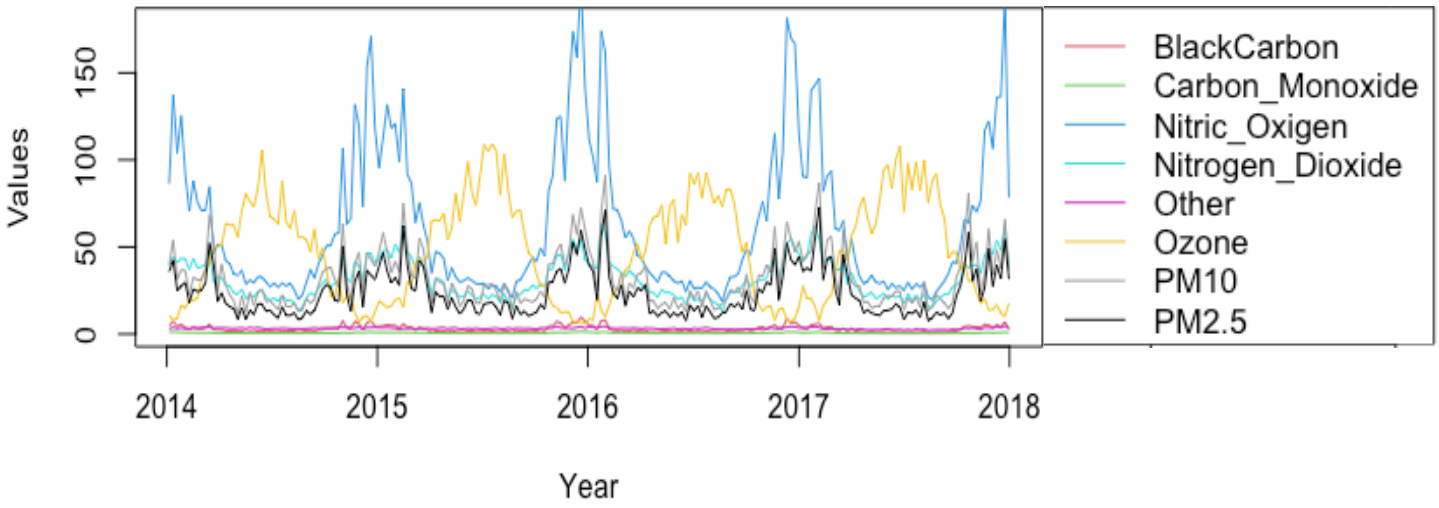
Forecasting air pollution in an Italian region

Introduction:

One of the concerns of governments around the world is the anthropogenic influence on air pollutant concentrations. The ability to model and predict future air pollutant concentrations is therefore of interest as it may be beneficial in controlling these air concentrations. The data set we have been provided records the concentration of 8 pollutions measured weekly, from 2014 – 2017. In this analysis, we will create a model for the time series data measuring one of these concentrations. We will then attempt to predict the concentrations from the year 2018 and measure the accuracy of our predictions.

Explanatory analysis:

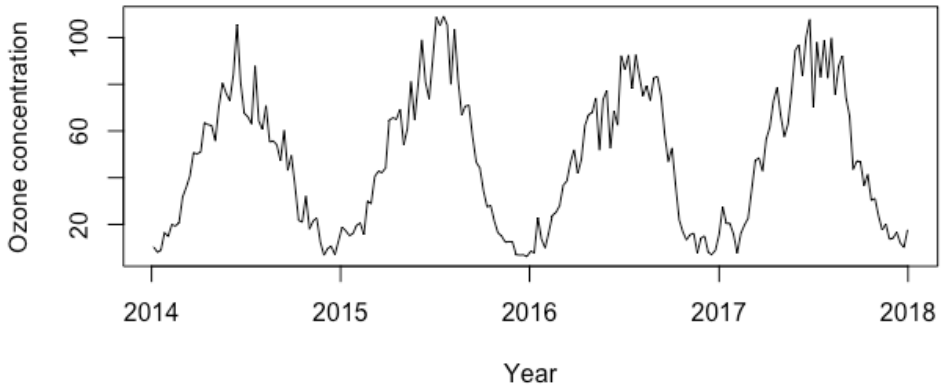
We will begin by visualising our time series data:



We can see from the time seres data that the concnetrations of Nitric oxeigen have the most variablity while the concentrations of Black carbon, other and Carbon monoxide have the least variablility. All variables show evidience of seasonal patters. The variables PM2.5, PM10 and Nitrogen dioxide seem to superimpose each other.

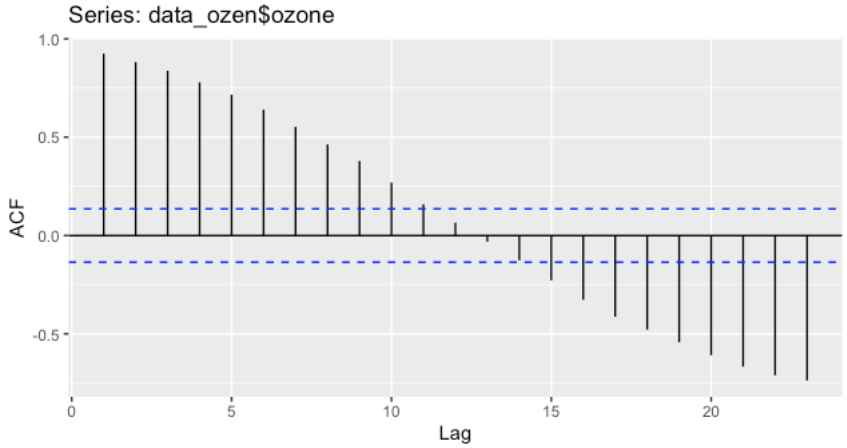
For this assignment, I have chosen to focus on the concentration of the ozone pollution. Ground level ozone is a secondary pollutant, meaning it is created in the atmosphere through a cycle of reactions of its precursor's nitrogen oxides and volatile organic compounds. These could be pollutants emitted from vehicles, factories and other industrial sources, fossil fuels, combustion, consumer products, evaporation of paints, and many other sources.

Let's look at the ozone time series alone:



We can see that the patters are very seasonal, with lowest values at the start of each year and highest values in the middle of each year. 2016 had the lowest peak. There doesn't appear to be any trend in this data.

We will now produce further plots to be able examine our data.



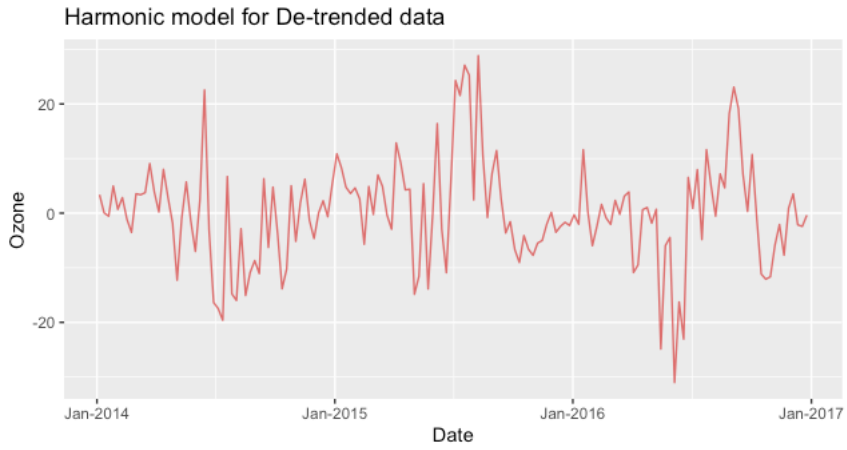
This correlogram displays either a non-linear trend or seasonal variation. We cannot tell as only 25 lags are produced. It is also not possible to determine if short-term correlation is present.

Model selection:

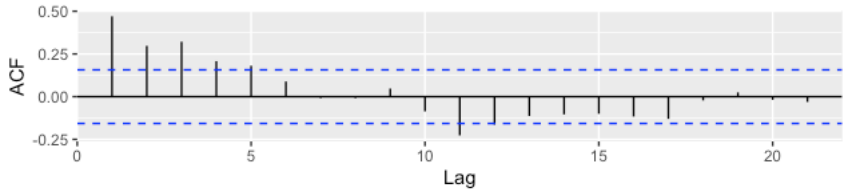
As with any model selection, we will split our data into training and validating data. The training data will be from the year 2014-2016 (3 years' worth of data) and the validating data will be for the year 2017. We have further data on the year 2018 which will be for the final testing set.

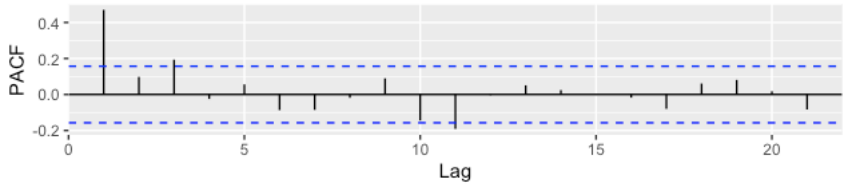
For a model to be fitted to this data, the trend or seasonal variation must be removed.

We will begin by fitting a Harmonic model to our data set. We will assume that our data follows the common seasonality model - sine and cosine regression. Below is the fitted data for the de-trended data.

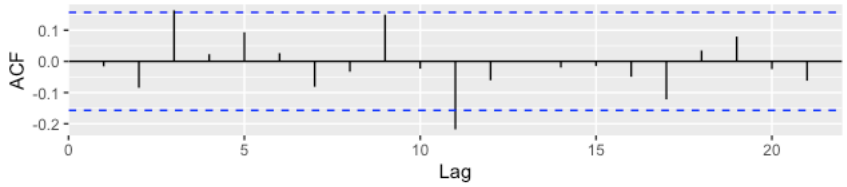


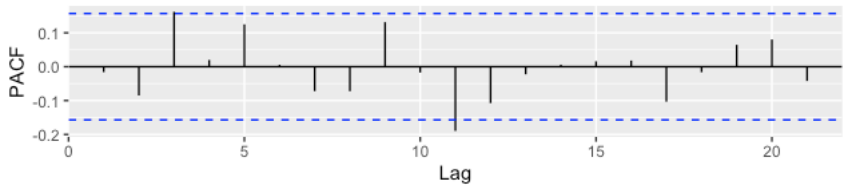
The seasonality does seem to be removed but is the data stationary?
We will not check the residuals of this harmonic model on our data set.





The partial autocorrelation function suggests that an AR(2) process may be appropriate. We will therefore fit an AR(2) model to our data.
Once the AR(2) model is fit to our data, we produce the following correlogram and PACF

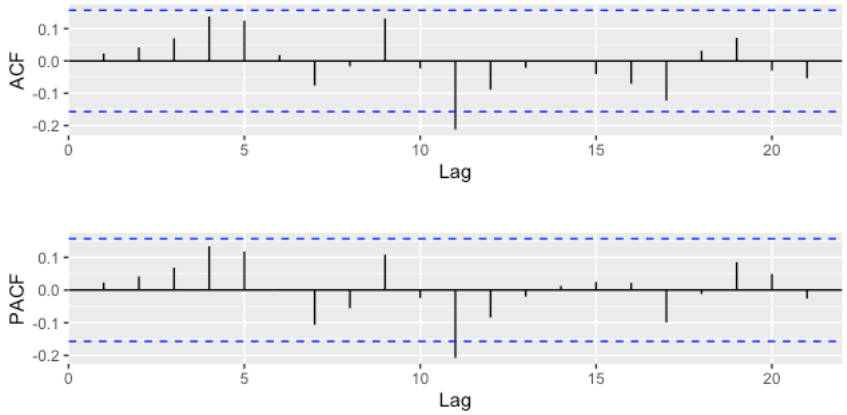




The residuals look independent, so the model we fitted is appropriate. All values for lag remain within the 95% confidence intervals aside for lag 11. There is no short-term correlation left.

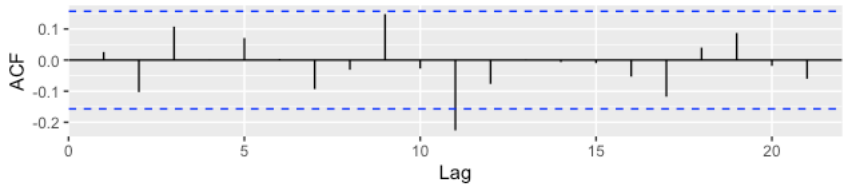
The second method we consider to removing trend and seasonal variation is called moving average smoothing. From the ACF plot produced from the de-trended data, we can see that a

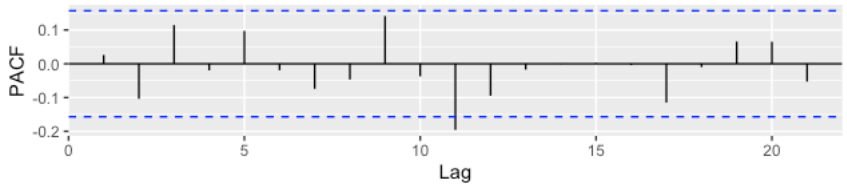
MA(3) model may be appropriate. After applying an MA(3) model to our data, we produce the following ACF and PACF:



Again, the ACF and PACF show a detrended data set with no short term correlation so the model used may be appropriate.

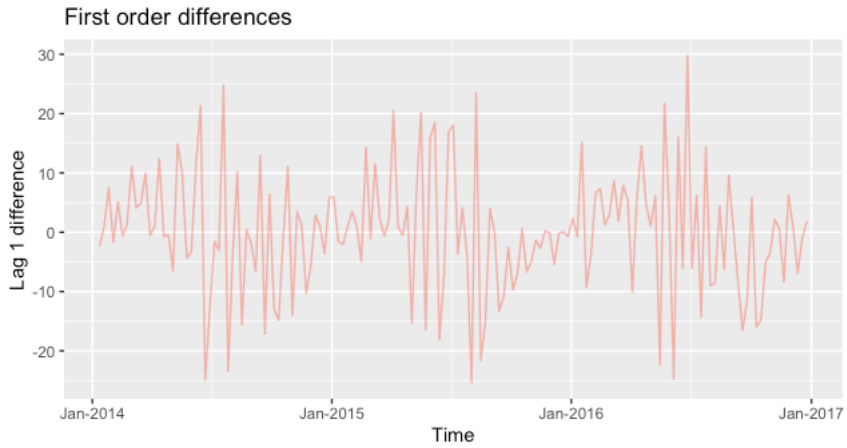
Another method we can fit is the ARMA model which is a combination of an AR and MA model. The, the model fitted will have an AR and MA value of 1



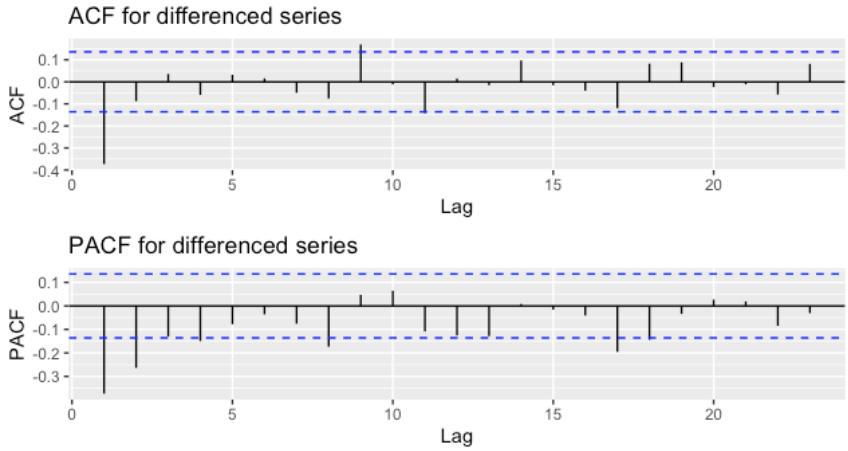


Again, the PACF and ACF show no trend so the ARMA model may be a good choice.

Trend and seasonal variation can also be removed by combining the difference operators. After applying first order differencing, the following de-trended data is achieved:

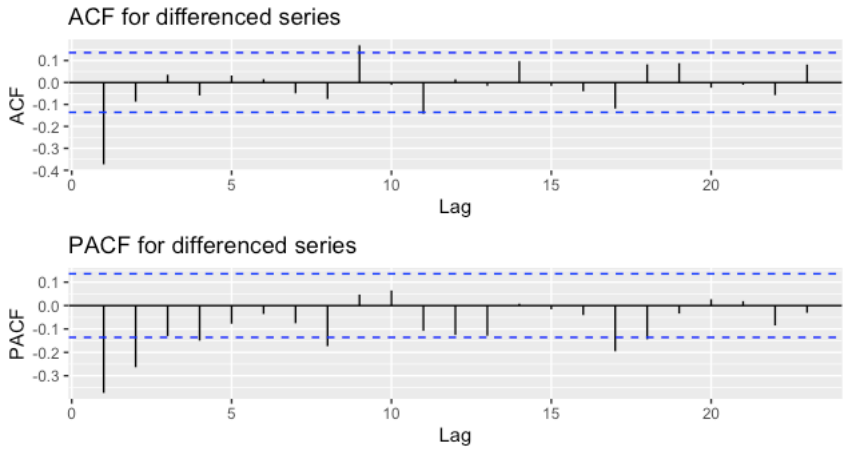


From this plot, it seems that first order difference has successfully removed all trend and variation.



The ACF and PACF have some concerning results, specifically at land 2 and 9. This suggests that using differencing may not be necessary.

We can try fitting an ARIMA model which combines that of the AR(1), MA(1), and the difference to see if it produces a good fit of our data.



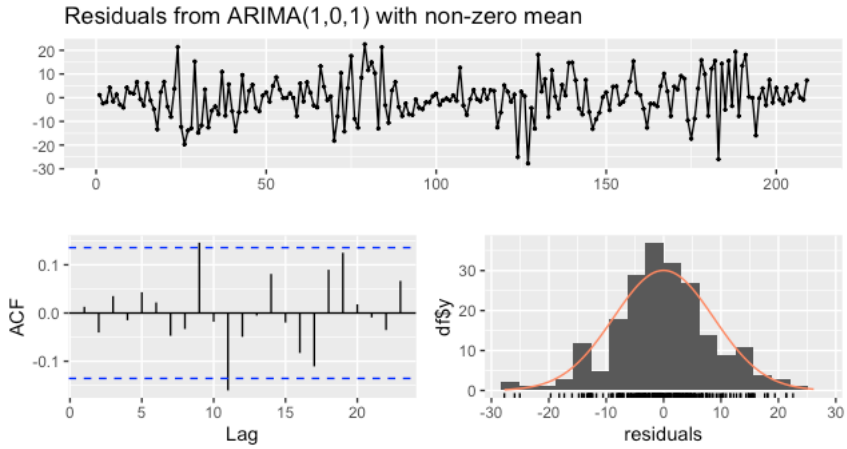
These plots again suggest that the fit may not be ideal as there are several lags beyond the 95% confidence intervals. Furthermore, the function auto.arima( ) in R allows us to locate the best arima model. We can see that the best model is without differencing - ARIMA(1,0,1) which is essentially the same as ARMA(1, 1).

To sum up, we have de-trended our data by applying a harmonic regression model to our time series data. We then applied 3 models to our data to predict the trend. AR(2), MA(3), and ARMA. All these models have shown a good possible fit for our data. We have then tried to de-trend our data by applying first order differencing to our data. This has not produced a detrended correlogram so we will not use differencing for this model.

Of the three models, which is best to use? We can look at the AIC of each of these models:

```
> model.ar$aic
[1] 1117.052
> model.ma$aic
[1] 1117.528
> model.arma$aic
[1] 1114.94
```

As lower AIC are preferred, we can see that the ARMA (ARIMA(1,0,1) model seems to be most efficient.
The residuals for this model are normal.

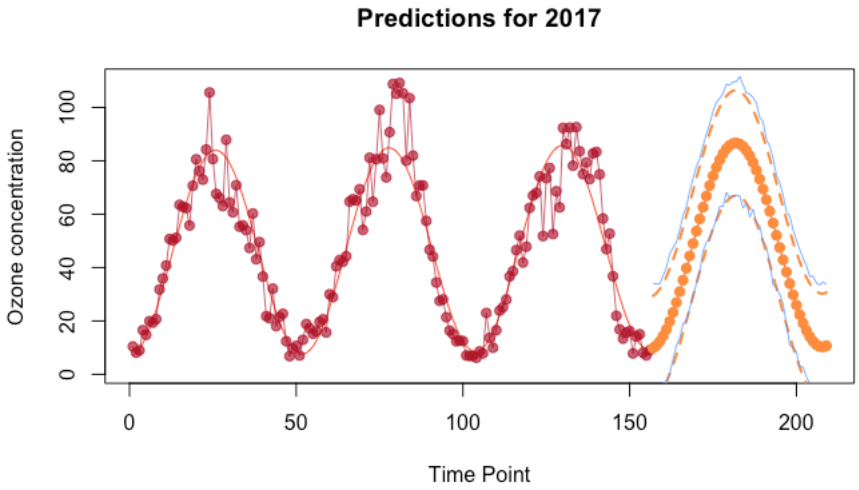


Predictions:

For the remainder of this analysis, we will predict future ozone concentrations. We will first test our models on the validation data (data for the year 2018). The model that performs best on our validation data will be used for prediction on the 2018 testing data set.

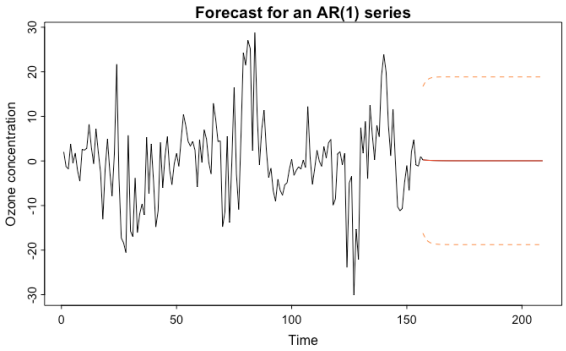
There are various methods for forecasting future time points such as regression (linear and non-linear) and exponential smoothing.

We have started to use a regression model to predict our data. We flitted the Hamonic model on our data set and produced the following predictions:



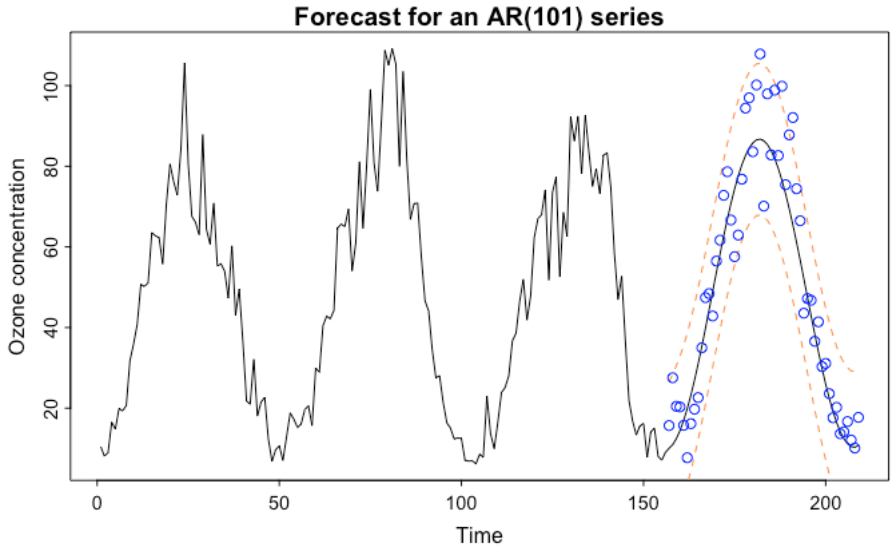
The model created a curve across our data and used that curve to predict future observations. However, the predictions may not be entirely accurate as they are based on seasonality alone, not trend.

Now, we will make predictions based on the Arima(1,0,1) model. We will first make predictions based on the residual's series. We obtain the following plot with prediction intervals.



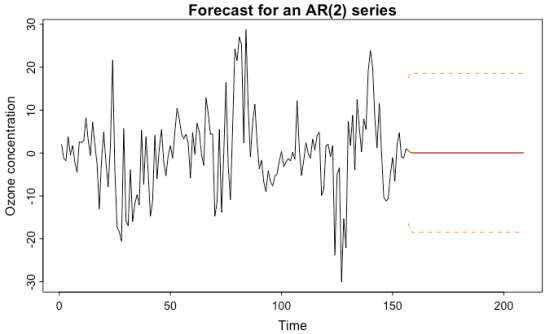
We are then going to take those predictions and add it to our model which captures seasonality. We will also create 95% prediction intervals for our predictions.

The curve is the predictions while the dotted lines are the confidence intervals. We can see that the predictions follow the overall trend seasoned pattern.

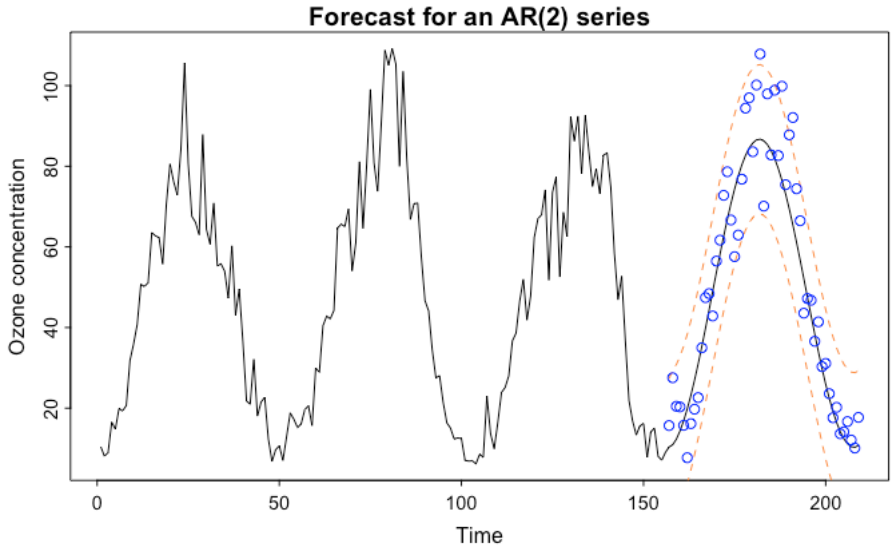


The blue values are the actual values so we can compare the predictions to the actual values. We can see that the predictions follow the pattern however underestimate the large values. This could be because the previous year was lower than the other 3.

We can forecast the next year based on the AR(2) model alone: Following the same steps as before, we produce the following residuals plot with the predictions and their intervals.



Again, these residuals must be added to the seasonality model for accurate predictions.



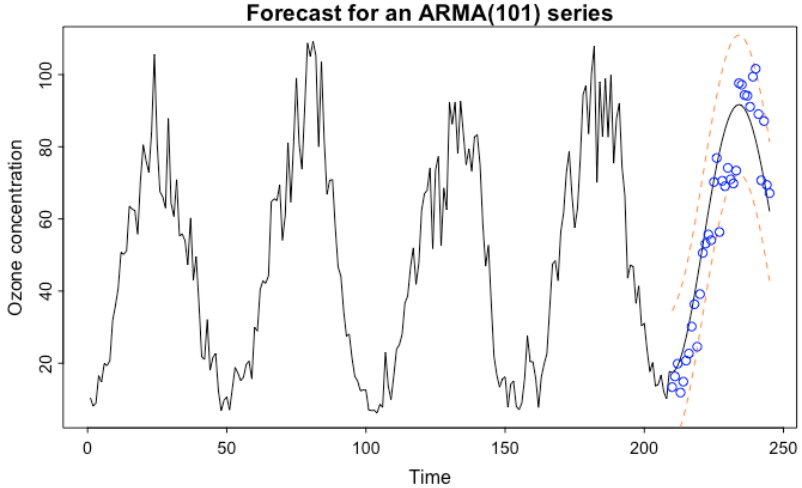
In this graph, the curve represents the predictions, the dotted lines are the 95% confidence intervals, and the blue dots are the original values. We can see that the predictions of ARIMA (101) and AR(2) are very similar.

Which model is preferred?
We will assess the bias, RMSE and coverage for both models:

	Arma(101)	AR(2)
Bias	-3.482474	-3.495396
RMSE	10.11672	10.11795
Coverage	94.33962	94.33962

The RMSE is quite large. This makes the validity of the model questionable. The coverage is quite high which suggests that most of the data is captured by the predictions.
We can see that the Bias and RMSE are lower (although very slightly) for the Arima model. The coverage is identical for the two models. Therefore, we will use the Arima model for on the testing data.

Our testing data contains the information for the years 2018-2020. We will first use our testing data set to predict the next 36 observations, (half a year). Using the same method we used above with an ARIMA(1,0,1) model, we create the following predictions:



The blue points are the actual values, the black curve is our prediction, and the dotted lines are the confidence intervals. We can see that our prediction goes approximately through all our points while there is a lot of variation. We can also see that the confidence intervals are tight when the concentration is between 40 and 80. This means we are sure about our predictions. However, the intervals are wide for high and low values. This shows that we are less confident with our predictions. This makes sense as the values for high concentration have huge variability.

How well did we do?
Bias: 3.504756
RMSE: 10.87359
Coverage: 91.66667
MAE: 8.845199
MAPE: 0.1956005

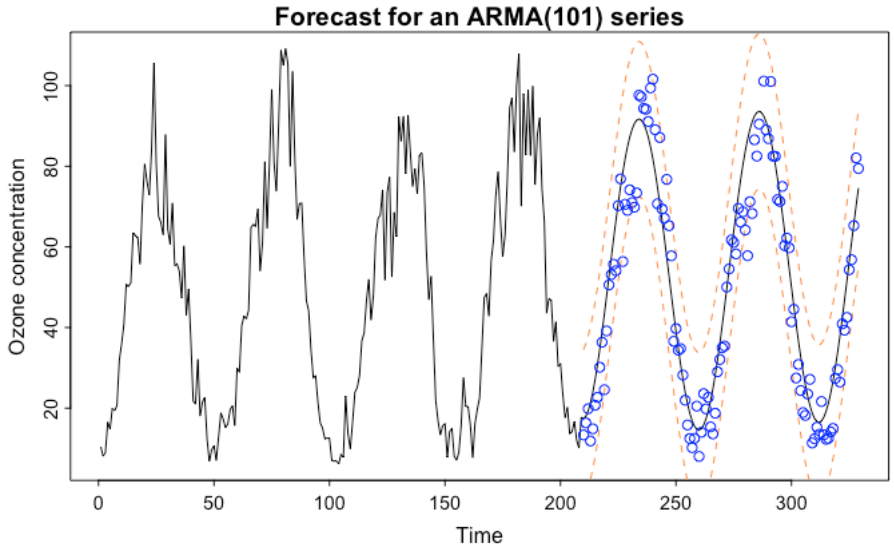
The RMSE is seemingly high, which means that we do have errors throughout our predictions. The MAE (Mean Average error) can be interpreted as follows: On average, the distance from the forecasts to the true values is 8.8. We can further understand the MAE by converting this to a percentage – MAPE. Our MAPE suggests that on average, the forecasts were 19% away from the actual values. This is considered a good prediction. The coverage of 91.6% shows us that our actual values fall within the 95% confidence intervals, 91.6% of the time. Considering the variability of the data, we can consider this to be good predictions.

We can evaluate the RMSE and MAE achieved here by comparing it to a threshold RMSE and MAE. We fit an arima model, without doing any work to it such as removing the harmonic trend. We then compute the RMSE and MAE:
RMSE: 39.00134
MAE: 32.43347

As we can see, the RMSE and MAE of our model is dramatically lower than that of the baseline arima model. This is very comforting as it proves that our model is so much more efficient in producing the predictions.

This model is limited as the predictions created a smoothed curve whereas the actual values do not follow a tidy pattern. Furthermore, our predictions do not consider any other predictors aside from previous ozone concentration patterns. To improve this model, it may be worth considering other predictors. One issue with our current model is that we ignored the large value for lag 11 on the ACF plot.

Following the same steps as above, we can calculate the next 120 values, from the year 2018-2020.



The curve shown is our predictions together with the confidence intervals. The graph shows us that the confidence intervals are very tight for the concentrations which are between 40 and 80 yet very high for the higher and lower values. This is because there is more variability for the smaller and large values. The blue points are the actual values. How well did we do?

Bias: 3.080703
RMSE: 9.514447
Coverage: 93.33333
MAE: 7.408387
MAPE: 0.2145798

The RMSE is seemingly high, which means that we do have errors throughout our predictions. On average, the forecasts distance from the true values is 7.4. On average, the forecasts were 21% away from the actual values. This is generally considered a good prediction. The coverage of 93% shows us that our actual values fall within the 95% confidence intervals, 93% of the time. Considering the variability of the data, we can consider this to be a good prediction.

Conclusion:

In this report, we created a model to predict the ground level ozone concentrations. We then tested out models on the validating data set and chose the best model based on the RMSE and MAE. The preferred model – ARIMA(1,0,1) was then used on an out of sample testing data set. Our predictions were reliable which suggested that our model will be able to make predictions for the future. However, it is possible that predictions based on data from the year 2014 – 2020 are unreliable and are not helpful for future for future predictions. This is because of the outbreak of coronavirus pandemic from the year 2020. It is known that the global lockdown dramatically reduced nitrogen oxides combustion. Ozone concentration is produced by NOx. The fact that the NOx concentration reduced resulted in a dramatic reduction in the Ozone concentrations. We can see from this that as pollutions concentration is dependable on many variables, it might not be reliable to predict concentrations solely on previous years.