STATS 4A03 Group Project
Winter 2020

Modelling Madrid Air Quality to Predict Future Pollution Levels

Emily Sanderson
Javeriah Waris
Jovana Zelenovic

August 20, 2020

Abstract

The purpose of this time series analysis is to examine, and ultimately forecast, the monthly averages of air pollutants such as Sulphur Dioxide and Nitrogen Dioxide for the city of Madrid. The basic design of our study included the examination and consequent transformation of the time series, followed by the selection of a seasonal ARIMA$(p, d, q)(P, D, Q)_m$ model. The model selection portion of our analysis was largely aided by the use of "auto.arima" function, ACF and PACF plots, and Akaike information criterion (AIC).

Contents

1 Introduction and Background

Some 3.8 million premature deaths annually are attributed to outdoor (ambient) air pollution according the World Health Organization [3]. In recent years, many cities have made it a priority to reduce pollution levels to improve the health of the community. In Madrid, Spain, authorities took drastic measures to reduce pollution within the city by making 472 hectares of the city center off-limits to traffic [1]. This was due to the dangerous levels of air pollution occurring during the dry seasons.

The city releases its hourly pollution levels collected from several stations across the city, all measuring a variety of contaminants in the air. Decide Soluciones has posted this data in an organized fashion to be used for public analysis [2]. We chose to aggregate the data into daily and monthly values for the purposes of our forecasting model.

One main obstacle of this data is that most of the stations did not collect every contaminant and the levels they collected changed over the years of collection (2001-2018). For the purposes of consistency, we chose the contaminants that were collected every year and created a 'Total Pollution' variable. Measured in $\mu g/m^3$, this total pollution variable is made of the levels of the following pollutants: sulphur dioxide, carbon monoxide, nitrogen dioxide, ozone, particles smaller than 10 μm,toluene (methylbenzene), benzene, ethylbenzene, total hydrocarbons and non-methane hydrocarbons. An in-depth definition and resulting health implications for each pollutant can be found in Appendix A.

2 Analysis

First we want to plot the data to get a better idea of it's general trend and to indicate what the next steps should be.

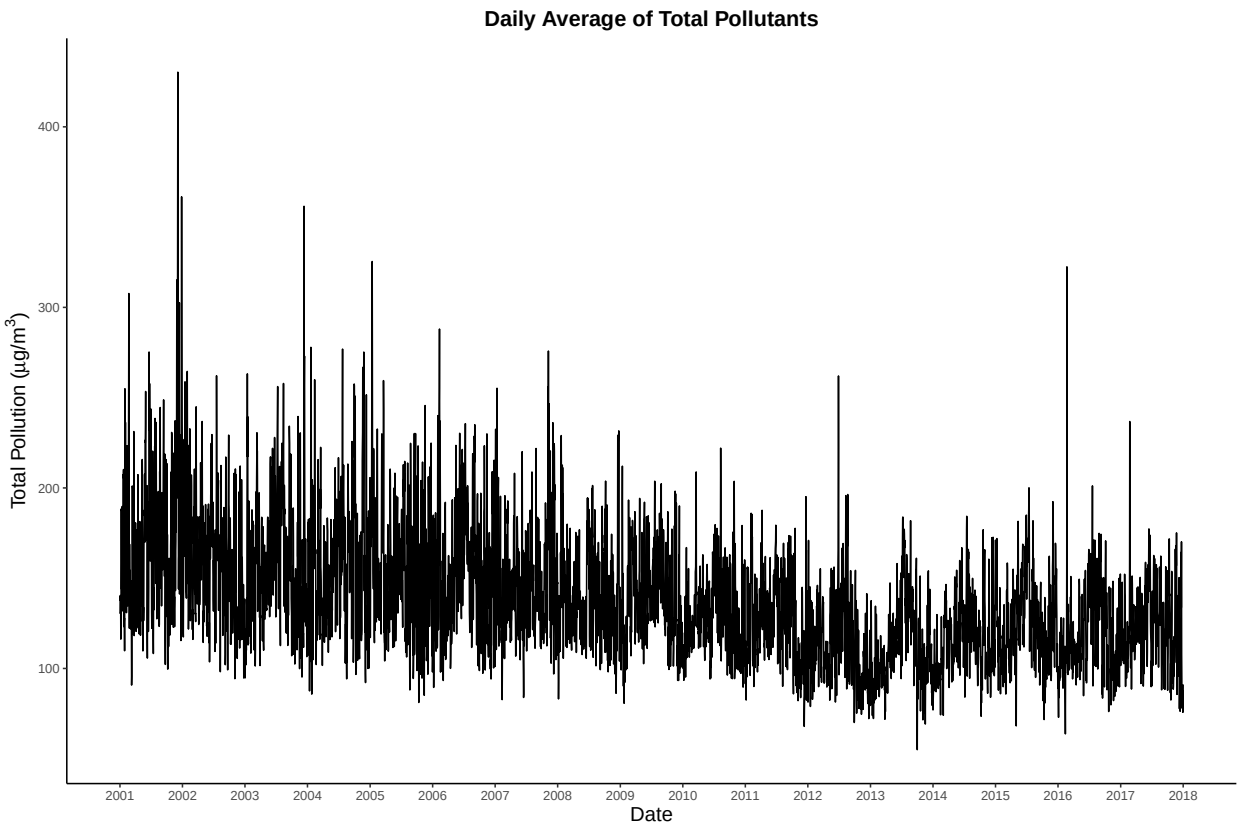

Figure 1: Simple plot showing the overall trend of the data

We can see from the above plot that the data displays an unstationary time series as seen with the downward trend and a variance that somewhat decreases as time goes on.
If we group the data by month, we can get a clearer picture of the trends and we can better visualize the data.

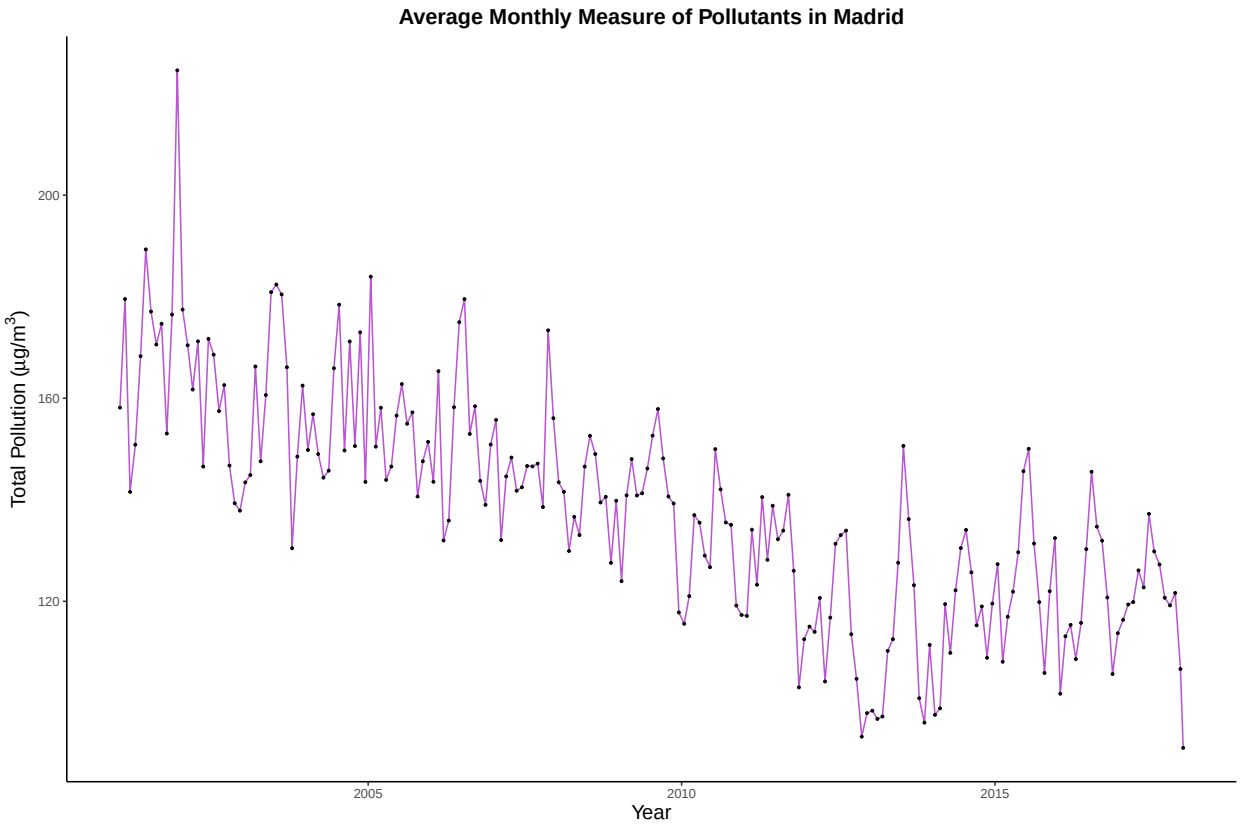

Figure 2: Simple plot showing the overall trend of the monthly data

From Figure 2 we can get a clearer picture of the data by taking the monthly mean which will reduce the impact of outlier days or weeks. We can more clearly see that the data is decreasing as time progresses with a variance that is somewhat unevenly dispersed.
Our next goal is to check for seasonality to see if the different times of the year have an impact on the levels of pollution in Madrid.

We will investigate the seasonality further using the following plots:
Looking at Figure 3, there appears to be seasonality so we will examine the ACF of this time series to confirm this, and then we will transform the data to get a model of the data.

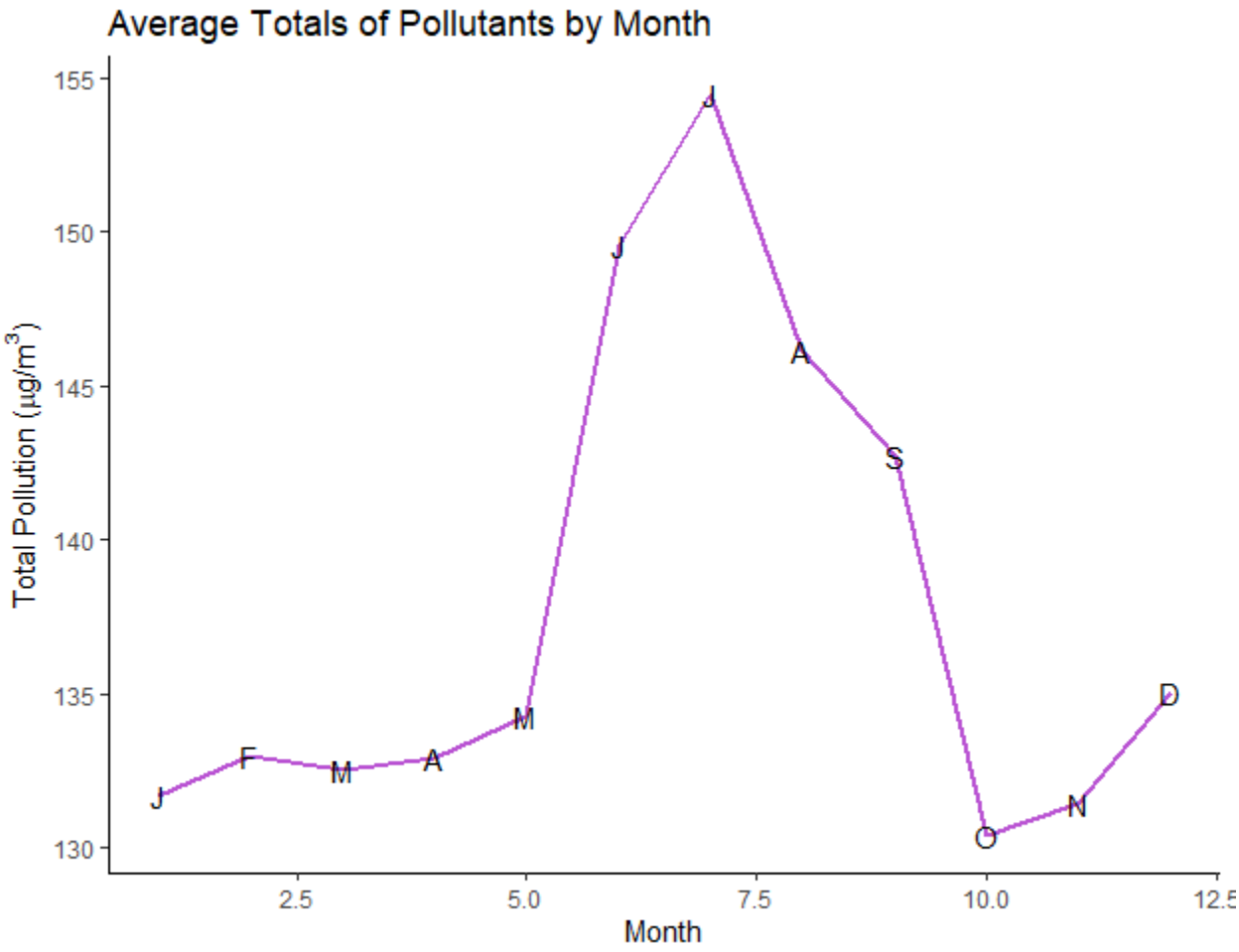

Figure 3: It's clear there's seasonality in the data: pollutants increase in quantity in the summer.

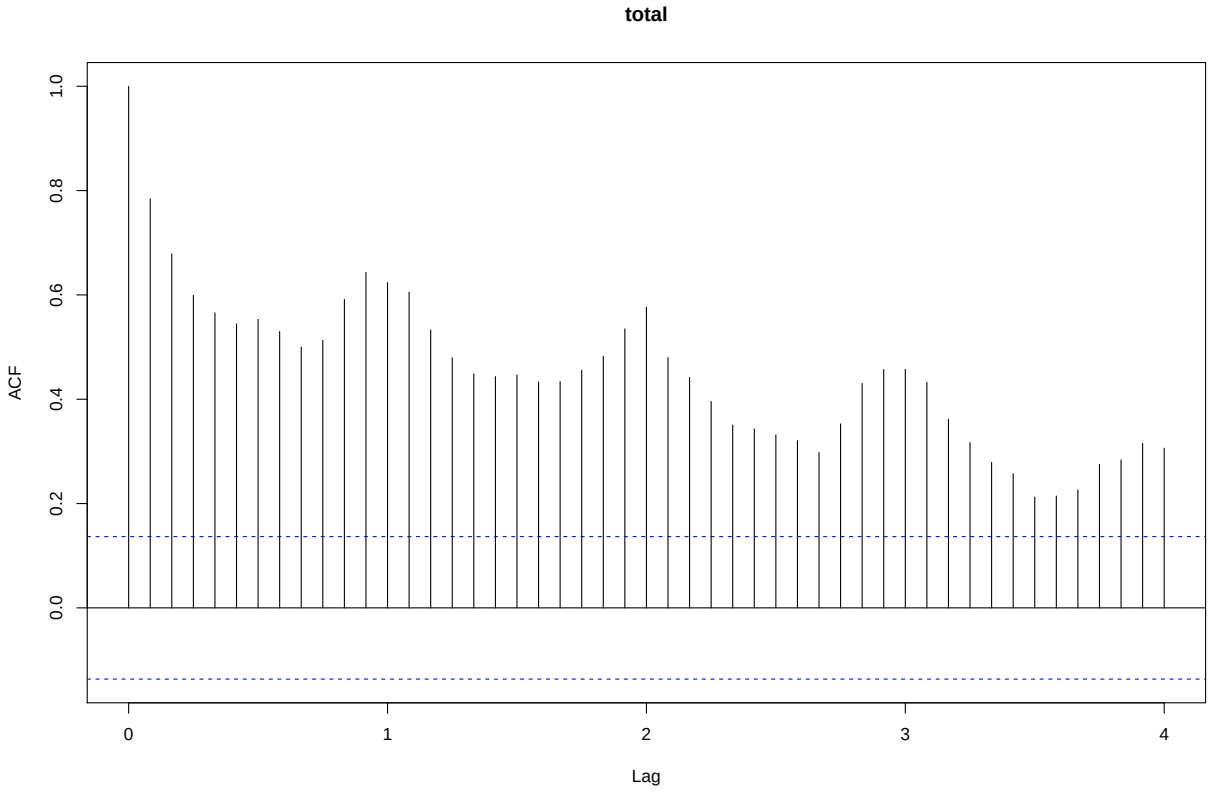

Figure 4: We can see clear increases of the ACF values at the 12th, 24th, 36th and 48th lags

It's clear from Figure 4 that our time series has a seasonal component because of the annual recurring nature of the increase values in the ACF. This means we need to transform the data to remove the seasonality, and also the non-constant mean.

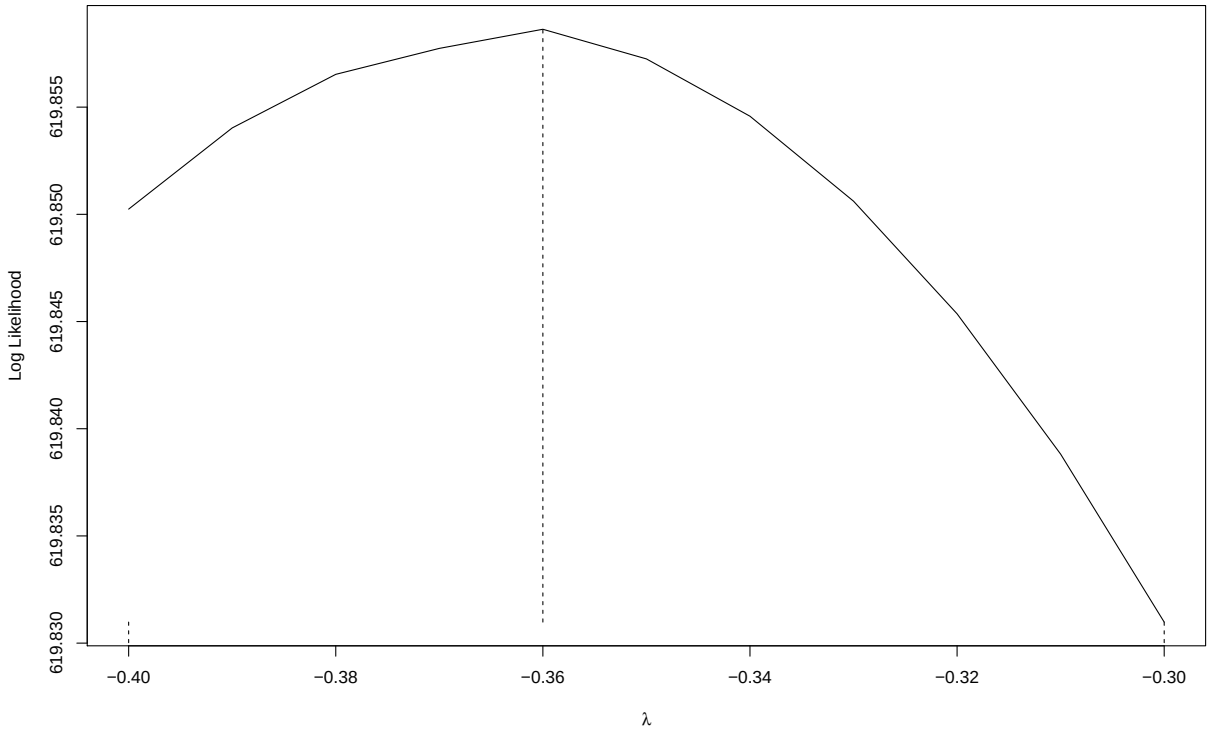

Figure 5: From the BoxCox plot we see we should do a power transformation of -0.36

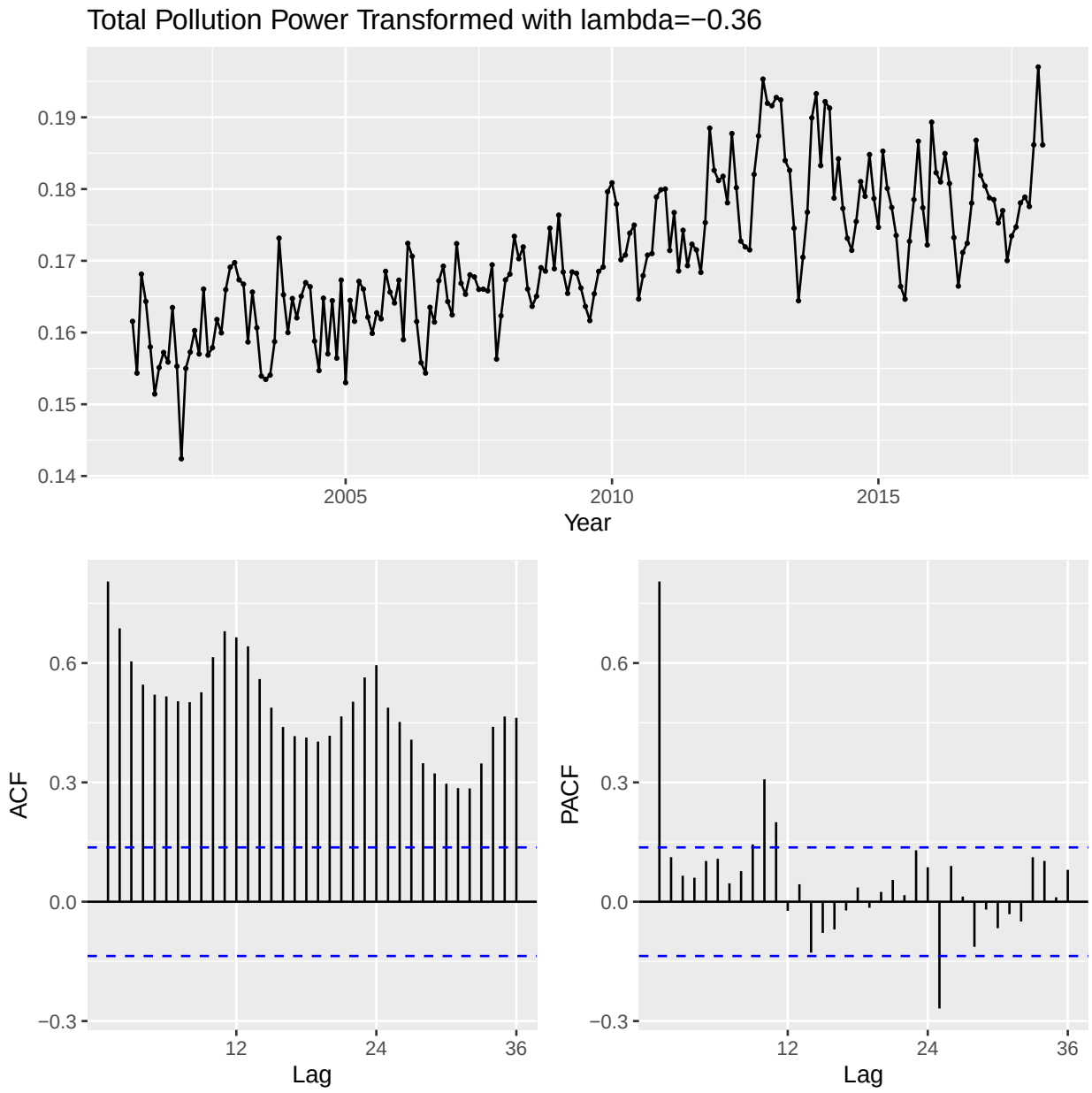

Figure 6: Trend is now increasing and ACF looks similar as above with PACF showing a large spike at lag 1

Following a power transformation, we can still see an increase in the trend of the data, so we will difference the data with lag 1 and lag 12, which is an extension of our previous conclusions of seasonality since the seasonality is related to month, we take the first difference with time lag $k = 12$:

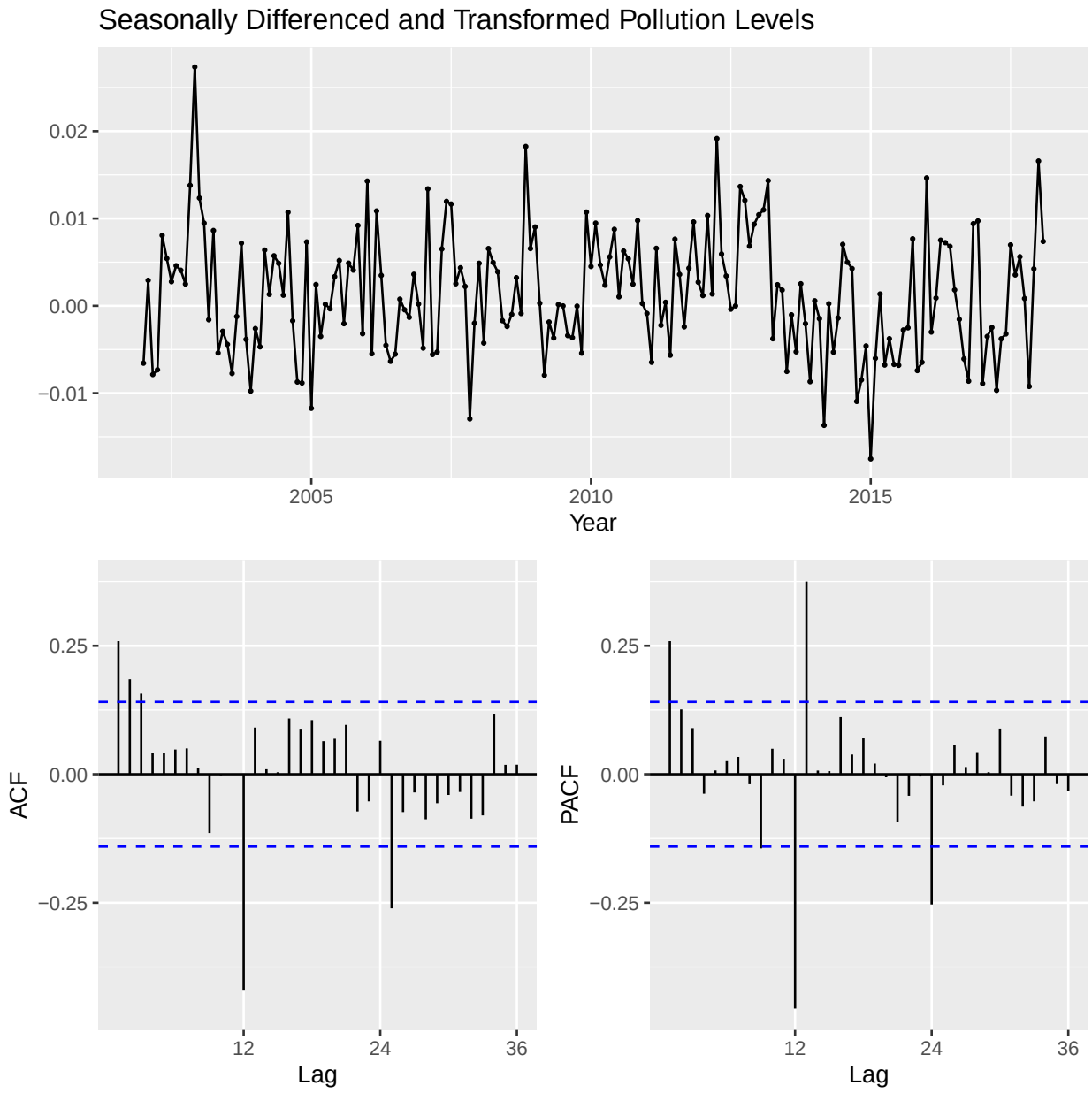

Figure 7: Seasonally Differenced and Transformed Data

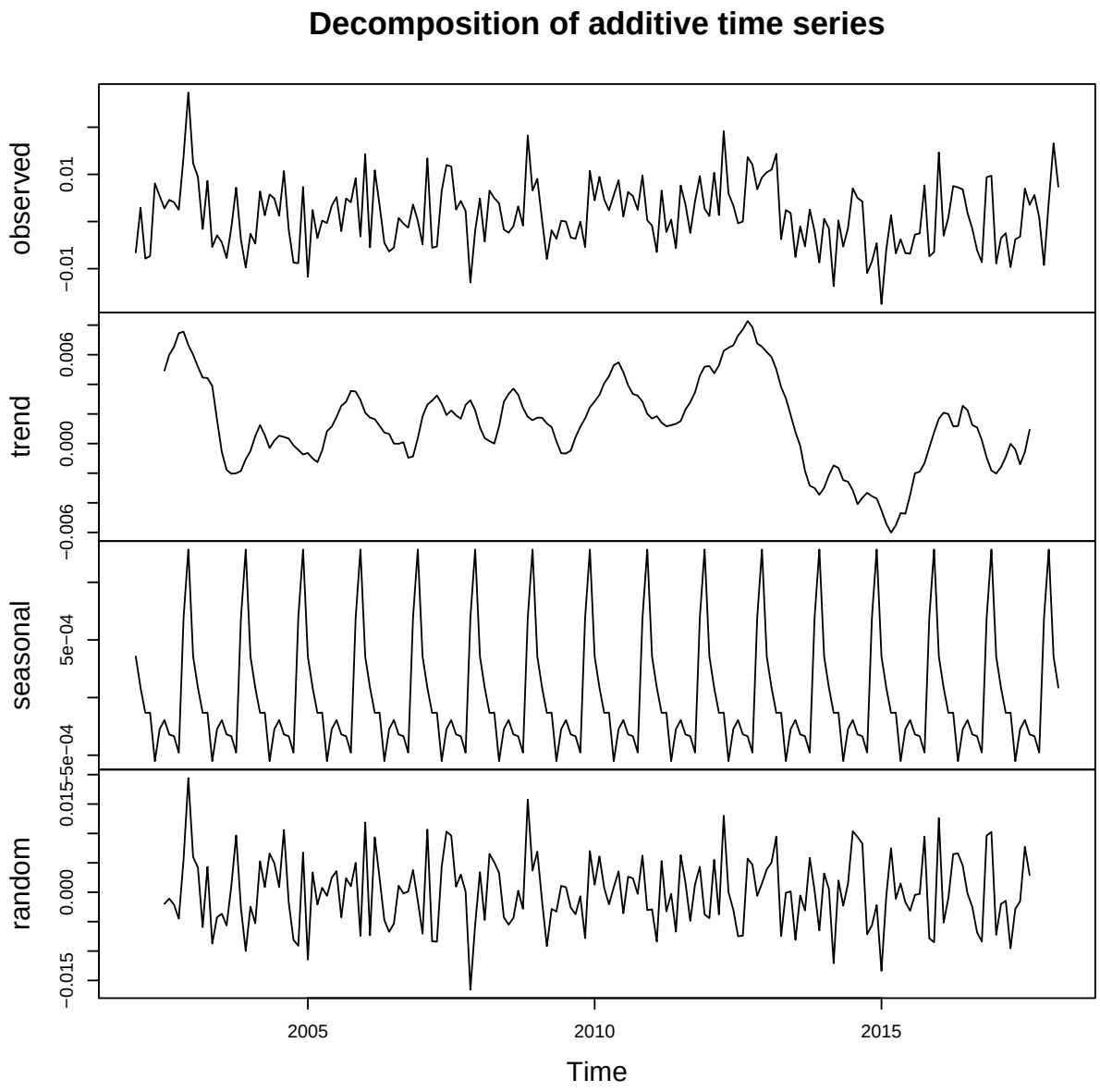

Figure 8: Decomposition of Seasonally Differenced and Transformed Data

By taking the seasonal difference at lag 12, we see a non-zero mean and fairly uneven variance even though the linear trend is reduced. Next we would try taking a non-seasonal difference to see if that can better regulate the variance of the data. One thing that this produces is the a clear indication of the seasonal auto-regressive and moving average components seen at spikes in the significance of lags of multiples of 12. There appears to be 2 significant seasonal auto-regressive components seen in the spikes at lag 12 and 24 on the PACF and 1 seasonal moving average component seen on the ACF at the spike at lag 12. We will use

this result later when fitting a model to the data.

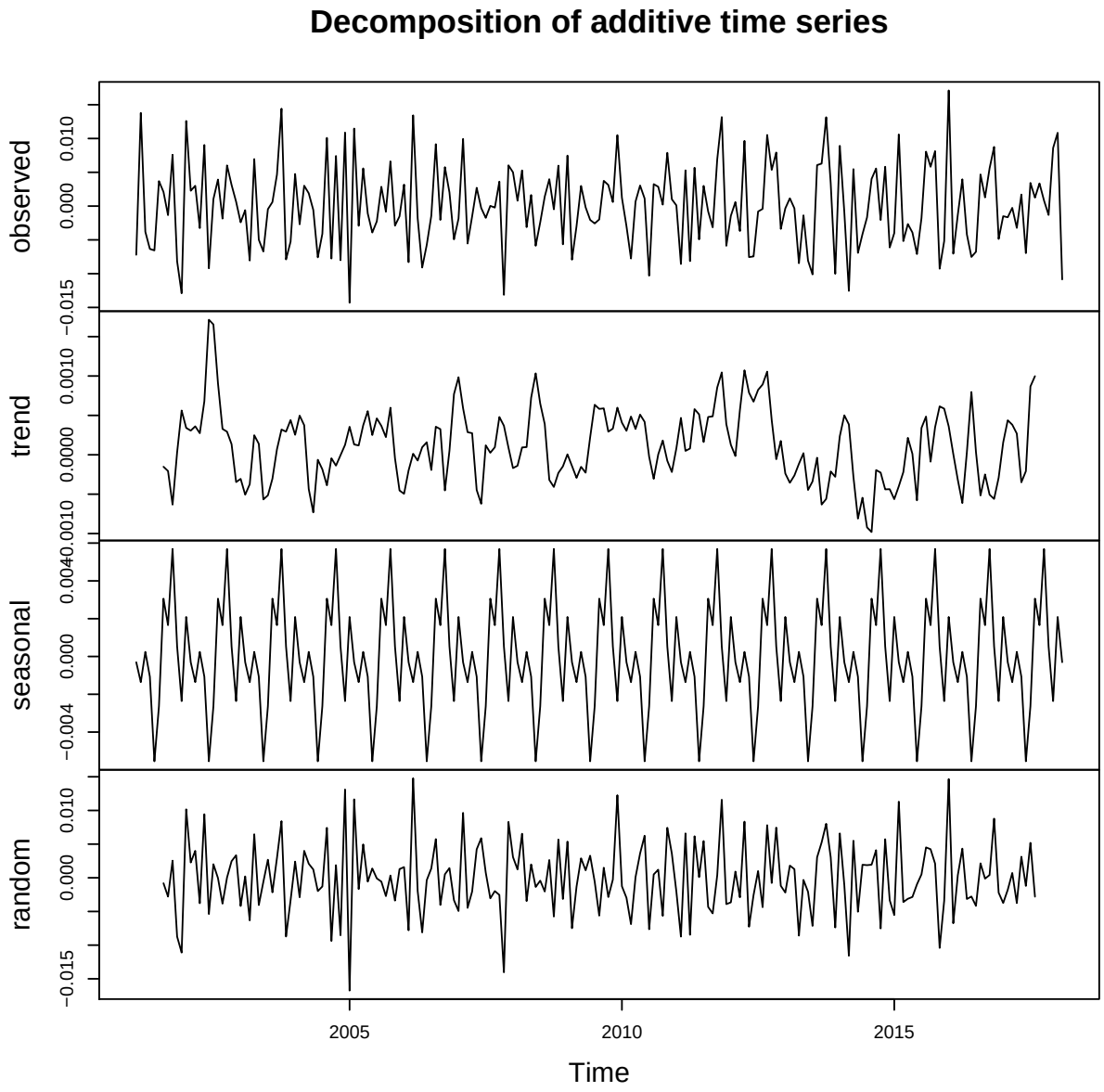

Figure 9: Decomposition of the First Difference of the power transformed data

We can see that all linear trends are now gone and and the remaining plot of the time series appears random with zero mean. This is an improvement over taking the seasonal difference. Taking a second difference could improve the model but is not shown to significantly improve the already random transformed time series. Also, we run the risk of over-differencing our time series.

3 Model Selection

From the above transformations we can see that our data follows a transformed SARIMA model, with non-seasonal coefficients (p,d,q) and seasonal coefficients (P,D,Q) with period of 12. We know from these transformations that d=1 and D=0 are suitable to use in our model.
Let's take a closer look at our PACF and ACF plots:

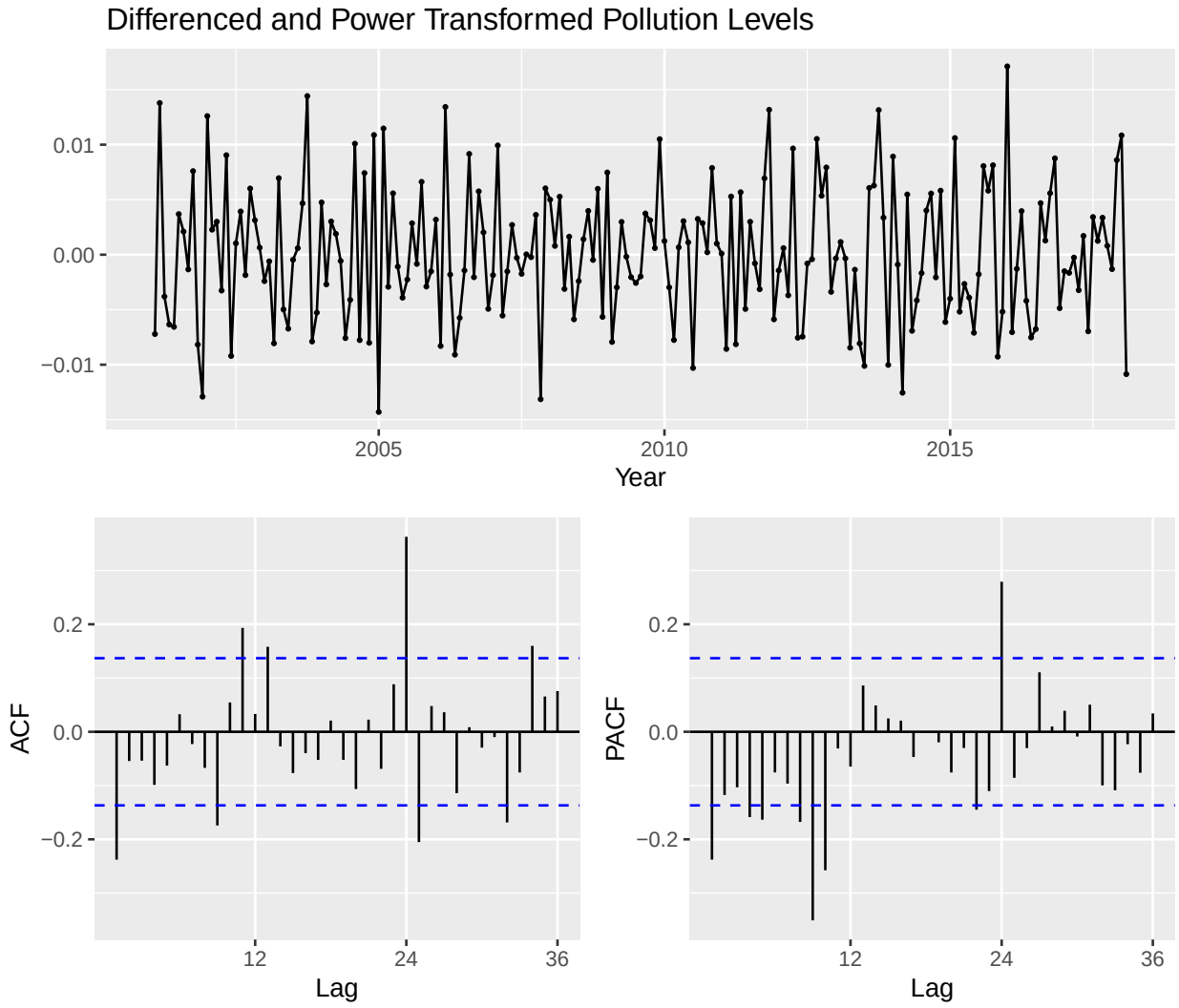

Figure 10: ACF and PACF of the First Difference of the power transformed data tell us the seasonal and non-seasonal components of the ARIMA model

From Figure 10 we can still see seasonality which will be modelled with a SARIMA model. For the non-seasonal AR(p) component, we can see from the PACF that the last "considerable" dropoff occurs around lag 4, maybe 5, which means an AR(4) or AR(5) may best fit

this data. For the MA(q) part, the ACF plot looks to be significant for multiple lags, so we should start with an MA(1) part and then change accordingly. To help with the process of choosing our parameters, we can fit a few different models with slight variations and then use AIC to choose among them:

```
fit1<-Arima(madrid_ts, order = c(4,1,1), seasonal = c(1,0,1), lambda = -0.36)
fit2<-Arima(madrid_ts, order = c(4,1,0), seasonal = c(2,0,1), lambda = -0.36)
fit3<-Arima(madrid_ts, order = c(5,1,1), seasonal = c(2,0,0), lambda = -0.36)
```

4 Testing the Model

We will calculate the AIC for three different models and then compare them. This will help with choosing a model.

Model	AIC
(4,1,1, 1,0,1)	-1165.02
(4,1,0, 2,0,1)	-1156.48
(5,1,1, 2,0,0)	-1167.49

The smallest AIC value relative to the other values is the most favorable, which comes from the third model. The next step would be to ensure that this model passes tests like the Ljung-Box test to show that the residuals have no remaining autocorrelations and that the series is stationary. Similarly, the Shapiro-Wilk's test will also help us confirm normality.

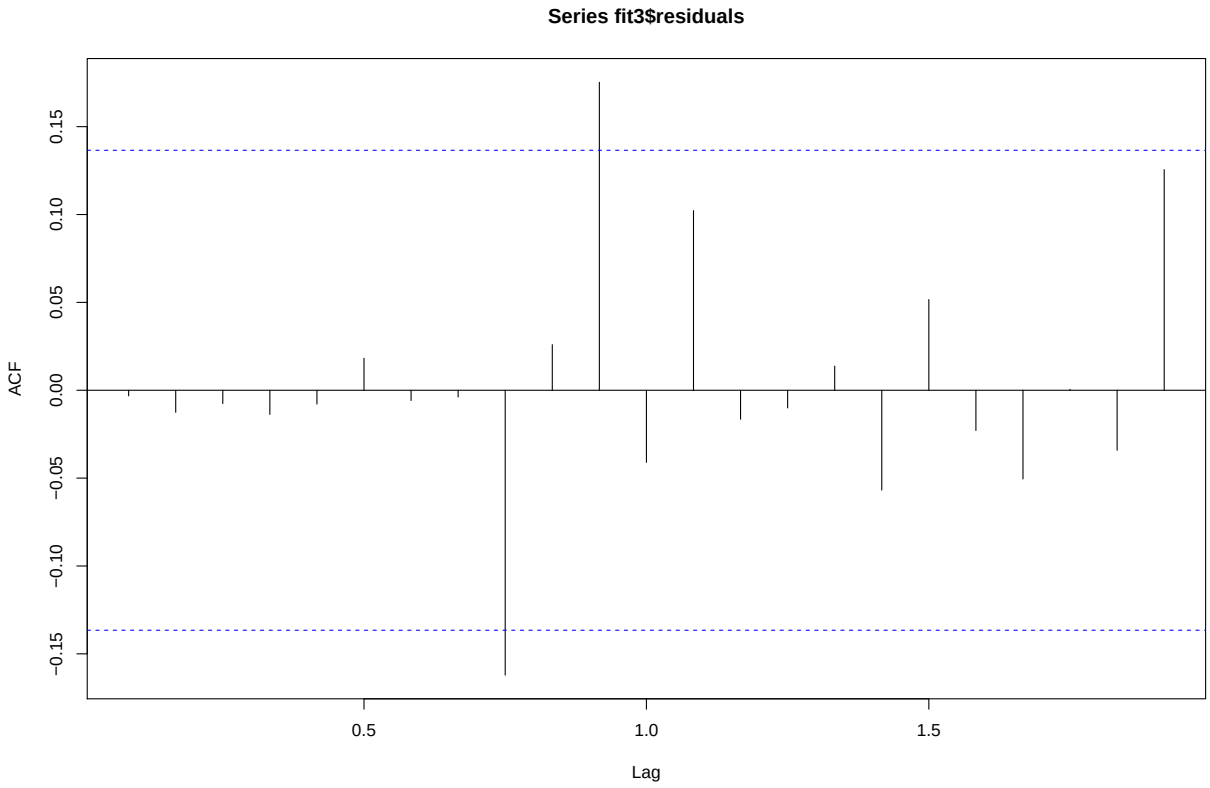

Figure 11: ACF plot of the residuals of the data.

Figure 11 suggests that the autocorrelation of the residuals is not particularly significant.

```
qqnorm(fit3$residuals)
qqline(fit3$residuals)
```

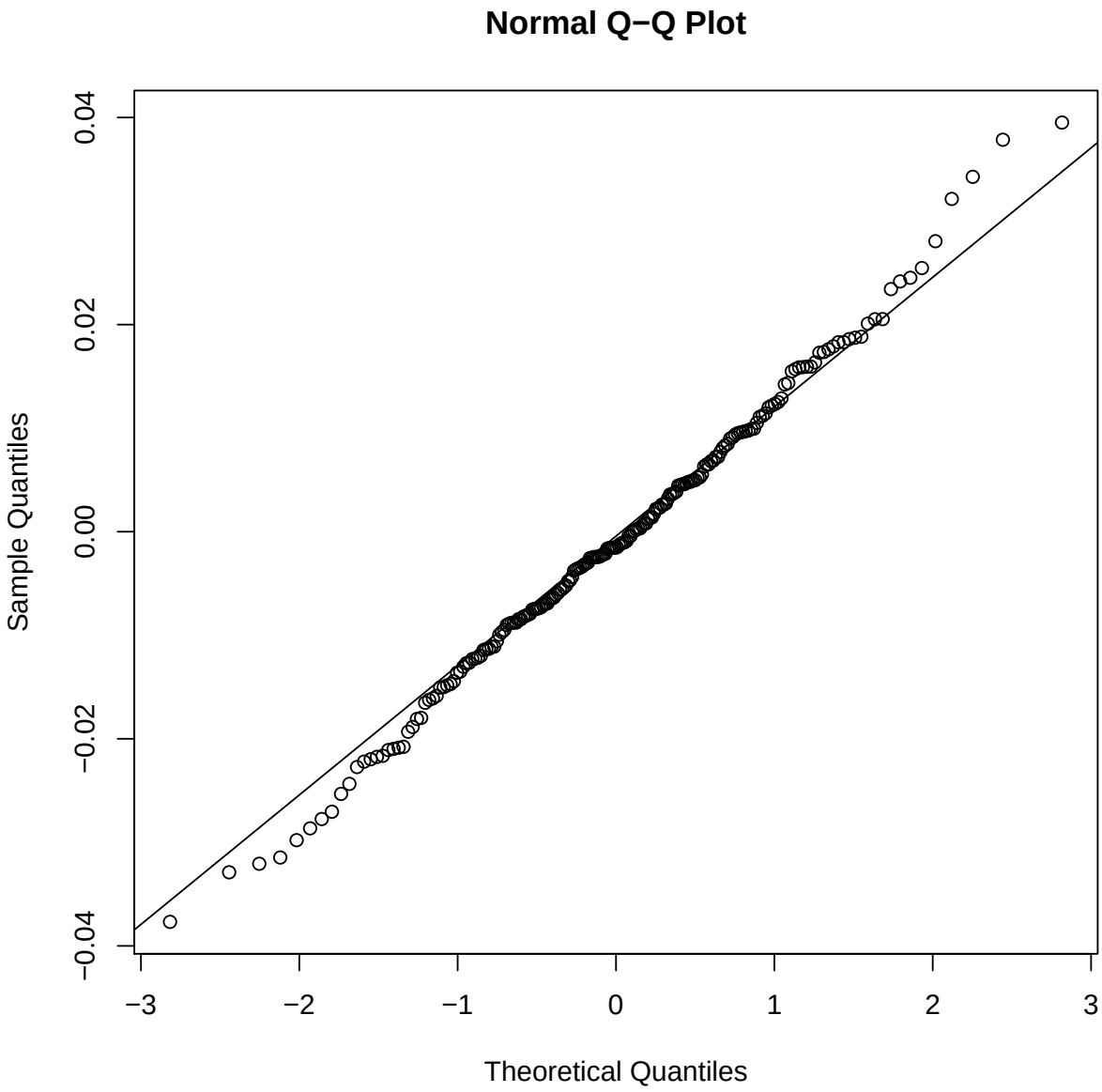

Figure 12: Q-Q Plot used to determine the normality of the residuals

```
shapiro.test(fit3$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  fit3$residuals
## W = 0.99547, p-value = 0.8001
```

```
Box.test(resid(fit3))

##
##  Box-Pierce test
##
## data:  resid(fit3)
## X-squared = 0.0021278, df = 1, p-value = 0.9632
```

The Ljung-Box test has a p-value greater than 0.05 which means we don't have enough evidence to reject the null hypothesis which is that the model does not exhibit a lack of fit. Similarily, the Shapiro Wilk's test has a p-value greater than 0.05 which helps confirm normality. This further confirms our "fit3" model is working well.
Given our analysis, we can conclude that an ARIMA(5,1,1, 2,0,0)[12] is an adequate model for our data.

5 Forecasting

Now we can look at this result and see how it it fares at forecasting future levels of air quality. First we will use a training set and a testing set to compare the model versus the actual data.

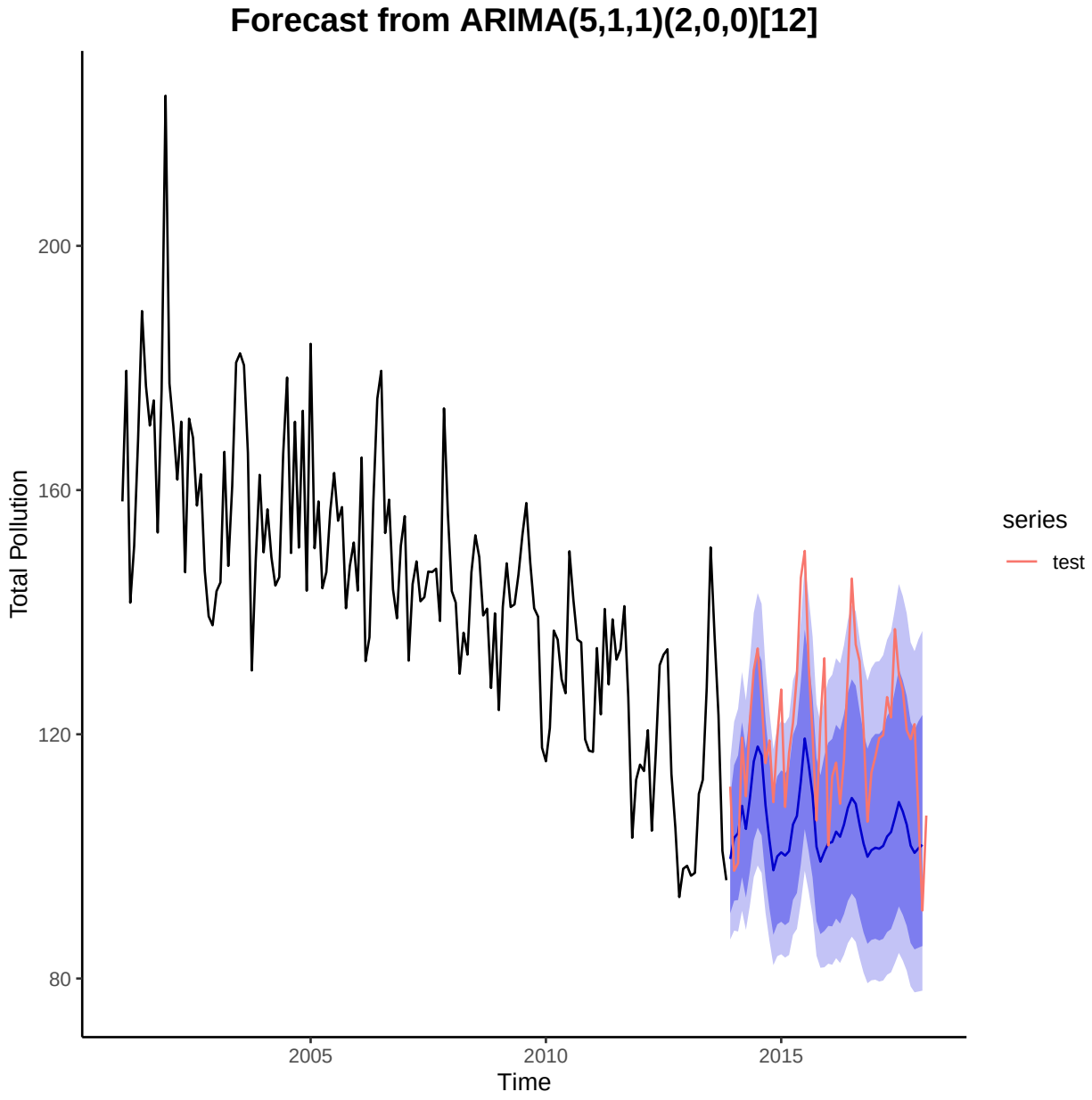

Figure 13: Forecasting the data using a training set and comparing the results to the test set

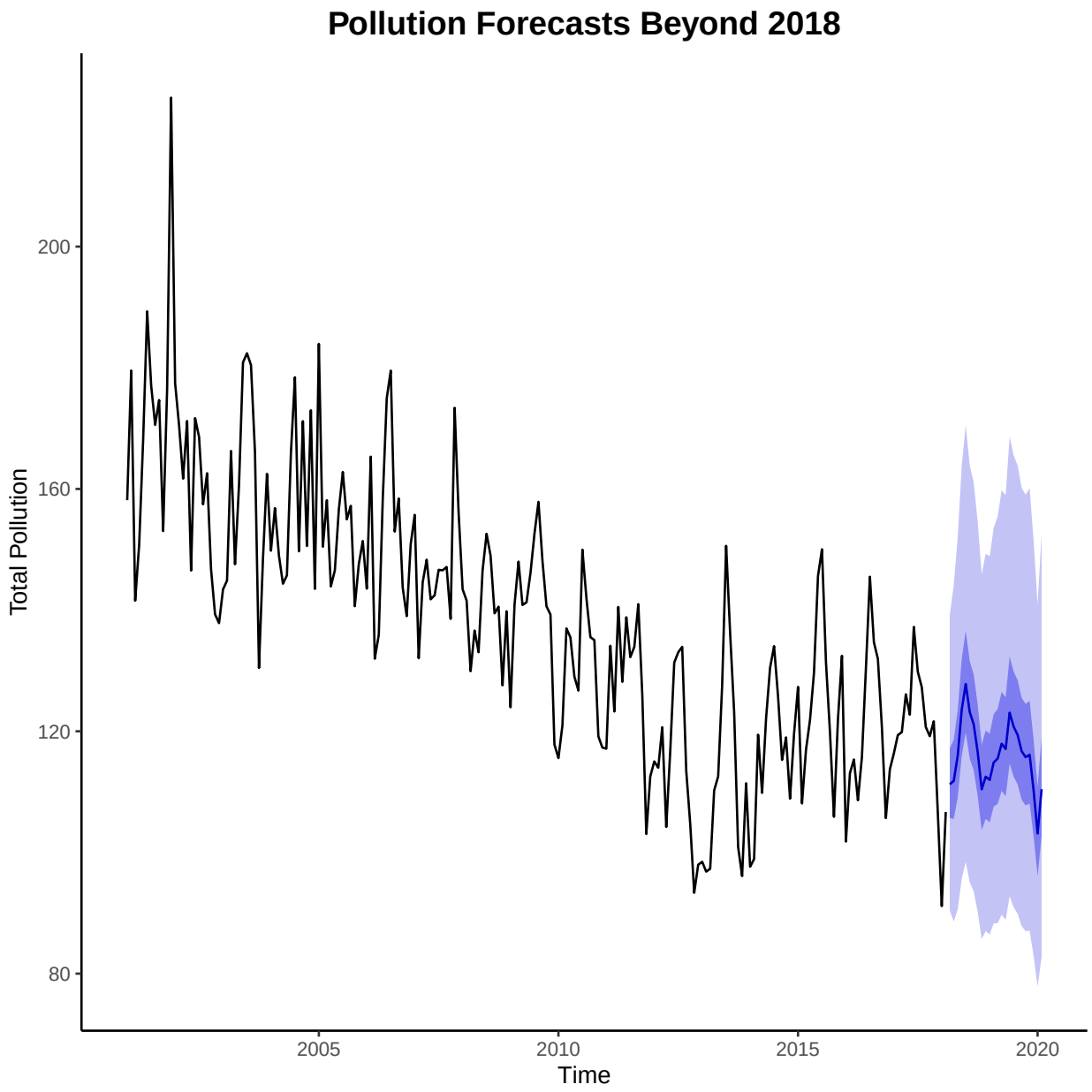

Figure 14: Forecasting the data using all the data as the training set

¡¡Q5a¿¿ set.seed(001432758)
arma.sim ¡- arima.sim(model = list(ar=c(0.65), ma=c(0.45)),n=100) arma.sim
acf(arma.sim)$acf[2]$$phi$
From Figure 13 we can see that the model produced for 2014 to 2018 that the test data is within in the 80% confidence interval of the predicted model but overall the model tends to underestimate the data. This is largely due to the plateau effect that is seen to begin around 2013 where pollution levels hover around 130. This plateau is largely in the test set

so it is not taken into account when calculating the coefficients from the training set. This problem is avoided when we plot the entire model using the entire dataset to plot and create the coefficients for model as the plateau is taken into account. We can see that in Figure 14 where the mean of the plotted data is around 120 with a slight linear decrease.

6 Conclusion

Based on our analysis above, the first step in choosing our model, and eventually forecasting future values, was ensuring our model was stationary. Our initial time series was clearly not stationary, based on it's changing mean, seasonality, and it's inconstant variance. We eventually achieved stationarity by transforming the series and also differencing. With the help of PACF and ACF plots, and AIC we were able to conduct model selection, and ultimately concluded that a seasonal ARIMA(5,1,1)(2,0,0)[12] would do a good job in modelling our monthly average values of air pollutants in Madrid.

After choosing our model, we then moved on to the forecasting portion of our analysis. We used a 75% training set and 25% test set to train and test our model, and ultimately we found that our model slightly underestimated the true data. This could be due to the fact that our data has variables that may be affected by other things that are unexplained by our model, and therefore are not captured by the model. Overall, we think that such research and analysis on similar topics concerning the environment and air quality levels are important given the current climate we live in and the issues we face surrounding the degradation of our air quality and environment. Future analysis on this time series could include a deeper dive into the seasonality we saw in the summer months, and possibly examine the pollutants and their effects on the air quality individually, rather than as a total.

A Appendix

SO_2: sulphur dioxide level measured in $\mu g/m^3$. High levels of sulphur dioxide can produce irritation in the skin and membranes, and worsen asthma or heart diseases in sensitive groups.
CO: carbon monoxide level measured in mg/m^3. Carbon monoxide poisoning involves headaches, dizziness and confusion in short exposures and can result in loss of consciousness, arrhythmias, seizures or even death in the long term.
NO_2: nitrogen dioxide level measured in $\mu g/m^3$. Long-term exposure is a cause of chronic lung diseases, and are harmful for the vegetation.
$PM10$: particles smaller than 10 μm. Even though the cannot penetrate the alveolus, they can still penetrate through the lungs and affect other organs. Long term exposure can result in lung cancer and cardiovascular complications.
O_3: ozone level measured in $\mu g/m^3$. High levels can produce asthma, bronchytis or other chronic pulmonary diseases in sensitive groups or outdoor workers.

TOL: toluene (methylbenzene) level measured in $\mu g/m^3$. Long-term exposure to this substance (present in tobacco smkoke as well) can result in kidney complications or permanent brain damage.
BEN: benzene level measured in $\mu g/m^3$. Benzene is a eye and skin irritant, and long exposures may result in several types of cancer, leukaemia and anaemias. Benzene is considered a group 1 carcinogenic to humans by the IARC.
EBE: ethylbenzene level measured in $\mu g/m^3$. Long term exposure can cause hearing or kidney problems and the IARC has concluded that long-term exposure can produce cancer.
TCH: total hydrocarbons level measured in mg/m^3. This group of substances can be responsible of different blood, immune system, liver, spleen, kidneys or lung diseases.
$NMHC$: non-methane hydrocarbons (volatile organic compounds) level measured in mg/m^3. Long exposure to some of these substances can result in damage to the liver, kidney, and central nervous system. Some of them are suspected to cause cancer in humans
[2]

References

[1] PABLO LEÓN, *At-risk madrid central scheme sees pollution fall to "historic" lows*, 2019.

[2] Decide Soluciones, *Air quality in madrid (2001-2018)*, 2018.

[3] World Health Organization, *Air pollution*, 2020.