**US Air Pollution Analysis**

***Introduction***

Earth’s atmosphere is primarily composed of Nitrogen, Oxygen, Argon, and Carbon Dioxide. Pollutants can enter the Earth's atmosphere in many different ways and damage the quality of air, water and land. The World Health Organization classifies harmful pollutants like ground-level ozone, sulfur oxides, carbon monoxide, nitrogen oxides, and lead.

***Problem***

Air pollution has been a leading concern in many areas all around the world. The new civilized world has many different emissions sources such as cars, planes, and industrial processes that have led to serious complications in human health and the environment. Pollutants such as carbon dioxide and methane are raising the earth’s temperature and have played a large role in climate change. Due to higher temperatures, smog, another form of air pollution, has increased throughout different countries and led to greater amounts of ultraviolet radiation. Exposure to pollutants also has toxicological impacts on human health including cardiovascular and respiratory diseases. High levels of air pollution have been considered a risk factor in the development of many diseases such as asthma, lung cancer, Alzheimers and Parkinsons disease. It is essential that we understand the impacts of air pollution and focus our efforts on reducing the production of harmful gasses for the wellbeing of ourselves and our community.

In this project, we will analyze the US Air Pollution data, focusing specifically on different regions within the United States. Our goal is to compare the changing air quality of Florida and Illinois. Our method will consist of finding the most appropriate forecasting model for this data by comparing applicable models. We will take model fit and model accuracy into consideration when finding the best model as well as provide a statistical evaluation of the chosen model. To understand the future trends, we will forecast the air quality measurements for the next 4 years and compare them to the original trends.

***Data Collection***

A data source for this project has been found from Kaggle (<https://www.kaggle.com/sogun3/uspollution>), which was originally scraped from the database of the United States Environmental Protection Agency. Since the original dataset downloaded from Kaggle was not able to fit in one single csv file, we used Tableau Prep to work with the dataset. First, we divided the dataset into three separate files by grouping 3-4 years together. Then we applied a filter to the ‘State’ column to grab only Florida and Illinois, and created a new column ‘YearMonth’ in ‘YYYY/MM’ format that grabs only year and month from the ‘Date Local’ column. By applying the state filter and aggregating the data by only Year and Month, we were able to reduce the size of the data drastically. All the measurements were averaged when we generated the aggregated data. (maybe a screenshot of Tableau Prep here)

The final dataset that’s going to be used for this project has 29 columns in total and 1771 rows. There are 16 decimal columns and 13 string columns (including the ‘YearMonth’ column). Observations on 4 pollutants (NO2, CO, SO2, and O3) are included with 4 columns each, which are the hour when the highest value for the day was taken (named ‘1st Max Hour’), the hour when the highest value for the day (named ‘1st Max Value’), the average value for the day (named ‘Mean’), and Air Quality Index for the day for the pollutant (named ‘AQI’). The definitions of each measurement were found in EPA AirData Documentation (<https://aqs.epa.gov/aqsweb/airdata/FileFormats.html>).

We are only using the ‘State’ column without ‘County’ and ‘City’ columns for a clearer analysis of each state rather than focusing on too granular region level. Out of the CO, NO2, SO2, and O3 compounds in this dataset, we will be focusing on the Air Quality Index columns of each compound. The reason for this selection is that 1st Max Value and Mean values were not only highly correlated to each other but also showed a straight linear line when plotted with AQI, indicating they may be a part of measuring AQI. To avoid any further confusion and clarify the analysis process, we thought focusing on index measures would be the best idea with the given data set.

Once we find the best model that fits the data the most for Florida and Illinois respectively, we are going to forecast NO2 Air Quality Index values of each state from 2016 to 2018 by using SO2, CO, and O3 Air Quality Indexes.

***Data Analysis and Results***

We looked at Florida and Illinois data separately and found the best model for each data set individually.

Time series plots and ACF plots for Florida look like below:

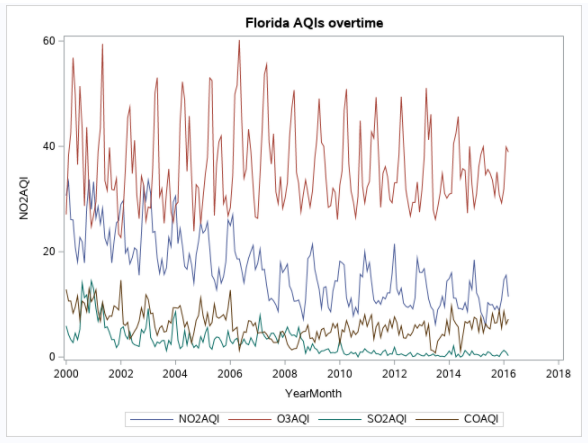


Figure Time Series Plot of Florida AQIs

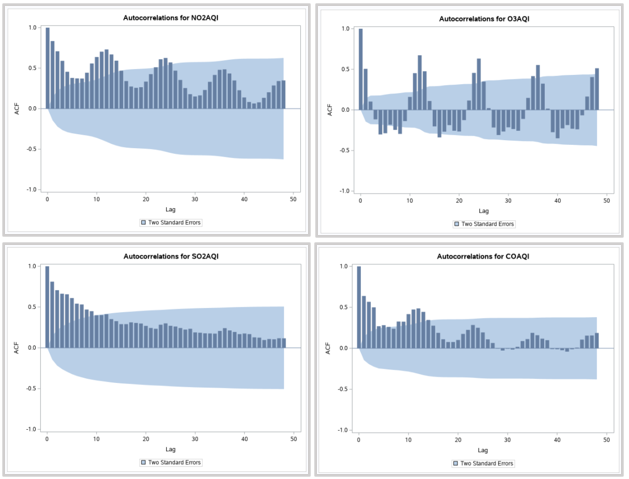


Figure ACF Plot of Florida AQIs

Since Florida data shows a trend and seasonality component, the following models were applicable.

According to the scatter plots and ACF plots, the NO2AQI and COAQI had a trend component. O3AQI has a seasonality component but not a trend and SO2AQI does not have seasonality but does have a trend component. Before creating the multiple regression model, we checked if the independent variables (x axis) have linear relationships with the dependent variable (y axis).

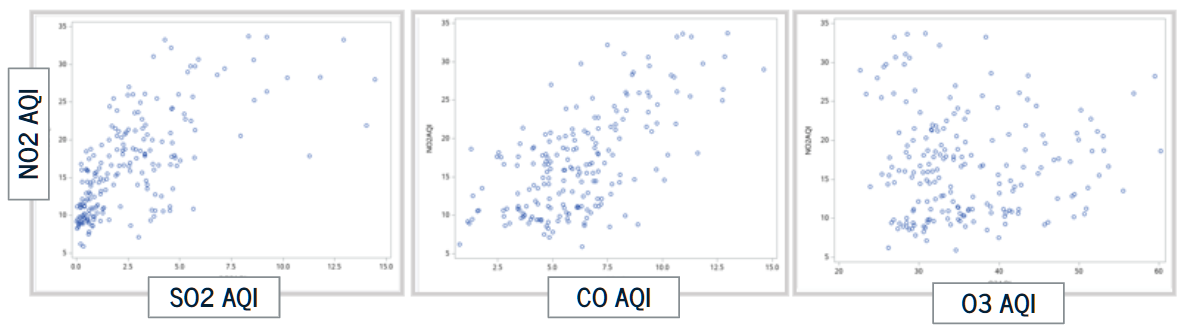


Figure Florida Relationships between variables

From the graphs above, we can see that O3AQI has a nonlinear relationship with NO2AQI. It is unclear whether SO2AQI and COAQI have a linear relationship with NO2AQI or not. Therefore we have decided to test both their linearity and nonlinearity in the models below.

The first model used was multiple regression.

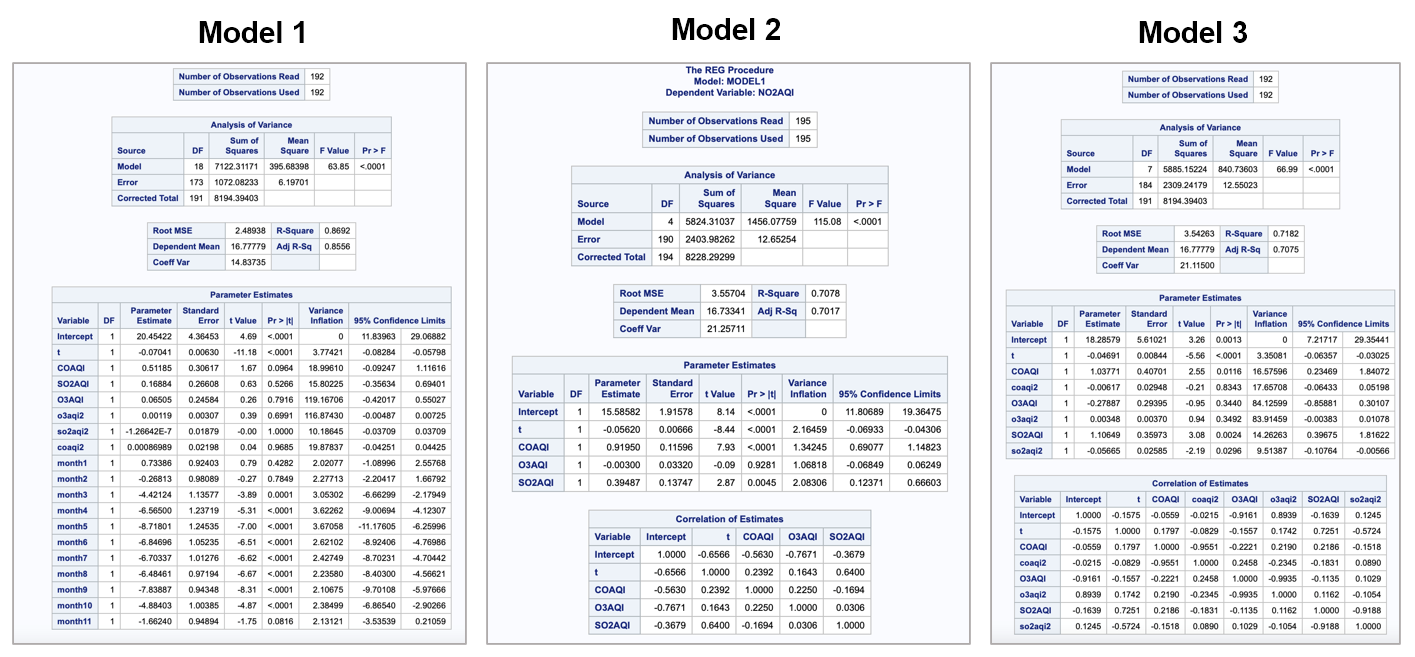


Figure Florida Regression Model Results

Model 1 accounts for the seasonality component by creating 11 dummy variables and a reference variable. In this model, month 12 was used as the reference variable and all other months were considered the dummy variables. Since O3AQI showed a nonlinear trend, we converted this to a linear model and used a new variable when computing for multiple regression. This was also accounted for in model 1. To compare the linear model with the non linear model, we created a model 2 which shows multiple regression with the original o3aqi values. For further analysis, a model 3 was created without dummy variables to compare to model 1 which used dummy variables and accounted for seasonality.

To find the best model among these three, the AIC, BIC, Adjusted r-squared values, mape fit and accuracy were compared.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Model 1** | **Model 2** | **Model 3** |
| **AIC** | 366.9 | 501.4 | 499.8 |
| **BIC** | 372.1 | 503.8 | 502.07 |
| **ADJ-RSQ** | 85.7% | 70.08% | 70.17% |
| **MAPE Fit** | 11.70% | 18.10% | 17.97% |
| **MAPE Accuracy** | 18.14% | 18.42% | 19.24% |

**Model 1** has the lowest AIC and BIC and has the highest Adjusted R-squared value, making it the best model. It should also be noted that model 1 also has the lowest mape fit and accuracy which contributes to choosing this model over the others. This makes sense logically since model 1 included the dummy variables which accounted for the seasonality in the data.

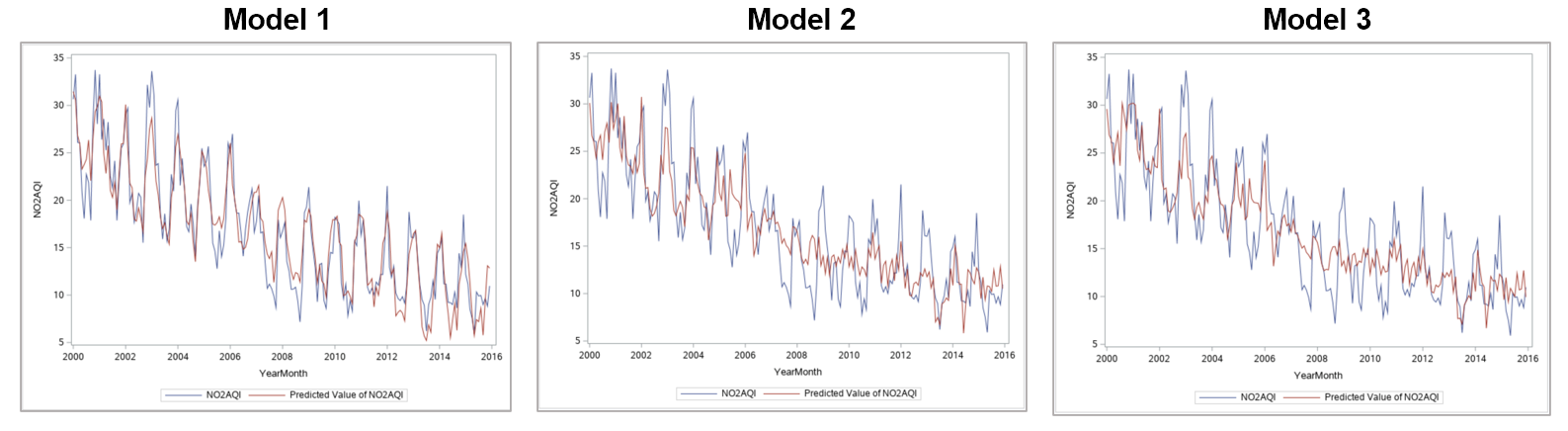


Figure Florida Original vs. Forecasted of Regression models

Since model 1 was chosen to be the best one, a statistical evaluation of model 1 was completed to understand the validity of the model. We found that the p value of COAQI, COAQI2, SO2AQI, SOAQI2, O3AQI, O3AQI2, Month 1 and Month 2 all were greater than alpha and therefore deemed to be significant. This shows that the slope of the mentioned variables is significant. The p value of the F test is also greater than alpha which shows that the model is significant. Looking at the adjusted r-squared value, it is clear that 85% or the variation in the dependent variable is explained by variation in the independent variables. This is a high percentage, further emphasizing the significance of the model. To determine if there was multicollinearity within the variables, we looked at the variance inflation factor. It was found that COAQI, COAQI2, SO2AQI, SOAQI2, O3AQI, O3AQI2 all have multicollinearity since they have a VIF of greater than 10.

Assumption testing was also done for model 1 to determine if we can correctly draw conclusions from the results. First, we checked for the normality assumption as shown in Figure 21. We found that the plot of the residuals showed a normal bell curve indicating that this assumption is true. Next, we checked for the constant variance assumption as shown in Figure 21. The predicted value vs residuals graph showed relatively random values indicating that this assumption is also true. Finally, we checked for the independent assumption. For this, we looked at the Durbin Watson test as shown in Figure 22, we found that the value for pr < DW was less than alpha showing that there is positive autocorrelation. However, the pr > DW value was greater than alpha showing that there was no evidence of negative autocorrelation. Since there was evidence of positive autocorrelation, we can say that the independent assumption is not true.

The second model used was Seasonal ARIMA. Since there are seasonal components in the data, seasonal differencing was performed. This data however still had a trend so regular differencing was performed after to make the data stationary. The following show the results of the ARIMA procedure:

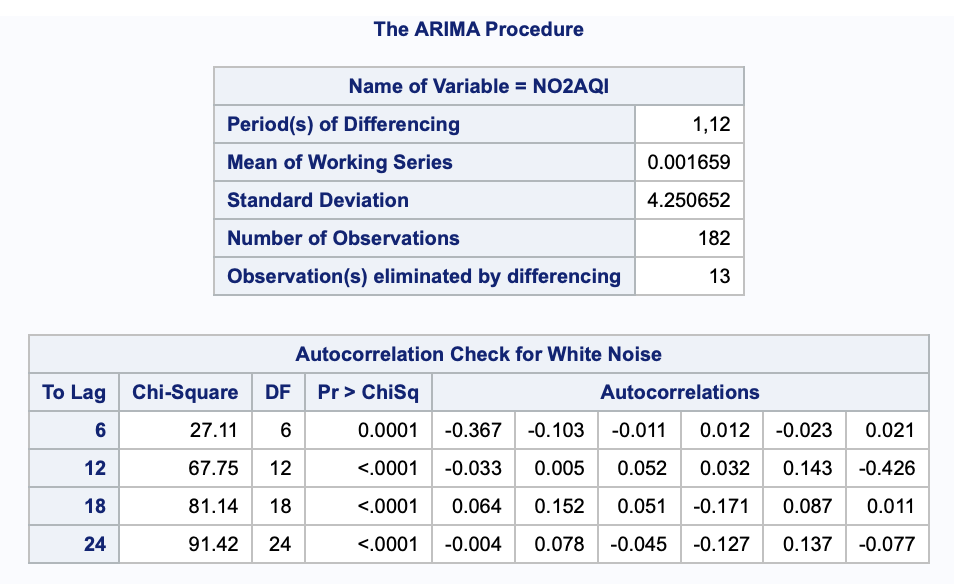


Figure Florida ARIMA 1

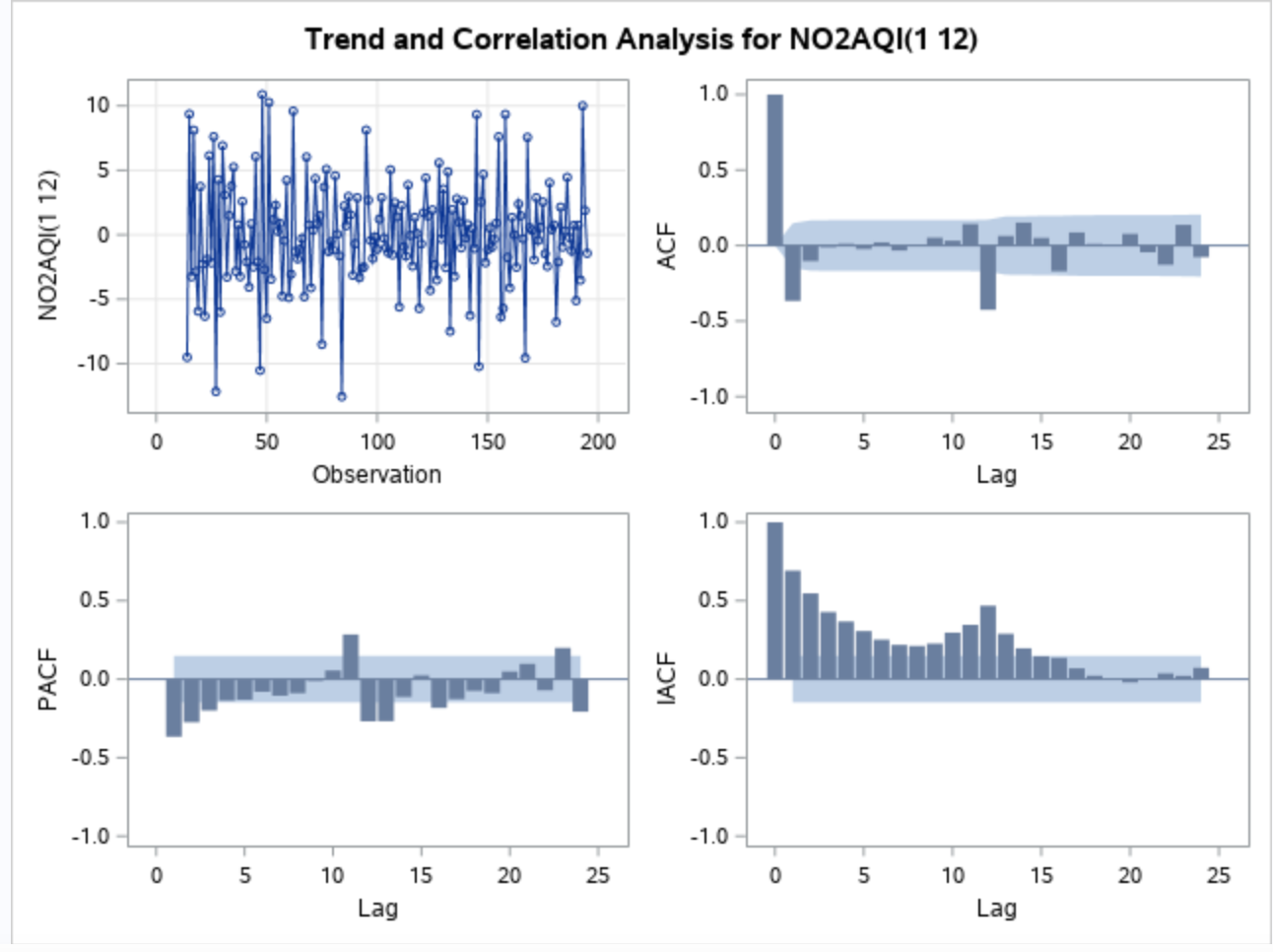


Figure Florida ARIMA 2

This data showed no white noise and was stationary which allowed us to move onto the next step. We came up with a few ARIMA models to try. These include ARIMA ( 0,1,1) (2,1,1)12

ARIMA (0,1,1) (1,1,1)12, ARIMA (1,1,1)(2,1,1)12.

According to Figure 23 for the model ARIMA(0,1,1) (2,1,1)12, the p values for the residuals are greater than alpha showing that the residuals are white noise and the model is adequate. By looking at the Figure 24 for the results for the model ARIMA (0,1,1) (1,1,1)12, it is clear that the residuals are greater than alpha, indicating that the data is white noise and showing that the model is adequate. According to Figure 25 for the model (1,1,1)(2,1,1)12, this model shows that the residuals are greater than alpha and the data is white noise.

To compare these three models and find the best match for this data, we compared the AIC and BIC. Below is a table showing these comparisons.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ARIMA(0,1,1) (2,1,1)** | **ARIMA(0,1,1) (1,1,1)** | **ARIMA (1,1,1)(2,1,1)** |
| **AIC** | 734.09 | 732.44 | 735.38 |
| **BIC** | 748.9 | 744.29 | 753.16 |
| **MAPE Fit** | 13.34% | 13.38% | 13.30% |
| **MAPE Accuracy** | 13.12% | 17.17% | 13.03% |

We have plotted the original and forecasted values for each of the models to look at the differences. Below shows a graph of the NO2AQI from 2000 to 2015.

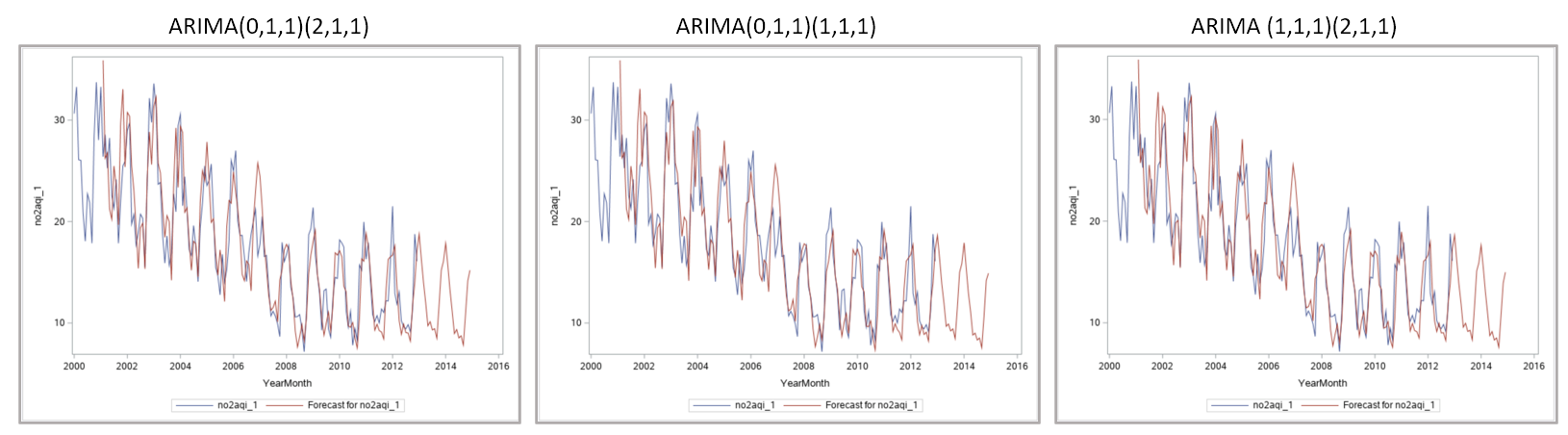


Figure Florida Original vs. Forecasted of ARIMA models

As shown in the above table, there is no drastic difference between the AIC, BIC, Mape fit and mape accuracy of the three models. However, the model ARIMA (0,1,1) (1,1,1)12 has the lowest AIC and BIC compared to the others. The model ARIMA (1,1,1)(2,1,1)12 has the lowest mape fit and accuracy but it can be shown that the difference in the mape fit and accuracy between the models is very small. Therefore, model ARIMA (0,1,1) (1,1,1)12 will be considered as the best model since it is the simplest model and has the lowest AIC and BIC.

Since we looked at both multiple regression models and seasonal arima models for the Florida data, we will compare these models to see which one will be better to forecast with. Out of the multiple regression models, we saw that model 1 was the best one according to the AIC and BIC. Out of the seasonal arima models, we saw that ARIMA(0,1,1)(1,1,1)12 was the best model. Now, we will compare the Mape fit and accuracy of the multiple regression model and the arima model (ARIMA(0,1,1)(1,1,1)12 )). Below is a table showing this comparison.

|  |  |  |
| --- | --- | --- |
|  | ARIMA(0,1,1)(1,1,1) | Multiple Regression Model 1 |
| MAPE Fit | 13.38% | 11.70% |
| MAPE Accuracy | 17.17% | 18.14% |

Looking at the data above, we can see that the multiple regression model has a lower Mape fit and the arima model has a lower mape accuracy. Since we are trying to find the best model to predict forecasts, we will be placing more weight on the Mape accuracy. The Mape accuracy uses the test set to generate forecasts and therefore a lower mape accuracy will show the most accruacte forecasts. Here, we can see that the arima model has a slightly lower mape accuracy than the multiple regression model. For this reason, we have chosen the arima model to be the best fit.

Next, we used this arima model to create the forecasts for the next three years. Below you will find a graph showing the NO2 predictions from 2016-2018.

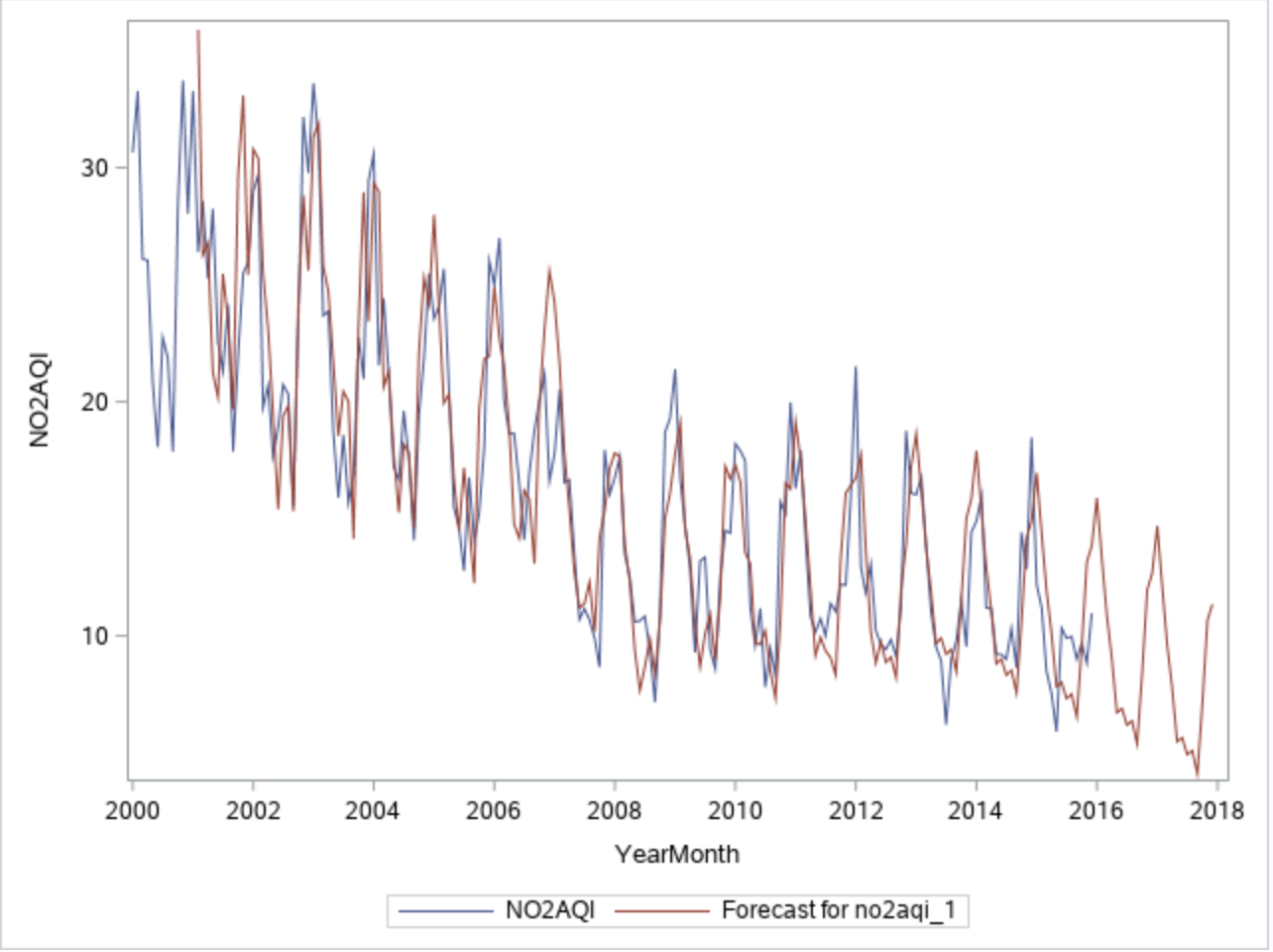


Figure Florida Forecasts with the best model

Let’s look at what Illinois’ data looks like this time. On the Illinois dataset, we eliminated the year 2000 at the beginning and 2016 at the end because these two years didn’t have a full cycle of 12 months. Time series plots and ACF plots for Illinois are shown below:

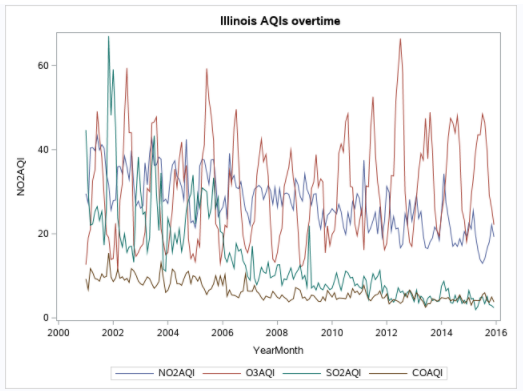


Figure Time Series Plot of Illinois AQIs

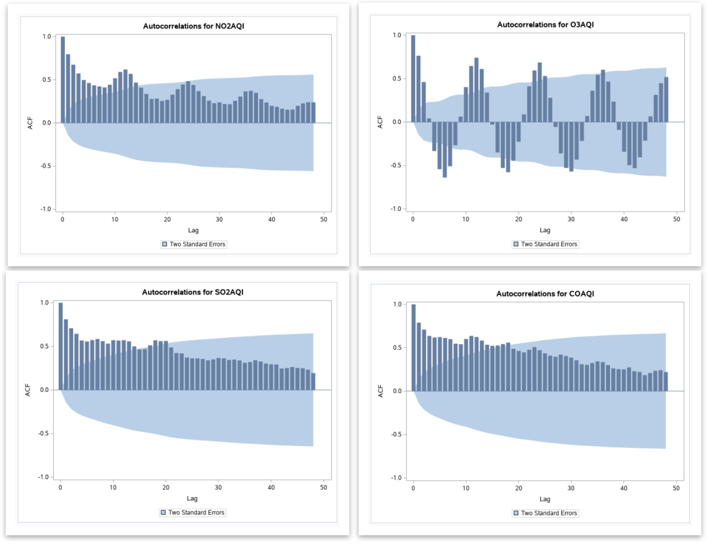


Figure ACF Plots of Illinois AQIs

Since Illinois’ NO2 dataset showed both Seasonality and Trend components, we decided to try Multiple Regression and Seasonal ARIMA models and find the best model that we can use to forecast.

For the Multiple Regression Model, we are having NO2 AQI as a function of O3, SO2, and CO AQIs. We used 11 dummy variables to account for seasonality in this dataset.

Before creating the multiple regression model, we also checked if the independent variables (x- axis) have linear relationships with the dependent variable (y-axis).

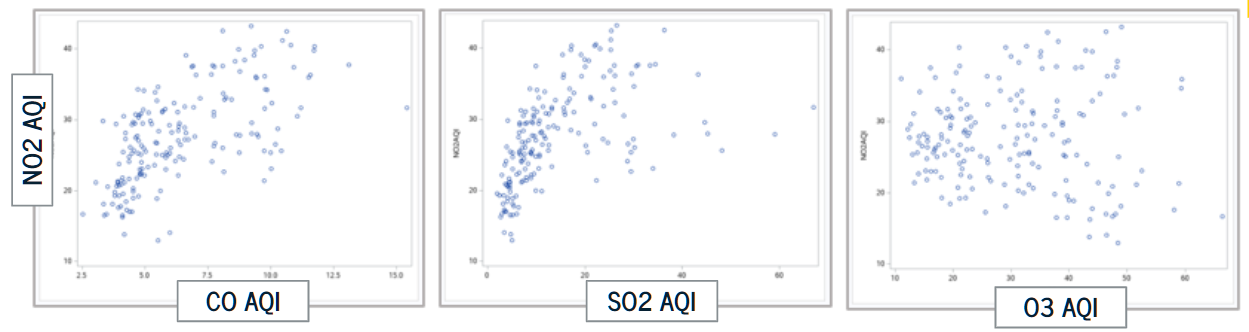


Figure Illinois Relationships between variables

It’s not very clear if SO2 AQI and CO AQI variables have either non-linear or linear relationships with NO2 AQI, our response variable. O3 AQI clearly seems to have a non-linear relationship with NO2 AQI. For this reason, we are going to test non-linearity for all the three independent variables by adding the square of them into a few of the regression models we’re generating.

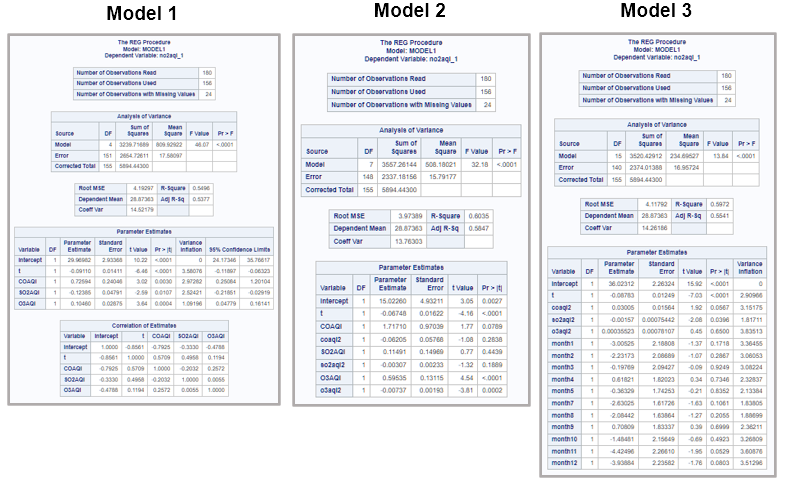
Model 1 includes 4 independent variables, which include the time variable, CO AQI, SO2 AQI, and O3 AQI. Model 2 includes the 7 variables with squared O3 AQI, squared SO2 AQI, squared CO AQI to test for nonlinearity on top of the 4 variables mentioned previously. Model 3 includes the 15 variables by having the 11 dummy variables to account for seasonality and the squared AQIs with the time variable as well. All of the models below are generated on the training sets that are from 2001 to 2013 and test sets that are 2014 and 2015 are set aside for finding accuracy later. 

Figure Illinois Multiple Regression models

Comparison of AIC, BIC, and Adjusted R-Square values of these three models are available in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Model 1** | **Model 2** | **Model 3** |
| **AIC** | 452.14 | 438.27 | 456.71 |
| **BIC** | 454.47 | 441.13 | 462.34 |
| **ADJRSQ** | 53.77% | 58.47% | 55.41% |
| **RMSE** | 4.19 | 3.97 | 4.12 |
| **MAPE Fit** | 12.12% | 11.20% | 11.56% |
| **MAPE Accuracy** | 22.73% | 22.68% | 22.89% |

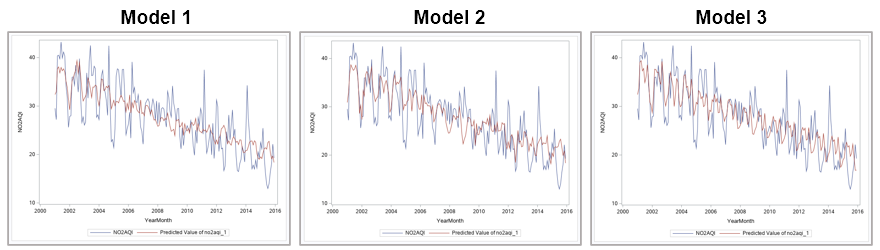


Figure Illinois Original vs. Forecasted of Regression models

Referring to the ACF plot of our response variable NO2 AQI, we found that this variable shows both trend and seasonality components. When we tried multiple regression models with different sets of independent variables, even though Model 2 didn’t account for seasonality by not having dummy variables, its AIC and BIC values were the lowest and its Adjusted R-Square was the highest. Model 3 that accounted for seasonality surprisingly showed the lowest Adjusted R-Square value and the highest AIC and BIC with differences more than 10 compared to those of Model 2. Model 3’s MAPE Fit and Accuracy are not really better than Model 2’s as well. We considered picking Model 3 as the best one since we know the data has seasonality, but it was hard to ignore the fact Model 2 showed a much better performance in terms of AIC, BIC, Adjusted R-Square, MAPE of fit and MAPE of accuracy. For these reasons, we are considering Model 2 as the best Multiple Regression model that fits the dataset the best.

Let’s statistically evaluate this model for its validity.

Model 2’s time-variable (t) coefficient slope is negative, which makes sense because we’ve noticed a decline in NO2 AQIs as time has gone by. Without further knowledge of the scientific mechanism of the Air Quality Index, it’s hard to tell if the negative slope of squared CO AQI, squared SO2 AQI, and squared O3 AQI, and the positive slope of CO AQI, SO2 AQI, and O3 AQI are logical. The p-values of the squared O3 AQI and squared O3 AQIs are less than alpha, meaning their slopes are statistically significant, but the rest of the independent variables’ p-values are bigger than alpha. The p-value of this model is less than 0.0001, indicating this model is statistically significant. The percentage of variation in the dependent variable that can be explained by variation in the independent variables is 58.5%, which is moderate. Lastly, since Model 2 included three squared AQIs as well as the original three AQIs to test their non-linearity, the Variance Inflation Factors (VIF) of all the independent variables were bigger than 10, so this model has independent variables that are highly correlated to each other and there’s multicollinearity (Figure 30).

As the result of non-linearity testing of CO AQI, SO2 AQI, and O3 AQI, the p-values of squared CO AQI and squared SO2 AQI were both bigger than 0.05, indicating CO AQI and SO2 AQI don’t have non-linear relationships with the response variable, NO2 AQI. As expected, O3 AQI does have a non-linear relationship with NO2 AQI. Even though Model 2 seems to be better here than the other two models, the non-linearity of O3 AQI makes us consider a non-linear model.

Let’s check if our assumptions on the regression model were correct for Model 2. First of all, the relationship between the forecasted variable NO2 AQI and the independent variables CO, SO2, and O3 AQI are not always linear. We checked that O3 AQI and NO2 AQI has a non-linear relationship.

According to Figure 26, by looking at the histogram of the residuals, the errors are normally distributed. By looking at the scatter plot of the residuals and predicted values, the errors are homoscedastic, meaning the residuals have constant variance. By looking at the Figure 27 for the last assumption check, if the errors are independent of each other, turns out to be not true since the Durbin-Watson statistics show that there is a positive serial correlation between residuals.

The second model used was Seasonal ARIMA. The steps for finding the best ARIMA model for the Illinois dataset are following.

When first plotted, p-values are less than alpha, meaning data is not white noise. But the ACF plot showed the data is non-stationary.

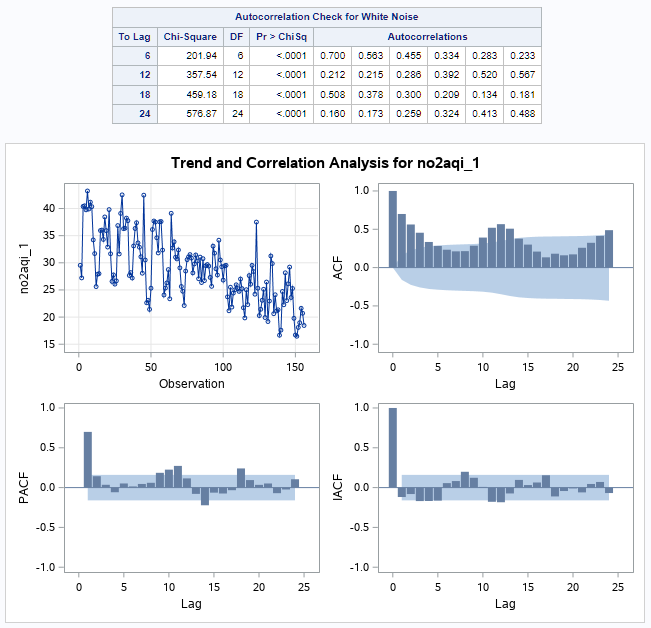


Figure Illinois ARIMA 1

We applied seasonal differencing first to make the data stationary.

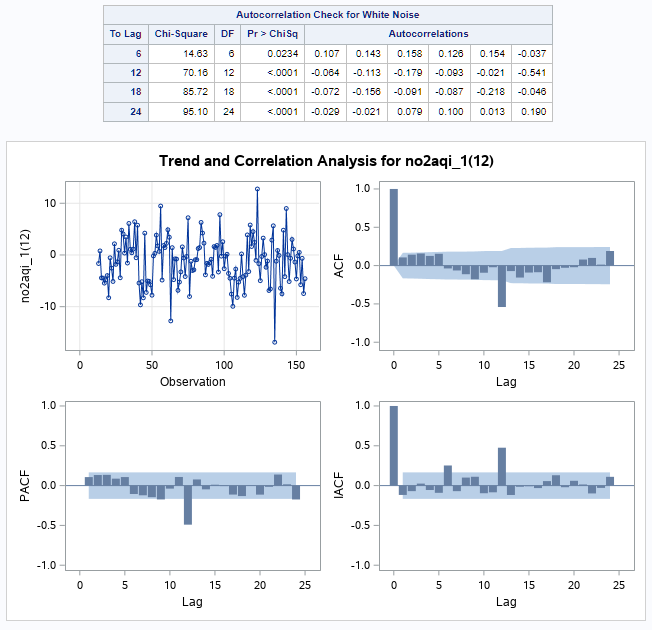


Figure Illinois ARIMA 2

Data is still not white noise and the ACF plot seems to show the data has now become closer to be stationary.

ACF and PACF plots both show one spike at the lag 12, indicating the order of seasonal AR terms is 1, the order of seasonal differencing is 1, and the order of seasonal MA terms is 1.

In terms of nonseasonal AR and MA terms, the ACF plot seems to almost decrease toward zero, and the PACF plot has 1 spike that’s almost reaching out of the interval at the 9th lag, meaning it can be an AR(1) model. Or, we can say there is one spike each in PACF and ACF and it’s ARMA(1,1). We have two models here which are ARIMA(1,0,0)(1,1,1) and ARIMA(1,0,1)(1,1,1). Since the spikes in ACF and PACF plots were not really clear, we also tried ARIMA(0,0,0)(1,1,1) model but it ends up with non-white noise residuals (Figure 31).

According to Figure 28 for the ARIMA(1,0,0)(1,1,1), p-values of Residuals are all bigger than alpha, meaning they are all white noise. We should note here that P-values of residuals up to 18th lag is 0.057, which is bigger than alpha but still very close to it. This model’s AIC is 793.37 and BIC is 805.25. Based on the Figure 29 for the ARIMA(1,0,1)(1,1,1), the p-values of residuals are all bigger than alpha, meaning they are white noise. This model’s AIC is 794.14 and BIC is 808.99.

|  |  |  |
| --- | --- | --- |
|  | **ARIMA(1,0,0)(1,1,1)** | **ARIMA(1,0,1)(1,1,1)** |
| **AIC** | 793.37 | 794.14 |
| **BIC** | 805.25 | 808.99 |
| **MAPE Fit** | 10.49% | 10.40% |
| **MAPE Accuracy** | 10.50% | 10.46% |

Both ARIMA models have similar AIC, BIC, MAPE of fit and MAPE of Accuracy values. Considering we are looking for the simplest model and the differences of AIC and BIC between the two models are not significant, we would consider the first model as the best model that fits the data.

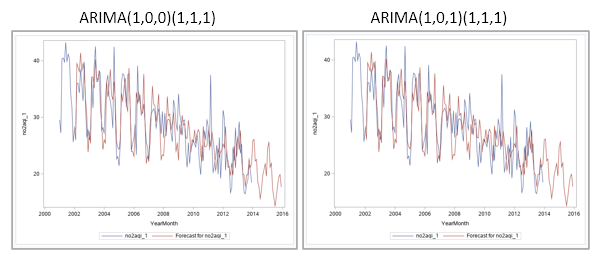


Figure Illinois Original vs. Forecasted of ARIMA models

So far, we have found the first-best model out of the Multiple Regression models and the second-best model out of the Seasonal ARIMA models. Model 2 Regression model had the MAPE of fit 11.20% and MAPE of Accuracy 22.68%, and ARIMA(1,0,0)(1,1,1) model had the MAPE of fit 10.49% and MAPE of Accuracy 10.50%. We can conclude that our ARIMA(1,0,0)(1,1,1) model will be the best fit to forecast the next 3 year(36 months)’s NO2 AQI values in Illinois.

By plotting the forecasted values with the original values we have, here’s the time series plot.

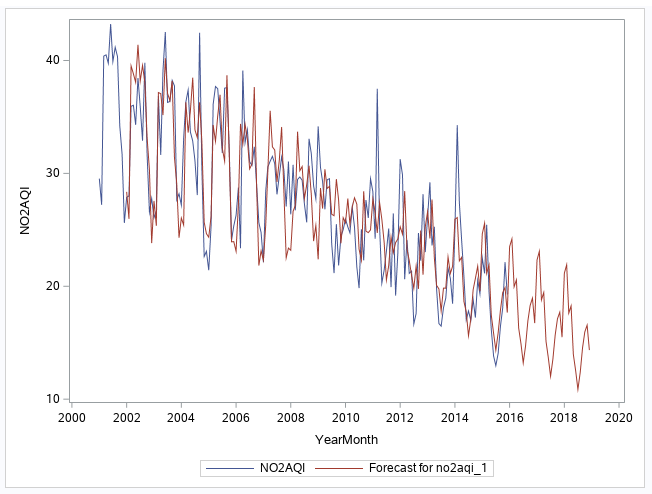


Figure Illinois Forecasts with the best model

***Discussion and Conclusions***

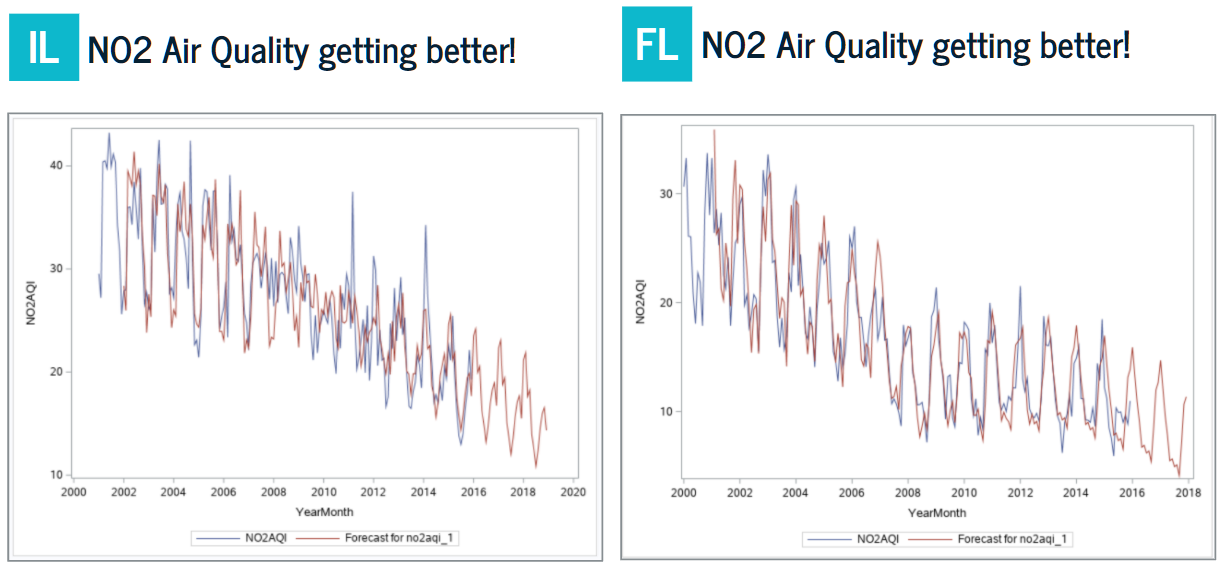


Figure NO2 AQI forecasts in both FL and IL

Both Florida and Illinois showed decreases in forecasted NO2 AQIs from 2016 to 2018. The first observation of Florida in January 2001 was 29.5 whereas the last forecast in December 2018 was 14.3. In Illinois, the first observation in January 2001 was 30.5 and the last forecast in December 2018 was 11.3, which also decreased a lot. It was generally assumed that air pollution has been becoming a bigger environmental issue around the world and the air quality indexes would be worse than the past. However, this was the opposite according to our results. To analyze why this may be happening, we have looked for reasons for lowered rates of NO2 in our environment.

The decreasing trend matched with the trend from the United States Environmental Protection Agency has published.

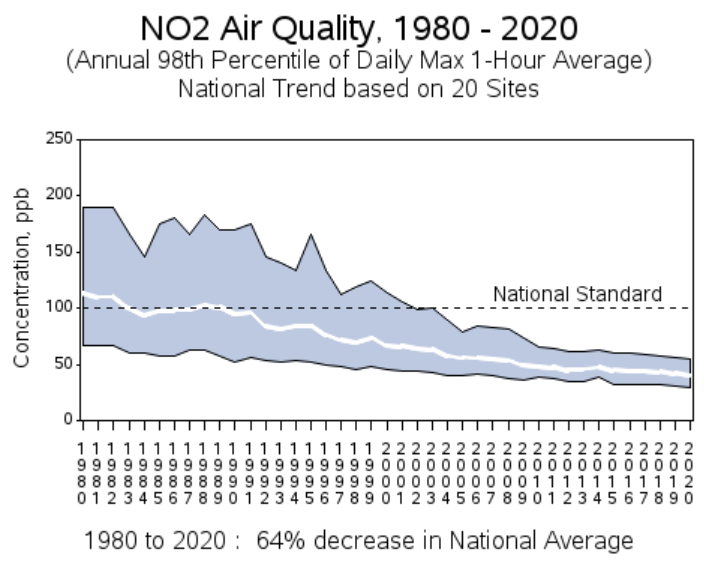


Figure NO2 AQI Trend by EPA

EPA has specified multiple national rules and regulations that exist to reduce emissions of NO2 and meet the National Ambient Air Quality Standards (NAAQS). They have two levels of standards, primary, health-based, and secondary, welfare-based, standards. According to the EPA, the primary standard rules were especially started in 2009 and have been reviewed and modified until recently, 2018. NAAQS for NO2 must have been effective to practically reduce its emissions and protect the public health as much as it can.

In future studies, we hope to use data from as far back as the 1980s. This study used specifically data from the past 16 years. We believe that adding more data to our analysis will make the predictions more accurate. Due to time constraints, we were not able to compare the values we have forecasted for the NO2 values from 2016-2018 with the actual values. In the future, we hope to compare and evaluate our margin of error. This will also show us how effective the model we have used was and can then work on improvements. This study also looked at only two methods of forecasting, seasonal arima and multiple regression. In the future, we hope to compare more models to ensure that the best model was found for this data. Furthermore, we would also like to use different variables (other than SO2,O2 and CO) to forecast NO2 and identify if any differences were made in the forecasts.

***Appendixes 1 - Sas Code***

/\* import data \*/

proc import out=pol datafile="/home/u59282323/sasuser.v94/US\_Pollution\_2000\_2016.csv" dbms=csv replace;

run;

/\* filter by state and create new outputs \*/

data mipol flpol txpol ilpol otherpol;

set pol;

if state = "Michigan" then output mipol;

else if state = "Florida" then output flpol;

else if state = "Texas" then output txpol;

else if state = "Illinois" then output ilpol;

else output otherpol;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Florida Cleansing \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* clean the column names and drop unnecessary columns \*/

options validmemname=extend;

data flpolclean;

set flpol;

drop address; drop county; drop city;

drop 'Table Names'n; drop 'State Code'n;

drop 'County Code'n; drop 'Site Num'n;

drop 'CO Units'n; drop 'SO2 Units'n; drop 'NO2 Units'n; drop 'O3 Units'n;

drop 'CO 1st Max Hour'n;

drop 'CO 1st Max Value'n;

drop 'CO Mean'n;

drop 'NO2 1st Max Hour'n;

drop 'NO2 1st Max Value'n;

drop 'NO2 Mean'n;

drop 'SO2 1st Max Hour'n;

drop 'SO2 1st Max Value'n;

drop 'SO2 Mean'n;

drop 'O3 1st Max Hour'n;

drop 'O3 1st Max Value'n;

drop 'O3 Mean'n;

rename 'NO2 AQI'n = NO2AQI;

rename 'SO2 AQI'n = SO2AQI;

rename 'O3 AQI'n = O3AQI;

rename 'CO AQI'n = COAQI;

run;

/\* group by Yearmonth \*/

proc sql;

create table flpolgrp as

select yearmonth,

avg(NO2AQI) as NO2AQI,

avg(COAQI) as COAQI,

avg(SO2AQI) as SO2AQI,

avg(O3AQI) as O3AQI

from flpolclean

where year(yearmonth) <= 2015

group by yearmonth

order by yearmonth;

quit;

/\* add time variable \*/

data flpolt; set flpolgrp;

t = \_n\_;

run;

proc sgplot data=flpolt; scatter x=coaqi y=no2aqi; run;

proc sgplot data=flpolt; scatter x=so2aqi y=no2aqi; run;

proc sgplot data=flpolt; scatter x=o3aqi y=no2aqi; run;

/\* non-linearity test \*/

data flpolt2; set flpolt;

o3aqi2=o3aqi\*o3aqi;

so2aqi2=so2aqi\*so2aqi;

coaqi2=coaqi\*coaqi;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Illinois Cleansing \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* clean the column names and drop unnecessary columns \*/

options validmemname=extend;

data ilpolclean;

set ilpol;

drop address; drop county; drop city;

drop 'Table Names'n; drop 'State Code'n;

drop 'County Code'n; drop 'Site Num'n;

drop 'CO Units'n; drop 'SO2 Units'n; drop 'NO2 Units'n; drop 'O3 Units'n;

drop 'CO 1st Max Hour'n;

drop 'CO 1st Max Value'n;

drop 'CO Mean'n;

drop 'NO2 1st Max Hour'n;

drop 'NO2 1st Max Value'n;

drop 'NO2 Mean'n;

drop 'SO2 1st Max Hour'n;

drop 'SO2 1st Max Value'n;

drop 'SO2 Mean'n;

drop 'O3 1st Max Hour'n;

drop 'O3 1st Max Value'n;

drop 'O3 Mean'n;

rename 'NO2 AQI'n = NO2AQI;

rename 'SO2 AQI'n = SO2AQI;

rename 'O3 AQI'n = O3AQI;

rename 'CO AQI'n = COAQI;

run;

/\* group by Yearmonth \*/

proc sql;

create table ilpolgrp as

select yearmonth,

avg(NO2AQI) as NO2AQI,

avg(COAQI) as COAQI,

avg(SO2AQI) as SO2AQI,

avg(O3AQI) as O3AQI

from ilpolclean

group by yearmonth

order by yearmonth;

quit;

/\* add time variable \*/

data ilpolt; set ilpolgrp;

t = \_n\_;

run;

proc sgplot data=ilpolt; scatter x=coaqi y=no2aqi; run;

proc sgplot data=ilpolt; scatter x=so2aqi y=no2aqi; run;

proc sgplot data=ilpolt; scatter x=o3aqi y=no2aqi; run;

/\* check for nonlinearity \*/

data ilpolt2; set ilpolt;

o3aqi2=o3aqi\*o3aqi;

so2aqi2=so2aqi\*so2aqi;

coaqi2=coaqi\*coaqi;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Time Series Plotting \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Florida plot the data AQIs \*/

proc sgplot data=flpolt;

series x=yearmonth y=no2aqi;

series x=yearmonth y=o3aqi;

series x=yearmonth y=so2aqi;

series x=yearmonth y=coaqi;

title "Florida AQIs overtime";

run;

proc timeseries data=flpolt plots=acf out=\_null\_;

var no2aqi o3aqi so2aqi coaqi;

corr acf/nlag=48;

run;

/\* Illinois plot the data AQIs\*/

proc sgplot data=ilpolt;

series x=yearmonth y=no2aqi;

series x=yearmonth y=o3aqi;

series x=yearmonth y=so2aqi;

series x=yearmonth y=coaqi;

title " Illinois AQIs overtime";

run;

proc timeseries data=ilpolt plots=acf out=\_null\_;

var no2aqi o3aqi so2aqi coaqi;

corr acf/nlag=48;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Multiple Regression Florida \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*account for seasonality with dummy variables\*/

data flpoltm; set flpolt2; month=month(yearmonth);

t=\_n\_;

if month=1 then month1=1; else month1=0;

if month=2 then month2=1; else month2=0;

if month=3 then month3=1; else month3=0;

if month=4 then month4=1; else month4=0;

if month=5 then month5=1; else month5=0;

if month=6 then month6=1; else month6=0;

if month=7 then month7=1; else month7=0;

if month=8 then month8=1; else month8=0;

if month=9 then month9=1; else month9=0;

if month=10 then month10=1; else month10=0;

if month=11 then month11=1; else month11=0;

run;

/\* training and test set data \*/

data flpolt\_1; set flpolt; no2aqi\_1=no2aqi;

if t > 156 then no2aqi\_1=.; run;

data flpolt\_2; set flpolt2; no2aqi\_1=no2aqi;

if t > 156 then no2aqi\_1=.; run;

data flpoltm\_1; set flpoltm; no2aqi\_1=no2aqi;

if t > 156 then no2aqi\_1=.; run;

/\* model 1 \*/

proc reg data=flpoltm\_1 outest=flpoltreg1; model no2aqi=t coaqi so2aqi o3aqi o3aqi2 so2aqi2 coaqi2 month1 month2 month3 month4 month5

month6 month7 month8 month9 month10 month11/clb corrb dwprob vif aic bic adjrsq;

output out=multiple1out p=multiple1hat r=multiple1resid;

run;

proc sgplot data=multiple1out;

series x=yearmonth y=no2aqi;

series x=yearmonth y=multiple1hat;

run;

/\* model 3 \*/

proc reg data=flpolt\_2 outest=flpoltreg2;

model no2aqi=t coaqi coaqi2 o3aqi o3aqi2 so2aqi so2aqi2/clb corrb vif aic bic adjrsq;

output out=multiple2out p=multiple2hat r=multiple2resid;

run;

proc sgplot data=multiple2out;

series x=yearmonth y=no2aqi;

series x=yearmonth y=multiple2hat;

run;

/\* model 2 \*/

proc reg data=flpolt\_2 outest=flpoltreg3;

model no2aqi=t coaqi o3aqi so2aqi/clb corrb vif aic bic adjrsq;

output out=multiple3out p=multiple3hat r=multiple3resid;

run;

proc sgplot data=multiple3out;

series x=yearmonth y=no2aqi;

series x=yearmonth y=multiple3hat;

run;

/\*mape fit and accuracy \*/

data multiple1out\_mapefit; set multiple1out;

mape\_fit = (abs(multiple1resid)/no2aqi\_1)\*100;

if t> 156 then mape\_accuracy = (abs(no2aqi - multiple1hat)/no2aqi)\*100;run;

data multiple2out\_mapefit; set multiple2out;

mape\_fit = (abs(multiple2resid)/no2aqi\_1)\*100;

if t> 156 then mape\_accuracy = (abs(no2aqi - multiple2hat)/no2aqi)\*100;run;

data multiple3out\_mapefit; set multiple3out;

mape\_fit = (abs(multiple3resid)/no2aqi\_1)\*100;

if t> 156 then mape\_accuracy = (abs(no2aqi - multiple3hat)/no2aqi)\*100;run;

proc means data=multiple1out\_mapefit mean;

var mape\_fit mape\_accuracy; run;

proc means data=multiple2out\_mapefit mean;

var mape\_fit mape\_accuracy; run;

proc means data=multiple3out\_mapefit mean;

var mape\_fit mape\_accuracy; run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Seasonal Arima Florida \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc sgplot data=flpolt;

series x=yearmonth y=coaqi;

series x=yearmonth y=so2aqi;

series x=yearmonth y=o3aqi;

series x=yearmonth y=no2aqi;

title "Florida NO2";

run;

proc timeseries data=flpolt2 plots=acf out=\_null\_;

var no2aqi coaqi coaqi2 so2aqi so2aqi2 o3aqi o3aqi2;

corr acf/nlag=48;

run;

/\*\*\* Three ARIMA models in one section\*\*\*/

proc arima data=flpolt\_2;

identify var=no2aqi\_1(1,12) whitenoise=ignoremiss;

\*estimate p=(12)(24) q=(1)(12) whitenoise=ignoremiss;/\*ARIMA ( 0,1,1 ) (2,1,1 )\*/

\*estimate p=(12) q=(1)(12) whitenoise=ignoremiss; /\*ARIMA (0,1,1) (1,1,1)\*/

estimate p=(1)(12)(24) q=(1)(12) whitenoise=ignoremiss;/\*ARIMA (1,1,1)(2,1,1)\*/

forecast lead=24 out=flpoltout id=yearmonth interval=month;

run;

/\* finding mape\*/

data flpoltout\_merged;

merge flpoltout flpolt\_2;

run;

data flpoltout\_mape; set flpoltout\_merged;

mape\_fit = (abs(residual)/no2aqi\_1)\*100;

if t> 156 then mape\_accuracy = (abs(no2aqi - forecast)/no2aqi)\*100;run;

run;

proc means data=flpoltout\_mape mean;

var mape\_fit mape\_accuracy; run;

/\*original vs forecasted data \*/

proc sgplot data=flpoltout;

series x=yearmonth y=no2aqi\_1;

series x=yearmonth y=forecast;

run;

/\*\*forecasting next three years\*\*/

proc arima data=flpolt\_2;

identify var=no2aqi\_1(1,12) whitenoise=ignoremiss;

\*estimate p=(12)(24) q=(1)(12) whitenoise=ignoremiss;/\*ARIMA ( 0,1,1 ) (2,1,1 )\*/

estimate p=(12) q=(1)(12) whitenoise=ignoremiss; /\*ARIMA (0,1,1) (1,1,1)\*/

\*estimate p=(1)(12)(24) q=(1)(12) whitenoise=ignoremiss;/\*ARIMA (1,1,1)(2,1,1)\*/

forecast lead=60 out=flpoltout1 id=yearmonth interval=month; /\* lead is 60 here because 24(2014-2015 for accuracy) + 36(2016-2018, what we actually want to forecast) \*/

run;

data flpolt\_2merge;

merge flpoltout1 flpolt\_2;

run;

proc sgplot data=flpolt\_2merge;

series x=yearmonth y=no2aqi;

series x=yearmonth y=forecast;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Illinois Multiple Regression \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*account for seasonality with dummy variables\*/

data ilpoltds; set ilpolt2; month=month(yearmonth);

t=\_n\_;

if month=1 then month1=1; else month1=0;

if month=2 then month2=1; else month2=0;

if month=3 then month3=1; else month3=0;

if month=4 then month4=1; else month4=0;

if month=5 then month5=1; else month5=0;

if month=7 then month7=1; else month7=0;

if month=8 then month8=1; else month8=0;

if month=9 then month9=1; else month9=0;

if month=10 then month10=1; else month10=0;

if month=11 then month11=1; else month11=0;

if month=12 then month12=1; else month12=0;

run;

/\* create new data sets by dividing training and test set. 156 is 194 (total count of rows) \* 80% \*\*/

data ilpolt\_1; set ilpolt; no2aqi\_1=no2aqi;

if t > 156 then no2aqi\_1=.; run;

data ilpolt2\_1; set ilpolt2; no2aqi\_1=no2aqi;

if t > 156 then no2aqi\_1=.; run;

data ilpoltds\_1; set ilpoltds; no2aqi\_1=no2aqi;

if t > 156 then no2aqi\_1=.; run;

/\*\*\*\*\* output1 - without o3aqi2 \*\*\*\*\*/

proc reg data=ilpolt\_1 outest=mreg1; model no2aqi\_1=t coaqi so2aqi o3aqi/clb corrb vif aic bic adjrsq;

output out=md1out p=md1hat r=md1resid;

run;

/\* Plot the original vs forecast \*/

proc sgplot data=md1out;

series x=yearmonth y=no2aqi;

series x=yearmonth y=md1hat;

run;

/\*\*\*\*\* output2 - with o3aqi2 but without dummy variables \*\*\*\*\*/

proc reg data=ilpolt2\_1 outest=mreg2; model no2aqi\_1=t coaqi coaqi2 so2aqi so2aqi2 o3aqi o3aqi2/aic bic adjrsq;

output out=md2out p=md2hat r=md2resid;

run;

/\* Plot the original vs forecast \*/

proc sgplot data=md2out;

series x=yearmonth y=no2aqi;

series x=yearmonth y=md2hat;

run;

/\*\*\*\*\* output3 - with o3aqi2 and with dummy variables \*\*\*\*\*/

proc reg data=ilpoltds\_1 outest=mreg3; model no2aqi\_1=t coaqi2 so2aqi2 o3aqi2 month1 month2 month3 month4 month5

month7 month8 month9 month10 month11 month12/vif aic bic adjrsq;

output out=md3out p=md3hat r=md3resid;

run;

/\* Plot the original vs forecast \*/

proc sgplot data=md3out;

series x=yearmonth y=no2aqi;

series x=yearmonth y=md3hat;

run;

/\*\*\*\*\* (not included in the report) Just to try deseasonalizing on the first training set (not included in the report) \*\*\*\*\*/

/\*\*\*\*\*\*\*\* using data ilpolt2\_1 to create reg model that still includes o3aqi2 \*\*\*\*/

proc timeseries data=ilpolt2\_1 outdecomp=ilpolt2\_sa out=\_null\_;

decomp sa; id yearmonth interval=month; var no2aqi\_1; run;

data ilpolt2\_merged; merge ilpolt2\_sa ilpolt2\_1;

si=no2aqi\_1/sa;

run;

proc sgplot data=ilpolt2\_merged;

series x=yearmonth y=sa;

series x=yearmonth y=no2aqi\_1;

run;

proc reg data=ilpolt2\_merged outest=mreg4;

model sa=t coaqi so2aqi o3aqi o3aqi2/aic bic adjrsq;

output out=md4out p=md4sahat r=md4saresid;

run;

proc sgplot data=md4out;

series x=yearmonth y=sa;

series x=yearmonth y=md4sahat;

run;

/\*find mape fit and accuracy with the sa data \*/

data md4out\_mape; set md4out;

mape\_fit = (abs(md4saresid)/sa)\*100;

if t > 156 then mape\_accuracy = (abs(sa-md4sahat)/sa)\*100;

no2aqi\_reseason=md4sahat\*si;

run;

proc means data=md4out\_mape mean; var mape\_fit mape\_accuracy; run;

proc sgplot data=md4out\_mape;

series x=yearmonth y=no2aqi;

series x=yearmonth y=no2aqi\_reseason;

run;

/\* for finding mape fit and accuracy \*/

data md1out\_mf; set md1out;

mape\_fit = (abs(md1resid)/no2aqi\_1)\*100;

if t> 156 then mape\_accuracy = (abs(no2aqi - md1hat)/no2aqi)\*100;run;

data md2out\_mf; set md2out;

mape\_fit = (abs(md2resid)/no2aqi\_1)\*100;

if t> 156 then mape\_accuracy = (abs(no2aqi - md2hat)/no2aqi)\*100;run;

data md3out\_mf; set md3out;

mape\_fit = (abs(md3resid)/no2aqi\_1)\*100;

if t> 156 then mape\_accuracy = (abs(no2aqi - md3hat)/no2aqi)\*100;run;

proc means data=md1out\_mf mean;

var mape\_fit mape\_accuracy; run;

proc means data=md2out\_mf mean;

var mape\_fit mape\_accuracy; run;

proc means data=md3out\_mf mean;

var mape\_fit mape\_accuracy; run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Illinois Seasonal ARIMA \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\* ARIMA(1,0,0)(1,1,1)12 \*\*\*\*\*/

proc arima data=ilpolt\_1;

identify var=no2aqi\_1(12) whitenoise=ignoremiss;

estimate p=(1)(12) q=(12) whitenoise=ignoremiss;

forecast lead=24 id=yearmonth interval=month out=ilpoltout1;

/\* lead is 24 here because we have to find mape accuracy first.

after that we'll run another arima with different lead number to generate 2016-2018 forecasts.\*/

run;

/\* Finding MAPE - merge datasets first to have no2aqi\_1 and no2aqi in the same dataset \*/

data ilpoltout1\_merged; merge ilpoltout1 ilpolt\_1;

run;

data ilpoltout1\_mape; set ilpoltout1\_merged;

mape\_fit = (abs(residual)/no2aqi\_1)\*100;

if t> 156 then mape\_accuracy = (abs(no2aqi - forecast)/no2aqi)\*100;run;

run;

proc means data=ilpoltout1\_mape mean;

var mape\_fit mape\_accuracy; run;

/\* output1 - Plot the original vs forecast \*/

proc sgplot data=ilpoltout1\_merged;

series x=yearmonth y=no2aqi;

series x=yearmonth y=forecast;

run;

/\*\*\*\*\* ARIMA(1,0,1)(1,1,1)12 \*\*\*\*\*/

proc arima data=ilpolt\_1;

identify var=no2aqi\_1(12) whitenoise=ignoremiss;

estimate p=(1)(12) q=(1)(12) whitenoise=ignoremiss;

forecast lead=24 id=yearmonth interval=month out=ilpoltout2;

/\* lead is 24 here because we have to find mape accuracy first.

after that we'll run another arima with different lead number to generate 2016-2018 forecasts.\*/

run;

/\* Finding MAPE - merge datasets first to have no2aqi\_1 and no2aqi in the same dataset \*/

data ilpoltout2\_merged; merge ilpoltout2 ilpolt\_1;

run;

data ilpoltout2\_mape; set ilpoltout2\_merged;

mape\_fit = (abs(residual)/no2aqi\_1)\*100;

if t> 156 then mape\_accuracy = (abs(no2aqi - forecast)/no2aqi)\*100;run;

run;

proc means data=ilpoltout2\_mape mean;

var mape\_fit mape\_accuracy; run;

/\* output1 - Plot the original vs forecast \*/

proc sgplot data=ilpoltout2\_merged;

series x=yearmonth y=no2aqi;

series x=yearmonth y=forecast;

run;

/\* generating forecasts with ARIMA(1,0,0)(1,1,1)12 \*/

proc arima data=ilpolt\_1;

identify var=no2aqi\_1(12) whitenoise=ignoremiss;

estimate p=(1)(12) q=(12) whitenoise=ignoremiss;

forecast lead=60 id=yearmonth interval=month out=ilpoltout3;

/\* lead is 60 here because 24(2014-2015 for accuracy) + 36(2016-2018, what we actually want to forecast) \*/

run;

/\* merge datasets first to have forecast and no2aqi in the same dataset \*/

data ilpoltout3\_merged; merge ilpoltout3 ilpolt\_1;

run;

/\* output1 - Plot the original vs forecast \*/

proc sgplot data=ilpoltout3\_merged;

series x=yearmonth y=no2aqi;

series x=yearmonth y=forecast;

run;

/\* just to try ARIMA(0,0,0)(1,1,1)\*/

proc arima data=ilpolt\_1;

identify var=no2aqi\_1(12) whitenoise=ignoremiss;

estimate p=(12) q=(12) whitenoise=ignoremiss;

forecast lead=24 id=yearmonth interval=month out=ilpoltout1;

/\* lead is 24 here because we have to find mape accuracy first.

after that we'll run another arima with different lead number to generate 2016-2018 forecasts.\*/

Run;

***Appendixes 2 – more figures***

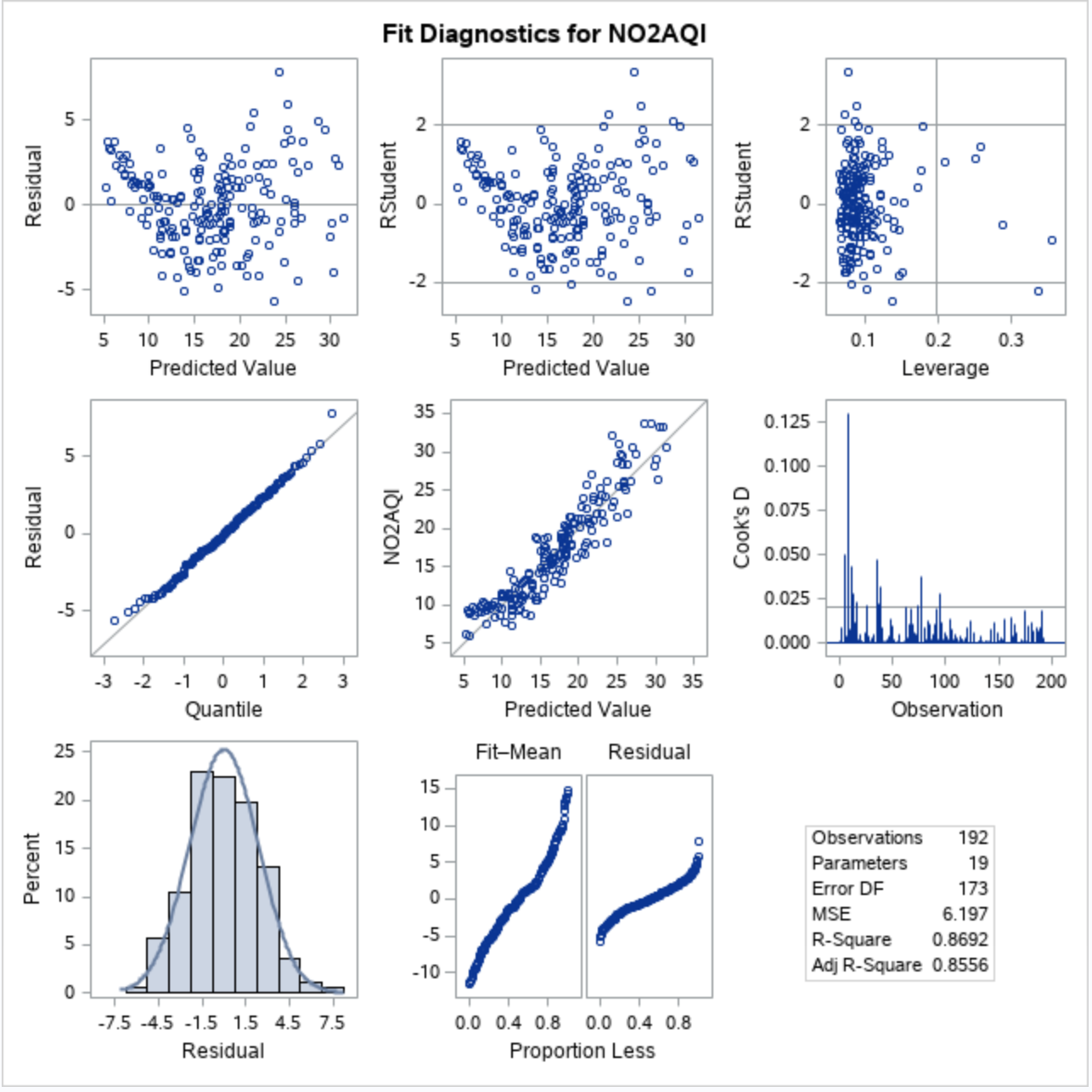


Figure Florida Model Assumption Check 1

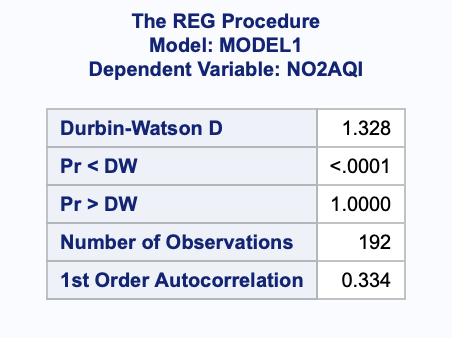


Figure Florida Model Assumption Check 2

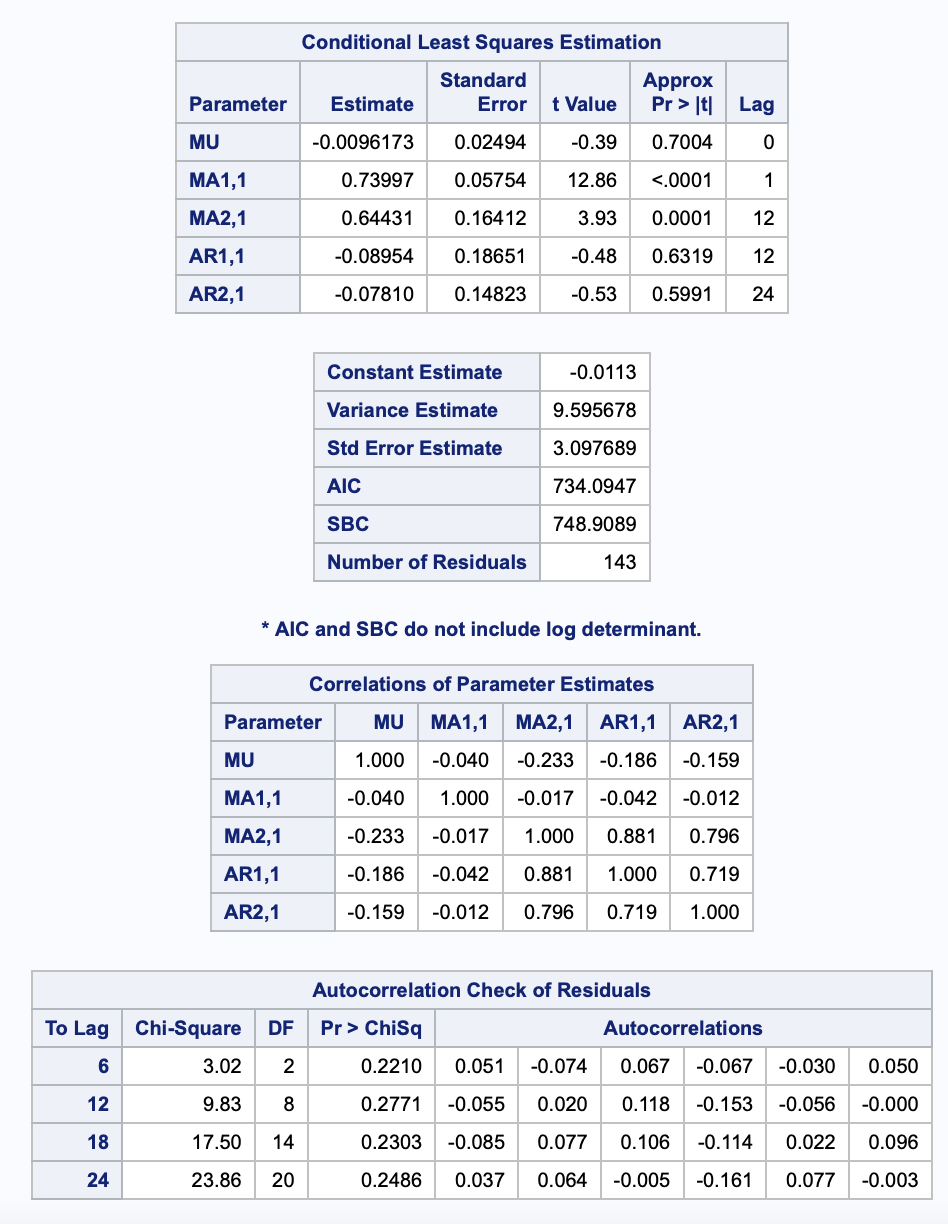


Figure Florida ARIMA(0,1,1)(2,1,1) Result

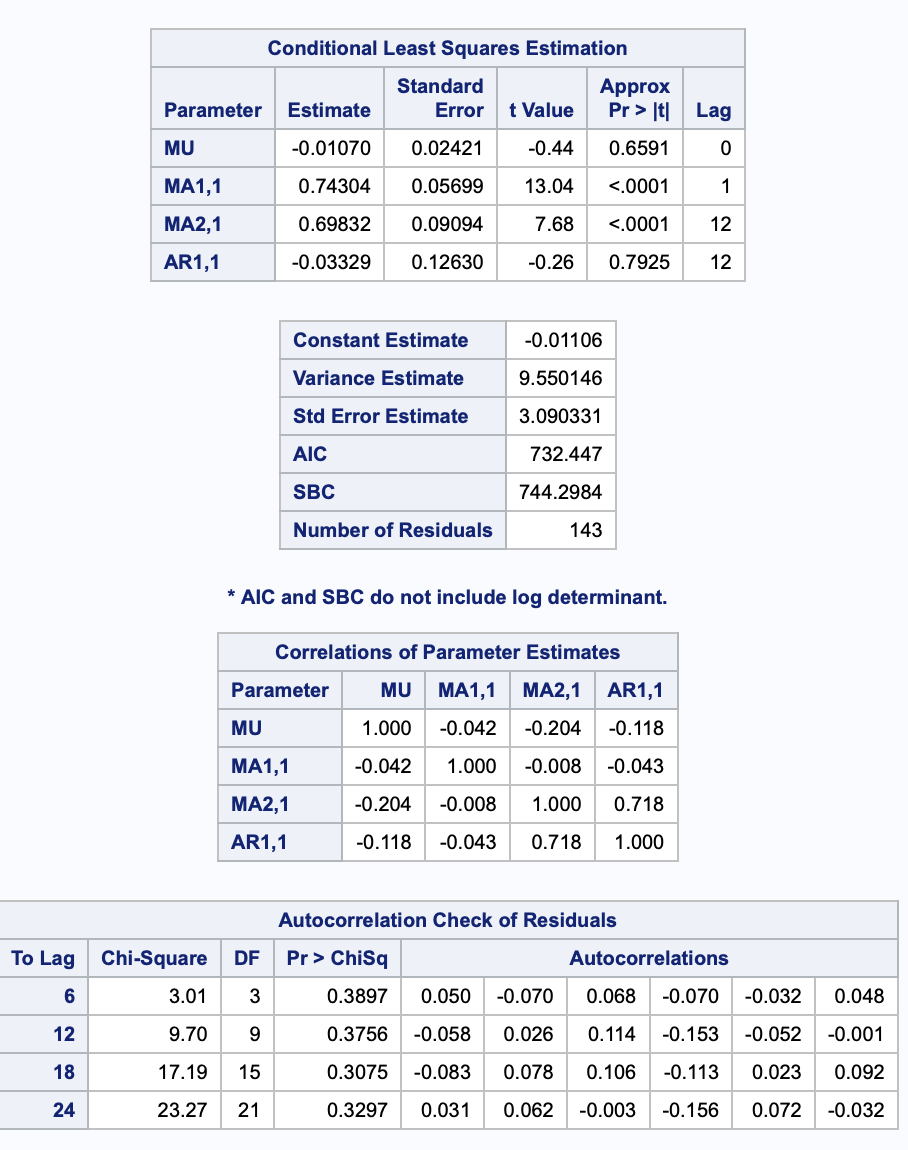


Figure Florida ARIMA(0,1,1)(1,1,1) Result

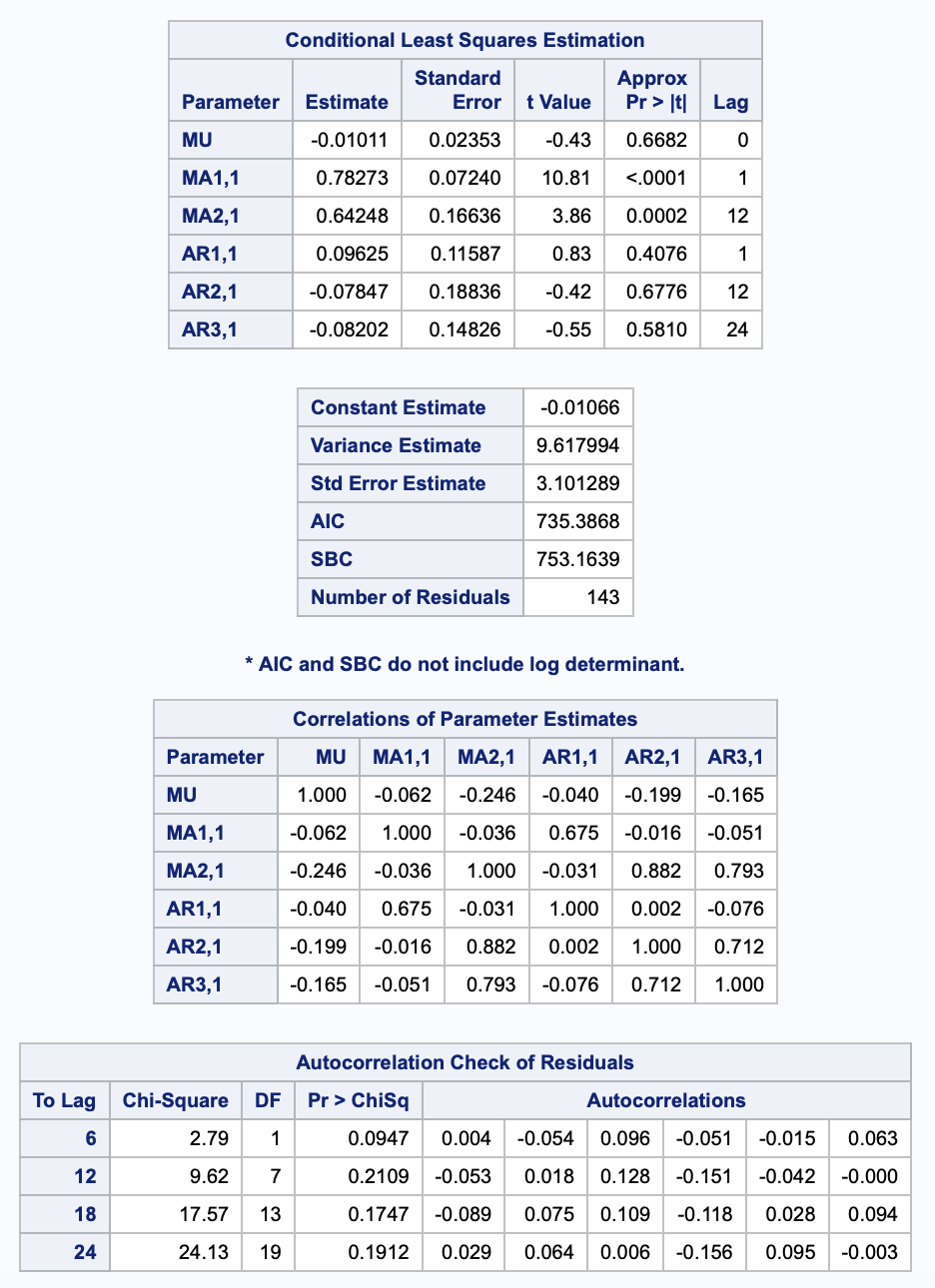


Figure Florida ARIMA(1,1,1)(2,1,1) Result

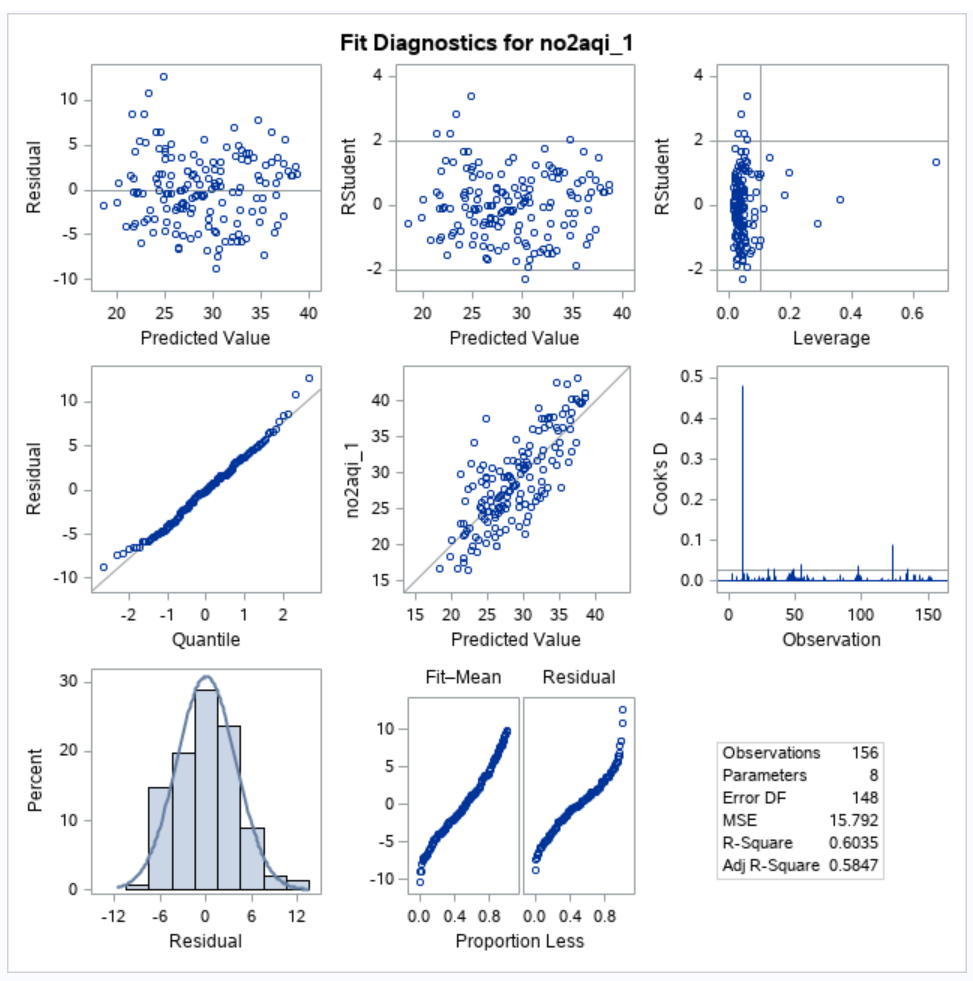


Figure Illinois Model Assumption Check 1

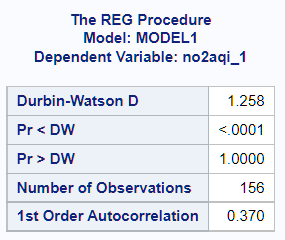


Figure Illinois Model Assumption Check 2



Figure Illinois ARIMA(1,0,0)(1,1,1) Result



Figure Illinois ARIMA(1,0,1)(1,1,1) Result

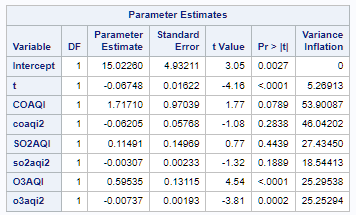


Figure Illinois Regression Model 2 VIFs for multicollinearity check

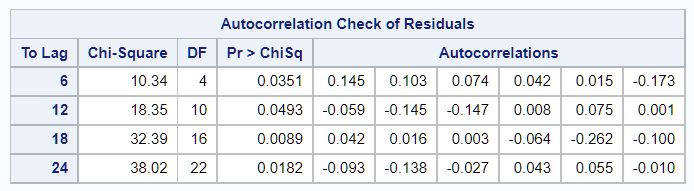


Figure Illinois ARIMA(0,0,0)(1,1,1) Result

***Source Links***

AQI

<https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf>

SO2

<https://www.epa.gov/so2-pollution/sulfur-dioxide-basics>

NO2

<https://www.epa.gov/no2-pollution/basic-information-about-no2#Effects>

CO

<https://www.epa.gov/co-pollution/basic-information-about-carbon-monoxide-co-outdoor-air-pollution#Effects>

O3

<https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics#effects>

EPA NO2 Trend

<https://www.epa.gov/air-trends/nitrogen-dioxide-trends>

EPA NO2 NAAQS regulations

<https://www.epa.gov/no2-pollution/basic-information-about-no2#Reduce>

Primary NAAQS for NO2

<https://www.epa.gov/no2-pollution/primary-national-ambient-air-quality-standards-naaqs-nitrogen-dioxide>

Secondary NAAQS for NO2

<https://www.epa.gov/so2-pollution/secondary-national-ambient-air-quality-standards-naaqs-nitrogen-dioxide-no2-and-sulfur>