Predicting The Number of Deaths from COVID-19

The aim of this project is to explore, model and compare the temporal pattern in the number of deaths from the coronavirus COVID-19 in Netherlands and Peru. Once the appropriate models are developed and trained, they will be used to forecast the number of COVID-19 associated deaths from the 15th of May until the 31st of May in both countries, using only information that was available prior to that date.

Over 14 million infections have been confirmed worldwide from the coronavirus COVID-19, and almost 600,000 deaths have been associated with it.

Data

The data set provided contains information on three countries: Netherlands, Peru and Romania. For the purpose of this project, only the observations corresponding to Netherlands and Peru have been used.

The data set contains the following variables:

- dateRep: Date of report
- day: a numeric variable for the day of the month of the report
- month: a numeric variable for the month of the report
- year: a numeric variable for the year of the report
- cases: number of confirmed COVID-19 cases on the date of report
- deaths: number of confirmed COVID-19 deaths on the date of report
- countriesAndTerritories: the country's name
- geoID: the geographic reference ID for the country
- countryterritoryCode: the country's three-letter code
- popData2018: the country's population in year 2018
- continentExp: the continent to which the country belongs
- Cumulative_number_for_14_days_of_COVID.19_cases_per_100000: the cumulative number of cases for 14 days per 100000

The initial data set has been split into the following sub sets:

Peru

peru.data - the data set contains 65 observations for Peru.

The peru.data data set has been further split into a training data set and test data set in the following way:

peru.training - contains 49 observations, recorded from 28th of March 2020 until (and including) the 15th of May 2020. This data set has been used for training the model.
peru.test - contains 16 observations, recorded from the 16th of May 2020 until the 31st of May 2020. This data set has been used for testing the prediction performance of the model.

The data sets contain the same variables as the initial data.set. One variable has been added to both nl.data and peru.data data sets:

day.no: an integer variable for the day of the report, counting the days since the reporting started.

Netherlands

nl.data - the data set contains 78 observations for Netherlands.

The nl.data data set has in turn been further split into a training set and a test set as follows:

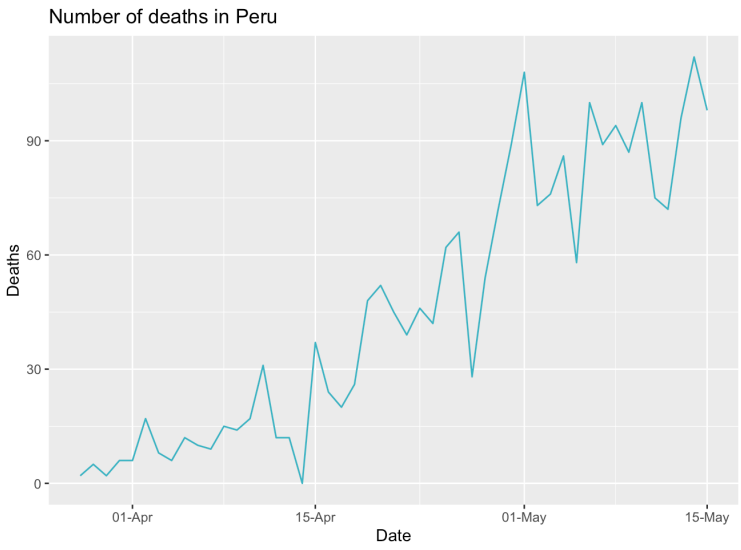
nl.training - contains 62 observations. The dates of the observations range from 15th of March 2020, up to and including the 15th of May 2020 This data set has been used for training the model.
nl.test - contains 16 observations, recorded from the 16th of May 2020 until the 31st of May 2020. This data set has been used as a test set.

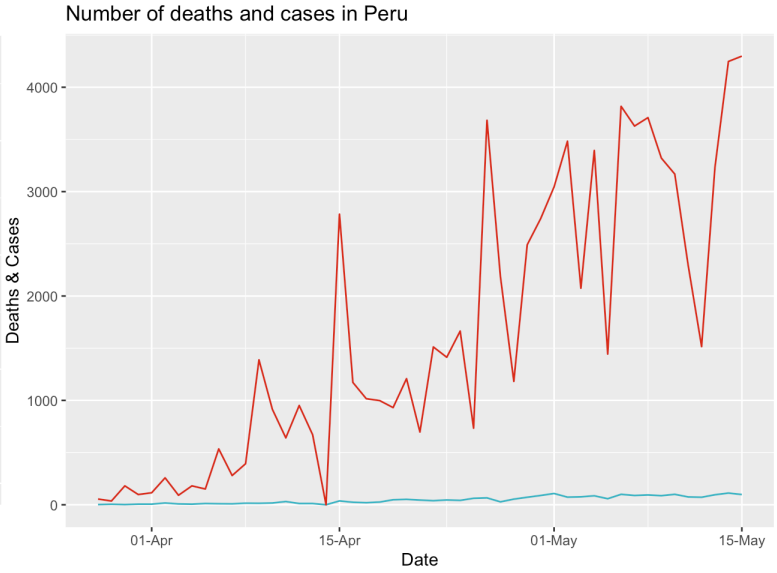
Exploratory Analysis

Peru

The first time plot below (in blue) reveals a steep increase in the number of deaths over time, suggesting the presence of a trend in the data.The relationship between the number of deaths and day appears linear. We also notice that the variance increases as the time increases, and then it slightly decreases again.

The second time plot illustrates the number of cases (blue) and deaths (red) over time and relative to each other. There seems to be more variability present in the number of cases. Also we can see that relative to the number of cases being reported, the number of deaths is low. At this point, both the number of cases and deaths are still increasing, indicating the infections have not reached a peak as of yet.





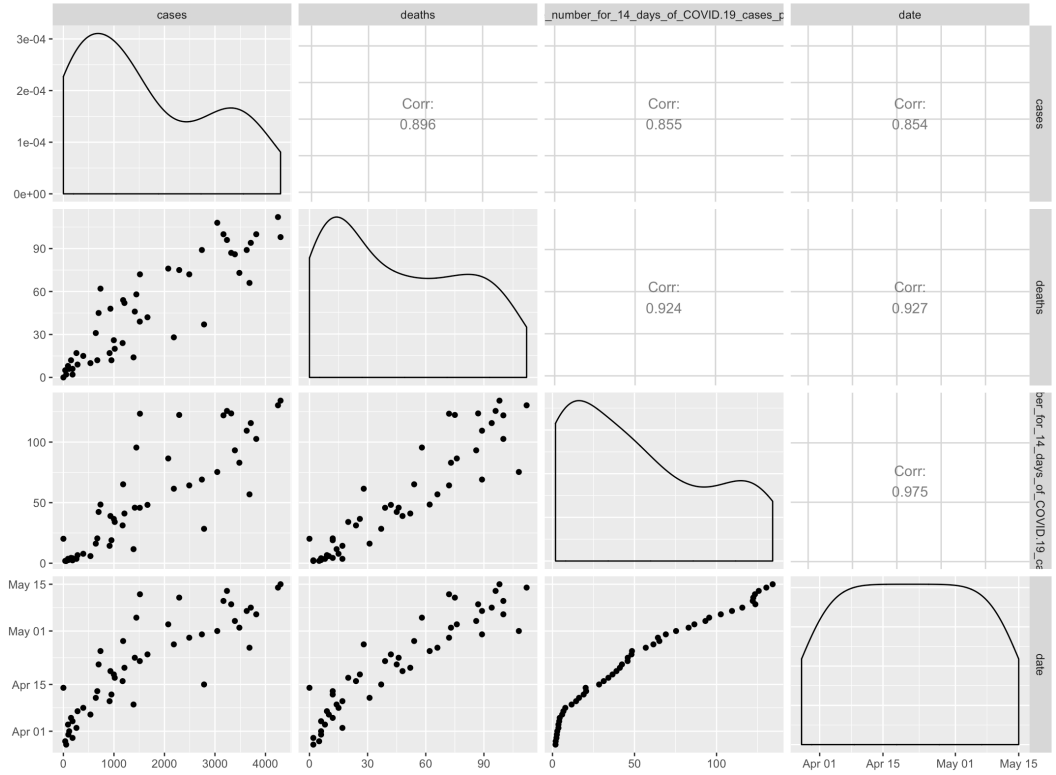
The summary for the relevant variables in the peru.training data set can be viewed below.

cases	deaths	Cumulative_number_for_14_days_of_COVID.19_cases_per_100000	day.no
Min. : 0	Min. : 0.00	Min. : 1.836	Min. : 1
1st Qu.: 535	1st Qu.: 12.00	1st Qu.: 11.639	1st Qu.:13
Median :1208	Median : 42.00	Median : 42.337	Median :25
Mean :1633	Mean : 46.08	Mean : 52.050	Mean :25
3rd Qu.:2784	3rd Qu.: 75.00	3rd Qu.: 86.477	3rd Qu.:37
Max. :4298	Max. :112.00	Max. :134.197	Max. :49

The observations for Peru were recorded over 49 days. The total number of cases up to that point was over 80,000 with nearly 2,300 dead. The cumulative number of cases over 14 days per 100000 people, is 112,000 at it's maximum.

Below we have the variables plotted against each other along with their correlations.
We can expect the relationship between the number of cases and the number of deaths to be a linear one, as the number of deaths depends on the number of cases, as illustrated by the plots below. The relationship between deaths and the cumulative number of cases over 14 days, per 100,000 people, is also linear. For the plot displaying the number of cases over 14 days per

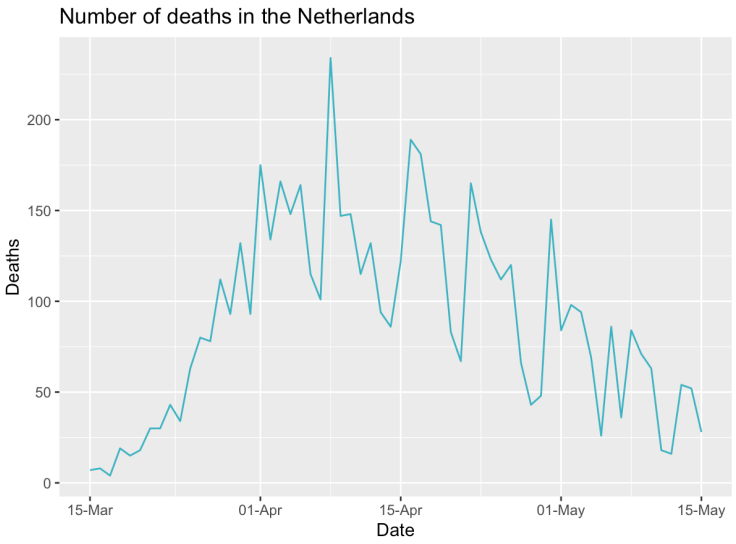
100,000 people, represented in time, we notice the points lie very close to straight line. Also, the correlation between all variables is extremely high - between cases and deaths being 0.9 and cumulative cases and deaths being 0.92.

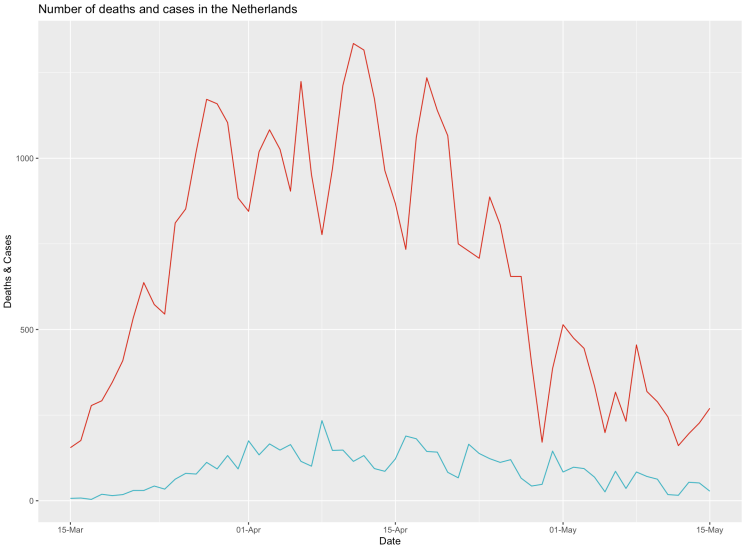


Netherlands

The first time plot below (in blue) displays how the number of deaths changes in time. There seems to be a non-linear trend present - the trend looks like a curve, at first the number of deaths increases, only to start decreasing again around the 15th of April. The variance is not constant and seems to be fluctuating as time increases.

The second time plot illustrates the number of cases (blue) and deaths (red) over time and relative to each other. The curve for cases has a similar shape to the one in our first plot, and the variability seems to fluctuate significantly.



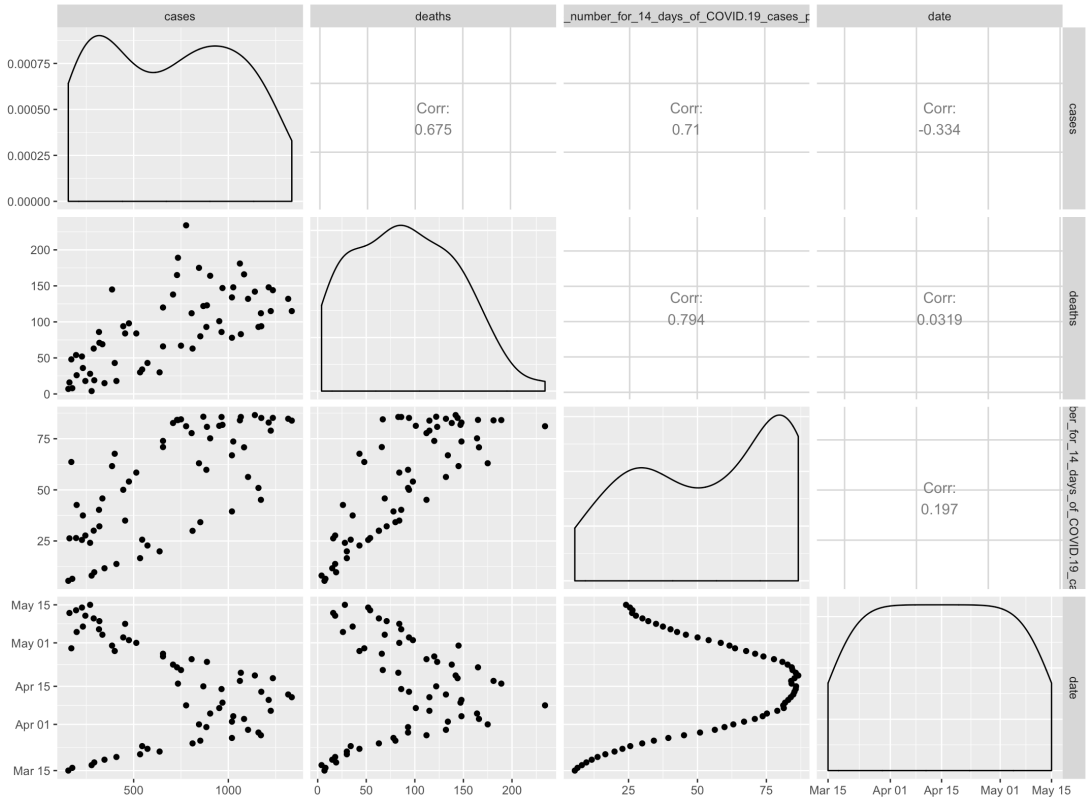


The summary for the relevant variables in the nl.training data set can be viewed below.

```
     cases           deaths       Cumulative_number_for_14_days_of_COVID.19_cases_per_100000     day.no      
 Min.   : 155.0   Min.   :  4.00   Min.   : 5.509                                             Min.   : 1.00  
 1st Qu.: 337.8   1st Qu.: 44.25   1st Qu.:29.982                                             1st Qu.:16.25  
 Median : 718.5   Median : 86.00   Median :59.133                                             Median :31.50  
 Mean   : 688.3   Mean   : 90.08   Mean   :54.291                                             Mean   :31.50  
 3rd Qu.:1006.5   3rd Qu.:132.00   3rd Qu.:81.272                                             3rd Qu.:46.75  
 Max.   :1335.0   Max.   :234.00   Max.   :86.575                                             Max.   :62.00  
```

We can see that the observations were recorded over 62 days.Netherlands has a significantly lower number of cases compared to Peru (nearly half), and also, the cumulative number of cases over 14 days per 100,000 people is lower too, indicating a lower infection rate relative to the size of the population. However, the ratio between deaths and cases in higher, as evident from the second time plot and summary statistics. Furthermore, the time plots suggest that the peak of the infection has passed, compared to Peru where the cases are still rising and there is no evidence of the infection rates slowing down.

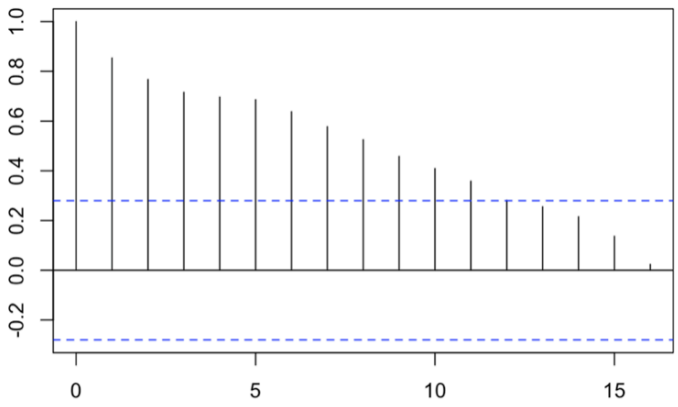
Below we have the variables plotted against each other, along with the correlations between them.
The plots suggest that, as expected, the relationship between the number of cases and number of deaths is linear - since the number of deaths very likely depends on the number of cases. The correlation between the two variables is high - 0.8 for deaths vs cumulative cases and almost 0.7 for deaths vs cases.
In the plot that displays the number of cumulative cases in time, we can see that points lie very closely to a curve. This is also the trend we had noticed when looking at the initial time plots.



Time Series Modelling: Description

Peru

In order to be able to model the data for Peru, we begin by looking at the correlogram for the number of deaths in the training data set. The ACF plot is therefore displayed below:



The ACF plot confirms the previous suspicion - a trend is indeed present in the data, which makes it impossible to tell if there is any short term correlation present. Therefore, we must first remove the trend before modelling the data further. Since the increase in the number of deaths appeared linear in the previous plots we have examined, I will use linear regression to model the trend present in the data.

The following model has been fit:

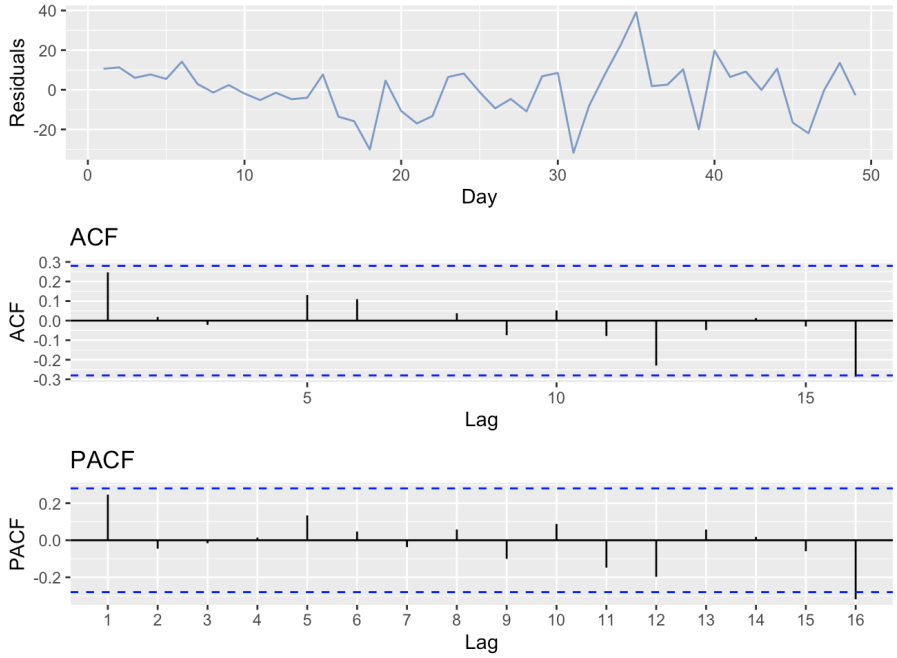
deaths_t = beta0 + beta1day.no_t + e_t

The summary of the model can be viewed below:

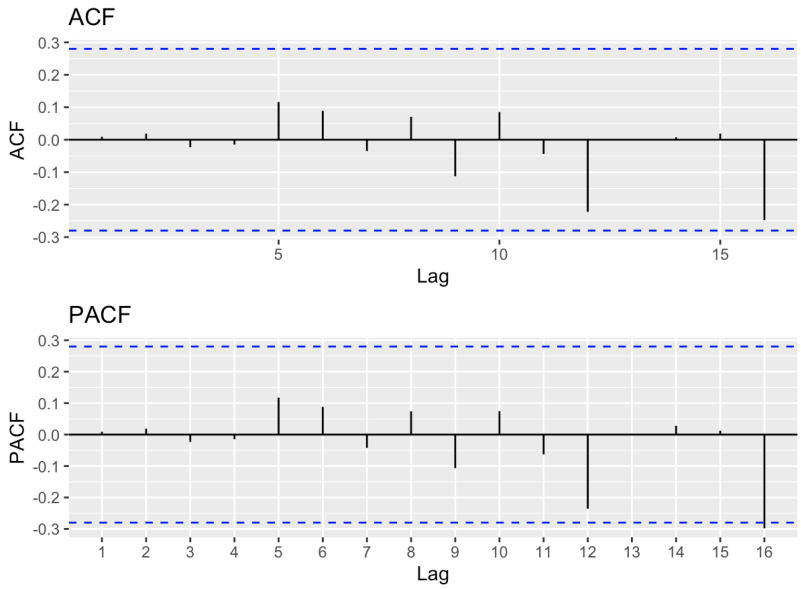
```
Call:
lm(formula = peru.training$deaths ~ peru.training$day.no)

Residuals:
    Min      1Q  Median      3Q     Max
-31.748  -8.025   1.864   7.749  39.142

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -10.8597     3.8729  -2.804  0.00731 **
peru.training$day.no   2.2777     0.1348  16.892  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.35 on 47 degrees of freedom
Multiple R-squared:  0.8586,    Adjusted R-squared:  0.8556
F-statistic: 285.3 on 1 and 47 DF,  p-value: < 2.2e-16
```

From the summary, the linear regression model describes the trend adequately. The coefficient for day number is positive, suggesting that with every day that passes there is a 2.2777 increase in the number of deaths on average. The adjusted R-squared value is 0.86, which means the model explained 86% of the variability in the data.

Next, the fitted values of the linear regression model were subtracted from the data, and the residuals were plotted. We can see below that we have heteroscedasticity present in the residuals, however the series appears stationary, possibly with The ACF and partial ACF plots exhibit very little short term correlation at lag 15 and 16, as displayed below.

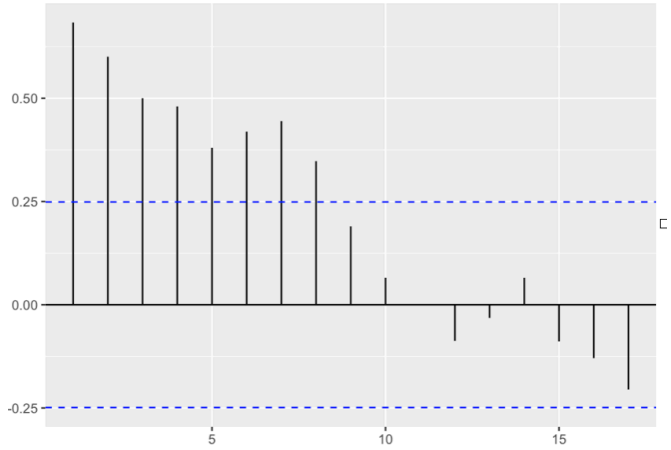


However, when I ran the auto.arima function, it suggested a MA(1) process. Therefore, I did fit a MA(1) series to the residuals and analysed the ACF and PACF plots, as displayed below. The plots have improved slightly, and they resemble a purely random process as they don't show evidence of correlation at any lags.



<u>Netherlands</u>

Before we begin modelling the data for Netherlands, we first check the correlogram of the number of deaths in the nl.training data set.



The plot displays significant correlation at the first eight lags, possibly indicating the presence of a trend. Therefore, I will remove the trend before checking the correlogram once again, in order to see if there is any short-term correlation exhibited.
Since the relationship between deaths and time doesn't appear linear, I will attempt to capture the trend in the data with a linear regression spline. In order to fit the spline, I have identified two points in time, which I will be using as knots, where the rate of change in the graph changes. For example, in the first 15 days there is a steep increase in the number of deaths, and afterwards the rate at which the number increases slows down for the next 15 days, almost plateauing. At around day 30, the number of deaths begins to decrease. Therefore, I will be using day 15 and day 30 as the knots of the spline.
In order to fit the regression spline, I have used the function tslm() from the library *forecast.* In order to handle the heteroscedasticity in the data, I have set *lambda = 0.*

The following model for the trend was fit to the data:

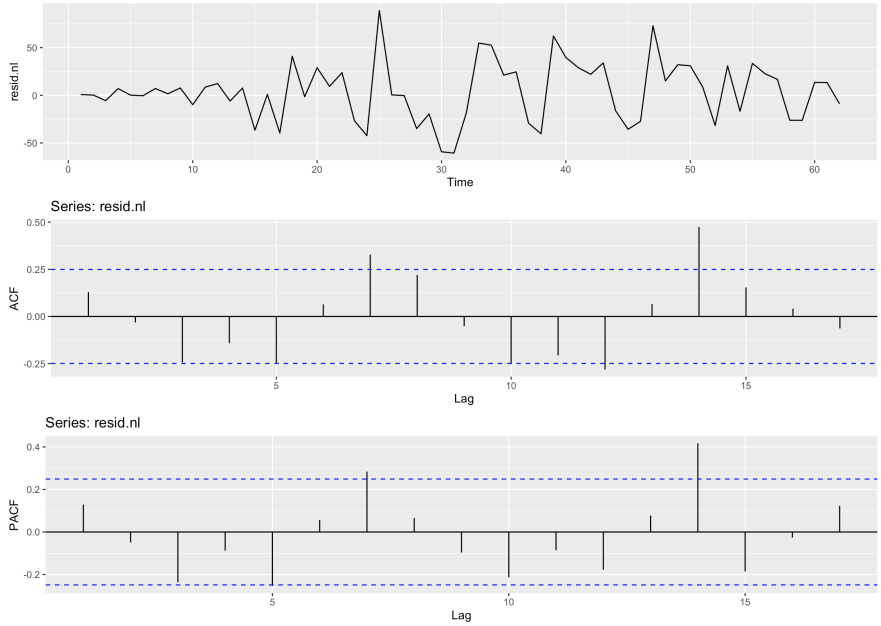
log(deaths_t) = beta0 + beta1day.no_t + beta2day.no_(t – 15)+ + beta3day.no_(t – 30)+ + e_t

The summary can be viewed below:

```
Call:
tslm(formula = deaths.ts ~ t1 + tb1 + tb2, lambda = 0)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97108 -0.26255  0.06791  0.28500  0.69974

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.61294    0.19803   8.145 3.50e-11 ***
t1           0.21679    0.01864  11.630  < 2e-16 ***
tb1         -0.20561    0.02862  -7.185 1.44e-09 ***
tb2         -0.05563    0.01742  -3.193  0.00228 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3892 on 58 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 2.879e+05 on 3 and 58 DF,  p-value: < 2.2e-16
```

The adjusted R-squared value is 0.99, indicating a very good fit to the data.

Once the trend has been removed from the time series, we plot the residuals along with the ACF and PACF.
We can see that the time series appears stationary now, but there is still a problem with heteroscedasticity. There is some significant autocorrelation at lag 14 and at lag 7 on both the ACF and PACF plots. Using the auto.arima function, it has indicated that a MA(1) process would be suitable to deal with the autocorrelation. Therefore, I have used a MA(1) process to model the series, and produced the residuals.



I have once again plotted the time series residuals against time, as well as the ACF and PACF. Unfortunately all three plots have remained unchanged, and they look identical to the plots provided above.

Comparison

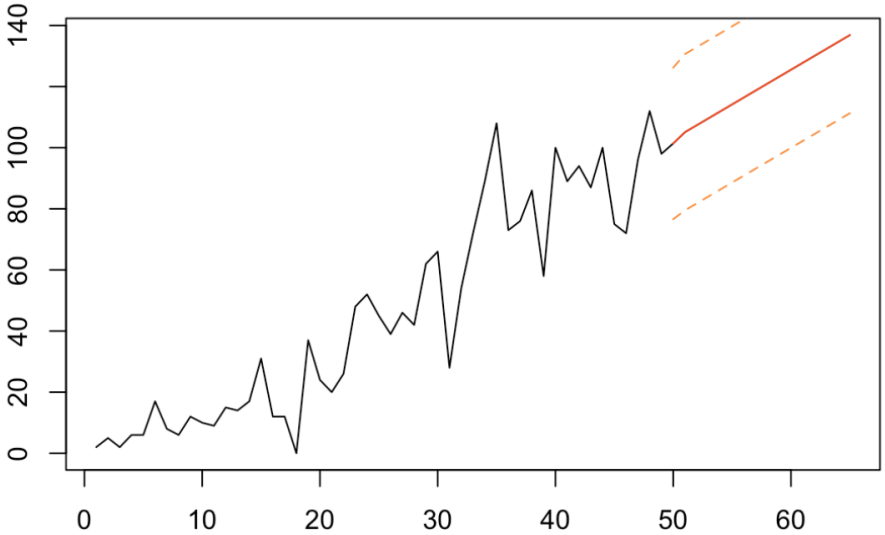
The two countries, Peru and Netherlands, are displaying different trends in the number of deaths over time - Netherlands seems to be ahead of Peru in what concerns the cycle of infections, therefore at this point Peru is only displaying a steep increase in the cases and deaths, whereas the cases and deaths in Netherlands are starting to tail off and therefore exhibiting a curve. As a result, I had to employ two slightly different approaches when describing and modelling the trends present - for Peru I was able to fit a linear regression model which described the data reasonably well, whereas for Netherlands I used a linear regression spline with two knots, and a log transformation in order to handle the heteroscedasticity.

In both cases the residual plots displayed heteroscedasticity. For Peru, I have successfully modelled the short-term correlation using a MA(1) process. In the case of Netherlands, some short-term correlation remained at lag 14, even after using a MA(1) process to model it.

Time Series Modelling: Forecasting and Assessment of Forecasts

Peru

In order to produce the 16 steps ahead forecasts for Peru, I have used a MA(1) series process with a linear trend. The plot below illustrates the predictions along with 95% confidence interval bands. We can see that the confidence bands are quite wide, indicating the uncertainty in prediction.
Comparing the predicted values to the observed values, we can tell than the model consistently under predicts - the predicted numbers are lower than the observed values.

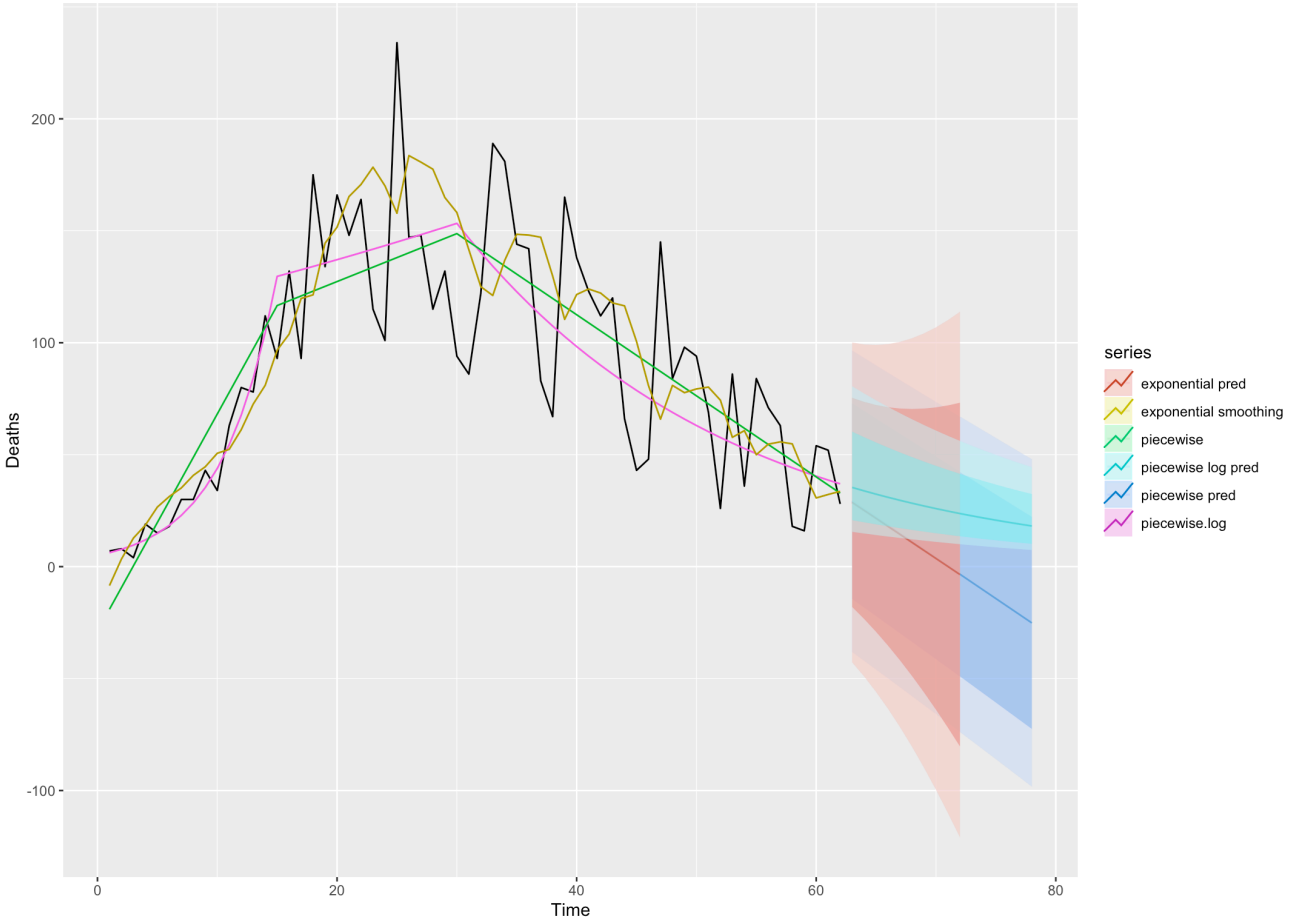


I have also used linear regression to predict the number of deaths, the plot can be viewed in the R file attached. This model also seems to predict lower values than the ones we observed and it has wide confidence bands.

Netherlands

In order to predict the number of deaths in the Netherlands, I have used the log transformed piecewise regression spline I had initially used to model the trend in the data, a non-log transformed piecewise regression spline with the same two knots at day 15 and 30, and Holt-Winters exponential smoothing with additive error and trend present, but no seasonality.

The below plot illustrates the fit of splines and exponential smoothing, along with the predicted lines and the 95% and 80% confidence bands.
We can see in the plot that the confidence bands for exponential smoothing are the widest, followed by the ones for the non-log transformed piecewise spline. The predictions are also rather similar for exponential smoothing and the non-log transformed spline. On the other hand, in the case of the log transformed spline, the predictions are quite different, and the line has a less steep slope. Also, the confidence bands are more narrow, indicating increased prediction accuracy.



Comparison

In this chapter we will compare the forecasting performance of all the models used, for both Peru and Netherlands. In order to formally assess the performance of the models, we will use the root mean square error. This value will be computed for each predictive model and the model with the root mean square error closest to zero will be selected as the best performing one.
The root mean square error (RMSE) is a measure for the difference between the values predicted by a model and the actual observed values. The RMSE can be obtained by taking the square root of the mean of squared differences between predicted values and observed values.

In the case of Peru, the RMSE has been manually computed in R using the procedure described above, using the observed values from the peru.test data set.
In the case of Netherlands, I have used the function *accuracy* from library *forecast* in order to compute the value of the RMSE for each of the regression splines and the exponential smoothing. The "observed" values were obtained from the nl.test data set.
The results can be viewed below:

Peru - MA(1) series with linear trend - RMSE: 28.18719

Linear regression - RMSE: 63.1121

Netherlands - Log - linear regression spline - RMSE: 11.37985

Linear regression spline - RMSE: 27.20981

Holt-Winters exponential smoothing ("AAN") - RMSE: 27.27199

As we can see above, the worst performing model is linear regression model used for Peru, as it has obtained the highest RMSE. The second lowest performance is the MA(1) series with the linear trend. The predictive models used for Peru seem to perform worse in general.

All three predictive models used for Netherlands are performing better, even though the linear regression spline and Holt-Winters are only seeing a marginal improvement. The best predictions by far are provided by the log-linear regression spline, for which we have obtained the lowest RMSE. The superior performance of the model has also been reflected in the confidence intervals plotted above - the bands are much more narrow compared to the other models.

Conclusion and Limitations

The projects's aim has been to explore the temporal pattern in the number of deaths for Peru and Netherlands and model it using regression models. Once modelled, the trend was removed from the data and the short correlation was modelled using an MA(1) series process in both cases. The regression models were then trained on the data up until and including the 15th of May and then used to generate 16 step ahead predictions. These predictions were then compared to the observed values and the best model was selected, based on the RMSE value. The model that generated the best performing predictions was the log -linear regression spline used for Netherlands, and it obtained an RMSE of 11.3798.

I believe there are various ways in which the models could be improved. First of all, I believe the non-constant variance present in both countries is quite problematic and it should be addressed. I have addressed the issue in the case of Netherlands by fitting a log-linear model, however, heteroscedasticity was still present in the residuals. The case is similar for Peru - we still had heteroscedasticity present after removing the trend from the data. I do presume that the non-constant variance is owed to the way data are reported, rather than actual fluctuations in the number of deaths and cases.

Furthermore, in the case of Netherlands there was still short-term correlation present at lag 14, even after fitting an MA(1) series process. It is unclear wether this is happening because the trend / seasonal effect was not captured and removed in an appropriate manner or whether I simply didn't handle the short term correlation properly. However I would like to add that I did fit an AR(1) and ARMA(1,1) processes as well, but the residual plots and ACF and PACF remained unchanged.

Another way in which the models could be improved is by adding additional covariates in the models such as the number of cases and the number of tests available. The number of cases would not be known in advance, therefore it would not be possible to use them for improving the prediction accuracy. However, other factors that affect the spread of the disease indirectly could be used in order to improve the prediction accuracy such as population density, quality of the health services, severity and duration of social distancing measures etc.