Universitat Oberta de Catalunya (UOC)

Master Degree - Data Science

Master Thesis

Area 5: Healthcare and Environment

COVID-19: Prediction of infections based on the mobility trends reported by mobile phone networks in Spain

PEC3 - Design and Implementation

Autor: Álvaro Rodríguez Sans

Tutor: Carlos Luis Sánchez Bocanegra

Tutor: Rafael Pastor Vargas

Professor: Albert Solé Ribalta

Delivery PEC3 - 23/05/2021

Copyright

WORK SHEET

Tittle of the master thesis:	COVID-19: Infection prediction based on the mobility reported by mobile phone networks in Spain
Autor's name:	Álvaro Rodríguez Sans
Tutor's name:	Carlos Luis Sánchez Bocanegra
Tutor's name:	Rafael Pastor Vargas
PRA's name:	Albert Solé Ribalta
Delivery date (mm/yyyy):	06/2021
Degree:	Master Degree – Data Science
Area of work:	Area 5 – Healtcare and Environment
Language:	English
Keywords:	COVID-19, Data-Mining, Deep-Learning, Mobility

Abstract

The pandemic caused by the virus called COVID-19, has lead our society to face new paradigms and challenges from healthcare and social life perspectives. Healthcare systems around the world face extreme situations and general population is impacted in their life-style / behaviours (e.g. mobility) since this new virus appears.

Bearing in mind that we live in a world completely connected on a physical and virtual levels, it is important to understand how people's mobility may or not affect a greater spread of this virus. Our society (individuals and public / private organizations) request tools and best practices that help in predicting a potential number or percentage of the population susceptible to being infected by this and other types of virus with similar behaviour. Planning human and material resources of health-care systems has become a priority.

There are a large number of data sources available that can be consumed, crossed and analysed to see the spread of the disease considering different dimensions. In Spain, the National Statistics Institute (INE - by its acronym in Spanish), as well as the different regional governments and other organizations, publish this type of data.

The present study / work seeks to understand if mobility is a relevant element in the spread of this virus in Spain using mobility data in combination with the virus evolution information. Data mining techniques, traditional statistical methods and neural networks are used to confirm this hypothesis and establish a predictive model of COVID-19 based on mobility.

Keywords: COVID-19, Data-Mining, Deep-Learning, Mobility

Resumen

La pandemia provocada por el virus COVID-19 ha llevado a nuestra sociedad a tener que afrontar nuevos paradigmas y retos desde la perspectiva de la salud y en normal desarrollo de la vida cotideanea. Los sistemas de salud de todo el mundo se enfrentan situaciones extremas y la población en general se ve afectada en su estilo de vida / comportamientos (por ejemplo, movilidad) desde que este nuevo virus irrumpio en nuestras vidas.

Teniendo en cuenta que vivimos en un mundo que está completamente conectado tanto a nivel físico como virtual, es importante comprender cómo la movilidad de las personas puede o no afectar a una mayor propagación de este virus. Nuestra sociedad (individuos y organizaciones tanto públicas como privadas) requieren de herramientas y buenas prácticas que ayuden a predecir el número o porcentaje potencial de población susceptible de ser infectada por este y otros tipos de virus con comportamiento similar. La planificación de los recursos humanos y materiales de los sistemas de salud se ha convertido en una prioridad para todos los entes sociales.

Existen una gran cantidad de fuentes de datos disponibles que se pueden consumir, cruzar y analizar para ver la propagación de la enfermedad considerando diferentes dimensiones. En España, el Instituto Nacional de Estadística (INE), así como los diferentes gobiernos autonómicos y otros organismos, publican este tipo de datos.

El presente estudio / trabajo busca comprender si la movilidad es un elemento relevante en la propagación de este virus en España. Se utilizan datos de movilidad en combinación con la evolución del virus. Se utilizan técnicas de minería de datos, métodos estadísticos tradicionales y redes neuronales para confirmar esta hipótesis y establecer un modelo predictivo de COVID-19 basado en la movilidad.

Palabras clave: COVID-19, Minería de Datos, Aprendizaje Profundo, Movilidad

Contents

Abstract v

Resumen vii

Table of contents ix

List of figures xi

List of tables 1

1 Definition and planning 3

1.1 Description, interest and relevance of the proposal . . . 3
1.2 Objectives . . . 4
1.3 Methodology to be used . . . 4
1.4 Planning . . . 5

2 State of the art 9

2.1 Machine learning process - Action plan . . . 9
2.1.1 Problem understanding . . . 10
2.1.2 Data understanding . . . 10
2.1.3 Data preparation . . . 10
2.1.4 Modelling . . . 11
2.1.5 Evaluation . . . 11
2.1.6 Deployment . . . 12
2.2 COVID-19 . . . 13
2.2.1 Global spread . . . 13
2.2.2 Spain - Spread and lessons learned . . . 15
2.2.3 Spread study based on mobility . . . 17
2.3 Mobility trends - Spain (INE / Google) . . . 19
2.4 COVID-19 - Machine and deep learning techniques . . . 20

2.4.1 Autoregressive Integrated Moving Average (ARIMA) 21
2.4.2 Long-Short Term Memory (LSTM) 22
2.4.3 Studies carried-out 26
2.5 Datasets to be used 27
2.6 Data-science IDE and language to be used 28

3 Methodology **29**
3.1 Steps followed 29
3.1.1 Domain Study - Bibliographic research 30
3.1.2 Data selection 30
3.1.3 Data preprocessing 30
3.1.4 Dimensionality reduction 32
3.1.5 Selection of the discovery goal 33
3.1.6 ARIMA 34
3.1.7 LSTM 37
3.1.8 Results assessment 37
3.1.9 Conclusions 37

Bibliography **37**

A Code used **45**

List of Figures

1.1 Droplets spread. Source: Morawska and Cao [35] . . . 3
1.2 Gantt diagram . . . 7

2.1 CRIPS-DM. Source: Wirth, [56] . . . 9
2.2 End to end process. Source: Treveil, [49] . . . 13
2.3 Covid-19 transmission. Source: Dhama et al., [14] . . . 14
2.4 Covid-19 transmission march 2021. Source: WHO [55] . . . 14
2.5 Covid-19 evolution until march 2021 Spain. Source: WHO [55] . . . 15
2.6 Mobility madrid (from - to). Source: Mazzoli et al., [33] . . . 16
2.7 Multivariate analysis. Source: Mazzoli et al., [33] . . . 16
2.8 Correlation (mobility - virus spread). Source: The Lancet [4] . . . 17
2.9 Spain mobility maps. Source: INE [22] . . . 19
2.10 Google mobility trends. Source: Google [19] . . . 20
2.11 ARIMA forecast. Source: Taylor and Letham [48] . . . 22
2.12 Neuron. Source: Ciaburro and Venkateswaran [9] . . . 23
2.13 Artifical neuron. Source: Vasilev et al., [50] . . . 23
2.14 Artifical neuron multilayer. Source: Vasilev et al., [50] . . . 24
2.15 LSTM. Source: Vasilev et al., [50] . . . 24
2.16 LSTM India prediction. Source: Rauf et al., [43] . . . 26
2.17 SLSTM India MAPE. Source: Devaraj et al., [12] . . . 27
2.18 LSTM comparative - 2. Source: Shahid, Zameer and Muneeb, [47] . . . 27

3.1 Methodology used . . . 29
3.2 Google - Change residential . . . 32
3.3 Correlation observed - Barcelona . . . 33
3.4 PCA - Variance explained - Barcelona . . . 33
3.5 Barcelona - Seasonality / Trend - STL . . . 35
3.6 Barcelona - Residuals - double difference . . . 35

3.7 Barcelona - Accuracy univariate (14 days) based on errors 36
3.8 Barcelona - Accuracy multivariate (14 days - All) based on errors 36
3.9 Barcelona - Multivariate (14 days - All) forecast plot 36
3.10 Sevilla - Accuracy univariate (14 days) based on errors 37
3.11 Sevilla - Univariate (14 days) forecast plot 37

List of Tables

1.1 Pec1 plan 5
1.2 Pec2 plan 5
1.3 Pec3 plan 6
1.4 Pec4 plan 6
1.5 Pec5 plan 6
1.6 Pec6 plan 6

Chapter 1

Definition and planning

1.1 Description, interest and relevance of the proposal

COVID-19 is an infectious disease caused by a newly discovered coronavirus. In one hand, the majority of people affected by this virus will experience from zero to slight-moderate respiratory symptoms / illness (similar to the seasonal flu) with, apparently, no side effects. On the other hand, people +65 years and those with previous / chronic medical diseases (e.g. diabetes, respiratory disease and / or cancer) it has been revealed as the risk group that can develop a serious illness and lead to death or important sequelae like altered cognition, pulmonary function abnormalities, etc. [8, 51].

The virus spreads primarily through two main ways. First, **droplets** (both large and small) of saliva or discharge from the nose when a person coughs or sneezes. Transmission through airborne of smaller droplets / particles is included into this category due to those small particles have the ability to remain suspended more time and go further compared with droplet transmission (Figure 1.1). Second, contact transmission, when direct contact with an infected person or contaminated surface takes place [34, 35, 36].

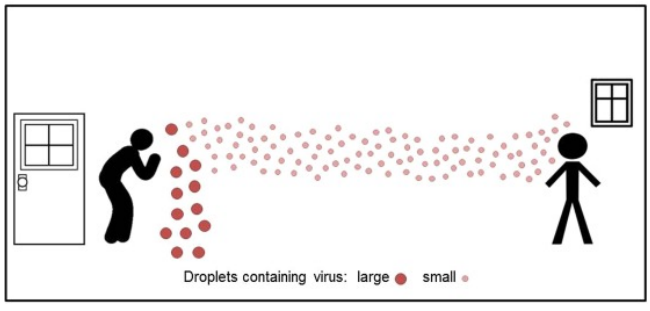


Figure 1.1: Droplets spread. Source: Morawska and Cao [35]

There is a real need to predict the virus spread based on new dimensions. Mobility has been considered as one of the most important due to restrictions adopted by governments around the world. Several data sources offers up-to-date information related to the number of infections, deceased and recovery population in Spain [10, 25], mobility data collected from mobile telephony companies [22] and big-tech providers that publish mobility trends thanks to the massive usage of their mobile applications and devices [3, 19].

The main objective of this work is to offer a series of best practises on how to analyse the importance of population mobility spreading the virus and how this patterns can be predicted thanks to the usage of data-mining and deep-learning techniques [2, 5, 13, 30, 31, 40].

1.2 Objectives

The primary objectives are:

- Understand the infective behaviour of the COVID-19.
- Identify relevant data, provided by mobile telephony providers, to predict the spread of COVID-19 in Spain.
- Use **data-mining and machine learning** techniques to assess the impact / influence of population mobility patterns spreading COVID-19 in Spain.
- Articulate the general best practices to be used in order to afford such scenarios.

The secondary objectives are to apply:

- The necessary **data-mining** techniques to find, create join and / or discard the raw data needed and provided by different organizations (public and private), performing the necessary **exploratory data analysis and transformation** to extract initial insights.
- Different **machine learning** techniques to assess which one produce the better performance / accuracy predictions.

1.3 Methodology to be used

Research into the appropriated academic and scientific databases, journals and books to understand COVID-19 behaviour and how it can be combined with Data-Science techniques in order to predict its spread based on mobility by:

- Looking for the relevant data related to mobility and COVID-19 in Spain (public and private sources of data), to understand its behaviour, how to combine it / use it, etc.
- Analysing / exploring the art state of the data-mining and machine-learning techniques and tools to extract the insights from data. The selection of tools / program languages to elaborate the necessary exploratory data analysis during development phase.
- The usage of **CRISP-DM** [17, 18, 37, 56], **PDSA** [15] and **Agile** [57] methodologies, among others, to meet previous stages from a project management perspective.

Based on the methods stated and thanks to the analysis / review of the models applied, we will confirm the initial hypothesis: "Mobility patterns affect / impact virus spread".

1.4 Planning

The following tables (Table 1.1, 1.2, 1.3, 1.4, 1.5 and 1.6) and Figure 1.2 shows a summary of the initial list of the different activities to be delivered in each PEC. This list will be enriched with further additions or modifications.

Name	Duration	Start	Finish
Pec1 - Definition and planning	12 days	Wed 17/02/21	Sun 28/02/21
- Topic selection and initial research	3 days	Wed 17/02/21	Fri 19/02/21
- Description, interest and relevance of the proposal	2 days	Sat 20/02/21	Sun 21/02/21
- Objectives and personal motivation	2 days	Mon 22/02/21	Tue 23/02/21
- Methodology to be used and planning	2 days	Wed 24/02/21	Thu 25/02/21
- Abstract preparation + Pec1 delivery	3 days	Fri 26/02/21	Sun 28/02/21

Table 1.1: Pec1 plan

Name	Duration	Start	Finish
Pec2 - State of the art / Market analysis of the project	21 days	Mon 01/03/21	Sun 21/03/21
- Research - COVID-19 spread	4 days	Mon 01/03/21	Thu 04/03/21
- Research - Data-Mining techniques (Clean-up, EDA, etc.)	4 days	Fri 05/03/21	Mon 08/03/21
- Research - Machine Learning (forecast time series, neural network, etc.)	3 days	Tue 09/03/21	Thu 11/03/21
- Research / review existing jobs that address the objective	3 days	Fri 12/03/21	Sun 14/03/21
- Incorporate new ideas / refine current approach - Pec2 delivery	7 days	Mon 15/03/21	Sun 21/03/21

Table 1.2: Pec2 plan

Name	Duration	Start	Finish
Pec3 - Design and implementation	63 days	Mon 22/03/21	Sun 23/05/21
- Program tool + lenguage selection (R / Python)	5 days	Tue 23/03/21	Sat 27/03/21
- Data gathering + Clean-up + EDA + Transformations	27 days	Sun 28/03/21	Fri 23/04/21
- Models review	5 days	Sat 24/04/21	Wed 28/04/21
- Models selection	17 days	Thu 29/04/21	Sat 15/05/21
- Prediction analysis	5 days	Sun 16/05/21	Thu 20/05/21
- Review + Pec3 delivery	3 days	Fri 21/05/21	Sun 23/05/21

Table 1.3: Pec3 plan

Name	Duration	Start	Finish
Pec4 - Report writing	14 days	Mon 24/05/21	Sun 06/06/21
- Review documentation generated	4 days	Mon 24/05/21	Thu 27/05/21
- Conclusions from of the results	5 days	Fri 28/05/21	Tue 01/06/21
- Write master thesis (official format)	5 days	Wed 02/06/21	Sun 06/06/21

Table 1.4: Pec4 plan

Name	Duration	Start	Finish
Pec5 - Presentation, exposure and defence of the project	8 days	Mon 07/06/21	Mon 14/06/21
- Generate video + presentation	6 days	Mon 07/06/21	Sat 12/06/21
- Exposure of the project	2 days	Sun 13/06/21	Mon 14/06/21

Table 1.5: Pec5 plan

Name	Duration	Start	Finish
Pec6 - Public exposure	1 day	Sat 19/06/21	Sat 19/06/21
- Public exposure of the project	1 day	Sat 19/06/21	Sat 19/06/21

Table 1.6: Pec6 plan

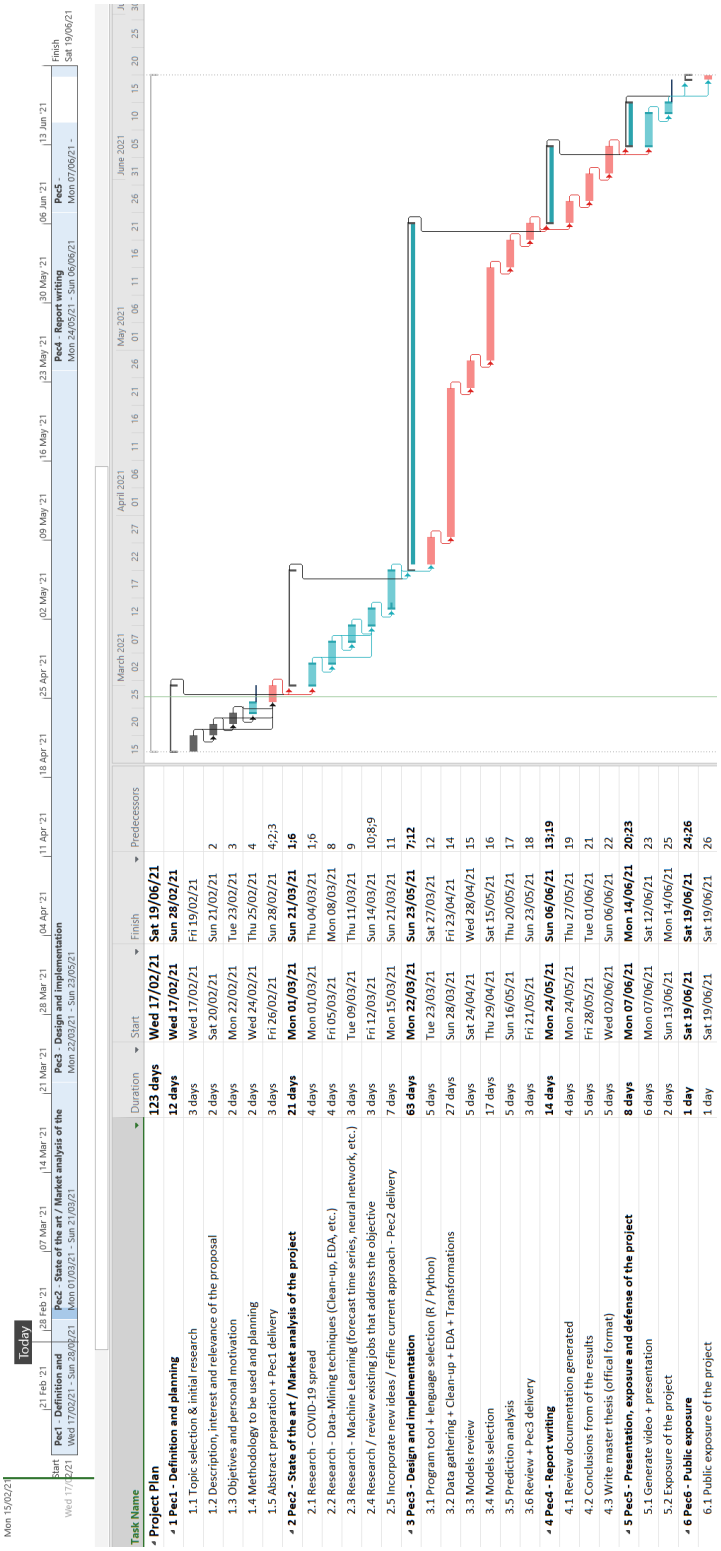


Figure 1.2: Gantt diagram

Chapter 2

State of the art

2.1 Machine learning process - Action plan

When using **Machine Learning**, it is convenient to follow a process to obtain good results, make better use of our time and have guidance on what to do in case our results are not as good as expected. Figure 2.1 and following points describes the different phases of the machine learning process and how they interact each other based on the **Cross Industry Standard Process for Data Mining** (CRISP-DM) standard [17, 18, 37, 56].

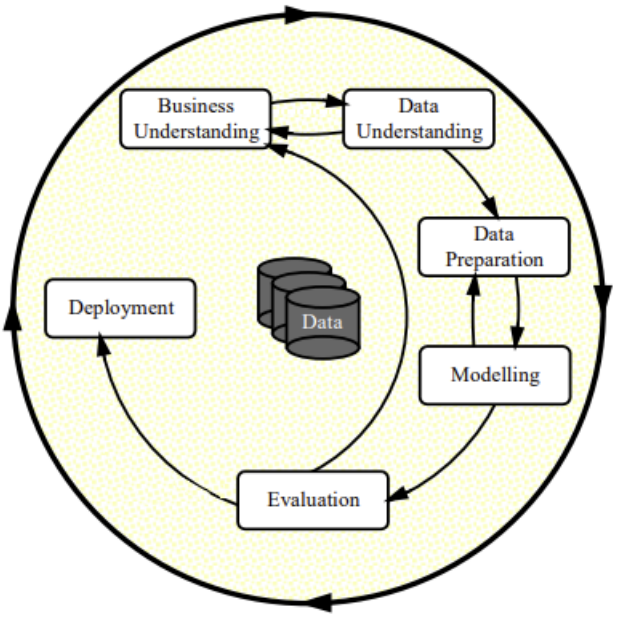


Figure 2.1: CRIPS-DM. Source: Wirth, [56]

2.1.1 Problem understanding

It is very important to understand the problem to solve. Understanding the problem usually takes a long time, especially if the problem comes from an industry or field where the analyst have little knowledge [2, 11, 18, 30, 31].

This task has a medium relative effort due to in this phase the idea is to collaborate with people who know about the problem (SME - Subject Matter Expert).

To get a better understanding of the problem is to ask why? several times until the answer satisfies the question. Knowing the why of things helps to understand the way of thinking / behaviour in that industry or field.

2.1.2 Data understanding

As important as understanding the problem is understanding the data we have available. It is common to do an **exploratory analysis of data** (EDA) to familiarize ourselves with them [2, 11, 18, 30, 31].

In exploratory analysis graphs, correlations and descriptive statistics are usually made to better understand what story the data is telling us. It also helps to estimate if the data we have is sufficient, and relevant, to build a model.

2.1.3 Data preparation

Data preparation is one of the phases in machine learning that requires a lot attention and effort. Main challenges to be addressed are the following [2, 11, 18, 30, 31]:

- **Incomplete data** - It is quite normal not having all the data needed (e.g. if is necessary to predict which customers are more likely to buy a product and the data comes from online surveys, there will be many of the fields needed not filled). In order to deal with incomplete data these are some of the strategies to be adopted:
  - **Eliminate** - Easy option it is just keep the complete data. This may be an option dealing with few incomplete data.
  - **Impute** - Reasonable value will be added / imputed if makes sense (e.g. if someone didn't add his / her age in a survey, the average age of the other participants can be used).

  - **Do nothing** and use some machine learning technique that can handle incomplete data.

- **Combine data from multiple / different sources** - Some data can come from databases, spreadsheets, flat files, etc. It is necessary to combine data to let machine learning algorithms consider / manage all the information.

- **Format data properly** - The usage of machine learning libraries implies to pass the data in a format the can understand. In general, these libraries expect the data to be in the form of a matrix or a tensor. A tensor is a generalization of a matrix. If the matrix has 2 dimensions, the tensor has a number "n" of dimensions.

- **Calculate relevant characteristics (features)** - Machine learning algorithms work much better if they can work only with relevant features instead of the raw data. This phase requires a lot of effort. It is necessary identify which features are going to be relevant to solve the problem and test it.

- **Data normalization** - It is useful to normalize the data to make learning easier for machine learning. Normalizing is the act of putting all the data on a similar / same scale. There are several ways to normalize the data.

2.1.4 Modelling

The phase of building a machine learning model, once we have the data ready, requires less effort. This is due to there are already several machine learning libraries available and a lot of them are free and open source.

During this phase, it is necessary choose the machine learning technique to be used. The machine learning algorithm will automatically learn to obtain the appropriate results with the historical data that has been prepared in previous phases. Of course, it will have an error [2, 11, 18, 30, 31].

2.1.5 Evaluation

The error analysis requires a medium relative effort. Analyzing errors (evaluation) is important to understand what to do to improve machine learning results. In particular the options will be:

- Use another more **complex** model.

- Use a **simpler** one.
- Review if more data and / or more **features** are needed.
- Review if a **better understanding** of the problem and / or the next steps needed during the whole process.

In the evaluation phase we will try to ensure that our model is capable of generalization. Generalization is the ability of machine learning models to produce good results when using new data.

In general, it is not difficult to achieve acceptable results using this process. However, to get really good results over the time, it is mandatory to iterate over the previous phases several times. With each iteration, the understanding of the problem and the data will increase. This allows to design better relevant features and reduce generalization error. A greater understanding will also offers the possibility to choose more precisely the machine learning technique / model that best suits the problem to be addressed [2, 11, 18, 30, 31].

The majority of the situations states that having more data helps to increase the performance of the model. In practice, more data and a simple model tend to perform better than a complex model with less data.

2.1.6 Deployment

Once the evaluation phase is completed and the model error is acceptable, it is a good practice to compare it with the error in the current solution (if is available). If it is better enough, the machine learning model will be deployed / integrated into the current system [49].

The integration phase of a machine learning model into a live system requires a relatively greater effort due to:

- It is needed to **automatically repeat** the data preparation stages. This requires the machine learning model have to communicate with other parts of the system and the results of the model have to be used by the system.
- In addition, **monitoring** the errors of the model and warn if model errors grow over time to rebuild the machine learning model with new data.

A considerable part of the effort it is consumed building the data interfaces necessary that provides data to the model in an automated way and then the system can use its prediction

automatically (Figure 2.2). For machine learning and artificial intelligence to be useful, in most cases they must be integrated into a larger system (e.g. an online translation system would not be very useful if had only built a machine learning model. This would be fine for academic proposes but will not have no commercial value unless it will be implemented. The real value is that the model is integrated into the system and allows multiple end users to use its machine learning model capabilities to translate sentences).

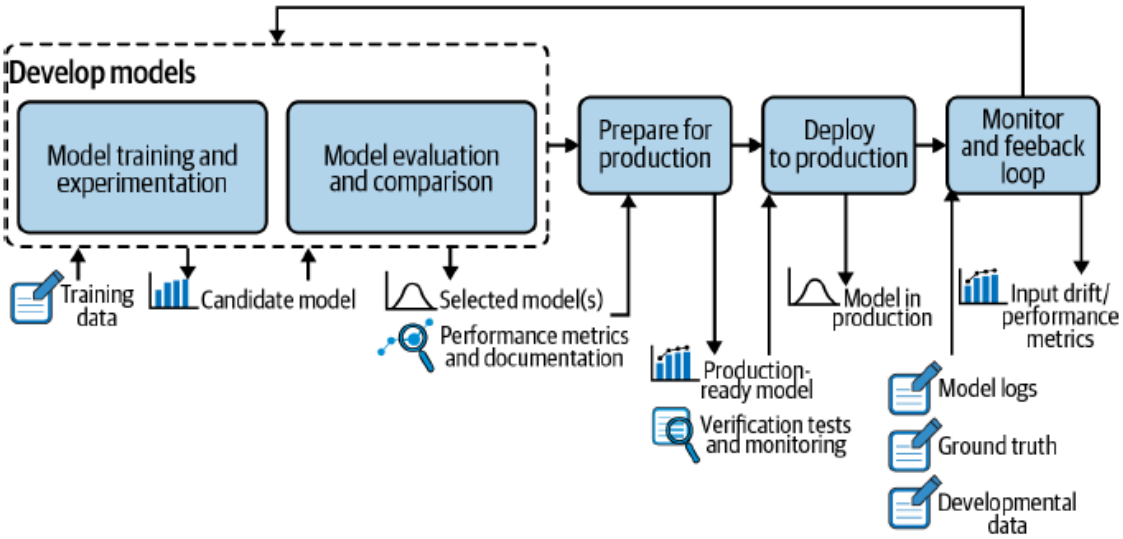


Figure 2.2: End to end process. Source: Treveil, [49]

2.2 COVID-19

2.2.1 Global spread

In early December 2019, local healthcare authorities in the city of **Wuhan** (China), were surprised by a new pneumonia disease of unknown origin. This new pneumonia had a great facility for its spread and consequently was easy to find parallelism with the previous epidemics, like the **Severe Acute Respiratory Syndrome Coronavirus** (SARS-CoV - 2003) [6] and the **Middle East respiratory syndrome** (MERS - 2012) [7]. It is important to remark that the new pneumonia is / was responsible to cause more deaths, although it is a virus with lower lethality rates [46] compare to other similar virus.

It is highly suspected that bats coronaviruses may have led to the evolution of COVID-19 and its introduction into humans [14]. Due to the fast spread of this new virus the **World Health Organization** (WHO) started to activate the different protocols available with the

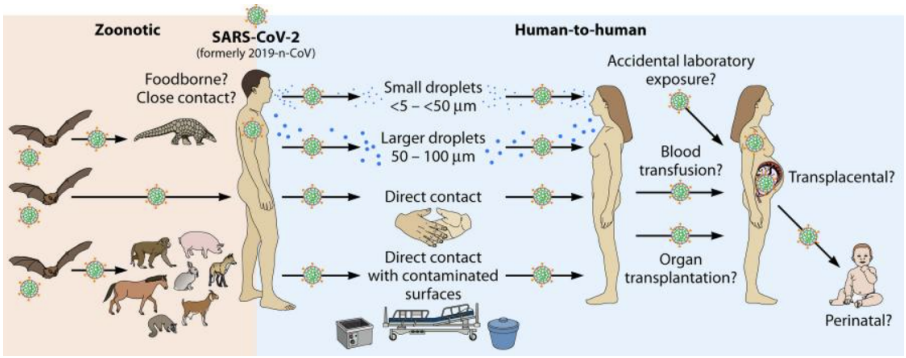


Figure 2.3: Covid-19 transmission. Source: Dhama et al., [14]

aim to contain its expansion in January 2020. During these days the number of new cases reported increases alarmingly [54]. **In March 2020 WHO** declare the global pandemic, reported a grand total of **132,000 cases** of COVID-19, from **123 countries** and territories. **In April 2020, 1.000.000 million of new cases** were reported and more than **50.000 deaths**. Unfortunately these figures experienced a grow until a grand total in **March 2021 of confirmed 115,653,459 cases and 2,571,823 of deaths** [55] around the world.

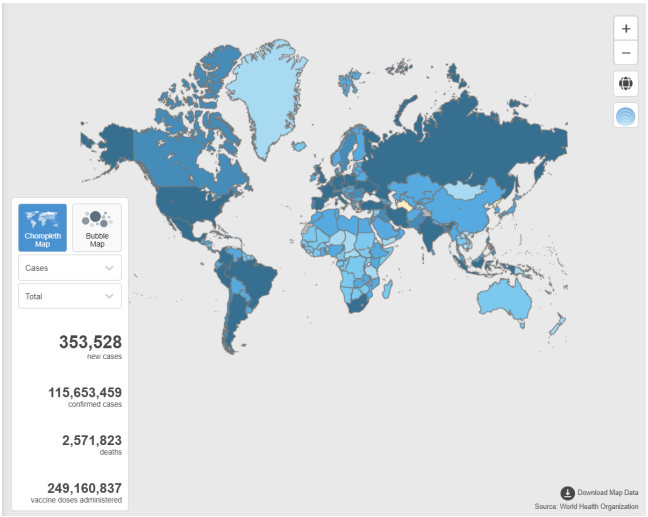


Figure 2.4: Covid-19 transmission march 2021. Source: WHO [55]

Along with the rapid spread of the virus worldwide, new variants have emerged (UK, South African, etc.). Even it has not been demonstrated these new variants present relevant higher levels of contagion, greater impact on the health and recovery of infected patients or more mortality rates, healthcare authorities consider that more investigation should be done [32].

To avoid these new variants adds more uncertainty to the current global / local situations, healthcare authorities recommends to all countries and parties involved in control and promote

local and international travels, to provide the necessary information related to the current status and the measures to follow in order to kept population safe [53].

2.2.2 Spain - Spread and lessons learned

The first case of COVID-19 in Spain was detected on the island of La Gomera - Canarias (January 2020), and it was considered as an imported case from Germany. In February 2020, the first cases were reported in mainland Spain. In May 2020, a grand total of 237,906 confirmed cases were reported and the virus was responsible of 27,119 deaths [46]. Figures in Spain experienced a grow until a grand total in March 2021 of confirmed 3,142,358 cases and 70,501 of deaths [55].

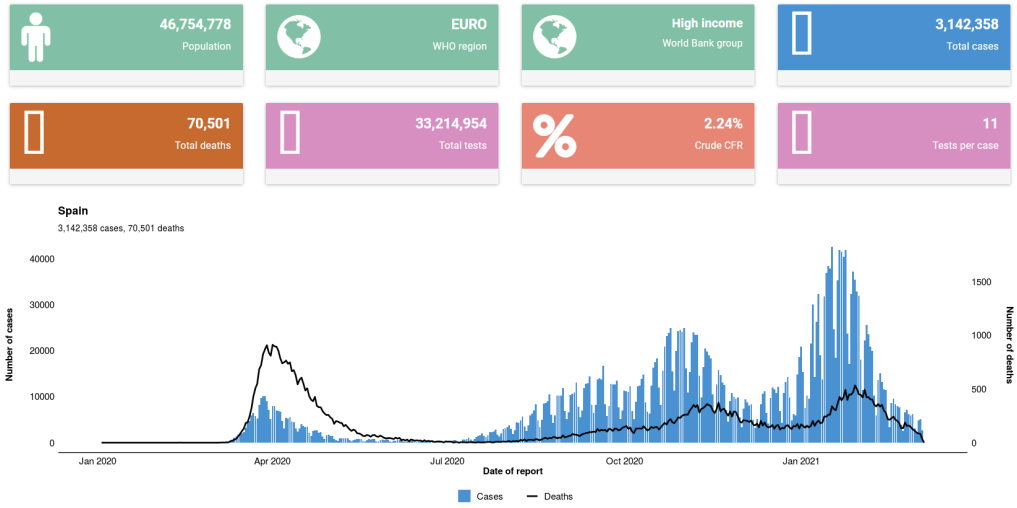


Figure 2.5: Covid-19 evolution until march 2021 Spain. Source: WHO [55]

It is necessary to take into consideration that until all the vaccines [52] will not be available for distribution worldwide, measures that helps to find a balance between daily behaviour and the maintenance of adequate self-protection measures that offers a rapid response to any new outbreak should be kept (usage of masks, social distance, etc.) [46].

A recent study [33], **focused on mobility trends in Spain, reveals that a proper assessment of mobility is crucial in order to understand the effects of measures** (quarantines, selective re-openings, etc.) containing the virus. High levels of mobility contribute to virus spread, where "multi-seeding" (consider as independent infected individuals arriving at a new region or city) can be consider a potential to boost new local infections, and could be impact negatively over the effectiveness of tracing measures. In Spain, multi-seeding, could be considered as a key player spreading COVID-19.

The study indicates that peaks in number of infections and mortality rates are highly correlated with mobility (from and to) occurred in Madrid, city consider the "hub" in Spain due to, as capital, attracts workers, students, visitors all around the country and daily commuters from neighbour provinces / cities. [33].

Figures 2.6 and 2.7 offers a graphical representation of the conclusions extracted by this study.

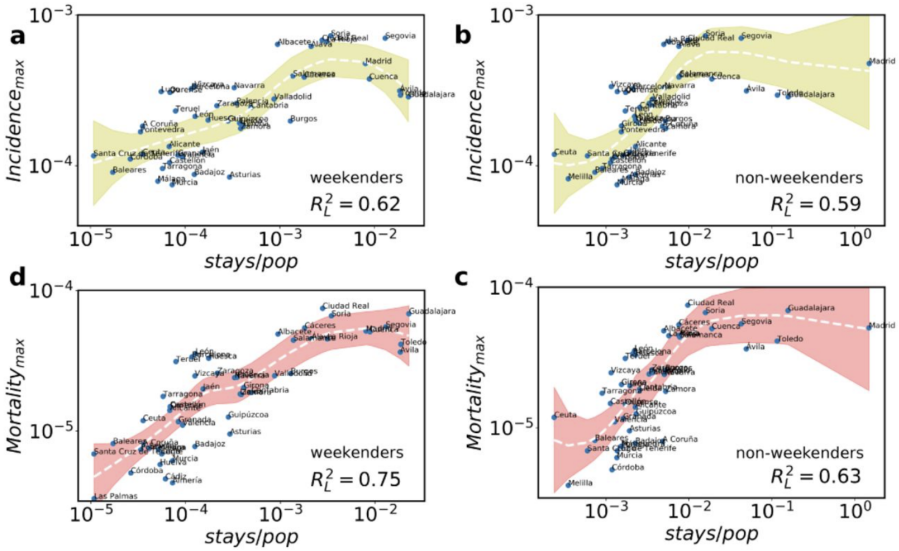


Figure 2.6: Mobility madrid (from - to). Source: Mazzoli et al., [33]

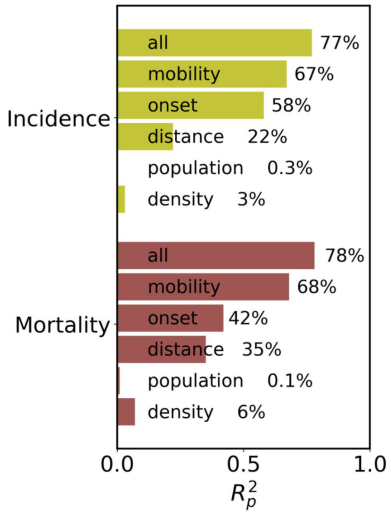


Figure 2.7: Multivariate analysis. Source: Mazzoli et al., [33]

2.2.3 Spread study based on mobility

In USA a study performed in 2020, based on daily mobility data (aggregated and anonymised) from January 2020 to April 2020, was used to retrieve trends in movement patterns for each US state under observation. Social distancing metrics were generated in combination with epidemiological data in order to compute COVID-19 growth rate, for a given state on a given day. The study was able to evaluate how social distancing (measured by relative change in mobility patterns), impacted the rate of new infections in all USA states under evaluation [4].

The **analysis concluded that mobility patterns have a strong correlation with the decreased of COVID-19 infections for the states with more incidence**. The effects of changes in mobility, are not visible until 9 to 12 days, which is consider the time needed to incubate and report the virus infection to the healthcare system [4].

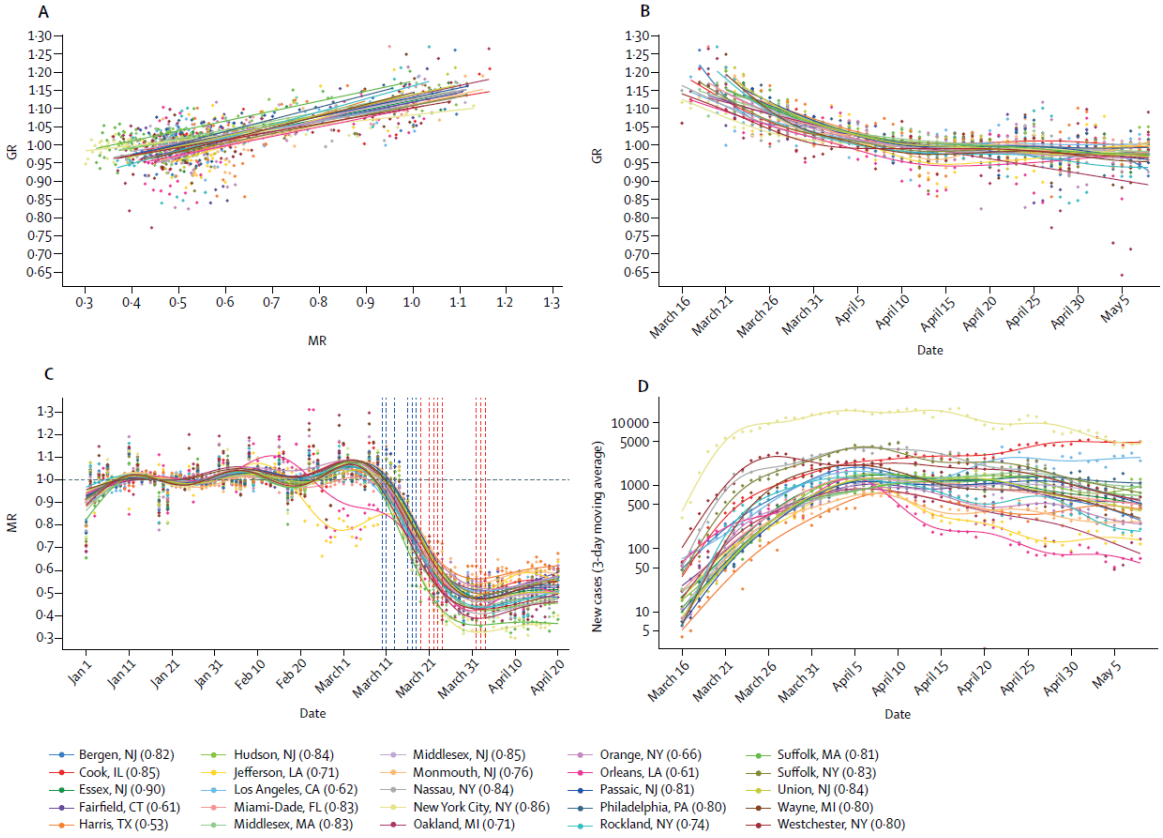


Figure 2.8: Correlation (mobility - virus spread). Source: The Lancet [4]

One of the key elements this study offer, is the inclusion of a quantitative dimension for analysis adding new metrics as MR (mobility ratio) and GR (growth ratio of the COVID-19). For MR (formula below), it was considered the difference between trips / movements (from - to and into the state) compared with a base reference date (before the virus outbreak) by a

given date. Here, i and j are the states, V is the number of commutes / trips, t is the given date and $t0$ is the reference date [4].

$$MR_j^t = \frac{\sum_{i \neq j} V_{ij}^t + \sum_{i \neq j} V_{ji}^t + V_{jj}^t}{\sum_{i \neq j} V_{ij}^{t0} + \sum_{i \neq j} V_{ji}^{t0} + V_{jj}^{t0}}$$

For the GR (formula below), it was considered the difference of cases reported in a period of 3 days compared with the cases reported in a period of 7 days. Here, j is the state, t is the given date, i is the number of cases reported, V is the number of trips / movements and $t0$ is the reference date [4].

$$GR_j^t = \frac{\log\left(\sum_{i-3}^{t} \frac{C_j^t}{3}\right)}{\log\left(\sum_{i-7}^{t} \frac{C_j^t}{7}\right)}$$

Thanks to the introduction of this approach and formulas, it was possible to demonstrate that there is a strong correlation between mobility and virus spread (Figure 2.10). Other studies in USA have adopted this approach trying to find out new indexes, like SDI (Social Distance Index), where mobility is also the foundation to discover correlations between mobility and virus spread. Here, five dimensions (*X1* - Percentage of residents staying home, *X2* - Daily work trips per person, *X3* - Daily non-work trips per person, *X4* - Distances travelled per person and *X5* - Out-of-county trips -in thousands-) were combined in a formula that states the level of social distance achieve by population in percentage [38].

$$SDI = [\beta 1 * X1 + 0.01 * (100 - X1) * (\beta 2 * X2 + \beta 3 * X3 + \beta 4 * X4)] * (1 - \beta 5) + \beta 5 * X5$$

In summary, according the authors, the formula above, was divided into two main blocks, the first ones it is focused on residential and inside trips and the second one on out of county trips. For the first one, it has been used the percentage of residents staying home (residents - trips less than 1.61km from home) so the weight is one ($\beta 1 = 1$). For the ones not staying at home, the percentage is 100 minus X1. For the individuals with more work and non-work trips and longer distances, the assumption is that the are practising less social distancing. Then the weights for each variable should sum up to one ($\beta 2 + \beta 3 + \beta 4 = 1$). That it is meant the resident travellers are comparable to residents staying at home [38].

The assignment of the appropriate weights to each variable, were based on actual observations and conceptual guidelines offered by federal agencies. The relative ratio between resident trips and out-of-county trips was four to one, it was assign a weight of 0.2 to $\beta 5$ [38].

2.3 Mobility trends - Spain (INE / Google)

In 2019, the Statistical National Institute (INE - by its acronym in Spanish), carried out an initial mobility study based on mobile telephony and due to the outbreak of COVID-19 (March 2020 in Spain), the study was extended to measure mobility during the alarm state and onwards [22, 23, 25].

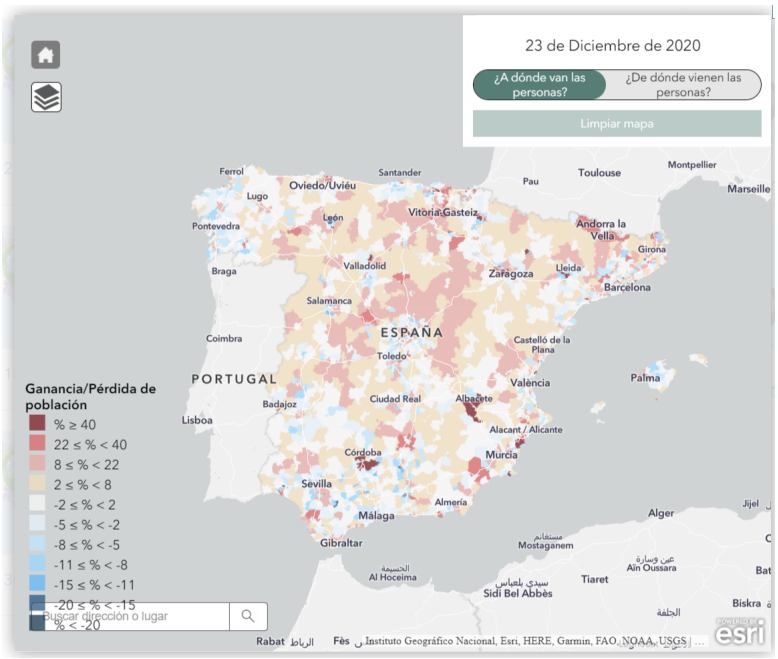


Figure 2.9: Spain mobility maps. Source: INE [22]

The study is based on mobile phones part of the national telephone numbering plan in Sapin (foreign phone lines are excluded). Data obtained is aggregated to population totals. The geographical scope is the entire national territory which is divided into 3,214 specific "mobility areas", each one consisting of a minimum of 5,000 inhabitants and an average of almost 15,000 inhabitants (Figure 2.8). The "mobility area" is a more homogeneous unit than the municipality, but less detailed than the coverage area of each antenna. For daily mobility data, the most frequent position of mobile phones at night (from 10 p.m. to 6 a.m.) is analysed for a given day, to determine the area of residence, and compared with the position in the schedule from 10:00 to 16:00, which determines the destination area [22, 23, 25].

The study observed that in 2019, weekdays (monday to friday) 30% of the population left their area of residence during the central hours (probably due to the need to go to work or study). The areas that received the most population on a daily basis in November 2019 were Madrid and Barcelona due to they are consider as city hubs in Spain [22, 23, 25]. The figures

observed in 2019 can be used as a reference for further comparisons due to the absence of mobility restrictions.

The analysis of 2020 figures, confirms the percentage of population that left their regular area of residence on weekdays during central hours, experienced a high reduction during the second half of the year compared with 2019 figures. The percentage observed was between 15% and 20%, compared to levels close to 30% in a "normal" week of 2019 (18 to 21 November 2019 was the reference week for this study)[22, 23, 25].

A similar approach was taken by Google, where local mobility reports are broken down by location and shows the number of visits to supermarkets, parks, etc. (Figure 2.10). The information in these reports is generated from aggregated and anonymized data sets, where users accept to have activated its location history. Anonymization technology to protect privacy and security is implemented, allowing Google and INE co-create valuable information while preserving the anonymity of any specific person [19, 23].

Figure 2.10: Google mobility trends. Source: Google [19]

The datasets provided by INE and Google are used in this work [19, 22, 23, 25].

2.4 COVID-19 - Machine and deep learning techniques

Multiple researchers and scientific teams related to statistics and machine learning are working in the development of solutions that helps to understand COVID-19 behaviour, patterns, char-

acteristics, future spread, and how the data available can help to implement the right decisions. There are several machine and deep learning techniques that can used in order to estimate the number of new possible inflections based on epidemic aspects (mortality, recovers, etc.) and social aspects (mobility in our case) [1].

The following subsections are going to define general concepts and techniques for better understanding.

2.4.1 Autoregressive Integrated Moving Average (ARIMA)

ARIMA (Autoregressive Moving Integrated Average) models are widely used to predict future values of a time-series. This is a statistical model for time series that take into account the dependence between the data, which is considered each observation at a given moment, and its modelling based on the previous values [11, 21, 28]. The relevant elements in ARIMA are:

- **Autoregressive Processes - AR (p)** - Autoregressive models are based on the idea the current value of the series, Xt, can be explained or predicted based on p past values $Xt-1$, ... $Xt-p$ plus an error term, where p determines the number of past values needed to forecast a current value.

- **Moving Average Processes - MA (q)** - A moving average model is one that explains the value of a certain variable in a period t as a function of an independent term and a succession of errors corresponding to preceding periods, appropriately weighted. These models are usually denoted by the initials MA, followed by the order in parentheses. All moving average processes are stationary processes but not all moving average processes are invertible.

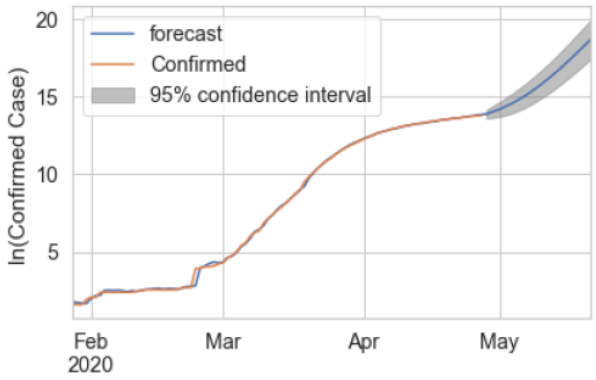
- **Autoregressive Process of Moving Averages - ARMA (p,q)** - A natural extension of the AR (p) and MA (q) models is a type of model that includes both autoregressive and moving average terms. The autoregressive models of moving averages, ARMA (p, q), are the sum of an autoregressive process of order p and one of moving averages of order q. It is very likely that a time series has characteristics of AR and MA at the same time and, therefore, is ARMA. An ARMA (p, q) process is stationary if its autoregressive component is stationary, and it is invertible if its moving average component is.

- **Integrated Process - I (d)** - Not all time series are stationary, some of them shows changes over time or the variance is not constant, so the series is differentiated d times to make it stationary. These types of processes are considered integrated processes, and an

ARMA (p, q) model can be applied to this differentiated series to give rise to an ARIMA (p, d, q) model.

In summary, ARIMA is an integrated autoregressive time series of moving average, where p denotes the number of autoregressive terms, d the number of times the series must be differentiated to make it stationary and q the number of terms of the invertible moving average [11, 21, 28].

A study performed to forecast COVID-19, based on time series in several countries [29], used ARIMA and its results stated that, compared with other similar models (like Facebook Prophet [48]), ARIMA was performing better when analysing the **Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Root Relative Squared Error (RRSE), and Mean Absolute Percentage Error (MAPE) error levels**. The dimensions forecasted were recoveries, deaths, confirmed and active cases of COVID-19 (Figure 2.11 - Confirmed cases).



(a) ARIMA Forecasting for US confirmed cases

Figure 2.11: ARIMA forecast. Source: Taylor and Letham [48]

In order to deal with multiple variables, ARIMA should be adopt the form of a VAR (Vector Autoregressive Model) for multivariate time series data.

2.4.2 Long-Short Term Memory (LSTM)

The **Artificial Neuron** (AN) is considered the minimum unit of calculation able to simulate the behaviour of a biological neuron and it is considered the basics of constitutes artificial neural networks (Figures 2.12, 2.13). The **Artificial Neural Networks** (ANN) are multiple AN connected each other with the aim to transmit signals / information that crosses over the

ANN. During these movements a series of mathematical operations happens and produces a series of output values.

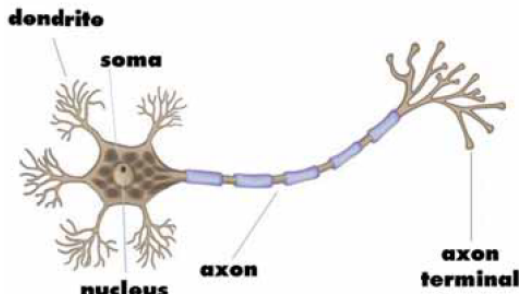


Figure 2.12: Neuron. Source: Ciaburro and Venkateswaran [9]

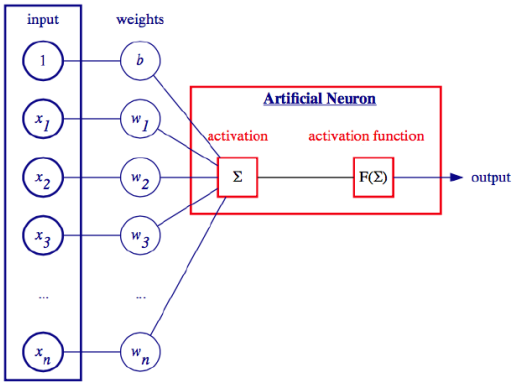


Figure 2.13: Artifical neuron. Source: Vasilev et al., [50]

As an example (Figure 2.14), we can observe the different connection possibilities. Here, we have one input layer, two hidden layers and one output Layer). The input layers receive the initial data, the hidden layers perform the mathematical calculations, and the output layer returns the prediction.

An **LSTM** it is a type of Recurrent Neural Network (RNN), with the ability to reuse valuable information (recent or not form other neurons) as input. RNNs are not able to learn from long-term memory dependencies, here LTSM solves this issue due to its ability to avoid the vanishing gradient problem from RNNs. The LSTM approach to face the vanishing gradients is via the introduction of gating mechanism able to remove or add information to cell state that control the information moved through them [20, 39, 50].

The gate system provided by LSTM (Figure 2.15) it is composed by the input state, output, and forget gates [50] as follows:

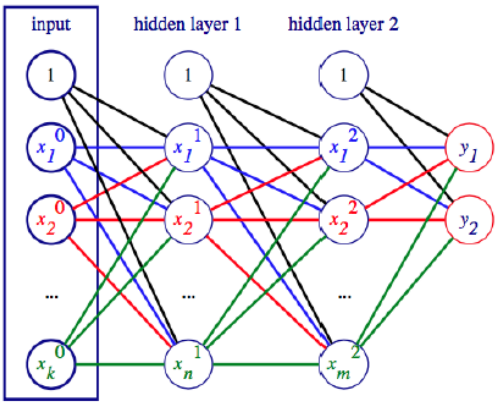


Figure 2.14: Artifical neuron multilayer. Source: Vasilev et al., [50]

- **Input state (write)** - Define the amount of information of the newly computed state that will be moved to the succeeding states for the current input, x_t.
- **Forget gate (reset)** - Control the amount of information from the previous state that will be moved to the next cell, c_t.
- **Output gate (read)** - Define the amount of internal state information that will be moved to the next state, h_t.

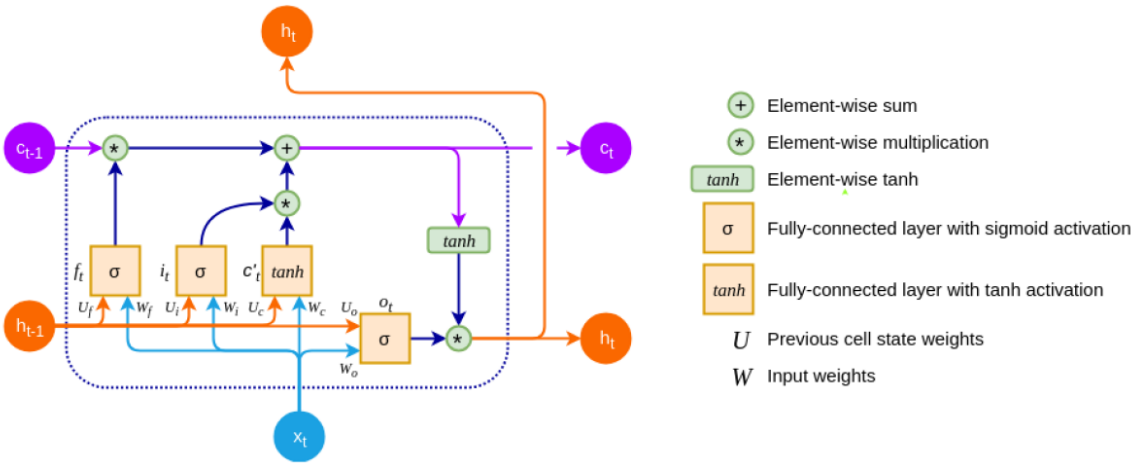


Figure 2.15: LSTM. Source: Vasilev et al., [50]

The figure above represents x_t as the LSTM input, c_t as the cell memory state and h_t as the output / hidden state in a given moment t. Here, x_t and previous h_{t-1} are connected to each gate. Candidate cell vector sets weights W and U. c_t is the cell state at moment t. f_t, i_t, and o_t are the forget, input, and output gates of the LSTM cell [50].

The first step in building an LSTM network is to identify the data / information that is not required and will be omitted / removed, forget gate. This identification is decided by the sigmoid function σ, which takes the output from the last LSTM unit h_{t-1} at time $t-1$ and the current input x_t at time t. The sigmoid function determines how much of the above output should be removed. This door is a vector with values ranging from 0 to 1, corresponding to each number in the cell state c_{t-1}. 0 value remove c_{t-1} the cell block and a value of 1 move on the information.

$$f_t = \sigma(W_f x_t + U_f h_{t-1})$$

The second step, input gate, decides the new information will be added to the memory cell based on h_{t-1} and x_t. Like in forget gate, this door is a vector with values from 0 or 1 for each cell. 0 means no information will be added to the cell block's memory.

$$i_t = \sigma(W_i x_t + U_i h_{t-1})$$

Then the a candidate input is added with a tangential, tanh, function that gives weight to the passed values, deciding their level of importance (-1 to 1).

$$c_t^{'} = \tanh(W_c x_t + U_c h_{t-1})$$

And now the forget and input gates the new cell state by selecting the old and new parts.

$$c_t = f_t * c_{t-1} + i_t * c_t^{'}$$

The third step, output gate, the output values (h_t) are based on the state of the output cell (o_t) but it is a filtered version. Here, a sigmoid layer decides which parts of the cell state get to the output.

$$o_t = \sigma(W_o x_t + U_o h_{t-1})$$

Finally, the output of the sigmoid gate is then multiplied by the new values created by the tanh layer from the cell state (c_t), with a value ranging from -1 to 1.

$$h_t = o_t * \tanh(c_t)$$

2.4.3 Studies carried-out

Several studies have been adopted the LSTM to forecast COVID-19 comparing with other models (ARIMA, etc.). The outcome from these studies is that LSTM can be consider a robust model forecasting COVID-19 due to the MAE, RSME, etc. offered [26, 27, 42, 43, 47].

One of these studies was focused to test the performance LSTM, RNN, and Gated Recurrent Units (GRU) to predict number of COVID-19 infections during a period of 10 days based on data published by WHO (World Health Organization). The outcome states that these models were able to obtain precisions rates close or greater than 90%. Countries under study were **Pakistan, India, Afghanistan, and Bangladesh** (Figure 2.16) [43].

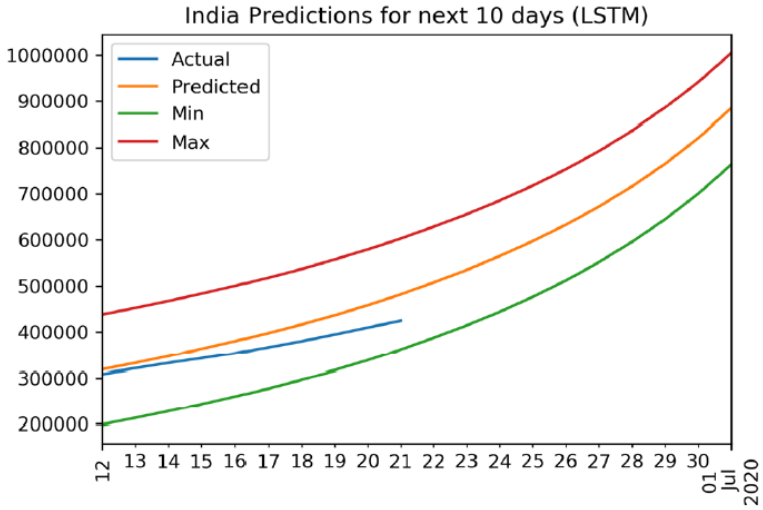


Figure 2.16: LSTM India prediction. Source: Rauf et al., [43]

A study carried-out in the city of Chennai (India) was able to forecast COVID-19 death and recovered cases using a Stacked LSTM model [12]. Figure 2.17 shows the performance of the Stacked LSTM model compared with others in different based on MAPE.

Other of these studies, with a global geographical scope [47], where ten countries were under observation. Again it was observed that the LSTM models were performing better compared with Support Vector Regression (SVR) and ARIMA. The outcomes of this study states clearly that ARIMA and SVR were not able to predict (Figure 2.18).

The information stated in these studies and the outcomes offered, leads us to adopt a similar approach in order to create a model that predict COVID-19 based on mobility dimension. We are going to compare ARIMA and LSTM.

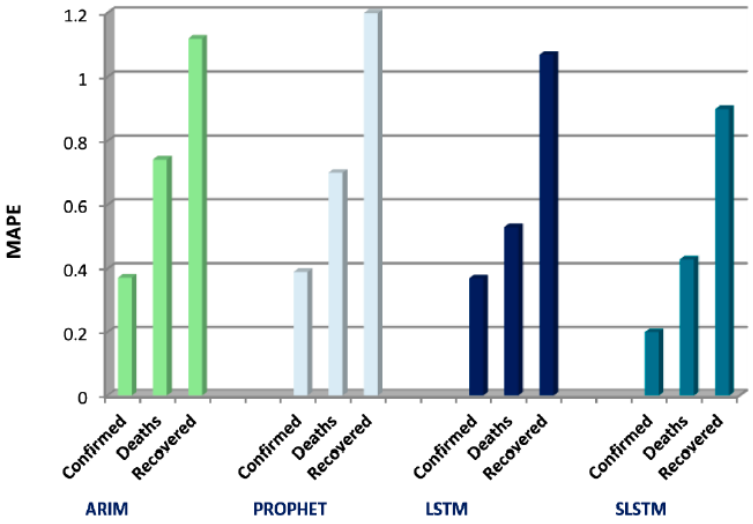


Figure 2.17: SLSTM India MAPE. Source: Devaraj et al., [12]

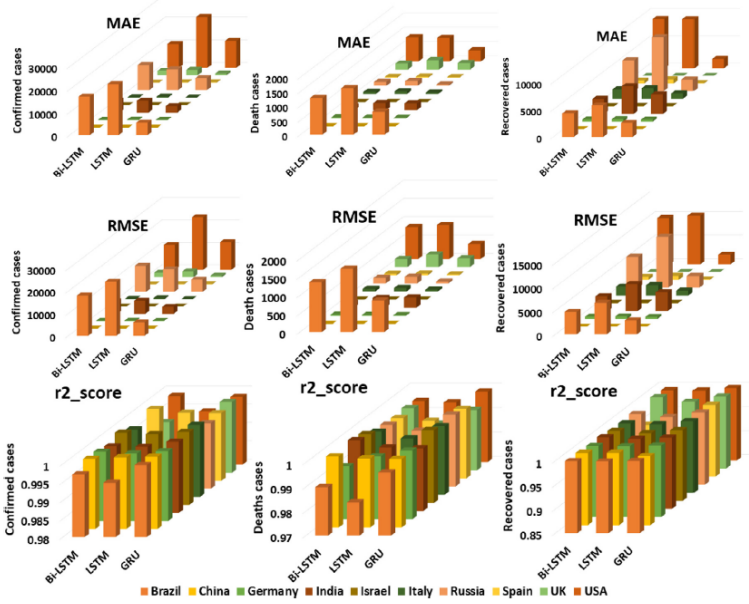


Figure 2.18: LSTM comparative - 2. Source: Shahid, Zameer and Muneeb, [47]

2.5 Datasets to be used

As it has been stated in previous sections, this study will use the data offered by CNE [10] Google [19] and INE [22]. CRISP-DM will be the foundation for data preparation, analysis and model selection.

- **CNE** - This dataset offers information related to infections, recoveries and deaths reported by local and regional governments in Spain.

- **Google** - This dataset offers information related to mobility using Google application services.
- **INE** - These datasets (there are several) offers the mobility of the population captured by the mobile telephony infrastructure system in Spain.

Due to all these datasets are compressed and offers a huge quantity of records, it will be necessary to perform a lot of previous data preparation activities as was mention when CRISP-DM was explained.

2.6 Data-science IDE and language to be used

R [41] is a high-level program language and environment for data analysis and graphics to perform statistical tasks. It is free and can be downloaded from the project site (CRAN - Comprehensive R Archive Network).

RStudio [45] is an integrated development environment (IDE) for the R programming (statistical computing language). It origins are based to statistical computing and graphics. It is open-source, includes a console syntax editor that supports code execution, as well as tools for plotting, debugging, and managing the workspace. It is available for **Windows, Mac, and Linux.**

This is the software to be used for this study due to is widely used in industry, is open-source and there are available a lot of predictive, explore and analysis libraries ready for use like **Tensorflow, Tidymodels, Tidyverse, ggplot2, dplyr, tidyr, etc.**

Chapter 3

Methodology

3.1 Steps followed

As has been stated at **"State of the art"** section and to establish a proper order in the elaboration of the work, **KDD** methodology (Knowledge Discovery in Databases) and **CRISP-DM** (Cross Industry Standard Process for Data Mining) were used, covering all the tasks and phases for the project [16, 17, 18, 37, 56] (Figure 3.1).

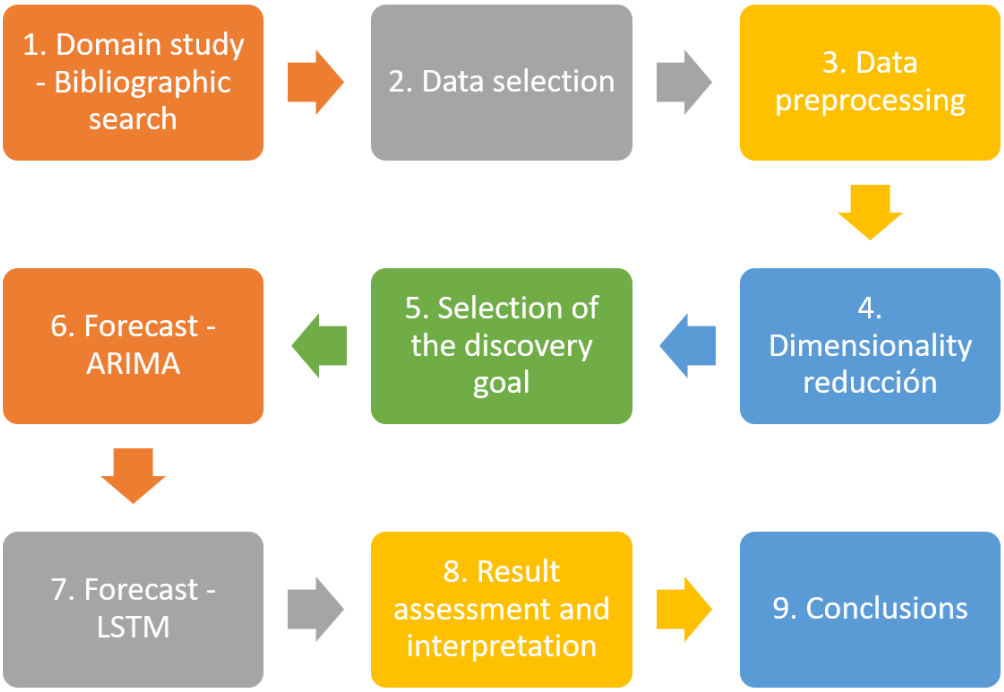


Figure 3.1: Methodology used

3.1.1 Domain Study - Bibliographic research

A bibliographic research was carried out in **"The Lancet", "Nature", "PUBMED",** etc. seeking for a better understanding of COVID-19 pandemic, measures taken by governments, open databases available to monitor COVID-19 and studies performed to predict its impact based on time series and neural networks (this research is contained at "State of the art" section).

3.1.2 Data selection

Due to the geographical focus of this project is Spain, as it has been stated in previous sections, this study will use the data offered by **CNE** [10] and **INE** [22] due to they are official institutions in Spain and offers open data. **Google** [19], is not and official institution, but the data offered it is widely used and as well as the CNE and INE, ensures anonymity and its data privacy policy, data gathering and methods applied to the data can be accessed and reviewed.

3.1.3 Data preprocessing

At this section we summarize only the most relevant data preparations accomplished. A separate report it is attached with all the different transformations, modifications, executions (pre and post with its results), etc. with specified and documented comments at the appropriated code block section [44].

The software used was **RStudio** [45] and the following libraries for manipulation, forecasting, visualization and document preparation were downloaded and installed (plus others): **corrplot, DescTools, fable, forecast, fpp3, ggplot2, imputeTS, keras, knitr, latex2exp, latexpdf, lubridate, psych, stats, tensorflow and tidyverse**.

- **CNE** - These datasets offers information related to infections, recoveries and deaths reported by local and regional governments in Spain. Two datasets were used:

  1. **cases_technic_province.csv** - Number of cases by diagnostic technique and province of residence, with **23426** obs. of **8** variables, start-date **01-01-2020**, end-date **17-03-2021**, **442** records per province and **0.01886792%** of missing values for column **"provincia-iso"** (those rows were omitted from the dataset).
  2. **cases_hosp_uci_def_sexo_edad_provres.csv** - Number of hospitalizations, number of ICU admissions and number of deaths by sex, age and province of residence, with **702780** obs. of **8** variables, start-date **01-01-2020**, end-date **17-03-2021, 13260** records per province (we have sub-age groups per province) and

0.01886792% of missing values for column **"provincia-iso"** (those rows were omitted from the dataset).

3. In both datasets we transformed column **"Fecha"** from **"character"** to **"date"**. Columns **"Grupo_edad"** and **"Sexo"** were eliminated from the dataset "CNE_casos" due to they are not adding value (mobility datasets do not include this variables). We changed **"NC"** values at iso code level to **"NA" (Navarra)** in both dataframes.

- **Google** - This dataset offers information related to mobility using Google application services. We downloaded the regional compressed dataset **"Google_Region_Mobility_Report_CSVs.zip"**, and we looked for the Spain one named as **"Google-2020_ES_Region_Mobility_Report.csv".**

1. This dataset has **24242** obs. of **15** variables, start-date **15-02-2020**, end-date **05-03-2021** and **385** records per province (we have groups per autonomous-communities). Several checks and transformations related to the appropriated **"province name"** and **"ISO code"** were carried-out.
2. Rows with **"na"** and **""** in **"sub_region_1"** and **"sub_region_2"** columns were eliminated. **"Date"** was transformed from **"character"** to **"date"**. Some columns were eliminated due to they are not adding value or they contain only blanks (**country_region_code, country_region, metro_area, census_fips_code,pace_id**). **"ES-"** characters were eliminated from **"iso_3166_2_code"** column.
3. For the missing values at the different areas of interest explored, as **"retail_and_recreation_percent_change_from_baseline"** we have used **"na_seadec()"** function from **"imputeTS"** package due to it takes into account seasonality for the time series (for this particular case we have generated one time series dataset per dimension of interest and finally we merged again into a cleaned data-frame).

- **INE** - This dataset offers the mobility population captured by the mobile telephony infrastructure system in Spain. We downloaded the dataset offered by the **INE web application** related only to the provinces in Spain [24].

1. **EM3-Movimiento de personas por provincias.csv** - This dataset has **9198** obs. of **3** variables, start-date **16-03-2020**, end-date **31-12-2020** and **146** records per province.
2. **"Total"** column was changed from **"character"** to **"numerical"** and **"Periodo"** column from **"character"** to **"date"**.

3. Due to the nature of this dataset, we have had to transpose it in order to analyse the missing values by province and impute them. As in the case of Google, we have used **"na_seadec()"** function from **"imputeTS"** package due to it takes into account seasonality for the time series (for this particular case we have generated a time series from the data-frame and finally we converted back to a data-frame).

- **Total** - This dataset is the result of a merge process carried-out for the previous ones (**CNE+INE+Google**) and it is the one used to perform our study, **ARIMA vs LSTM** [21]. This dataset has **15080** obs. of **20** variables, start-date **16-03-2020**, end-date **31-12-2020** and 290 **records per province**.

  The dataset has been also converted to a time-series and a CSV file version (Total.csv) it is provided for its review and usage, if necessary, at the Github [44] repository generated for this project. This dataset was converted to time-series with a daily frequency (Total_ts and Total_ts_b -this one with 5 provinces and 1450 rows-). The following are some figures extracted from Total dataset (Figures 3.2).

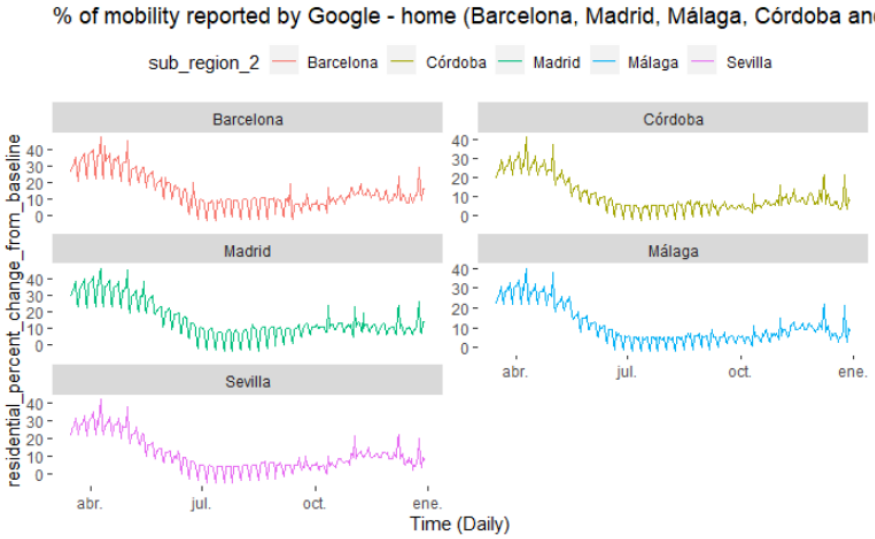


Figure 3.2: Google - Change residential

3.1.4 Dimensionality reduction

Correlation and **PCA** (Principal Component Analysis) were carried-out to address which are the mos important variables for our analysis (Figures 3.3, 3.4).

For our case we consider that up to **PC3, which explain the 84%** of the accumulated variance is enough and we have eliminated **"num_casos_prueba_test_ac", "num_casos_prueba**

_ag", "num_casos_prueba_elisa" and "num_casos.y" columns. The case of "num_casos.y" is due to we can consider it as duplicated against "num_casos.x". This means we count with 15 columns that it is suppose will offer valuable information for further stages (15080 total rows - 290 per province). This approach was based on **Barcelona** and was extrapolated to the rest of provinces.

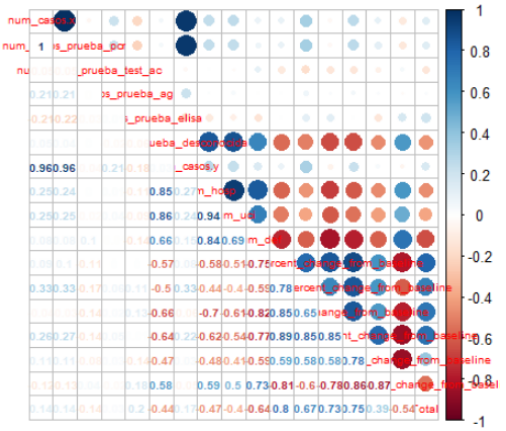


Figure 3.3: Correlation observed - Barcelona

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9     PC10    PC11    PC12    PC13    PC14    PC15    PC16      PC17
Standard deviation     3.0590 1.8914 1.12142 1.02409 0.83504 0.57825 0.46266 0.41006 0.33620 0.28445 0.21986 0.17734 0.16911 0.15372 0.1369 0.01274 1.826e-17
Proportion of Variance 0.5504 0.2104 0.07398 0.06169 0.04102 0.01967 0.01259 0.00989 0.00665 0.00476 0.00284 0.00185 0.00168 0.00139 0.0011 0.00001 0.000e+00
Cumulative Proportion  0.5504 0.7609 0.83485 0.89655 0.93756 0.95723 0.96982 0.97971 0.98636 0.99112 0.99397 0.99582 0.99750 0.99889 1.0000 1.00000 1.000e+00
                                                                PC1          PC2          PC3          PC4          PC5          PC6          PC7          PC8          PC9
num_casos.x                                          -0.1354826901 -0.469400471 -0.002875288 -0.028445983 -0.15653475  0.16549669  0.11538168  0.07515692 -0.204798939
num_casos_prueba_pcr                                 -0.1356873829 -0.469237915 -0.012292151 -0.026795983 -0.15256567  0.17028866  0.11048811  0.07713501 -0.206074966
num_casos_prueba_test_ac                             -0.0782958432  0.237502756 -0.392712472 -0.268237834 -0.81534671 -0.10365823 -0.08510797 -0.12117026  0.031408747
num_casos_prueba_ag                                   0.0005323686  0.008790618 -0.034460138  0.950253303 -0.25969423 -0.01360833 -0.03926649 -0.14832473 -0.027397924
num_casos_prueba_elisa                                0.0495912937  0.004385553  0.811470962 -0.060011278 -0.34805371 -0.37185910  0.24859208  0.01374512  0.085159377
num_casos_prueba_desconocida                         -0.2907430443  0.113686282  0.019674417 -0.078105637  0.20580692 -0.11429059  0.19845835 -0.75630827 -0.349092565
num_casos.y                                          -0.1546876637 -0.450080493  0.041668217 -0.036946432 -0.15827091  0.15297764  0.04117476 -0.02860659 -0.130984317
num_hosp                                             -0.2956764064 -0.199498016 -0.051419434 -0.010413059  0.03855580 -0.11688342 -0.11367668 -0.07611704  0.283065654
num_uci                                              -0.2788863801 -0.230508957 -0.018126448  0.006323566  0.06172358 -0.06173530 -0.02723992 -0.20583241  0.717067401
num_def                                              -0.3153112111 -0.030904089 -0.031133681  0.009747098  0.08791553 -0.26888200 -0.20890547  0.15436921 -0.009119052
retail_and_recreation_percent_change_from_baseline    0.3077977291 -0.116093786  0.033200491 -0.075398568 -0.02731712  0.18700422 -0.10130686 -0.31700087  0.099531350
grocery_and_pharmacy_percent_change_from_baseline     0.2553316609 -0.261732823  0.027494995 -0.030995413  0.01445198 -0.34534708 -0.63305922 -0.04093499 -0.110976209
parks_percent_change_from_baseline                    0.3128404870  0.012776948  0.029299658 -0.011018427 -0.01705528  0.39173313  0.15076652  0.04293438  0.283874181
transit_stations_percent_change_from_baseline         0.2855716515 -0.218931952 -0.099387689 -0.018021579  0.03520761 -0.26003250 -0.06852096 -0.19335324  0.094114251
workplaces_percent_change_from_baseline               0.2622358364 -0.170838668 -0.303553292  0.028981871  0.10943614 -0.48832683  0.37588850  0.19906444 -0.125028631
residential_percent_change_from_baseline             -0.2895648649  0.166332144  0.211975965  0.003369575  0.03104587  0.13040337 -0.39494991  0.23079899 -0.178700468
Total                                                 0.2993175108 -0.081611685  0.173800838 -0.065471155 -0.05468582  0.19337305 -0.26415699 -0.26912855 -0.118124812
```

Figure 3.4: PCA - Variance explained - Barcelona

3.1.5 Selection of the discovery goal

Data were processed using the following regression and neural network forecasting methods: **"ARIMA" and "LSTM"** [21].

The goal for this study is to predict infections caused by COVID-19 based on mobility data (quantitative time-series forecast). We have selected "num_casos.x" as target / dependent

variable and the rest of mobility variables are considered as the independent ones. The **forecast horizon will be 7, 14 and 21 days** (our frequency is daily - 365 based on the data obtained).

3.1.6 ARIMA

As stated by Hyndman [21], "...

- **ARIMA** models aim to describe the autocorrelations in the data...
- **Trend** exists when there is a long-term increase or decrease in the data...
- **Seasonal** pattern occurs when a time series is affected by seasonal factors such as the time of the year...
- **Cyclic** occurs when the data exhibit rises and falls that are not of a fixed frequency...
- **A stationary** time series is one whose statistical properties do not depend on the time at which the series is observed...
- When a decomposition of a time series happens, trend and cycle are combined into a single trend-cycle component (called "the trend"). So, a time series is composed by three components: trend-cycle, seasonal component, and remainder (this one contains anything else in the time series)."

In our case and for simplicity reasons we have analysed 5 provinces (Barcelona, Madrid, Málaga, Cádiz and Sevilla) from a univariate and multivariate perspective (in this report we have focused on Barcelona results due to the other provinces offers a similar behaviour). We have followed some of the general steps stated by Hyndman [21] for the analysis of our time series.

- Plot the data.
- If necessary, transform the data (stabilise the variance).
- If the data are non-stationary, take first differences till its became stationary.
- Examine the ACF/PACF.
- Try chosen model(s) / Use the AICc to search for a better model.
- Check residuals from chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals (if no white noise is observed try a modified model).

- If residuals can be consider white noise, calculate forecasts.

We have observed a week seasonality thanks to the Seasonal and Trend decomposition using Loess (STL) test performed, and thanks to difference method applied we were able to convert our time series to a stationary one (Figures 3.5 and 3.6), but the **ACF** plots shows that more transformations are necessary. In our case we have selected the Box-Cox transformation to control variance and due to the libraries used in **fpp3** package performs the reverse conversion when the forecast is done.

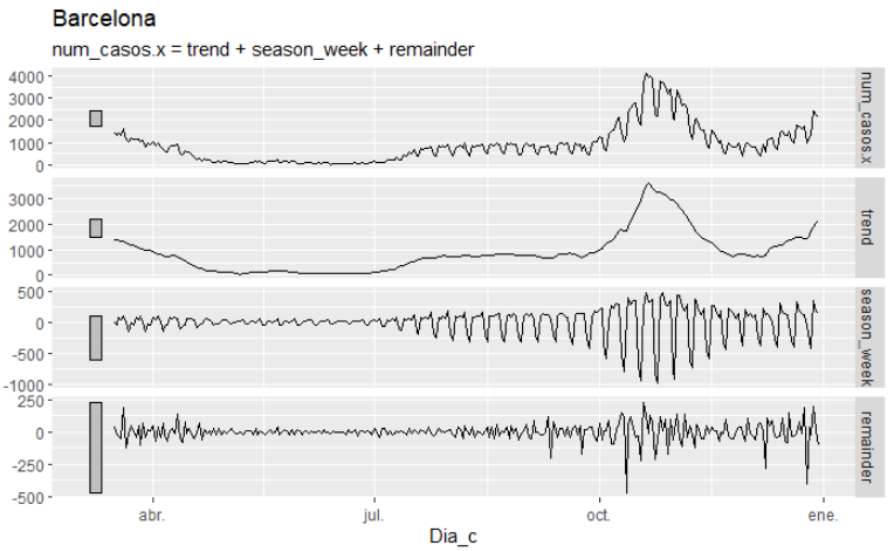


Figure 3.5: Barcelona - Seasonality / Trend - STL

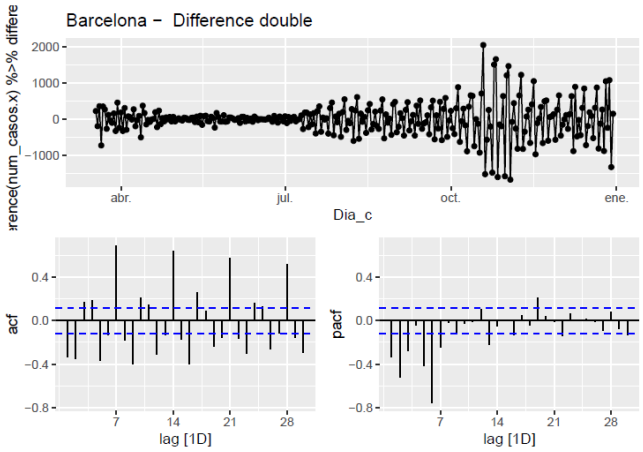


Figure 3.6: Barcelona - Residuals - double difference

Univariate and multivariate (this one using the dynamic regression using errors approach, using **fable()** function from **fpp3 library**) were performed. The two approaches were

carried out for Barcelona with similar results and the univariate model was performed over Madrid, Málaga, Cádiz and Sevilla.

ARIMA models performs better compare with **SNaive** in all cases (Barcelona use case). But ARIMA automated option with extensive search of parameters performs better under the uni-multivariate options (**arima_at2**). For the manual (**arima_mnX**) and easy automate (**arima_at1**) approaches, results were similar between them. The inclusion of externals variables increases the performance of the forecast (Figures 3.7, 3.8 and 3.9).

.model <chr>	sub_region_2 <chr>	.type <chr>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>
arima_at1	Barcelona	Test	505.5713	547.2573	505.5713	38.05364	38.05364
arima_at2	Barcelona	Test	463.7301	495.0512	463.7301	35.19974	35.19974
arima_mn1	Barcelona	Test	494.7994	532.9852	494.7994	37.28315	37.28315
arima_mn2	Barcelona	Test	463.6181	494.9163	463.6181	35.19252	35.19252
arima_mn3	Barcelona	Test	460.2675	491.6425	460.2675	34.93691	34.93691
SNaive	Barcelona	Test	553.6429	612.4804	553.6429	41.78377	41.78377

Figure 3.7: Barcelona - Accuracy univariate (14 days) based on errors

.model <chr>	sub_region_2 <chr>	.type <chr>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>
arima_at1	Barcelona	Test	229.8860	368.8146	312.3555	12.74038	23.12511
arima_at2	Barcelona	Test	216.3219	253.4041	216.3219	15.85733	15.85733
arima_mn1	Barcelona	Test	373.3955	419.6146	373.3955	28.03662	28.03662
arima_mn2	Barcelona	Test	367.9690	409.2342	367.9690	27.72285	27.72285
arima_mn3	Barcelona	Test	364.7335	405.2355	364.7335	27.52434	27.52434
SNaive	Barcelona	Test	553.6429	612.4804	553.6429	41.78377	41.78377

Figure 3.8: Barcelona - Accuracy multivariate (14 days - All) based on errors

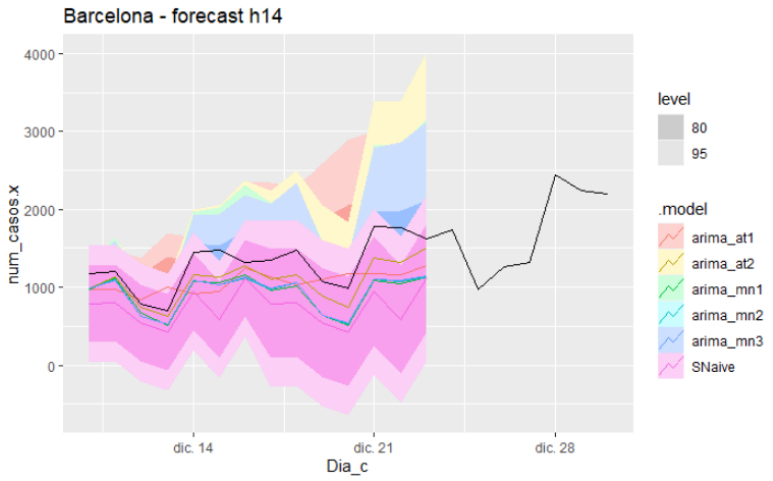


Figure 3.9: Barcelona - Multivariate (14 days - All) forecast plot

In this exercise we have notice that for Málaga, Cádiz and Sevilla provinces, SNaive model performed better for the univariate analysis (Figures 3.10 and 3.11).

.model <chr>	sub_region_2 <chr>	.type <chr>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>
arima_at1	Sevilla	Test	60.71952	67.85314	60.71952	38.59897	38.59897
arima_at2	Sevilla	Test	60.01494	67.35877	60.01494	38.08009	38.08009
arima_mn1	Sevilla	Test	66.78203	72.98320	66.78203	44.29884	44.29884
arima_mn2	Sevilla	Test	61.38012	68.49618	61.38012	39.02890	39.02890
arima_mn3	Sevilla	Test	60.71952	67.85314	60.71952	38.59897	38.59897
SNaive	Sevilla	Test	25.78571	50.24298	38.64286	13.51818	24.09845

Figure 3.10: Sevilla - Accuracy univariate (14 days) based on errors

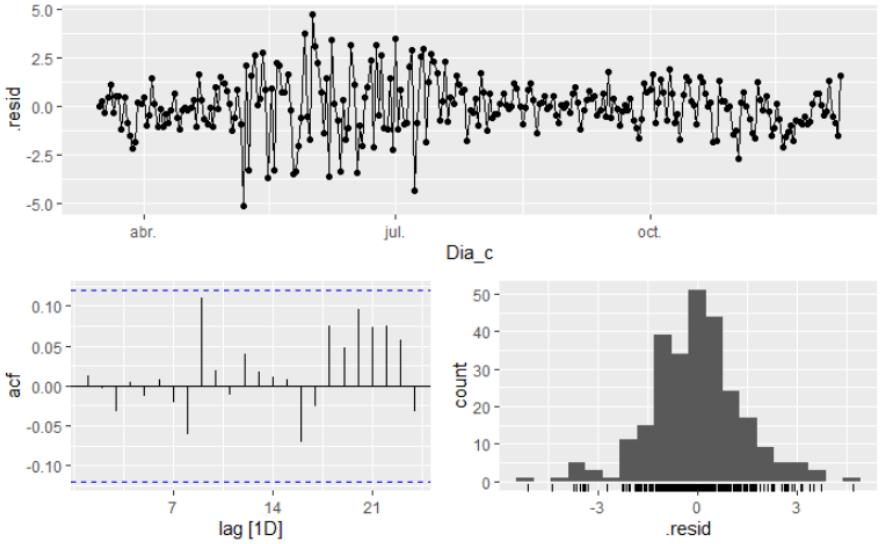


Figure 3.11: Sevilla - Univariate (14 days) forecast plot

This situation leads us to consider that, for ARIMA models, it is necessary to manage each time series as a unique study case (by province, choosing each ARIMA parameter and the size of each series) even if the area of study and variables used are the same and the residuals ar considered white noise.

3.1.7 LSTM

3.1.8 Results assessment

3.1.9 Conclusions

Bibliography

[1] Mobility-based prediction of sars-cov-2 spreading. https://arxiv.org/abs/2102.08253. [Online; accessed 25-Feb-2021].

[2] Charu C Aggarwal. *Data Mining*. Springer, 2015.

[3] Apple. Covid-19 - mobility trends repors. https://covid19.apple.com/mobility, 2021. [Online; accessed 22-Feb-2021].

[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] Taweh Beysolow II. *Introduction to Deep Learning Using R*. Apress, 2020.

[6] CDC. Severe acute respiratory syndrome (sars). https://www.cdc.gov/sars/about/index.html, 2013. [Online; accessed 6-Mar-2021].

[7] CDC. Middle east respiratory syndrome (mers). https://www.cdc.gov/coronavirus/mers/about/index.html, 2019. [Online; accessed 6-Mar-2021].

[8] CDC. Late sequelae of covid-19. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/late-sequelae.html, 2020. [Online; accessed 21-Feb-2021].

[9] Giuseppe Ciaburro and Balaji Venkateswaran. *Neural networks with R*. Packt Publishing, 2017.

[10] CNE. Covid-19. https://cnecovid.isciii.es/covid19/#documentaci%C3%B3n-y-datos, 2021. [Online; accessed 21-Feb-2021].

[11] Gergely Daróczi. *Mastering data analysis with R*. Packt Publishing, 2015.

[12] Jayanthi Devaraj, Rajvikram Madurai Elavarasan, Rishi Pugazhendhi, G.M. Shafiullah, Sumathi Ganesan, Ajay Kaarthic Jeysree, Irfan Ahmad Khan, and Eklas Hossain. Forecasting of covid-19 cases using deep learning models: Is it reliable and practically significant? *Results in Physics*, 21:103817, 2021.

[13] Roger Devine and Michael Pawlus. *Hands-On Deep Learning with R*. Packt Publishing, 2020.

[14] Kuldeep Dhama, Sharun Khan, Ruchi Tiwari, Shubhankar Sircar, Sudipta Bhat, Yashpal Singh Malik, Karam Pal Singh, Wanpen Chaicumpa, D. Katterine Bonilla-Aldana, and Alfonso J. Rodriguez-Morales. Coronavirus disease 2019–covid-19. *Clinical Microbiology Reviews*, 33(4), 2020.

[15] Deming Edwards. Pdsa cycle - the w. edwards deming institute. https://deming.org/explore/pdsa/, 2021. [Online; accessed 22-Feb-2021].

[16] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37, Mar. 1996.

[17] Jordi Giroes. *Metodologías y estándares*. Editorial UOC, nd.

[18] Jordi Gironés, Jordi Casas, and Julià Minguillón. *Minería de datos modelos y algoritmos*. Editorial UOC, 2017.

[19] Google. Covid-19 community mobility report. https://www.google.com/covid19/mobility/, 2021. [Online; accessed 22-Feb-2021].

[20] Swarna Gupta, Rehan Ali Ansari, and Dipayan Sarkar. *Deep learning with R cookbook*. Packt Publishing, 2020.

[21] R.J. Hyndman and G. Athanasopoulos. Forecasting: Principles and practice (3rd ed). https://otexts.com/fpp3/, 2021. [Online; accessed 5-Mar-2021].

[22] INE. Estadística experimental. https://www.ine.es/experimental/movilidad/experimental_em.htm, 2020. [Online; accessed 21-Feb-2021].

[23] INE. Estudio de movilidad a partir de la telefonía móvil durante el periodo julio-diciembre 2020 (em-3). https://www.ine.es/experimental/movilidad/exp_em3_proyecto.pdf, 2020. [Online; accessed 21-Feb-2021].

[24] INE. Estudio de movilidad a partir de la telefonía móvil durante el periodo julio-diciembre 2020 (em-3) - provincias. https://www.ine.es/jaxiT3/Tabla.htm?t=37812, 2020. [Online; accessed 21-Feb-2021].

[25] INE. Información estadística para el análisis del impacto de la crisis covid-19. https://www.ine.es/covid/covid_inicio.htm, 2020. [Online; accessed 21-Feb-2021].

[26] Shwet Ketu and Pramod Kumar Mishra. A hybrid deep learning model for covid-19 prediction and current status of clinical trials worldwide. *Computers, Materials and Continua*, 66(2):1896–1919, 2021.

[27] Erdinç Koç and Muammer Türkoğlu. Forecasting of medical equipment demand and outbreak spreading based on deep long short-term memory network: the covid-19 pandemic in turkey. *Signal, Image and Video Processing*, 2021.

[28] Rami Krispin. *Hands-On time series analysis with R.* Packt Publishing, 2019.

[29] Naresh Kumar and Seba Susan. Covid-19 pandemic prediction using time series forecasting models, 2020.

[30] Daniel T Larose. *Data Mining Methods and Models.* John Wiley and Sons, 2 edition, 2015.

[31] Johannes Ledolter. *Business analytics and data mining with R.* John Wiley and Sons, 2013.

[32] Smriti Mallapaty. What's the risk of dying from a fast-spreading covid-19 variant? *Nature*, 590(7845):191–192, 2021.

[33] Mattia Mazzoli, David Mateo, Alberto Hernando, Sandro Meloni, and José J. Ramasco. Effects of mobility and multi-seeding on the propagation of the covid-19 in spain. *medRxiv*, 2020. [https://doi.org/10.1101/2020.05.09.20096339].

[34] The Lancet Respiratory Medicine. Covid-19 transmission—up in the air. *The Lancet Respiratory Medicine*, 8(12):1159, 2020.

[35] Lidia Morawska and Junji Cao. Airborne transmission of sars-cov-2: The world should face the reality. *Environment International*, 139:105730, 2020.

[36] Lidia Morawska, Julian W. Tang, William Bahnfleth, Philomena M. Bluyssen, Atze Boerstra, Giorgio Buonanno, Junji Cao, Stephanie Dancer, Andres Floto, Francesco Franchimon, and et al. How can airborne transmission of covid-19 indoors be minimised? *Environment International*, 142:105832, 2020.

[37] Braulio Nuria and Josep Curto. *Customer analytics.* Editorial UOC, nd.

[38] Yixuan Pan, Aref Darzi, Aliakbar Kabiri, Guangchen Zhao, Weiyu Luo, Chenfeng Xiong, and Lei Zhang. Quantifying human mobility behaviour changes during the covid-19 outbreak in the united states. *Scientific Reports*, 10(1), 2020.

[39] Michael Pawlus and Rodger Devine. *Hands-On Deep Learning with R*. Packt Publishing, 2020.

[40] PKS Prakash and Achyutuni Sri Krishna Rao. *R deep learning cookbook*. Packt Publishing, 2017.

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[42] Firda Rahmadani and Hyunsoo Lee. Hybrid deep learning-based epidemic prediction framework of covid-19: South korea case. *Applied Sciences*, 10(23):8539, 2020.

[43] Hafiz Tayyab Rauf, M. Ikram Ullah Lali, Muhammad Attique Khan, Seifedine Kadry, Hanan Alolaiyan, Abdul Razaq, and Rizwana Irfan. Time series forecasting of covid-19 transmission in asia pacific countries using deep neural networks. *Personal and Ubiquitous Computing*, 2021.

[44] Alvaro Rodriguez S. Github repository. https://github.com/arodriguezsans/TFM_PEC3, 2021. [Online; accessed 21-Feb-2021].

[45] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020.

[46] A. Serrano-Cumplido, P.B. Antón-Eguía Ortega, A. Ruiz García, V. Olmo Quintana, A. Segura Fragoso, A. Barquilla Garcia, and Á. Morán Bayón. Covid-19. la historia se repite y seguimos tropezando con la misma piedra. *Medicina de Familia. SEMERGEN*, 46:48–54, 2020.

[47] Farah Shahid, Aneela Zameer, and Muhammad Muneeb. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. *Chaos, Solitons and Fractals*, 140:110212, 2020.

[48] Sean J Taylor and Benjamin Letham. Forecasting at scale. 2017. [Online; accessed 5-Mar-2021].

[49] Mark Treveil and the Dataiku team. *Introducing MLOps*. OReilly Media, Inc., 2020.

[50] Ivan Vasilev, Daniel Slater, Gianmario Spacagna, Peter Roelants, and Valentino Zocca. *Python Deep Learning - Second Edition*. Packt Publishing, 2 edition, 2019.

[51] WHO. Coronavirus. https://www.who.int/health-topics/coronavirus#tab=tab_1, 2021. [Online; accessed 21-Feb-2021].

[52] WHO. Covid-19 - vaccines. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines, 2021. [Online; accessed 6-Mar-2021].

[53] WHO. Sars-cov-2 variants. https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/, 2021. [Online; accessed 6-Mar-2021].

[54] WHO. Timeline: Who's covid-19 response. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#!, 2021. [Online; accessed 6-Mar-2021].

[55] WHO. Who coronavirus (covid-19) dashboard. https://covid19.who.int/, 2021. [Online; accessed 6-Mar-2021].

[56] Rüdiger Wirth. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pages 29–39, 2000.

[57] Wysocki and K. Sybex Brown. *Effective Project Management*. John Wiley and Sons, 8 edition, 2019.

Appendix A

Code used