UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques
i Informàtica

GRAU DE MATEMÀTIQUES I ENGINYERIA INFORMÀTICA

Treball final de grau

Comparison of predictive methods of the COVID-19 pandemic evolution

Author: Albert Solà Roca

Supervisor:	Jordi Vitrià Marca
Conducted in:	Departament de Matemàtiques i d'Informàtica
Barcelona,	June 20, 2021

Contents

Abstract

The appearance of the COVID-19 pandemic has urged governments from around the world to resort to the scientific community in order to predict how the spread of the virus would evolve. In Catalonia, the need to compare the two most used models arose to decide on whether restrictions should be applied or not. The Catalan government has mainly used the Gompertz model to predict the evolution of the infected population but some scientists have shown concern for this model as compared to the more traditional deterministic compartmental models. This created the need to compare the effectiveness of these two models and determine which of them is more useful for predicting future pandemics.

In this project, we will present and analyse both types of models. In the case of compartmental models we will summarise only the three most basic types, SIR,SEIR and SEIRS. We will determine which of these models is most effective in predicting the spread of the virus by retrospectively comparing the predicted values and the real ones obtained. We will use data from the database provided by the Catalan government, Dades Obertes, and train our models with data from 30 days in order to predict two weeks in advance.

Our results have shown that SEIR is the best of the initially proposed methods but is clearly affected by noise in our data, especially after restrictions were applied. A clear improvement was shown by averaging the values from the Gompertz model together with the values of the SEIR model. We clearly obtained the best possible model from combining both of them as it is smoother than predicting with SEIR and yields better results than the Gompertz model.

In future work, we plan to include more complex compartmental models, such as the ones including geographical and transportation factors, given the necessary data. In addition, a larger average of models can be used, in which more than 2 models is applied, for an even more precise prediction of the evolution of pandemics.

Resum

L'aparició de la pandèmia de la COVID-19 ha instat a governs de tot el món a recórrer a la comunitat científica per predir com evolucionaria la propagació del virus. A Catalunya ha sorgit la necessitat de comparar els dos models més utilitzats per tal de decidir amb facilitat quan aplicar restriccions. El govern català ha utilitzat principalment el model de Gompertz per a predir l'evolució de la pobla-

2020 Mathematics Subject Classification. 03C65,34C60,62J05

ció infectada, però alguns científics han mostrat preocupació per aquest model en comparació amb els models compartimentals deterministes més tradicionals. Això ha creat la necessitat de comparar l'eficàcia d'aquests dos models i determinar quin d'ells és més útil per a predir futures pandèmies.

En aquest projecte, presentarem i analitzarem els dos tipus de models. En el cas dels models compartimentals, resumirem els tres tipus més bàsics: SIR, SEIR i SEIRS. Determinarem quin d'aquests models és més eficaç per predir la propagació del virus comparant retrospectivament els valors predits i els reals obtinguts. Utilitzarem dades de la base de dades facilitada pel govern català, Dades Obertes, i entrenarem els nostres models amb dades de 30 dies per tal de predir l'evolució de la pandèmia en les dues setmanes següents.

Els nostres resultats han demostrat que el SEIR és el millor dels mètodes proposats inicialment, però que està clarament afectat pel soroll de les nostres dades, especialment dels dies posteriors en que s'apliquen restriccions. Es va demostrar una clara millora mitjançant la mitjana dels valors del model de Gompertz juntament amb els valors del model SEIR. Hem vist com clarament hem obtingut el millor model possible combinant-los tots dos ja que obtenim un model més suau que l'obtingut amb el SEIR i obtenim valors més precisos que amb el model de Gompertz.

En futurs treballs, tenim previst incloure models compartimentals més complexos, que tinguin en compte factors geogràfics i de transport, donat que tinguem accés a les dades necessàries. A més, es pot utilitzar una mitjana amb més models, en comptes d'utilitzar-ne dos, per a una predicció encara més precisa de l'evolució de les pandèmies.

Resumen

La aparición de la pandemia de la COVID-19 instó a gobiernos de todo el mundo a recurrir a la comunidad científica para predecir cómo evolucionaría la propagación del virus. En Cataluña surgió la necesidad de comparar los dos modelos más utilizados para decidir con facilidad cuando aplicar restricciones. El gobierno catalán ha utilizado principalmente el modelo de Gompertz para predecir la evolución de la población infectada, pero algunos científicos han mostrado preocupación por este modelo en comparación con los modelos compartimentales deterministas más tradicionales. Esto ha creado la necesidad de comparar la eficacia de estos dos modelos y determinar cuál de ellos es más útil para predecir futuras pandemias y en qué situaciones.

En este proyecto, presentaremos y analizaremos los dos tipos de modelos. En el caso de los modelos compartimentales, resumiremos los tres tipos más básicos: SIR, SEIR y SEIRS. Determinaremos cuál de estos modelos es más eficaz para predecir la propagación del virus, comparando retrospectivamente los valores predichos y los reales obtenidos. Utilizaremos los datos de la base de datos facilitada por el gobierno catalán Dades Obertes y entrenaremos nuestros modelos con datos de 30 días a fin de predecir la evolución de la pandemia en las dos semanas siguientes.

Nuestros resultados han demostrado que el SEIR es el mejor de los métodos propuestos inicialmente, pero que está claramente afectado por el ruido de nuestros datos, especialmente los días después de que se apliquen restricciones. Se demostró una clara mejora tomando la media de los valores del modelo Gompertz junto con los valores del modelo SEIR. Hemos obtenido el mejor modelo posible combinando ambos dado que obtenemos un modelo más suave que el de SEIR y más preciso que el de Gompertz.

En futuros trabajos, tenemos previsto incluir modelos compartimentales más complejos, que tengan en cuenta factores geográficos i de transporte, si tenemos acceso a los datos necesarios. Además, se puede utilizar una media con más modelos, en vez de utilizar dos, para una predicción aún más precisa de la evolución de las pandemias.

Chapter 1

Introduction

As the new SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) pandemic started, governments from around the world resorted to the scientific community in order to predict how the spread of the virus would evolve. This led the Catalan government to compare various methods from multiple research groups but, in the end, they mostly took advice from the BIOCOMSC group at UPC. This research group uses a mathematical model named Gompertz model which predicts the evolution of the pandemic over the next few days and even weeks. This curve was first proposed by Benjamin Gompertz [1] and was based on the idea that the mortality rate increases exponentially as a person ages. This model is one of the most frequently used sigmoid functions to fit cumulative growth due to its asymmetric nature and has been used to model many kinds of growths such as tumour growth, plant and animal growth, and bacterial growth.

Despite its widespread use, other researchers, such as URV's Alex Arenas [2], have doubted the effectiveness of this model as compared to the one proposed by them, the more traditional compartmental models, due to the simplicity of the Gompertz curve and the non-determinism involved in it. These compartmental models are based on the idea of dividing the population into compartments or categories such as Susceptible, Infected and Recovered (SIR) and had already been used in previous pandemics such as the spread of measles in the UK in 1986, or the description of the Ebola epidemic in 1995 [3] and 2014 [4] or even the previous SARS epidemic of 2003 [5]. These use differential equations to model the evolution of the pandemic and can be expanded to take into account the incubation period of the disease by simply adding a new compartment named Exposed (SEIR). Moreover, you can take into account temporal immunity by feeding back the model to the Susceptible compartment again (SEIRS). We can even expand it further as more information becomes available. For example, Alex Arena's group uses SEAIHR, which takes into account Asymptomatic cases and Hospitalised patients.

Nonetheless, this pandemic has been marked by the continuous use of government measures and involvement, such as the application of lockdowns and curfews. These measures affect the evolution of the disease differently based on the strength and type of restrictions applied on mobility and social life. Of course, the evolution of the disease will be very different in countries such as New Zealand, where a long total lockdown was applied at the start as compared to other countries with less restrictive measures such as the United States of America. Spain and Catalonia fall somewhere in between, as multiple lockdowns have been applied, which varied in both strength and restrictions. This creates situations in which the choice of model becomes more complex. Simple compartmental models are not designed to deal with this type of situation where the network of contacts and its changes due to government policies are key [6].

In the mathematical modeling of a disease, as in most others areas of mathematics there is always a trade-off between simple models, which omit most details and are designed only to highlight general qualitative behaviour, and detailed models, usually designed for specific situations including short-term quantitative predictions [7] and this case is no exception. On the one hand, the Gompertz model simplifies the evolution of the pandemic to a single equation and is useful to determine statistically what the evolution of the pandemic will be in the future. On the other hand, compartmental models can be as detailed as possible and try to simulate reality in a more accurate way.

The main difference between these two models is that the Gompertz curve is a stochastic model and compartmental models are deterministic models. Given enough information compartmental models can model the evolution of a pandemic almost perfectly whereas the Gompertz curve can give us a rough idea of how the disease will evolve. That being said, the main problem compartmental models present is the effect noise has on the prediction of the data. Compartmental models are based on differential equations which can lead to chaotic systems. In addition, there is a lot of uncertainty in predicting the evolution of the pandemic. Cases such as the New Year illegal party in Llinars del Vallès with over 2000 participants can create a lot of variance which compartmental models cannot adapt to properly. In addition, data available contains a lot of this noise because of the weekend effect, which lowers infected cases during weekends and accumulates the expected difference into weekdays, especially Mondays and Tuesdays [9].

With regards to tackling future pandemics more efficiently, the need to properly evaluate the two most important methods used in the prediction of the coronavirus pandemic becomes clear. In this project, we will compare the simpler

Gompertz model to the more complex compartmental models in order to determine which of these would have been more valuable to predict the evolution of the pandemic in Catalonia.

First of all, we will present and define both the Gompertz Curve and compartmental models, and explore some of the simpler subtypes of compartmental models such as SIR, SEIR and SEIRS. We will also explore the advantages and disadvantages each of these models present. In addition, in Chapter 3 we will explain the implementation of all of the models, focusing especially on the optimisation of the initial parameters for the Gompertz model. This initial parameters are a fundamental part of this project as it is important not to fit the data with local minimums for the parameter values but instead fit with global minimums. This is hard to accomplish as a single initial parameter has been used for all of the days of the year. In this Chapter we have also explained the reasoning behind the choice of initial parameters.

We will evaluate the effectiveness of all of these models by training them with 30 days of data and predicting the cumulative number of cases in two weeks time for every day of this pandemic. In order to evaluate which of these methods is best we will use a number of metrics, we will observe how well they predict the curve using the coefficient of determination R^2, we will present and study qualitatively the variance of the predictions and, finally, we will evaluate the cumulative relative error rate. This way, we will have a better understanding of how the error varies, the effect noise has on our predictions and the effect restrictions have on the evolution of our models. Finally, we will also take the average of the best compartmental model and the Gompertz model in order to provide with a new possible model and we will analyse the improvement this model provides. The results obtained will be of great use in order to manage future possible pandemics as a better understanding of all of these models will lead to a better response from the governments.

Chapter 2

State of the art

In this chapter, we discuss the two most used models to predict the evolution of pandemics. Section 2.1 discusses the Gompertz model, whereas section 2.2 summarises the simplest compartmental models SIR, SEIR and SEIRS.

2.1 The Gompertz model

We employ the Gompertz model for growing processes to model the cumulative cases of COVID-19. The Gompertz model is one of the most frequently used sigmoid models and is fitted to growth data and other data, perhaps only second to the logistic model [8]. The main difference between the logistic model and the Gompertz function is the replacement of the saturation of the growing factor from linear in the logistic model to an exponential decrease in the Gompertz model. This modifies the symmetric function the logistic equation produces to an asymmetric function with fast growth of new cases combined with a slow decrease, which is closer to the distribution of new cases observed in different countries. It has been shown [6] that this asymmetric nature of the Gompertz model is the proper framework to study epidemics in which control measures are part of the evolution of the disease. The simplicity of this function is what allows us to create a simple model which easily approximates the cumulative cases for the COVID-19 epidemic. The Gompertz equation reads as follows:

$$N(t) = Ke^{-\ln(\frac{K}{N_0})e^{-at}} \tag{2.1}$$

where the parameter K represents the final number of cases, N_0 is the initial number of cases at time t=0, a is the rate of decrease in the initially exponential growth and $N(t)$ is the cumulative cases of the disease at time t. This equation was originally proposed as a means to explain human mortality curves [10] but it has also been used in many other cases, including the evolution of bacterial colonies [11] and tumours [12]. This function can also be interpreted as the solution to the following pair of differential equations:

$$\begin{aligned} \frac{dN}{dt} &= \mu N \\ \frac{d\mu}{dt} &= -a\mu \end{aligned} \tag{2.2}$$

Which corresponds to an exponential growth with a growing rate μ which exponentially decreases with rate a [6]. This exponential rate μ provides us with the relationship between K and a, as we have that:

$$\mu = a\ln(K/N_0) \tag{2.3}$$

In Figure 2.1 we can observe the behaviour of this function for different values of a. As a note, we remind the reader that the Gompertz curve represents the evolution in the cumulative cases. As observed in Figure 2.1 an increase in the value of a relates to a slower decrease of the initial exponential growth. In Figure 2.2 we can also observe the behaviour of this function for different values of K. As said before, K represents the final number of cases, which results in a horizontal asymptote at the value of K. This Figure clearly illustrates the horizontal asymptote this function has at K.

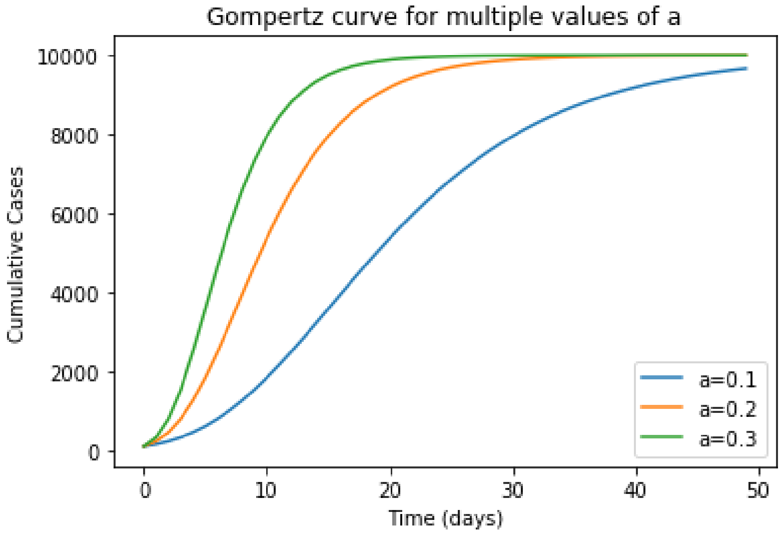


Figure 2.1: Parameter a varies the rate of decrease in the initially exponential growth

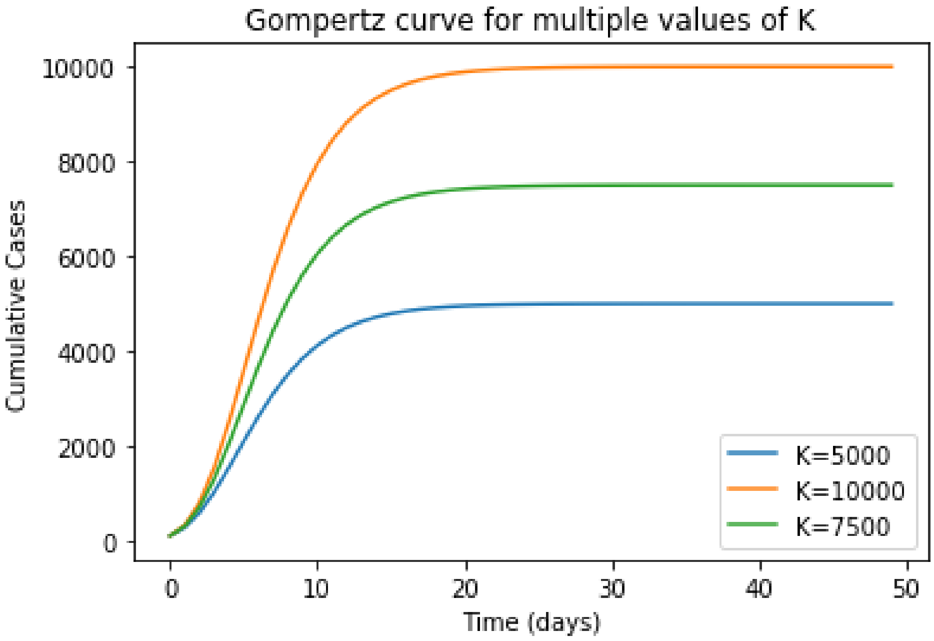


Figure 2.2: Parameter K corresponds to the final number of cases

In order to more clearly illustrate the effect varying the parameter a has, we need to observe the derivative of the function. We know that the Gompertz curve models cumulative cases, therefore its derivative will model the daily cases. By calculating the derivative with respect to time we obtain that the evolution of the daily cases is modelled by the following equation:

$$N_n = \frac{dN}{dt} = aKe^{-\ln(\frac{K}{N_0})e^{-at}}(\ln(\frac{K}{N_0})e^{-at}) \tag{2.4}$$

And by plotting this we obtain Figure 2.3. In this case we can clearly observe how larger values of a lead to a larger, sooner, and higher peak in daily infections. We can also observe in Figure 2.3 and in Figure 2.1 what politicians refer to as smoothing the curve. Despite the total number of cases remaining the same a more disperse peak is formed for lower values of a. This can be the difference between saturation of ICUs and a more relaxed environment for doctors.

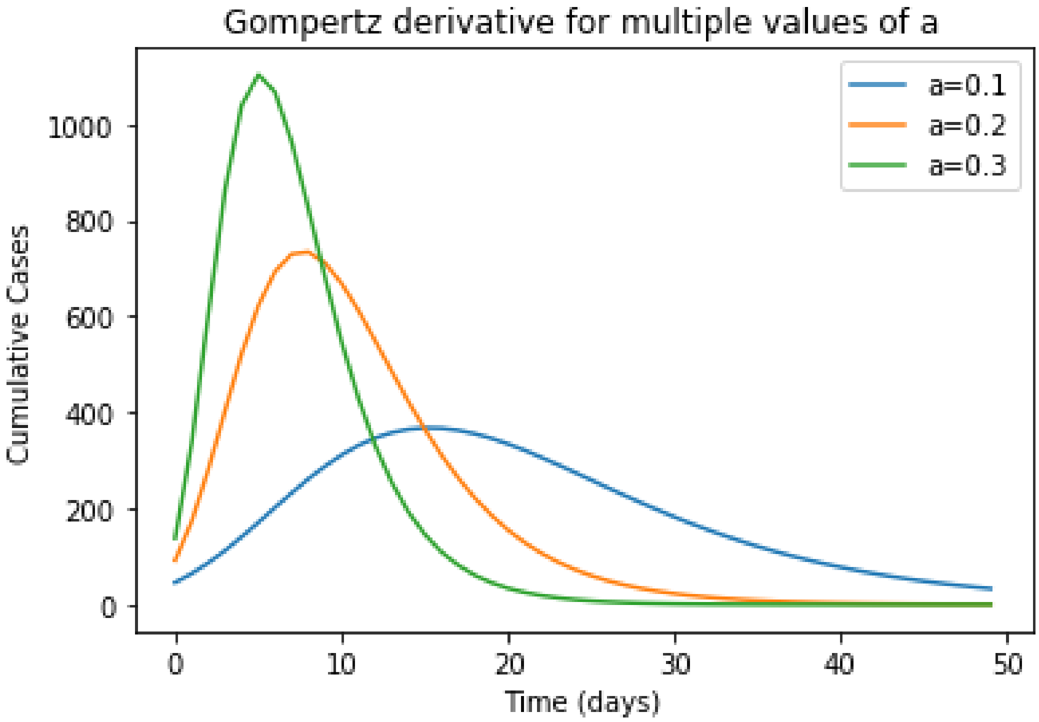


Figure 2.3: Variation of the parameter a on the derivative of the Gompertz curve

The peak observed in Figure 2.3 corresponds to the inflection point in the cumulative cases curve in Figure 2.1 and can be calculated from equation 2.4 as follows:

$$\begin{aligned}
\frac{d^2N}{dt^2} = \frac{dN_n}{dt} &= aK(ae^{-at}(\ln(\frac{K}{N_0})e^{-at})^2 + e^{-\ln(\frac{K}{N_0})e^{-at}}(-a\ln(\frac{K}{N_0})e^{-at})) \\
0 &= a^2Ke^{-\ln(\frac{K}{N_0})e^{-at}}(\ln(\frac{K}{N_0})e^{-at})^2 - a^2Ke^{-\ln(\frac{K}{N_0})e^{-at}}(\ln(\frac{K}{N_0})e^{-at}) \\
e^{at} &= \ln(K/N_0) \\
t_p &= \frac{1}{a}\ln(\ln(\frac{K}{N_0}))
\end{aligned} \tag{2.5}$$

In this equation, we can more clearly see the effect a has. The larger the value of a, the faster the appearance of the peak, as we have seen in Figure 2.3.

2.2 Compartmental models

Compartmental models are based on the idea that we can divide the population into different compartments representing the different stages of the disease and use the relative size of each compartment to model how the numbers evolve in time. The population is assigned labels for a better understanding, such as **S**usceptible, **E**xposed, **I**nfected and **R**ecovered, in our case. This population can transition from one compartment to another as defined by differential equations. Different sets of differential equations result in the different types of models. In this project we will only explore the simpler models as these compartmental models can become as large as necessary, given enough information is known. Compartmental models are especially useful when taking into account a very large population, as deterministic models become more useful than stochastic models at this point.

2.2.1 SIR

The simplest compartmental model consists of three compartments with two transitions, **S**usceptible, **I**nfected and **R**ecovered. Population can go from being susceptible to catching the disease and being infected and then to being recovered, but it doesn't take into account incubation periods of the disease or the possibility of reinfection. In terms of transitions, this model can be visualised as:

$$
\begin{gathered}
S + I \xrightarrow{\beta} I + I \\
I \xrightarrow{\mu} R
\end{gathered}
$$

This model can be interpreted as follows. The first transition represents that a susceptible person will become infected with a probability of β when interacting with someone infected and the second one represents that an infected person will recover spontaneously with a probability of μ at any moment. That being said, these probabilities become rates when taking into account the entire population. Therefore, we could also describe the previous system of transitions as follows. The susceptible population is becoming infected at a rate of β and the infected population is recovering at a rate of μ. This model can be better represented mathematically as the following set of differential equations:

$$\begin{aligned}
\frac{\delta}{\delta t} S_t &= -\beta S_t \frac{I_t}{N} \\
\frac{\delta}{\delta t} I_t &= +\beta S_t \frac{I_t}{N} - \mu I_t \\
\frac{\delta}{\delta t} R_t &= +\mu I_t
\end{aligned}$$

Where N is the total population size and I_t/N represents the fraction of infected population. S_t, I_t and R_t are the populations that are susceptible, infected and recovered at a time t after the beginning of the epidemic. First of all, we can observe that:

$$\frac{\delta}{\delta t} S_t + \frac{\delta}{\delta t} I_t + \frac{\delta}{\delta t} R_t = 0 \tag{2.6}$$

from which we can obtain that $S_t + I_t + R_t = constant$, and, because we initially had the entire population N divided into the three compartments, we have that $S_t + I_t + R_t = N$. In addition, compartmental models allow us to calculate easily the value of R_0, which is a very important metric in order to understand the evolution of a pandemic. Given $R_0 < 1$ then the pandemic will slowly extinguish, but an $R_0 > 1$ results in an exponential increase in number of cases. From this specific model we can calculate R_0 as:

$$R_0 = \frac{\beta}{\mu} \tag{2.7}$$

In Figure 2.4 we can observe an example of the evolution of a disease with 100 initial infected, a population N of 10000, β is 0.2 and μ is 0.1. This, as calculated with Equation 2.6, results in an R_0 of 2. As we can see, there is a single large peak of infected population, but it quickly decreases and the epidemic ends. This peak is in daily cases, and results in around 20000 cases at the peak of the pandemic, which represents about 2% of the population being infected in a single day. Another important detail is the fact that not all the population becomes infected. This is because after herd immunity is reached, the pandemic slowly dies down and the rest of the population is not infected. The value this takes depends on β and is βN. Therefore, in this case, with a population of 100000 and a β of 0.2, we have that 20000 people won't be infected thanks to heard immunity.

Despite being a very simple model, the behaviour displayed in Figure 2.4 represents what the evolution of the pandemic would have looked like if the disease was not lethal, had no incubation period and without any type of interventions from the government.

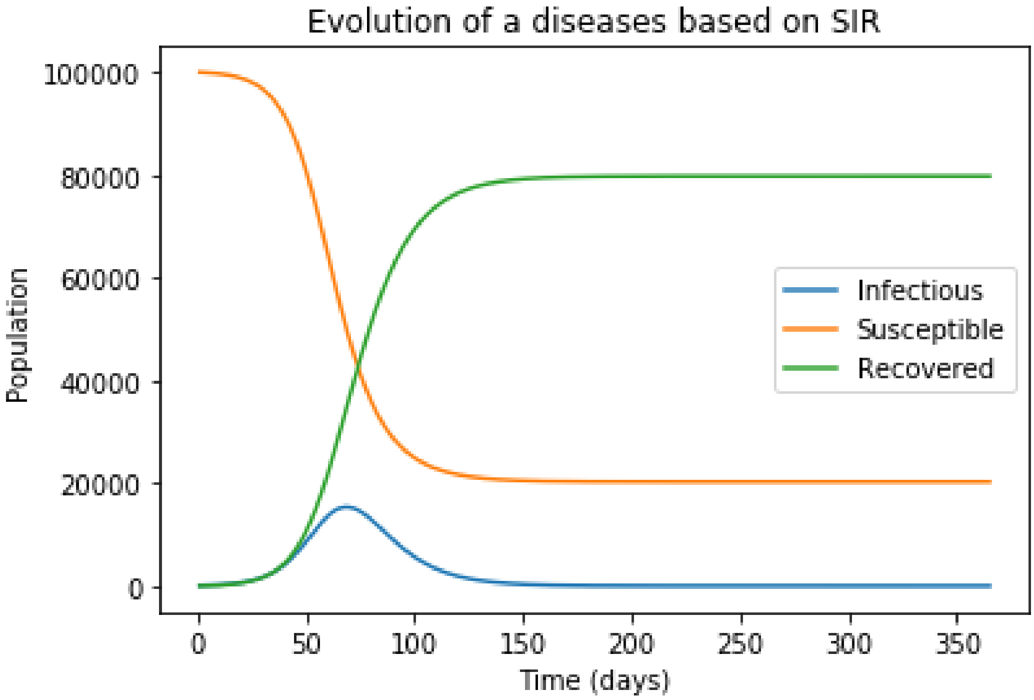


Figure 2.4: Evolution of the different compartments of SIR

2.2.2 SEIR

Compartmental models can be easily expanded to cover more cases. SEIR is one example of this behaviour in which, by simply adding the **E**xposed compartment to SIR, the model now includes an incubation period in which an individual is infected but cannot transmit the disease. In terms of transitions, this model can be visualised as:

$$S + I \xrightarrow{\beta} E + I$$
$$E \xrightarrow{\epsilon} I$$
$$I \xrightarrow{\mu} R$$

This model changes the interaction of Susceptible with Infected to an Exposed instead of an Infected status, maintaining the probability of β, and adds a spontaneous interaction from Exposed to Infected at a rate ϵ which is proportional to the duration of the incubation period. As said before, this intermediate step allows for a representation of the population that has been infected but cannot yet infect others. We can also represent these transitions as the following set of differential equations:

$$\begin{aligned}
\frac{\delta}{\delta t} S_t &= -\beta S_t \frac{I_t}{N} \\
\frac{\delta}{\delta t} E_t &= +\beta S_t \frac{I_t}{N} - \epsilon E_t \\
\frac{\delta}{\delta t} I_t &= +\epsilon E_t - \mu I_t \\
\frac{\delta}{\delta t} R_t &= +\mu I_t
\end{aligned}$$

Once again, we have that

$$\frac{\delta}{\delta t} S_t + \frac{\delta}{\delta t} E_t + \frac{\delta}{\delta t} I_t + \frac{\delta}{\delta t} R_t = 0 \tag{2.8}$$

Which means the total population stays constant. For this model, the basic reproduction number remains $R_0 = \frac{\beta}{\mu}$. In Figure 2.5 we can see an example of the behaviour of the pandemic with the same base values as with SIR but with the addition of the **E**xposed compartment. We initially start with 0 individuals in the Exposed compartment and we assign an ϵ of 0.4. This models a pandemic with a short incubation period, similar to what we see with COVID-19. That being said, we can observe that introducing ϵ does not have an effect on R_0, as seen before, it remains 2. We can observe how the total number of cases remains the same but the infectious curve becomes slightly smoother. In this case, the peak of the pandemic is reached at around day 100 instead of day 60 as with SIR. This is a clear consequence of the incubation period. Of course, if we had an ϵ of 0 we would obtain the same values as with SIR, as it would mean that the incubation period is none.

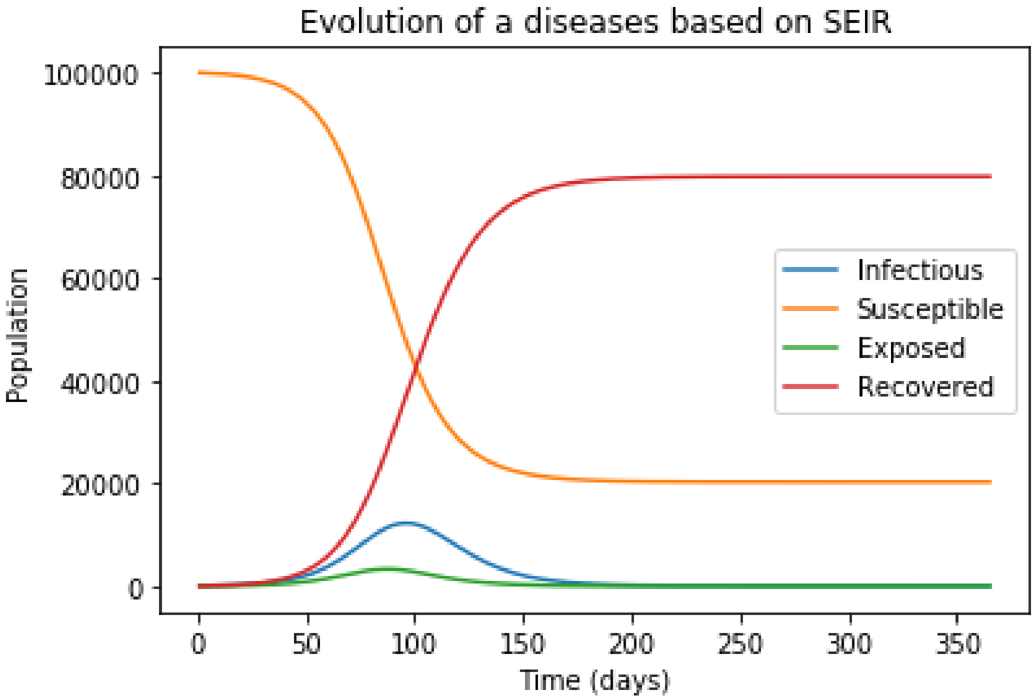


Figure 2.5: Evolution of the different compartments of SEIR

2.2.3 SEIRS

The final model we will be using is named SEIRS and allows us to model a period of immunity. In terms of transitions, this model can be visualised as:

$$
\begin{aligned}
S + I &\xrightarrow{\beta} E + I \\
E &\xrightarrow{\epsilon} I \\
I &\xrightarrow{\mu} R \\
R &\xrightarrow{\rho} S
\end{aligned}
$$

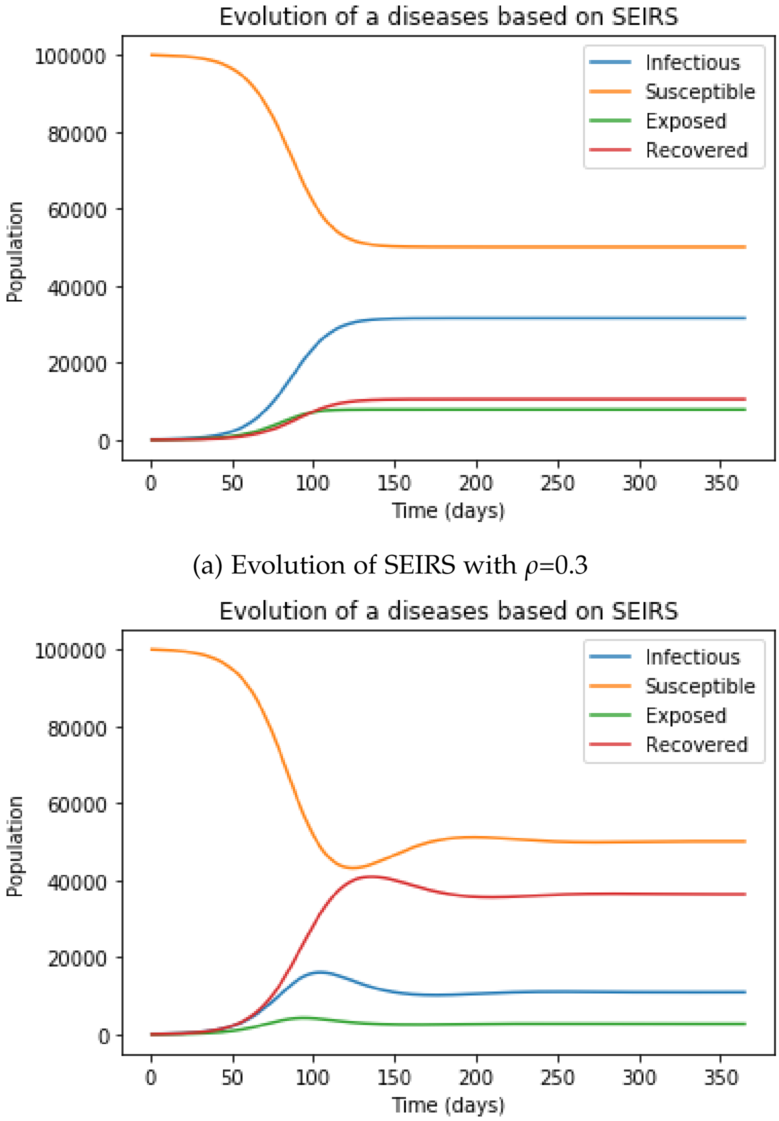
Or, as the following set of differential equations:

$$
\begin{aligned}
\frac{\delta}{\delta t} S_t &= +\rho R_t - \beta S_t \frac{I_t}{N} \\
\frac{\delta}{\delta t} E_t &= +\beta S_t \frac{I_t}{N} - \epsilon E_t \\
\frac{\delta}{\delta t} I_t &= +\epsilon E_t - \mu I_t \\
\frac{\delta}{\delta t} R_t &= +\mu I_t - \rho R_t
\end{aligned}
$$

Once again, we have that

$$\frac{\delta}{\delta t} S_t + \frac{\delta}{\delta t} E_t + \frac{\delta}{\delta t} I_t + \frac{\delta}{\delta t} R_t = 0 \tag{2.9}$$

These more complex compartmental models have the slight drawback that R_0 becomes harder to calculate and would need to be approximated. Therefore, we don't have a simple formula to calculate this value. That being said, we present the evolution of a pandemic given the same values as with SEIR but adding a ρ of 0.3 in Figure 2.6a and a ρ of 0.03 in Figure 2.6b. We can observe how the behaviour becomes very different to the ones proposed before as there is an equilibrium after about 120 days, therefore, the epidemic reaches an endemic state. This is because the total number of population infected to be smaller because some population that has already been infected stops being immune and catches the disease again. The rate ρ at which immunity is lost has a determinant effect in the progress of the epidemic and the rise of endemicity. If ρ is sufficiently small (immunity is longer lasting) we can even have several epidemic peaks before the steady state of a fixed fraction of the population is reached [13]. In our case, a very high ρ in Figure 2.6a is shown so we don't see a clear peak in infections but we can already observe how a ρ of 0.03 produces two peaks before stabilising. In Figure 2.6a we can also observe how the Recovered compartment is way smaller as compared to 2.6b, which is because a larger ρ results in a lower immunity and individuals can become susceptible again a lot faster.



(a) Evolution of SEIRS with ρ=0.3

(b) Evolution of SEIRS with ρ=0.03

Figure 2.6: Evolution of the different compartments of SEIRS for different values of ρ

Chapter 3

Implementation

In this chapter we present the implementation and optimisation of our models, together with the process of fitting.

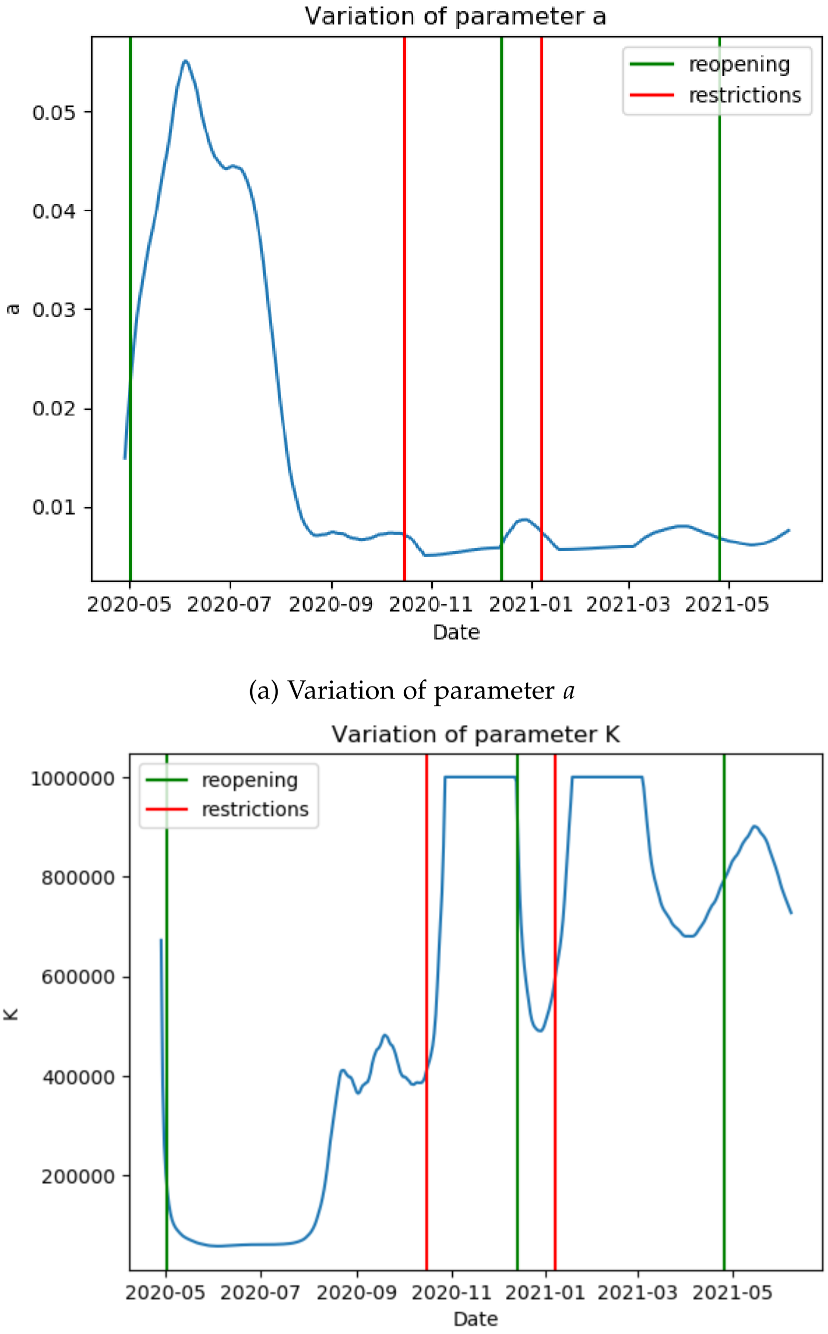
3.1 Models

For this project we wanted the models to predict 14 days in the future because we believe that it would be an acceptable amount of time required by hospitals to prepare for future peaks in infections. In order to predict this accurately we decided to train our models with data from the past 30 days, and repeat that for every day of the pandemic. This would then give us the most optimal parameters[1] together with some data that we cannot obtain such as the Exposed population. This would simulate how models have been used during the pandemic and allow for a better understanding of which of these models is better.

Given the complexity of the models, non-linear least squares regression needs to be applied. For this reason, we will be using **Scipy**'s *optimize.curve_fit*, which allows us to fit a function f to a set of data d. This method also allows us to apply bounds to both of our functions. In Gompertz's case, we have $K \in [0, 1000000]$, i.e. the maximum number of total infected population is the entire population of Catalonia and $a \in [0, 1]$. On the other hand, SIR, SEIR and SEIRS models have each compartment bounded by [0,7556000] and each parameter is bounded by [0,1]. That being said, we made sure that the population stayed constant at 7556000, as we need that S+E+I+R=7556000 as seen before.

It is important to mention that finding the optimal initial parameters is a very important part of this project as the fitting of the functions can easily tend to local minimums instead of global minimums. For this reason, our initial values in the

[1] $\beta, \epsilon, \mu, \rho$ in case of compartmental models and a,K in the case of the Gompertz model



(a) Variation of parameter *a*

(b) Variation of parameter *K*

Figure 3.1: Optimisation of the parameters in the Gompertz model

base model are the following. For Gompertz, we chose a=0.01 as we found that our model converged to this value after enough time has passed, N_0 is the number of cumulative cases 30 days before the first prediction of our model and we take K to be the number of cumulative cases for the day before we are currently doing the prediction from. We chose these initial values because, as seen in Figure 3.1a a tends to 0.01 after the pandemic settles but we don't have a clear indicator of which value K tends to, as it more easily fluctuates depending on the current conditions, as seen in Figure 3.1b. Nonetheless, we can clearly see an increasing tendency for the value of K, which we can model by using the number of cumulative cases for the day before we are currently doing the prediction from. Finally, we capped the value of K to 1000000 because we retrospectively know that in Catalonia we have not yet reached this value of cumulative cases and this should allow us to obtain a better prediction. For this reason, Figure 3.1b presents maximums at 1000000. Uncapping this value could potentially lead to really high predictions in certain days which don't correlate to the other predictions.

We also plot in Figure 3.2 the evolution of μ which, as said in Chapter 2 represents the relationship between this two values. As we can see, it follows a pattern similar to the one in Figure 3.1a and, despite the variation in the parameter K, it still converges to around 0.04.

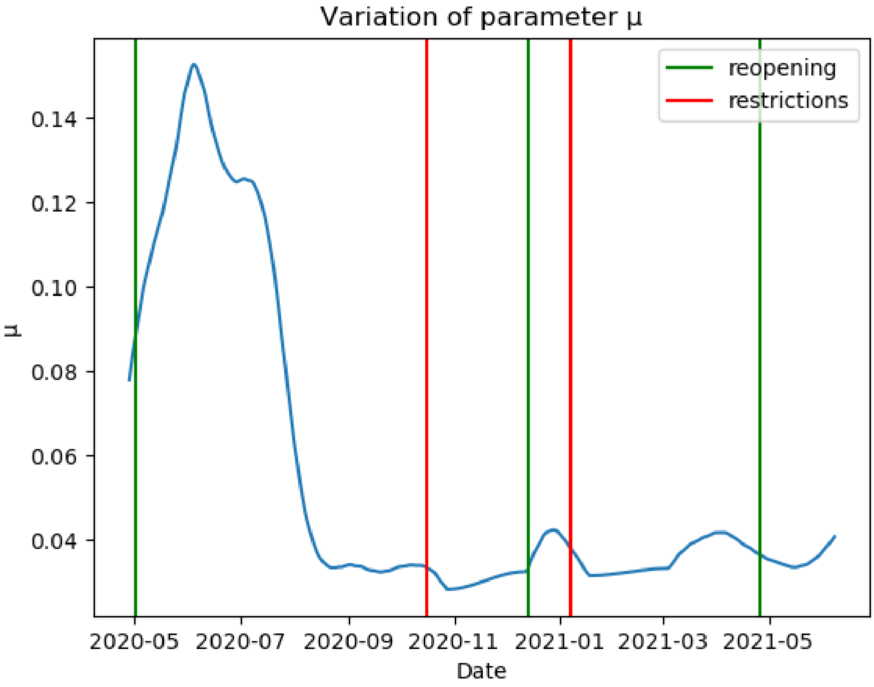


Figure 3.2: Variation of parameter μ

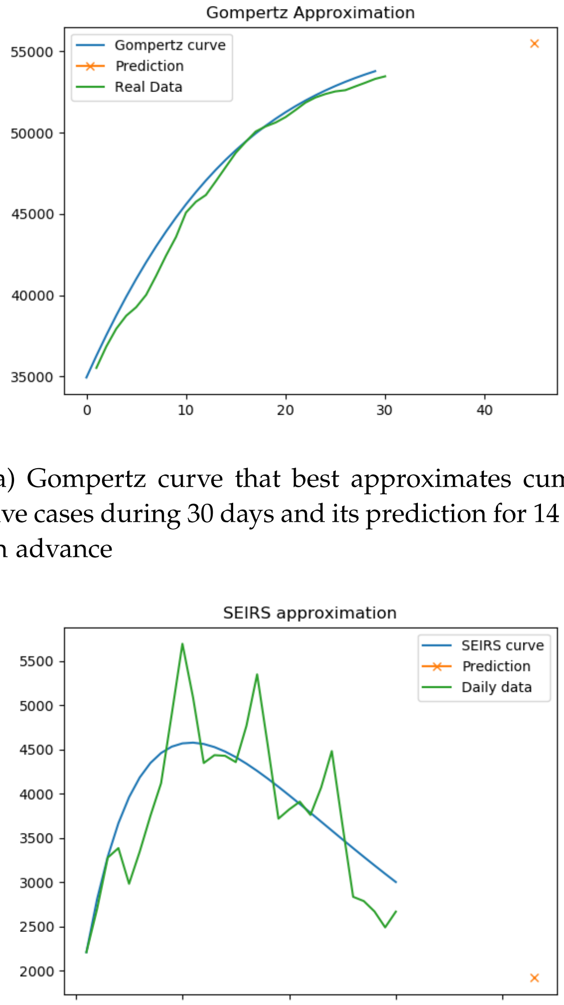
In the case of SIR we use $\beta = 0.2$, $\mu = 0.1$ and for SEIR we also have $\epsilon = 0.3$ and we suppose that the Exposed compartment initially has a population of 25000. Finally, for SEIRS we also suppose an initial ρ value of 0.01. This election of the parameters was reached by trying multiple values of all of the parameters for every day and choosing which one yielded the lowest error in the fitting of the data. Given the complexity of these model, finding a single optimal parameter that works for every day is complicated and in real life should be recalculated in a daily basis.

It is also important to note that we train the compartmental models with daily data and the Gompertz model with the cumulative data. This is because Gompertz model was made to approximate cumulative cases but compartmental models approximate daily cases in an easier way. This brought us a big challenge because we had to eliminate noise from the data, especially the daily data. The main problem our data had was the weekend effect. Infected cases were detected less during the weekends and, instead, would be accumulated on Mondays and Tuesdays. Cumulative data does not have this problem as the variation in daily cases is relatively small as compared to the total cumulative cases.

Another important note is that with Gompertz model a single prediction can be made, but, because we are fitting compartmental models to daily data, we have to add up all the values predicted for the 14 days by our model. Therefore, despite seeing in Figure 3.3b a single predicted value, all of the previous predicted values will also be added. Therefore, a bigger error is expected with compartmental models, given that error can accumulate more easily.

Finally, we had to find a balance between number of iterations and precision of the fitting. We found that a maximum of 100000 iterations with a tolerance for the termination by the change of the cost function of 10^{-7} allowed us to fit both models easily and precisely. Anything that could create a prediction with less time than a night would be acceptable for us, as we need models that can predict 14 days into the future for every single day. Figure 3.3a and 3.3b present the fitting of the data. In Figure 3.3a we have fitted the Gompertz curve to the cumulative data whereas in Figure 3.3b we have fitted the SEIRS model to the daily data. In this second Figure it is more clear how the weekend effect clearly affects compartmental models whereas the cumulative data is not affected by it.

The provided code in Annex 1 is our implementation of all that has been explained during our work. We chose Python as our Programming Language as it provides many libraries that are very useful for data analysis. Specifically, we use libraries that are useful for data management, such as **Numpy**, **Pandas** and **Scipy**. Another library we have used in this project is Bruno Gonçalves' **EpiModel.py**, which facilitates the modelling of the compartmental models. We make especial use of its integrate method instead of the simulate method because we are not interested in the statistical variation of these models but rather the deterministic part of them. This library uses **Scipy**'s *integrate.odeint()* to calculate the necessary differential equations. In the case of the Gompertz model, no additional libraries were used as we implemented Equation 2.1 in Python and the fitting can be done directly on the function.



(a) Gompertz curve that best approximates cumulative cases during 30 days and its prediction for 14 days in advance

(b) SEIRS approximation of daily cases during 30 days and its prediction for 14 days in advance

Figure 3.3: Examples of approximations of the models

Chapter 4

Experiments

After a short note about data acquisition, we describe the methodology followed to evaluate the experiments.

4.1 Data acquisition and management

All the data employed in this project has been accessed through the public database of the Catalan government[1]. The Catalan government uses a portal named Dades Obertes which contains all kinds of data that is public and free to use. The Catalan administration receives, generates and manages a large amount of data, which is stored in a wide variety of information systems. The main purpose of Dades Obertes is to make data managed by the Catalan administration available to society, so that any person or organisation can use them. With this service, the Catalan administration increases transparency. In addition, the reuse of open data by companies, entities, associations and the general public allows the development of new products and services that provide value, innovation, knowledge and business opportunities [14].

In our case, we are interested in the database that contains daily data on COVID-19 divided by counties, "**Dades diàries de COVID-19 per Comarca**". In order to access this database we need to use the **Socrata** API which we will obtain from the **sodapy** library. The main objective of this dataset is to facilitate in a public and transparent way the epidemiological monitoring of the COVID-19 pandemic by monitoring the key indicators defined in the control plan for the transmission of COVID-19 in Catalonia. It contains data of the confirmed cases of COVID-19, positive PCR tests, hospital admissions and deaths of the population, further divided by age group, sex and whether or not the cases occurred in

[1] All of the available databases related to COVID-19 can be found at https://analisi.transparenciacatalunya.cat/browse?q=covid&sortBy=relevance

a nursing home. Given that our main interest was on the cumulative cases, this was the best dataset to use. In addition, if further compartmental methods have to be applied, we can expand the ones we have implemented by adding hospitalised cases and deaths.

It is important to note that this dataset starts on the 1st of March and is updated two days after the current date. That being said, the first detection of COVID-19 in Catalonia was on the 25th of February 2020. In addition, data from the first few weeks of the pandemic was very inconsistent and unreliable, therefore, the first day we will start to predict is the 30th of April. This is because we need 30 days of reliable data in order to predict accurately and we believe that after the 15th of March 2020, once the pandemic was more set, data becomes more reliable. We end our predictions on the 8th of June 2021, 450 days after the start of the pandemic. This allows us to have a large sample of predictions in order to better understand where the models fail to act correctly and improve these predictions in the future.

In addition, the most important days regarding predictions have been marked in all of the figures. These correspond to the 2nd of May 2020, where the restrictions started to get lifted due to an improvement in the evolution of the pandemic, the 16th of October 2020, where restrictions where reapplied previous to the Christmas campaign, the 14th of December 2020, where this campaign started and restrictions where temporarily lifted, the 7th of January 2021, where restrictions where reapplied after a failed Christmas campaign with a high increase in cases and a lot of pressure from the medical standpoint and finally, the 26th of April of 2021, where restrictions started to get lifted again, in a de-escalating manner.

From this database we will be grouping by days to obtain the total number of infected individuals each day because for this project we are not interested in dividing the data by counties. We will be making use of **Pandas**' dataframes to facilitate the treatment of this data. It is important to note that we are still missing some information that we must use for this project. Some of it is easy to find, for example, the cumulative cases needed for the modeling of the Gompertz curve can be calculated by doing the cumulative sum of the daily cases we have, which we do when parsing the data.

Other data is harder to find but we were able to find approximations. This is the case of the recovered cases that we need for the compartmental models. No data on the number of recovered patients is included in this or other databases so we needed to approximate this value. We found that the average time it took for an individual to recover from COVID-19 was 21 days so, in order to find the initial approximation of this value, we took the cumulative cases registered 21 days ago.

Finally, it is necessary to mention that there is some data which cannot be found nor approximated, and we had to guess the initial value. This is the case for the Exposed compartment. It is virtually impossible to know how many people have been exposed to the virus at any point in time, as we can only guess this number by later knowing how many infected individuals there are two or three days after.

4.2 Methodology

We have implemented all of the models described in Chapter 2. We have implemented the Gompertz function directly in **Python**, whereas for compartmental methods we have implemented them using the **EpiModel.py** library. After this implementation, we train them with 30 days of data and predict for day 44, exactly two weeks later. We repeat this for every day of this pandemic and evaluate the resulting curve as compared to the real values in Catalonia. We evaluate all of the results based on two metrics:

- The coefficient of determination

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})}$$

- The mean difference of consecutive days

$$\text{Mean diff} = \frac{\sum_i |y_i - y_{i+1}|}{n}$$

These metrics will give us an idea of which of the models better approximates the real values as we will understand how well our functions predicts the real values, how much variance it has and how much peaks affect our predictions. In order to facilitate the analysis of the variance, we also provide qualitative graphs that we will study to determine whether or not the model groups errors in a single day or distributes them throughout the entire pandemic. In addition to implementing all of the models and evaluating them with these parameters, we will also be implementing an average of the Gompertz model together with the compartmental model that provides us with the best results. Finally, we will evaluate the cumulative relative error rate. This metric will allow us to better understand where the error is concentrated and reads as follows:

$$\text{Cumulative relative error} = \sum_i \frac{|y_i - f_i|}{f_i}$$

Chapter 5

Results and discussion

Gompertz Model

We now present and analyse the results obtained. We begin by showing the predictions made by our models, which use a normal 5 day average on the daily data to train the models and a static initial value. Every Figure shown shows our predictions made with each model in orange compared to the real values in blue. We also include the days restrictions were applied and lifted to show the effect this has on our predictions.

We start by presenting the predictions made by fitting the Gompertz curve. In Figure 5.1 we can observe how we obtain a very smooth curve to begin with. We can also observe how the model adjusts very well when no restrictions are applied but deviates slightly at both restrictions. This is because the model fails to realise fast enough that changes have been made and that the number of cases is no longer increasing rapidly as it used to. This creates a valley in the prediction which quickly increases to the more normal values. The stronger the restrictions are, the lower this valley is as compared to the real values. We find that these predictions by the model adjust to the curve with an R^2 of 0.99323[1]. This is already a very promising result as we can already see how the base model of the Gompertz curve adjusts very well to the real cumulative cases curve.

[1] All the values presented in this section have been summarised in table 5.1 for a better understanding

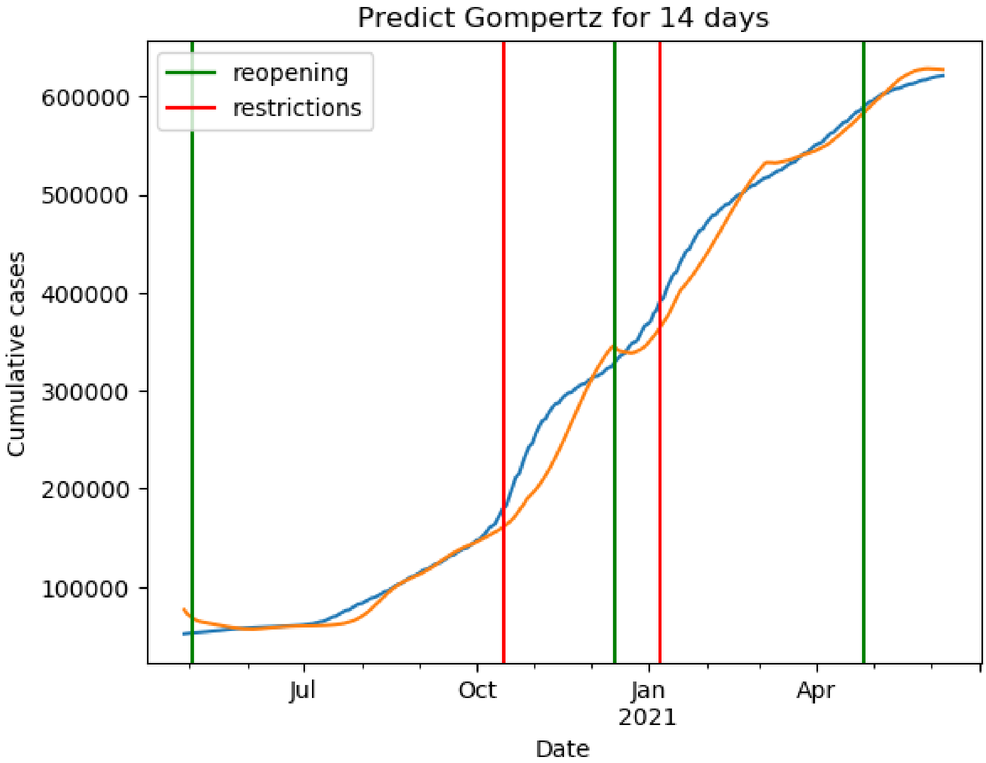


Figure 5.1: Base model Gompertz predictions

We can also observe how the Gompertz curve has a mean difference of consecutive days of 1499. It is important to note that to make sense of this metric we have to compare it to the value obtained with the real cases curve, which is of 1436. As we can see, the value obtained with the Gompertz model is very close to the real value and we will also see how it translates to it being the smoothest of the obtained curves. It is also worth noting that this model tends to predict lower values than the real obtained number of cases as compared to the future compartmental models.

We finally make a qualitative study of the evolution of the error of the Gompertz model in order to better understand if the error is accumulated in certain days or if it is distributed throughout the entire pandemic. First of all, we can observe in Figure 5.2 the evolution of the daily error created by our prediction. We can see how the highest difference in cases is of around 60000[2], which is attributed to the peak we see in December in Figure 5.1.

[2] It is important to mention that this is a difference of 60000 cumulative cases, distributed over 14 days, which means an average daily error of infections of around 4285. That being said, these functions are not linear, so a higher error is expected the further away we predict.

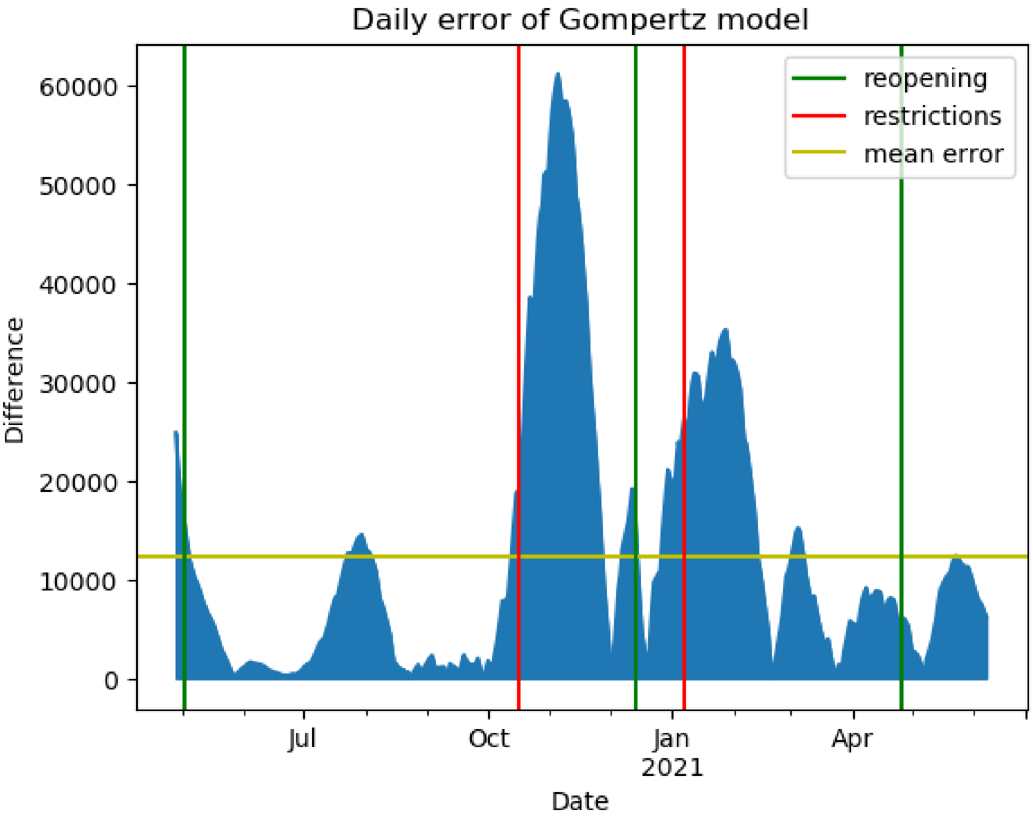


Figure 5.2: Absolute value of the daily error of the Gompertz model

That being said, the Gompertz model presents many smaller peaks in difference. We can see that until June 2021 there are at least 6 peaks of at least 10000 cases in absolute difference and 3 of these peak with more than 20000. We can also more clearly observe the effect restrictions have on our data. As we can better observe in Figure 5.2 peak errors come after these restrictions are applied, and are higher the more strict these restrictions are. Finally, we also plot the mean error from the Gompertz model, which is of about 12000 cumulative cases. This represents an average daily error of around 850 cases.

Compartmental Models

A similar behaviour can be seen in compartmental models, as we can observe in figures 5.3, 5.5, and 5.7. However, the more complex the model is, the quicker and easier it adjusts to changes in restrictions. We can already observe in Figure 5.3 a clear effect of noise in our model. That being said, we still obtain that our SIR model adjusts to the real values with an R^2 of 0.99319. We can see an improvement in the predicted values in figures 5.5 and 5.7. SEIR and SEIRS models seem to be less affected by changes in restrictions and react faster to these than SIR, but are still worse than the Gompertz model in the sense that they present larger peaks.

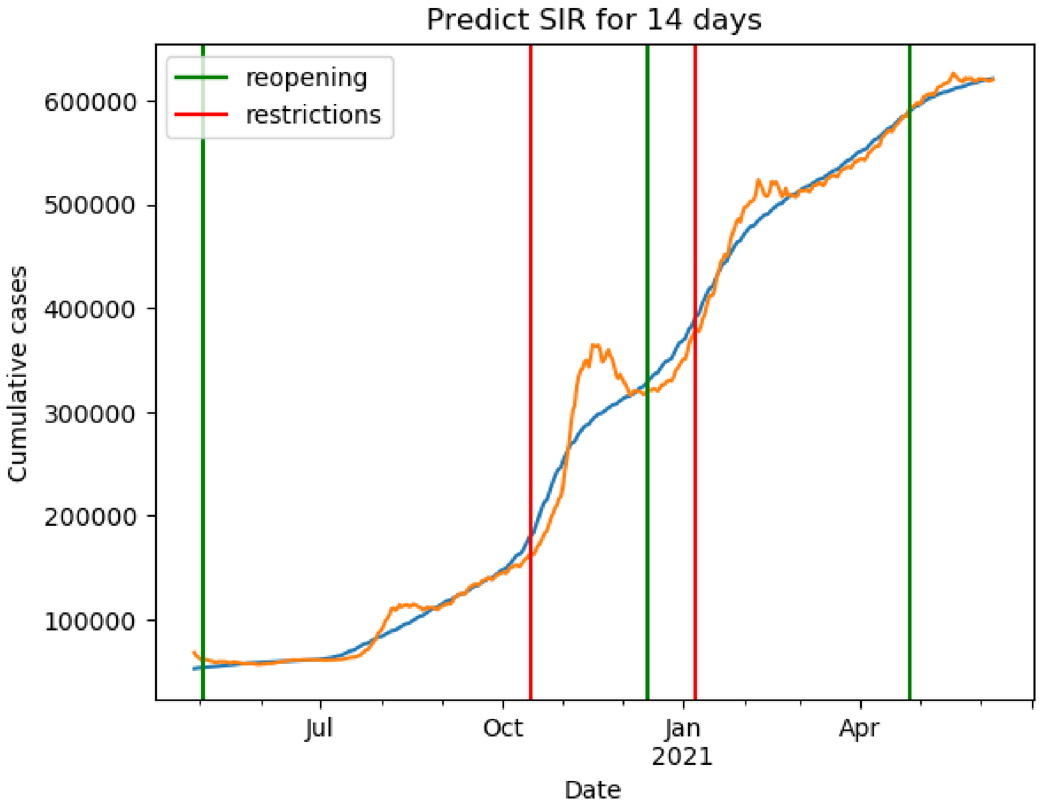


Figure 5.3: Base model SIR predictions

This can be easily observed when comparing error rates. In Figure 5.4 we can observe how the peak in error is higher than with Gompertz, at around 70000. Despite this, it seems like this model adjusts better in the other days as compared to Gompertz, as it presents less peaks of a difference of more than 10000. Similarly, we can observe how 4 peaks are formed with more than 10000 cases in absolute difference and 2 of more than 20000. In addition, we can observe a slight improvement from the Gompertz model in the mean error rate. In this case we have around 11000 cases as compared to Gompertz' 12000, which results in a daily average error of around 785.

It is interesting to observe how in both cases, peaks and valleys are formed. We believe that this is because the model takes time to correct itself and it overcorrects to better predict the results. This generates a pattern in which some days the model is extremely precise for the 14 day prediction but others where it is very far from the actual value. As we have seen before, this pattern also appears with the Gompertz model, and we will also see how it appears again with SEIR and SEIRS.

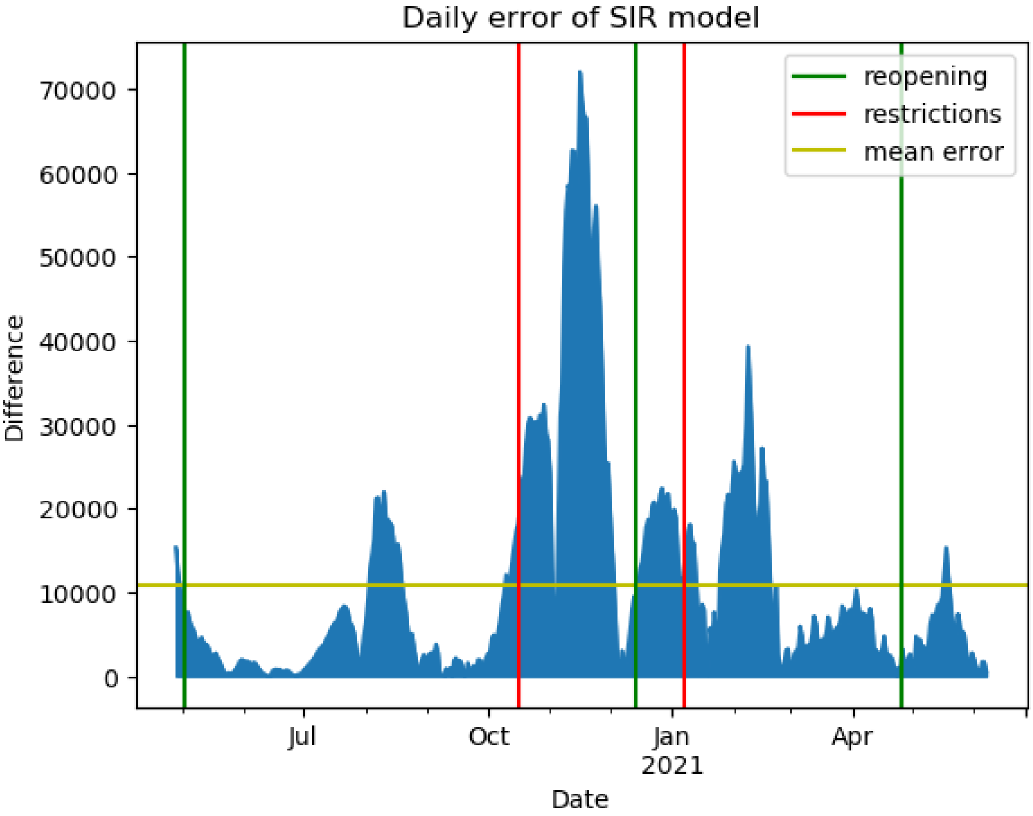


Figure 5.4: Absolute value of the daily error of the SIR model

In Figure 5.5 we can observe the evolution of the predictions for the SEIR model. As we said before, it improves the SIR model by reducing the peak after the restrictions and smoothing the curve overall. We can also see how it adjusts better to the real values, as we have an R^2 of 0.99598 as compared to the previous 0.99319 from SIR. We can clearly observe some improvement from SIR to SEIR. Finally, we can also see how this curve is smoother than the one obtained by SIR by comparing consecutive days. We obtain that the average difference between days reduces to 2121 as compared to SIR's 2331, which proves that this curve is slightly smoother on average. Nevertheless, we can still observe the effect restrictions have on this model. In this case, a more clear peak can be observed but the correction to adjust to the new values happens more quickly.

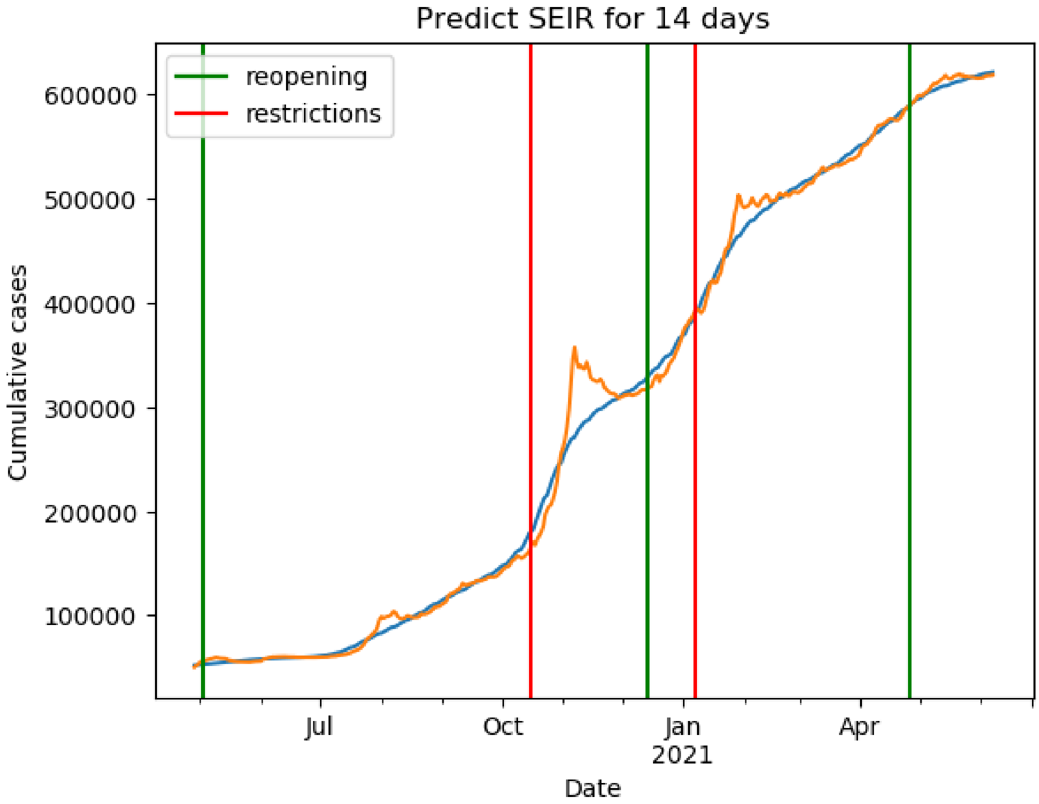


Figure 5.5: Base model SEIR predictions

By analysing the error rate in Figure 5.6, we can see some similarities and differences presented by this new model as compared to SIR. First of all, this compartmental model also presents a very large peak after the restrictions. This peak is larger than with SIR but also lasts less time. We can clearly observe how most of the error this function brings is concentrated in that period of time, whereas the other smaller peaks are less pronounced than with SIR. It is especially important to notice how this model corrects itself quicker than SIR when restrictions are applied, specially when these are less strict. We again have an improvement in the average error rate. This time, we have a mean error rate of around 9000, which represents an average daily error rate of around 650 cases.

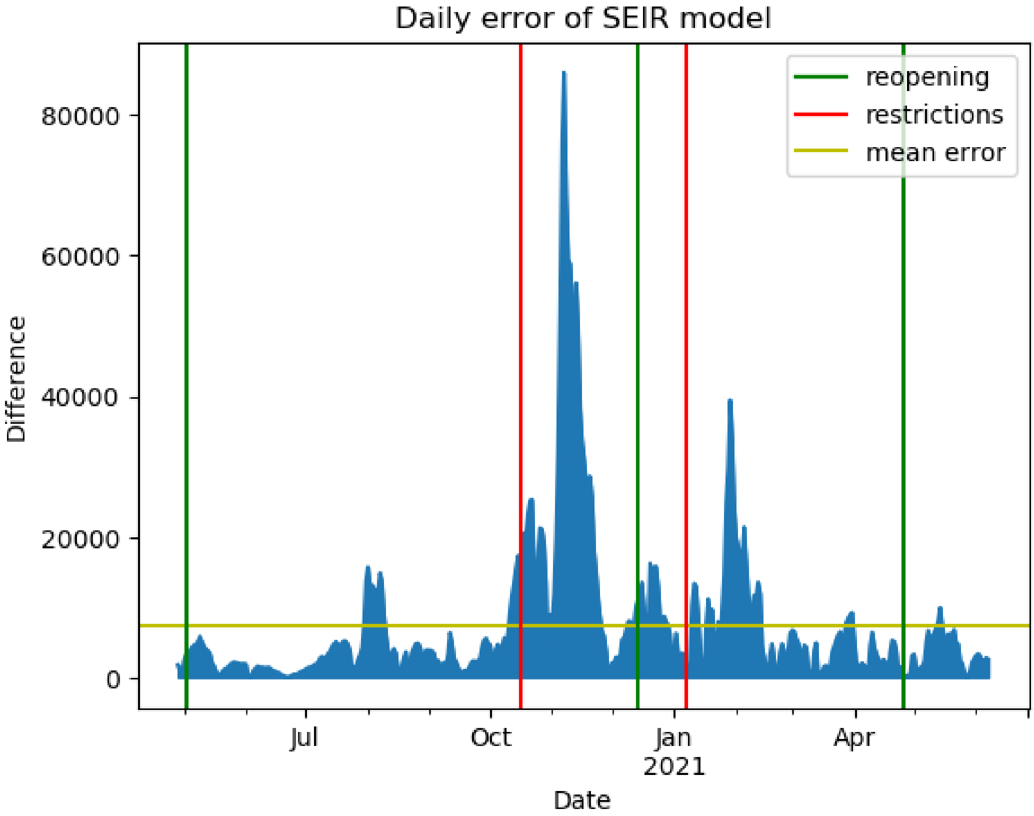


Figure 5.6: Absolute value of the daily error of the SEIR model

Finally, we present the results obtained by SEIRS. We can clearly see in Figure 5.7 a very similar result to the one obtained with SEIR. The main difference being an increase in smaller peaks, due to SEIRS being more affected by noise. We can also see this with our mean difference of days metric, which is of 2354 in SEIRS as compared to the 2121 obtained with SEIR.

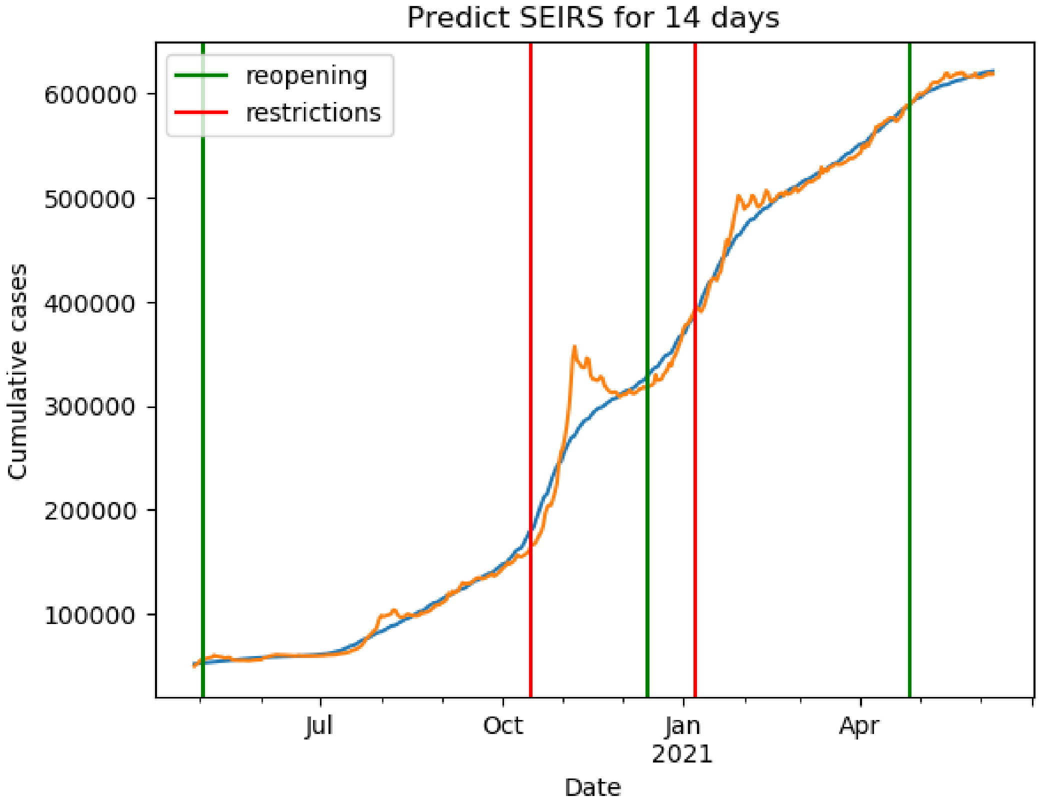


Figure 5.7: Base model SEIRS predictions

By analysing Figure 5.8 we can more clearly see how this error graph really resembles the one created when predicting with SEIR. We again see a clear peak in error after the restrictions but it realises faster than the Gompertz model. This creates the more clear peak of 80000 cases in difference. Finally, we do not see a clear improvement in average error rate in this model as compared to SEIR, which follows the pattern we have been seeing.

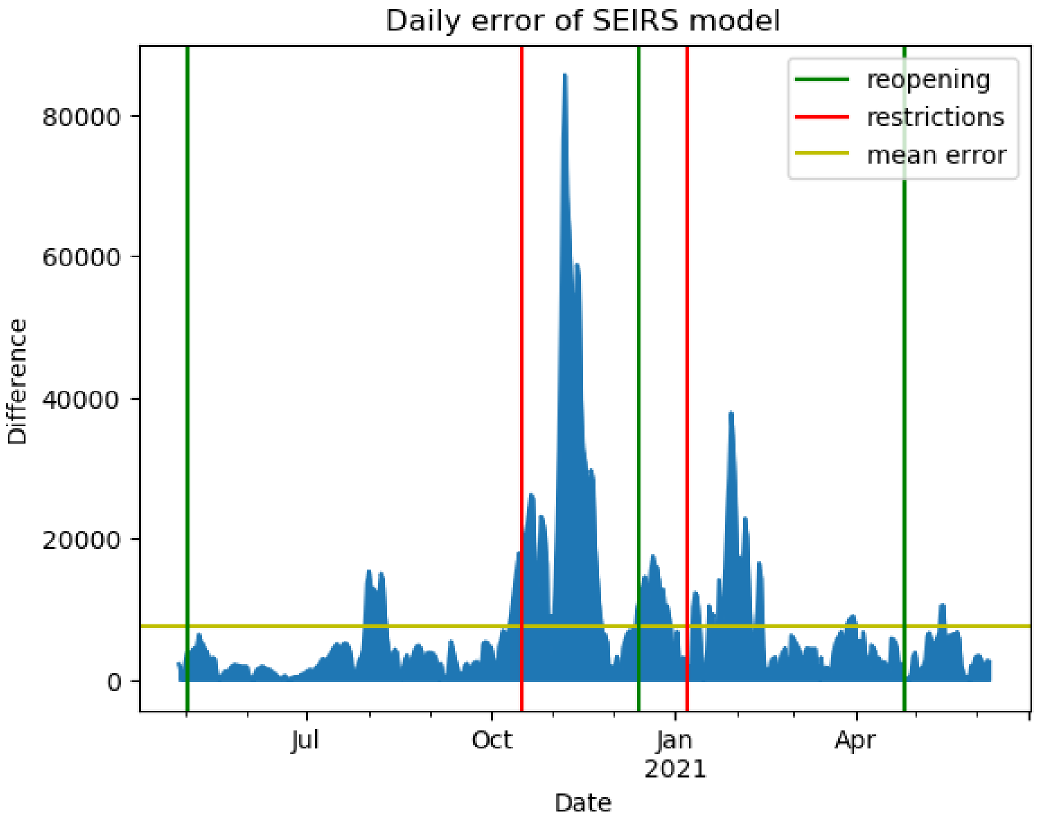


Figure 5.8: Absolute value of the daily error of the SEIRS model

For now, we have seen that SEIR produces the best model. It's the model with the highest R^2, together with the lowest variance in the daily change of the prediction. Having said this, Gompertz produces the smoothest of curves. Of course, we know that this is probably due to the fact that Gompertz fits cumulative cases and compartmental models fit daily cases. Therefore, the effect noise has is more clear in compartmental models as compared to the Gompertz model.

Average of Gompertz and SEIR

With all of this in mind, we finally propose a single new model. We have taken the averages of the values obtained with SEIR and the Gompertz model to create a new model. We have decided to try this model because the Gompertz model had a tendency to predict lower values than expected whereas compartmental models had the opposite tendency. We also decided to use the SEIR model instead of SEIRS because we believe it produced the best approximation of the real values, together with it being simpler than SEIRS. We only used two of the models because using more compartmental models would influence the new curve as they all have similar tendencies of focusing the error during the periods of restrictions. This would have led to similar results with a multiple average of the values as to the compartmental models. We now present the results in Figure 5.9.

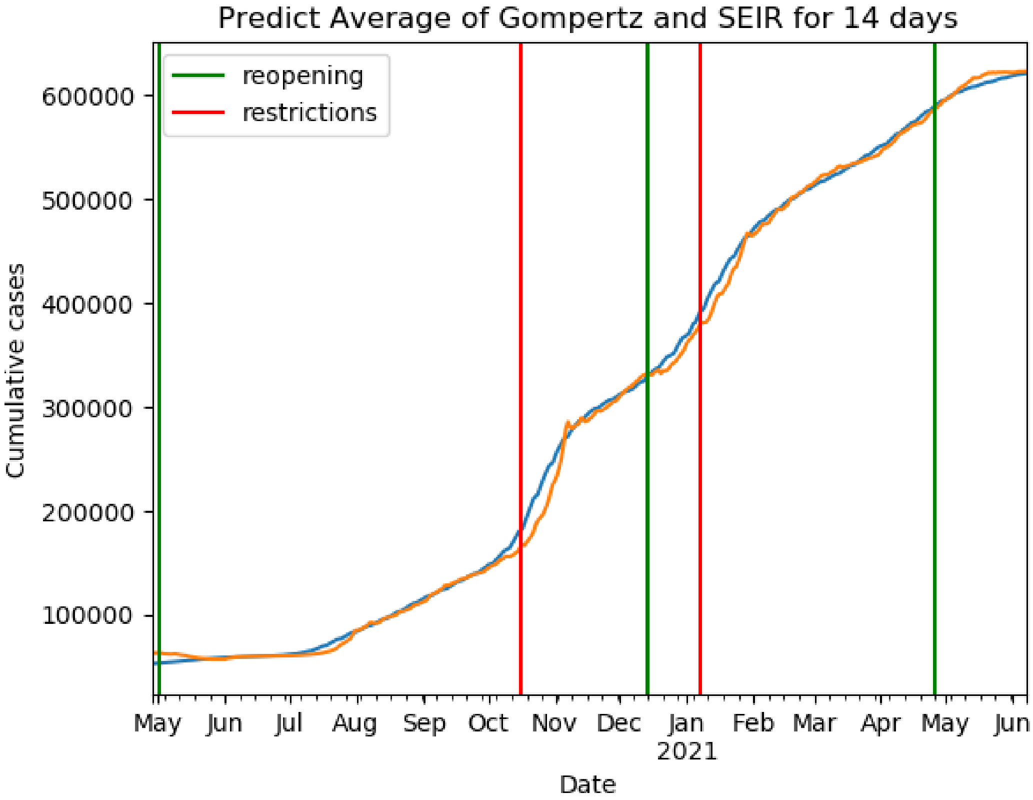


Figure 5.9: Predictions made with the average of SEIR and Gompertz models

We can already clearly observe how this average model is an improvement of all of the previous ones presented. This model deviates very little when restrictions are applied and presents the smoothest of curves. We can observe this trend more clearly by studying the metrics mentioned previously. This curve presents an R^2 of 0.9985 and the mean difference between consecutive days is of 1538, which is much closer to the real value of 1422 as compared to compartmental models. This curve clearly presents the advantages of both of the models.

We can more clearly observe the improvement this method provides in Figure 5.10. We can already observe how the peak value in error is of around 35000 cases which, compared to SEIR's 80000 or Gompertz's 60000 is a great improvement. In addition to this, we obtain only 2 peaks of more than 20000 cases in difference and only 4 of more than 10000. Clearly, the error is well distributed and not focused on single days, like with compartmental models, but also a more precise value is obtained daily as compared to the Gompertz model. This can also be seen by observing an average error rate of around 5500 cases, the best we have seen. This represents an average daily error of around 390 cases, which results in a very precise model.

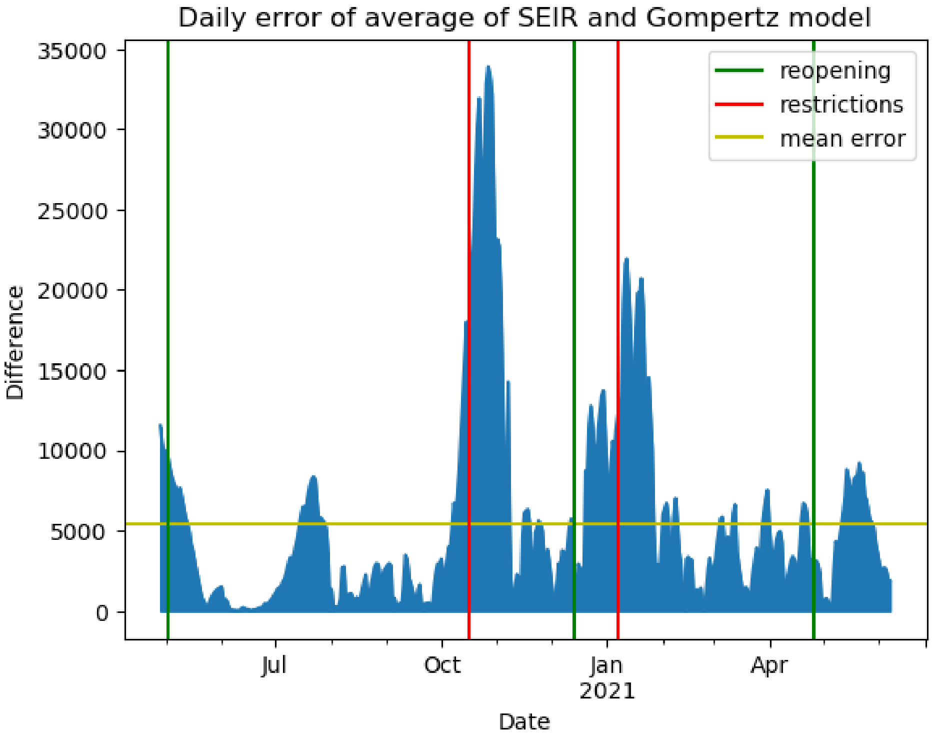


Figure 5.10: Absolute value of the daily error of the Average of SEIR and Gompertz model

Finally, we also compare the relative error rates of all of the models. In this case, we add the daily relative error to compute a cumulative error. In Figure 5.11 we can observe the obtained values. In this case, it is normal that the Gompertz model produces higher values given that it distributes its error evenly. For this reason, when lower real values are made, the Gompertz model has some more error than compartmental models and this leads to it having a total cumulative relative error higher than the one obtained with the real models. We can also observe more clearly how, by a slight margin, SEIR model is the best of the compartmental models. Nonetheless, even by using this metric, the average model is still by far the best of all of those mentioned above.

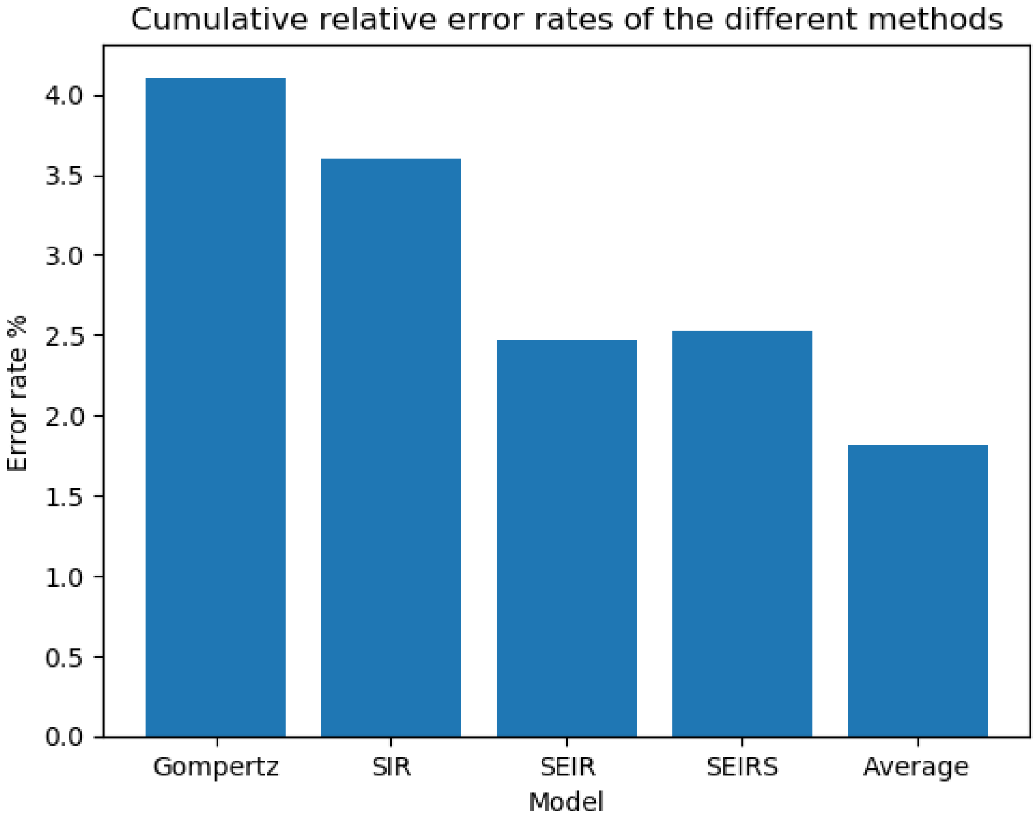


Figure 5.11: Relative error rates of the models

Summary

We summarise the obtained values in table 5.1 for a better understanding of the different models. As we have seen, the Gompertz curve results in the smoothest of the original models but only results in a more accurate curve than SIR. We have also seen how SEIR and SEIRS present less smooth curves but more accurate than the ones mentioned before and how the average curve of Gompertz and SEIR improves all of the previous ones, it now becomes the smoothest of the curves and the most accurate to predict the evolution of the disease. This is due to the fact that the Gompertz model tends to predict less than the real values whereas compartmental models tend to predict higher number of cases. Therefore, the accumulated error from both cancels out and results in the best obtainable curve.

	R^2	Mean difference	Mean Error Rate
Gompertz	0.99323	1499	12000
SIR	0.99319	2331	11000
SEIR	0.99598	2121	9000
SEIRS	0.99579	2354	9000
Average	0.99855	1538	5500

Table 5.1: Metrics used to evaluate the effectiveness of the models

Chapter 6

Conclusion

The main goal of this project was to learn about the effectiveness of different predictive methods for the evolution of the COVID-19 pandemic. These included the Gompertz curve and compartmental methods such as SIR, SEIR, and SEIRS. We mainly wanted to compare which of these methods predicts best the evolution in cumulative cases.

We first presented and analysed the behaviour of all of the models. Learning that the Gompertz curve presented a stochastic approach to predicting the evolution of the pandemic, whereas compartmental models use a deterministic approach to do so. Both models have advantages and disadvantages, on the one hand, the stochastic approach estimates the parameters based on previous data in order to predict the future. On the other hand, the deterministic approach uses a set of differential equations to do so with parameters adjusted to the real data.

We presented different metrics to analyse which of the methods was most effective, and obtained data from the Catalan government's database Dades Obertes. We used this data to train our models with 30 days of data and predict 14 days in the future. We did a retrospective analysis of the evolution of the pandemic by repeating this prediction for every day of this pandemic and comparing it to the real values. We also described how we treated our data and we justified the initial parameters used to train our predictions.

By analysing the results obtained, we are able to conclude that Gompertz produces the smoothest of curves and SEIR produces the best overall model, resulting in the most accurate predictions. That being said, we decided that trying to average both values would yield a better prediction and therefore we applied the same process in order to obtain an average predictor. This model produced the best results overall, predicting the evolution of the pandemic very closely and being smoother than the Gompertz curve obtained. We have also been able to observe how restrictions and re-openings affect all of our models. Clearly, all of the models take some time to realise the change in cases that these restrictions create and tend to have a higher error some time after these restrictions.

In future work, we plan to include more complex compartmental models, such as the one used by Alex Arenas, given the necessary data. In addition, a larger average of models can be used, in which more than 2 models is applied, or we can adjust the weights of the different models, for an even more precise prediction of the evolution of pandemics.

Bibliography

[1] Gompertz, Benjamin (1825). *On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies.* Philosophical Transactions of the Royal Society of London. 115: 513–585. doi:10.1098/rstl.1825.0026. S2CID 145157003

[2] Àlex Arenas web page*https://deim.urv.cat/ alexandre.arenas/*

[3] Lekone PE,Finkenstädt BF. *Statistical inference in a stochastic epidemic SEIR model with control invention: Ebola as a case study.*, Biometrics. 2006; **62(4)**:1170-1177.

[4] Althaus CL., *Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa*, PLoS currents. (2014);6.

[5] Ng TW, Turinici G, Danchin A, *A double epidemic model for the SARS propagation*. BMC Infectious diseases. 2003; 3(1):19.

[6] Català M, Alonso S, Alvarez-Lacalle E, López D, Cardona P-J, Prats C (2020) *Empirical model for short-time prediction of COVID-19 spreading.* PLoS Comput Biol 16(12): e1008431. https://doi.org/10.1371/journal.pcbi.1008431

[7] Brauer F. (2008) *Compartmental Models in Epidemiology.* In: Brauer F., van den Driessche P., Wu J. (eds) Mathematical Epidemiology. Lecture Notes in Mathematics, vol 1945. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78911-6_2

[8] Verhulst P-F. *Notice sur la loi que la population suit dans son accroissement*. Correspondance mathématique et physique. 1938;10:113–21.

[9] Hany Aly, MD *The Weekend Effect and COVID-19 Mortality* https://consultqd.clevelandclinic.org/the-weekend-effect-and-covid-19-mortality/

[10] Gompertz B. XXIV. *On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies.* In a letter to Francis Baily, Esq. FRS &c. Philosophical transactions of the Royal Society of London. 1825;(115):513–583.

[11] Zwietering M, Jongenburger I, Rombouts F, Van't Riet K. *Modeling of the bacterial growth curve*. Appl Environ Microbiol. 1990;56(6):1875–1881.

[12] Gerlee P. *The model muddle: in search of tumor growth laws*. Cancer research. 2013;73(8):2407–2411.

[13] Bruno Gonçalves (2020) *Epidemic Modeling 102: All CoVID-19 models are wrong, but some are useful* https://medium.data4sci.com/epidemic-modeling-102-all-covid-19-models-are-wrong-but-some-are-useful-c81202cc6ee9

[14] Dades Obertes Webpage *http://governobert.gencat.cat/ca/dades*$_o$*bertes/*