Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

MASTER THESIS

Area 5: Healthcare and Environment

COVID-19: Outbreak prediction combining meteorological, mobility and demographic data

---

Autor: Jaime Pérez Ordieres

Tutor: Carlos Luis Sánchez Bocanegra

Pola de Siero, June 5, 2022

Copyright

WORK SHEET

Title of the master thesis:	COVID-19: Outbreak prediction combining meteorological, mobility and demographic data
Author's name:	Jaime Pérez Ordieres
Tutor's name:	Carlos Luis Sánchez Bocanegra
Delivery date:	05/06/2022
Degree:	Data Science
Area of work:	Area 5: Healthcare and Environment
Lenguage:	English
Keywords	COVID-19, Time series forecasting, Meteo, Mobility

Dedication/Citation

To all my family, and especially to my wife Olaya, without whom all this would not have been possible.

Acknowledgments

To my tutor Carlos and to the whole university.

Abstract

Two years after the start of the pandemic caused by the COVID-19 virus, the Spanish health system has been on the verge of collapse on several occasions, forcing an adaptation of the system and professionals and highlighting some of the structural organizational shortcomings.

From the scientific and educational fields, the need arises to alleviate these deficiencies through innovation.

Unlike to what could happen in the past, a vast amount of data and information is currently available. Given such an amount of data, and in order to alleviate the effects of the pandemic on society, it is vital to identify relevant factors that help to identify situations of high spread of the virus in advance.

The present work seeks to understand if the meteorological, mobility and demographic factors are relevant in the spread of the virus. To do this, public data combined with machine learning techniques applied to the prediction of time series will be used.

The ultimate goal will be to provide tools that make it possible to predict coronavirus outbreaks, thus being able to optimize the available health resources.

Keywords: COVID-19, Time series forecasting, Meteo, Mobility

Resumen

Tras dos años desde el inicio de la pandemia provocada por el virus COVID-19, el sistema sanitario español ha estado al borde del colapso en varias ocasiones, forzando una adaptación del sistema y de los profesionales y poniendo de manifiesto algunas de las carencias organizativas estructurales y profesionales del mismo.

Desde el ámbito científico y docente, surge la necesidad de paliar dichas carencias mediante la innovación.

Al contrario de lo que podía ocurrir en el pasado, actualmente se dispone de una cantidad inmesa de datos y de información. Ante tal cantidad de datos, y de cara a paliar los efectos de la pandemia en la sociedad, resulta vital la identificación de factores relevantes que ayuden a identificar con antelación situaciones de alta propagación del virus.

El presente trabajo busca comprender si los factores meteorológicos, de movilidad y demográficos resultan relevantes en la propagación del virus. Para ello, se utilizarán datos públicos combinados con técnicas de machines learning aplicadas a la predicción de series temporales.

El objetivo último será dotar de herramientas que posibiliten la predicción de brotes de coronavirus, pudiendo de esta forma optimizar los recusos sanitarios disponibles.

Keywords: COVID-19, Predicción series temporales, Meteo, Movilidad

Contents

List of Figures

List of Tables

Chapter 1

Definition and planning of the work

1.1 Description, interest and relevance of the proposal

The Spanish National Microbiology Center (CNME) registered the first case of SARS-CoV-2 (COVID-19) in January 2020 at the Virgen de Guadalupe Hospital in La Gomera [1]. Since then thousands of infected and deceased have come to collapse the national health system.

Two years after the start of the pandemic, Spain is facing the sixth wave with a cumulative total of 11 million infected and almost 100,000 deaths [2]. The high vaccination rate has cushioned the impact of the new variants of COVID-19. However, the relative ignorance of the virus, as well as the inability of administrations to predict its outbreaks, continue to put the health system at risk.

Transmission of the virus occurs primarily through exhalation of very small respiratory droplets and particles that contain the virus even when the infected person has no symptoms. According to the World Health Organization (WHO), infected people are apparently most contagious just before symptoms appear (about two days before) and in the first phase of the disease [3]. Said respiratory particles can be inhaled by people and/or deposit on their eyes, nose or mouth.

The current measures that governments are taking to mitigate the socioeconomic impact of COVID-19 and support the economic recovery of the countries seem to have a tendency towards coexistence with the virus. This strategy is mainly supported by the high vaccination rate that minimizes the potential health effects of the virus. Given this scenario, obtaining a prediction model that includes factors considered key in transmission will be vital for proper optimization of the health system's resources.

[1] https://gacetamedica.com/investigacion/espana-confirma-su-primer-caso-de-coronavirus/

[2] https://www.rtve.es/noticias/20220225/mapa-del-coronavirus-espana/2004681.shtml

[3] https://www.who.int/es/news-room/questions-and-answers/item/coronavirus-disease-covid-19-how-is-it-transmitted

The main objective of this work will be the use of machine learning techniques applied to time series for the prediction of COVID-19 outbreaks. For this, meteorological data and mobility data will be used, detailing demographic factors as far as possible.

1.2 Objectives

The main objectives are:

- Identify whether meteorological and mobility factors are important in predicting COVID-19 outbreaks
- Use of machine learning and time series prediction techniques for modeling the evolution of the virus

For which, it will be necessary to achieve a series of secondary objectives:

- Identification of relevant data and study of the possibilities they offer
- Understand the behavior of the virus by analyzing the medical literature
- Identification of machine learning and prediction techniques that best suit the purpose of our project

1.3 Methodology to be used

During the course of the project, a quantitative methodology will be adopted that allows us to analyze the data numerically with the aim of generalizing and objectifying the results from the available data [6].

A search of the medical literature will be carried out to understand the propagation behavior of COVID-19. This behavior will be combined with:

- COVID-19 data obtained from the National Epidemiological Surveillance Network (RENAVE) through the SiViES (Surveillance System of Spain) computer platform managed by the National Epidemiology Center (CNE)
- Meteorological data provided by the AEMET and by global climatological models
- Mobility data from the study carried out by the Ministry of Transport, Mobility and Urban Agenda (MITMA) during the COVID-19 pandemic, as well as data from large technology companies (Google, Apple, ...)

- Demographic data from the National Institute of Statistics.

Furthermore, it will be necessary to carry out an exploratory analysis of the state of the art of machine learning techniques applied to time series, as well as the toolkit and programming languages that best adapt to them.

The progress of the project will be controlled by the Kanban flow method[4] where work flows continuously through 'To-do', 'Doing' and 'Done' cards reviewed weekly.

1.4 Planning

PEC 1 - Definition and planning	**Planning**		
	Duration	**Start**	**Finish**
Topic selection and initial research	4 days	16/02/2022	19/02/2022
Description, interest and relevance of the proposal	3 days	19/02/2022	21/02/2022
Objectives and personal motivation	3 days	22/02/2022	24/02/2022
Methodology to be used and planning	2 days	24/02/2022	25/02/2022
Abstract preparation + PEC 1 delivery	2 days	26/02/2022	27/02/2022

Table 1.1: PEC-1: Planning

PEC 2 - State of the art / Market analysis	**Planning**		
	Duration	**Start**	**Finish**
Search bibliography on COVID-19 transmission	3 days	28/02/2022	02/03/2022
Study previous works on the field	3 days	03/03/2022	05/03/2022
Identify techniques for data cleaning/transformation	2 days	06/03/2022	07/03/2022
Identify machine learning techniques	3 days	08/03/2022	10/03/2022
Refine and adapt PEC 1 achievements	2 days	11/03/2022	12/03/2022
PEC 2 delivery	1 days	13/03/2022	13/03/2022

Table 1.2: PEC-2: Planning

[4]https://todoist.com/productivity-methods/kanban

PEC 3 - Work design and implementation	**Planning**		
	Duration	**Start**	**Finish**
Toolkit and programming language (R / Python)	5 days	14/03/2022	18/03/2022
Extract, transform and load data	25 days	19/03/2022	12/04/2022
Models review	5 days	13/04/2022	17/04/2022
Models selection	3 days	18/04/2022	05/05/2022
Prediction analysis	2 days	06/05/2022	14/05/2022
PEC 3 delivery	1 days	15/05/2022	15/05/2022

Table 1.3: PEC-3: Planning

PEC 4 - Official report writing	**Planning**		
	Duration	**Start**	**Finish**
Review documentation generated	2 days	16/05/2022	17/05/2022
Conclusions from of the results	3 days	18/05/2022	20/05/2022
Write master thesis (official format)	9 days	21/05/2022	29/05/2022
Adaptation of content according to recommendations	6 days	30/05/2022	05/06/2022

Table 1.4: PEC-4: Planning

PEC 5 - Public defense	**Planning**		
	Duration	**Start**	**Finish**
Preparation of the defense	7 days	06/06/2022	12/06/2022
Public defense	12 days	13/06/2022	24/06/2022

Table 1.5: PEC-5: Public defense

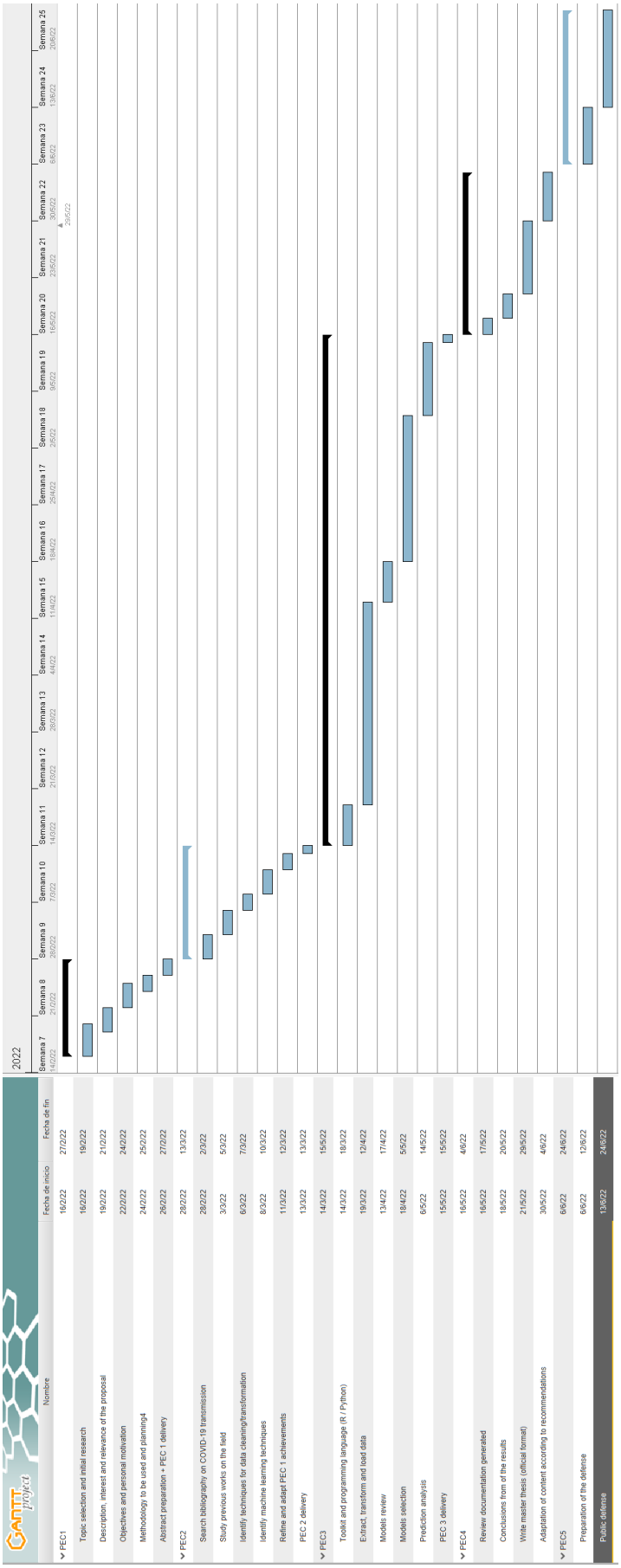

Figure 1.1: Gantt project diagram

Chapter 2

State of the art

2.1 Brief introduction to coronaviruses

2.1.1 History

Coronaviruses are a family of enveloped, single-stranded-RNA viruses that can affect both humans and animals.

The first known pandemic caused by this kind of pathogens was at the beginning of the 21st century. On that occasion, an atypical pneumonia, later named Severe Acute Respiratory Syndrome (SARS-CoV-1 or SARS), emerged in Foshan, Guangdong Province, mainland China, in November 2002 [19] [23].

Its infectiousness, with some 8000 patients and a total of 774 deaths in 26 different countries, opened a source of debate about the need to coordinate a global response to contain such threats [19].

Later in 2012, a respiratory pathology caused by another coronavirus (Middle East respiratory syndrome coronavirus, or MERS-CoV) was detected for the first time in Saudi Arabia. According to the World Health Organization (WHO), the mortality rate during that outbreak was approximately 35% of reported patients.[1]

Seventeen years after the outbreak of SARS and seven years since the first case of MERS, the World Health Organization declared a global pandemic due to the spread of COVID-19 (also known as SARS-CoV-2). Globally, as of March 11, 2022, 452,052,304 confirmed cases of COVID-19, including 6,027,059 deaths, have been reported to WHO.[2]

[1] https://www.who.int/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov)

[2] https://covid19.who.int/

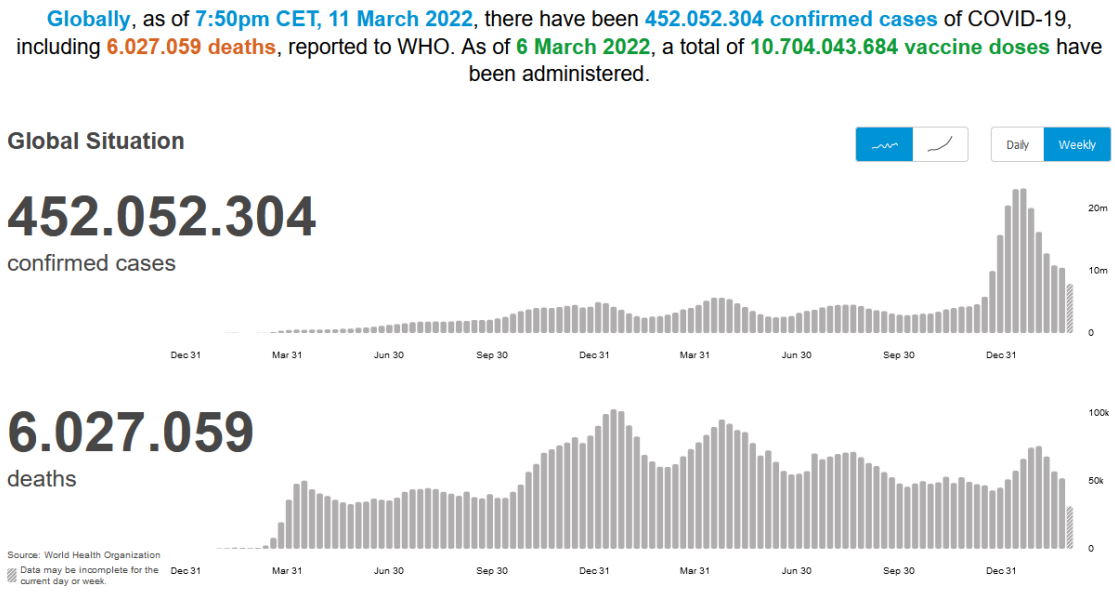

Figure 2.1: Global covid-19 situation March, 2022

In all cases, there are trends that identify its origin in animals. However, the highest rate of transmission was identified among humans.

2.1.2 Symptoms

The initial manifestations of coronaviruses are not specific, and cannot be clinically differentiated from other acute community-acquired pneumonias [2]. Typical symptoms are fever, cough and respiratory distress. Pneumonia is also frequent, but not always present [22]. But it is not only limited to these symptoms. Muscle or body aches, headache, sore throat and gastrointestinal symptoms, including diarrhea, have also been reported.[3]

The infection is not only manifested in the respiratory tract, but is also present in respiratory secretions, feces, urine, and tissue specimens from lung biopsy [19].

SARS and MERS were more aggressive and lethal than COVID-19. However, the latter spreads more rapidly, sometimes with hidden symptoms, allowing each infected person to infect several others.

2.1.3 Vaccines

Studies of SARS-CoV-1 vaccines were initiated and tested in animal models. Inactivated whole virus was used in ferrets, non-human primates and mice. All vaccines resulted in protective immunity, but there were complications; the vaccines resulted in immune disease in the animals.

[3]https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

No human studies were conducted, nor were the vaccine studies taken further because the virus disappeared. [4] [5] [17]

Considering the rapid spread of COVID-19 around the world, the academia, industry and government collaborated closely to develop and test various vaccines at an unprecedented pace [14]. As a result, in December 2020 the first vaccine has been available under EUA licensure for humans. As of March 6, 2022, a total of 10,704,043,684 doses of vaccines have been administered. [6]

Figure 2.2: Spanish vaccination roadmap [https://www.vacunacovid.gob.es/]

2.2 Spread based on mobility

While prior health crises, such as SARS, impacted in the mobility, the COVID-19 pandemic is unprecedented, resulting in exceptional impacts on the mobility trends and transportation sector [1].

The COVID-19 pandemic has mobilised science communities across all over the world sparking a great deal of data exchange and collaboration.

One of the strategies implemented by governments that has had the best results in curbing the virus has been mobility restrictions. The main idea behind them is that high levels of

[4] https://theconversation.com/the-mysterious-disappearance-of-the-first-sars-virus-and-why-we-need-a-vaccine-for-the-current-one-but-didnt-for-the-other-137583
[5] https://pubmed.ncbi.nlm.nih.gov/27076136/
[6] https://covid19.who.int/

mobility contribute to the spread of the virus. Based on this factor, many mobility-related studies have been published over the past few months.

Kraemer et al [12], in their localized study in the provinces of origin of the virus, concluded that travel restrictions were particularly useful in the early stage of an outbreak when it is confined to a certain area that acts as major source. Carteni et al [5] estimated that showed that during the first outbreak on Italy mobility habits represent the variable that mainly explains (from a statistical perspective) the number of COVID-19 infections. Furthermore, research results showed that the number of new COVID-19 cases in one day was directly related to the trips performed three weeks (21 days) before.

On the other hand, Nikos Askitas et al [3], studied how the 'lockdown policies' affects the daily incidence of COVID-19 and mobility patterns finding that cancelling public events and enforcing restrictions on gatherings, which restrict mobility in numerous and dense locations, have the largest effect on curbing the pandemic in terms of statistical significance and levels of effect. Interestingly, they also pointed out that restrictions on internal movement, public transport closures and international travel controls do not lead to a significant reduction of new infections.

More recently, Badr et al [4] examined mobility changes in 25 US counties and found evidence that reductions in mobility reduced growth in cases. Ilin et al [11] found evidences that mobility data alone were sufficient to meaning-fully forecast COVID-19 infections 7-10 days ahead at all geographic scales. Nouvellet et al [18] concluded that for 52 countries having experienced, or still experiencing, substantial active SARS-CoV-2 transmission, there was a strong link between mobility measures and transmissibility, supporting the implementation of population-wide social distancing interventions to control the epidemic.

In Spain, the high heterogeneity in incidence between similar areas despite the uniform mobility control measures taken suggests that multi-seeding (several independent (non-clustered) infected individuals arrive at a susceptible population) could have played an important role in shaping the spreading of the disease [15]. Mattia Mazzoli and cia addressed the question of how relevant is multi-seeding for the epidemic indicators in a population. They found that local peaks of incidence and mortality strongly correlate with mobility occurred in the early-stage weeks occurred in Madrid, city consider the "hub" in Spain due to economic and social reassons [15].

2.3 Spread based on meteorology

Viruses can be transmitted through the influence of several factors. It has already been suggested that meteorological factors, such as temperature and humidity, are related to the spread

of certain infectious diseases [16].

Early in the pandemic, some studies indicated that mean temperature correlated significantly with the spread of COVID-19 [21]. At the same time, another study focusing on major Turkish cities concluded that the climatic factors that best correlated with the spread of COVID-19 were the mean temperature on the day of positive measurement and the wind speed 14 days before positive detection [20]. Hypotheses that were later validated by another study conducted in Russia [13].

However, the conclusions of these studies should be analyzed with caution. Other studies carried out in other countries, such as the study conducted in South Asian countries [8], have obtained results that are not in line with the above.

In Spain, Fernández-Ahúja et al [7] found in their study that mean temperature is the atmospheric variable that best correlates with virus spread. In their case they used the density of positives in PCR tests. Being more specific, they indicate that low minimum temperatures correlate better than high minimum temperatures. This may be due to the fact that the spread of the virus in cold environments is accelerated, although it may also be due to the tendency of people to stay indoors in adverse weather conditions. Similarly, they indicate that atmospheric pressure may also be a relaxing factor in the density of positives in PCR tests. While they found no evidence that ambient humidity, daily sunshine hours, or precipitation substantially influenced it.

The effects of meteorological variability on COVID-19 transmission is an emerging area of interest. However, the conclusions are highly varied and depend on the location and quantity/quality of data used. In conclusion, from both the analysis of the available literature and the review of the article by McClymont et al [16], temperature appears to be the most promising variable for studying virus infectivity. On the other hand, humidity presents different results according to the different articles analyzed. Finally, both wind speed and precipitation show inconsistent results, so they will not be taken into account during our analysis.

2.4 Data sources

Since the beginning of the pandemic, governments, universities and some technology companies have made available to the community data related to infections, deaths, mobility, etc...

In Spain, the Ministry of Transport, Mobility and Urban Agenda put out to contract a mobility study using Big Data technology. The objective of the study is to provide a characterization of mobility at national, autonomic, provincial and local levels, to support the monitoring of the evolution of the disease, to evaluate the effectiveness of the mobility restriction measures

adopted, as well as to support decision making during this stage of the COVID-19 pandemic.[7]

The analysis compares daily mobility with that of an equivalent typical week prior to the crisis. The week chosen was from 14-February-2020 to 20-February-2020, which had a normal mobility behavior, as there were no public holidays in any autonomous community.

Similarly, the National Institute of Statistics (INE) has been carrying out population mobility studies from mobile telephony since 2019. [8] Currently, the INE is publishing mobility data every week, relating to two specific days of the previous week (Wednesday and Sunday)

In the wake of the pandemic, both Google and Apple began publishing mobility reports based on anonymized data from those users who have location history enabled on their Android and Apple devices, respectively. [9] [10]

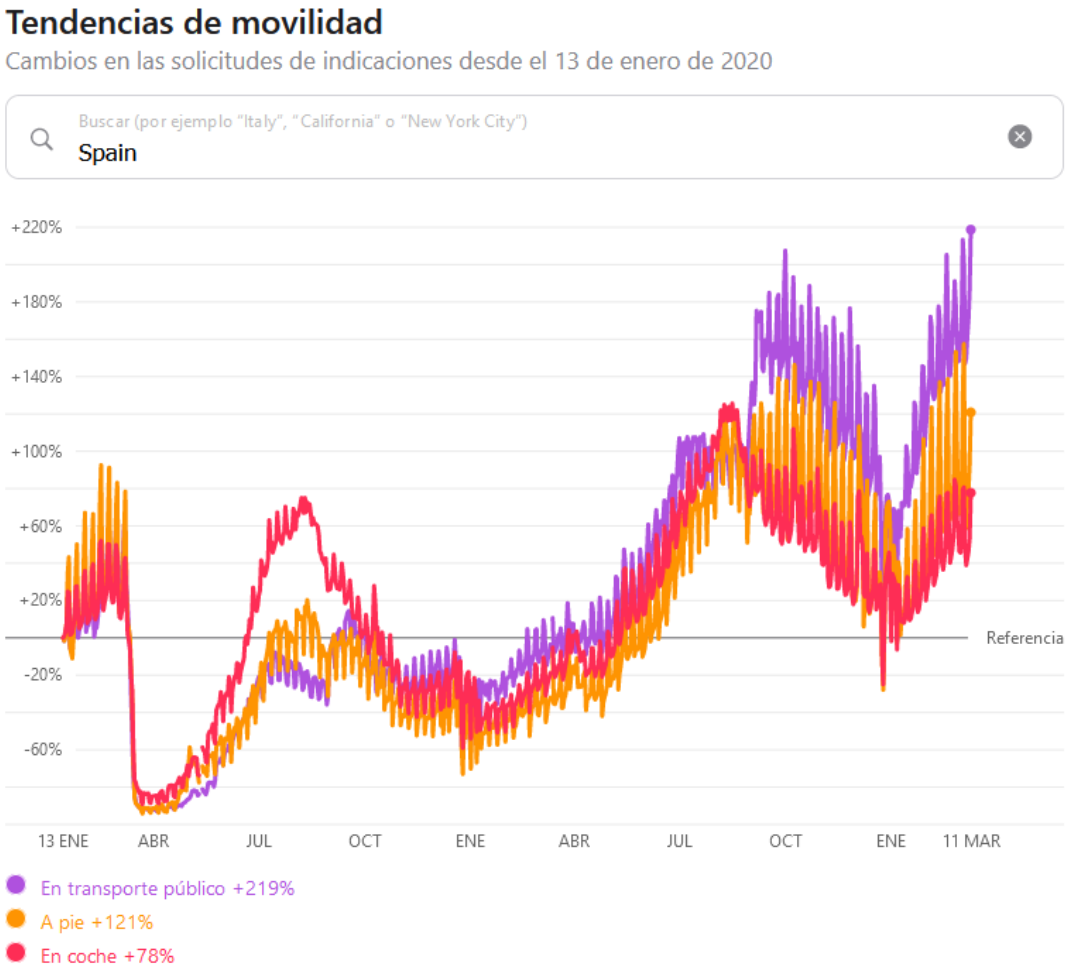

Figure 2.3: Apple's mobility report for Spain

[7] https://www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data
[8] https://www.ine.es/experimental/movilidad/experimental_em4.htm
[9] https://www.google.com/covid19/mobility/
[10] https://covid19.apple.com/mobility

The objective of these Local Mobility Reports is to provide valuable information on the changes that have occurred in people's mobility as a result of the policies that have been put in place to combat COVID-19. In the case of Google, these reports show movement trends over time sorted by geographic area and classified into various categories of locations, such as stores and entertainment spaces, supermarkets and pharmacies, parks, transportation stations, workplaces, and residential areas. Moreover Apple aggregates the data into the following categories: public transportation, walking and driving.

As for meteorological data, as they are not specific to the current pandemic, they have been collected for many years. These can be consulted, both in national pages, such as the Agencia Estatal de Meteorológica (AEMET)[11], in international models such as the European Centre for Medium-Range Weather Forecasts (ECMWF), the Icosahedral Nonhydrostatic (ICON), The Global Forecast System (GFS) which is a National Centers for Environmental Prediction (NCEP) weather forecast model that generates data for dozens of atmospheric and land-soil variables, including temperatures, winds, precipitation, soil moisture, and atmospheric ozone concentration. [12]

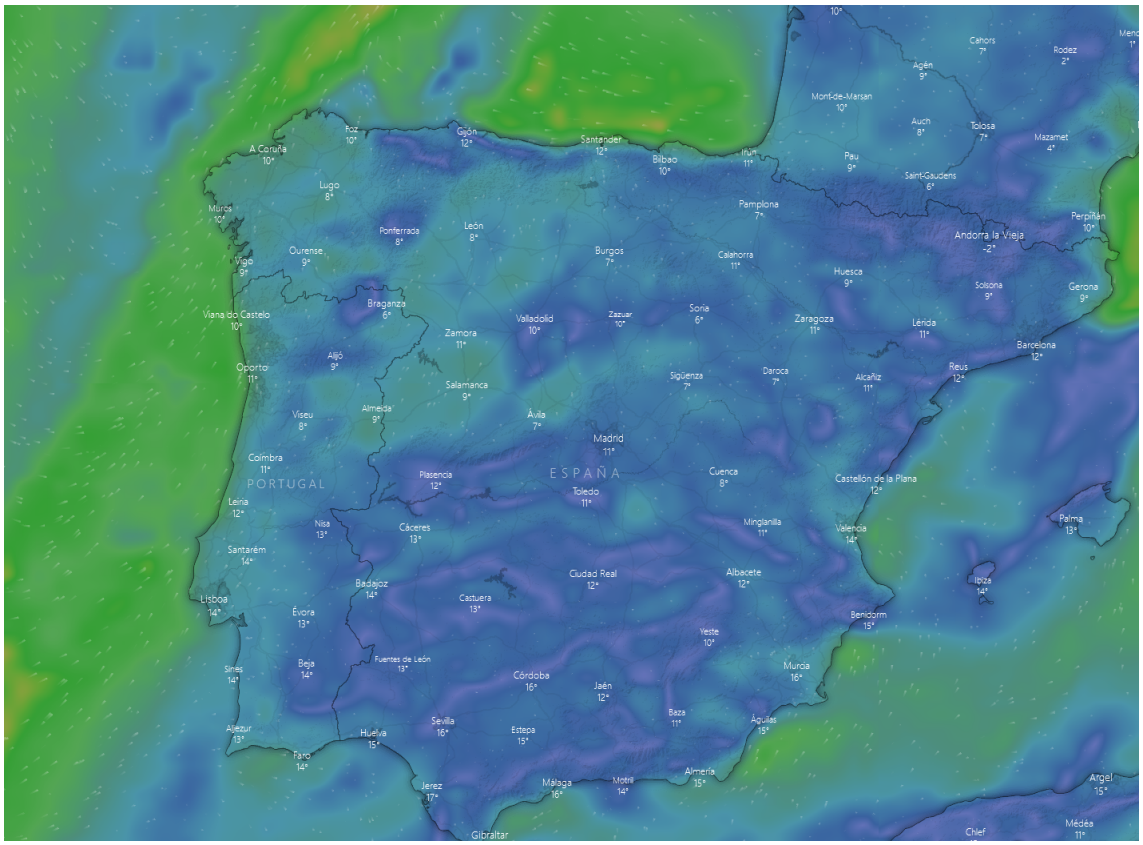

Figure 2.4: Data from ECMWF [https://www.windy.com/]

[11] http://www.aemet.es/es/portada
[12] https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast

Covid data will be extracted from periodical publications by the Ministry of Health [13], from the National Center of Epidemiology (CNE) [14] and from the repository published by the Johns Hopkins University of Medicine. [15]

2.5 Machine learning techniques

Using algorithms or techniques to make predictions is not new. For example, exponential smoothing (ETS) was proposed in the late 1950s. However, the nature, volume and complexity of data availability is changing and that changes bring with them complexity in analysing these data.

Machine learning techniques are relatively new. They bring the ability to learn and improve its performance without being explicitly programmed in advance.

Relevant to the present project, time series analysis are methods used for analysing time series data in order to extract meaningful statistical information from the data. Time series forecasting however, is all about predicting future values based on previously observed values over time.

There are four general components that a time series forecasting model is comprised of [16]:

- Trend: Increase or decrease in the series of data over longer a period.
- Seasonality: Fluctuations in the pattern due to seasonal determinants over a period.
- Cyclical variations: Occurs when data exhibit rises and falls at irregular intervals.
- Random or irregular variations: Instability due to random factors that do not repeat in the pattern.

Plenty of methods for time series forecasting have been introduced over the years, being some of the most successful ones motivated by ETS methodology.

Forecasts produced using ETS are weighted averages of past observations, the weights of which decay exponentially as the observations age [9]. The simplest of the exponential smoothing methods is naturally called Simple Exponential Smoothing (SES). This method is suitable for forecasting data without a clear trend or seasonal pattern. Increasing in complexity, double exponential smoothing can model trend components and level components for univariate

[13] https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/home.htm
[14] https://cnecovid.isciii.es/
[15] https://coronavirus.jhu.edu/map.html
[16] https://www.advancinganalytics.co.uk/blog/2021/06/22/10-incredibly-useful-time-series-forecasting-algorithms

time series data while Holt-Winters Exponential Smoothing, also known as triple exponential smoothing, can model seasonality, trend, and level components for univariate time series data.

On the other hand, ARIMA models are among the currently most widely used approaches for time series forecasting providing with exponential smoothing a complementary approach to the problem. While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data [9] and can be classified into two different formats: non-seasonal and seasonal.

Non-seasonal ARIMA model are a combination of:

- an autoregression model, where we forecast the variable of interest using a linear combination of past values of the variable (autoregression indicates that it is a regression of the variable against itself) [AR(p) model].
- a difference in the nonseasonal observations [I(d)].
- a moving average model, where rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model [MA(q) model].

The seasonal format adds to the non-seasonal part of the model terms that imply seasonal period reversals.

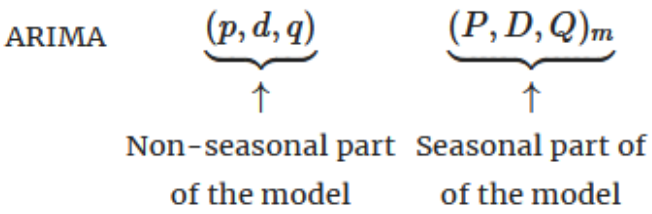

Figure 2.5: Seasonal ARIMA Model

Both ARIMA and ETS allow the inclusion of information from past observations of a series, but not the inclusion of other information that may also be relevant. Dynamic regression models, on the contrary, allows for the inclusion of a lot of relevant information from external predictor variables, which may explain some of the historical variation, but do not allow for the subtle time series dynamics that can be handled with ARIMA models [9]

There are also other advanced forecasting methods such as the neural network models. Artificial Neural Networks (ANNs) are a set of algorithms inspired by the communication mechanism of the biological neuron. They have proven to be a good approach to problems where knowledge is imprecise or time-varying. Their ability to learn makes neural networks

both adaptive and elaborate algorithms allowing complex nonlinear relationships between the response variable and its predictors.

Starting from the basic unit (the neuron) and its most basic application (the perceptron), neural networks can be increased in complexity by the organization of neurons in different layers. Each architecture can be valid for different applications. Additionally, it is possible to add feedback mechanisms to modify the activation of each neuron. These networks are known as Recurrent Neural Networks (RNN). This type of connections can be very useful for dealing with sequences of data, as in the case of time series.

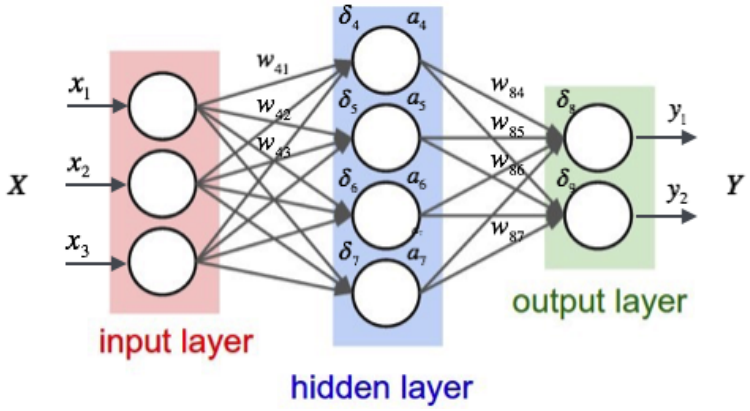

Figure 2.6: Artificial neural networks architecture

Based on ANNs, there are some algorithms often used to solve time series forecasting problems[17].

- Prophet: Prophet, which was released by Facebooks Core Data Science team, is an open-source library developed by Facebook and designed for automatic forecasting of univariate time series data.

- LSTM: Long Short-Term Memory (LSTM) is a type of recurrent neural network that can learn the order dependence between items in a sequence.

- DeepAR: DeepAR developed by Amazon is a probabilistic forecasting model based on autoregressive recurrent neural networks.

- N-BEATS: N-BEATS is a custom Deep Learning algorithm which is based on backward and forward residual links for univariate time series point forecasting.

[17]https://www.advancinganalytics.co.uk/blog/2021/06/22/10-incredibly-useful-time-series-forecasting-algorithms

- Temporal Fusion Transformer (Google): A novel attention-based architecture which combines high-performance multi-horizon forecasting with interpretable insights into temporal dynamics.

Chapter 3

Work design and implementation

3.1 Hosting and code access

In order to facilitate the presentation of the results and the development of the project, a book has been hosted on a public Github repository.

The book was written with the Quarto[1] publishing system and hosted at https://jperezord.github.io/ using Github pages.

Throughout that book, it is exposed the acquisition, cleaning and exploration process of the starting data. It also includes the development of the predictive models used for outbreak prediction and the results obtained.

All the cleaning, the visual analysis and the prediction using ARIMA was developed in R while the LSTM was developed using Python.

[1] https://quarto.org/

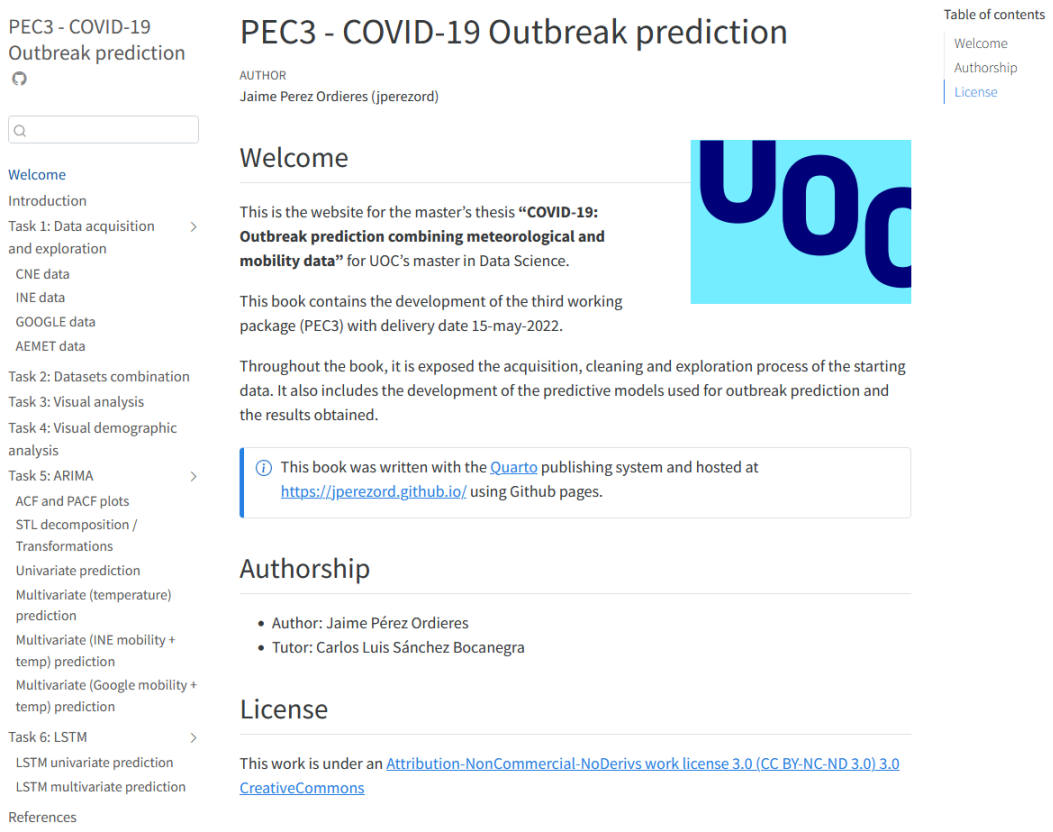

Figure 3.1: Project book hosted on Github Pages [https://jperezord.github.io/]

3.2 Data preprocessing

Since our study is focused in Spain all the source information has been chossen from official institutions when possible.

- CNE (National Epidemiology Center) offers information related to infections, recoveries and deaths reported by local and regional governments in Spain.
- INE (Spanish National Institute of Statistics) provides measurent mobility between areas during the period starting in March-2020 and ended in 2021.
- GOOGLE, even when it is not an official source, provides information related to mobility using the Google ecosystem application services such as Android widely implemented all over the world. It also provides daily data with more detail than INE.
- AEMET, which is the Meteorological State Agency, provides meteorological services under the competence of the State and support for the exercise of other public policies and private activities.

In order to limit our scope, we have selected a series of spanish provinces, which are considered representative, both in terms of volume and location. The provinces listed in alphabetical order are as follows:

- Asturias
- Barcelona
- Madrid
- Málaga
- Sevilla

3.2.1 CNE data[2]

The data published in the CNE COVID-19 Panel offers information related to infections, recoveries and deaths reported by local and regional governments in Spain. It comes from the individualised declaration of COVID-19 cases to the National Epidemiological Surveillance Network (RENAVE) through the SiViEs computer application.

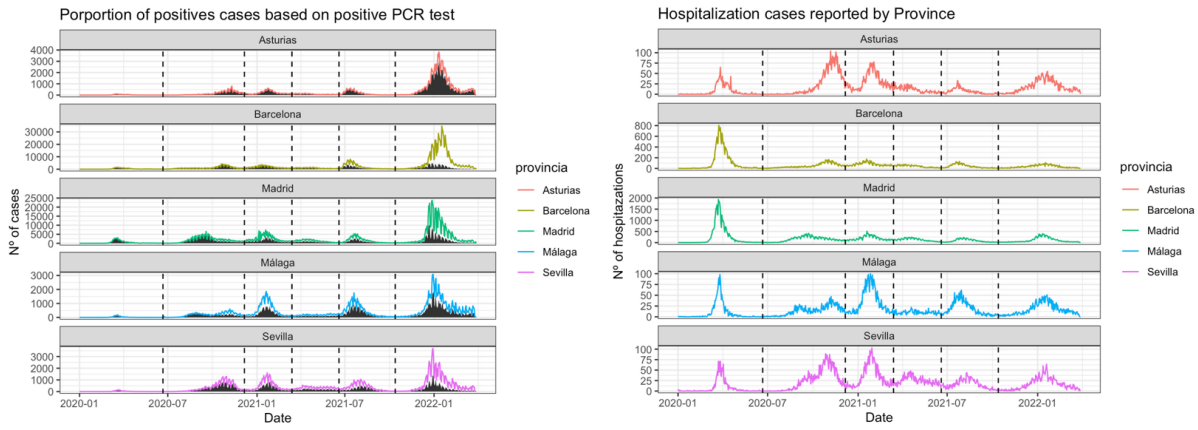

Figure 3.2: Example of hospitalization and cases reported by province

In SiViEs, all reported cases are accounted for, following the surveillance strategy in force at the time.

In the current work, we will use two datasets from CNE:

- casos_tecnica_provincia.csv: Number of cases by diagnostic technique and province (of residence). After the cleaning process we get 42588 rows with information from January 1, 2020 to March 29, 2022.

[2] https://jperezord.github.io/cne_data.html

  - provincia_iso: ISO code of the province of residence.
  - fecha: date considering onset of diagnosis minus 6 days when possible. If this is not possible, the date of the symptoms, or diagnostic for asymptomatic, is considered for the calculation.
  - num_casos: number of cases by diagnostic technique and province of residence.
  - num_casos_prueba_pcr: number of cases with PCR laboratory test or molecular techniques
  - num_casos_prueba_test_ac: number of cases with laboratory test of rapid antibody test.
  - num_casos_prueba_ag: number of cases with laboratory test of antigen detection test antigen test.
  - num_casos_prueba_elisa: number of cases with high-resolution serology laboratory testing.
  - num_casos_prueba_desconocida: number of cases without information on laboratory testing.

- casos_hosp_uci_def_sexo_edad_provres.csv: number of hospitalisations, number of admissions to ICU and number of deaths by sex, age and province of residence. After the cleaning process we get 1277640 rows with information from January 1, 2020 to March 29, 2022.

  - provincia_iso: ISO code of the province of residence.
  - fecha: data. For cases the date of diagnosis is used and for Hospitalisations, ICU admissions and deaths, the cases are represented by date of hospitalisation (failing this, date of diagnosis, and if not, date of death).
  - sexo: sex of cases: H (male), M (female), NC (not stated).
  - grupo_edad: age group to which the case belongs: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79 , $\geq$ 80 years. NC: not stated.
  - num_casos: number of reported cases confirmed as having a positive diagnostic test for active infection (PDIA) as infection (PDIA) as set out in the Early Detection Strategy, surveillance and control strategy for COVID-19 and in addition cases notified before 11-May that required hospitalisation, ICU admission or required hospitalisation, ICU admission or died with a clinical diagnosis of COVID19, according to the case definitions in force at the time.

- num_hosp: number of hospitalised cases.
- num_uci: number of cases admitted to ICU.
- num_def: number of deaths.

3.2.2 INE data[3]

The main objective of the INE dataset is to measure mobility between areas during the period starting in March-2020 and ended in December-2021.

The population scope is made up of the mobile telephones of the resident population in Spain of the three main mobile operators. Foreign numbered phones are excluded, usually mobiles in the hands of tourists which operate in Spain while roaming.

Data is available in the Spanish National Institute of Statistics grouped by autonomous communities, provinces and even islands. However, the information provided is not daily, so we have opted to interpolate the missing information to obtain daily details.

Once downloaded in csv format and cleaned we get 34008 rows with information for all the provinces from March 16, 2020 to December 29, 2021. As to information columns, it is provided:

- fecha: date.
- province: province name.
- flujo: mobility percentage.

3.2.3 Google data[4]

The data published by Google offers information related to mobility using the Google ecosystem application services such as Android. The dataset information is available worldwide, but in our case, only the information relating to Spain was extracted.

The Google Mobility Reports aim to provide insights into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, supermarkets and pharmacies, parks, public transport, workplaces and residential.

Each Community Mobility Report is broken down by location and displays the change in visits to places.

[3] https://jperezord.github.io/ine_data.html
[4] https://jperezord.github.io/google_data.html

Figure 3.3: Visits to workplaces and workplaces vs infections by province

Data provided by Google are splitted in different files by year. For spanish data, files with format "20XX_ES_Region_Mobility_Report.csv" were extracted being "XX" the year contained inside.

Once cleaned, google data contains 40250 rows with information related to all the provinces in Spain. The information ranges from February 12, 2020 to April 29, 2022. It provides detail information about:

- CA: autonomous communities codes.
- province: province names.
- iso_3166_2_code: province iso code.
- fecha: date.
- mob_grocery_pharmacy: Mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug, stores, and pharmacies.
- mob_parks: Mobility trends for places like national parks, public beaches, marinas, dog parks, plazas, and public garden
- mob_residential: Mobility trends for places of residence.
- mob_retail_recreation: Mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters.
- mob_transit_stations: Mobility trends for places like public transport hubs such as subway, bus, and train stations.
- mob_workplaces: Mobility trends for places of work.

3.2.4 AEMET data[5]

Spanish State Meteorological Agency's data are available throug its service AEMET OpenData.

AEMET OpenData is a REST API (Application Programming Interface. REpresentational State Transfer) through which data can be downloaded free of charge. It allows the dissemination and reuse of the Agency's meteorological and climatological information, in the sense indicated in Law 18/2015, of July 9, amending Law 37/2007, of November 16, on the reuse of public sector information.

AEMET OpenData allows two types of access where both allow access to the same data catalog and data download in reusable formats:

- General Access: It is a graphical access, intended for the general public. Its purpose is to allow access to data for users in a friendly way.
- AEMET OpenData API: it allows another type of interaction with the data. This interaction is characterized by the possibility of being periodic and even programmed from any programming language.

For the present study, we used last method to download data from:

- Asturias airport
- Barcelona airport
- Madrid airport
- Málaga airport
- Sevilla airport

Once cleaned, it provides information from January 1, 2020 to March 31, 2022 related to:

- fecha: date.
- provincia: province name.
- tmed: average temperature (ºC).
- prec: precipitations (mm).
- tmin: minimum temperature (ºC).

[5] https://jperezord.github.io/aemet_data.html

- tmax: maximum temperature (ºC).
- wd: wind direction (º).
- ws: wind speed average (m/s).
- ws_max: wind speed maximum (m/s).
- sol: hour of sun light (hr).

3.3 Relevant variables

From visual, correlation and PCA analysis the variables that contribute the most to explain the variance of the data are the average temperature from the meteorological data and transit/grocery/retail information from the mobility provided by Google.

INE mobility data in addition to being less detailed than Google, do not contribute as much as Google data did.

It is also important the number of hospitalizations in Asturias, Málaga and Sevilla, but since that variable is not usefull to predict covid cases in advanced, we will not feed future models with it.

Both correlation and PCA analysis shows that in Madrid the average temperature is important to explain the variance of data.

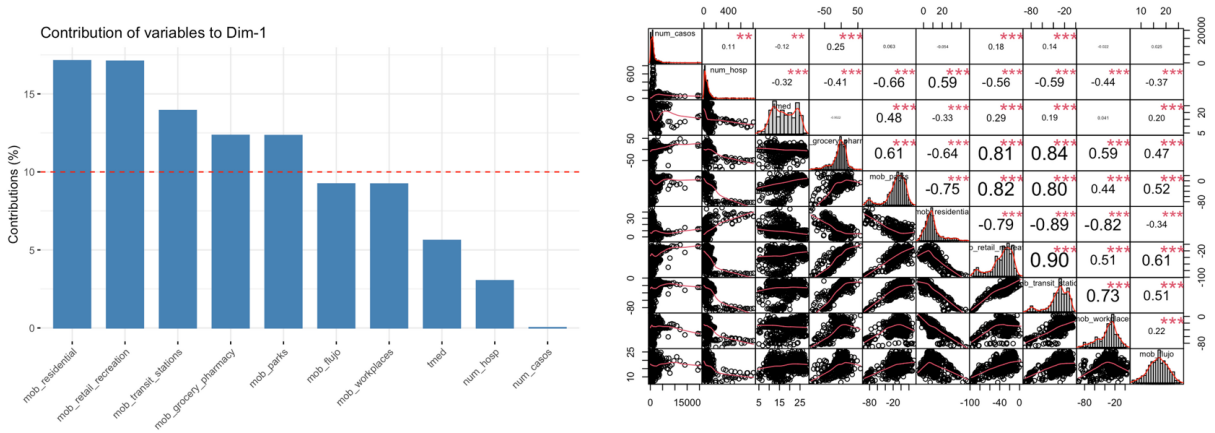

Figure 3.4: Example of correlation and PCA analysis

3.4 Preparation for modelling

Firstly, although information is available from the beginning of 2020, it has been decided to train the model with data after June 14, 2020. Data on the onset of the coronavirus pandemic must be treated with particular caution due to the level of uncertainty. In addition, during this initial period, the population was totally confined, so any relevant information that can be extracted cannot be extrapolated to later situations.

Secondly, we adopted two approaches to select train and test data. As the latest wave of the coronavirus has had a much higher incidence than the rest, it has been decided to analyze two different periods. On the one hand, information prior to the start of the latest incidence wave. On the other hand, information related to the deceleration period of the occurrence corresponding to the last wave.

- First approach: train data from June 14, 2020 to October, 14 2021.
- Second approach: train data from June 14, 2020 to January, 31 2022.

In both approaches the forecast horizon was 7, 14 and 21 days from the final date of the train data. We also did a 90 days horizon in order to test the quality of the predictions in long terms.

A similar procedure has been applied to LSTM modelling, in which an attempt has been made to predict both the last wave, the peak and the period of contagion decline of the last wave. For the latter method, training based on the last 90 days of incidence has been chosen for the prediction of a future incidence value.

3.5 ARIMA

As discussed in the analysis of the state of the art, time series are composed of a number of common patterns: trend, seasonal, cyclic and white noise or irregular variations [9].

The Seasonal and Trend decomposition using Loess (STL) is a versatile and robust method for decomposing time series on those patterns which will be useful to analyse the typology of the data we are dealing with.

Our modelling procedure has followed the basis set out by R.Hyndman and G.Athanasopoulos [9]. When fitting an ARIMA model to a set of (non-seasonal) time series data, the following procedure provides a useful general approach.

1. Plot the data and identify any unusual observations.

2. Apply STL decomposition and transform the data stabilizing the variance by using a Box-Cox transformation.

3. If the data are non-stationary, take first differences of the data until the data are stationary.

4. Examine the ACF/PACF

5. Try the chosen model(s), and use the AICc to search for a better model.

6. Check the residuals from the chosen model by plotting the ACF of the residuals testing the residuals.

7. Once the residuals look like white noise, calculate forecasts.

We did both an univariate and multivariate forecast perspectives in order to look at the differences between variables considered relevant in the infection levels.

We used the ARIMA() function from fable R package that uses a variation of the Hyndman-Khandakar algorithm [10], which combines unit root tests, minimisation of the AICc and MLE to obtain an ARIMA model.

To quantify the quality of the models, we used scale-dependent measures based on the absolute errors and squared errors:

- Mean absolute error: $MAE = mean(|e_t|)$
- Root mean squared error: $RMSE = \sqrt{mean(|e_t^2|)}$

3.6 LSTM

In ARIMA models it is required to first remove the trend and seasonality, for example by computing the difference between the value at each time step and the value one year earlier. After the model is trained and makes predictions, you would have to add the seasonal pattern back to get the final predictions.

On the contrary, when using recurrent neural network (RNNs) no such action is needed, improving performance in some cases since the model will not have to learn the trend or the seasonality. What is necessary on the other hand is to normalize the input data to the network.

The RNN used in our work was the Long short-term memory (LSTM) which uses a sequences of data to predict the next one. LSTM networks are well-suited to making predictions based on time series since there can be lags of unknown duration between important events in a time series.

For the application of this methodology, it has been decided to apply a slightly different approach to that applied for the ARIMA case. On the one hand, a prediction has been made with all available information. Since the sequence is 90 days, this approximation helps us to predict whether the model is capable of analysing the maximum peak of infection in the sixth wave, as well as its decline curve.

On the other hand, information prior to the start of the sixth wave has been analysed. The objective in this case is to see if the model is able to predict both the onset and the peak and trough of the sixth wave.

As to the configuration of the LSTM networks, we have choosen to use the following structure:

- The first layer is the input of the data. We decided to use a 90 days sequence.
- Three hidden layers would be in charge of learning with 50, 25 and 5 neurons respectively.
- A final dense layer with ReLu activation which returns the forecasted number of new covid infections.

The parameters used to train the previous networks has been:

- Adam as the optimization functions due to its efficiency.
- A batch size of 1000
- An epoch of 30

Results were analyzed using RMSE as in the ARIMA models exposed previously.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 90, 90)            33120

 lstm_1 (LSTM)               (None, 90, 50)            28200

 lstm_2 (LSTM)               (None, 25)                7600

 dense (Dense)               (None, 5)                 130

 dense_1 (Dense)             (None, 1)                 6

=================================================================
Total params: 69,056
Trainable params: 69,056
Non-trainable params: 0
_________________________________________________________________
```

Figure 3.5: Model summary of LSTM network

Chapter 4

Results

4.1 Visual analysis

In the section 'Task 3: Visual analysis' of the repository hosted on Github Pages https://jperezord.github.io/visual.html, the representation of each of the variables by itself and against the variable number of cases is shown in detail.

Each of the graphs groups the information by provinces, where the different incidence of covid in each of the regions can be seen.

Due to the disparity between the incidences of the population centres Madrid and Barcelona, and in order to see in detail the information of each of the provinces, the scaling of the y-axis has been released.

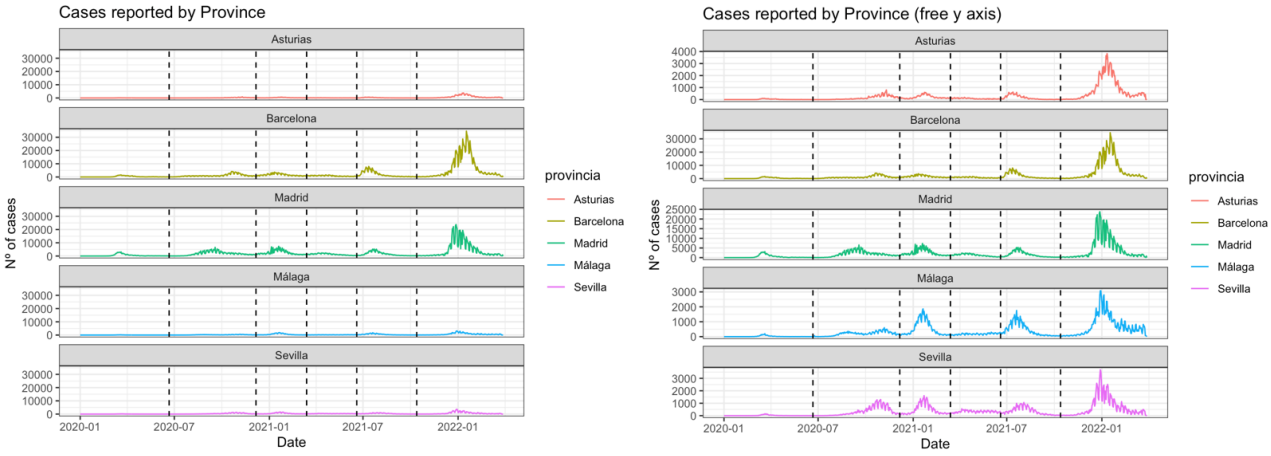

Figure 4.1: Covid incidence by province: fixed vs free Y axis

From the visual analysis the idea is drawn that it is possible that the data reported by the

different provinces may not have the same consistency. For example, while the number of covid cases identified in the last wave in Asturias are supported by PCR testing, in Barcelona the proportion is minimal. This could indicate that many cases in Asturias have not been reported because PCR testing were not available or that in Barcelona more cases have been counted than could actually be associated with Covid-19.

However, the incidence of the latest wave, which has been much greater than in the rest, also suggests that the cases reported at the beginning of the pandemic may have been underestimated. Similarly, an analysis of the incidence of each of the covid-19 mutations in each province could help explain this behaviour.

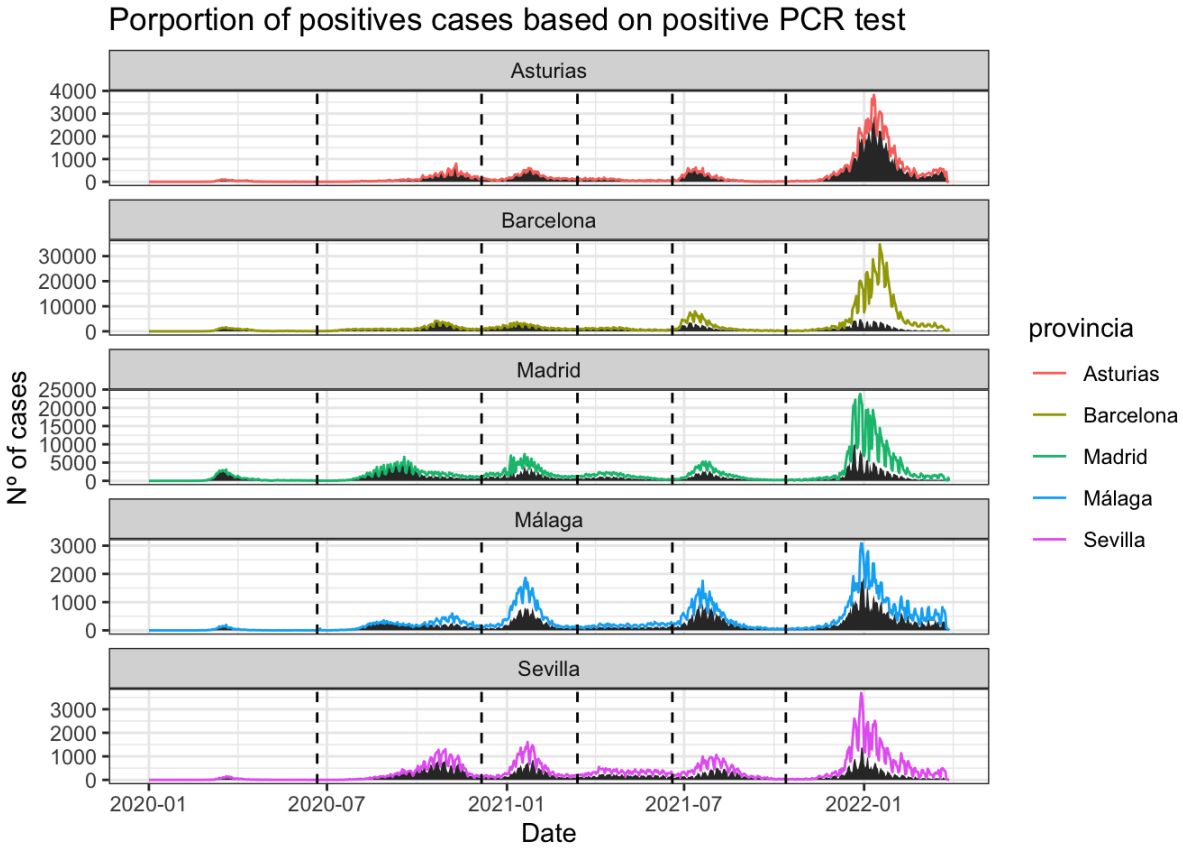

Figure 4.2: Covid incidence supported by PCR testing

4.2 ARIMA results

4.2.1 Univariate

Univariate = ARIMA models using number of cases

Model	Province	RMSE	MAE
arima_at1	Asturias	10.3	7.00
arima_at2	Asturias	9.97	7.28
Snaive	Asturias	14.0	9.98
arima_at1	Barcelona	64.7	54.8
arima_at2	Barcelona	68.1	55.2
Snaive	Barcelona	79.0	67.8
arima_at1	Madrid	163.0	144.0
arima_at2	Madrid	163.0	144.0
Snaive	Madrid	124.0	105.0
arima_at1	Málaga	22.6	21.8
arima_at2	Málaga	22.9	22.3
Snaive	Málaga	22.9	19.8
arima_at1	Sevilla	11.4	8.80
arima_at2	Sevilla	11.4	8.93
Snaive	Sevilla	16.0	13.7

Table 4.1: Univariate ARIMA: 7 days forecasts before 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	14.3	11.1
arima_at2	Asturias	14.2	11.5
Snaive	Asturias	16.0	11.4
arima_at1	Barcelona	73.8	57.1
arima_at2	Barcelona	76.8	58.4
Snaive	Barcelona	87.1	76.0
arima_at1	Madrid	178.0	155.0
arima_at2	Madrid	178.0	155.0
Snaive	Madrid	126.0	105.0
arima_at1	Málaga	27.7	24.0
arima_at2	Málaga	29.0	25.2
Snaive	Málaga	22.7	18.7
arima_at1	Sevilla	9.82	7.83
arima_at2	Sevilla	9.88	7.91
Snaive	Sevilla	19.3	16.4

Table 4.2: Univariate ARIMA: 14 days forecasts before 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	13.5	10.5
arima_at2	Asturias	13.7	11.0
Snaive	Asturias	15.6	11.5
arima_at1	Barcelona	107.0	75.5
arima_at2	Barcelona	103.0	73.6
Snaive	Barcelona	119.0	90.4
arima_at1	Madrid	210.0	177.0
arima_at2	Madrid	210.0	177.0
Snaive	Madrid	140.0	117.0
arima_at1	Málaga	33.3	29.2
arima_at2	Málaga	34.8	30.7
Snaive	Málaga	22.0	18.6
arima_at1	Sevilla	11.3	8.54
arima_at2	Sevilla	11.4	8.61
Snaive	Sevilla	22.1	19.1

Table 4.3: Univariate ARIMA: 21 days forecasts before 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	1211	704
arima_at2	Asturias	1221	714
Snaive	Asturias	1213	704
arima_at1	Barcelona	9027	5072
arima_at2	Barcelona	8907	4964
Snaive	Barcelona	8890	4939
arima_at1	Madrid	8744	5008
arima_at2	Madrid	8744	5008
Snaive	Madrid	8564	4792
arima_at1	Málaga	1100	680
arima_at2	Málaga	1099	677
Snaive	Málaga	1049	628
arima_at1	Sevilla	1161	682
arima_at2	Sevilla	1161	682
Snaive	Sevilla	1161	677

Table 4.4: Univariate ARIMA: 90 days forecasts before 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	272	251
arima_at2	Asturias	334	317
Snaive	Asturias	586	570
arima_at1	Barcelona	4824	4569
arima_at2	Barcelona	4383	4145
Snaive	Barcelona	8139	7870
arima_at1	Madrid	788	629
arima_at2	Madrid	788	629
Snaive	Madrid	2070	1836
arima_at1	Málaga	190	152
arima_at2	Málaga	190	152
Snaive	Málaga	240	191
arima_at1	Sevilla	78.6	56.6
arima_at2	Sevilla	78.8	55.9
Snaive	Sevilla	368	311

Table 4.5: Univariate ARIMA: 7 days forecasts end of the 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	332	310
arima_at2	Asturias	447	424
Snaive	Asturias	812	769
arima_at1	Barcelona	7100	6582
arima_at2	Barcelona	6534	6038
Snaive	Barcelona	10606	9985
arima_at1	Madrid	612	467
arima_at2	Madrid	612	467
Snaive	Madrid	3386	2914
arima_at1	Málaga	290	229
arima_at2	Málaga	290	229
Snaive	Málaga	223	170
arima_at1	Sevilla	223	156
arima_at2	Sevilla	224	157
Snaive	Sevilla	338	299

Table 4.6: Univariate ARIMA: 14 days forecasts end of the 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	411	381
arima_at2	Asturias	543	514
Snaive	Asturias	980	920
arima_at1	Barcelona	8400	7823
arima_at2	Barcelona	7750	7199
Snaive	Barcelona	12018	11275
arima_at1	Madrid	851	664
arima_at2	Madrid	851	664
Snaive	Madrid	4584	3903
arima_at1	Málaga	267	203
arima_at2	Málaga	267	203
Snaive	Málaga	279	227
arima_at1	Sevilla	208	154
arima_at2	Sevilla	209	155
Snaive	Sevilla	430	381

Table 4.7: Univariate ARIMA: 21 days forecasts end of the 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	631	589
arima_at2	Asturias	735	702
Snaive	Asturias	1359	1282
arima_at1	Barcelona	12454	11642
arima_at2	Barcelona	11449	10693
Snaive	Barcelona	15796	14809
arima_at1	Madrid	2217	1701
arima_at2	Madrid	2217	1701
Snaive	Madrid	7446	6507
arima_at1	Málaga	278	190
arima_at2	Málaga	278	190
Snaive	Málaga	525	443
arima_at1	Sevilla	208	157
arima_at2	Sevilla	207	156
Snaive	Sevilla	850	738

Table 4.8: Univariate ARIMA: 57 days forecasts end of the 6th epidemiological period

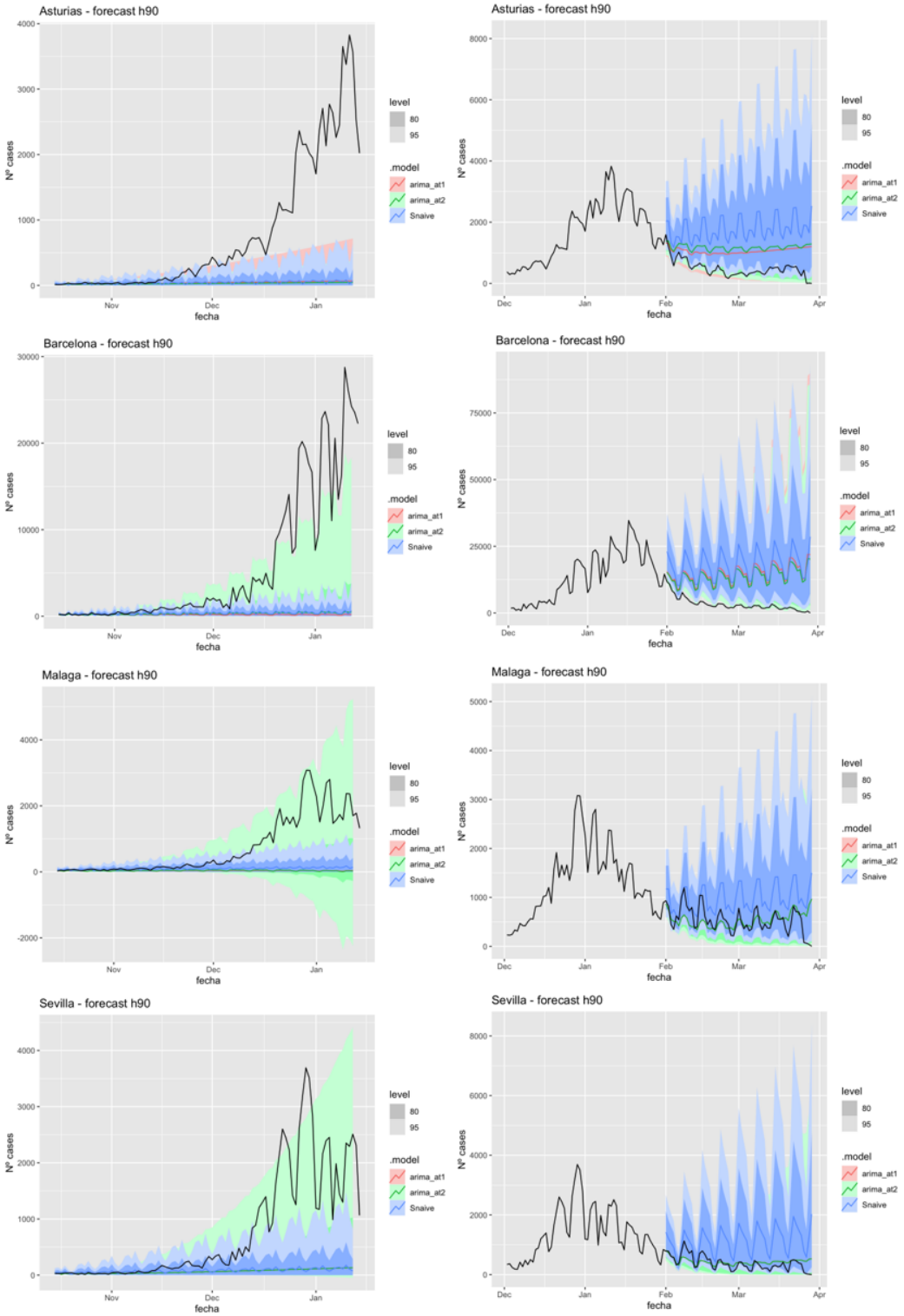

Figure 4.3: ARIMA Univariate results. Left column predictions prior to the start of the last wave. Right column predictions during the deceleration of the incidence of the last wave.

4.2.2 Multivariate

Multivatiare = ARIMA models using number of cases + average temperature + Google mobility.

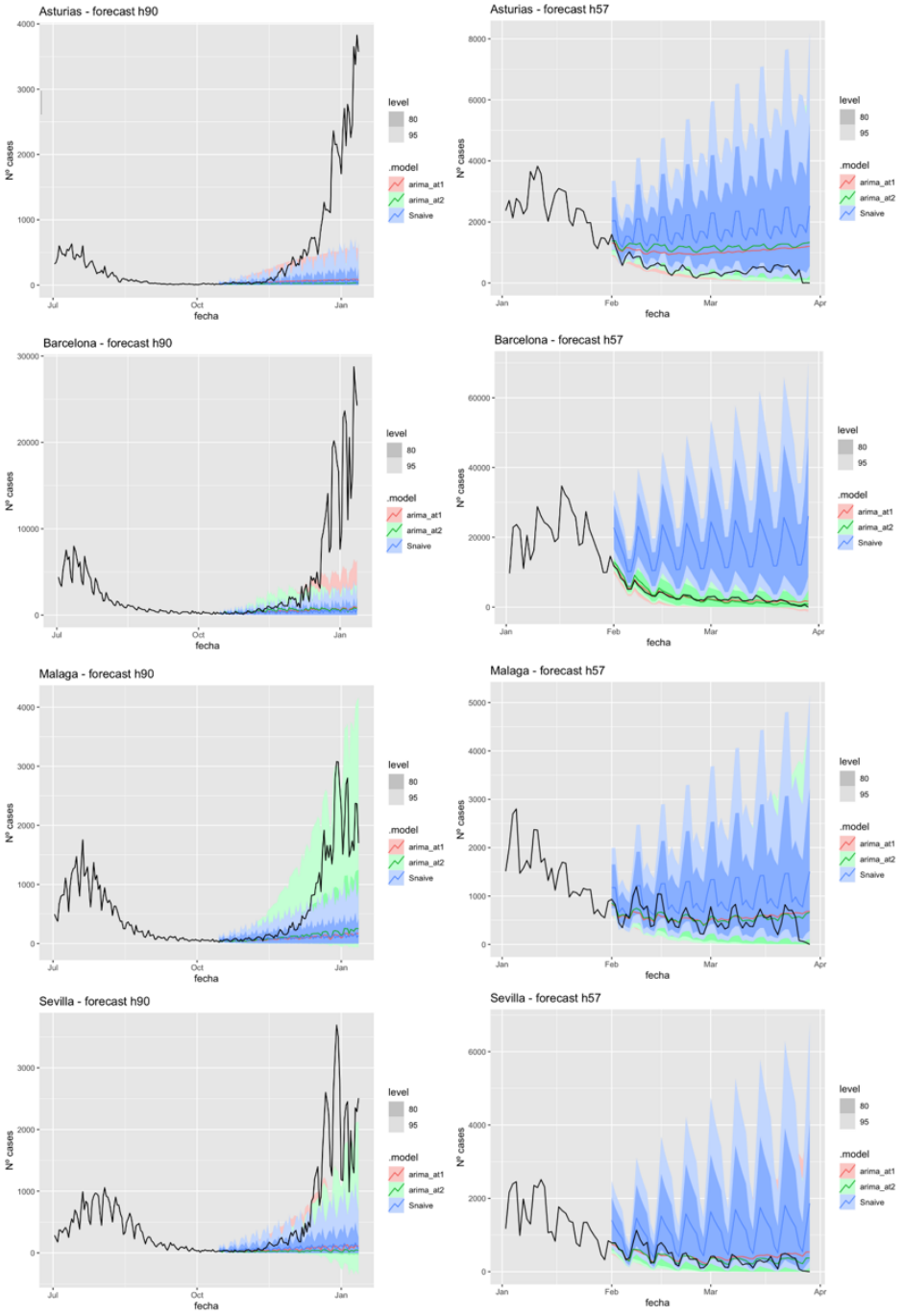

Figure 4.4: ARIMA Multivariate results. Left column predictions prior to the start of the last wave. Right column predictions during the deceleration of the incidence of the last wave.

Model	Province	RMSE	MAE
arima_at1	Asturias	9	5.65
arima_at2	Asturias	9.84	6.95
Snaive	Asturias	14.0	9.98
arima_at1	Barcelona	75.9	67
arima_at2	Barcelona	88.5	79
Snaive	Barcelona	80.0	67.1
arima_at1	Madrid	114.0	98.7
arima_at2	Madrid	114.0	98.7
Snaive	Madrid	123.0	102.0
arima_at1	Málaga	18.0	16.2
arima_at2	Málaga	18.3	16.3
Snaive	Málaga	22.9	19.8
arima_at1	Sevilla	9.18	7.23
arima_at2	Sevilla	10.6	8.11
Snaive	Sevilla	15.6	13.3

Table 4.9: Multivariate ARIMA: 7 days forecasts before 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	11.7	8.37
arima_at2	Asturias	15.0	12.1
Snaive	Asturias	16.0	11.4
arima_at1	Barcelona	72.1	63.5
arima_at2	Barcelona	91.4	80.9
Snaive	Barcelona	88.0	75.0
arima_at1	Madrid	118.0	102.0
arima_at2	Madrid	118.0	102.0
Snaive	Madrid	124.0	102.0
arima_at1	Málaga	21.1	17.6
arima_at2	Málaga	21.7	17.9
Snaive	Málaga	22.7	18.7
arima_at1	Sevilla	8.67	7.40
arima_at2	Sevilla	10.4	7.82
Snaive	Sevilla	18.4	15.6

Table 4.10: Multivariate ARIMA: 14 days forecasts before 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	10.5	7.85
arima_at2	Asturias	14.8	12.1
Snaive	Asturias	15.6	11.5
arima_at1	Barcelona	82.2	67.8
arima_at2	Barcelona	91.9	83.2
Snaive	Barcelona	119.0	90.1
arima_at1	Madrid	148	121.0
arima_at2	Madrid	148.0	121.0
Snaive	Madrid	137.0	112.0
arima_at1	Málaga	26.2	22.0
arima_at2	Málaga	26.0	22.0
Snaive	Málaga	22.0	18.6
arima_at1	Sevilla	10.7	8.64
arima_at2	Sevilla	14.6	10.9
Snaive	Sevilla	22.1	18.2

Table 4.11: Multivariate ARIMA: 21 days forecasts before 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	1200	691
arima_at2	Asturias	1229	721
Snaive	Asturias	1213	704
arima_at1	Barcelona	8726	4788
arima_at2	Barcelona	5750	4736
Snaive	Barcelona	8865	4913
arima_at1	Madrid	8366	4665
arima_at2	Madrid	8366	4665
Snaive	Madrid	8525	4752
arima_at1	Málaga	1042	629
arima_at2	Málaga	1007	601
Snaive	Málaga	1049	628
arima_at1	Sevilla	1168	689
arima_at2	Sevilla	1200	717
Snaive	Sevilla	1161	680

Table 4.12: Multivariate ARIMA: 90 days forecasts before 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	272	248
arima_at2	Asturias	336	319
Snaive	Asturias	586	570
arima_at1	Barcelona	374	304
arima_at2	Barcelona	1213	1168
Snaive	Barcelona	7943	7681
arima_at1	Madrid	509	401
arima_at2	Madrid	569	462
Snaive	Madrid	1869	1649
arima_at1	Málaga	169	128
arima_at2	Málaga	162	117
Snaive	Málaga	240	191
arima_at1	Sevilla	75.6	58.3
arima_at2	Sevilla	72.4	52.1
Snaive	Sevilla	356	300

Table 4.13: Multivariate ARIMA: 7 days forecasts end of the 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	335	311
arima_at2	Asturias	461	434
Snaive	Asturias	812	769
arima_at1	Barcelona	432	365
arima_at2	Barcelona	1217	1123
Snaive	Barcelona	10297	9702
arima_at1	Madrid	586	460
arima_at2	Madrid	553	449
Snaive	Madrid	3064	2634
arima_at1	Málaga	245	194
arima_at2	Málaga	220	176
Snaive	Málaga	223	170
arima_at1	Sevilla	224	157
arima_at2	Sevilla	211	147
Snaive	Sevilla	321	282

Table 4.14: Multivariate ARIMA: 14 days forecasts end of the 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	410	380
arima_at2	Asturias	555	524
Snaive	Asturias	980	920
arima_at1	Barcelona	381	322
arima_at2	Barcelona	1032	886
Snaive	Barcelona	11601	10898
arima_at1	Madrid	911	709
arima_at2	Madrid	830	657
Snaive	Madrid	4142	3530
arima_at1	Málaga	227	170
arima_at2	Málaga	201	153
Snaive	Málaga	280	228
arima_at1	Sevilla	211	155
arima_at2	Sevilla	196	140
Snaive	Sevilla	404	358

Table 4.15: Multivariate ARIMA: 21 days forecasts end of the 6th epidemiological period

Model	Province	RMSE	MAE
arima_at1	Asturias	623	581
arima_at2	Asturias	749	715
Snaive	Asturias	1359	1282
arima_at1	Barcelona	598	482
arima_at2	Barcelona	912	777
Snaive	Barcelona	14803	13931
arima_at1	Madrid	1478	1249
arima_at2	Madrid	1762	1461
Snaive	Madrid	6384	5635
arima_at1	Málaga	246	182
arima_at2	Málaga	219	161
Snaive	Málaga	527	445
arima_at1	Sevilla	207	154
arima_at2	Sevilla	155	113
Snaive	Sevilla	784	682

Table 4.16: Multivariate ARIMA: 57 days forecasts end of the 6th epidemiological period

4.3 LSTM results

4.3.1 Univariate

Results for LSTM models using all the data available and a sequence of 90 days. These corresponds to the study of the peak and deceleration of the 6th covid wave.

	Asturias	Barcelona	Madrid	Malaga	Sevilla
MAE	1074	5228.7	4748.2	353.5	335.1
RMSE	1479.1	7609.8	6975.8	481.1	474

Table 4.17: Univariate LSTM model results with all the data

Results for LSTM models using data before the beginning of the 6th wave and a sequence of 90 days. The aim of these models is to analyze if LSTM are able to identify the 6th wave from the previous available data.

	Asturias	Barcelona	Madrid	Malaga	Sevilla
MAE	208.7	1674.3	2521.6	197.7	329.2
RMSE	387	3525.8	5345.2	331.7	595.9

Table 4.18: Univariate LSTM model graphic 90 days results using data before the beginning of the 6th wave

4.3.2 Multivariate

All the data:

	Asturias	Barcelona	Madrid	Malaga	Sevilla
MAE	738.5	6872.2	4531.8	322.4	1622.4
RMSE	1169.8	11041.0	7047.1	508.6	1889.7

Table 4.19: Multivariate LSTM model results with all the data

Just before the 6th wave:

	Asturias	Barcelona	Madrid	Malaga	Sevilla
MAE	263.9	2657.6	2755.3	301.5	502.2
RMSE	465.3	4156.6	5897.2	408.5	535.4

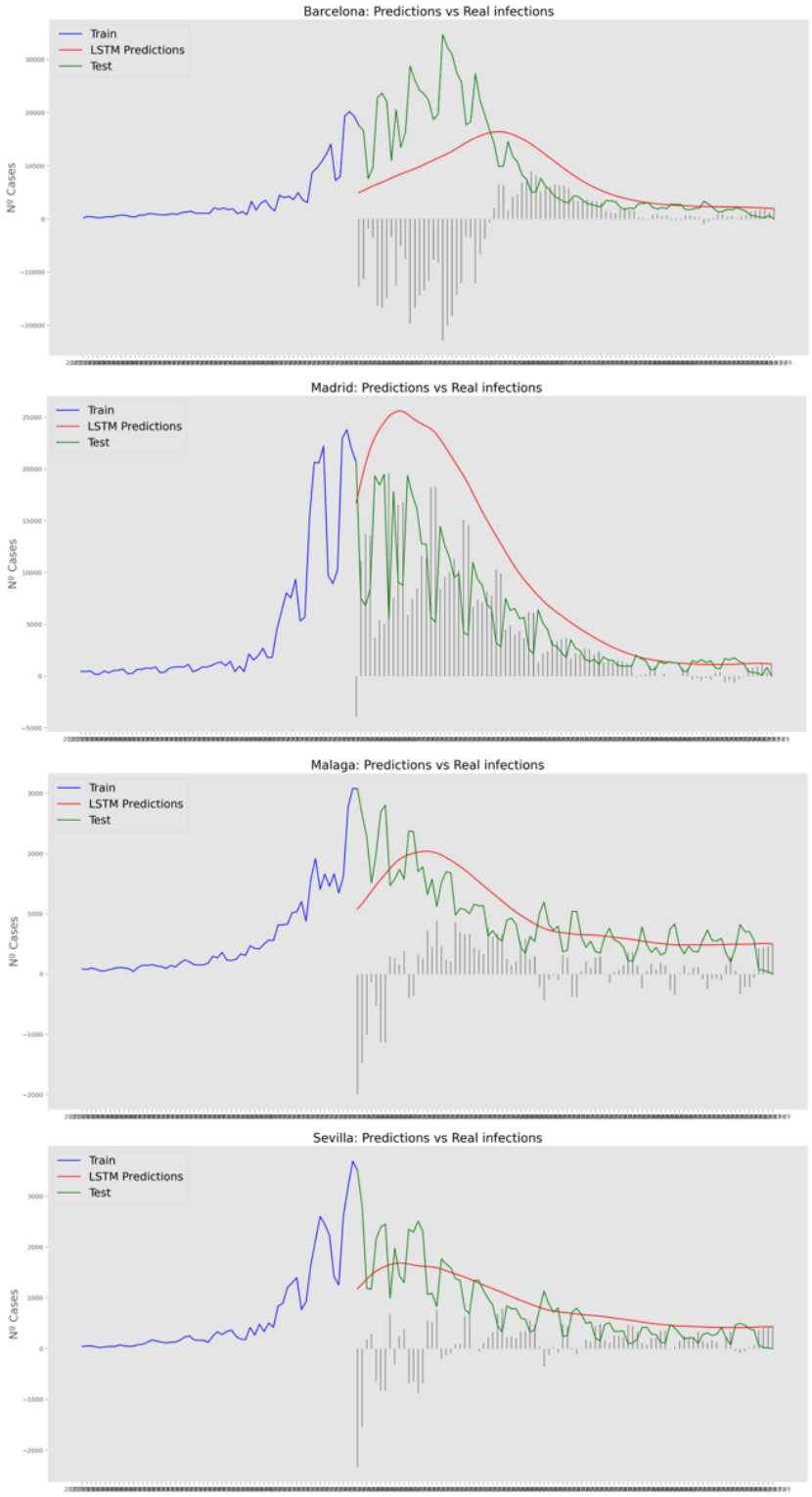

Figure 4.5: Univariate LSTM model graphic 90 days results with all the data

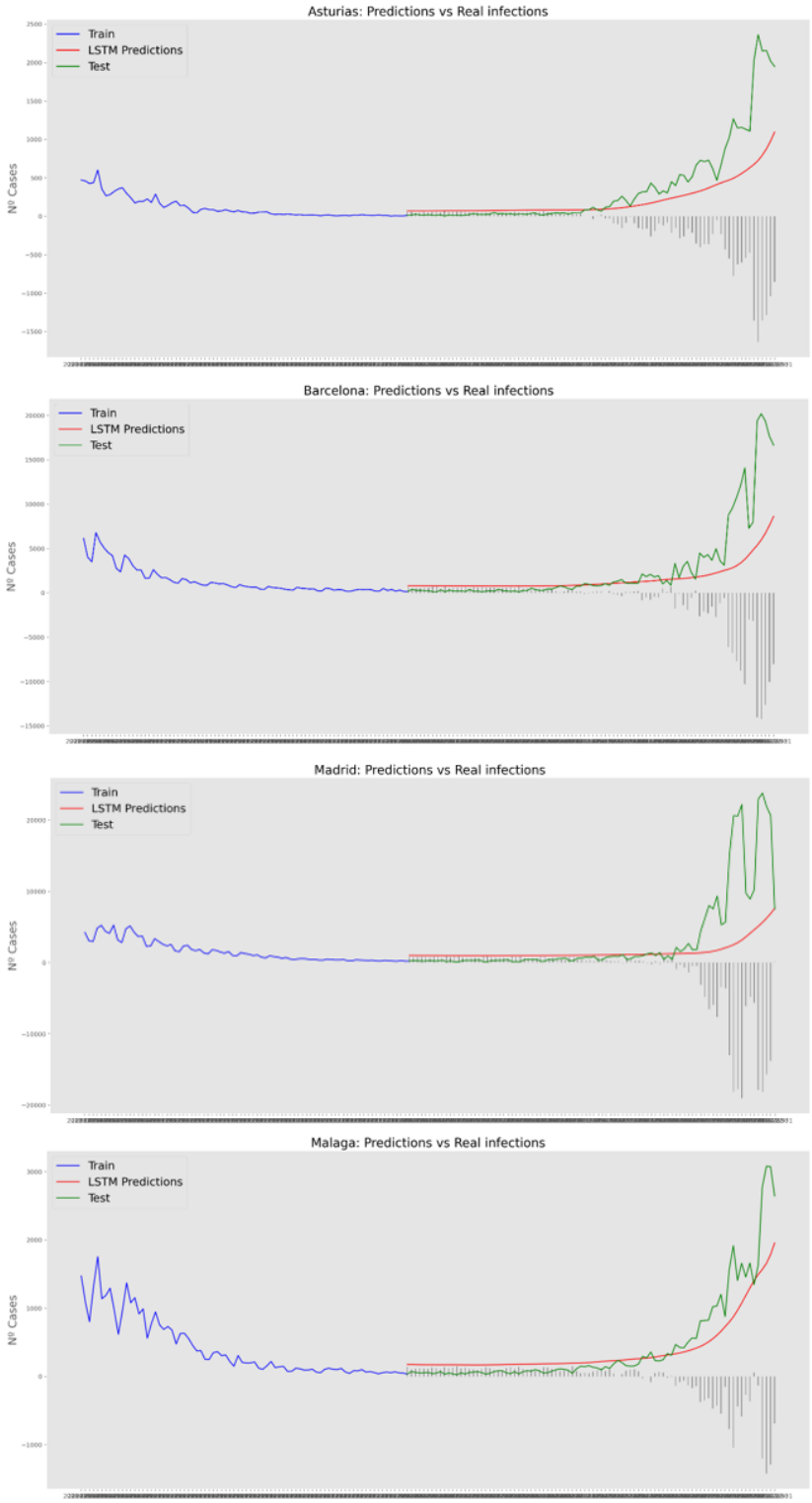

Figure 4.6: Univariate LSTM model graphic 90 days results just before the 6th wave

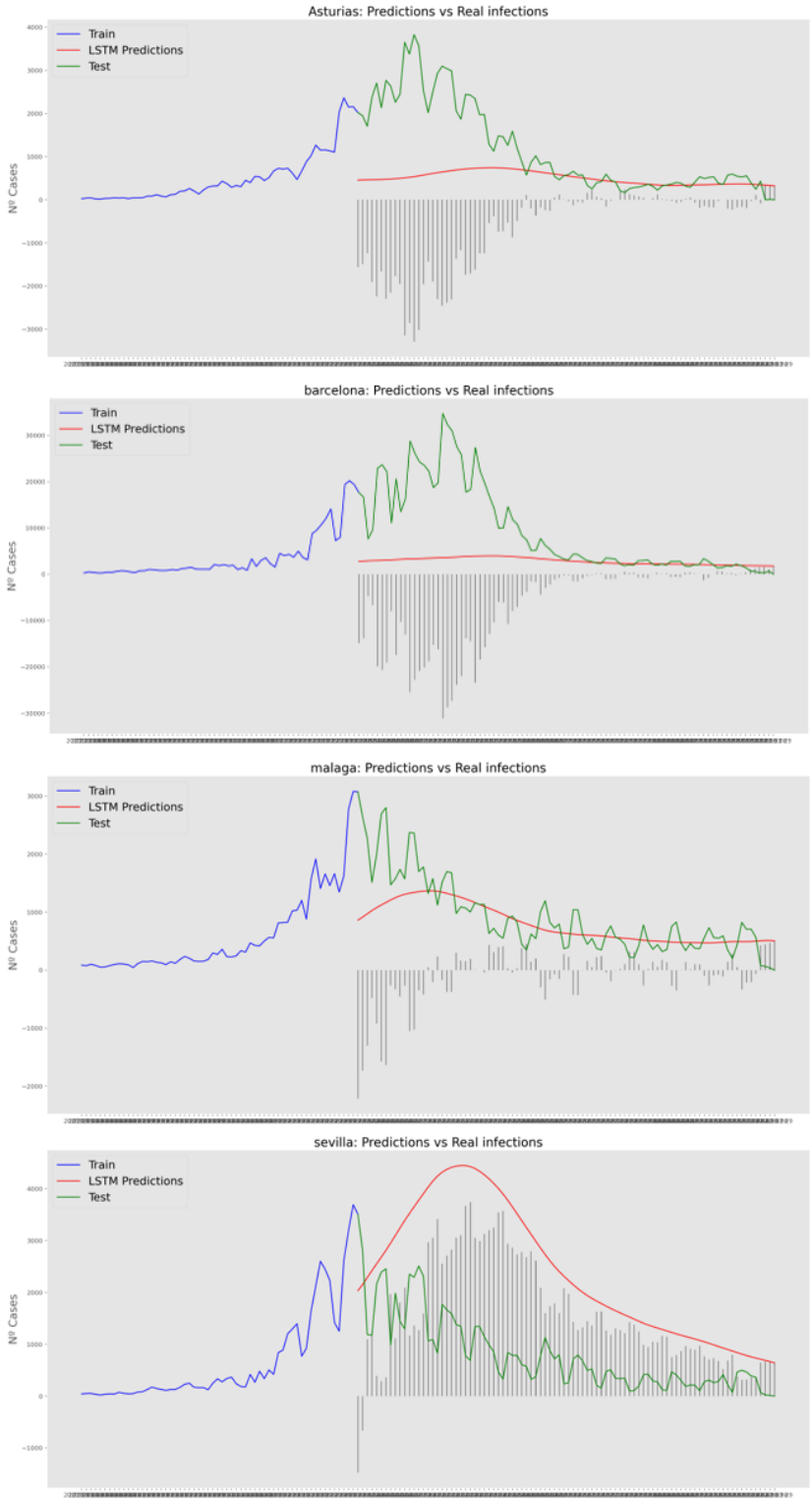

Figure 4.7: Multivariate LSTM model graphic 90 days results with all the data

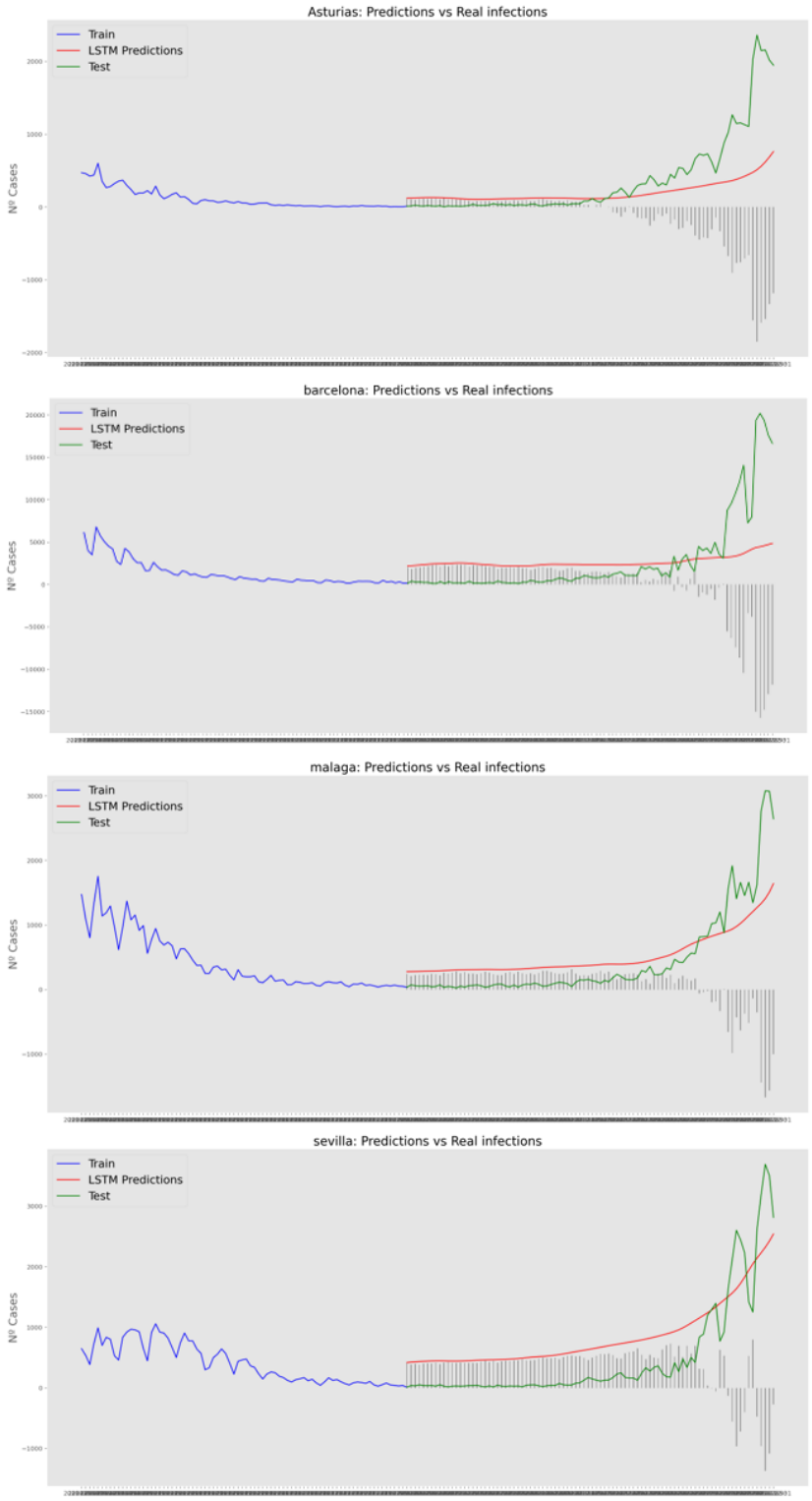

Figure 4.8: Multivariate LSTM model graphic 90 days results just before the 6th wave

Chapter 5

Conclusions

During the course of this project, the objectives set at the beginning have been met.

Firstly, a gathering and cleaning of information has been achieved by constructing a database prepared for further analysis.

In second place, relevant information has been identified regarding the evolution of the pandemic in five Spanish provinces through visual analysis and the application of specific methodologies.

Lastly, through the creation of machine learning models, it has been identified that the training of models considering meteorological data (average temperature) and mobility data, returns slightly more accurate predictions in general. However, the small margin of improvement obtained does not allow us to accept/generalise the assumption initially sought.

Of the machine learning models created, LSTM returns better results than ARIMA for identifying future peaks of infection during the pandemic. However, none of the models return reliable results, so it will be necessary to adjust the parameters of the models to improve the training process.

In conclusion, with the help of machine learning models it is possible to identify periods susceptible to a high level of contagion. However, their results have to be treated with caution due to their lack of precision. A sufficiently detailed mobility analysis, such as that provided by Google, should help to identify anomalous periods, although in the present analysis there has been no substantial improvement in the models created.

5.1 Future work

The application of the models to predict future waves has been compromised due to the change in policy towards Covid-19.

New measures for living with the virus have changed the situation. The treatment of covid

as a common disease will result in an unprecedented increase in the level of infections, however, thanks to the high level of vaccination many of the infections will not be counted due to the lack of symptoms. Many others will be unsupported due to lack of testing and PCR and some others will be camouflaged or confused with previous common diseases.

Therefore, a different approach to the problem will be necessary if we are to provide a tool that allows professionals to anticipate periods that may lead to a collapse of the system. Future studies are needed to adapt to the new situation.

Bibliography

[1] Azzam Abu-Rayash and Ibrahim Dincer. Analysis of mobility trends during the covid-19 coronavirus pandemic: Exploring the impacts on global aviation and travel in selected cities. *Energy research & social science*, 68:101693, 2020.

[2] García-Garrigós E Arenas-Jiménez JJ, Plasencia-Martínez JM. Cuando la neumonía no es covid-19. *Radiologia (Engl Ed)*, 63:180–192, 2021.

[3] Nikolaos Askitas, Konstantinos Tatsiramos, and Bertrand Verheyden. Lockdown strategies, mobility patterns and covid-19. 2020.

[4] Hamada S Badr, Hongru Du, Maximilian Marshall, Ensheng Dong, Marietta M Squire, and Lauren M Gardner. Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11):1247–1254, 2020.

[5] Armando Cartenì, Luigi Di Francesco, and Maria Martino. How mobility habits influenced the spread of the covid-19 pandemic: Results from the italian case study. *Science of the Total Environment*, 741:140489, 2020.

[6] Irida da Cunha. *El trabajo de fin de grado y de máster: redacción, defensa y publicación.* Editorial UOC, 2016.

[7] José María Loché Fernández-Ahúja and Juan Luis Fernández Martínez. Effects of climate variables on the covid-19 outbreak in spain. *International journal of hygiene and environmental health*, 234:113723, 2021.

[8] Md Sabbir Hossain, Sulaiman Ahmed, and Md Jamal Uddin. Impact of weather on covid-19 transmission in south asian countries: an application of the arimax model. *Science of The Total Environment*, 761:143315, 2021.

[9] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice.* OTexts, 2018.

[10] Rob J. Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 27:1–22, 2008.

[11] Cornelia Ilin, Sébastien Annan-Phan, Xiao Hui Tai, Shikhar Mehra, Solomon Hsiang, and Joshua E Blumenstock. Public mobility data enables covid-19 forecasting and management at local and global scales. *Scientific reports*, 11(1):1–11, 2021.

[12] Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Open COVID-19 Data Working Group†, Louis du Plessis, Nuno R Faria, Ruoran Li, et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

[13] Taiwo Temitope Lasisi and Kayode Kolawole Eluwole. Is the weather-induced covid-19 spread hypothesis a myth or reality? evidence from the russian federation. *Environmental Science and Pollution Research*, 28(4):4840–4844, 2021.

[14] Yen-Der Li, Wei-Yu Chi, Jun-Han Su, Louise Ferrall, Chien-Fu Hung, and T-C Wu. Coronavirus vaccine development: from sars and mers to covid-19. *Journal of biomedical science*, 27(1):1–23, 2020.

[15] Mattia Mazzoli, David Mateo, Alberto Hernando de Castro, Sandro Meloni, and José J Ramasco. Effects of mobility and multi-seeding on the propagation of the covid-19 in spain. may 2020.

[16] Hannah McClymont and Wenbiao Hu. Weather variability and covid-19 transmission: a review of recent research. *International journal of environmental research and public health*, 18(2):396, 2021.

[17] Clifton McPherson, Richard Chubet, Kathy Holtz, Yoshikazu Honda-Okubo, Dale Barnard, Manon Cox, and Nikolai Petrovsky. Development of a sars coronavirus vaccine from recombinant spike protein plus delta inulin adjuvant. In *Vaccine Design*, pages 269–284. Springer, 2016.

[18] Pierre Nouvellet, Sangeeta Bhatia, Anne Cori, Kylie EC Ainslie, Marc Baguelin, Samir Bhatt, Adhiratha Boonyasiri, Nicholas F Brazeau, Lorenzo Cattarino, Laura V Cooper, et al. Reduction in mobility and covid-19 transmission. *Nature communications*, 12(1):1–9, 2021.

[19] Joseph SM Peiris, Kwok Y Yuen, Albert DME Osterhaus, and Klaus Stöhr. The severe acute respiratory syndrome. *New England Journal of Medicine*, 349(25):2431–2441, 2003.

[20] Mehmet Şahin. Impact of weather on covid-19 pandemic in turkey. *Science of the Total Environment*, 728:138810, 2020.

[21] Ramadhan Tosepu, Joko Gunawan, Devi Savitri Effendy, Hariati Lestari, Hartati Bahar, Pitrah Asfian, et al. Correlation between weather and covid-19 pandemic in jakarta, indonesia. *Science of the total environment*, 725:138436, 2020.

[22] Lin-Man Weng, Xuan Su, and Xue-Qiang Wang. Pain symptoms in patients with coronavirus disease (covid-19): A literature review. *Journal of Pain Research*, 14:147, 2021.

[23] Z Zhao, F Zhang, M Xu, K Huang, W Zhong, W Cai, Z Yin, S Huang, Z Deng, M Wei, et al. Description and clinical treatment of an early outbreak of severe acute respiratory syndrome (sars) in guangzhou, pr china. *Journal of medical microbiology*, 52(8):715–720, 2003.