The War on Air Pollution

Time Series Analysis

Juliet Maharaj

CUNY Hunter College

18 May 2021

In this report, we will analyze a time series dataset on air quality levels. First, we will plot the dataset to visualize the observations. Next, we will make the dataset stationary in order to create a model. Once we find the optimal parameters, we will create a model for the dataset. Once we have a model for the dataset, we can make predictions and forecast.

I. Introduction

Air pollution is defined as the presence of solid and gas particles in the atmosphere that can be dangerous to the health of humans and other living organisms. Air pollution can also cause detrimental damage to the environment. In 1955, The Air Pollution Control Act of 1955 was passed, becoming the first federal air pollution legislation. The Air Pollution Control Act was created to fund research for the sources and analysis of air pollution. One of the many reasons why Congress felt the need to create air quality standards is because of a town called Donora in Pennsylvania. In October 1948, over the span of five days, almost half of the town's 14,000 residents experienced severe respiratory or cardiovascular problems. Steel and zinc smelters had left the town with polluted air that had nowhere to escape. This was an extreme event, but it exhibited the newly rising issue of air quality in the United States. As a result, in 1970, a milestone year, Congress passed the Clean Air Act Amendments which led to the establishment of national air quality standards. The objective for these amendments is to establish a new source of performance standards for new and modified stationary sources, to establish a National Emission Standards for Hazardous Air Pollutants and to establish requirements for control of motor vehicle emissions.

The dataset being used contains 9358 observations gathered from March 2004 to April 2005. The observations gathered were an array of five metal oxide chemical sensors embedded in an Air Quality Chemical Multisensory Device. The device was placed in a field, in Italy, known for being significantly polluted. In this report, we will focus on the carbon monoxide level on Earth. The Air Quality Chemical Multisensory Device records the true hourly averaged concentration of carbon monoxide in mg/m3 (one milligram per cubic meter). For reference, 1000 mg/m3 is equal to 1 ppm (parts per million). This dataset is the longest freely available recording of a field deployed air quality chemical sensor devices responses (De Vito).

Air pollution consists of multiple different components, such as gasses, particulates, and other biological molecules. Some air pollutants can be toxic to humans and/or other living organisms on Earth. According to the U.S. National Ambient Air Quality Standards, the highest level of carbon monoxide that can be present in outdoor air is 9 ppm (40,000 micrograms per meter cubed) for 8 hours, and 35 ppm for 1 hour (EPA). These are maximum levels and should only be encountered once per year. The average level of carbon monoxide in homes without gas stoves vary from 0.5 to 5 parts per million (ppm). Carbon monoxide levels near properly adjusted gas stoves are often 5 to 15 ppm and those near poorly adjusted stoves may be 30 ppm or higher (EPA). The Air Quality Chemical Multisensory Device used to track the level of carbon monoxide was placed outdoors. We will use the EPA's maximum carbon monoxide levels to compare and analyze our values. One question that we can keep in mind throughout our analysis is whether or not air quality is improving or worsening given the applied regulations.

II. Model Specification

To start in the time series analysis of air quality, we can examine the correlogram of the observations from the dataset. Figure 2.1 is visually very dense. The correlogram is visually dense because the observations were recorded hourly, and that creates a lot of observations to include in a

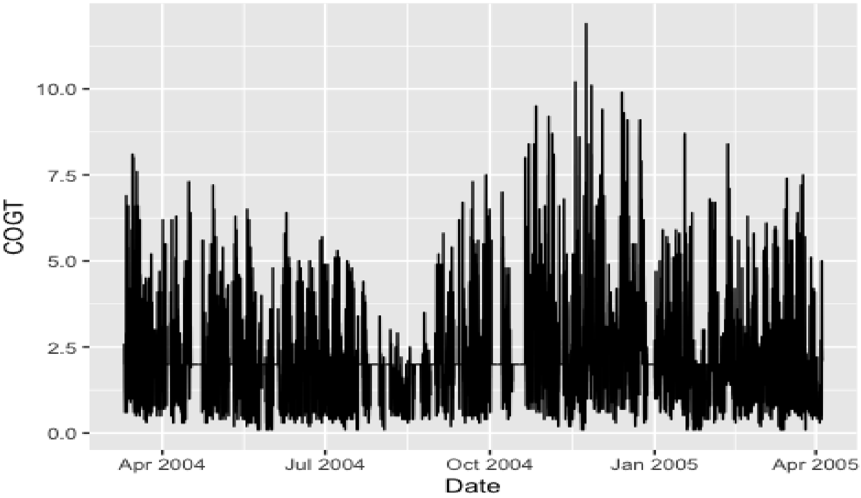


Figure 2.1: The hourly observed values of carbon monoxide in the air from March 2004 to April 2005. CO(GT) levels recorded in parts per million (ppm).

plot. The range of the carbon monoxide levels in our dataset goes from 0.100 to 11.900 being the maximum value observed. From table 2.1, we can see that 75% of the carbon monoxide observation values are below 2.600. We know that 2.600 ppm is a realistic level of carbon monoxide for the air on Earth, being that 9 ppm is the maximum level. We can say, that only about 25% of our observations are between 2.600-11.900 ppm. By Figure 2.1, we can visually see that our dataset is not stationary. There is no constant mean. There is no constant variance either. If we look at the plot after October 2004, the data is sparser than after January 2005. This could be a result of missing observations from the dataset being replaced with the mean value. In Figure 2.2, we cut a portion of the correlogram to analyze further. The plot ranges from March 11th to March 19th. We can observe in Figure 2.2 that the data is seasonal on a daily basis where carbon monoxide levels rise throughout the day and fall at night. During the week of March 11th to March 19th, the level of carbon monoxide reaches up to 7 and 8 ppm. We know, proven from the EPA's research, that most people should only encounter levels this high once per year. Although these high levels are only present during the peak of the day, they are still dangerous levels of carbon monoxide for people to be inhaling

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   1.200   2.000   2.127   2.600 11.900
```

Table 2.1: The interquartile range of the air quality dataset.

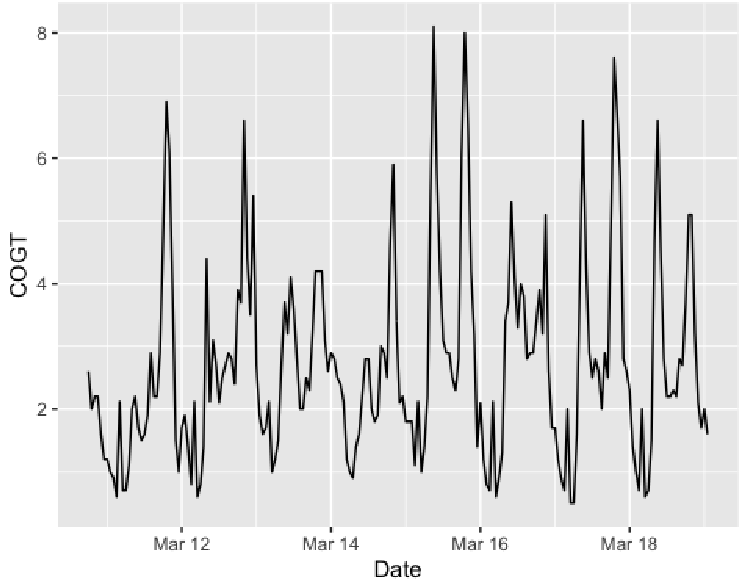
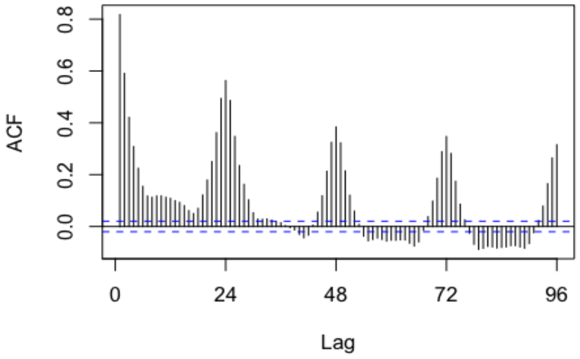


Figure 2.2: A portion of Figure 2.1, specifically the dates range from March 11th, 2005, to March 19th, 2005.

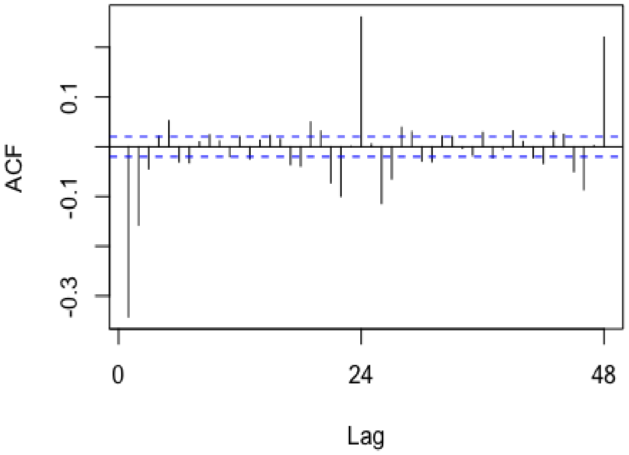
Figure 2.3: The ACF plot of the dataset. Every 24 lags represents a new day.



daily. Also, according to Figure 2.1, March 2005 was not even the month with the highest record of carbon monoxide levels. Other months on the plot look like they have experienced even higher levels of carbon monoxide. Visually, from Figure 2.1, it looks like November 2004 has faced some of the highest carbon monoxide levels, reaching near 11 ppm.

Next, we will examine the autocorrelation function of the dataset in Figure 2.3. The time lag is hourly. We can see a cycle that agrees with our comment that carbon monoxide levels rise throughout the day and fall at night. This supplements the fact that the time series is not stationary. The ACF plot ranges over three days so that we can see the cycle is continuous daily, every 24 lags the cycle generally repeats. The PACF plot portrays the same concept, that the carbon monoxide levels have daily seasonality.

Figure 2.4: The ACF plot of the dataset after being 24-hour seasonally differenced.



In order to create a model for our dataset, we must start by making the dataset stationary by differencing. First, we will try by taking a twenty-four-hour seasonal difference. Differencing can help stabilize the mean of the dataset by removing changes in the levels of the dataset, in return eliminating or reducing the trend and seasonality. Seasonality causes the series to be nonstationary because the average values at some particular times within the seasonal span may be different than the average values at other times. In Figure 2.4, we can see the 24-hour seasonal differenced ACF plot of the dataset. We can see that the daily cycle of fluctuating values has been reduced. There are still significant spikes every 24 lags, we will take this into consideration when trying to create a model that fits our time series data. To verify that our time series is stationary, we ran an Augmented Dickey Fuller Test on our 24-hour seasonally differenced data. The results are shown in Table 2.2.

```
## Augmented Dickey-Fuller Test
##
## data:  dcogt
## Dickey-Fuller = -21.455, Lag order = 72, p-value = 0.01
## alternative hypothesis: stationary
```

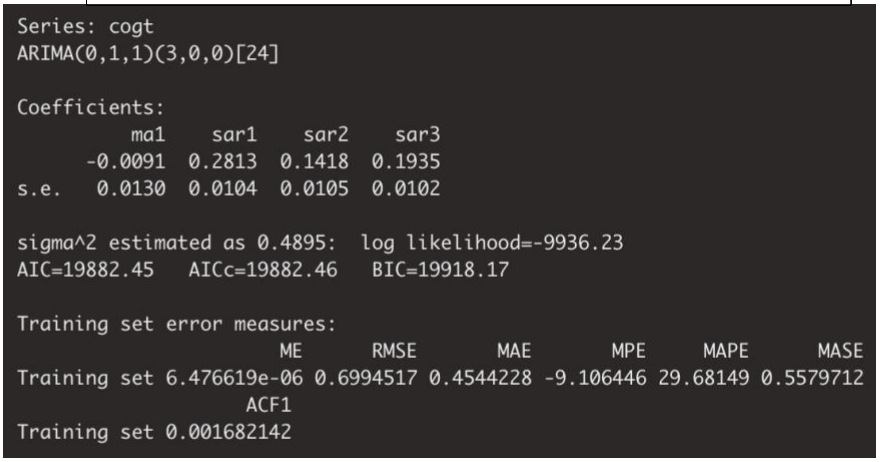
Table 2.2: The Augmented Dickey Fuller test on the 24-hour seasonally differenced data. The ADF states that our time series is now stationary based on its p-value.

To conclude, we can say that our data is stationary after one 24-hour seasonal difference. The data is now ready to be fitted to a model.

III. Model Fitting

Since we had to difference our model to make it stationary, we determined we needed to use an ARIMA model. We started by running an auto ARIMA function to see which values are adequate for our model. The results are shown in Figure 3.1. The first observation made was that the auto ARIMA function did not run a seasonal difference on the data. The AIC of the model is 19,882.45. The standard deviation is 0.4895. This

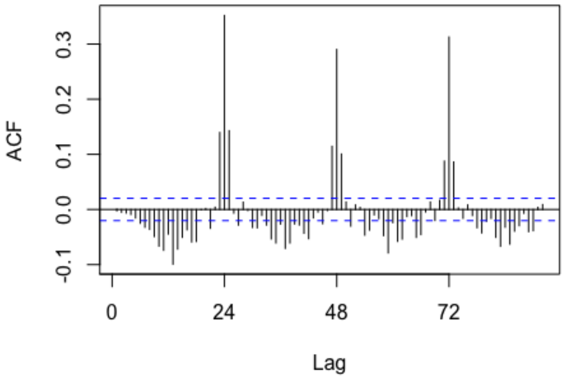
Figure 3.1: The summary output of running an auto ARIMA model.



```
Series: cogt
ARIMA(0,1,1)(3,0,0)[24]

Coefficients:
          ma1     sar1    sar2    sar3
      -0.0091   0.2813  0.1418  0.1935
s.e.   0.0130   0.0104  0.0105  0.0102

sigma^2 estimated as 0.4895:  log likelihood=-9936.23
AIC=19882.45   AICc=19882.46   BIC=19918.17

Training set error measures:
                       ME      RMSE       MAE       MPE     MAPE      MASE
Training set 6.476619e-06 0.6994517 0.4544228 -9.106446 29.68149 0.5579712
                    ACF1
Training set 0.001682142
```

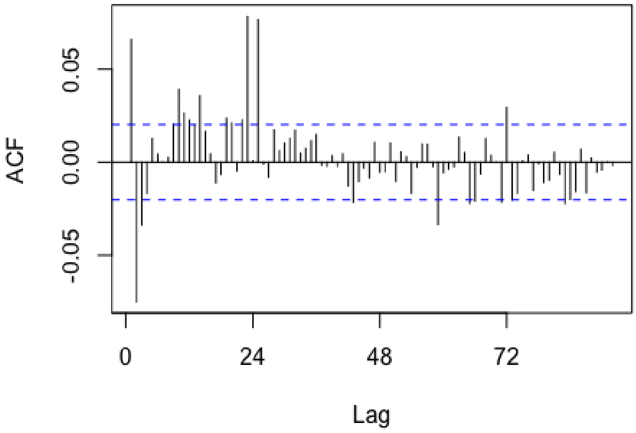
model seems to be significant, but we need accompanying models to compare it to. So, the next model we ran was an ARIMA (0,1,1) (3,1,0) [24]. The notion behind adding one 24-hour seasonal difference to the model we already have is to stabilize the daily seasonality that we established exists. The summary of ARIMA (0,1,1) (3,1,0) [24] produced an AIC of 20,445.52. The lower the AIC the better the model. The RMSE of ARIMA (0,1,1) (3,1,0) [24] is 0.7217506. The lower the root square mean error (RMSE) the better the model. At this point, the best model that we have created is ARIMA (0,1,1) (3,0,0) [24].

Figure 3.2: The ACF plot of the residuals for a ARIMA (1,1,1) (1,1,1) [24



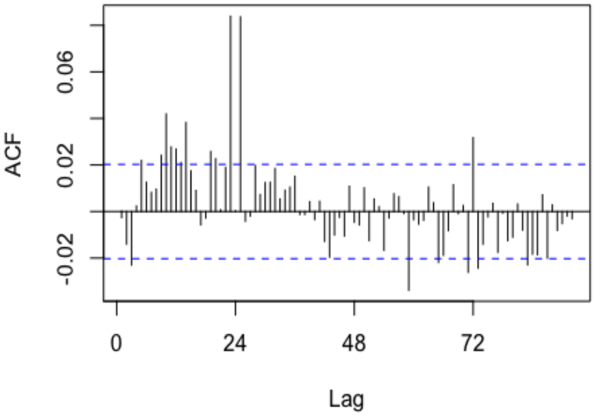
Next, we will create additional models to use as a comparison to our chosen model. We want to find the model with the lowest AIC. The first model created for comparison was an ARIMA (1,1,1) (1,1,1) [24]. The AIC is 18,001.56. This AIC value is lower than both models that we already examined. But, in Figure 3.2, the plot of the ACF of the residuals shows us that the model still has some daily seasonality. So, the next model that was created was ARIMA (2,1,2) (1,1,1) [24]. The AIC is 17,916.03, the lowest AIC yet. The RMSE is 0.6286638, which is significantly lower than ARIMA (0,1,1) (3,1,0) [24]. To the left, in Figure 3.3, is the ACF plot of the

Figure 3.3: The ACF plot of the residuals for ARIMA (2,1,2) (1,1,1) [24].



residuals of ARIMA (2,1,2) (1,1,1) [24]. Initially, we considered the ACF plot in Figure 3.3 to not be white noise. But it could be white noise because the ACF values only range from 0.05 to -0.05, meaning that most of the data is actually between the blue lines. A few of the lags cross over the blue lines but that does not automatically imply that the model is not white noise.

Figure 3.4: The ACF plot of the residuals for ARIMA (2,1,2) (2,1,1) [24].



The next model ran was ARIMA (2,1,2) (2,1,1) [24]. The AIC is 17,918.02, only a few values off from the previous model. The RMSE is 0.6286648. The ACF of the residuals for ARIMA (2,1,2) (2,1,1) [24] in plotted in Figure 3.4. The ACF plot of the residuals in Figure 3.4 is visually very similar to ARIMA (2,1,2) (1,1,1) [24]. There are a few lags that go out of the blue line boundary, but that does not necessarily imply that the model is not white noise.

One of the final models that was created was ARIMA (4,0,4) (3,0,0) [24]. The AIC is 18689.61 the RMSE is 0.6556302 In Figure 3.5, we can see the residuals from this model. The residuals for this model are free from deterministic trends. The residuals seem to be constant and normally distributed with the exception of a few outliers.

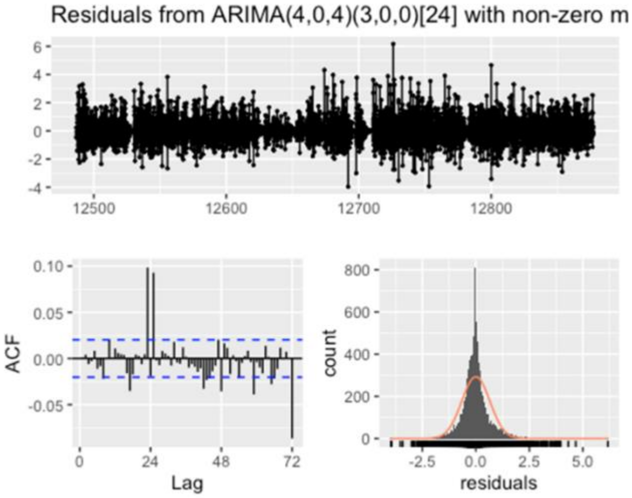


Figure 3.5: The plot of the residuals from the ARIMA (4,0,4) (3,0,0) [24].

IV. Model Diagnostics

At this point in our analysis, we are in between two models, ARIMA (2,1,2) (1,1,1) [24] and ARIMA (2,1,2) (2,1,1) [24]. In Figure 4.1, we can see the residuals for ARIMA (2,1,2) (1,1,1) [24]. The standardized residuals do not show a trend, there are a few outliers, but nothing too significant. In general, no changing variance across time, the plot looks visually dense throughout. The ACF plot has a few significant lags, but we cannot say that the model is not white noise just by visually looking at the ACF plot.

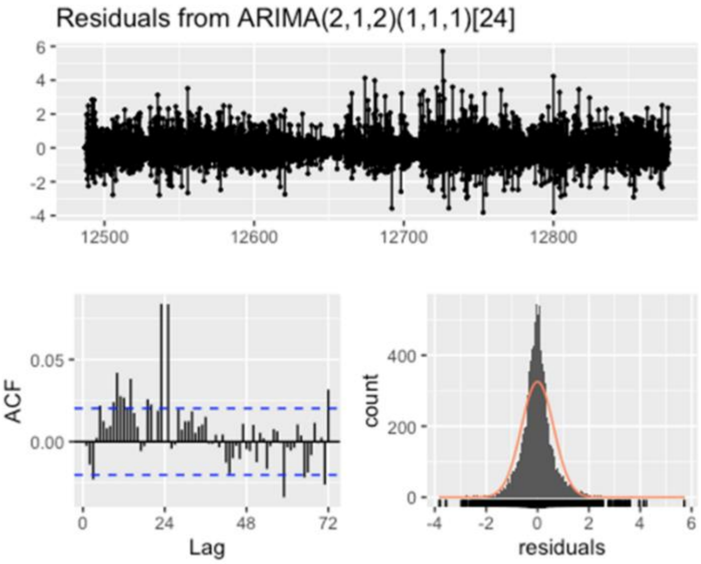
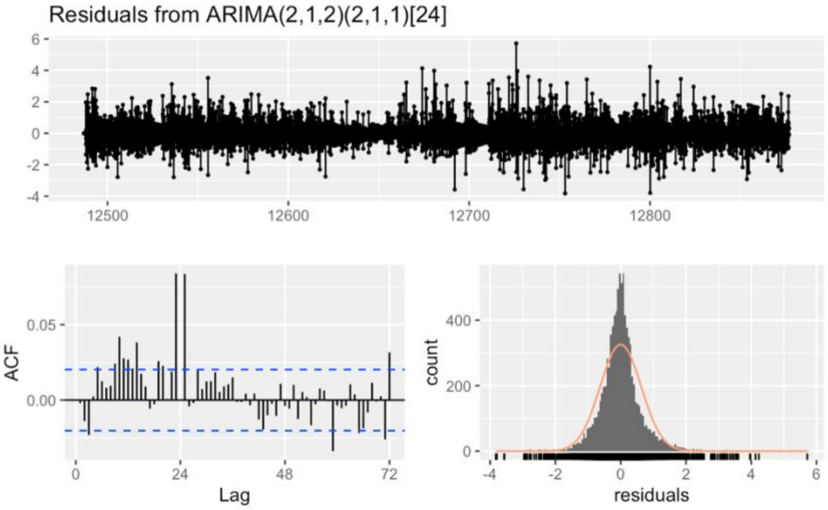


Figure 4.1: The plot of the residuals for ARIMA (2,1,2) (1,1,1) [24].

The residuals for ARIMA (2,1,2) (2,1,1) [24] are very similar to ARIMA (2,1,2) (1,1,1) [24]. The residuals are normally distributed with a few exceptions. And we have an ACF plot that does not directly imply that the model is white noise. It makes sense as to why the models are so similar, because they are only one term different.

Figure 4.2: The plot of the residuals for ARIMA (2,1,2) (2,1,1) [24].



In the end, the model chosen as the best fit for this time series data is the ARIMA (2,1,2) (1,1,1) [24]. We chose this model because it has the lowest AIC.

The equation for ARIMA (2,1,2) (1,1,1) [24] is (1-0.4386 B)(1-0.2446 B)(1-0.1321 B^24)(1-B)(1-B^24)Yt = (1-0.5721B)(1-0.4255B)(1-0.9477B^24)e^t. To confirm that we had chosen the correct model, we ran one last test. The Box-Ljung test states the null hypothesis to be, our model does not show lack of fit and the alternative to be that the model does show a lack of fit. A significant p-value in this test rejects the null hypothesis that the time series isn't autocorrelated. In Table 4.1, we can see the results of the Box-Ljung test for ARIMA (2,1,2) (1,1,1) [24], our p-value is significant. We can conclude that the ARIMA (2,1,2) (1,1,1) [24] is the best fit for this time series dataset.

Table 4.1

```
   Box-Ljung test
##
## data:  model6$residuals
## X-squared = 0.021043, df = 1, p-value = 0.8847
```

V. Conclusion

To finalize that the model we created is best, we forecasted. Forecasting is a technique that uses historical data as inputs to make informed estimates that are predictive in determining the direction of future trends. In Figure 5.1, we see a plot similar to the initial correlogram of the time series. On the right hand side, you can see that the plot turns blue, that is the forecasted portion of the dataset. We only forecasted an additional day, because forecasting too far ahead can lead to unrealistic results. Our forecasted portion of the plot visually looks like it belongs with the rest of the data. This tells us that the ARIMA model we picked is a good model to use for future analysis.

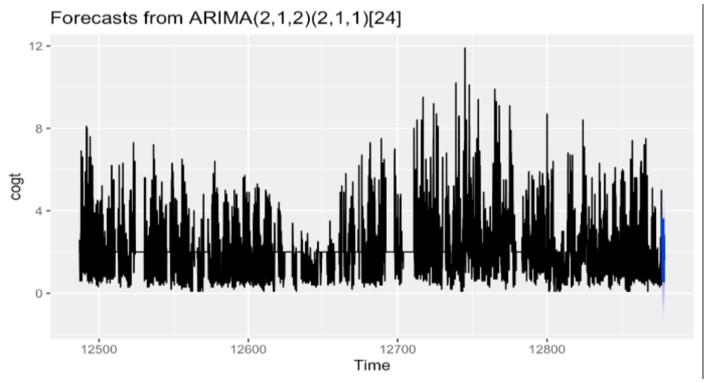


Figure 5.1: The plot of the ARIMA model forecasted.

In the end, it is essential to monitor the level of carbon monoxide in the air. When we breathe, these pollutants get into our lungs. They can cause severe health problems such as asthma, cardiovascular diseases and even cancer and they reduce the quality and number of years of our lives.

Is air quality actually improving because of the regulations put into order? The answer to our question is to be determined. We would need more data to determine whether or not the air quality is improving. We would need air quality levels before and after the regulations to compare and analyze. Research of this nature is imperative for all living organisms, in order to improve the air quality instead of letting it worsen. The future of our air quality relies on how we preserve it today.

References

S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical, Volume 129, Issue 2, 22 February 2008, Pages 750-757, ISSN 0925-4005

"What Is the Average Level of Carbon Monoxide in Homes?" EPA, Environmental Protection Agency, 1 Aug. 2019, www.epa.gov/indoor-air-quality-iaq/what-average-level-carbon-monoxide-homes.

"NAAQS Table." EPA, Environmental Protection Agency, 10 Feb. 2021, www.epa.gov/criteria-air-pollutants/naaqs-table.