Prediction of number of daily deaths from covid-19 in Luxembourg, Egypt and Serbia

Preparer: Alexey Pankratov | June 2020

Introduction

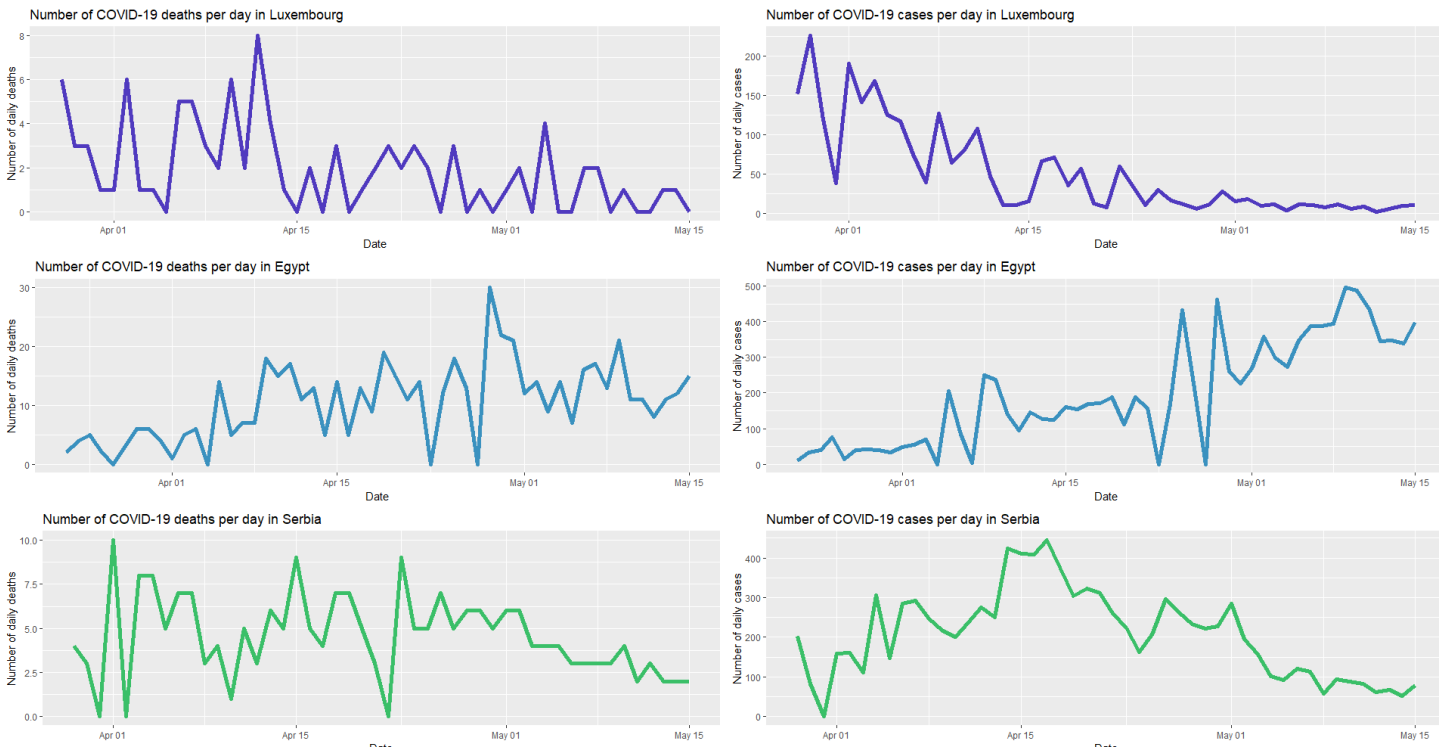
The aim of the assignment is to build the model that forecasts number of daily confirmed deaths in Luxembourg, Egypt and Serbia (separately for each of the beforementioned countries). The models will be built on the dataset that includes the following variables:

- Date (same information is included in several variables that duplicate each other). Beginning of records is different for each country but the earliest data available is from March 23rd. The records are available until May 31str.
- Number of daily deaths from covid-19
- Number of daily new cases of covid-19
- Country (same information is included in several variables that duplicate each other)
- Population of the country

In this report I will explore the data and determine which models to build based on characteristics of the time series (stationarity). Then I will build the models, compare them and conclude on the most appropriate model for each country.

Exploratory analysis

First, I will split the data into training (before May 15th, including) and test data (after May 15th). I will perform exploratory analysis of the training data only as though the data from May 15th had not existed at the date of the analysis.

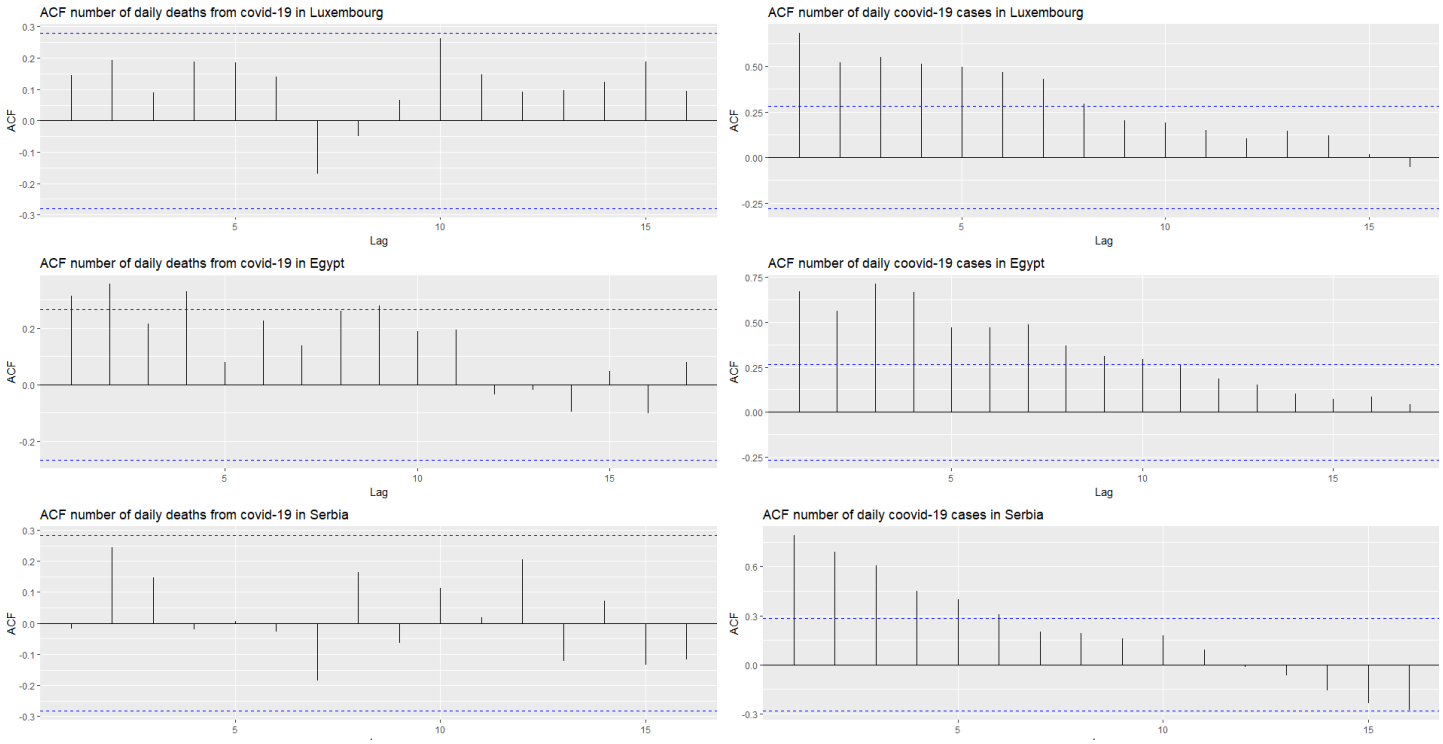


The data in each of the countries does not show seasonality. It is expected as the period of recording of the data is too short. Regarding the trend, there is a strong evidence of the trend in the number of daily cases (plots on the right).

In Egypt, the number of cases is on the rise, whereas the trend is opposite for the number of daily covid-19 cases in Luxembourg. In Serbia the daily cases increase up until mid-April and then starting to decrease after.

At the same time, although plots of daily covid-19 death on the left slightly resemble those on the right, the trend there is not clearly visible. In order to check if the number of deaths plot includes trend or not, I will use as autocorrelation function to see if this is the case.

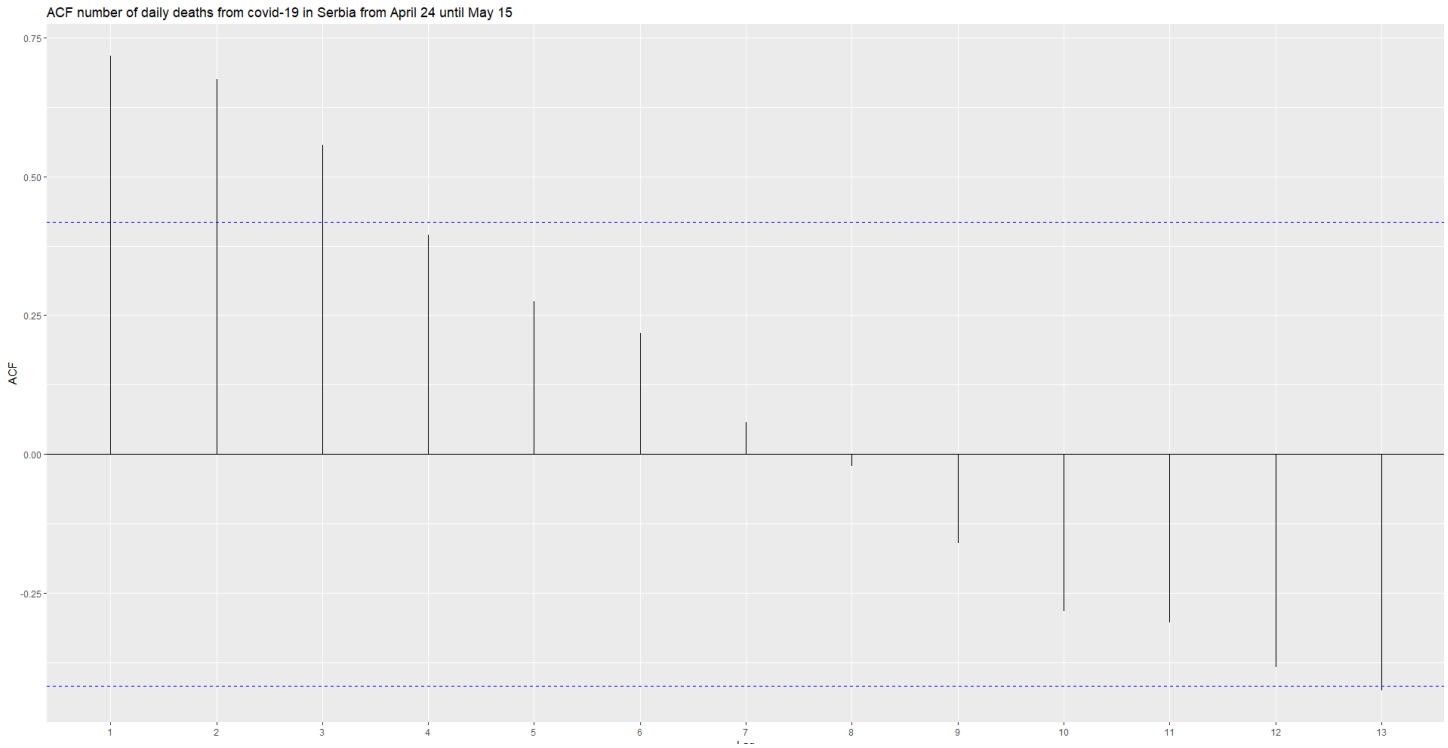
ACF plot is more robust technique of trend identification than visual observation of the daily deaths and cases plots. **It should be noted that autoplot function for plotting correlogram below omits the first lag that always equals to 1.**



Number of daily deaths appear to be stationary and show no trend and seasonality in Luxembourg and Serbia as no lags exceed the confidence interval of the ACF (blue dashed line). It looks like that the number of deaths in Egypt show patterns of trend as lags on the ACF plot decrease gradually although the trend is not as strong as the trend for the number of covid-19 daily cases. Looking at the plot of number death over the training data period in Egypt, it does not show any inflection points or any other sign that the trend is not linear, so I will use linear trend in the model.

In comparison, number of daily cases in almost all lags exceed the confidence and the lags subside gradually, which indicates a presence of strong trend in the number of daily cases. Qualitatively, I can see and conclude from the ACFs that the number of daily deaths are stationary for Luxembourg and Serbia, considering full period of training data (see below additional considerations for Serbia deaths in starting from April 24th) while deaths in Egypt and daily cases in all countries are not stationary since lags exceed the confidence interval and decrease gradually, which is an indication of existence of trend.

At the same time although correlogram of Serbia time series shows that the daily deaths are stationary if to consider full period of training data, the general plot shows trend downwards trend in daily deaths in Serbia starting from approximately April 24th. Looking at the ACF plot of the data from this date it is clear that there strong trend exists starting from this date. Considering this, I will build exponential smoothing model that would incorporate additive trend and use the smoothing factor alpha closer to 1 in order to give more value to recent observations.



I will finalize exploration of data by checking correlation between number of daily cases and daily death. It is not expected to be strong for Luxembourg and Serbia as I noted above that number of daily deaths is stationary time series whereas number of daily cases present strong trend, but it may be worth it to have a closer look at the figures.

The correlation is not strong as it was expected. It is stronger (0.676) in Egypt where the number of daily cases increases and where I saw trend pattern, and getting lower in Serbia (0.464) where the number of daily covid-19 infections started to subside recently and is even lower in Luxembourg (0.308) where the decreasing trend was present from the beginning of the period under analysis. However, the correlation is not strong enough for number of daily cases to be the main predictor of the number of daily deaths from covid-19.

Time Series Modelling: Description, Forecasting and Assessment of Forecast(s)

After exploring the data, I will perform time series modelling for covid-19 deaths in each country. Considering that the number of daily deaths appears stationary for Luxembourg and Serbia (looking at the whole period of testing data) I can build ARIMA model. I will also build exponential smoothing models without trend and seasonality for Luxembourg and with trend for Serbia (under assumption that trend started in April 24th will continue) and Egypt. And for comparison reason I will also build exponential smoothing models using ets() function of forecast package in R without specification of seasonality and trend so that the functions would build the model it considers the most optimal and compare it with our selected model in case they are different from the one I selected manually.

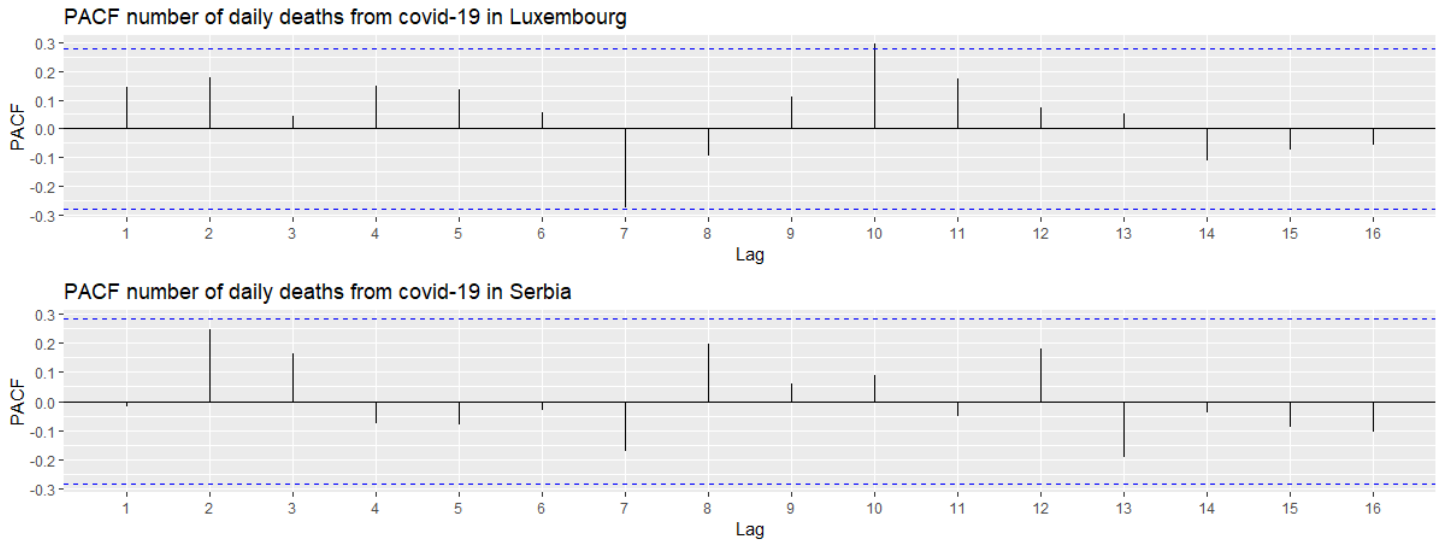
After I compare forecasting accuracy of three models by plotting it and calculating root mean square error.

ARIMA MODELS
I will first select the best ARIMA model for each country based on smallest AIC using the auto.arima() function from library(forecast).

Using auto.arima() I received the following optimal models for each country:

- Luxembourg – ARIMA(0,0,0)
- Serbia ARIMA(0,0,0)

The data shows no short term correlation for Serbia and Luxembourg so it may be reasonable that the best fit ARIMA model is ARIMA(0,0,0) for these countries. I have already looked at ACF plots that support MA(0) model, now I will look at partial ACF plots to double check if AR(0) is appropriate as well.



No Lags are above confidence interval in Serbia and Luxembourg, therefore AR(0) looks appropriate. In other words I can say that ARIMA does not identify any short term correlation in the time series data.

EXPONENTIAL MODELS

Is it was mentioned above, I will prepare exponential smoothing models without trend and seasonality for Luxembourg and with trend for Serbia and Egypt, where for Serbia the model with trend will be based on the assumption that the trend started in April 24th is expected to continue after.

Exponential smoothing models were built using R function ets()

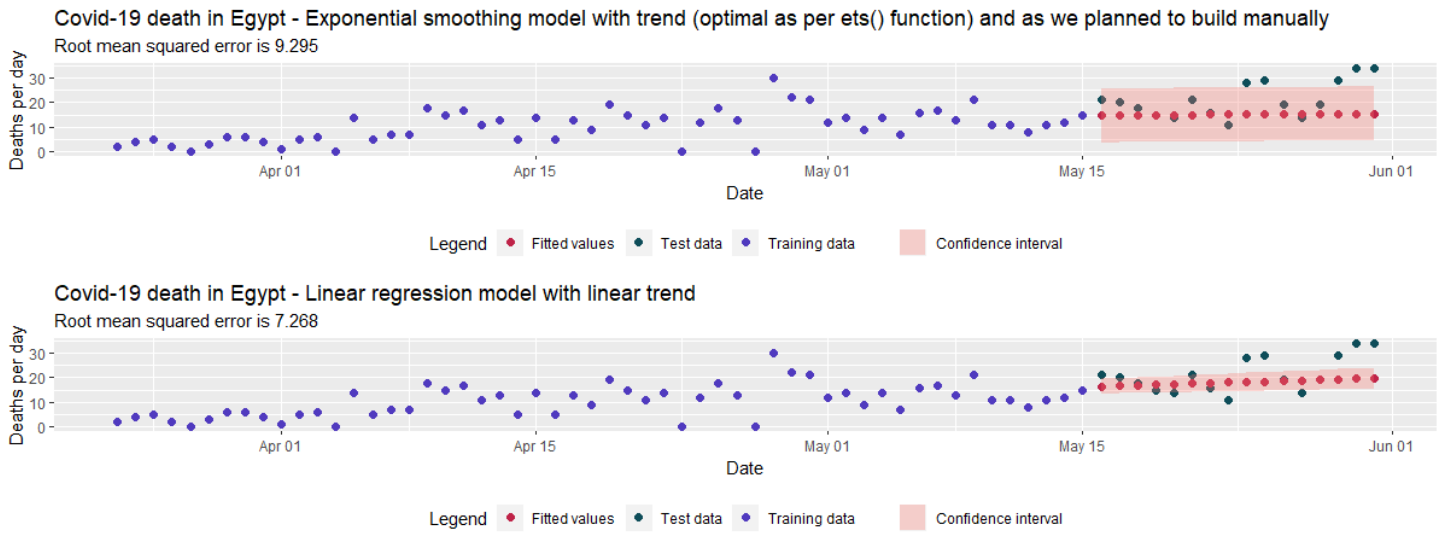
In order to build the exponential model with additive trend for Serbia I need to estimate the most appropriate smoothing parameter alpha that a will put more weight on recent observations. To do that, I performed iterative construction of the models with different parameter alpha (from 0.1 to 0.99 with step of 0.01) and estimated root mean squared error for each model. It should be noted that RMSE was calculated for observation from April 24th the end of training data, which is May 15th as the assumption is related to the trend that started only on April 24th. It turned out that the best RMSE is when the alpha parameter is 0.49.

And also I have not specified model in R making the function ets() to prepare optimal model based on its own criteria and compare with our model provided the optimal model will be different from assumed by us at exploratory analysis stage. It turned out that optimal models in accordance with ets() are models different from those selected by us for Serbia and Luxembourg. For Luxembourg, the optimal model is with additive trend whereas I manually selected without trend and for Serbia it is vice versa. For Egypr optimal model in accordance with ets() function is with additive trend, which is the same as I selected manually.

LUXEMBOURG. Start building form Luxembourg

As it can be seen from the plots above exponential smoothing without trend and seasonality (i.e. the assumption that the time series are stationary that I considered to be true based on analysis of ACF plots) provides model with the leased root mean squared error (0.948). The exponential smoothing model with trend that was considered optimal by ets() function in R provides worse prediction. However, the model with the least accurate prediction is ARIMA model with assumption that there is no short-term correlation in the variable.

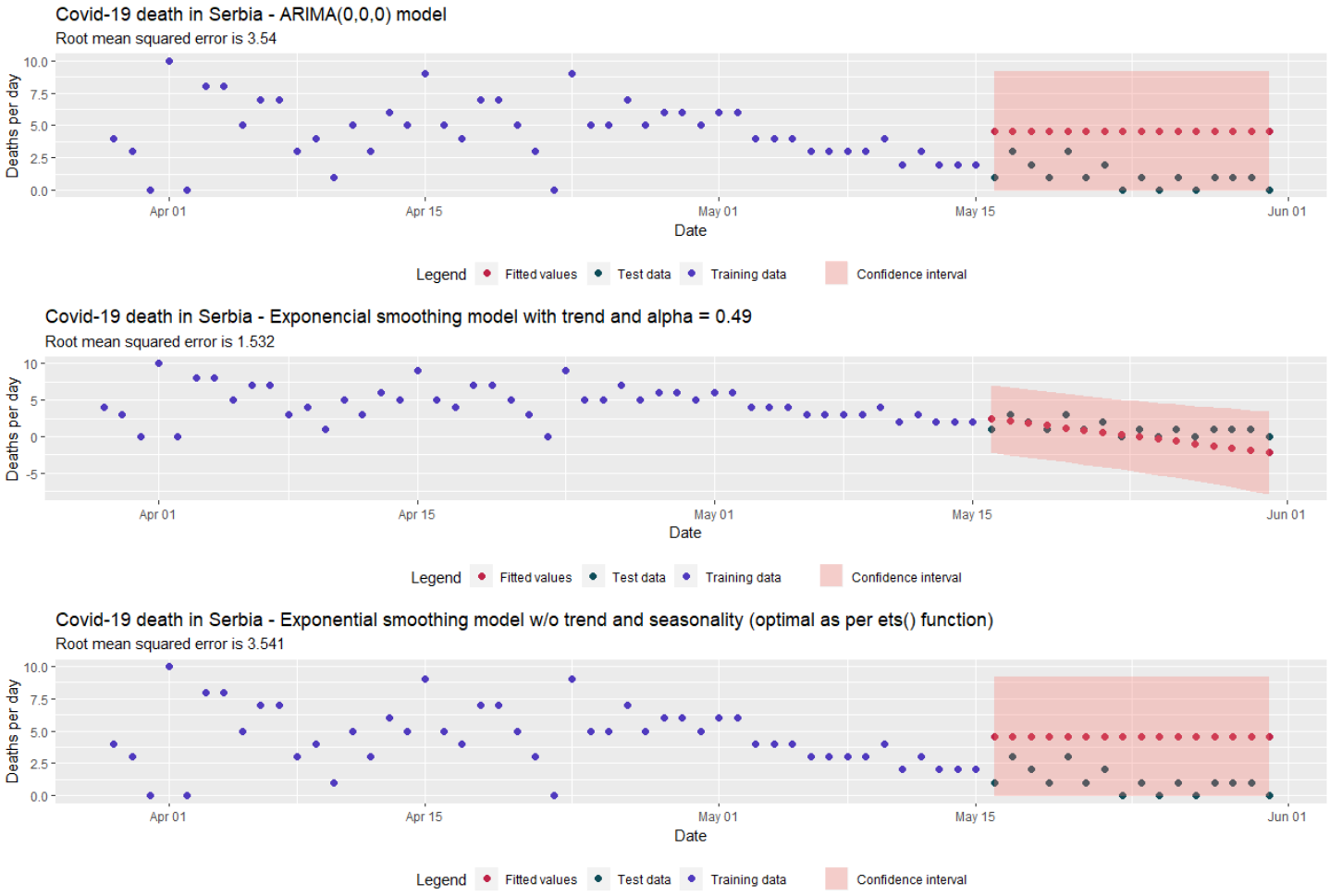
EGYPT. Now let's look at Egypt forecast



Looking at the test data, the trend is clearly visible as I correctly identified during the exploratory analysis. So, the models without consideration of the trend would have shown worse result then the ones I have selected. At the same time, I can see that linear trend model predicts more steep increase in deaths number over time then exponential smoothing model. Linear model more accurately captured the actual evolution of covid-19 situation in Egypt (in the

short term) and therefore is more preferable to the exponential smoothing model. Lower RMSE value in linear trend model (7.268) compared to exponential smoothing (9.295) supports this conclusion.

SERBIA. And finally, for Serbia the plot is as follows:



With regards to Serbia, both ARIMA model and exponential smoothing model without trend and seasonality provide much poorer fit than the exponential model with our assumption that the trend started from end of April will continue to be there after May 15th. RMSE of the model built on this assumption (1.532 for the exponential smoothing with trend) is more than twice as low as for two other models (3.54 for both ARIMA and exponential smoothing w/o trend and seasonality).

As a general comment I can tell in Luxembourg and Serbia exponential smoothing models manually selected by us on the assumption of stationarity and trend starting from April 24th respectively were better than optimal model selected by function ets(). Considering these results. it is important to also visually explore the data as some trends may be spotted there that may not be identified mathematically.

And as a final word I would like to discuss limitations of the model what can be done in terms of enhancing of goodness of fit of the model.

1. Very short period for analysis. Analysis of the short period from March 23rd to May 15th showed that the data is very volatile if to look at the short period, which makes the models build on this data less accurate. In case the data was available for more than a year and even for several years it would have been more likely that the data showed clearer trend and seasonality patterns.
2. Data collection. Large amounts of data on covid-19 are collected manually and prone to errors. The data is collected by Health agency in every country, where method of its collection as well as criteria for considering that a death is from covid-19 or distinction between active cases and closed cases may differ from country to country, which makes the information less reliable. So the data collection and registration needs to be performed in a standardised way that is consistent in every country.

3. Reliability of data. Some countries may understate figures intentionally in order not to spread panic among population or for any other reasons. Or some countries with less well-established health care systems may fail to properly record covid-19 cases and deaths. The data thus needs to be more closely inspected for signs of potential errors, whether intentional or not.

Thank you for your attention!