INTRODUCTION

The models described in this report were created with the aim to forecast the total number of deaths during the pandemic of COVID-19 in the last two weeks of May 2020 in three countries: France, Japan and Philippines. The model for Philippines will not be described in order to respect the maximum number of pages. However, the forecasting error will be discussed at the end of this report.

The data set is composed of 11 variables. *Deaths* is the number of deaths reported by each country and it represents the response variable. The other variables are the explanatory variables and they give information about the date when deaths has been reported, like *dateRep, day, month* and *year; Cases* is the number of people confirmed positive to the COVID-19; then there is information about countries name (*countriesAndTerritories*), two different type of code that identify the country (*geoID* and *countryterritoryCode*); finally, there is information about country's population in year 2018 (*popData2018*) and the continent to which the country belongs (*continentExp*).

The second part of this report is the exploratory analysis which provides information about the general features of the time series of the three countries in the data set. Then for each country starting with France, there will be an analysis of the trend and seasonality, if they are present, followed by an analysis of the residuals. The same procedure will be then replied for Japan. Finally, there will be discussion and conclusion about the best model forecasting the number of deaths for the last two weeks of May 2020.

EXPLORATORY ANALYSIS

The data set has been divided in order to have one data set for each country, in this way, it is possible to perform separate analysis and describe characteristics of each nation. Furthermore, the three data set obtained have been divided into training set, which is composed of data from March to the 15th of May 2020 and is used to create statistical models, and test set which is composed of data from 16th to 31st of May 2020 and is used to check the perform of forecasting.

Figure 1 shows the time series of the three countries from training data set. France has an upward trend until the beginning of April when the curve appears to flatten out, until around the 15th of April where it starts a downwards trend. This behaviour could be explained with the measurement taken by the French government to contrast the spread of the virus. In fact, a total lockdown was official from the 17th of March and the effect of this action has repercussions two weeks later, as experts declared to expect. No seasonality is observable for the three time series.

In Japan the number of deaths seems to weakly and constantly increase for the whole period. In Philippines instead, between 20th and 25th of March some measurements were taken in the country such as more tests trough the population and travel restrictions. Therefore, the curve shows an increment of number of death until 10th of April and then a weak and constant decrement.

More generally it is possible to notice that all of them are non-stationary time series because the means are not constant and also some variations in values are present indicating a non-stationarity. Another distinguishable characteristic is that the variance does not seem to be the same through the period in examination. Therefore, a logarithmic transformation has been applied to the data, and figure 2 shows that variances, in this way, is more homogeneous.

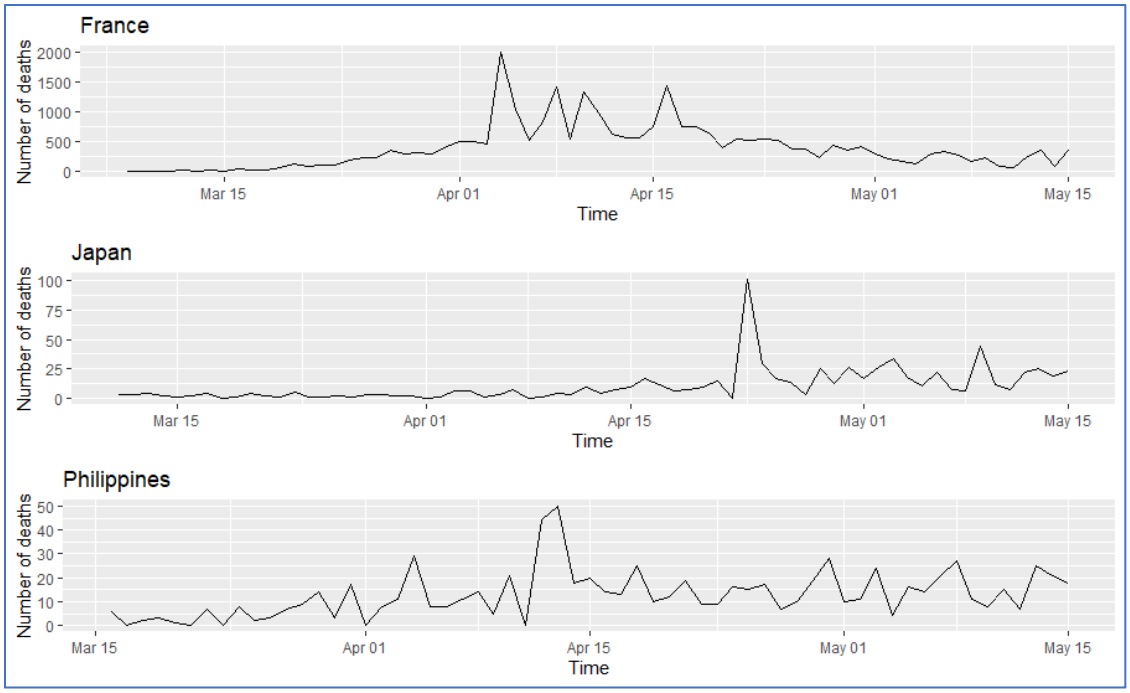


Figure 1

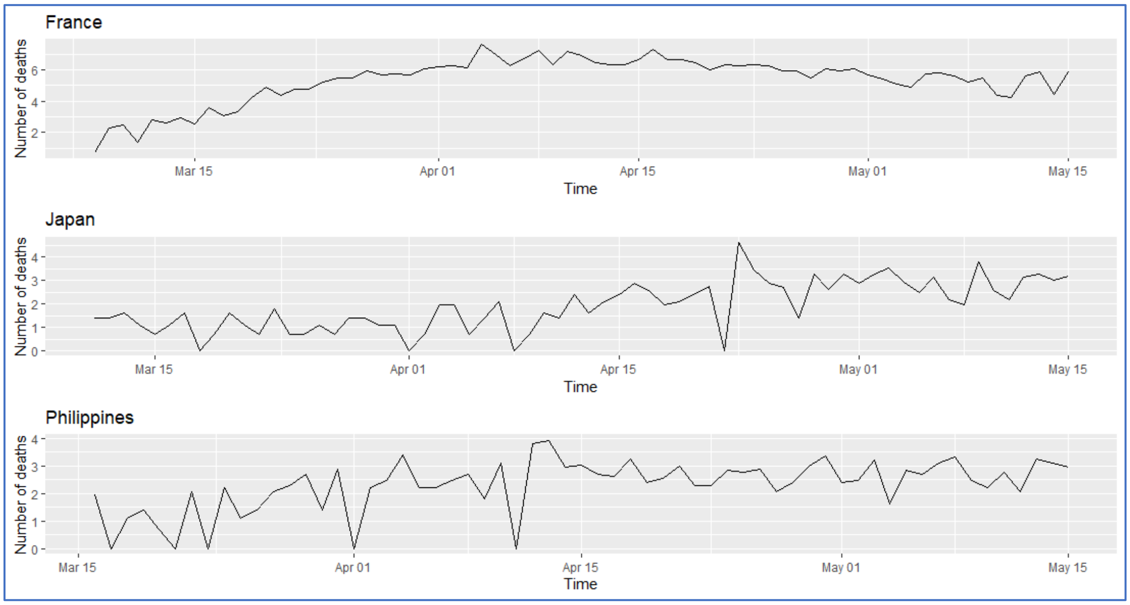


Figure 2

ANALYSIS - FRANCE

The first step in the analysis is to create a model that describes, in this case, only the trend. Initially, a simple linear regression has been created but it did not seem to be appropriate because this model can describe only 26% of the data as indicated by R squared and adjust R squared of the model summary. Therefore, a non-linear regression technique like natural spline has been used and it is shown in figure 3 where it is possible to see the changing form positive to negative slope.

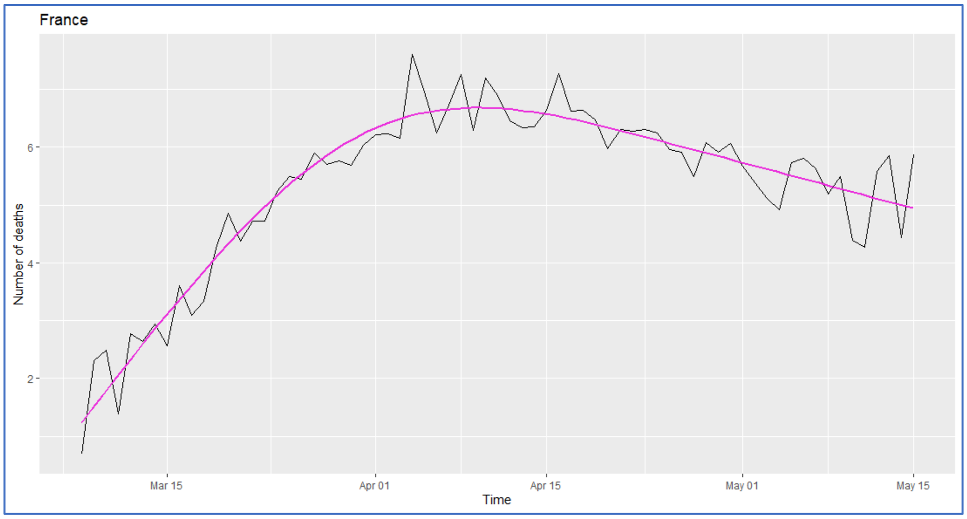


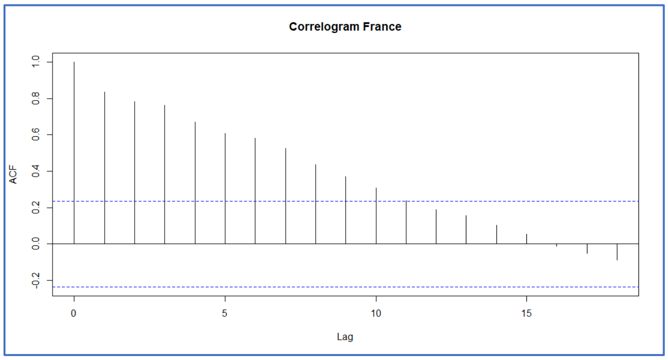
Figure 3:

The summary for this model is as follow:

```
Call:
lm(formula = trainFrance$deaths ~ France_trend)

Residuals:
       Min         1Q     Median         3Q        Max 
-3.428e-15 -6.753e-16 -1.192e-16  5.727e-16  1.607e-14 

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)    
(Intercept)  6.931e-01  9.283e-16 7.467e+14   <2e-16 ***
France_trend 8.619e+00  1.644e-15 5.244e+15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.299e-15 on 67 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 2.75e+31 on 1 and 67 DF,  p-value: < 2.2e-16
```

It seems to be too perfect as indicated by R-squared and Adjusted R-squared but it explains the trend better than the linear one.

In figure 4 it is reported the correlogram and it clearly indicates the presence of a trend and no signs of seasonality.



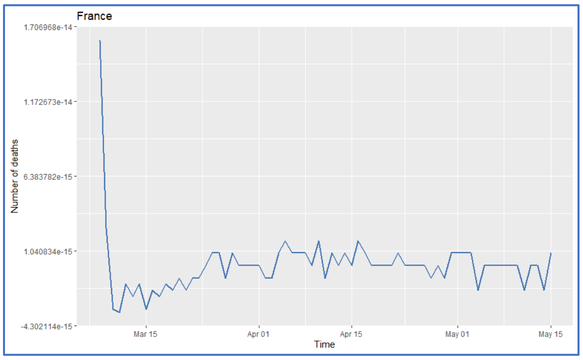


Figure 4

Figure 5

After subtracting the fitted value to the original data, the series seems to be stationary as illustrated in figure 5. The plot looks like a random process with mean equal to zero.

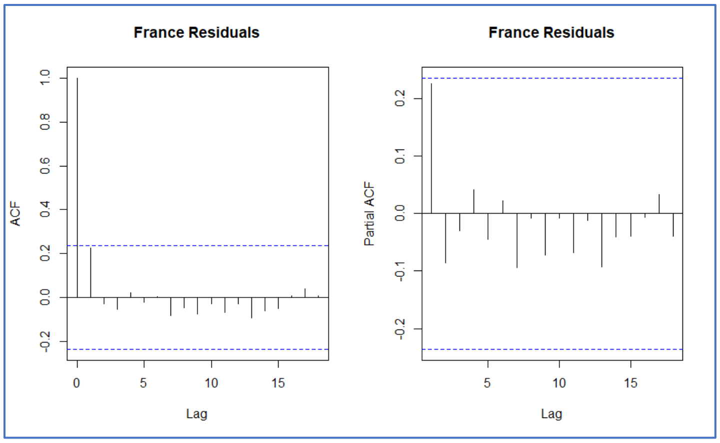


Figure 6

Figure 6 shows the auto correlation function and partial auto correlation function which suggest a short-term correlation has been removed by the above described process. In this case, there is no need to adjust the errors of the parameter estimates of the mean part of the model and it is possible to proceed with the forecasting section.

FORECASTING – FRANCE

The next step is forecasting the trend of the series during the last two weeks of May 2020 in France. Two technique have been used in this report. The first is Exponential Smoothing. In order to obtain the best forecasting model, it is necessary to choose the most appropriate value of alpha which is the smoothing parameter. Figure 7 shows the best value calculated minimizing the root mean square prediction error, indicated by the pink dot and corresponding to a value of 0.1.

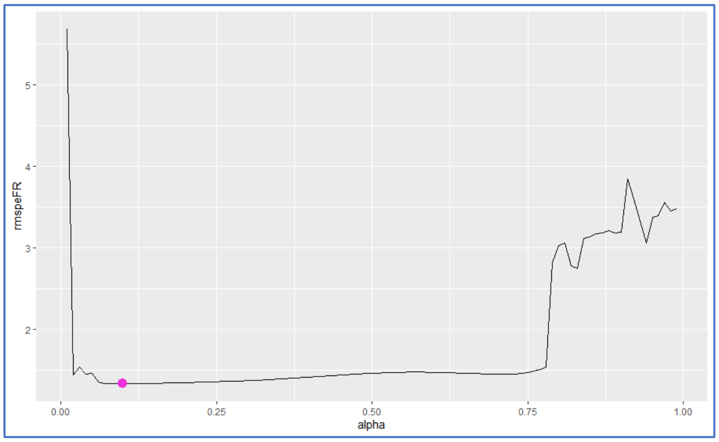
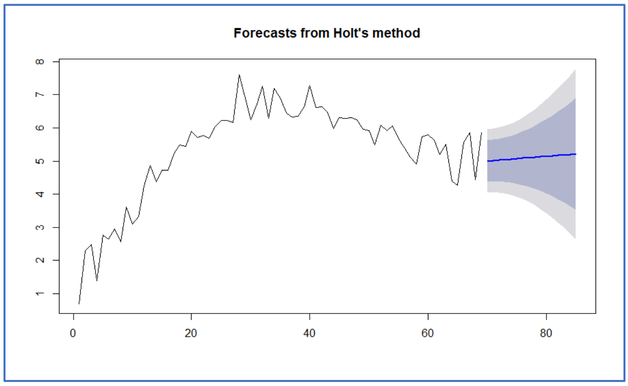


Figure 7

This value of alpha has been used in the function *holt* of the package *forecast*. This function has been chosen because the time series does not have seasonality and it is not a stationary series.

A second technique used is the function *auto.arima* of the same package, which automatically choose the best parameters to be used. The graph with the forecasting of the last two week of May 2020 for both techniques are shown in figure 8 and 9 respectively. The dark and light blue represent the 80% and 95% prediction intervals respectively.



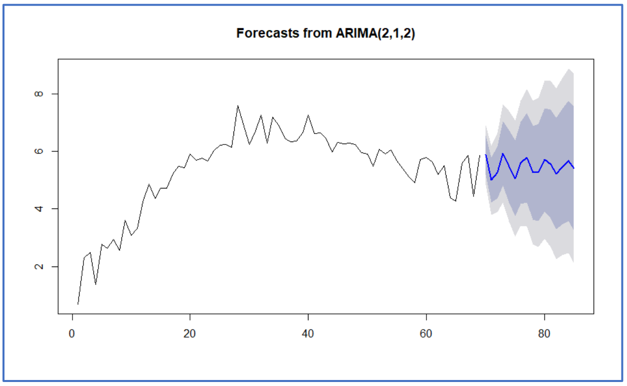


Figure 8

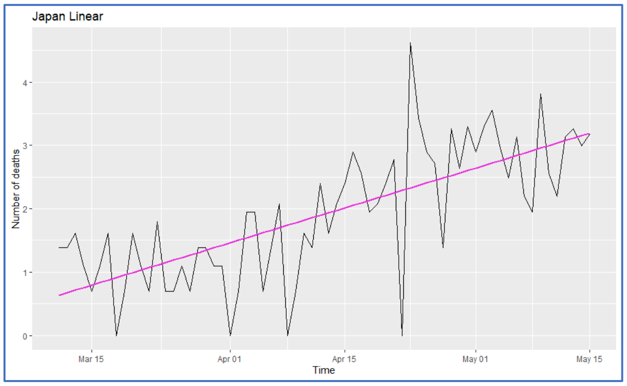
Figure 9

The root mean square prediction error has been calculated in logarithmic scale for both method and they are reported in the following table. The Exponential Smoothing seems performing better than Auto ARIMA.

Forecasting Technique	RMSPE
Exponential Smoothing	0.904
Auto ARIMA	1.225

ANALYSIS – JAPAN

The same procedure described above has been used to analyse the time series of Japan. A linear model which describes only 50% of the data, has not been taken into consideration and a non-linear model has instead been created. The two models are reported in figures 10 and 11 respectively.



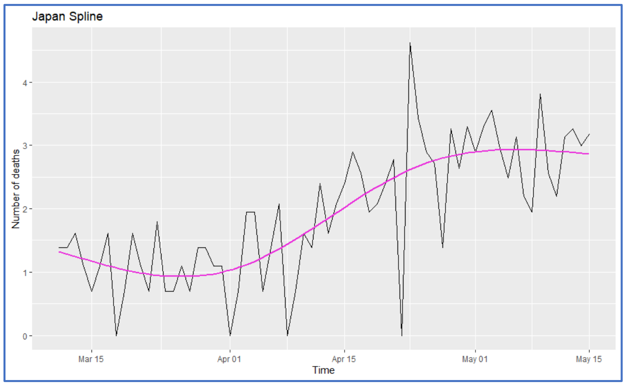


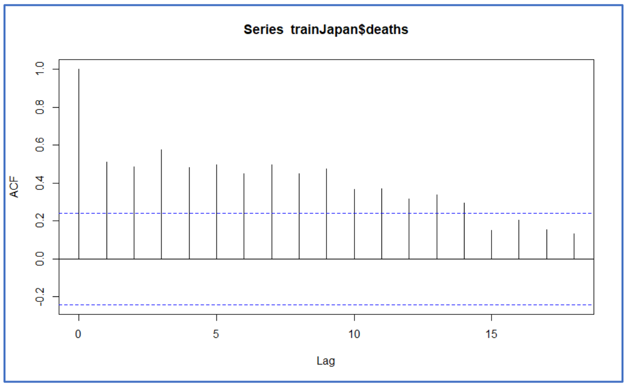
Figure 10

Figure 11

Comparing the two plots it is clear that the one created with spline describes better the whole time series and that is confirmed by the values of R squared and Adjust R squared from the summary of the two models. This time series as well, does not seem to be stationary for the same reasons explained previously. The summary of the spline model is as follow:

```
Call:
lm(formula = trainJapan$deaths ~ Japan_trend)

Residuals:
       Min         1Q     Median         3Q        Max 
-1.644e-15 -1.490e-16 -8.040e-18  1.006e-16  2.510e-15 

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)    
(Intercept)  8.746e-16  2.151e-16 4.065e+00 0.000137 ***
Japan_trend1 2.369e+00  2.548e-16 9.296e+15  < 2e-16 ***
Japan_trend2 4.990e+00  5.365e-16 9.301e+15  < 2e-16 ***
Japan_trend3 3.971e+00  3.894e-16 1.020e+16  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.878e-16 on 62 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1 
F-statistic: 1.002e+32 on 3 and 62 DF,  p-value: < 2.2e-16
```

The correlogram shows a weak trend and no seasonality. After subtracting the fitted value to the historical data, the time series looks stationary with mean equal to 0 as shown in figures 12 and 13 respectively.



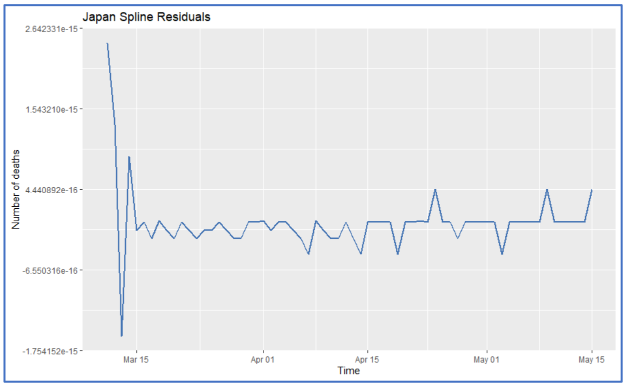


Figure 12

Figure 13

Also in this case, the acf and pacf plots suggest that any short-time correlation has been removed by the process of subtracting the spline fitted value to the original data, as shown in figure 14.

It is therefore possible to proceed to the forecasting section.

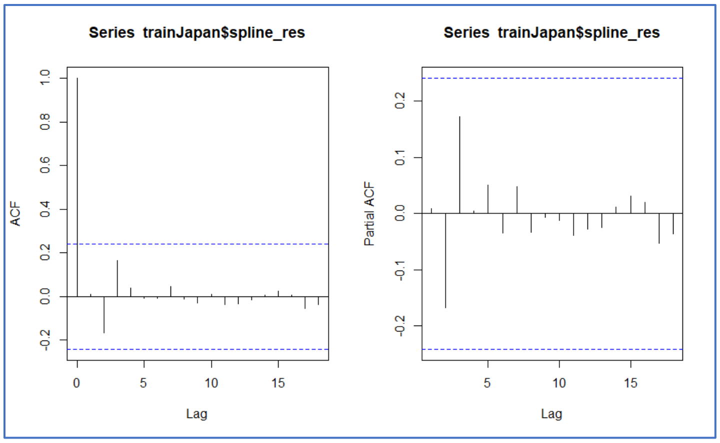


Figure 14

FORECASTING – JAPAN

In figure 14 the pink dot indicates the best value of alpha to use in the function *holt* that is equal to 0.04.

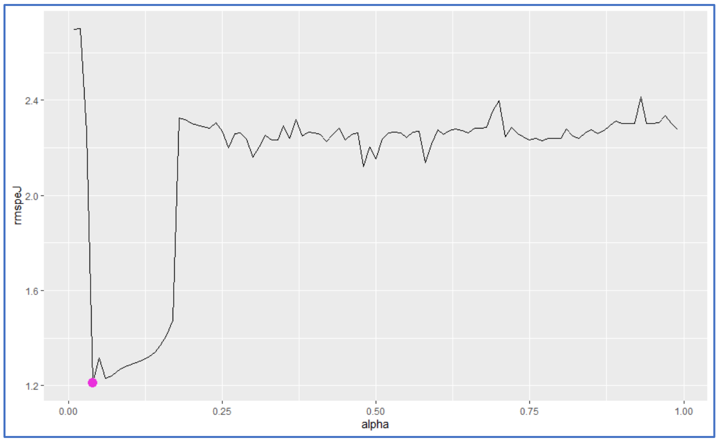
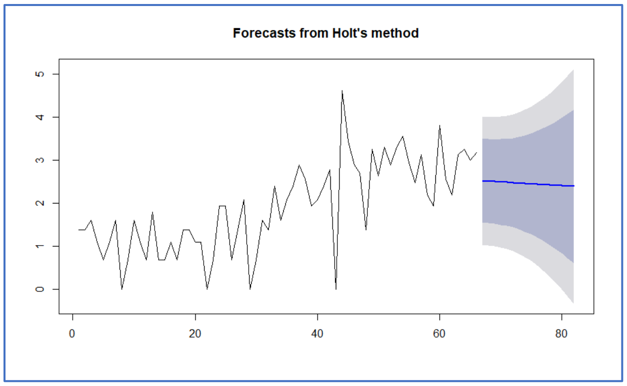


Figure 15

Figure 15 and 16 respectively show the forecasting on the number of deaths in Japan for the last two weeks of May 2020, using also in this case *holt* function for the same reason as above mentioned and *auto.arima.*



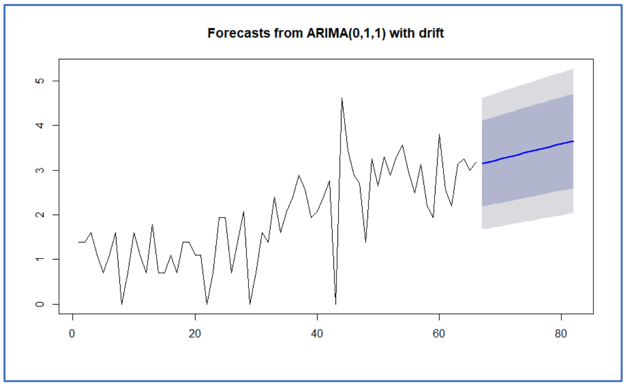


Figure 16
Figure 17

The two forecasting models above have root mean square prediction error shown in logarithm scale in the following table. Here as well, Exponential Smoothing seems to do a better job than Auto ARIMA.

Forecasting Technique	RMSPE
Exponential Smoothing	0.735
Auto ARIMA	1.352

DISCUSSION

Comparing the root mean square prediction error of the three countries analysed and reported in the following table (all in logarithm scale), it is possible to notice that of the two technique, Exponential Smoothing seems to predict better the number of deaths in France and Japan, whereas, in Philippines, the two technique seem to be equivalent with Auto ARIMA performing slightly better than Exponential Smoothing.

Country	Forecasting Technique	RMSPE
France	Exponential Smoothing	0.904
France	Auto ARIMA	1.225
Japan	Exponential Smoothing	0.735
Japan	Auto ARIMA	1.352
Philippines	Exponential Smoothing	1.081
Philippines	Auto ARIMA	1.068

It is interesting to compare the original data and the predicted one in order to have a better idea of the performance of the models created. Figure 17 shows the original trend of France and the red line separates the observations that are used as training data, on the left, and those used as test set on the right. The forecasted models are showed in figure 18 and 19

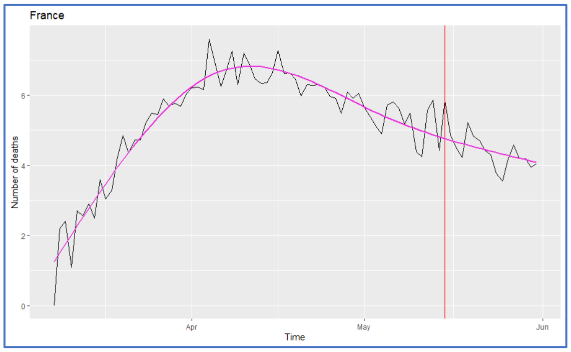
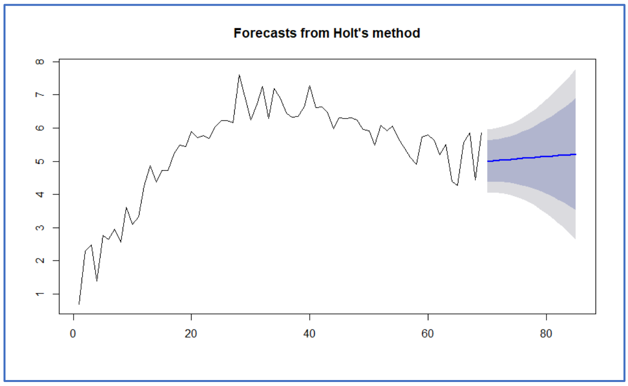


Figure 18

It is possible to notice that the original data have a distinguishable negative slope, instead both the predicted models have null slope (figure 20, ARIMA (2, 1, 2) model) or slightly positive slope (figure 19, Exponential Smoothing).



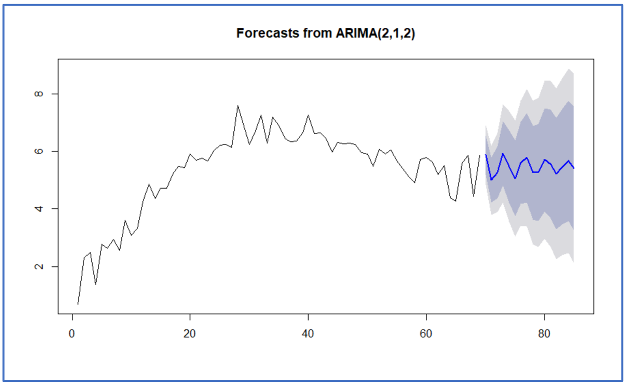


Figure 19

Figure 20

The original data of Japan as well, after the 15th of may have a negative trend (figure 21) while the two predicted models have slightly negative slope for Exponential Smoothing (figure 22) and positive for ARIMA (0, 1, 1) (figure 23)

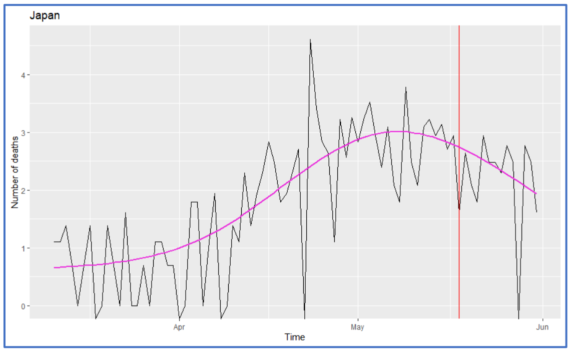
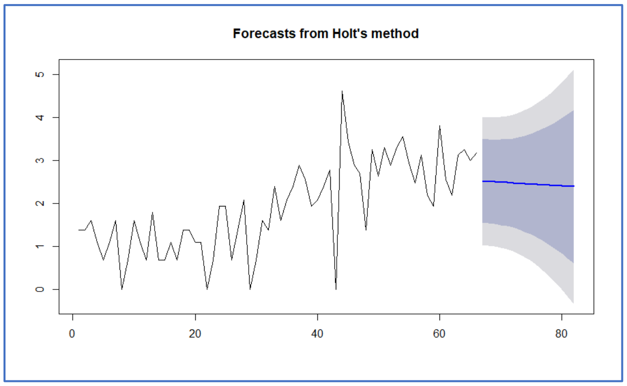


Figure 21



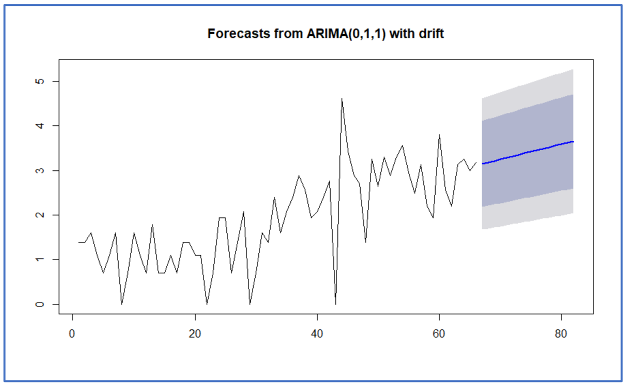


Figure 22

Figure 23

FUTURE IMPROVEMENT

As seen in the previous paragraph the models described in this report perform well but they can obviously be improved. There are some points that could be better developed in the whole process, starting with the modality of collecting data. For example, adding information about any delay in reporting casualties, or adding any other information about the performance of different hospitals in the country and thus incorporate other techniques like mixed models.

Other techniques of forecasting, such as regression or ARMA(p,q) models, can be used and then compared against each other in order to find the best model that better forecast the future number of deaths.

It would be also interesting to mix with other disciplines like Machine Learning and maybe develop a forecasting model using neural network.

Furthermore, more predictors could be added to the models in order to obtain more reliable forecasts.