give a greater incidence the longer the vector that represents it. Conversely, those in the first quadrant will give a smaller incidence the longer the vector. That is, the higher the altitude and the lower the minimum temperature, the higher the COVID-19 incidence rate. On the other hand, variables whose vectors are almost perpendicular to the incidence of COVID-19 have little or no effect.

The bubbles illustrated the areas covered by 85% of the points of an AC. Different colours are given to each AC. As one can observe, the only one that does not include the centre is the orange bubble (Valencian Community), whose incidence and altitude data are lower and the minimum temperature is higher, consistently compared to the rest of the regions.

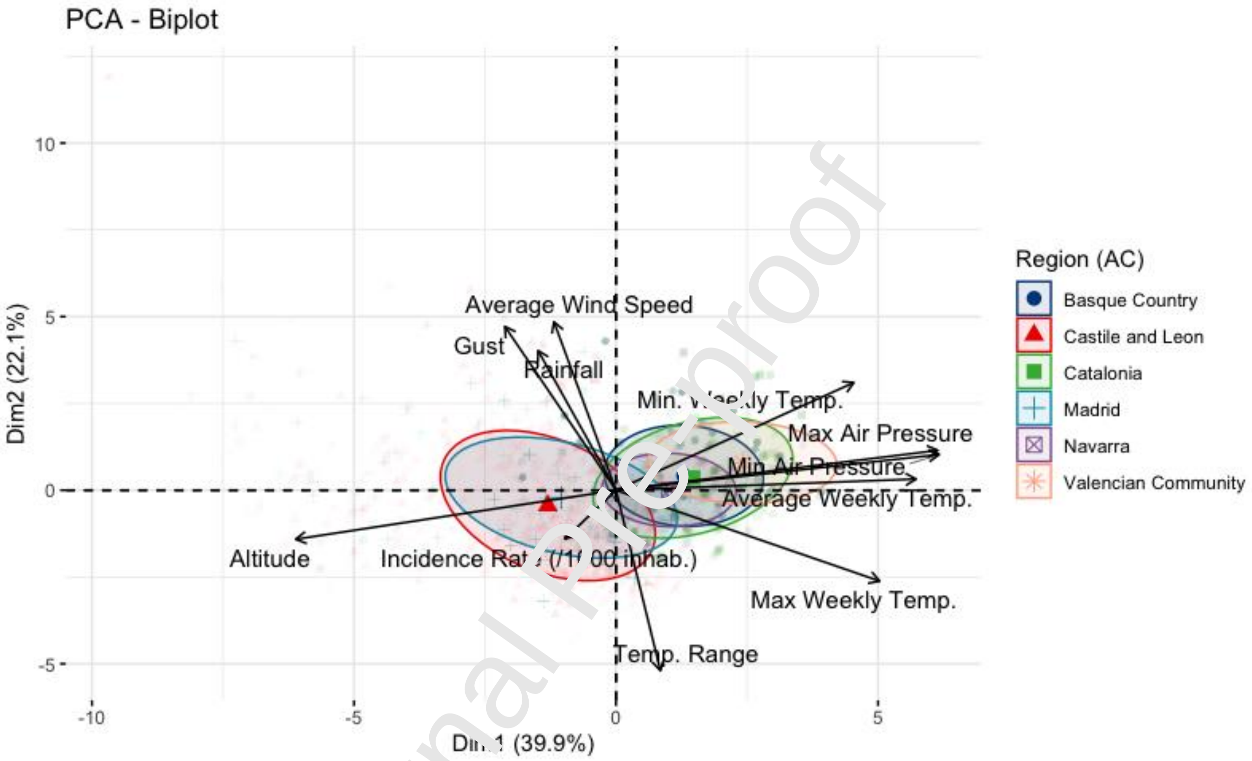

Figure 7. First two dimensions of the PCA for the districts (coloured by AC) and atmospheric variables and COVID-19 incidence represented by vectors.

The possible segmentation according to altitude was also studied and represented in Figure 8. Areas with altitudes over 1500m tended to have a rising weekly incidence rate with respect to lower altitude districts. The higher altitude cluster (in blue) presents points with a greater COVID-19 rate and is clearly differentiated from the other two altitude groups.

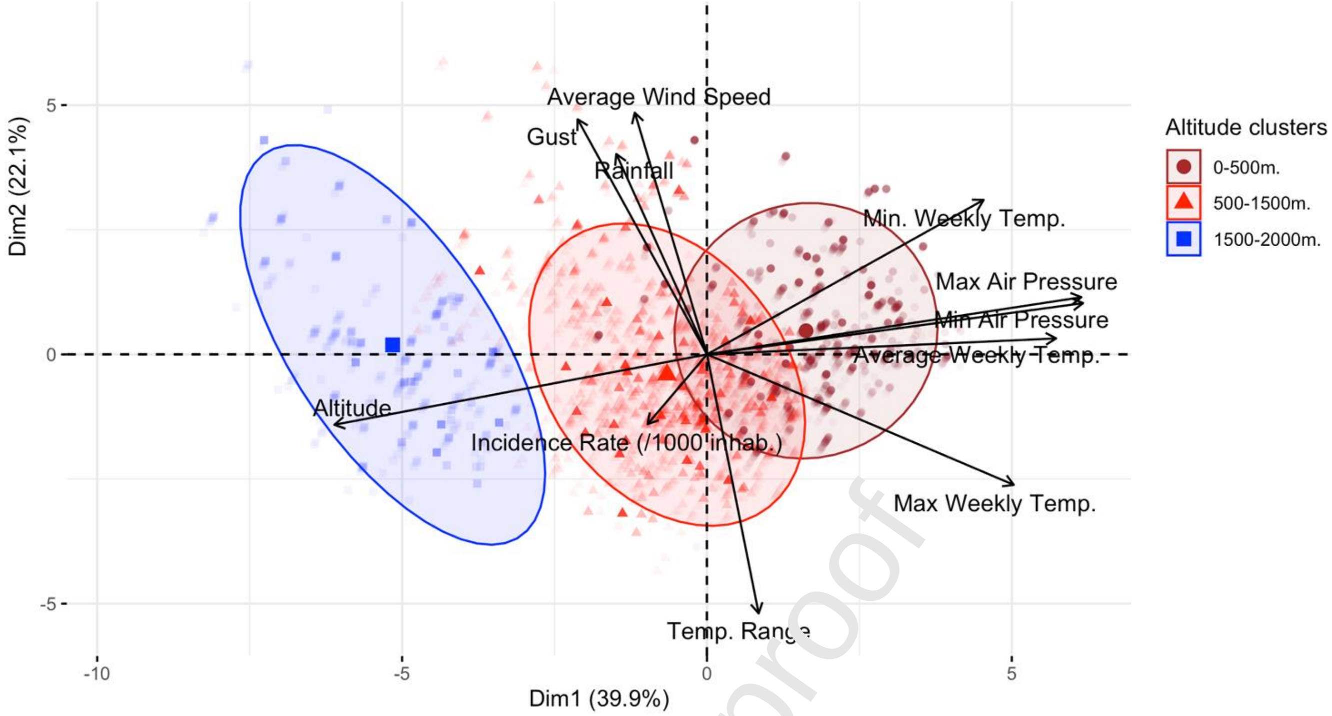

Figure 8. First 2 dimensions of PCA with altitude clustering.

4.3 Modelling of the weekly behaviour of the incidence rate of COVID-19 based on atmospheric and geographic variables.

Based on the conclusions obtained in the PCA analysis, the focus was on the most relevant climatic variables, e.g. the minimum and maximum daily difference in temperatures registered each week, available in DB2. In order to avoid presumably unreal delay effects, in the fitted models below, incidence rate data is related to the climatic values in a synchronous way, that is, each cumulative incidence weekly rate is regressed on the climatic weekly value registered for that week. All available data, disaggregated by district, is included in the fit.

We proposed two approaches to predict the COVID-19 incidence over time (a weekly effect), according to geographic variables (region, province, altitude, etc.) and to climatic variables relative to extreme temperatures, as found in the PCA. Lower minimum daily temperature was identified to be associated with higher COVID incidence. Moreover, as an alternative way to identify colder districts, the maximum daily temperature variation was proposed as an explanatory factor of the incidence rate.

Figure 9 displays the relationship between minimum weekly temperature and the incidence rate on a log scale, separated by AC in order to visualise possible differences in the observed trend. The linear effect of temperature in the log-rate is very clear, despite the high variability of the available data. As a result, the smaller the minimum weekly temperature, the more severe the incidence of the pandemic would be.

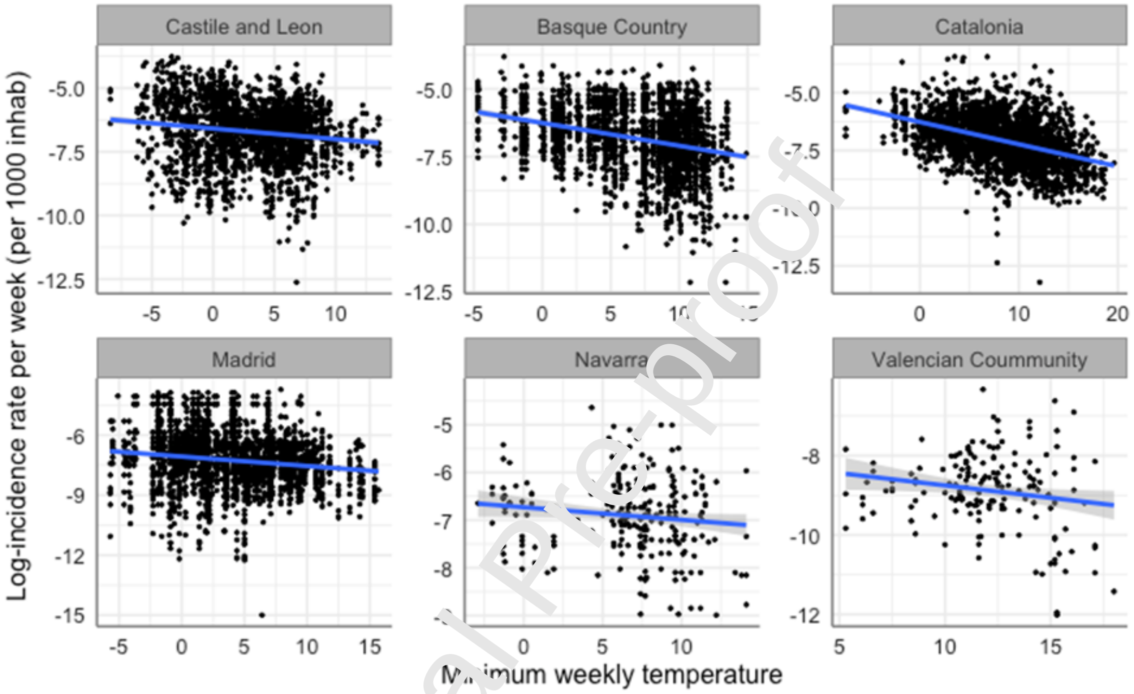

Figure 9. Relationship between minimum weekly temperature and COVID-19 incidence (on a log scale).

Figure 10 displays a log scale showing the relationship between the maximum daily temperature per week and the incidence rate by AC, to visualise possible differences in the observed trend. Again, a linear trend is perceivable, this time with a positive effect of temperature on the incidence rate, except in the autonomous community of Madrid. Again, the data is highly variable. Thus, a greater variation in daily temperatures would be associated with a greater incidence of the disease. Regarding the results observed in Madrid, other variables such as mobility may have had greater weight, since even if there is confinement for reasons of supply and services, a certain amount of mobility is necessary, which in densely populated areas can have greater effects than other environmental factors. However, a pattern can be observed in the rest of the regions.

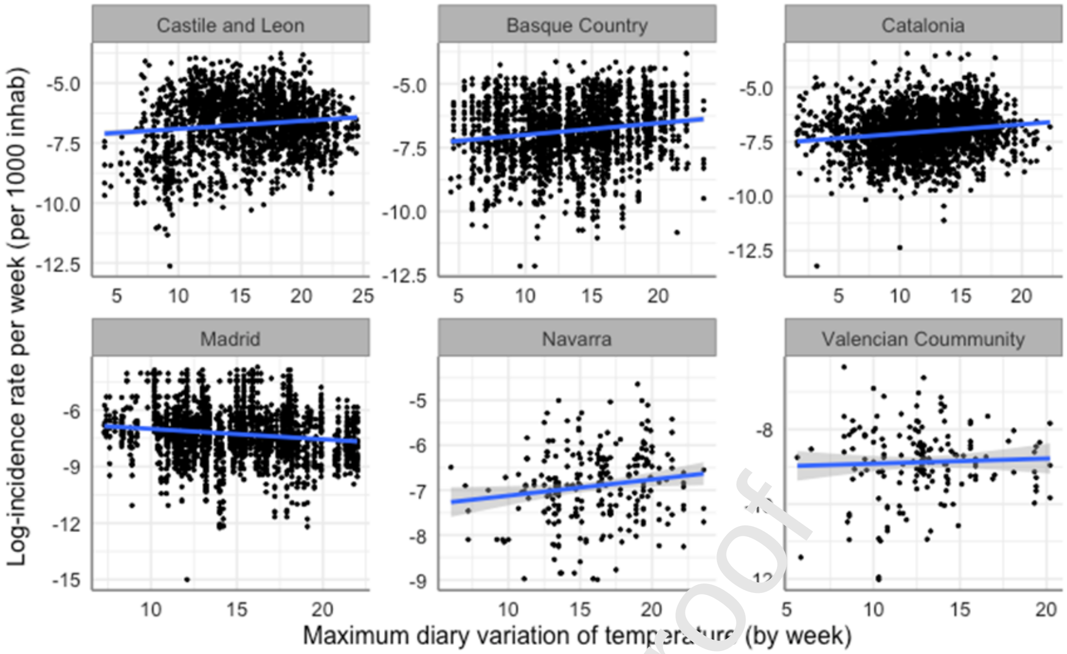

Figure 10. Relationship between maximum daily temperature variation per week and COVID-19 incidence (on a log scale).

Given Spain's configuration, the political aggregation itself, i.e. by province or by Autonomous Community (AC) clearly delimits the climatic and even geographical nature of the areas included within them regarding basic conditions, such as continentality and altitude. The AC or province is therefore understood to have a genuine climatic-geographic effect in the prediction model, and to better explain the COVID data variability. The temporal (weekly) component, accounting for variation over time, was included as a factor effect.

Finally, the proposed models predicted the weekly incidence rates for each district (on the log scale), in the following way:

- **Model 1**: the minimum weekly temperature registered by district, a geographic effect given by the AC and possibly affecting temperature and a nested effect of time (week) given the corresponding AC. The model can be written as:

$$\log(rate_{week}) \sim \alpha \cdot tmin_{week} + \gamma_{AC} + \alpha_{AC} \cdot tmin_{week} + \delta_{week}^{AC}$$

  Where $rate_{week}$represents the weekly incidence rate of contagion per 1,000 inhabitants, increased by one unit, $tmin_{week}$ is the minimum temperature recorded in each district over each week and that acts as a regressor, with possible variations in each AC (α_{AC}), γ_{AC}, is a geographic effect of each AC, and finally δ_{week}^{AC} increases the temporal effect of each week elapsed from the beginning of the records, to the contagion incidence, taking into account that not all ACs had records available for all the weeks under study.

- **Model 2** takes into account the maximum weekly thermal oscillation recorded in each district, a geographic effect that takes into account the province and altitude of each district, as well as a temporal effect from the beginning of the records:

$$\log(rate_{week}) \sim \alpha \cdot dtem_{week} + \beta \cdot alt + \gamma_{prov} + \ \alpha_{prov} \cdot dtem_{week} + \beta_{prov} \cdot alt + \delta_{week}^{prov}$$

where $rate_{week}$ represents the weekly incidence rate of contagion per 1,000 inhabitants, increased by one unit, $dtem_{week}$is the maximum thermal oscillation (minimum temperature - maximum temperature) recorded in each district over each week and acting as a regressor, with possible variations in each province, (α_{prov}), alt is the altitude of each district that can affect the contagion rate differently in each province, (β_{prov}), γ_{prov} is a geographic effect of each province, and finally δ_{week}^{prov} quantifies the temporal effect of each week elapsed from the beginning of the records, to the contagion incidence, taking into account that not all ACs had records available for all the weeks under study.

The goodness of fit statistic gives an adjusted R-squared of 0.38 for Model 1 and 0.34 for Model 2. The high variability of the available data on districts, weeks and autonomous communities implies that these values are reasonable, as simpler models rather than more sophisticated ones were sought for this study. Normal linear models for the weekly incidence rate log proved to be simple but satisfactory.

4.3.1 Discussion from Model 1

In Model 1, the higher the minimum temperature registered during a week, the lower the incidence of the pandemic (see Figure 11, with the fitted trend according to minimum temperature). This effect was different depending on the AC, as shown by the significant interaction. With respect to the intrinsic variability due to the region effect (AC), Catalonia stands out for the very high variability on its fitted effect (Figure 12). Navarra and the Valencian Community also presented high variability, probably due to a lack of data for the whole period between weeks 9 and 22. The quality of the data, and therefore that of the fitted effects, were greater in the case of Castile and Leon, the Basque country and Madrid, whose incidence rates were satisfactory based on Model 1.

Model 1 gives an efficient estimate of the pandemic's evolution in mean terms, over the available weeks (see Figure 12), the peak of the pandemic having been reached in Spain in week 13. Infections began to decline thereafter, with a slight uptick in week 19.

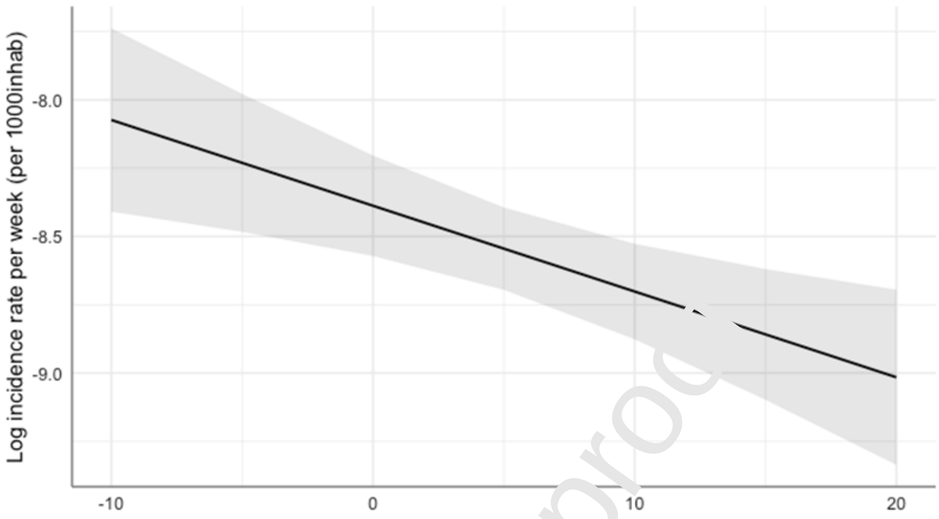

Figure 11. Prediction of the incidence rate (on a log scale) according to minimum temperature, from Model 1.

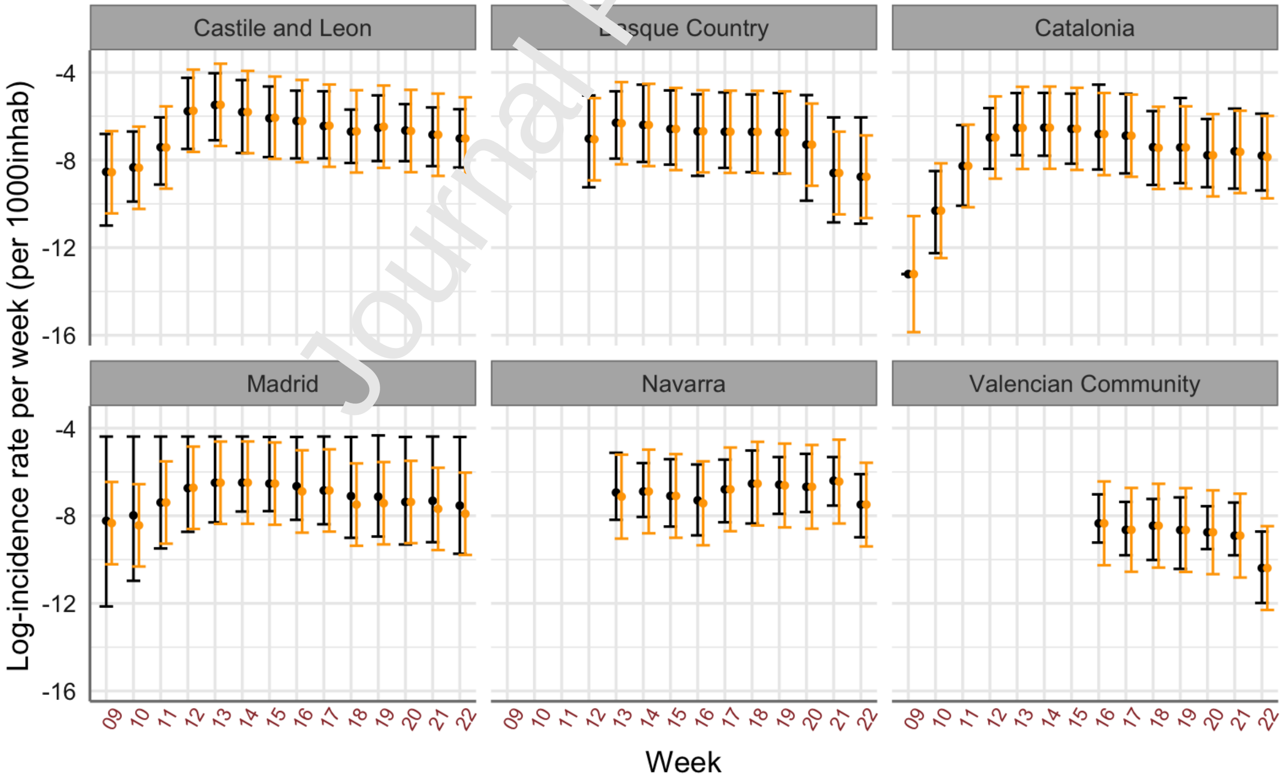

Figure 12. Prediction of incidence rate (on a log scale) for each week from Model 1. Black points represent observed means and black lines represent the interquantile range for a 95% probability interval. Orange ones represent, respectively, predicted values and 95% confidence intervals.

4.3.2 Discusssion from Model 2

In Model 2, the higher the maximum daily difference in temperature registered per week, the higher the incidence of COVID-19 (see Figure 13, with the fitted trend for daily difference). This effect varied depending on the province, as shown by the significant interaction. Altitude also affected incidence, but differently depending on the province, according to its features and continental nature. In mean terms, the higher the altitude, the higher the incidence (see Figure 14), as coldest districts are related to higher altitudes. Again, the model is able to efficiently estimate the pandemic's evolution in mean terms, over the available weeks (see Figure 15), the peak of the pandemic having been reached in Spain in week 13. Infections began to decline thereafter, with a slight uptick in week 19. The effect of the province on the incidence is not shown as there are too many estimated effects.

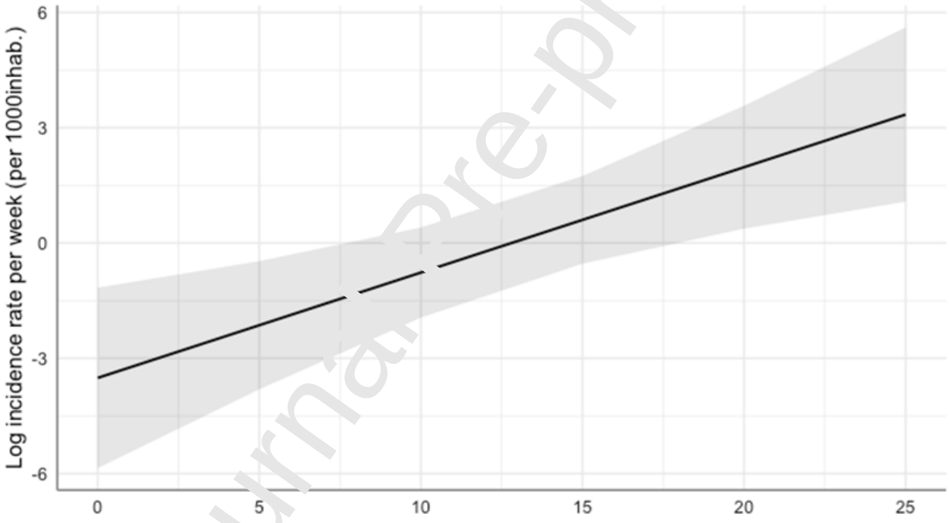

Figure 13. Prediction of the incidence rate (on a log scale) according to maximum daily temperature variation from Model 2.

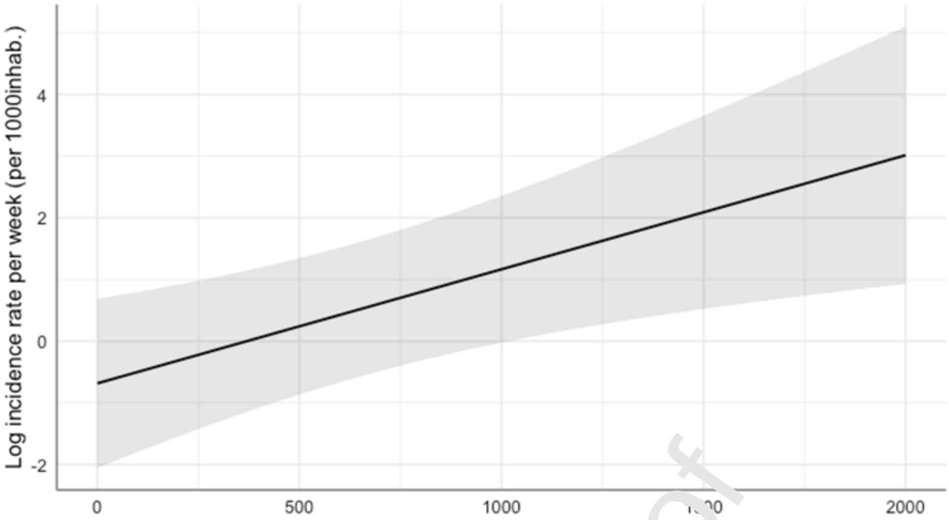

Figure 14. Prediction of the incidence rate (on a log scale) according to altitude, from Model 2.

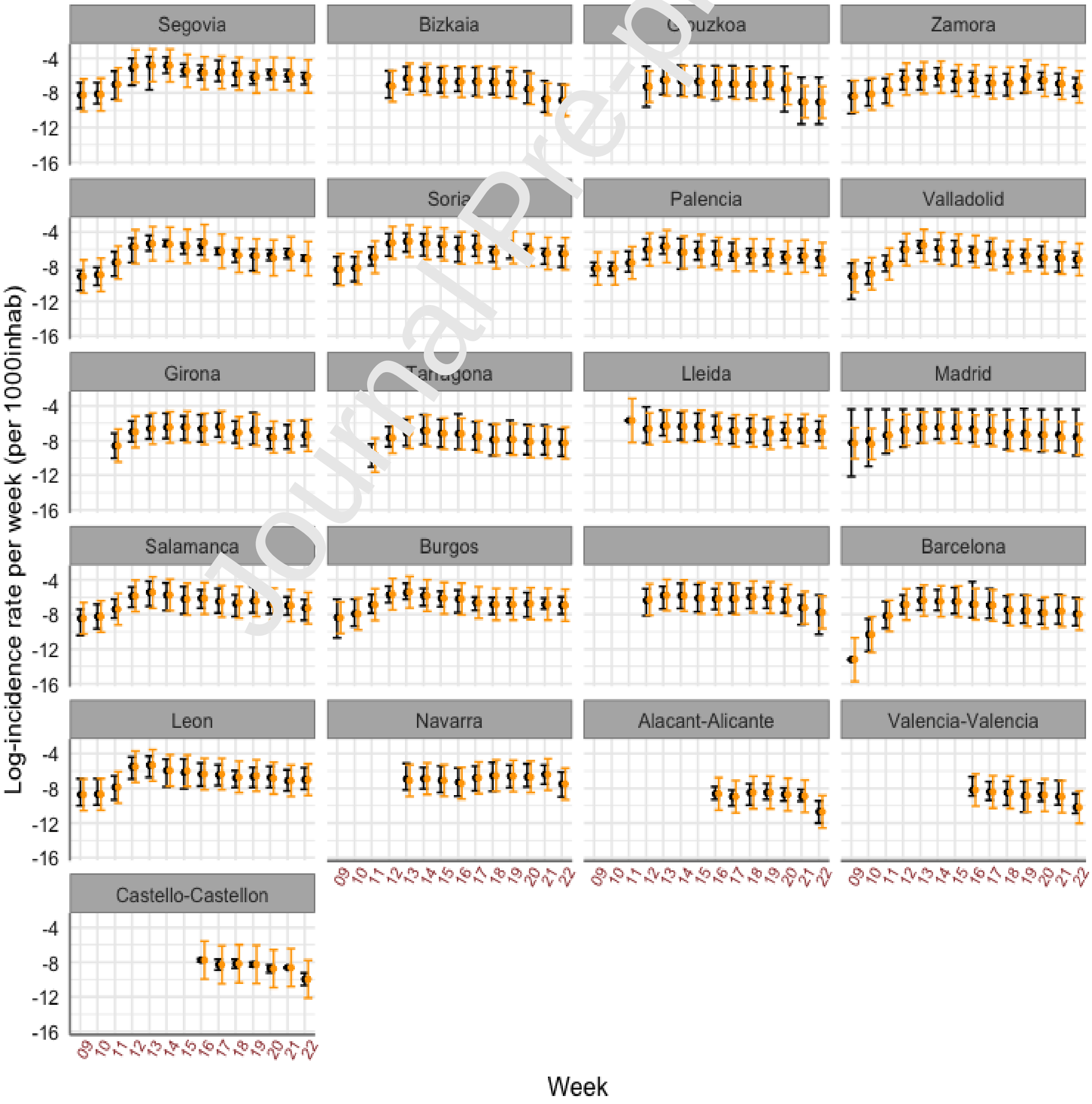

Figure 15. Prediction of the incidence rate (on a log scale) per week from Model 2. Black points represent observed means and black lines represent the interquantile range for a 95% probability interval. Orange ones represent, respectively, predicted values and 95% confidence intervals.

5 Conclusions

The present study addressed the usefulness of open government data to determine the impact of the SARS-CoV-2 virus and the COVID-19 disease based on the atmospheric conditions according to different space-time scales during Spain's 2020 lockdown.

The conclusions can be summarised as follow:

- A relationship between temperature and COVID-19 contagion was observed. There was a slight trend of lower incidence with minimum temperature increases i.e. the lower the minimum temperature, the higher the number of cases.
- A relationship was also found between altitude and spread of the disease. Higher contagion rates were recorded in areas with a higher average altitude (the central plateau).
- In coastal areas, with higher minimum temperatures, the pandemic's impact was lower.
- No relationship with precipitation or wind was encountered based on the study data.

Importantly, our work was based entirely on open data sources available on government portals, which highlights the key role of good data management following the FAIR (Findable, Accessible, Interoperable and Reusable data) principles to enable open government data to be useful in conducting scientific studies of interest to society that are also fully reproducible (Wilkinson et al., 2016).

Worthy of note, the conditions under which the COVID-19 incidence data were measured allowed comparing the different areas under lockdown conditions, because all areas underwent the same mobility restrictions, as established by the regulations in force (Spain's 2020 state of alert). Mobility was reduced to only 10% of the normal rate (according to the Apple report). These are optimal conditions for a study of this kind, focused on the analysis of the effect of atmospheric variables on the incidence of COVID-19.

Furthermore, we wish to stress that this study was possible thanks to the availability of open data on COVID-19 cases, which were updated daily at the basic level of health districts or municipalities. We must specify, however, that only 6 out of Spain's 17 autonomous communities offered this information in their open data portals. Furthermore, not even in those ACs was the data entirely available (for example in Navarra and in the Valencian Community). The rest of the ACs provided case data aggregated by autonomous community or province, or a web application to view the data, but did not enable access (for example, via download or an API), thereby preventing its reuse for other types of analysis. The availability of open data of a geographic and meteorological nature is also worthy of note.

The space-time scale addressed in the present work makes this study an unprecedented one in Spain and it has allowed to highlight the ineffectiveness of current Spanish information systems. There is a critical need to design efficient, uniform and coordinated systems for the collection and publication of open data. With centralised and standardised open information, society, as a whole, would benefit from the contributions of the scientific community on critical events such as the current pandemic.

Finally, contribution of our study, as well as its limitations and future research directions, are summarized. Contribution of our paper is twofold: (i) development of models to determine incidence and evolution of COVID-19 based on atmospheric and geographic factors, and (ii) using models to forecast whether open data would be useful for mitigating crisis scenarios, such as COVID-19 pandemic. Limitations of our study rely on availability of open data sources and government commitment with open data publication. As a matter of fact, only 6 regional governments in Spain offered required open data. Also, other important pitfall we found is the lack of homogeneity among open data sources (each government published open data related to COVID-19 pandemic with different levels of granularity, different formats, etc.). As future work we plan to focus on other approachable relationships, using data on disease incidence from other sources of open data (e.g. demographic or sociological data), as well as extrapolating our approach to other crisis situations (such as natural disasters), thus studying thoroughly the supportive role of open data, so that governments around the world are motivated to make their data open.

References

Akaike, H. (1974), A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (6): 716-723, MR 0423716, doi:10.1109/TAC.1974.1100705.

Apple (2020). Informe de tendencias de movilidad por países (escala diaria*)*. https://www.apple.com/covid19/mobility

Araújo, M. B., & Naimi, B. (2020). Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. *MedRxiv*. https://doi.org/10.1101/2020.03.12.20034728

Baker, R. E., Yang, W., Vecchi, G. A., Metcalf, C. J. E. & Grenfell, B. T. Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic. Science. https://doi.org/10.1126/science.abc2535 (2020).

Brassey, J., Heneghan, C., Mahtani, K. R., & Aronson, J. K. (2020). COVID-19: Do weather conditions influuence the transmission of the coronavirus (SARS-CoV-2)? Oxford COVID-19 Evidence Service. https://www.cebm.net/do-weather-conditions-influence-the-transmission-of-the- coronavirus-sars-cov-2/

Briz-Redón, A., & Serrano-Aroca, Á. (2020). A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. Science of The Total Environment, 728, 138811. https://doi.org/10.1016/j.scitotenv.2020.138811

Bu, J., Peng, D.-D., Xiao, H., Yue, Q., Han, Y., Lin, Y., ... & Chen, J. (2020). Analysis of meteorological conditions and prediction of epidemic trend of 2019-nCoV infection in 2020. MedRxiv. https://doi.org/https://doi.org/10.1101/2020.02.13.20022715

Bukhari, Q. & Jameel, Y. (2020). Will Coronavirus Pandemic Diminish by Summer?. *SSRN*. http://dx.doi.org/10.2139/ssrn.3556998

Carlson, C.J., Gomez, A.C.R., Bansal, S. et al. Misconceptions about weather and seasonality must not misguide COVID-19 response. Nat Commun 11, 4312 (2020). https://doi.org/10.1038/s41467-020-18150-z

Chatterjee, S., Hadi, A. S. and Price, B. (2000). Regression Analysis by Example. 3rd. John Wiley & Sons, New York.

Chin, A., Chu, J., Perera, M., Hui, K., Yen, H., Chan, M., & Poon, L. (2020). Stability of SARS-CoV-2 in different environmental conditions. MedRxiv. https://doi.org/10.1101/2020.03.15.20036673

Courtemanche, C., Garuccio,J., Le, A., Pinkston, J., Yelowitz, A. (2020) Strong Social Distancing Measures in the United States Reduced the COVID-19 Growth Rate: study evaluates the impact of social distancing measures on the growth rate of confirmed COVID-19 cases across the United States. Health Aff., 39

Dezan, S. (2020) “China Coronavirus Updates: latest developments and business advisory” China Briefing. Available on: https://www.china-briefing.com/news/china-coronavirus-updates-latest-developments-business-advisory-part-2/

Dbouk, T. and Drikakis, D. (2020) Weather impact on airborne coronavirus survival, Physics of Fluids. DOI: 10.1063/5.0024272

Dubey, R., Gunasekaran, A., Childe, S. J., Papadopoulos, T., Luo, Z., Wamba, S. F., & Roubaud, D. (2019). Can big data and predictive analytics improve social and environmental sustainability?. Technological Forecasting and Social Change, 144, 534-545.

Ficetola, G.F. & Rubolini, D. (2020). Climate affects global patterns of COVID-19 early outbreak dynamics. MedRxiv. https://doi.org/10.1101/2020.03.23.20040501

Forster, P.M:; Forster, H.I.; Evans, M.J.; Gidden, M.J.; Jones, Ch. D.; Keller, Ch.A.; Lamboll, R.D.; Le Quéré, C.; Rogelj, J.; Rosen, D.; Schleussner, C-F.; Richardson, T.B.; Smith Ch.J. & Turnock, S.T. 2020. “Current and future global climate impacts resulting from COVID-19”. Nature Climate Change (07 August 2020). https://www.nature.com/articles/s41558-020-0883-0

Gallagher, J., Orcutt, J., Simpson, P. et al. (2015). Facilitating open exchange of data and information. Earth Sci Inform 8, 721–739. https://doi.org/10.1007/s12145-014-0202-2

Gutierrez-Corea, FV., Manso-Callejo, MA. & Vázquez-Hoehne, A. (2013) Assessment of the availability of near-real time open weather data provided by networks of surface stations in Spain. Earth Sci Inform 6, 145–163. https://doi.org/10.1007/s12145-013-0120-8

Gutiérrez-Hernández, O., & García, L.V. (2020). ¿Influyen tiempo y clima en la distribución del nuevo coronavirus (SARS CoV-2)? Una revisión desde una perspectiva biogeográfica. *Investigaciones Geográficas,* (73), 31-55. https://doi.org/10.14198/INGEO2020.GHVG

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. Applied Statistics, 28, 100–108. doi: 10.2307/2346830.

Holtmann, M., Jones, M., Shah, A., & Holtmann, G. (2020). Low ambient temperatures are associated with more rapid spread of COVID-19 in the early phase of the endemic. Environmental Research, 109625. https://doi.org/10.1016/j.envres.2020.109625

Ianevski, A.; Zusinaite, E.; Shtaida, N.; Kallio-Kokko, H.; Valkonen, M.; Kantele, A.; Telling, K.; Lutsar, I.; Letjuka, P.; Metelitsa, N.; Oksenych, V.; Dumpis, U.; Vitkauskiene, A.; Stašaitis, K.; Öhrmalm, C.; Bondeson, K.; Bergqvist, A.; Cox, R.J.; Tenson, T.; Merits, A.; Kainov, D.E. Low Temperature and Low UV Indexes Correlated with Peaks of Influenza Virus Activity in Northern Europe during 2010–2018. Viruses 2019, 11, 207.

ISCIII, & AEMET (2020). Primeros indicios de correlación entre variables meteorológicas y propagación del coronavirus y la COVID-19 en España. Recuperado de: https://www.isciii.es/Noticias/Noticias/ Paginas/Noticias/AcuerdoISCIIIAEMETEstudioTemperaturasCOVID19.aspx

ISS, Instito Superiore di Saniti (2020) COVID-19 integrated surveillance data in Italy. Avaliable on: https://www.epicentro.iss.it/en/coronavirus/sars-cov-2-dashboard

Jones, N.R., Quresshi, Z.U., Temple, R.J., Larwood, J.P.J., Greenhalgh, T., Bourouiba,L. (2020) Two metres or one: what is the evidence for physical distancing in covid-19?. BJM, 370. doi: https://doi.org/10.1136/bmj.m3223

Liu, S.; Ermolieva, T.; Cao, G.; Chen, G.; Zheng, X. (2021) Analyzing the Effectiveness of COVID-19 Lockdown Policies Using the Time-Dependent Reproduction Number and the Regression Discontinuity Framework: Comparison between Countries. *Eng. Proc*. 5, 8. https://doi.org/10.3390/engproc2021005008

Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., ... & Luo, B. (2020). Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Science of the Total Environment, 724, 138226. https://doi.org/10.1016/j.scitotenv.2020.138226

Mecenas, P., Bastos, RTdRM., Vallinoto, ACR., Normando, D. (2020) Effects of temperature and humidity on the spread of COVID-19: A systematic review. PLoS ONE 15(9): e0238339. https://doi.org/10.1371/journal.pone.0238339

Miranda, L. C. M., & Devezas, T. (2022). On the global time evolution of the Covid-19 pandemic: Logistic modeling. Technological Forecasting and Social Change, 175, 121387.

Morawska, L., Milton, D.K. (2020) “It is Time to Address Airborne Transmission of COVID-19”, Clinical Infectious Diseases, ciaa939, https://doi.org/10.1093/cid/ciaa939 .

Noland, R.B. (2021) Mobility and the effective reproduction rate of COVID-19. Journal of Transport & Health, Volume 20, ISSN 2214-1405. DOI: 10.1016/j.jth.2021.101016.

Olcina Cantos, J., Biener Camacho, S. y Martí Talavera, J. (2020). Aspectos atmosféricos y climáticos en la expansión de la pandemia (COVID-19) en la provincia de Alicante. *Investigaciones Geográficas,* (73), 275-297. https://doi.org/10.14198/INGEO2020.OCBCMT

Olsen SJ, Azziz-Baumgartner E, Budd AP, et al. (2020). “Decreased Influenza Activity During the COVID-19 Pandemic — United States, Australia, Chile, and South Africa, 2020”. MMWR Morb Mortal Wkly Rep 69:1305–1309. DOI: http://dx.doi.org/10.15585/mmwr.mm6937a6external

Oto-Peralías, D. (2020). Regional correlations of COVID-19 in Spain. OSF Preprints. https://doi.org/ https://doi.org/10.31219/osf.io/tjdgw

Pan, A., Liu, L., Wang, C., Guo, H., Hao, X., Wang, Q., Huang, X., He, N. Yu, X. Lin, X. (2020) Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China. JAMA, 323, pp. 1915-1923

Pincombe, M.; Reese, V.; Dolan, C.B. (2021) The effectiveness of national-level containment and closure policies across income levels during the COVID-19 pandemic: An analysis of 113 countries. *Health Policy and Planning*, Volume 36, Issue 7, August, 1152–1162,

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Ramadhan, T., Joko, G., Devi, E. La, A., Hariati, E., Hartati, B., Pitrah, A. (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. Science of The Total Environment. 725. 138436. 10.1016/j.scitotenv.2020.138436.

Rohde, R. (2020) The relationship between coronavirus (COVID-19) spread and the weather, Berkeley Earth, august. Available on: http://berkeleyearth.org/coronavirus-and-the-weather-new/

Rosario, D.K.A, Mutz, Y.S., Bernardes, P.D., Conge-Junios, C.a. (2020) “Relationship between COVID-19 and weather: case stuty in a tropical country”. International Journal of Hygiene and Environmental Health, vol. 229, august, 113587.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. URL: http://www.rstudio.com/

Saez, M., Tobías, A., Varga, D., Barceló, M. Effectiveness of the measures to flatten the epidemic curve of COVID-19. The case of Spain. Sci. Total Environ., 727 (2020), Article 138761, 10.1016/j.scitotenv.2020.138761

Sahoo B.K., Sapra B.K. (2020) A data driven epidemic model to analyse the lockdown effect and predict the course of COVID-19 progress in India, Chaos Solitons Fractals. DOI: 10.1016/j.chaos.2020.110034

Sajadi, M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., & Amoroso, A. (2020). Temperature, Humidity and Latitude Analysis to Predict Potential Spread and Seasonality for COVID-19. *SSRN*. http://dx.doi.org/10.2139/ssrn.3550308

Shi, P., Dong, Y., Yan, H., Li, X., Zhao, C., Liu, W., Xi, S. (2020). The impact of temperature and absolute humidity on the coronavirus disease 2019 (COVID-19) outbreak - evidence from China. MedRxiv. https://doi.org/10.1101/2020.03.22.20038919

Simionescu, M., & Raišienė, A. G. (2021). A bridge between sentiment indicators: What does Google Trends tell us about COVID-19 pandemic and employment expectations in the EU new member states?. Technological Forecasting and Social Change, 173, 121170.

Sun, Z.; Zhang, H.; Yang, Y.; Wan, H.; Wang, Y. (2020) Impacts of geographic factors and population density on the COVID-19 spreading under the lockdown policies of China, *Science of The Total Environment*, Volume 746, 141347.

Takagi, H., Kuno, T., Yokoyama, Y., Ueyama, H. Matsushiro, T., Hari, Y., Ando, T. The higher temperature and ultraviolet, the lower COVID-19 prevalence–meta-regression of data from large US cities. American Journal of Infection Control (2020). ISSN 0196-6553. https://doi.org/10.1016/j.ajic.2020.06.181.

Wang, J., Tang, K., Feng, K., & Lv, W. (2020). High Temperature and High Humidity Reduce the Transmission of COVID-19. *SSRN*. http://dx.doi.org/10.2139/ssrn.3551767

Ward, M., Xiao, S., and Zhang, Z. (2020). Humidity is a consistent climatic factor contributing to SARS-CoV-2 transmission. Transboundary and Emerging Diseases. https://onlinelibrary.wiley.com/doi/abs/10.1111/tbed.13766

WHO (2020) Transmission of SARS-CoV-2: implications for infection prevention precautions. Scientific Brief. 9 July 2020. Avaliable on: WHO/2019-nCoV/Sci_Brief/Transmission_modes/2020.3

Wilkinson, M. et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

WMO (2020) Q&A: Climate change and COVID-19. Available on: https://www.who.int/westernpacific/news/q-a-detail/q-a-on-climate-change-and-covid-19 . Consulted: august 2020.

Yao, Y., Pan, J., Liu, Z., Meng, X., Wang, W., Kan, H., Wang, W. No association of COVID-19 transmission with temperature or UV radiation in Chinese cities. Eur Respir J, 55 (2020), Article 2000517

Xu, R., Rahmandad, H., Gupta, M., Digennaro, C., Ghaffarzadegan, N., Amini, H., Jalali, M.S. (2020). The Modest Impact of Weather and Air Pollution on COVID-19 Transmission. SSRN Electronic Journal. 10.2139/ssrn.3593879.

Highlights

- Open data is used to model the incidence of weather conditions on COVID-19
- The lower the minimum temperature, the greater the number of cases
- There is a link between COVID-19 incidence, altitude and the proximity to the sea
- No relationship with other climatic or geographical variables has been discovered
- It is shown the utility and importance of open data to conduct scientific research

Author Statement

Jose Jacobo Zubcoff: Conceptualization, Methodology, Formal Analysis, Investigation, Writing - review & editing.
Jorge Olcina: Conceptualization, Methodology, Formal Analysis, Investigation, Writing - review & editing.
Javier Morales: Conceptualization, Methodology, Formal Analysis, Investigation, Writing - review & editing.
Jose-Norberto Mazón: Conceptualization, Methodology, Data Curation, Investigation, Writing - review & editing.
Asunción M. Mayoral: Conceptualization, Methodology, Formal Analysis, Investigation, Writing - review & editing.

Authors BIO

JOSÉ ZUBCOFF is Professor at the University of Alicante since 2000 in applied statistics subjects in Biology and Marine Sciences Degrees, Masters and PhD. He has more than 100 publications including journal articles, books and conference papers. He has supervised 5 doctoral theses and more than 30 master's and bachelor's theses. He has participated in more than 50 projects. His research pursues what he has defined as "democratisation of knowledge" which aims to facilitate access to information and knowledge to anyone.

ASUNCIÓN MARTÍNEZ MAYORAL holds a PhD in Mathematics, specializing in Statistics, from the University of Valencia (Spain). She has been working at the Miguel Hernández University of Elche (Alicante, Spain) since 1997. She is a full professor in the area of Statistics and Operations Research since 2008. Between 2007 and 2020 she has performed several university management tasks related to educational innovation and technology. She has worked with numerous researchers in different areas applying statistics. She is currently collaborating with a cardiology team working on survival.

JAVIER MORALES SOCUÉLLAMOS holds a PhD in Mathematics, specializing in Statistics, from the University of Valencia (Spain). He has been working at the Miguel Hernández University of Elche (Alicante, Spain) since 1997. He has worked with numerous researchers in different areas applying statistics: ecology, agronomy, medicine, climate change. He is currently collaborating with a cardiology team working on Bayesian survival models.

JORGE OLCINA CANTOS is Professor at Alicante University, where teaches on Spatial Planning, Climatology and Natural Hazards. He has focused his research on various geographical issues. Author of more than a hundred publications (scientific papers, book chapters and book editor). He has participated in several research projects on geographical and historical issues developed at the University of Alicante. Main-speaker at the International Year of Planet Earth (2008), declared by UNESCO. Member of Editorial Board of various scientific journals on geographical and environmental issues. President of Spanish Association of Geography (2017-21).

JOSE-NORBERTO MAZÓN is currently an Associate Professor with the Department of Software and Computing Systems, University of Alicante, Spain. He is also member of the University Institute for Computer Research (IUII) of the University of Alicante, as well as the Chair of the Torrevieja Venue of the University of Alicante. He is the author of more than 100 scientific publications in international conferences and journals. His research interests include open data, business intelligence in big data scenario, design of data-intensive web applications, smart cities, and smart tourism destinations.