EDA and SIR prediction model in R for COVID-19 cases and deaths

1st Alessandro Cinti
Università degli Studi di Perugia
Ingegneria Informatica e Robotica
curriculum Data Science
Perugia, Italy
alessandro.cinti@studenti.unipg.it

2nd Ciro Rapone
Università degli Studi di Perugia
Ingegneria Informatica e Robotica
curriculum Data Science
Perugia, Italy
ciro.rapone@studenti.unipg.it

3rd Mattia Rengo
Università degli Studi di Perugia
Ingegneria Informatica e Robotica
curriculum Data Science
Perugia, Italy
mattia.rengo@studenti.unipg.it

***Abstract*—In this document we'll see the evolving epidemic curve of COVID-19 starting from the first months of pandemic and reaching out to nowadays. We'll use the updated WHO COVID-19 dataset to analyze and predict the cases and deaths curves day by day. First we analyzed and corrected the dataset and then proceeded to visualize the plots of the most significant data. We then studied the distribution of the data and performed the normality tests. Finally we focused on the creation of a predictive SIR model to predict the trend of COVID-19 in the near future.**

***Index Terms*—COVID-19, WHO, EDA, SIR Prediction model**

I. Introduction

This document is an exploratory analysis and SIR model about the COVID-19 epidemic in R.

II. About the Dataset

A. WHO COVID-19 DATASET

The COVID-19 dataset made by WHO [1] contains all the data about deaths and infections for every country from the start of the pandemic. The available columns are "Date reported", "Country code", "Country", "WHO region", "New cases", "Cumulative cases", "New deaths" and "Cumulative deaths".

```
> summary(data)
 Date_reported         Country_code          Country           WHO_region          New_cases
 Min.   :2020-01-03   Length:201924      Length:201924      Length:201924      Min.   : -32952
 1st Qu.:2020-08-02   Class :character   Class :character   Class :character   1st Qu.:     0
 Median :2021-03-03   Mode  :character   Mode  :character   Mode  :character   Median :    24
 Mean   :2021-03-03                                                           Mean   :  2535
 3rd Qu.:2021-10-02                                                           3rd Qu.:   509
 Max.   :2022-05-03                                                           Max.   :1252940
 Cumulative_cases     New_deaths       Cumulative_deaths
 Min.   :       0   Min.   :-2448.00   Min.   :     0
 1st Qu.:     200   1st Qu.:    0.00   1st Qu.:     2
 Median :   11649   Median :    0.00   Median :   150
 Mean   :  629103   Mean   :   30.91   Mean   : 11813
 3rd Qu.:  164899   3rd Qu.:    6.00   3rd Qu.:  2771
 Max.   :80598784   Max.   :11447.00   Max.   :986437
```

Fig. 1: Summary of the WHO dataset used

Note that some of the contagion data values are negative, this is because the WHO has decided to report false positives as negative values to permit values compensation. [2] To always obtain updated data within the code, we have decided to load the dataset dynamically via the URL provided by the WHO; that way, every time the code run again, new data will be included.

B. CODE USEFUL PARAMETERS

To make the most of this ever updating data functionality within the plots and calculations, we have introduced two time variables "start" and "end" that allow you to select the time window of the data to be used. By default, the values range from February 2020 to today's date.

A "steps" variable has also been introduced that allows you to change the temporal granularity of the data. In this way you can manage the temporal resolution of the data used and therefore of the plots created and at the same time you can adapt the computational burden to which your machine is subjected.

You can also set a "delay" parameter to cut-off the last n days from the time window and data. This is very useful when there's a delay in the WHO dataset's daily data submission causing missing reports on latest date. This behaviour could cause missing information on some plots.

In the code there's also a "prediction window" parameter, you can change it to choose the time window showed in the prediction plots.

C. TMAP WORLD MAP DATASET

To view the covid data on a world map and to obtain from this further data to focus our analysis on a specific country we used the tmap library and in particular the shapes of its world package. In addition to data like the shape of the country and the name of the country, it contains other countries related data such as estimated population. We joined the information from the WHO dataset with the tmap world dataset by country (called "sovereignt" in tmap world dataset) to obtain a unified dataset containing all the information for each country that we'll use later.

III. Exploratory Data Analysis

In this phase we have visualized and analyzed the data. First we checked for the presence of null values, converted the date from char type to date format and printed the most important information on the screen. Then we started our EDA phase: Firstly we wanted to see how the pandemic develops in different areas of the world and then we focused our efforts on the Italian situation.

During the previous join operation since some countries was entered in the two datasets using different names, like ”Syrian Arab Republic” instead of ”Syria”, NA values appeared in some resulting rows. We fixed this changing the mismatched names prior of running the join instruction. Now our dataframe contained all the joined information without errors or NA values.

A. WORLD STATUS REGARDING COVID-19

After fixing the dataset, we graphed the global pandemic situation both daily and in the long term. We started plotting the curve for daily cases and deaths worldwide. To know in which countries the COVID-19 spread most we visualized a plot about cumulative cases of the seventh countries with higher cases during this pandemic to compare them. The resulting countries are: United States of America, India, Brazil, France, Germany, Russia and Italy.

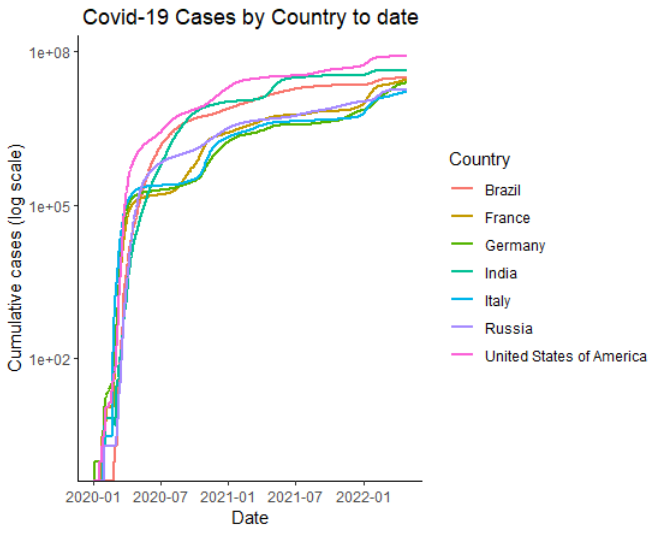

Fig. 2: COVID-19 World cumulative Cases to date.

This countries are situated in different world regions, so we also wanted to highlight in which WHO Region the virus attacked the most, in a Continents-like bar plot.

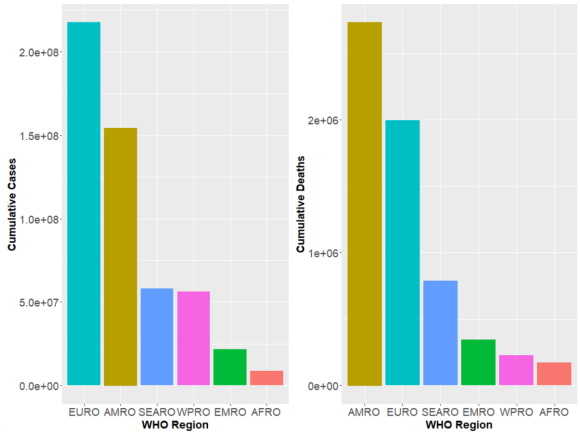

Fig. 3: COVID-19 World Cases and Deaths to date.

Looking at the graphs (Fig.3) it is possible to see that there is a notable difference between the amount of cases and deaths between the various regions. This may be due to the size and density of the population or the way the pandemic situation has been handled. To verify this we decided to do the same plot using the Infections Ratio index grouped by WHO region. As we can see Europe is actually the region most afflicted by the virus. From the new bar-plot it can be seen that, although the American’s territories have obtained the highest number of deaths of all, Europe has been the territory with the most infections in the world, which confirms the result of the bar-plot in Fig. 3. As further proof we have verified that our data are in line with the WHO official data [3].

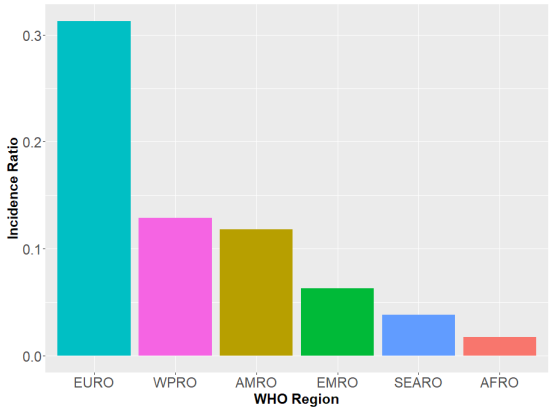

Fig. 4: Incidence Ratio of virus in the WHO Regions.

B. WORLD MAP COVID PLOT

To have an even broader view of the world scenario, we decided to display a world map with a chromatic scale where each country has a darker color tone as the number of cases increases. As we said earlier, all this has been possible to achieve through a join operation between the COVID-19 dataset of the WHO and the World dataset of the tmap library. The scenario just described is represented according to the current date.

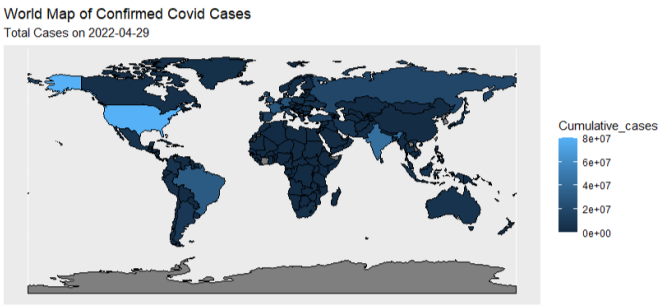

Fig. 5: COVID-19 Cases in the World on current day.

Notice how in the plot (Fig.4) on the new cases we can see that the derivative function is increasing. This is and indicator that the power of the virus is decreasing but its ability to infect is rising in spite of the approaching of the summertime

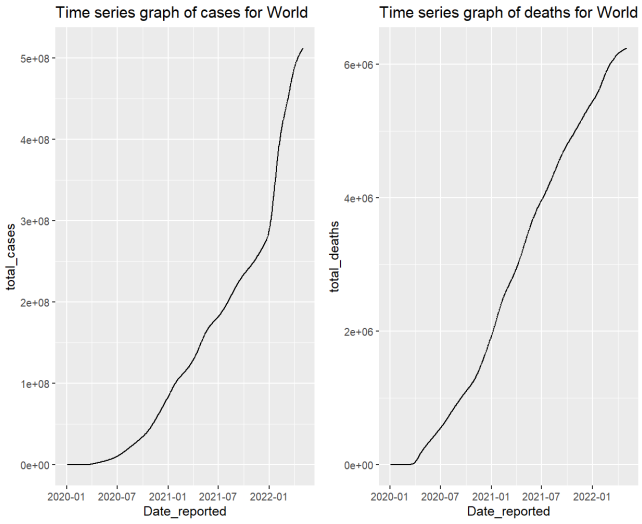

Fig. 6: COVID-19 World Cases to date.

and with vaccines still in place. We can see it graphically comparing the derivative functions (slope of the plot) from the infections plots versus the deaths plots. This trend is valid both worldwide and in Italy as we'll see later.

C. FOCUS ON ITALY STATUS REGARDING COVID-19

In addition to the world situation, we focused on the Italian territory by graphing the new deaths and new cases day by day.

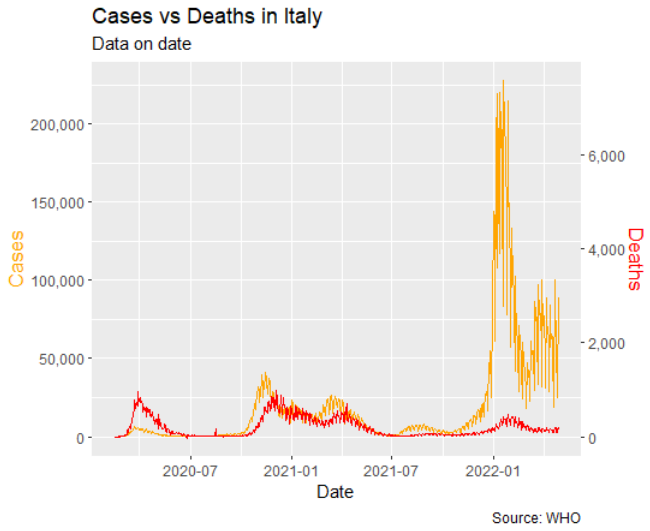

Fig. 7: COVID-19 Cases and Deaths in Italy day by day.

IV. NORMALITY TEST

With the help of the Normality Tests we can evaluate the normality of the data based on the value of the p-value, which is the result of the algorithm applied to verify this property. For this analysis, we used the Shapiro-Wilk test, which is one of the most powerful tests for verifying normality.The data studied with this test are daily cases and deaths contained in the Covid-19 dataset. Below are the results of the test conducted on the two data: what can be seen by looking at the value of the p-value is the non-normality of the data, since the value is well below the minimum acceptance threshold of the hypothesis for which the aforementioned data are to be considered normal.

TABLE I: Shapiro-Wilk Test results.

Shapiro-Wilk Normality Test	W	p-value
Daily cases	0.58206	2.2e-16
Daily deaths	0.45687	2.2e-16

The same conclusions can also be drawn by looking at the qq-plots of the daily data, where in both curves it is possible to note that the data are distributed far from the interquartile line.

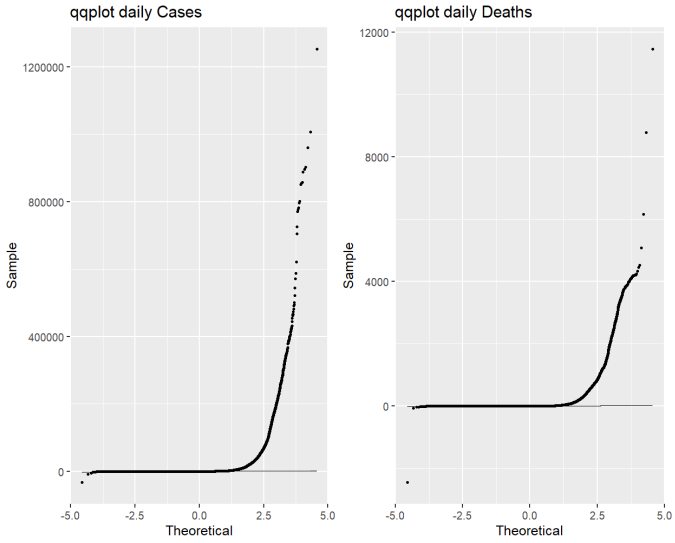

Fig. 8: qq-plot of daily cases and deaths.

V. SIR PREDICTIVE MODEL FOR COVID-19

To make prediction about the next two weeks Covid-19 curve we used the combined informations from the WHO Covid-19 dataset and the tmap World dataset, put them in a SIR model to predict how the pandemic will develop. Note that for this section, we focused on a single country, both for computational needs and for make tailored predictions on a specific country without affecting its predictions with data from countries that have different population density, habits, anti-covid normatives and different climate.

We focused on Italy but you can change it in the code using the variable "target country".

A. DATA SELECTION BY FOCUSED COUNTRY

First of all we extracted the country specific data from the joined datasets mentioned above. These data are ordered by date because we'll use them to train the model and predict the pandemic curve evolution day by day.

B. FIRST SIR MODEL USING EPIMODEL LIBRARY

This library was initially developed for the HIV SIR MODEL but later it has been upgraded and expanded as generic SIR model and then specialized again for the COVID-19. It use a pre-trained SIR model (others models are available too) and specific initial conditions to simulate the covid spreading over time.

The initial conditions are set using the recent data available for the focused country, like estimated susceptible, infected and recovered population [4] other than Infection Ratio, Infection probability [5], recovery medium time and other factors called "params" modelling people encounters, travels, births and deaths [6]. All these datas are country specific and are present in the dataset or derived using specific columns like for the Infection Ratio.

The simulation is carried away using the initial conditions like a day zero, and the SIR model doesn't permit reinfections so its data predictions are limited by this factors. We used this SIR model also to predict the case curve in the next few weeks.

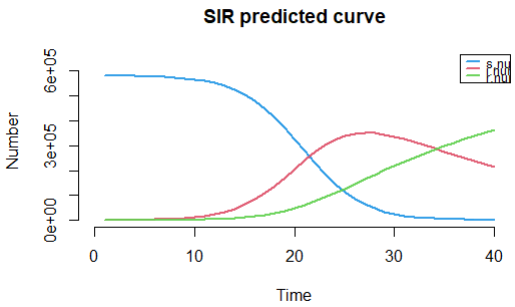

Fig. 9: Predicted SIR Curves for COVID-19 using Epimodel

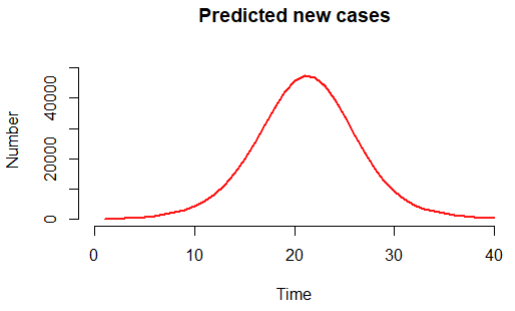

Fig. 10: Predicted Curve for COVID-19 using Epimodel SIR model

C. SECOND SIR MODEL: CUSTOM TRAINED ONE

The previous SIR model couldn't show us how the deaths curve evolves. this is a dramatic yet important factor to consider; So we decided to train a SIR model using our data and a sliding time window. We started by the mathematical equations, defining a model for a specific country and then training and testing it against the data of the WHO covid dataset. We achieved it using the lag function to create a sliding window to train the model using a certain window's data and the testing its prediction accuracy using the next window and reiterating this process until a best r squared error is reached and then using the resulting model to predict the next weeks of pandemic.

We used this to predict the pandemic curve and specifically the deaths predicted. The predict model is more regular the the raw data thank to the mean smoothing process used during the training using the WHO data.

We can observe that death cases curve tends to regularize now that we're approaching the summer and with more vaccinated/recovered people surviving to the covid without serious consequences.

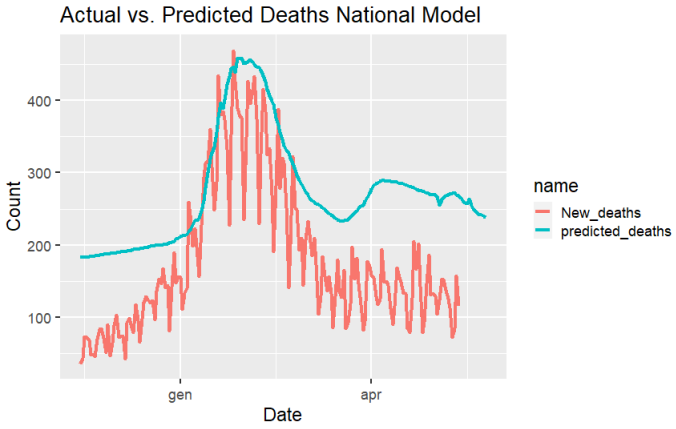

Fig. 11: Real and Predicted Curve for COVID-19 using custom trained SIR model

This also confirm the behaviour that we noticed in the EDA phase where the data showed a greater infectious capacity of the virus but a lower aggressiveness and virulence. This appears to hold for the next few weeks according to our sir model prediction.

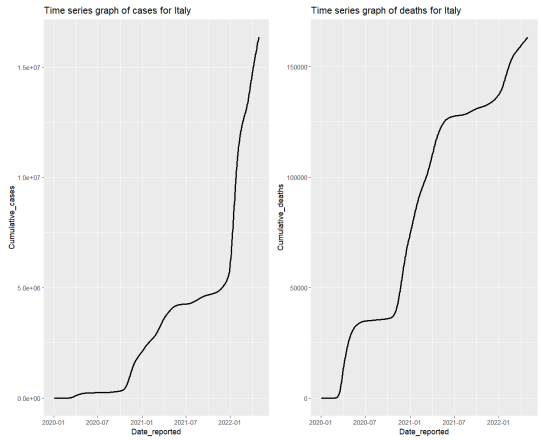

Fig. 12: COVID-19 Italy Cases to date.

VI. CONCLUSION

To summarize the purpose of this analysis in R was to monitor the development of the epidemic curve from the point of view of cases of infection and deaths throughout the world with a subsequent focus on the Italian scenario and in a time horizon that goes from the first days of pandemic to date.

Furthermore, an attempt was made to create a predictive model to be able to estimate the trend of the curve of infections and deaths in the days to come. It was possible to see from the graphical representations that the peak of infections occurred during the winter and autumn seasons, while during the hottest periods of the year the trend of the epidemic curve showed a decline, both on a global scale and on the Italian front. This is due to the fact that with higher temperatures the virus has less chance of proliferating.

Given the great spread of the virus due to the increase in variants circulating in the world, it is difficult to predict with certainty the future development of the pandemic situation, so with the two predictive models implemented in the analysis we tried to hypothesize a possible evolutionary scenario based on of the epidemiological data acquired so far.

REFERENCES

[1] https://covid19.who.int/WHO-COVID-19-global-data.csv

[2] As stated in https://covid19.who.int/data "Data sources": "Due to the recent trend of countries conducting data reconciliation [...] such data may reflect as negative numbers in the new cases / new deaths counts as appropriate."

[3] https://covid19.who.int/table?tableChartType=heat

[4] https://www.worldometers.info/coronavirus/country/italy/

[5] https://www.science.org/doi/10.1126/science.abg6296

[6] Immigrazione https://www.istat.it/it/archivio/245466, Popolazione residente per sesso, nati vivi, morti, saldo naturale, saldo migratorio, saldo totale e tassi di natalità, mortalità, di crescita naturale e migratorio totale https://seriestoriche.istat.it/fileadmin/documenti/Tavola_2.3.xls , Emigrazione https://seriestoriche.istat.it/fileadmin/documenti/Tavola_2.10.1.xls