uc3m | Universidad Carlos III de Madrid

Master in Telecommunication Engineering

2020-2022

Master Thesis

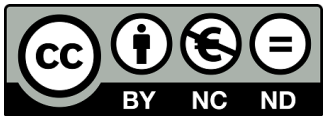
Analysis of the Evolution of COVID-19 Cases Based on Traffic Data

---

Antonio Carrera Maestro

Mario Muñoz Organero

Madrid, September 8th 2022



SUMMARY

The COVID-19 outbreak in the beginning of 2020 brought a wide range of changes in many areas of life. The world has had to adapt to unseen circumstances and take decisions under a high degree of uncertainty. Artificial Intelligence has played a key role to help the investigation and the decision making processes, due to its capability to adapt to any kind of conditions.

This project focus on the particular counter measurements that were applied on the region of Madrid to contain the spread of the virus after the initial lockdown. Open-source data is available online, so starting from there, the traffic data influences over the total number of active cases is studied. The project covers the data extraction and pre-processing, the preliminary analyses to determine the structure of the data and its characteristics, and applies Machine Learning models to evaluate the predictability of the number of COVID-19 cases.

The restrictions were applied to concrete zones were the number of cases was of concern. Three different scenarios are independently studied, including three different zones and their corresponding traffic data. Distinct Machine Learning models are applied to the data after a graphical analysis is carried out. The results show that traffic data may be useful to refine predictions for certain cases, and that the predictions change significantly depending on the zone under consideration.

Keywords: Machine Learning, COVID-19, spread, traffic, prediction, time series, regression

DEDICATION

Thanks to my supervisor Mario Muñoz, for offering me this opportunity and helping me during the development of this project.

Thanks to Marina and my parents for supporting and encouraging me to keep going now and always. Without your support, it would have been much harder.

Thanks to my friends and classmates. You made of this master a rewarding experience.

CONTENTS

1. INTRODUCTION 1
1.1. Context 1
1.2. Objectives 1
1.3. Structure of the report 2
2. STATE OF THE ART 4
2.1. Artificial Intelligence and COVID-19 4
2.2. Virus spread prediction 5
2.3. The impact of COVID-19 on traffic and mobility 8
3. METHODOLOGY 10
3.1. The Dataset 10
3.1.1. Traffic Data 11
3.1.2. Cumulative Incidence 13
3.1.3. Closed Health Zones 15
3.1.4. Data Selection 16
3.2. Graphical Analysis 17
3.3. Machine Learning Analysis 20
4. RESULTS 23
4.1. Andrés Mellado and Guzmán el Bueno 23
4.1.1. Machine Learning Analysis 23
4.2. Chopera and Legazpi. 27
4.2.1. Graphical Analysis 27
4.2.2. Machine Learning Analysis 29

4.3. Virgen de Begoña and Fuencarral 33
4.3.1. Graphical Analysis 33
4.3.2. Machine Learning Analysis 35
5. CONCLUSIONS 38
5.1. Final Results 38
5.2. Future Work 39
BIBLIOGRAPHY 41

LIST OF FIGURES

2.1 Cumulative (up) and daily (down) COVID-19 cases [8] 5

2.2 COVID-19 data collection, processing and publication scheme [15] 7

3.1 Example of traffic point location and description in Google Earth 12

3.2 Raw (left) and processed (right) traffic data for point with ID 10499 . . . 13

3.3 Traffic measurement points between the zones of Las Tablas and Sanchinarro 14

3.4 Raw (left) and processed (right) cumulative incidence data for Andrés Mellado health zone 15

3.5 Traffic sensors for Andrés Mellado scenario 18

3.6 Cumulative incidence for Andrés Mellado scenario 19

3.7 Correlation for Andrés Mellado variables 19

4.1 14-day prediction for Andrés Mellado with traffic dataset 26

4.2 Traffic sensors for Chopera scenario 27

4.3 Cumulative incidence for Chopera scenario 28

4.4 Correlation for Chopera variables 28

4.5 14-day prediction for Chopera with traffic dataset 31

4.6 14-day prediction for Chopera with cumulative incidence and traffic dataset 32

4.7 Traffic sensors for Virgen de Begoña scenario 33

4.8 Cumulative incidence for Virgen de Begoña scenario 34

4.9 Correlation for Virgen de Begoña variables 34

4.10 14-day prediction for Virgen de Begoña with cumulative incidence and traffic dataset 37

LIST OF TABLES

4.1 Benchmark results for Andrés Mellado 24

4.2 Traffic sensors results for Andrés Mellado zone 24

4.3 Traffic sensors and cumulative incidence results for Andrés Mellado zone 25

4.4 Benchmark results for Chopera 29

4.5 Traffic data results for Chopera 30

4.6 Traffic data and cumulative incidence results for Chopera 30

4.7 Benchmark results for Virgen de Begoña 35

4.8 Traffic data results for Virgen de Begoña 35

4.9 Traffic data and cumulative incidence results for Virgen de Begoña 36

1. INTRODUCTION

1.1. Context

The outbreak of the COVID-19 pandemic has supposed a big change in several aspects of life since the beginning of 2020. Unseen situations have been faced and the world has had to give responses to those scenarios, knowing that they could threaten the global health and economic conditions. One of the greatest difficulties that has been confronted is to determine how does the virus spread to estimate its potential impact. For that purpose, Artificial Intelligence has played a key role to give relevant insights about the virus and to complement science and epidemiology discoveries and decisions.

Machine Learning and Deep Learning techniques have the advantage of being adaptable to any kind of situation, and the uncertainty has been a common characteristic along the past years. Hence, they have been widely employed under many difference circumstances. This project focuses on predicting the future number of active cases of COVID-19 in a particular region, but some other predictions have been made about the virus structure, the sequences of symptoms, the vaccine development and other multiple areas.

Data availability has also been an important problem to be faced during this period. Initially, there were almost no data about similar past cases, so the datasets have been generated as the pandemic advanced. The collaboration between institutions and countries has been key for compiling, processing and publishing relevant datasets to study certain aspects of the pandemic.

1.2. Objectives

The breakdown of COVID-19 in Spain took place in March 2020, where a lockdown was decreed, lasting for up to two months in the region of Madrid, which is the region under study. After that, once the summer was over, the regional government took preventive measures to control the spread of the virus in the further waves. One of those measures was to restrict the traffic in the zones where the concentration of active cases of the virus

was bigger. The primary objective of this project is to study the relationship between the traffic data and the number of COVID-19 cases after the aforementioned measures were applied. This major objective can be broken down into a list of minor objectives, which are described below.

The project starts from a dataset containing the proper information to carry out this analysis. Data related to the traffic and the number of active cases needs to be compiled, in conjunction with information about the closed zones and the closure date. The only available sources in this case are open-source datasets. Hence, an adequate pre-processing structure needs to be designed for each of the variables, to set all of them into a common format to be subsequently combined for the analyses.

The study that is carried out does not cover the present days, as it focuses on the period between September 2020 and May 2021, where the closures of specifical areas were produced. Hence, all information is available, including the results. For this reason, it is crucial not to focus on what is known and try not to overfit the model to the final result, as the objective is to extract insights about the process instead of maximizing the final accuracy. Preliminary analyses can be carried out in the whole dataset, but the knowledge about the test information should not be used in the model training.

The relationship between traffic data and COVID-19 cases is modelled through Machine Learning models. These models need to be adapted to work with a time series structure. Different algorithms are compared, and multiple parameters can be tested, such as which is the optimal prediction horizon and which are the best input parameters for the model. Nevertheless, it must be stated that the COVID-19 spread is influenced by multiple factors that are not directly considered and that the availability of data is limited, as only nine months of data compose the dataset. Therefore, the results need to be put into context, as the objective is to extract the most relevant insights as possible instead of maximizing the prediction accuracy.

1.3. Structure of the report

The chapters that compose the report are briefly described in the following paragraphs, to ease the reading of the document.

Chapter 2 defines the role that Artificial Intelligence has played when dealing with the COVID-19 pandemic and compares the different algorithms and variables that have been employed to solve the raised issues. Some concrete examples and its results are also commented.

Chapter 3 describes the methodology that has been followed and justifies the decisions that are made along the process. The dataset obtention, the pre-processing and the structures of the analysis that are applied are described with examples. The different possibilities that are tested during prediction are commented. The zones under study and their characteristics are presented.

Chapter 4 includes a section for each of the zones, showing the appliance of the methodologies that are described in Chapter 3 separately for each of them and comparing the obtained results. The accuracies of the models are displayed in tables and the resulting optimal parameters for each of the models are commented with respect to those tables.

Chapter 5 summarizes the obtained results in Chapter 4 and compares it to the previously defined objectives. Finally, some future work proposals are briefly commented.

2. STATE OF THE ART

2.1. Artificial Intelligence and COVID-19

The uncertainty associated to COVID-19 left many questions without a clear answer given by science, so Artificial Intelligence emerged as a potential tool both to support the fight against the disease and to provide explanations that science may not be able to reach by itself [1].

The use of Artificial Intelligence models to tackle COVID-19 has led to the publication of many different papers covering multiple topics related to the disease. These topics relate to several applications, such as spread prediction, early symptoms detection, vaccine development or genomes evolution [2]. One of the greatest barriers to further research has been the quality and quantity of available data. As the pandemic has evolved, the applications of Big Data have eased the data collection from several countries, but the lack of standard datasets is still a challenge to be addressed [3].

The main areas where artificial intelligence is being used can be categorized as it follows [4], [5]:

- Spread forecasting. Prediction of the number of future infections in order to prevent further outbreaks of the virus. These predictions are made based on previous data and additional factors such as the news or the people sentiment.
- Medical evaluation. Optimization of the detection by PCR or antigen test and proper identification of the symptoms and their sequences of appearances.
- Drug development. Employment of past information about similar diseases to speed up the development and production of different drugs and vaccines.
- Contact tracing. Alternative to the traditional tracing method carried out by professionals, in order to handle the high number of cases and offer an alternative perspective to the obtained results.

Multiple algorithms can be employed, relating to different families. Generally, the

distinction is made between Machine Learning algorithms and Deep Learning algorithms. Some articles as [6] review the different academic papers that have been published, concluding that generally Deep Learning algorithms are more commonly used, reaching values around 75% of total. Machine Learning algorithms focus on Decision Trees and SVMs, while Deep Learning is more focused on Convolutional and Recurrent Neural Networks.

2.2. Virus spread prediction

The focus of this project is placed on the prediction of cumulative incidence, which is a measurement that is directly related to the number of active cases on a certain region. Epidemiology defines several concepts to measure the impact of a disease within different populations. Hence, to properly understand the terminology, the concept of cumulative incidence is defined. According to [7], cumulative incidence "expresses the risk of contracting an illness in a given population within a timespan", as it measures the number of active cases over a group of individuals that were healthy when the study started.

The analyses in this case are mostly based on time series forecasting, as the data is a temporary sequence. There are mainly two ways to approach the prediction of the number of active cases of the disease. The first way focuses on predicting the cumulative cases since the start of the disease, which hence focuses on predicting the slope of a curve that grows with time. The second way focuses on predicting the daily new infections, without considering the total sum [8]. Figure 2.1 shows these two approaches to predict the number of COVID-19 cases.

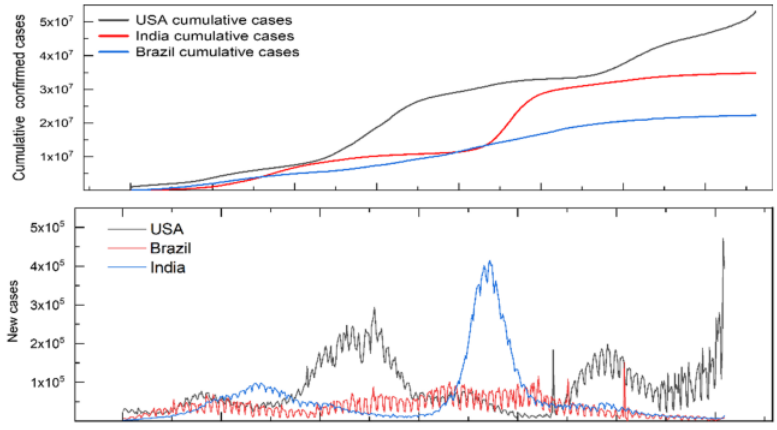


Fig. 2.1. Cumulative (up) and daily (down) COVID-19 cases [8]

Time series forecasting generally relies on two types of models, which are the statistic and econometric models and the aforementioned ML and DL methods. The previously commented survey papers barely mention the use of statistical models. However, there are some papers that test these kind of models to check whether they are useful or not to predict future COVID-19 cases. These models usually focus on predicting the cumulative cases since the pandemic start instead of predicting the daily evolution of cases.

Statistical models, even not having been widely used for COVID-19 prediction as it was thought that the situation was too complex and too dynamic to be predicted, still give good results when the parameters are properly configured [9]. Other good example can be found on [10], where ARIMA is compared to an LSTM. In this case, the error is greater for the ARIMA models, but it is considered as a reasonable result to support the LSTM predictions, as both predict an upwards trend in the number of cases. Hence, statistical and econometric models have been useful for COVID-19 prediction and to contrast the results of other models, or to give a baseline for the prediction of another model.

Artificial Intelligence methods offer different possibilities for prediction and estimation, as there is a great variety of algorithms with different natures that can give many different solutions to a unique problem. Nevertheless, those methods should be adapted to a time series forecasting scheme. The work in [11] studies the dynamic incorporation of information to Machine Learning models by including lagged data to the algorithm as a predictor, resulting in a clear improvement with respect to models that do not include it. Many distinct algorithms are found on literature, among which Support Vector Machines, Elastic Nets and Decision Trees can be found, fundamentally.

Among Artificial Intelligence algorithms, Deep Learning methods have been the most widely used, due to their capabilities of modeling dependencies on data and to deal with non-linearities [12]. Neural Networks have different structures, which have been carefully studied to determine which is the optimal model to deal with COVID-19 predictions. Recurrent Neural Networks (RNNs), and concretely Long-Short Term Memory networks (LSTMs) usually achieve the best performance in comparison with other types. [13] shows an example of the application of LSTMs and CNNs to the cases evolution in different countries, where LSTMs achieve the minimum MAE values. Several combinations of neural networks are evaluated also in [14], such as Gated Recurrent Units (GRUs), which

turn out to be a viable alternative to LSTMs.

As it has been commented before, the availability of datasets was limited, specifically at the pandemic start. As time has passed, historical data from different regions of the world has been collected and published in open-source datasets. The availability of information has increased due to the collaboration of different entities that have developed methodologies to gather data from different sources in order to unify it and place it in open-source repositories. One of these processes is described in [15], where data from different natures is compiled, processed and integrated in order to make them comparable. Figure 2.2 shows the flow of data collection, processing and publishing:

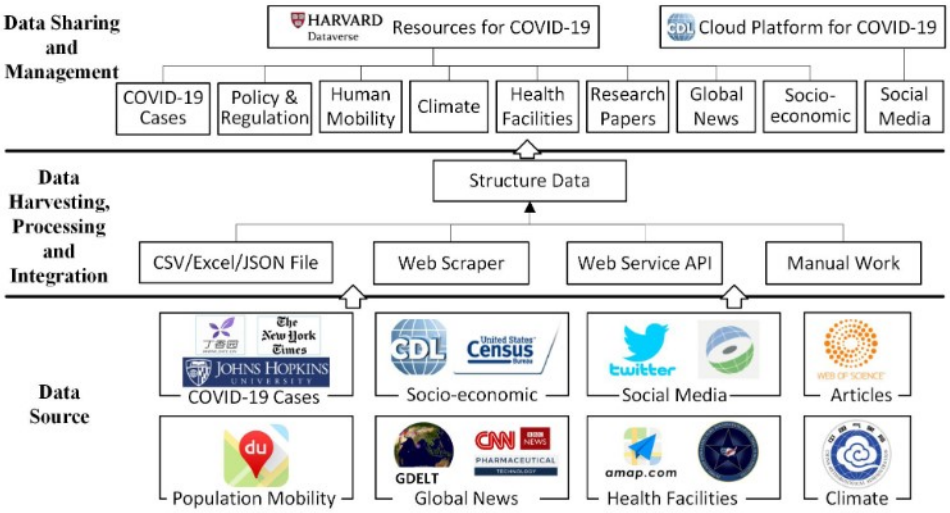


Fig. 2.2. COVID-19 data collection, processing and publication scheme [15]

Another example of information collection is the Coronavirus Resource Center developed by Johns Hopkins Univeristy [16]. The web site shows various dashboards including the current available information for different parts of the world to track the evolution of the pandemic in multiple regions of the world in real time. Data is also available open-source, including datasets for cases and deaths, vaccination, testing and other relevant topics.

Finally, an important aspect to be considered are the external factors that can not be easily modeled. These factors may also be dependant on every region separately, and some examples can be the appliance of counter-measures from the governments or the public sentiment. Hence, the cases evolution can be different depending on the region

under study. Some studies as [17] select countries with different dynamics to check how the prediction with the same algorithms can change depending on the particular characteristics of each of the regions under study.

2.3. The impact of COVID-19 on traffic and mobility

Mobility has been deeply affected by the outbreak of COVID-19. Lock-downs were one of the first measures to be applied to fight the virus spread, causing a drastic traffic and mobility reduction. Focusing on Spain, multiple datasets about mobility were published and analyzed, having different sources, such as public entities like the Ministry of Transport and private entities like Google and Apple. The European Comission published a report [18] where these open-source datasets were analyzed to study the mobility changes in Spain, concluding that there was a drop of 40%-50% in the first lockdown. After that, the following months did not reach the levels that were usual before the pandemic. For the following waves, lock-downs were not implemented anymore, but some alternative restrictions were imposed. Focusing on Comunidad de Madrid, which is the region under study in this project, the local government imposed the restrictions for the following waves based on Basic Health Zones, which are the regions associated to a unique public health center.

The academic papers have gone one step further, trying to relate the mobility data to the number of COVID-19 cases through Machine Learning models. [19] proves that including weather and mobility data as exogenous predictors for predicting the virus spread in Japan helps increasing the precision of results. For the case of big cities, mobility data seems to be enough by itself, while for smaller cities meteorological data is also relevant.

The estimation can be stated in the opposite way, by trying to predict how did the restrictions affect the mobility. Traffic volume estimation using Artificial Intelligence is a common prediction problem, and the models up to 2020 were not designed to handle such a great reduction. The models proved to adapt better to short-term predictions than to long-term predictions after a model and feature selection process [20].

During the pandemic, several new models have been created to try to adapt these previous models to the COVID-19 scenario [21]. These models try to find correlations

between the impact of the imposed restrictions and the traffic flows. The models allow carrying simulations of potential future scenarios that can be adapted to multiple situations that can happen on the future.

3. METHODOLOGY

3.1. The Dataset

The dataset that is used for this project comes from three different sources. Each one of them allows obtaining different pieces of information to enrich the content of the dataset. The sources are the Community of Madrid and the Madrid City Hall, which publish different datasets concerning multiple aspects that are open source. The three parts that compose the dataset for this project are the following:

- **Traffic data**. Quarter-hourly data for each of the traffic points of the city of Madrid. Several measurements are included, such as intensity, occupation or load. The points are distinguished depending on if they are on the city or in the bypass roads.
- **Cumulative incidence**. Weekly data for each of the basic health zones of the Community of Madrid. The measurements include the 14-day cumulative incidence and the total number of COVID-19 cases per zone.
- **Closed health zones**. Half-monthly data mined from the official PDF journals that the Community of Madrid published regularly to inform about the COVID-19 measurements updates at that time.

By looking at the description of the data, it is appreciated that all sources have different natures. Hence, pre-processing is needed to transform all data into a common format, so that it can be properly used to achieve the aforementioned goals. The objective is to have a daily instance for each of the studied zones, so the frequency of the data needs to be changed, either by upsampling or by downsampling. None of the mentioned data sources has this target structure, so several transformations will be applied to each of them. The pre-processing strategy for each of the parts and its intrinsic structure is described on the sub-sections below.

3.1.1. Traffic Data

The dataset is retrieved from the Madrid City Hall page [1], which contains the historical of traffic data measured by the city sensors since 2013. Each month of data is stored in a different CSV file, as there is a great number of measurements. Hence, to extract the desired dataset, the period between September 2020 and May 2021 is downloaded one by one.

The sensors are static, and they are placed on strategic points to extract different variables related to the traffic statistics. These variables are the following:

- **Intensity**. Number of vehicles that have passed through the measurement point over the measurement time, expressed in vehicles per hour.
- **Occupation**. Percentage of time that the sensor is occupied by a vehicle. This measurement can give an overview of how smooth has the traffic been at that period of time.
- **Load**. Estimation of the congestion in percentage. A high value implies heavy traffic while a lower value implies lower concentration of vehicles.

Another important parameter to be considered is the type of point. In this case, the dataset distinguishes between URB and M30 points, referring to the urban points and the bypass-roads points. The bypass roads concentrate different traffic flows from different neighborhoods, so they may not be a good option to measure the traffic flow between two districts, as the measurement will surely include traffic that is not related to these districts. Hence, only points labeled as URB are finally employed, as they can give more precise measurements of the traffic flow between two concrete districts.

The points are identified by a numerical ID, so it is not trivial to assign them to a district. To solve this problem, the Madrid City Hall page includes a group of files containing geo-spatial data related to the location of each of the points. This information can

[1] https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=33cb30c367e78410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

be easily overlaid on Google Earth, so that the location of the points is now easily identified. These files are updated monthly to include the newly added points for each month, so the file for September 2020 is downloaded, in order to ensure that the selected points have existed for all the desired months. Figure 3.1 shows an example of the structure on Google Earth:

Fig. 3.1. Example of traffic point location and description in Google Earth

The target of this project is to use traffic data to study the evolution of cumulative incidence of COVID-19. The measurement of cumulative incidence does only make sense for periods equal to or longer than a day. The studied dataset contains one sample per each 15 minutes, so these measurements are excessively fine-grained for the purpose of this project.

In order to adapt the traffic data to the cumulative incidence structure, the samples will be aggregated by day, obtaining the average traffic intensity for each of the days. The occupation and load are left as an alternative possibility in case the intensity does not add any relevant information to the problem or is not precise enough by itself. Figure 3.2 shows the result of the transformation.

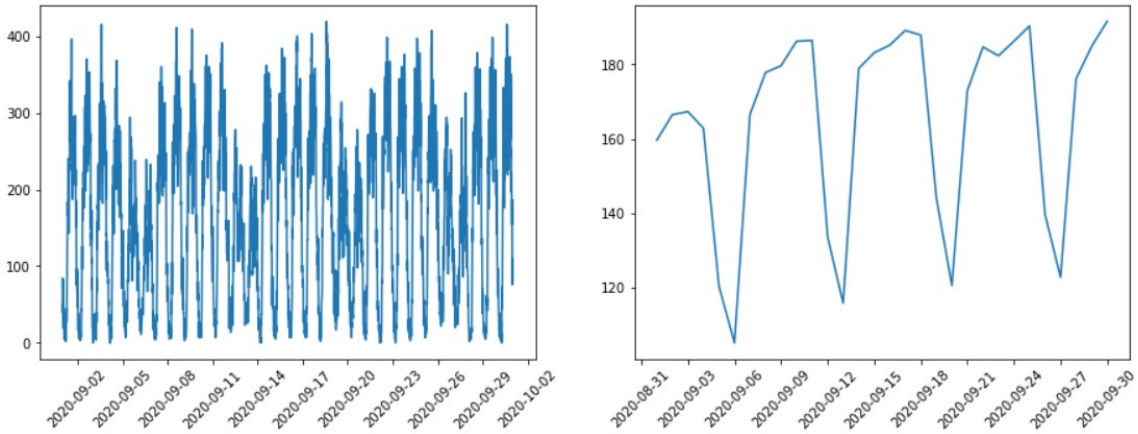


Fig. 3.2. Raw (left) and processed (right) traffic data for point with ID 10499

The sequence is now clearly smoothed, eliminating the daily seasonality that is present on the data, corresponding to the day and night periods. The lower peaks are related to the weekends, where the traffic volume is reduced. As an example for this particular case, the Sundays were the days 6, 13, 20 and 28. This shape does not hold exactly for all data points, but is appreciated that it is a common structure for urban points. The series can be smoothed in a greater way by taking the moving averages, to appreciate the general trend of the series, that for the case of the shown point seems to increase along September.

3.1.2. Cumulative Incidence

The cumulative incidence is available at the Community of Madrid official data page [2]. This dataset contains the weekly evolution of the cumulative incidence for the different health zones of the community, which are those units of territory that are associated to a public health center. Therefore, these health zones are not the same as the districts or neighborhoods of the city, because they depend on the location of the health center.

The impact of COVID-19 is evaluated by comparing the different health zones, so these zones need to be linked to the traffic points to evaluate the impact of traffic intensity over the cumulative incidence of the zone. To do so, the Community of Madrid has available the same files that were available for traffic data, so that the shape of the zones can be

[2] https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_zonas_basicas_salud

overlaid on Google Earth. By overlapping both layers, the traffic points associated to the frontier of two zones are easy to identify. Figure 3.3 shows the traffic points that can be identified as the frontier points from the zone of 'Las Tablas' to the zone of 'Sanchinarro', surrounded by a white circle. The road separating both basic health zones is considered as a bypass road, so no points related to it are considered.

Fig. 3.3. Traffic measurement points between the zones of Las Tablas and Sanchinarro

This procedure can be systematically is applied in this project to different health zones, to assess the influence of traffic data over the cumulative incidence within different scenarios.

The peak of the pandemic was reached on March and April 2020, so the cumulative incidence was monitored on a daily basis. The available data is hence daily until July 1st 2020. Since then, the dataset changes to a weekly frequency, which is the structure that will be used as the starting point for this project. The common format that has been established for all pieces of data has a daily frequency, so in this case interpolation is needed to oversample the original series, as oppose to the case of traffic data.

The initial dataset is relatively small, as the period under study comprises nine months, so there are only 39 samples, one per week. The initial shape is hence very sharp, as the cumulative incidence can grow in a really fast way. The interpolation, thus, is used both to retrieve more samples from the original series and to smooth its structure. The interpolation method that has been selected is *pchip*, which relates to a cubic Hermite

polynomic interpolation. This method is smoother than *linear* interpolation, but not as smoother as a *spline* interpolation can be. The problem that has found with *spline* is that it gives negative values to the cumulative incidence, which is not correct, as it is defined to have a minimum value equal to zero. Hence, *pchip* is selected as it is able to smooth the function without reaching negative values. Figure 3.4 shows the appliance of the mentioned interpolation method to a concrete health zone.

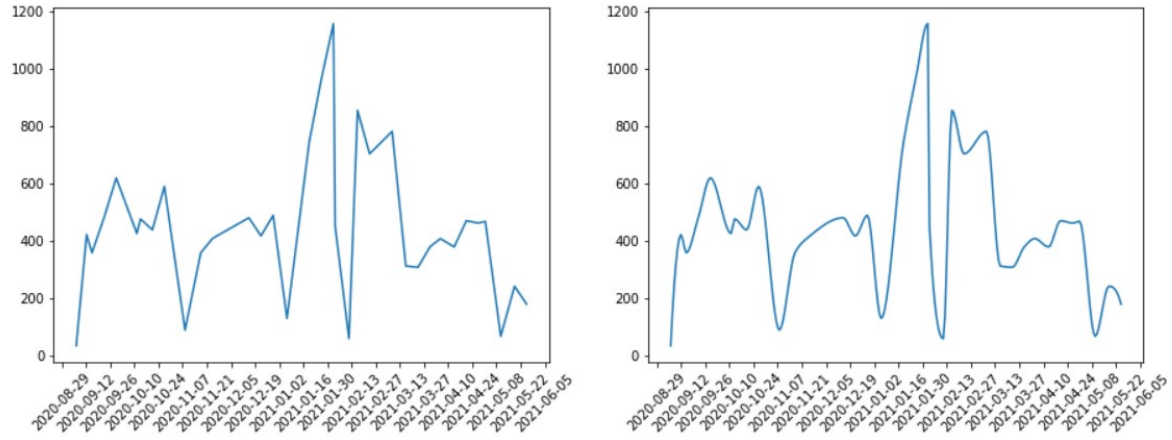


Fig. 3.4. Raw (left) and processed (right) cumulative incidence data for Andrés Mellado health zone

The peaks are smoothed, but preserving the original structure of the data, and the limits are respected as no points reach values under zero. The function is ready to be consistently applied to all health zones. The other interpolating methods are left as an alternative to explore in the case that *pchip* does not work well.

3.1.3. Closed Health Zones

During the last months of 2020 and the first half of 2021, several measurements were taken by the regional governments in Spain to try to reduce the impact of COVID-19 along its different waves. In the case under study, which is the city of Madrid, the Community of Madrid decided to implement restrictions on the movement of the citizens across different health zones. These restrictions allowed closing certain health zones, depending on their cumulative incidence. The restrictions were updated every two weeks, so the minimum period of closure was equal to two weeks.

There is not an official dataset containing the exact dates of these restrictions. The

only possibility then is to mine this information from the Internet, either from the news of that days or looking for official sources. The closed zones were regularly published on the Official Bulletin of the Community of Madrid [3]. Hence, by downloading the corresponding PDF documents which are available online, the dataset can be easily mined. All the documents have a common pattern, so with a simple PDF mining Python function, the dataset is generated.

The frequency in this case is half-monthly, so it needs to be oversampled to match the characteristics of the global dataset. In this case, no interpolation is needed, as only two states are considered: zone without restrictions (0) and zone with restrictions (1). This data is useful to check whether the restrictions did actually help to lower the cumulative incidence in the affected health zones, and if the traffic did actually decrease from the usual values.

3.1.4. Data Selection

The complete dataset has a great size, as it contains the data from all urban traffic sensors and all the basic health zones. As it was commented before, the sensors and the health zones are easily linked thorugh the KML files that are displayed over Google Earth. Working with such a large quantity of data can be very time-consuming and would require an in-depth analysis. For this project, three different situations are studied, related to three different health zones characterized by various conditions in terms of closure dates. Three frontiers between two basic health zones, the frontier traffic sensors and the corresponding cumulative incidence data is selected from each of the three following situations:

- **Guzmán el Bueno to Andrés Mellado**. Urban districts separated by a narrow street, with multiple frontier points and four frontier sensors. Guzmán el Bueno zone was closed prior to the Andrés Mellado closure, so a potential cumulative incidence transfer from the first zone to the second can be studied. The closure was decreed in the month of December 2020.

- **Legazpi to Chopera**. Urban districts separated by great streets, with a longer and irregular frontier. Legazpi zone was never closed, whereas Chopera zone was closed

[3] https://www.bocm.es/

from March 2021 and May 2021. This situation can help evaluating if there could be a potential cumulative incidence transfer between a non-closed zone and a closed one.

- **Fuencarral to Virgen de Begoña**. Suburban districts separated by a street and surrounded by a bypass road. The characteristic fact of this scenario is that Virgen de Begoña was closed twice, the first time in October 2020 and the second time in March 2021. In this case, it can be studied how the cumulative incidence is impacted by a single point of contact, or how being close to a bypass road can impact the cumulative incidence of a zone.

The traffic sensors are selected manually over the map, as they are not related by any information field to the basic health zone that they are close to. Two types of sensors can be distinguished depending on their relative position with respect to the zone under study. External sensors are those sensors located outside of the zone, and where the traffic flows towards the district, and internal sensors are those ones located inside the zone. Combining the information from both types of sensors, the incoming and the internal traffic flows can be monitored. Pictures of the frontiers and the criteria to select the points are included in Appendix 1.

3.2. Graphical Analysis

The first step after pre-processing is to carry out a visual analysis of the data, to get a preliminary idea of its structure and analyze its behavior. All the analysis that is described in this section will use the data from Andrés Mellado zone as the reference, to have a consistent analysis. The analyses for the remaining zones is included in Chapter 4.

The analysis consists on three basic steps. The most important result to be determined in this section is if the traffic volume related to a concrete zone was affected by the restrictions, and then assess if this impact was also reflected on the cumulative incidence. The compliance of the first step is essential for the second step to ensure that the second step is meaningful.

There are four sensors in this case that have been identified as incoming traffic from Guzmán el Bueno to Andrés Mellado. These sensors are plotted in a common graph,

where the closure period is shadowed in a darker color to be distinguished from the rest of the days where there were no restrictions. As it was previously mentioned, the traffic has weekly seasonality, so a moving average for ten days has been applied to the series in order to retrieve the trend that is present on the data. In this way, the analysis is easier. The results are shown in Figure 3.5.

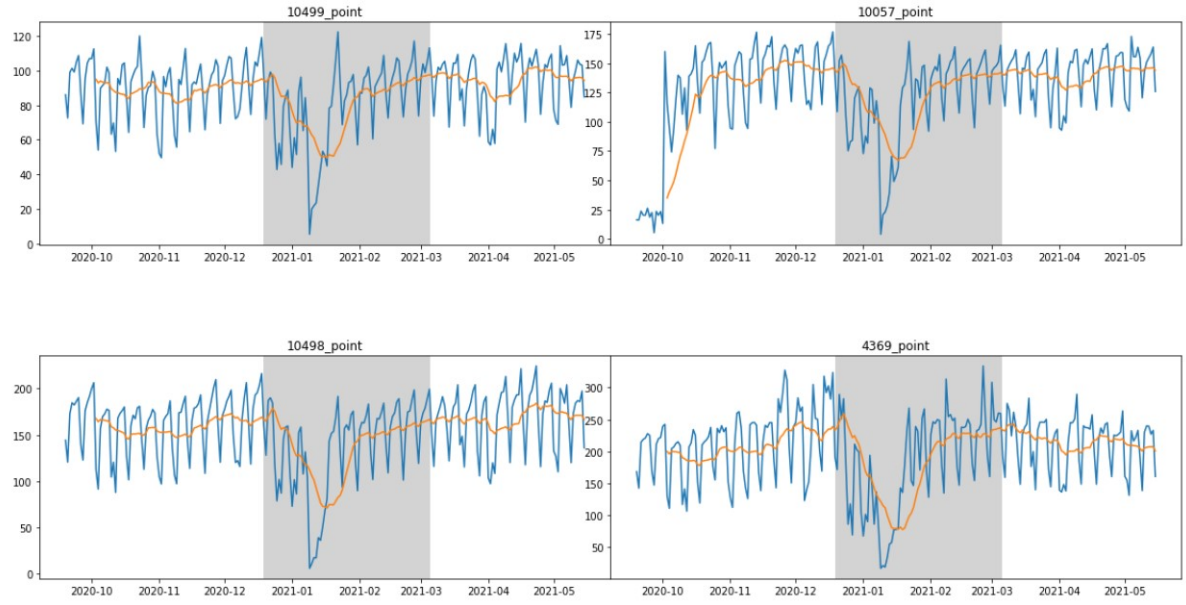


Fig. 3.5. Traffic sensors for Andrés Mellado scenario

The pictures have to be carefully analyzed, as at first glance it seems that the restrictions were clearly respected. Nevertheless, around January 10th 2021 there is a huge down peak, associated to Filomena squall, which caused continuous snowfalls that cut off the streets for a few days. This effect is observed for all districts, so it is something to be taken into account. However, it is appreciated that in general the average traffic tended to be reduced when the restrictions started for this particular case, so this effect could be useful as an input for the cumulative incidence prediction. Then, the second step can be carried out.

The other type of series to be analyzed is the cumulative incidence. As the objective is to measure a potential influence of the cumulative incidence of a zone in its neighbour zone, Guzmán el Bueno and Andrés Mellado are both studied for the following example. Figure 3.6 shows the interpolated cumulative incidence curves for both health zones.

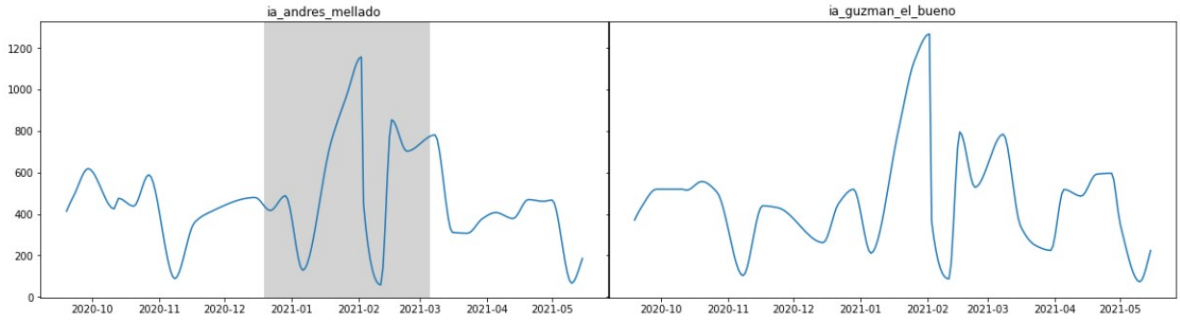


Fig. 3.6. Cumulative incidence for Andrés Mellado scenario

The picture shows that both curves have a very similar shape, with a great peak on January 2021, where the incidence reached its maximum peak of cumulative incidence since the pandemic was declared. Although Guzmán el Bueno zone was closed earlier than Andrés Mellado, it seems that both zones had a similar level of cumulative incidence. These similarities can imply that the Guzmán el Bueno curve can be a good predictor for Andrés Mellado cumulative incidence.

Finally, the correlations between the different time series can be measured. The results can give an idea about whether there is any relationship between the series, specially between traffic data and cumulative incidence data. Figure 3.7 shows a heat map with the correlations for all variables.

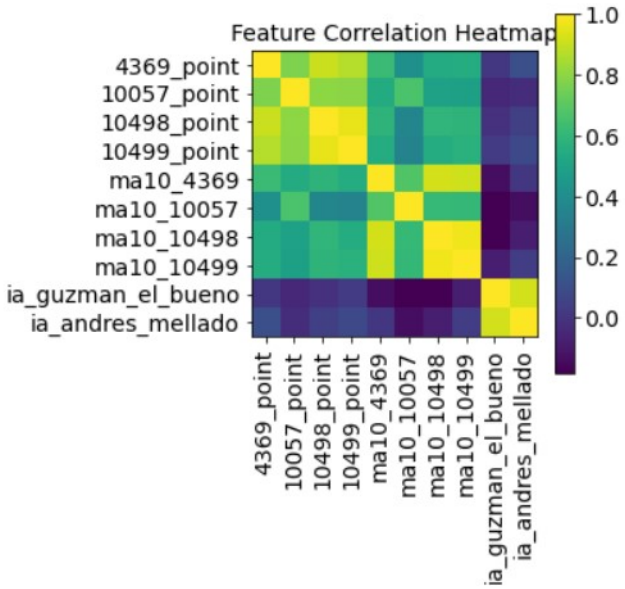


Fig. 3.7. Correlation for Andrés Mellado variables

For this particular case, it is appreciated that the traffic series have a high correlation between themselves, as all of them have a similar seasonal behavior. The cumulative

incidences are correlated between them too, as they are very likely as observed on the previous step of the analysis. Nevertheless, the correlation between traffic series and cumulative incidence is close to zero, specifically in the case of the moving averages of the series. These results will be considered for the next section to take decisions about which predictors to include or not.

3.3. Machine Learning Analysis

The following analysis to be carried out consists on using Machine Learning models to test how predictable is the cumulative incidence by the traffic data. To do this, several tests are carried out in order to check whether the prediction error is lowered or not depending on the variables that are under consideration. The tested models can be categorized into three sub-categories for each of the zones under study:

- **Benchmark**. Univariate model that uses the cumulative incidence series itself to predict its futures values. As no additional information is added, this model is useful to test if traffic predictors are useful to refine the predictions or not.
- **Traffic predictors**. Multivariate model that uses the traffic predictors as exogenous variables to predict the cumulative incidence, including all the frontier points.
- **Contiguous zone cumulative incidence**. Multivariate model that adds the cumulative incidence from the contiguous zone to the traffic predictors to assess a potential performance increase.

The problem can be tackled in many different ways. The chosen option for this project is by using the *skforecast* [4] library, which is a *Python* library that allows adapting the usage of the common *sklearn* predictors in a time series framework. Hence, the cross-validation methods are adapted to a temporal data structure, so that the samples are not randomly shuffled. The number of past samples to be taken is also tunable, so the importance of different past instants can be compared to determine which is the best window to predict the cumulative incidence. Therefore, the hyperparameter tuning consists on both tuning the algorithm hyperparameters and the number of lags to be used to make the predictions.

[4] https://joaquinamatrodrigo.github.io/skforecast/0.4.3/index.html

The data is firstly normalized, in for the results to be compared in an easier way. The goal is to simulate a real scenario that could have happened during the pandemic, so that predictions need to be made at least five days ahead, as one-day-ahead predictions may not give relevant information. Several prediction steps are evaluated, starting by one day just to use it as a benchmark and then see from there how the accuracy worsens as the number of days ahead increase. The evaluation metric that will be used to compare all models is the Mean Absolute Error (MAE). The predictions are evaluated for five, seven and fourteen days ahead. The results are displayed at Section 4.

The dataset has a total of 240 samples, so it can be considered as small. The problem is hence complex, as the synthetical generation of samples is not a feasible option, because the data for the following months is no longer affected by traffic restrictions. The structure of the data that is used as the input for the algorithm varies depending on the selected zone. All columns have 240 samples as stated before, covering from September 15th 2020 to May 15th 2021. The number of columns for each of the datasets is composed by the traffic intensities for the selected traffic points and the moving averages for these points. In addition, the cumulative incidence for the zone under study is also included in conjunction with the cumulative incidence for the contiguous zone. Finally, the periods of closing are included as a binary variable, but they are not included in the input for the algorithms.

The cross-validation methods that will be employed are based on backtesting strategies, where only past samples are employed for making predictions. Two different cross-validation models are evaluated:

- **Backtesting without refit**. The model is trained, and after the training stage finishes, the predictions are made sequentially. The immediately following samples to the training test can be normally predicted, but the predictions may be downgraded for further predictions, as no information is added. This method is lighter in terms of computational cost than the following one, so it could be set as a benchmark to compare with the refit scheme.

- **Backtesting with refit**. The model is re-trained in every single step before making predictions, so that it is likely to be more accurate. The method is heavier computationally talking in comparison with the previous option. This strategy would also

have been adequate for carrying out these predictions during the pandemic, as the dataset is small and more information could have been added as weeks passed by.

The confidence intervals are also calculated, as it is known that the problem has a high degree of complexity. Then, in such an uncertain scenario, it is important to estimate how the signal will behave, and having the 95% confidence intervals helps estimating its range of movement. This approach is not as precise as estimating the signal itself but it can give more useful results.

Finally, several regressors are compared, as their results can give different insights about the structure of the data. The intrinsic structures of these regressors are different in some of the cases, to give different perspectives to the problem. The employed regressors relate to two families: support vector machines and gradient boosting. The employed regressors are XGBoost, LightGBM and SVR. The results are shown and discussed on the following section.

4. RESULTS

The results are divided in three sections, corresponding to each of the defined scenarios. Within each of these three sections, the results for each of the scenarios is independently commented. The procedures that are followed are those described in the preceding chapter. The algorithm parameters are the same for all cases, as the goal is to test all scenarios under the same conditions.

4.1. Andrés Mellado and Guzmán el Bueno

The graphical analysis for this zone is not included in this chapter, as it has been used in Section 3 to describe the methodology that is used.

4.1.1. Machine Learning Analysis

The first step before the data is entered to the algorithms is to define the train-test strategy. The dataset is split into the train and test sets. As the number of samples is low, the size of the test set needs to be carefully chosen in order to avoid either having too little information to train the model or to having too small a test set that does not give a meaningful result. The chosen option is to leave two months as the test set, while the others are defined as the train set. This results in a 75%-25% scheme, with the train set starting on September and the test set starting on March. The same strategy is applied for the two scenarios left.

Once the data is split into train and test sets, the next step is to evaluate the model. The first approach consists on defining a benchmark for the rest of the models to compare against. As described on the previous section, this approach is an univariate time series model where the input is the cumulative incidence of the model itself. The Table 4.1 shows the MAEs obtained with the different algorithms with and without refit.

	Without refit				With refit			
Period	1 day	5 days	7 days	14 days	1 day	5 days	7 days	14 days
XGBoost	0.0976	0.6209	0.6414	0.6484	0.0438	0.1595	0.2871	0.6545
LightGBM	0.2661	0.4460	0.4584	0.4595	0.4595	0.3576	0.3891	0.4170
SVR	0.0385	0.1490	0.2564	0.2714	0.0380	0.1554	0.2487	0.3227

Table 4.1. BENCHMARK RESULTS FOR ANDRÉS MELLADO

The results are as expected, as the error increases for wider periods, and as re-fitting the model for every prediction step seems to lower the error, but at the expense of a higher computational cost. Support Vector Regression is the best-performing algorithm by far with the defined parameters, and it is the only one that is not able to improve the results when re-fitting the model. These results are set as the reference to compare against the results displayed on Table 4.2, which shows the obtained results when the cumulative incidence is predicted only considering the traffic sensors and its moving averages:

	Without refit				With refit			
Period	1 day	5 days	7 days	14 days	1 day	5 days	7 days	14 days
XGBoost	0.1210	0.2917	0.4307	0.6486	0.0639	0.1893	0.6352	0.6432
LightGBM	0.2656	0.4269	0.4428	0.4588	0.1022	0.3638	0.2870	0.4188
SVR	0.0439	0.1717	0.2381	0.2699	0.0404	0.1723	0.2383	0.2338

Table 4.2. TRAFFIC SENSORS RESULTS FOR ANDRÉS MELLADO ZONE

The results are not consistent for all algorithms, as there are some cases in which there is some improvement in comparison with the benchmark, while there are other results that worsen. Generally speaking, it seems that the traffic data inclusion is misleading in this case, and this may be due to the aforementioned extraordinary conditions suffered during the closure of the district. The effect of the snowfalls produced an atypical traffic intensity drop, while the effect of Christmas holidays caused the highest peak of cumulative incidence along the pandemic. This impact can only be modelled a posteriori, so it would not be consistent to introduce a correction in data, as it would modify the characteristics of the dataset. Nevertheless, some positive results are extracted from this approach, such as the 7-day prediction for the refit cases, where it is a general improvement with respect

to the previous scenario. Finally, Table 4.3 shows the results for the prediction including the cumulative incidence of the contiguous health zone:

	Without refit				With refit			
Period	1 day	5 days	7 days	14 days	1 day	5 days	7 days	14 days
XGBoost	0.1721	0.4252	0.4266	0.3762	0.1810	0.3138	0.2724	0.3817
LightGBM	0.2566	0.3588	0.4667	0.5001	0.0905	0.2113	0.2687	0.3570
SVR	0.0471	0.1985	0.2732	0.2881	0.0450	0.1730	0.2659	0.3013

Table 4.3. TRAFFIC SENSORS AND CUMULATIVE INCIDENCE RESULTS FOR ANDRÉS MELLADO ZONE

The results are generally better in this case for gradient boosting algorithms, while SVR loses performance in comparison with the previous case. Yet the SVR performance is the best among all models.

An interesting way to analyze how the algorithms behave is to monitor the parameters that have been chosen by each one to get the displayed results. As commented before, the hyperparameters for this project are integrated by the parameters of the algorithm and the number of lags that are chosen to make the prediction. The algorithm hyperparameters can give an overview of how complex is the solution that is being chosen. On the other hand, the selected lags can be used to determine how important are each of the past samples for the algorithm. If the algorithm selects a high number of samples from the past and the error is low, it may indicate that the further past information is useful. If a lower number of past samples is selected, it could tell that the signal changes in such a way that it is only predictable by considering the most recent samples. For this particular case, the best-performing models associated to SVR generally consider only the five latest lags to make their predictions.

The confidence intervals are also an useful tool to determine how certain are the predictions that are being made about the result. The situation at that time was highly unpredictable, so the expectations needed to be adapted to an achievable target. Hence, although predictions may not be really precise, it is important to accurately predict the range of variation of the cumulative incidence. A brief example is studied for a good-performing model related to this health zone. The selected model is the SVR with refit

for a 14-day prediction with the traffic dataset, and the predictions are shown in Figure 4.1

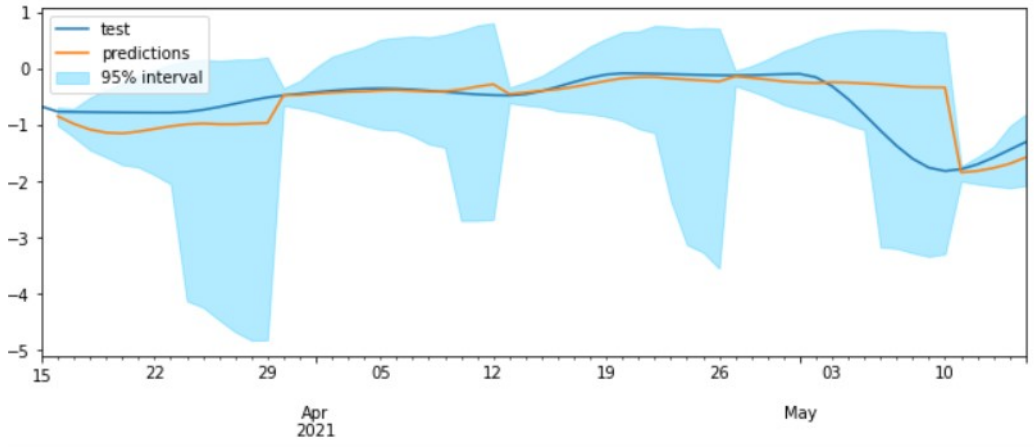


Fig. 4.1. 14-day prediction for Andrés Mellado with traffic dataset

The picture shows that the prediction line follows approximately the trend of the cumulative incidence. Hence, the predictions are good. Considering the confidence intervals, it is appreciated that they are really wide, so that they contain the range of movement of the signal, but at the expense of giving a very big potential range of movement. This implies that the algorithm is not really certain about the actual value of the cumulative incidence. It is also appreciated that, due to the refit, the predictions are more certain for the first days of the 14-day period and the intervals are wider for the last days of each period.

4.2. Chopera and Legazpi

4.2.1. Graphical Analysis

The first step is to describe the situation to be studied. Chopera is the zone under study and Legazpi is its contiguous zone. The frontier between both zones is wide and irregular, so several streets need to be monitored. The Legazpi zone did never have restrictions during this period, so the objective in this case is to check if there was a potential cumulative incidence transfer to its contiguous zone, which had restrictions during April 2021 and May 2021. These restrictions took place on the last part of the restriction period as opposed to the Andrés Mellado zone. Hence, the difference between early and late restrictions can also be assessed. Figure 4.2 shows the traffic evolution during the selected period:

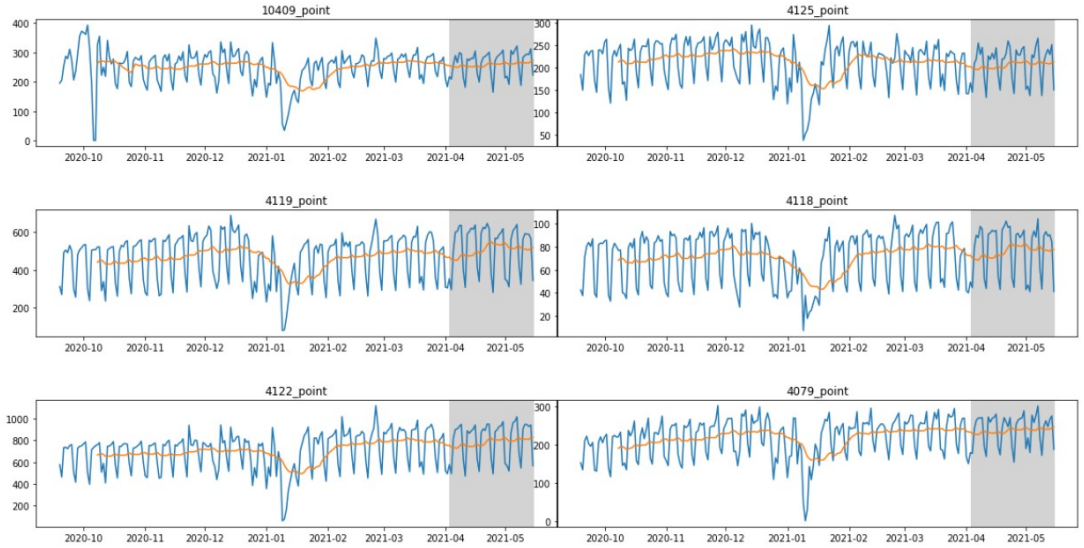


Fig. 4.2. Traffic sensors for Chopera scenario

The traffic data evolution in this case is different in comparison with the Andrés Mellado scenario. Nevertheless, common patterns are observed, such as the data seasonality on weekends and the impact of Filomena squall in January 2021, which causes a significative traffic drop. The average traffic intensity during the period with restrictions is similar to the average on the previous period, except for the previously mentioned situations. At the start of the restrictions period, it is slightly lowered for most of the cases, but no remarkable impact is appreciated. The objective of the predictive analysis is to determine if this small change in the average traffic impacted the cumulative incidence and is useful to predict it. The evolution of cumulative incidence is shown in Figure 4.3:

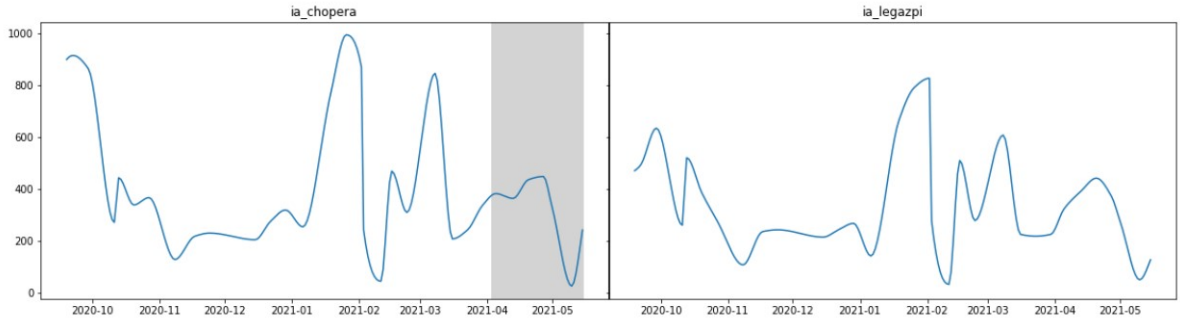


Fig. 4.3. Cumulative incidence for Chopera scenario

The cumulative incidence evolution is similar to the rest of the zones, as at the end of the day, all are a part of the same city. The different waves can be observed in all graphics, but in this case it is seen that the peak values are lower than in the case of Andrés Mellado. This may be the reason why there were no restrictions in these zones in January, as there were other districts with higher cumulative incidence values. The period of restrictions is now commented. It is appreciated that immediately after the restrictions started, the cumulative incidence in Chopera started to be stabilized, unlike in Legazpi, where it kept constantly growing until the peak was reached. These changes correspond to the initial moment of the restrictions, where the traffic volume was slightly reduced. Finally, Figure 4.4 shows the correlation among all the shown series.

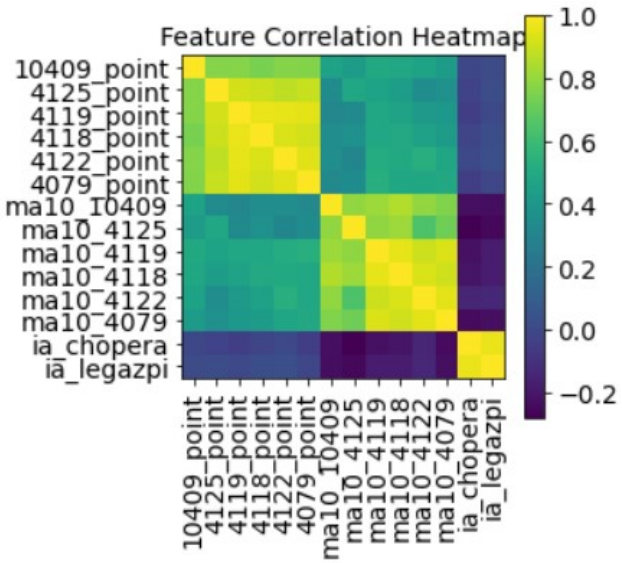


Fig. 4.4. Correlation for Chopera variables

The correlation between the cumulative incidence and the raw traffic data is slightly higher than in the Andrés Mellado case, and the moving averages are almost uncorrelated

with the cumulative incidence. The traffic points have a strong correlation between them, so a trade-off between simplicity and information can be assessed, as eliminating some points can be beneficial in terms of computational cost, and at the expense of losing very little information. The correlation between both cumulative incidence series is also remarkably high. The predictive analysis can now be carried out.

4.2.2. Machine Learning Analysis

The data is split into train and test with the previously defined 75%-25% distribution. The benchmark model is now trained. Table 4.4 shows the results for this first model.

	Without refit				With refit			
Period	1 day	5 days	7 days	14 days	1 day	5 days	7 days	14 days
XGBoost	0.0829	0.1912	0.2737	0.3127	0.0699	0.1276	0.2343	0.3584
LightGBM	0.0997	0.2390	0.2459	0.3125	0.0856	0.2113	0.2865	0.3409
SVR	0.0429	0.1802	0.2107	0.3260	0.0398	0.1855	0.1869	0.3664

Table 4.4. BENCHMARK RESULTS FOR CHOPERA

The results seem to improve in comparison with the Andrés Mellado scenario, as all the algorithms are able to achieve better results for most of the cases. Re-fitting the model does also improve most of the results, reaching some reasonably good MAEs, such as a 0.1276 error for the 5-day case with XGBoost. By checking the selected hyperparameters for each of the algorithms, it is appreciated that most of the models choose the same number of lags for predicting the cumulative incidence. The models with shorter horizons focus on the five past lags, while some of the models choose the fifteen previous lags to make the predictions for the longer horizons. This behaviour is now studied for the following models, to check whether there is a match between their results and the benchmark. The results for the predictions with the traffic data are displayed on Table 4.5:

	Without refit				With refit			
Period	1 day	5 days	7 days	14 days	1 day	5 days	7 days	14 days
XGBoost	0.1443	0.2170	0.4778	0.5137	0.0907	0.1846	0.2743	0.4345
LightGBM	0.1271	0.2703	0.2459	0.2705	0.1120	0.2637	0.2791	0.2263
SVR	0.0938	0.3029	0.3882	0.4609	0.0521	0.2462	0.5308	0.3319

Table 4.5. TRAFFIC DATA RESULTS FOR CHOPERA

The results are in general worse than in the previous case. The introduction of traffic data and its moving averages seems to confuse the algorithms, that tend to significantly increase the error for the 5-day and 7-day scenario specially. The refit helps stabilizing the results but still they are higher than in the previous case. The positive point from this section is that LightGBM achieves a good prediction for a 14-day scenario, with an error equal to 0.2263. The models tend now to focus on further lags to make the predictions, in general taking up to lag fifteen, and in some cases taking up to five-spaced lags from lag five to lag thirty. It is remarkable that the aforementioned better-performing model does only consider the previous lag to make predictions. This result can indicate that the best way to predict this cumulative incidence may be using very few lags and eliminating information from the further past. This hypothesis can be contrasted by inspecting Table 4.6, which contains the results for the models including the cumulative incidence and the traffic data.

	Without refit				With refit			
Period	1 day	5 days	7 days	14 days	1 day	5 days	7 days	14 days
XGBoost	0.0891	0.2732	0.2944	0.2820	0.0749	0.2178	0.2410	0.2788
LightGBM	0.1349	0.1612	0.1772	0.1929	0.1120	0.2637	0.2791	0.2263
SVR	0.1229	0.2196	0.2107	0.2174	0.0711	0.1467	0.2507	0.2535

Table 4.6. TRAFFIC DATA AND CUMULATIVE INCIDENCE RESULTS FOR CHOPERA

The results generally improve in comparison with the case only considering traffic data. Refit is still useful to improve the results for some cases, but LightGBM stands out for the opposite reason, as the results worsen when re-fitting the model. Nevertheless,

its results without refit are much better in comparison with the rest of the algorithms. With respect to the taken lags, the situation holds with respect to the previous cases. The best-performing algorithm overall is LightGBM without refit, which takes only one lag as input data. Meanwhile, the other models do not improve and they generally take more lags as predictors, eventually reaching lag 30 in five-lag steps.

Once all models have been separately analyzed, some general conclusions can be drawn from this scenario. It seems that including past information does not help the algorithms improving their predictions, as the best-performing models for each of the datasets are based at most on the five previous lags. Hence, focusing on a small period of past time may be enough to get a good result for this scenario.

The 95% confidence intervals for the predictions are now studied. To determine if the objective of properly predicting the range of variation of the signal has been achieved, the best models can be analyzed. This can allow double checking their precision and if their results are really useful. Firstly, the results of the 14-day-ahead LightGBM model with refit and using only the traffic data are shown in Figure 4.5.

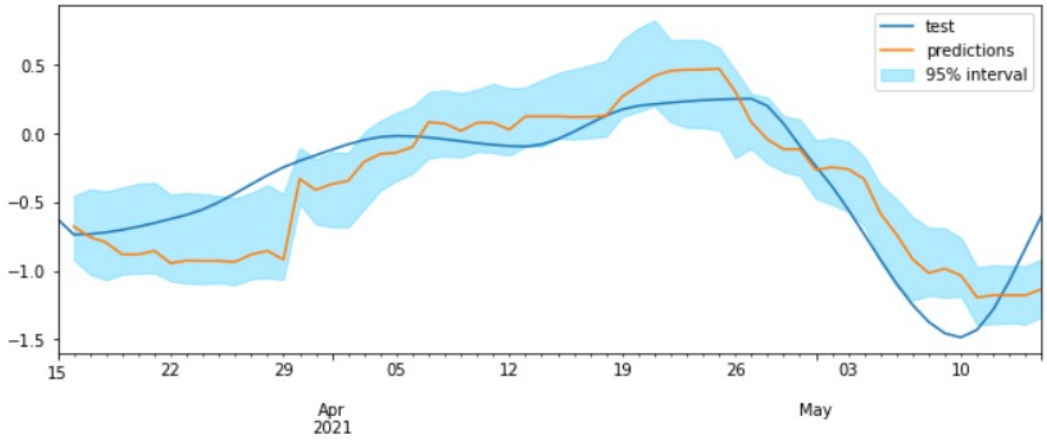


Fig. 4.5. 14-day prediction for Chopera with traffic dataset

The Figure shows that the original signal is contained within the confidence intervals for most of the inspected time. Although these instants seem to be higher than the 5% of time which is expected from a 95% confidence interval, the result is positive considering the high variability of the problem. To have another example, the situation can be evaluated for the dataset including the cumulative incidence, where there is more information than in this particular case. Figure 4.6 shows the curve for the 14-day ahead XGBoost

model with refit, which has a greater error in comparison with the previous case but it is not far from it.

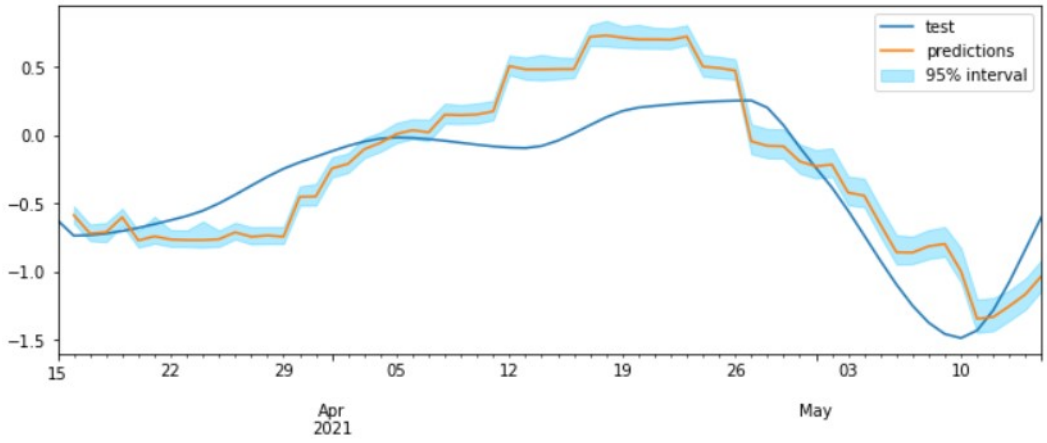


Fig. 4.6. 14-day prediction for Chopera with cumulative incidence and traffic dataset

This case shows the opposite example to the previous one, as the error is acceptable, but the confidence intervals are really narrow, so the actual values of the test data are not contained within its range. Although the shape is similar to the curve, the range of variation of the signal is not properly bounded.

4.3. Virgen de Begoña and Fuencarral

4.3.1. Graphical Analysis

The analyzed situation differs clearly from the two previous cases. This scenario has two particularities. The first one is that the closure was decreed twice, one at the start of the period and the second one near the end. This allows checking whether there was or not any difference on the impact of closures according to the time that they were imposed. The second particularity is that these districts are in the outskirts of Madrid, close to a ring road. In this way, the impact of external traffic on the zone's cumulative incidence can be assessed. The frontier between both zones is a straight street, where the points of contact are easy to identify. Now, Figure 4.7 shows the evolution of the traffic sensors measurements along the chosen period:

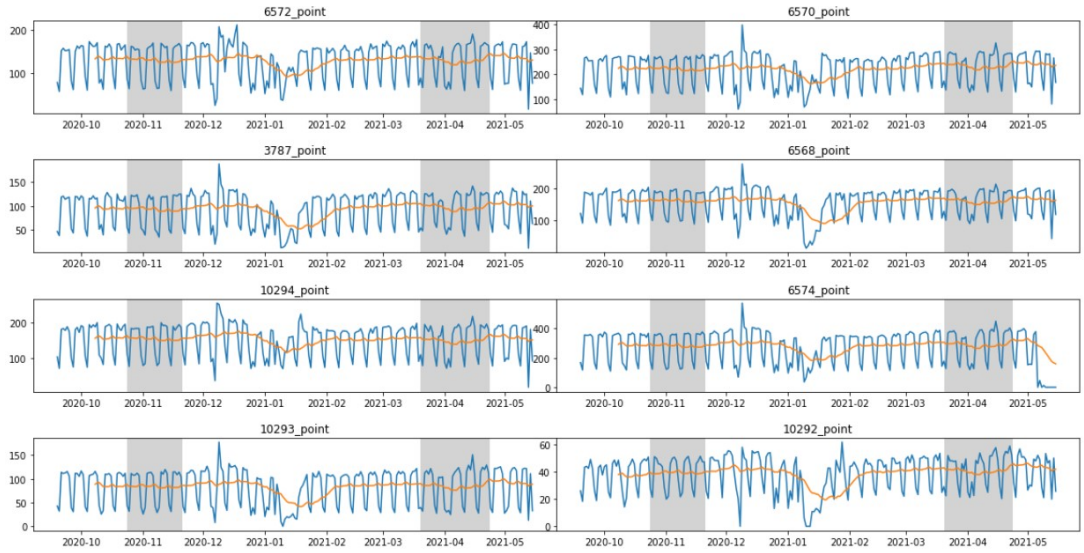


Fig. 4.7. Traffic sensors for Virgen de Begoña scenario

The traffic data shows a very regular behavior along the studied period, as it keeps almost constant excepting for the aforementioned special situations, such as the snowfall from January 2021. The earliest restrictions seemed not to have impacted the traffic, since it remains almost constant. The latest restrictions seem to have implied a small drop, that was corrected after a couple of weeks. The scenario is confusing, as the restrictions did not seem to have a clear downward effect on the traffic data. The machine learning study helps determining if these variations are enough to predict the cumulative incidence. The cumulative incidence curves are shown in Figure 4.8.

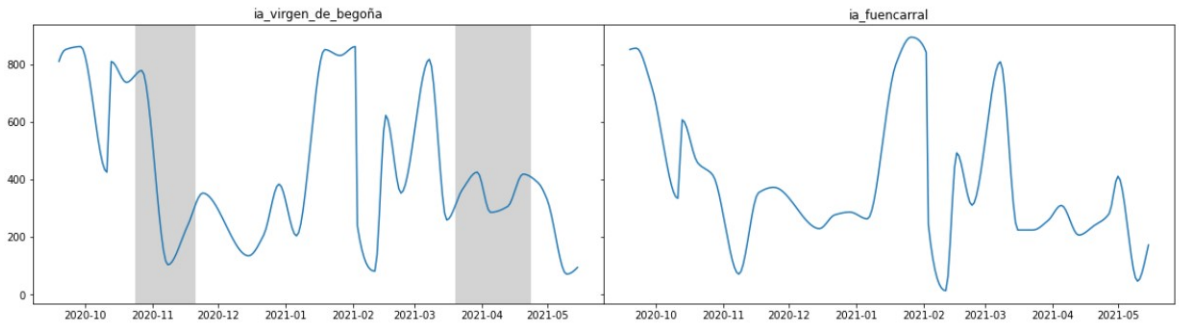


Fig. 4.8. Cumulative incidence for Virgen de Begoña scenario

The cumulative incidence curves are similar to each other, as in the previous cases. The first period of restrictions shows that there was an upturn in Virgen de Begoña that did not affect Fuencarral, so this may be the reason why one district was closed while the other was not. The January peak is lower in comparison with the other two scenarios, reaching values around 800 while for the other zones under study these curves eventually reached values near 1000-1200. The same thing happens for the second period of restrictions, where the peak is higher again for Virgen de Begoña, and the restrictions are applied for the second time. In addition, it is observed that the shape of the curves does not differ significantly from the other curves associated to the city of Madrid, as it was expected. The only important variation in this case is the magnitude of the cumulative incidence value. The correlations are now shown in Figure 4.9.

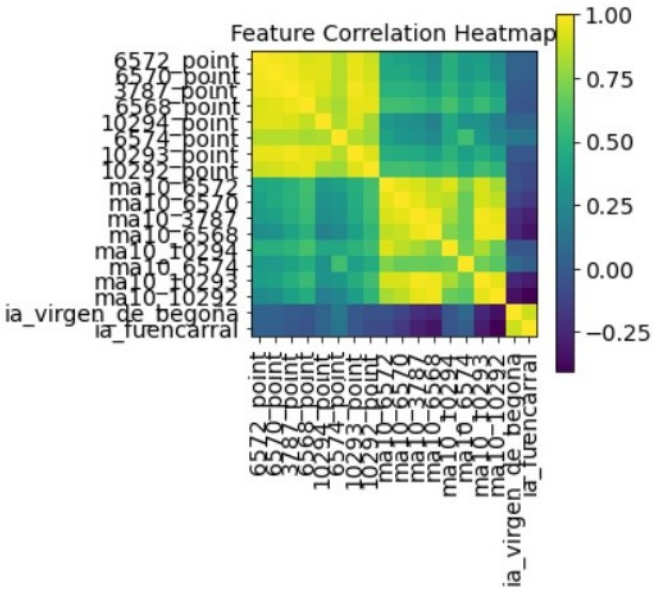


Fig. 4.9. Correlation for Virgen de Begoña variables

The heat map has a similar structure to the previously studied ones. The correlations in this case between the raw traffic points and the cumulative incidence are relatively high for some cases, as for example the point with ID 6574 eventually reaches a value close to 0.5. The moving averages, in exchange, are very weakly correlated with the cumulative incidence. As in the previous scenario, the number of traffic points is very high, so maybe they can be reduced after the first study is made to obtain a simpler representation.

4.3.2. Machine Learning Analysis

The results for the benchmark dataset are shown in Table 4.7:

	Without refit				With refit			
Period	1 day	5 days	7 days	14 days	1 day	5 days	7 days	14 days
XGBoost	0.1484	0.8132	0.9102	0.5967	0.1094	0.7790	0.8451	0.3547
LightGBM	0.1088	0.2747	0.4802	0.7123	0.0830	0.2428	0.3986	0.5638
SVR	0.0362	0.1963	0.2021	0.3706	0.0403	0.1924	0.1905	0.3364

Table 4.7. BENCHMARK RESULTS FOR VIRGEN DE BEGOÑA

The results look similar to those from the Andrés Mellado zone, with SVM being the best-performing algorithm overall, XGBoost suffering from a huge error increase and LightGBM behaving as in the rest of the scenarios. The number of lags chosen by each of the algorithm is more variate, as it changes depending on the prediction horizon and and regardless of the algorithm being used, unlike in the previous scenario. Hence, the results are open to potential improvements with the following datasets. Table 4.8 shows the results for the traffic data approach.

	Without refit				With refit			
Period	1 day	5 days	7 days	14 days	1 day	5 days	7 days	14 days
XGBoost	0.1428	0.8392	0.9721	0.4513	0.0987	0.7382	0.7848	0.4967
LightGBM	0.1363	0.3497	0.4961	0.7318	0.0821	0.2240	0.3728	0.7648
SVR	0.0498	0.1432	0.1660	0.3472	0.0407	0.1712	0.3690	0.3078

Table 4.8. TRAFFIC DATA RESULTS FOR VIRGEN DE BEGOÑA

The errors are similar to the benchmark, even worsening for the gradient boosting algorithms. Nevertheless, SVR is able to improve the results from the previous approach and without needing refit, as the obtained results for the refit approach are worse in comparison to the approach without refit. Like in the Andrés Mellado scenario, it is appreciated that generally the error grows as the time horizon increases, obtaining hence the worst result for a 14-day ahead prediction. The worst prediction on previous scenarios is generally obtained for the 7-day scenario. In terms of lags, gradient boosting algorithms choose different combinations depending on the scenario, and SVR is more consistent and focuses on lags up to 15 days before in most of the cases, obtaining in this way the best predictions. Hence, it may be thought that for this scenario it is better to choose a greater number of lags. These insights can be contrasted comparing the results with Table 4.9, which includes the cumulative incidence of the contiguous health zone.

	Without refit				With refit			
Period	1 day	5 days	7 days	14 days	1 day	5 days	7 days	14 days
XGBoost	0.2527	0.4480	0.4829	0.5317	0.1216	0.2128	0.3379	0.4009
LightGBM	0.2781	0.3935	0.3935	0.3946	0.1452	0.3020	0.3402	0.3469
SVR	0.1440	0.2718	0.3324	0.3471	0.0876	0.1931	0.3202	0.2771

Table 4.9. TRAFFIC DATA AND CUMULATIVE INCIDENCE RESULTS FOR VIRGEN DE BEGOÑA

The results in this case tend to stabilize in comparison with the previous table. The improvements are not remarkable, but the errors are generally lower, specially in the case of XGBoost. The long-term predictions are clearly improved, specially for the 14-day-ahead horizon. Following with the lag analysis, again SVR focuses on long-term lags while gradient boosting algorithms generally use short-term lags. Considering the global results, it can be stated that for this particular case, the models considering longer periods of time are likely to obtain lower errors.

The average MAE for this scenario is greater than in the previous cases, specially for the long-term predictions. The addition of new predictors to the models seems not to add relevant information, or seems not to be properly captured by the algorithms. The benchmark generally gives the best results, unless for particular cases. The hypothesis that are raised in the graphical analysis can be now supported, as the traffic changes were

so minor that the algorithms seem not to have captured them properly. Nevertheless, the SVR results are acceptable considering the obtained results in previous sections.

Finally, the confidence intervals are checked for the best-performing model for a long-term prediction, which in this case is the SVR with refit for the 14-day ahead prediction and the dataset with traffic data and cumulative incidence. Figure 4.10 shows the predictions and its confidence intervals.

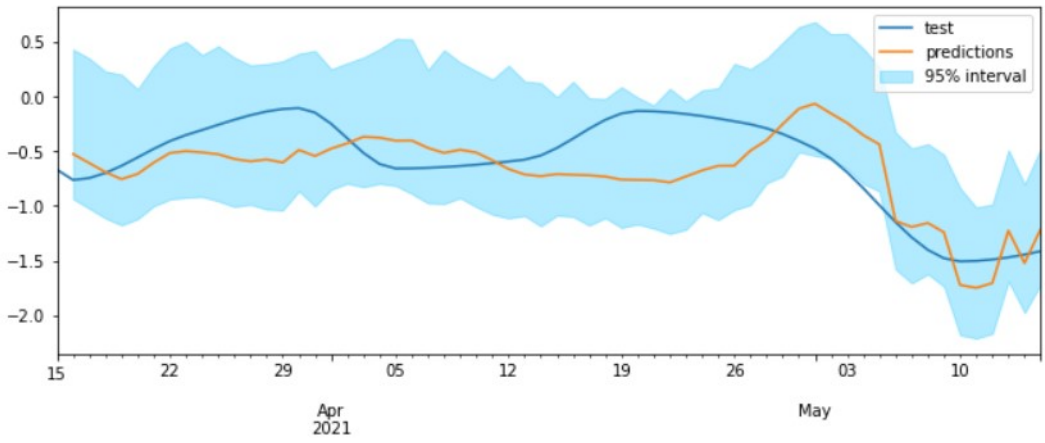


Fig. 4.10. 14-day prediction for Virgen de Begoña with cumulative incidence and traffic dataset

The error is the lowest one for all models, but it is not able to properly follow the distinct trends that are present on the cumulative incidence curve. However, the range of movement of the curve is properly delimited by the confidence intervals, so the results can be considered as positive. The only drawback about the confidence intervals is that they may be excessively wide, but considering the uncertainty of the predictions for this zone, the result is acceptable.

5. CONCLUSIONS

5.1. Final Results

The project has been successfully completed, as the datasets have been correctly generated, and then the analyses have revealed several insights about the studied situation. The employed methods are not the most common ones in literature, and dealing with results that can not be compared against external sources has been a challenge. The problem is complex, as it is influenced by many different external factors apart from mobility changes, so the results need to be properly contextualized.

The dataset preparation and processing has been a challenging task. Multiple transformations were applied to data previous to the final approach. One-day samples were finally selected, as it was an intermediate point between the frequencies of the traffic data and the cumulative incidence. The closing data acquisition was also challenging, as no datasets were found and it needed to be mined from the official bulletins.

The most accurate models for each of the zones are able to properly follow the movement of the cumulative incidence curve without having a great deviation. The confidence intervals for the shown cases are also satisfactory, as the range of movement is correctly captured, giving estimations that would have been useful at the time where the restrictions were active. The time series structure for the evaluation methods have made this possible, as no information from the months of April and May 2021, which integrate the test set, has been leaked in the training stages.

Several prediction horizons have been tested. The fourteen-day-ahead scenario has obtained better performances in some cases in comparison with the five-day and seven-day scenarios. The models change significantly depending on the zone under study, so the option of finding a global model to predict the cumulative incidence independently of the zone seems to be an even more challenging task. The addition of traffic data as exogenous predictors has helped refining the predictions in some of the cases, but not clearly improving the benchmark results in any specific case. The same happens for the contiguous cumulative incidence data, which has only contributed to improve the results in particular

cases. These results can mean different things depending on the interpretation. One option would be concluding that the traffic data is not really useful to refine significantly the cumulative incidence predictions. Other possibility would be concluding that the restrictions were not totally effective, because either the traffic was not really affected by them or they were not applied at the right time.

Personally talking, this project has been very useful to apply the knowledge that I have acquired during the master. Facing a real scenario with open source data is always a demanding task, as decisions have to be made on how to handle the data and how to use it. Initially the results were not as expected, and multiple approaches were tested. Although not all of them have been included in the report, they have given me more experience in dealing with problems of this kind.

5.2. Future Work

The objectives of the project have been successfully implemented, but some improvements can be considered, both in order to enhance the precision of the results and to extract further insights. Several state-of-the art papers have been reviewed for the development of this project, employing different tools and different data sources to solve similar problems. These options are also contemplated for the future work. Among the possibilities that can be tested, the following can be the most relevant for this project:

- Consideration of meteorological factors. The approach in [19] adds meteorological information to the traffic data, that results in useful predictions for small cities. This data could be tested for the city of Madrid, as it has been seen that the models do not hold even for areas from a common city, so they could be useful. Some situations as the described snowfall that happened in Madrid in January 2021 are not included in this analysis, as they are not usual and unpredictable from a long-term estimation.

- Statistical analysis to validate the hypothesis. The analyses can give many different conclusions, and these conclusions can be tested through statistical tests. An example for these procedures can be found in [22], where various mathematical models are employed to estimate future cumulative incidence and to estimate the correlations between the different zones.

There is still much research related to COVID-19, to understand past behaviors and evaluate the efficacy of the taken measurements in some of the cases. Artificial Intelligence is constantly growing, so surely new methods to tackle this problem will be developed. Although Deep Learning has been the most recurrent way to tackle these problems, other algorithms can also be useful to give relevant results about this topic.

BIBLIOGRAPHY

[1] C. of Europe, *Artificial intelligence and the control of covid-19*, 2022. [Online]. Available: https://www.coe.int/en/web/artificial-intelligence/ai-covid19.

[2] N. Arora, A. K. Banerjee, and M. L. Narasu, "The role of artificial intelligence in tackling covid-19," *Future Virology*, vol. 15, no. 11, pp. 717–724, 2020. DOI: 10.2217/fvl-2020-0130.

[3] Q.-V. Pham, D. C. Nguyen, T. Huynh-The, W.-J. Hwang, and P. N. Pathirana, "Artificial intelligence (ai) and big data for coronavirus (covid-19) pandemic: A survey on the state-of-the-arts," 2020. DOI: 10.20944/preprints202004.0383.v1.

[4] F. Kamalov, A. K. Cherukuri, and F. Thabtah, "Machine learning applications to covid-19: A state-of-the-art survey," *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, 2022. DOI: 10.1109/aset53988.2022.9734959.

[5] M. M. Islam *et al.*, "Application of artificial intelligence in covid-19 pandemic: Bibliometric analysis," *Healthcare*, vol. 9, no. 4, p. 441, 2021. DOI: 10.3390/healthcare9040441.

[6] Y. Meraihi, A. B. Gabis, S. Mirjalili, A. Ramdane-Cherif, and F. E. Alsaadi, "Machine learning-based research for covid-19 detection, diagnosis, and prediction: A survey," *SN Computer Science*, vol. 3, no. 4, 2022. DOI: 10.1007/s42979-022-01184-z.

[7] M. Suchmacher and M. Geller, *Practical biostatistics: A step-by-step approach for evidence-based medicine*. Academic Press, 2021.

[8] Y. Wang *et al.*, "Prediction and analysis of covid-19 daily new cases and cumulative cases: Times series forecasting and machine learning models," *BMC Infectious Diseases*, vol. 22, no. 1, 2022. DOI: 10.1186/s12879-022-07472-6.

[9] H. Alabdulrazzaq *et al.*, "On the accuracy of arima based prediction of covid-19 spread," *Results in Physics*, vol. 27, p. 104 509, 2021. DOI: 10.1016/j.rinp.2021.104509.

[10] M. A. Rguibi, N. Moussa, A. Madani, A. Aaroud, and K. Zine-dine, "Forecasting covid-19 transmission with arima and lstm techniques in morocco," *SN Computer Science*, vol. 3, no. 2, 2022. DOI: 10.1007/s42979-022-01019-x.

[11] Y. Alali, F. Harrou, and Y. Sun, "A proficient approach to forecast covid-19 spread via optimized dynamic machine learning models," *Scientific Reports*, vol. 12, no. 1, 2022. DOI: 10.1038/s41598-022-06218-3.

[12] J. Devaraj *et al.*, "Forecasting of covid-19 cases using deep learning models: Is it reliable and practically significant?" *Results in Physics*, vol. 21, p. 103 817, 2021. DOI: 10.1016/j.rinp.2021.103817.

[13] L. Xu, R. Magar, and A. Barati Farimani, "Forecasting covid-19 new cases using deep learning methods," *Computers in Biology and Medicine*, vol. 144, p. 105 342, 2022. DOI: 10.1016/j.compbiomed.2022.105342.

[14] K. ArunKumar, D. V. Kalaga, C. M. Kumar, M. Kawaji, and T. M. Brenza, "Forecasting of covid-19 using deep layer recurrent neural networks (rnns) with gated recurrent units (grus) and long short-term memory (lstm) cells," *Chaos, Solitons amp; Fractals*, vol. 146, p. 110 861, 2021. DOI: 10.1016/j.chaos.2021.110861.

[15] T. Hu *et al.*, "Building an open resources repository for covid-19 research," *SSRN Electronic Journal*, Jun. 2020. DOI: 10.2139/ssrn.3587704.

[16] J. H. University, *Coronavirus resource center*, 2022. [Online]. Available: https://coronavirus.jhu.edu/.

[17] D. Chumachenko, I. Meniailov, K. Bazilevych, T. Chumachenko, and S. Yakovlev, "Investigation of statistical machine learning models for covid-19 epidemic process simulation: Random forest, k-nearest neighbors, gradient boosting," *Computation*, vol. 10, no. 6, p. 86, 2022. DOI: 10.3390/computation10060086.

[18] P. Christidis and M. Radics, *Impact of the covid-19 pandemic on mobility in spain.* Apr. 2022. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/82ae69be-c1e6-11ec-b6f4-01aa75ed71a1/language-en.

[19] E. A. Rashed and A. Hirata, "One-year lesson: Machine learning prediction of covid-19 positive cases with meteorological data and mobility estimate in japan," *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, p. 5736, 2021. DOI: 10.3390/ijerph18115736.

[20] D. Haputhanthri and A. Wijayasiri, "Short-term traffic forecasting using lstm-based deep learning models," *2021 Moratuwa Engineering Research Conference (MERCon)*, 2021. DOI: 10.1109/mercon52712.2021.9525670.

[21] M. S. Ghanim, D. Muley, and M. Kharbeche, "Ann-based traffic volume prediction models in response to covid-19 imposed measures," *Sustainable Cities and Society*, vol. 81, p. 103 830, 2022. DOI: 10.1016/j.scs.2022.103830.

[22] G.-G. David *et al.*, "Perimeter confinements of basic health zones and covid-19 incidence in madrid, spain," 2021. DOI: 10.21203/rs.3.rs-646779/v1.

APPENDIX I: HEALTH ZONES MAPS AND TRAFFIC POINTS IDENTIFICATION

Andrés Mellado and Guzmán el Bueno

Chopera and Legazpi

Virgen de Begoña and Fuencarral

Click here to go back to Section 3.1.4