Predicting SARS-COV2 growth through maximum likelihood estimation

Nikolaos Gialitsis
DIT, NKUA
BRFAA

Abstract

Today, the world is facing a major crisis, caused by the novel coronavirus SARS-COV2, recently renamed to COVID19. Institutes from all over the globe, are rigorously collecting and sharing their data on public repositories in order to assess the potential spread and effect of the pandemic in order to minimize humanitarian, economical, and environmental losses. In this pursuit, machine learning can be an important ally for governments and other decision-making entities in order to make the best out of the current situation. In this study, we examine the growth curves of the confirmed cases in various countries, and by building and testing different families of linear polynomial models, we make predictions on how the cases might evolve with time. Our results give insight into the different types of viral spread across countries, as well as observing how different political decisions have influenced the current landscape. Last but not least, the predictions can function as worst-case-benchmarks for humanity, and as a community we should try the best in our power to disprove these predictions.

1 Introduction

1.1 The pandemic and its current state

Coronaviruse disease (COVID-19), identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China, outbreak has created an emergency globally, and social distancing and isolation is the only solution to prevent its spread. Several countries have announced fully locked on to tackle this pandemic. The recent COVID-2019 has shaken the globe with incidence cases of more than half-million cases, and a mortality toll of more than twenty thousand to date. . Early on, many of the patients in the outbreak in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring. At this time, it's unclear how easily or sustainably this virus is spreading between people. [2]

At the time this article was written, in 5th of April, WHO reported that one new territory, the Falkland Islands located in the South Atlantic Ocean, was added to the list of those affected by the novel coronavirus [1]. This just shows the terrifying scale of the pandemic as these islands only host a few thousand residents.

On the other hand, China terminates its quarantine and tries to slowly bring its society to where it was before the quarantine. In a recent Q&A between Dr. Anthony Fauci, lead member of the White House Coronavirus Task Force, and the founder of facebook Mark Zuckerberg, [2] Dr. Fauci stressed that judging by how the situation in China will progress, we will be able to get a better understanding of what the future will be in the rest of the world. So, the current days will be very crucial for the spreading of the coronavirus.

1.2 Supervised Machine Learning

The goal of any machine model is to extract knowledge or inference from data, by making various assumptions, based on what it has 'learned'

[1] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200405-sitrep-76-covid-19.pdf?sfvrsn=6ecf0977_4

[2] https://www.youtube.com/watch?v=p1jLKNPxJuQ

Figure 1: Falkland Islands (British Territory)

Figure 1. Epidemic curve of confirmed COVID-19, by date of report and WHO region through 5 April 2020

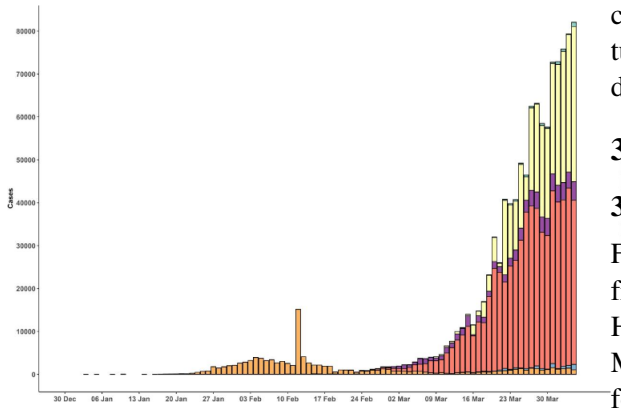


Figure 2: Spread of the virus per continent as of March 5

about the world. "Learning" is conducted by training a model on a dataset and then evaluating its performance against new data (a testing set). In many cases, this process repeats with parameter tuning until it is deemed that the model's performance is adequate for the task, or that the model stops learning . Learning can be thought of as inferring plausible models that are able to explain the data [1].

Within ML there exist three major divisions in the approaches used: supervised learning, unsupervised learning, and semi-supervised learning [3]. In supervised learning, the classes are known a-priori and the training dataset contains class labels for each instance.

A machine learning model is expected to be able to form predictions about unobserved data, having been trained on observed data. This process encompasses a lot of uncertainty. The unobserved data might differ significantly from the data on which the model has been trained on, especially since all real-measurements are subject to noise.

Thus, models should be flexible in capturing many aspects of the trained data, in order to make more confident predictions, when encountered with data that differ in some, but not all, aspects.

2 Motivation

In this study, our goal was to investigate whether simple linear models will be able to efficiently predict the growth rate of the pandemic in different countries. Another factor that is of interest, is to identify similar tendencies between different countries, in order to predict, for example, the future viral spread in Greece by the current available data from Italy.

3 Proposed Method

3.1 Dataset Description

For our analysis, we used publicly available data from the Kaggle Platform, supplied by the Johns Hopkins University.[3] and updated on a daily basis. More specifically, we utilized the time series data for confirmed cases for each country. The table contains information related to the location of the incident (GPS. coordinates) and has a column for each day after the first identified viral occurrence

[3] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

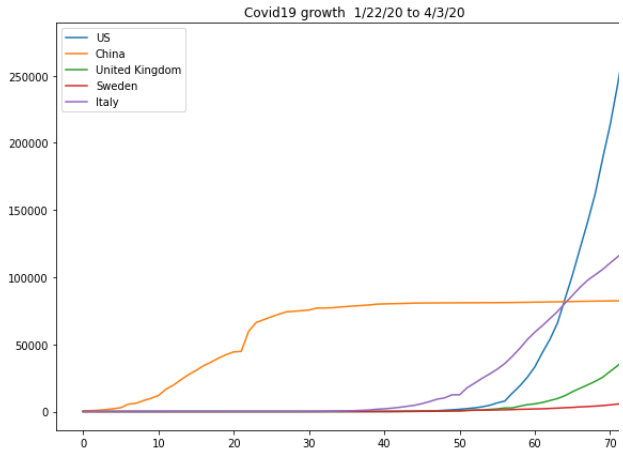


Figure 3: Quantitative comparison between countries that have adopted different strategies for combating the crisis

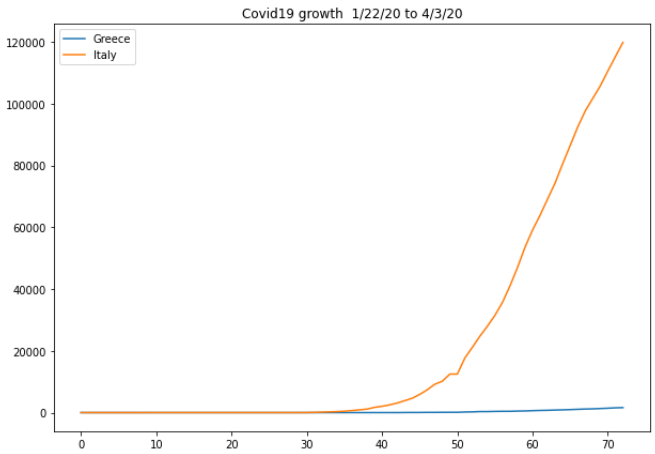


Figure 4: Quantitative comparison between Italy and Greece

in Wuhan. The dataset is sanitized and does not contain missing values or malformed input. Since this resource is in compliance with the FAIR principles, we selected it for training and validating our machine learning model.

3.2 Country Groupings

A few changes needed to be implemented in order to keep only the information that is relevant to the questions we am trying to answer. First of all, even though the GPS coordinates provided would be certainly welcomed if we wanted to visualize as a map the spreading of the virus, in this case it is unnecessary for the analysis and we drop the corresponding columns as well as information about the specific province or territory within larger countries. Finally, for countries that have more than one entry in the table (e.g China which has multiple regions) we group together all their entries and we sum them up. At the end of this procedure, we have a table containing #rows equal to the number of countries and #columns equal to the number of days since the first occurrence + 1 corresponding to the country names.

3.3 Data Exploration

Next we visualize the growth of the virus for each country in order to get a grasp of the "bigger" picture as of today. The plots we construct are useful to verify that the data agrees with indisputable facts regarding the number of cases; for example we would know that something is wrong with the data if the number of cases in China was lower than that of Greece for example.

We only take into account a subset of all countries, including those that have been hit the hardest by the crisis and those that have adopted different strategies for countering the virus, such as the United Kingdom and Sweden.

Greece is compared directly with Italy by investigating their growth-curves. We were interested in assessing their similarities and dissimilarities both quantitatively and qualitatively, with the underlying question of whether the cases in Greece can reach the magnitude of those in Italy.

3.4 Further preprocessing

After observing the data, we noticed that for countries in which the virus arrived late compared to others, the vectors were quite sparse as the entries for first couple of days were zero. Intuitively, this is expected to negatively influence the predictive power of our models since the zeros do not contribute information related to the growth rate of the viral cases.

So, in order to compact the vectors and set an equal footing between different countries, we removed entries for which the virus had not yet affected.

3.5 Model Definition

In order to build a model on the corpus of confirmed cases, we employ a simple linear polynomial model in the form of

$$y = f(x|\theta) = \theta_1 * x + \theta_2 * x^2 + \theta_3 * x^3 + \theta_n * x^n \quad (1)$$

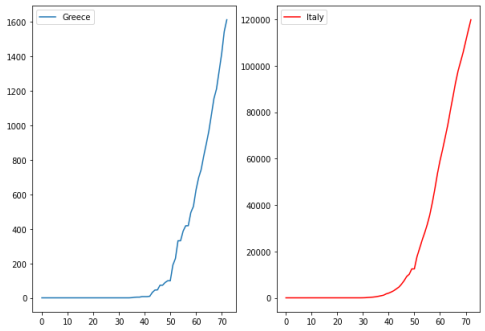


Figure 5: Qualitative comparison between Italy and Greece

where θ is the parameter vector learned in training and x the available data . It is important to note that such a model is linear relatively only to the parameters θ and not the actual data x.

3.6 Partitioning Data into Training and Testing sets

In order to develop an efficient model, we kept only a portion of the dataset for training and we used the remaining data points for validating our model.

More precisely, we performed the following steps:

1. define date X as the critical date, after which, the model does not observe any data. The date is supplied by the user in the form of an integer, corresponding to the number of days after the first viral occurrence.

2. using random sampling, problematically select a portion of the days before date X as the training set. In our analysis, we considered 65% of the dates before date X. The cases recorded for those dates consist the training set for our model.

3. use the dates after X and the dates not selected before X as the testing set.

Of-course if X is close to the current date, the model receives more information in the training phase but then, we might not be able to assess its' performance accurately on the testing set as the testing set will become very limited. So, it is important to select X in such a way that balance is achieved between training and testing.

3.7 Learning Procedure

For each country, we construct different models iteratively assuming $n \in [1, 9]$. As n increases, the model becomes more complex and tends to overfit the training data, resulting in worse predictions on the testing set.

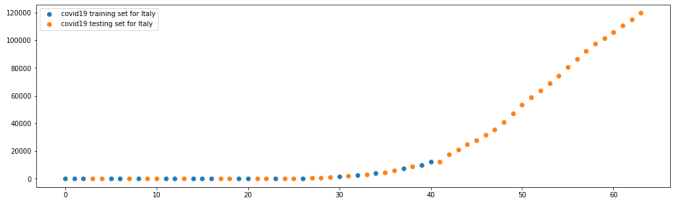


Figure 6: Dataset partitioning for Italy: xaxis='days since first occurrence in Italy' , yaxis='number of confirmed cases'

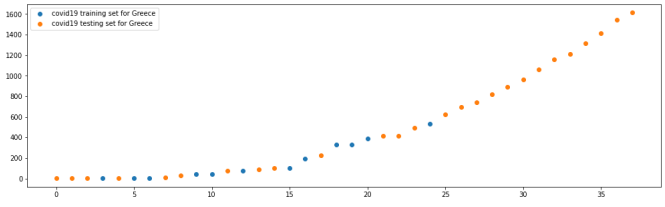


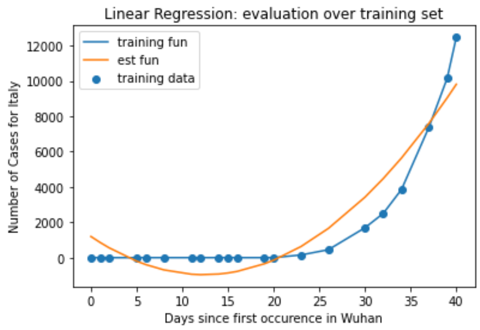
Figure 7: Dataset partitioning for Greece : xaxis='days since first occurrence in Greece' , yaxis='number of confirmed cases'

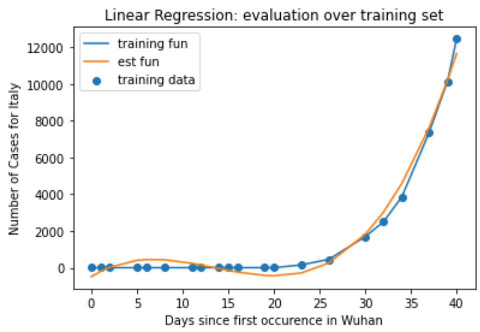
For each iteration, we estimate the optimal parameters according to the Maximum Likelihood Estimation (MLE) . According to MLE, the answer can be given in a closed form in terms of linear algebra as [4]:

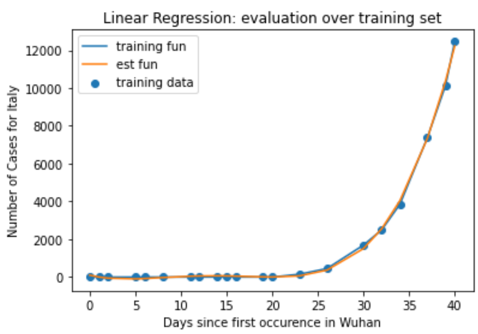
$$\theta_{ML} = (X^T X)^{-1} X^T \tag{2}$$

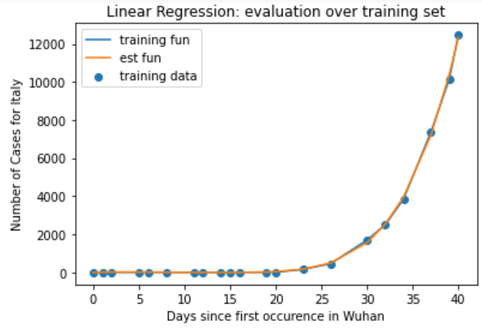
3.8 Model Validation

For each model, we perform predictions on both the training and testing set and as a metric of performance we use the Mean Square Error (MSE).









MSE is defined as

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - {y^*}_i) \quad (3)$$

and is a very common way to measure the distance between observations and predictions.

3.9 Model Selection

In order to speed-up the learning process, for each country, we stop iterating over polynomial degrees, after reaching an elbow point in the loss function. An elbow point is the point at which, the loss at the testing set stops decreasing significantly and either remains unchanged or increases. By doing so, we respect both the underfitting-overfitting criteria by skipping configurations which result in non-optimal losses on the testing set, and the 'Occam's razor' empiric principle which when applied to machine learning, signifies that between two equal models we should choose the simpler one. The simpler model in our example is the one with the lower degree.

When an elbow point is reached, we keep track of the polynomial degree and the estimated θ vector which is considered the optimal.

3.10 Future Predictions

The goal of this analysis essentially is, to be able to predict how the virus will continue to spread in different countries. So without further ado, we

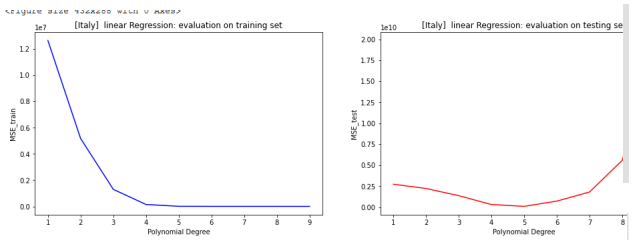


Figure 8: Elbow point for Italy

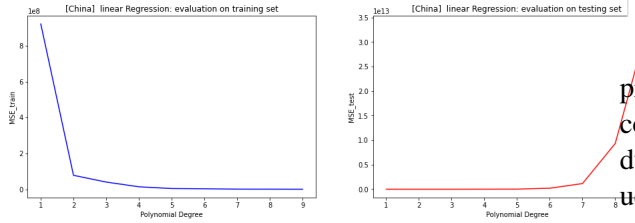
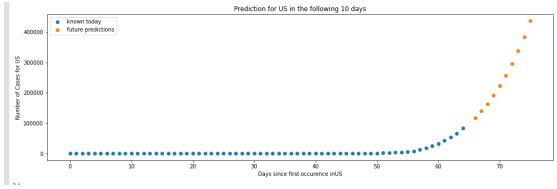
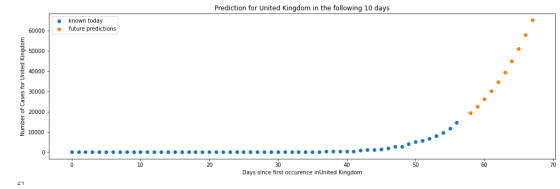
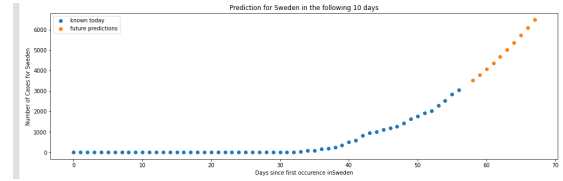
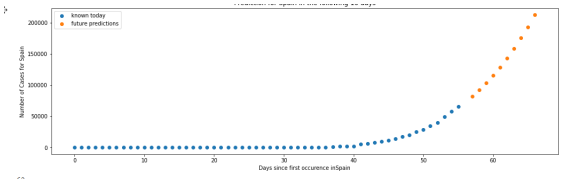


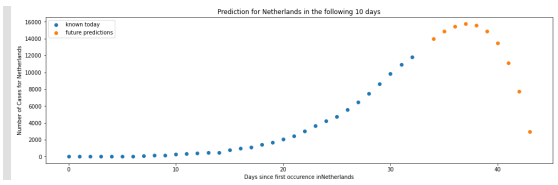
Figure 9: Elbow point for China

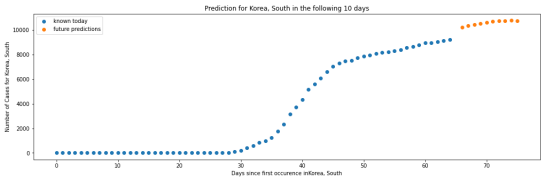


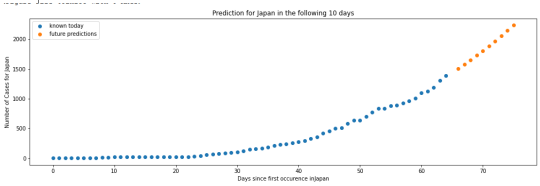












proceed with the predictions by using for each country the best model and parameters estimated during the training phase. We predicted the values of the "covid19" function for the next 10 days after the maximal date at which the training set stopped. This parameter along with the ones previously mentioned can be tuned easily per user request.

The results are depicted in Figures References-fig:1 and are discussed below.

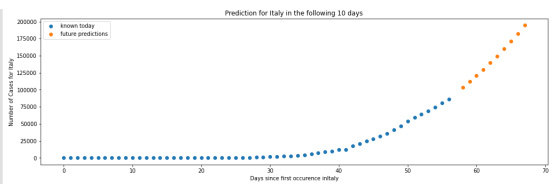
4 Discussion

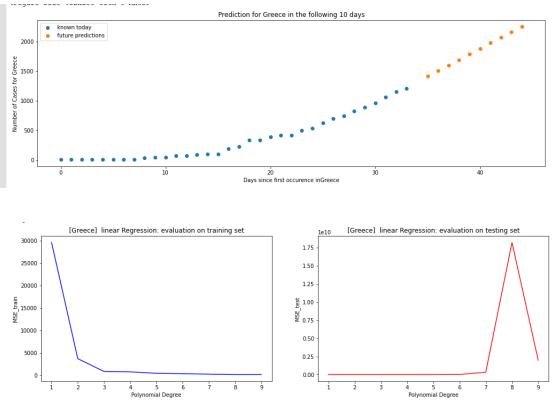
In the previous sections, we explained how our data was gathered and processed and how we build predictive models trained on confirmed cases of covid19 using families of linear models. In this section, we will proceed with discussing the results obtained.

4.1 Greece-Italy comparison

First of all, let's start with Greece, a country that has not yet experienced a dramatic increase in the number of cases. It is important to consider that its neighboring country, Italy, is one of the main players in the global covid19 map since the cases are increasing exponentially, exceeding 100K. On the other hand, as of now (5th March), the cases in Greece are less than 2K. We can observe this significant difference in Figure 3.4.

We also plotted the cases for the countries independently 3.4, and thus being able to observe the qualitative growth of the curves. In that aspect, we were surprised to observe that the curve structure





is very similar. So, potentially even though Italy's cases exceed those in Greece by orders of magnitude, in theory if the virus continues spreading in Greece with the same intensity, it is expected to reach the same status as Italy's so the population should be really cautious. The good news is that since the two curves are similar, it is easier for a machine learning model to predict accurate results in both cases, in contrast to having two radically different data generating functions.

4.2 The global landscape

In Figure 3.1 we observe the differences between the spreading of the virus in key-countries that have either been severely damaged by the virus (China, US, Italy) or countries that have followed a counter-intuitive approach such as Sweden (no strict quarantine) and the United Kingdom (herd immunity) .

The first thing that automatically draw attention is the immense growth of cases in US, which undoubtedly did not implement drastic enough measures to prepare for the spreading of the virus (NY city is a viral hub).

China on the other hand, which is the source of the pandemic seems like it has reached a plateau in cases with less than 100K cases as of today. Italy on the other, which resides into a different continent has exceeded the cases of china and looks like it is continuing increasing rapidly.

The growth rate in the United Kingdom and in Sweden does not compare with the that in the previous countries but especially in UK, the curve is becoming very steep (uphill) and that is a worrying sign.

Sweden on the other hand, for now, seems like it does not face any imminent threat even though its measures were much less strict than those implemented by Italy for example which is currently

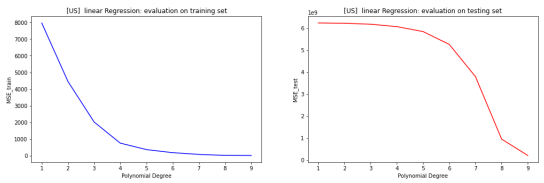


Figure 10: There is no elbow point for US

under total quarantine.

4.3 Curve Complexity per Country

In table 7 we have grouped the countries based on the polynomial degrees in which they reached an elbow point. This does not propose that a country with a higher degree is more or less vunerable, only that the growth function of the virus is similar. So, this table can be thought of as a clustering of countries based on their viral spread. We see that the least complex models were recorded for Italy,Greece,Japan,the Netherlands,Sweden and China whereas the most complex model was recorded for US. The results are in agreement with what we previously observed between Greece and Italy, since their growth patterns looked pretty similar. For US, on the other hand, this complexity is also evident in Figure 4.3 since even at the highest degree (D=9) the loss function keeps reducing in the testing set, hence there is no elbow point.

4.4 What does the future await?

Everything signifies that the threat is far from over, as the predictions for most countries we analysed do not show positive signs that the virus is fading away.

The only countries for which the model predicted that the number of cases will stop increasing are China and the Netherlands. However, the plots show that the curve is predicted to go down, but that is impossible since the yaxis reflects the total number of confirmed cases. The way that we interpreted those results was that the prediction signifies a plateau. However, these results should be taken with a grain of salt since we need to check the models again; the prediction for the Netherlands is especially counter-intuitive.

For the majority of the countries, though, the plots seem logical as the predictions are mostly in line with the previous measurements. The prediction is very alarming for US ($\geq 400k$ cases in the next 10 days) , Germany (160K), Spain ($\geq 200k$)

and Italy ($\geq 200k$).

Japan and Greece, for the time being, show great promise as the curves tend to look more like straight lines, so they do not follow the exponential paradigm of most countries. Both Japan and Greece are predicted to have $\approx 2k$ cases in the following 10 days.

5 Estimating the number of true cases by the number of available data

Randomized Clinical Trials [2] can be used in the population in order to get an intuition into the state of whole population, healthy or not, for the development of drugs to combat the covid19 coronavirus. Through statistics (and hypothesis testing), we can estimate the minimal necessary sample that is needed in order to draw conclusions for the state of the whole country, and assess the specificity and sensitivity of any proposed treatment. In doing so, qPCR can be employed, since by multiplying genetic material and recording its' quantity, after a number of X samples we will be able to observe the genetic material of the virus. Of-course, since this is a retrovirus, and thus it is made of RNA we have to use reverse-transcriptase to convert its material to DNA in order to use any form of PCR.

6 Future Work

This study conducted was a preliminary investigation of the data. A very simple model was built for predicting the growth tendencies of the viral cases for various countries. In the future, more countries can be taken into account, and more model architectures can be tested out, such as non-linear models. Heterogeneous data from other sources, such as data from hospitals, tourism, biomedical literature and experiments, and data from past epidemics or pandemics can be embedded into the pipeline to result in more evidence-based predictions. Last but not least, we can run the models multiple times with different samples of the training set and record their average or median performance. It is also needless to say that as time passes, more data will become available to the scientific community.

7 Conclusion

The world is in a state of crisis and we need to act quickly. Data is becoming available at a record-high rate and thus, machine learning models can be built to take advantage of this rich information in order to help governments make the optimal decisions. By training polynomial models on the available data, we were able to make predictions about the future state of the pandemic in terms of total cases. We also identified groupings between different countries using as a criteria the models' complexities. However, more experimentation needs to take place in order to arrive at more well-founded conclusions.

Polynomial Degree

Countries	**D=3**	**D=4**	**D=5**
	Italy	South Korea	UK
	China		Spain
	Greece		Germany
	Japan		
	Netherlands		
	Sweden		

Growth curve complexities across countries

References

[1] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452, 2015.

[2] Andre C. Kalil. Treating COVID-19—Off-Label Drug Use, Compassionate Use, and Randomized Clinical Trials During Pandemics. *JAMA*, March 2020.

[3] Maxwell W. Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, June 2015.

[4] Sergios Theodoridis and Konstantinos Koutroumbas. Pattern recognition. 2003. *Google Scholar Google Scholar Digital Library Digital Library*, 2009.