DePaul University

DSC 425 Final Project, Winter 2019

Time Series Analysis of Carbon Monoxide Levels in New York City

Final Project Technical Report

by:

Dhaval Delvadia
Lei Lao
Pravika Chitagi
Steven Jordan

Contents

ABSTRACT

In this paper, we perform a time series analysis on Carbon Monoxide (CO) parts-per-million (ppm) levels in New York City using data collected from 2000 to 2016. We explore models using an Autoregressive–Moving-Average (ARMA) analysis, Autoregressive Integrated Moving Average (ARIMA) analysis, Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Analysis, and Neural Network analysis. ARIMA and GARCH models were tested using backtesting and evaluated by the Akaike information criterion (AIC) and mean absolute percentage error (MAPE). We compared the results of each of these time series analysis techniques and found the best model which combined maximum parsimoniousness with minimal error, resulting on an ARIMA (0,0,1) model with GARCH effects.

Key Words: Time Series Analysis, Exploratory Data Analysis of Time Series Analysis, CO Pollutant in New York, Neural Network Forecasting, GARCH, ARIMA, ARMA

1. INTRODUCTION

Carbon monoxide (CO), while not a major greenhouse gas itself, contributes indirectly due to its role in the production of tropospheric surface-level ozone (O3), which has adverse greenhouse effects - without the protective effects of stratospheric high-altitude ozone (Climate & Clean Air Coalition, 2019). Furthermore, both carbon monoxide and ozone are primary components of smog, which has a negative health effects on a city's population, plants, and animals (Canada.ca, 2018). Tracking the levels of CO in the air of a city is useful because it provides insight into the progress and future of curbing greenhouse gas emissions, but also a predictive model can be used by citizens to determine whether to spend time outside, wear masks, etc. - a practice that is becoming increasingly common in large metro areas globally (Smedley, 2018).

New York City, once infamous for its incredible smog, has introduced many regulations to improve air quality and reduce greenhouse gas emissions (NYC Health, 2019). Therefore, in this study we are analyzing, modeling, and forecasting a time series of CO levels in New York City. In order to obtain the best model, we tested numerous types of time series analyses, including: Autoregressive-Moving Average (ARMA) Analysis, Autoregressive Integrated Moving Average (ARIMA) Analysis, Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Analysis, and Neural Network Analysis. Furthermore, we also analysis cross correlation relationship between Nitrogen Oxide (NO) on the main time series CO. We found a suitable model for forecasting carbon monoxide in city of New York, which, when backtested (using previous data), only had an error rate of 5.03%.

This research provides background information on our data sets, preprocessing and feature engineering, an exploratory analysis of the CO attribute, and results of different time series modeling techniques. It provides our conclusions and an explanation of what we believe to be the best model, and recommendations of work that can be done in the future.

2. DATA SETS, PREPROCESSING, AND FEATURE ENGINEERING

Our initial dataset, which was obtained from the US Environmental Protection Agency (EPA Website, 2018) consisted of pollutant measurements in different US cities from 2000 to 2016. It consists of total

28 variables and 1,746,661 observations. The variables included city, state, county, and site location information, as well as daily measurements in parts-per-million (ppm) of NO2, O3, SO2, and CO. Because of the enormous dataset size, and because different locations had inconsistent records for the entire time period, we decided to limit the scope of the analysis to carbon monoxide (CO) from 05/05/2000 to 04/30/2016 in New York City. This reduced the number of observations to 31,887 (the number is still high as there were multiple measurements per day).

We further narrowed down our calculation by aggregating observations from multi-daily to the weekly and monthly mean values, which reduced the observation counts to 866 and 196 respectively. While we conducted data exploration and time series analyses on both the weekly and monthly mean time series, we ultimate determined to only do extensive model testing for the weekly means because not only are there more data points, but also a weekly predictive model is more useful to society as it enables citizens to plan more effectively.

3. EXPLORATORY DATA ANALYSIS

In this exploratory analysis, we first started by plotting the weekly mean CO time series from year 2000 to 2016. This is shown in Fig. 3.1 below. We can observe that the initially CO PPM stated to trend up, but on Dec 2002 (160 week) it picked and started downward trend. It is unclear if this is due to intervention policy by New York City or the trend had been going down even before the year 2002. If there were more time series data collected prior to the year 2000, it would have been easier to determine.

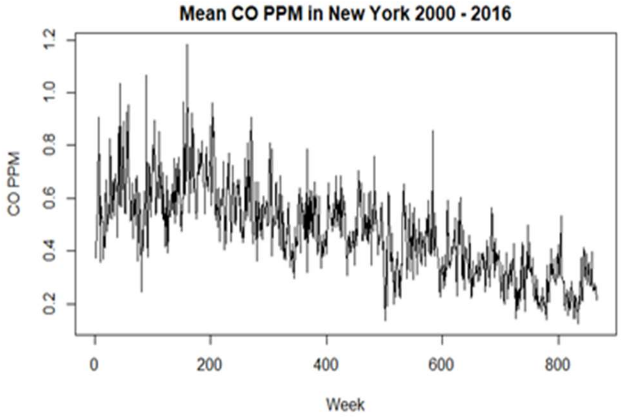


Figure 3.1

Next, we reviewed the distribution of the CO using the histogram and the Normal quantile-quantile plot. Based on the histogram and the QQ plot, we can observe weekly mean CO series is slight right skew with skewness of 0.45 and kurtosis of 0.14. Furthermore, Fig. 3.2 shows histogram with 5-point summary statistic points and fig. 3.5 was actual calculated numbers.

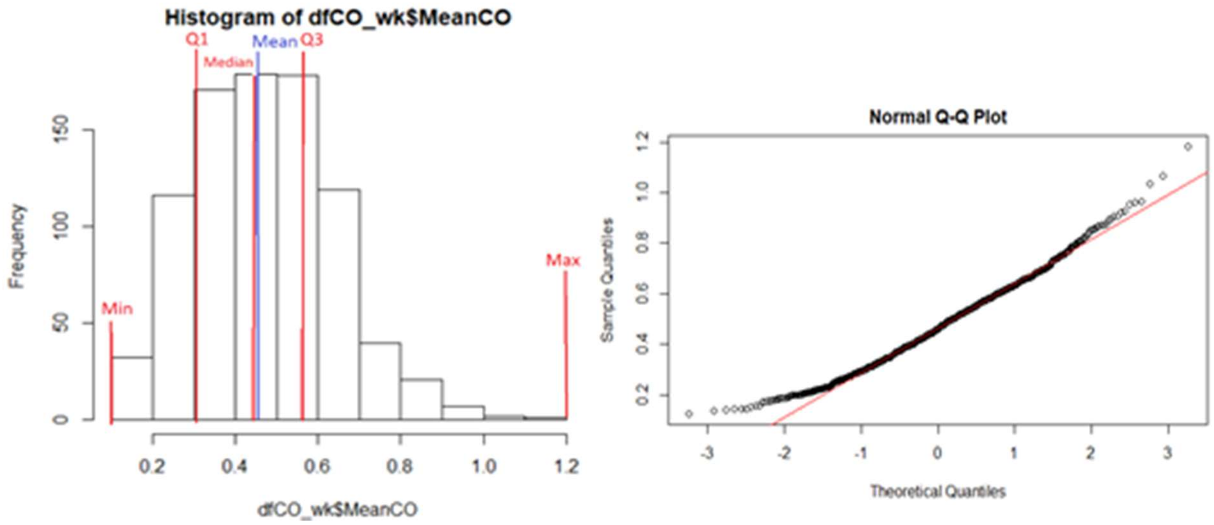
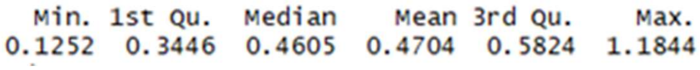


Figure 3.2



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.1252  0.3446  0.4605  0.4704  0.5824  1.1844
```

Figure 3.3 5-Point Summary Statistic

Next, investigation into the series also revealed the series is not stationary. We determined this using the autocorrelation function. See Figure 3.4 for details. We also plotted the Partial Autocorrelation Function to observe the behaviors of the lags. We observed, as shown in fig. 3.4, lags quickly goes to zero exhibiting the Autoregressive (AR) behavior.

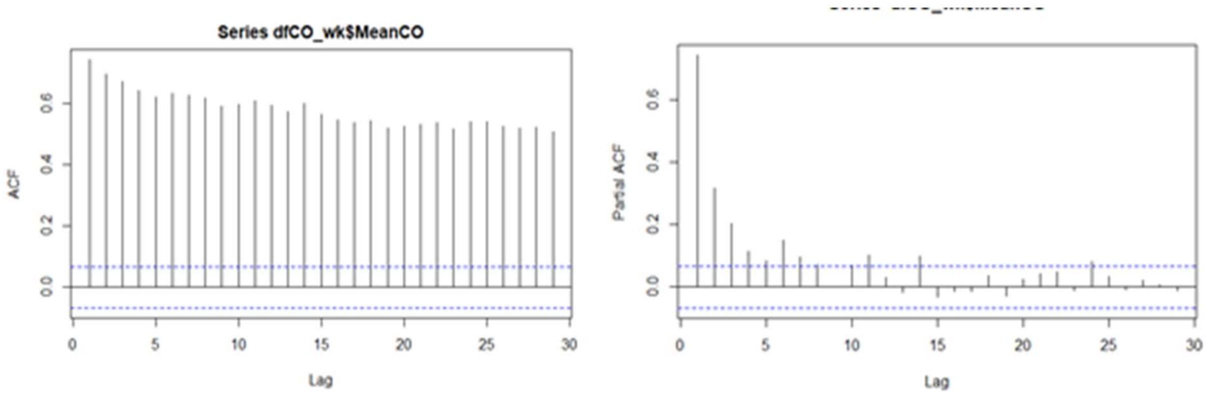
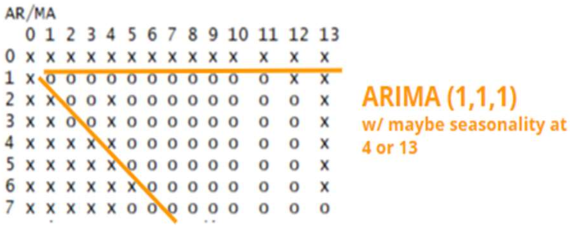


Figure 3.4

We also explored the Extended Autocorrelation Function (EACF) of the original series to observe if there was any seasonality and if we need to take any difference for the arima model. Based on the output shown in Fig. 3.5, it looks like there may be seasonality at frequency 4 and 13. Also, due to AR row 0 being all x's or non-significant, we will need to difference first row in the arima model on the original time series.



```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x o o o o o o o o o o  o  x  x
2 x x o o x o o o o o o  o  o  x
3 x x o o x o o o o o o  o  o  x
4 x x x x x o o o o o o  o  o  x
5 x x x x x o o o o o o  o  o  x
6 x x x x x x o o o o o  o  o  x
7 x x x x x o o o o o o  o  o  o
```

ARIMA (1,1,1)
w/ maybe seasonality at
4 or 13

Figure 3.5 - EACF Output of Original Series

Based on the output of the EACF function on the original series, we took the weekly mean rate (difference return). We plotted this return, see fig. 3.6, and observed the series is stationary. There is no trend. Although there was lot of variations in the output. We also plotted the Normal QQ plot. See figure 3.7 for details. The normal QQ plot also showed that it's normal with fat tails.

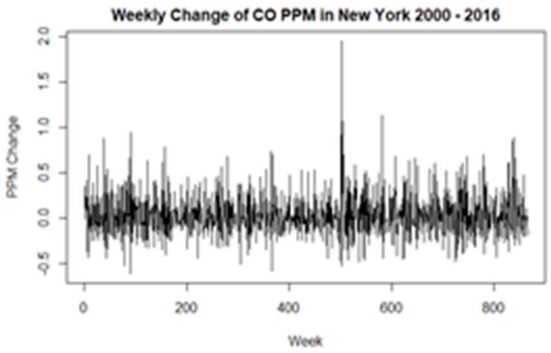


Figure 3.6

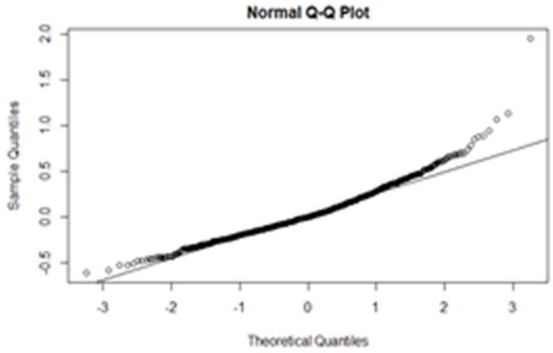


Figure 3.7

Moreover, when ACF and PACF were plotted, we were able to also confirm the stationary output since ACF and PACF did not decay slowly. They both decayed fast. Both are plotted and shown in Fig. 3.8. This was also confirmed with EACF function in Fig. 3.9. Finally, in these plots and in EACF function, we did not observe any seasonality. However, we observed that the possible arima model will be arima (1,1,1) on the original series.

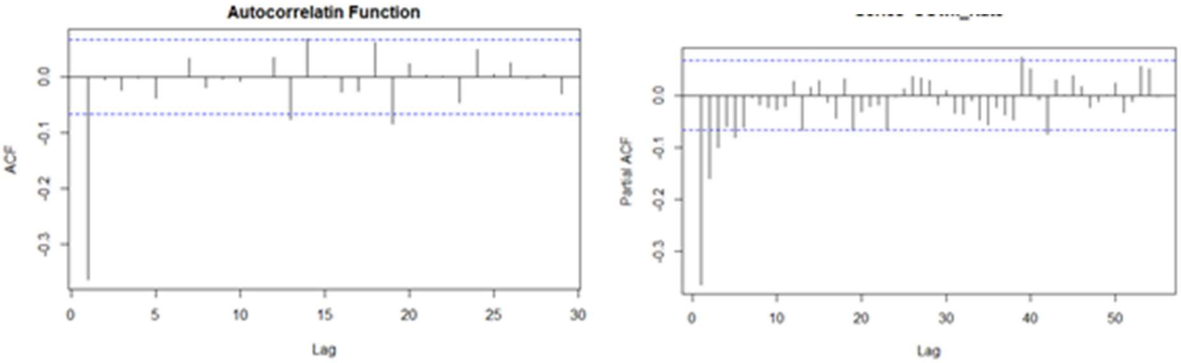
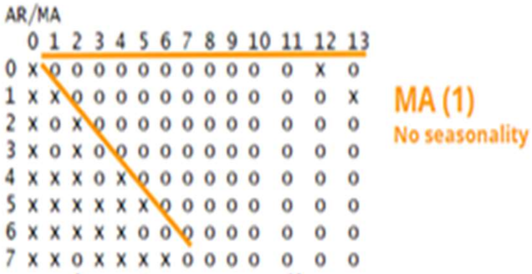


Figure 3.8



```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  o  x  o
1 x x o o o o o o o o o  o  o  x
2 x o x o o o o o o o o  o  o  o
3 x o x o o o o o o o o  o  o  o
4 x x x o x o o o o o o  o  o  o
5 x x x x x x o o o o o  o  o  o
6 x x x x x o o o o o o  o  o  o
7 x x o x x x x o o o o  o  o  o
```

MA (1)
No seasonality

Figure 3.9 EACF Output of Rate (Difference return) Series

Finally, we explored the log of the weekly series and found the outcome of the log returns was not different from the output of the original series. The only visible difference was the volatility was less than the original series. Fig. 3.10 shows this outcome. The Normal QQ plot in Fig. 3.11 also showed normal output except for the flaring out ends of the 45-degree line. The EACF output of rate indicates best model is possibly arima (0,0,1).

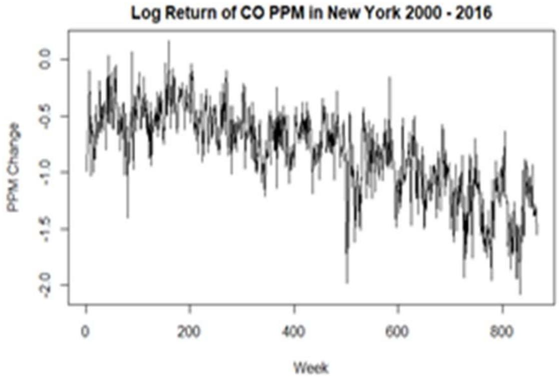


Figure 3.10

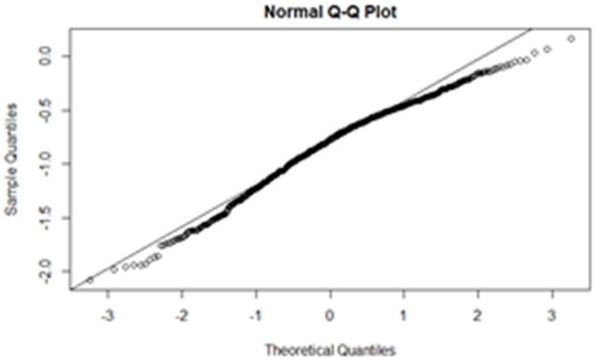


Figure 3.11

The ACF plot for the log series did not show stationary behavior, shown in Fig. 3.12. The PACF of the log series did show stationary behavior, shown in Fig. 3.13. Conducting the EACF function evaluation showed similar output as EACF of the original series shown in Fig. 3.5 above. Since the outcome of the EACF is same as the difference of rate in Fig. 3.5 above, it will also be arima (1,1,1) model.

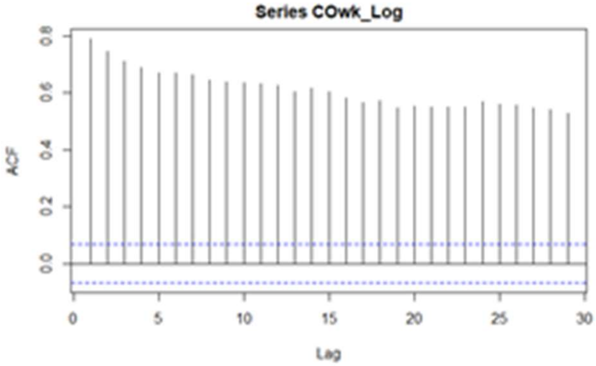


Figure 3.12

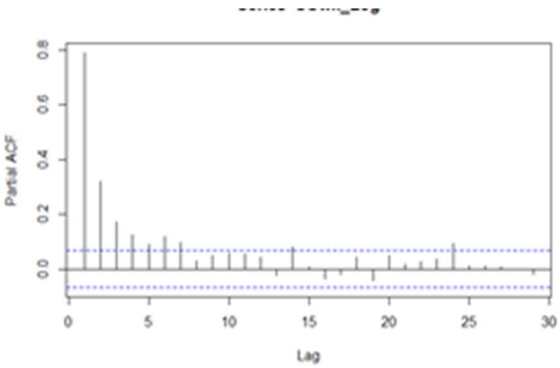


Figure 3.13

Based on these outcomes, next we analyzed the data employing time series analysis techniques noted in section 4.

4. TIME SERIES ANALYSIS TECHNIQUES

4.1. AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) ANALYSIS

Next our analysis consisted of finding the model based on the ARIMA model building technique. We employed R software's built in ARIMA function. We fitted the model based on the exploratory data analysis in section 3 of this report on the original CO time series where the order of ARIMA was determined to be 1,1,1. We then conducted the hypothesis test using Ljung-Box test and found the residuals to be white noise. We also confirmed which coefficients were significant using the built-in R test called coeftest and found all coefficients to be significant. This model was noted as model M1a. We plotted the residuals at this point and noticed GARCH effect which was also analyzed later in the analysis to come up with better model.

We wanted to compare our model order of p=1, d=1, and q=1 to the auto.arima function, so we passed the original weekly mean time series with BIC and stationarity criteria into the function and found a model. We conducted the hypothesis test using Ljung-Box test and found the residuals to be white noise. We also confirmed which coefficients were significant using the coeftest and found all to be significant. Based on the auto.arima, we found the best model with arima order of p=2, d=1, q=1. This model was noted as model M1b.

We also conducted auto.arima model on the weekly rate series. We found the best model to be p=0, d=0, and q=1. Just like before, we conducted all the tests to confirm adequacy of the model.

Finally, we investigated using the log return of the CO weekly series and tried the arima model with p, d, q parameters to be 1,1,1. We found the model using these parameters and then we conducted all the tests noted earlier to confirm adequacy of the model. We further evaluated this same log return of the CO weekly series using the auto.arima function in R. We found the best model to be at the parameter p=5, d=0, and q=0. Just like before we conducted all the tests noted earlier to confirm the adequacy of the model.

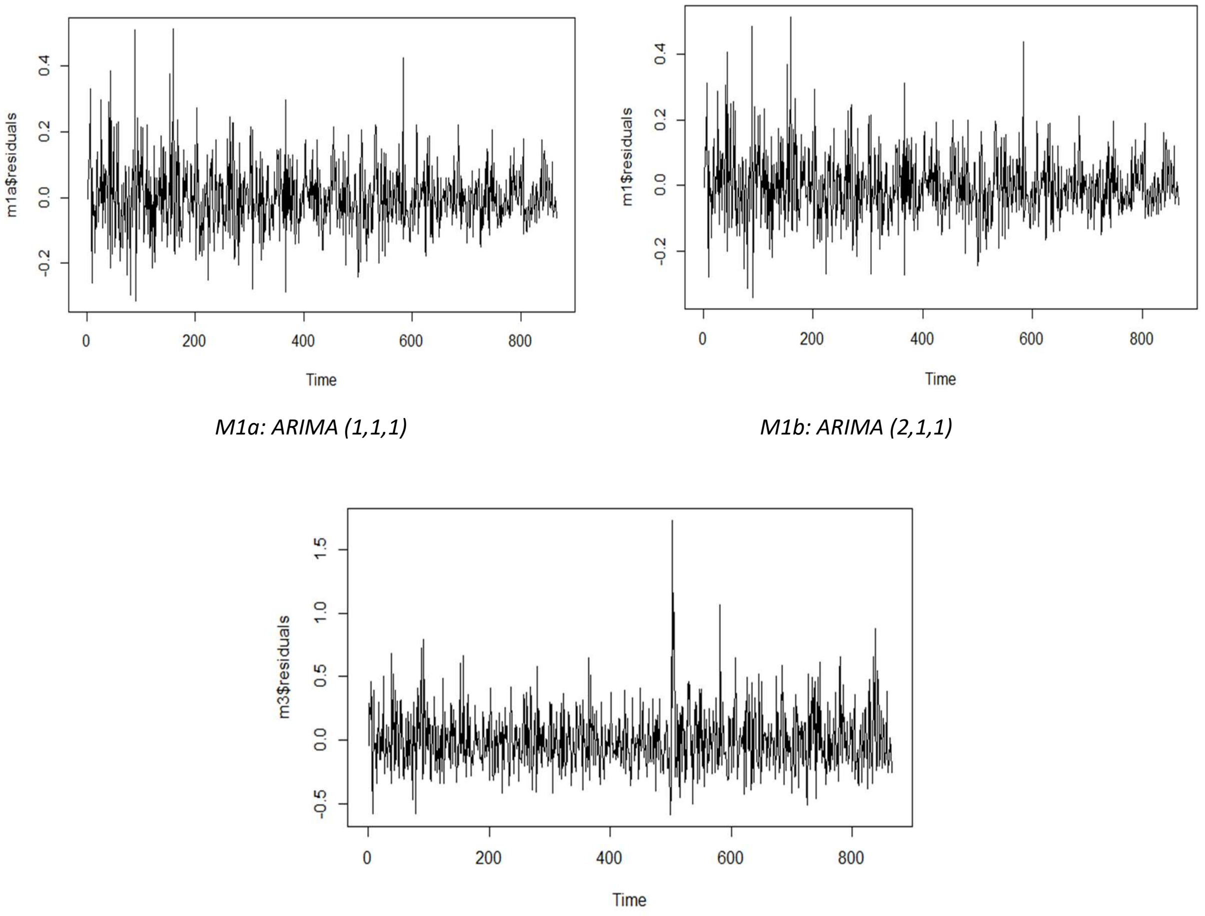
Models	p, d, q	AIC	BIC
M1a: arima	1,1,1	-1511	---
M1b: auto.arima	2,1,1	-1514	-1495
M2: auto.arima	0,0,1	-25	-11
M3a: arima	1,1,1	-210	---
M3b: auto.arima	5,0,0	-177	-144

Table 4.1.1

Next seasonal effects were investigated. We investigated the seasonal effect at frequency 4 and 13 based on the EACF output in the Fig. 3.5 above. We conducted seasonal effect modeling with arima function at various p, d, q values investigated earlier. However, at all different values, seasonal effects were not significant. Based on these outcomes, we did not think there was any seasonality in the model.

4.2. GENERALIZED AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTICITY (GARCH) ANALYSIS

In the ARIMA analysis of the previous section, we noticed in each of the generated models the residuals were clumping close to each other and then spreading apart and this type of behavior was being repeated. It's also shown in Fig. 4.2.1 below. Based on this, we conducted GARCH analysis on the residuals of each model to find best model which incorporate ARIMA plus GARCH.



M1a: ARIMA (1,1,1)

M1b: ARIMA (2,1,1)

M2: ARIMA (0,0,1)

Figure 4.2.1

We begin by employing the arima models obtained earlier and noted as M1a to M3b and the garch 1,1 model to output new models. To do this, we employed the built in garchfit function in R to build the models. Based on these fits, we recorded following outputs. Furthermore, upon conducting the back testing on each of these models following Mean Absolute Percentage Error (MAPE).

Models	p, q	Log likelihood	Back testing, MAPE
gm1a: arma(p,q) + garch(1,1)	1,1	789	19.3%
gm1b: arma(p,q) + garch(1,1)	2, 1	790	---
gm2: arma(p,q) + garch(1,1)	0, 1	51	5.03%
gm3a: arma(p,q) + garch(1,1)	1, 1	116	14.7%

Table 4.2.1

Based on the above results, we narrowed down our best model to gm1a and gm2. We picked gm1a since the AIC for gm1a is -1511, its log likelihood is the highest and tied with gm1b model at 789, and the back testing MAPE error is 19.3%. We picked the other model gm2 since it has AIC of -25, log likelihood of 51, and back testing MAPE error of 5.03%. Based on these two final models and since we wish to find the parsimonious model, we also need to look at the number of independent variables of the model determining the output. In this case, since gm1a has 5 variables and gm2 has only 4 with lower error than gm1a, therefore, our final model was gm2. Our model equation is as follow:

$$x_t = 0.584a_{t-1}$$

$$r_t = 0.0271 + a_t$$

$$\sigma_t^2 = 0.0155a_{t-1}^2 + 0.641\sigma_{t-1}^2$$

4.3. CROSS CORRELATION

Since our data set consist of other pollutant, we want to see if we can utilize time series regression model to forecast the Carbon Dioxide. We checked the correlation and plot between the Carbon Dioxide and the other three pollutants. We decided to use Nitrogen Dioxide as our predictor, because it has the highest correlation (0.79) and a nice linear relationship with Carbon Dioxide.

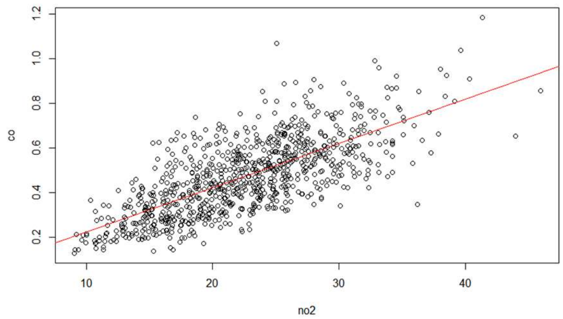


Figure 4.3.1

We used the auto.arima on the regression model residual to determine our ARIMA(3,0,1) model. We can now forecast CO using NO2. To check our forecasting performance, we did a 90/10 split on the dataset, then we check the one step predicted value and actual value. The actual CO value at step 781 is 0.29, and our predicted value is 0.28. Our prediction is not too far off from the actual value, so our forecast model is pretty good.

4.4. NEURAL NETWORK - EXTRA CREDIT

We looked in Neural Network model to forecast our time series. We used Long Short Term Model (LSTM) to analyze our time series data. LSTM is a Recurrent Neural Network model that is capable of learning long term dependency, which is perfect for time series data. LSTM consists a network of memory cells. Each memory cell contains three gates: Input Gate, Forget Gate, and Output Gate. The Input Gate takes the previous output together with the new input and passes them through another layer. The Forget Gate takes the previous output and current input and combine them into a tensor. The Output Gate controls how much of the internal state is passed to the output and works in a similar manner to the other gates.

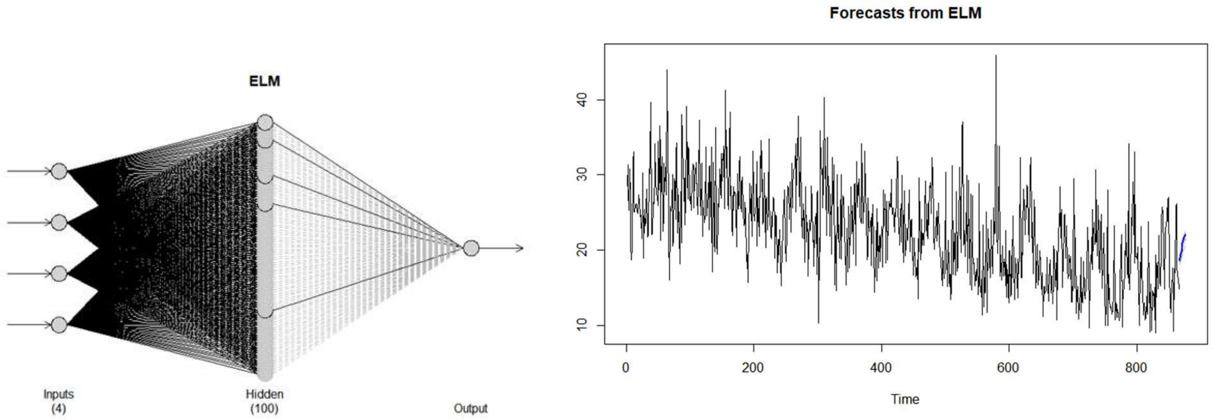


Figure 4.4.1

We used the library(nnfor) package in Rstudio to make our neural network forecast. To check our forecast, we used the 90/10 split on dataset to compare the one-step predicted value and actual value. The actual CO value at step 781 is 0.29, and our predicted value is 0.33. Our prediction is far off from the actual value, so our forecast model is not very good. For neural network modeling, it will require a lot of parameters tuning to improve the model.

4.5. INTERVENTION ANALYSIS

The plot of the CO time series showed an upward trend until Dec 2002 and then it started to move downward until the year 2016. It was difficult to tell why there were two trends. We could not find any policy change or significant event that caused the trend to reverse. However, we wanted to figure out the model that included the effects of interventions on time series' normal behavior. The general model for time series {Yt} is given as Yt = mt + Nt where mt is the change in the mean function and Nt is modeled as some ARIMA process. And if the time series is subjected to the intervention, then at time before intervention is referred to as pre-innervation data and Nt can be used to specify the model. In order to figure out the if intervention was permanent or temporary, we employed the step function and the pulse function on the series. We plotted both the step function and the pulse function. Results with the step function and pulse function are shown in Fig. 4.5.1 and 4.5.2, respectively.

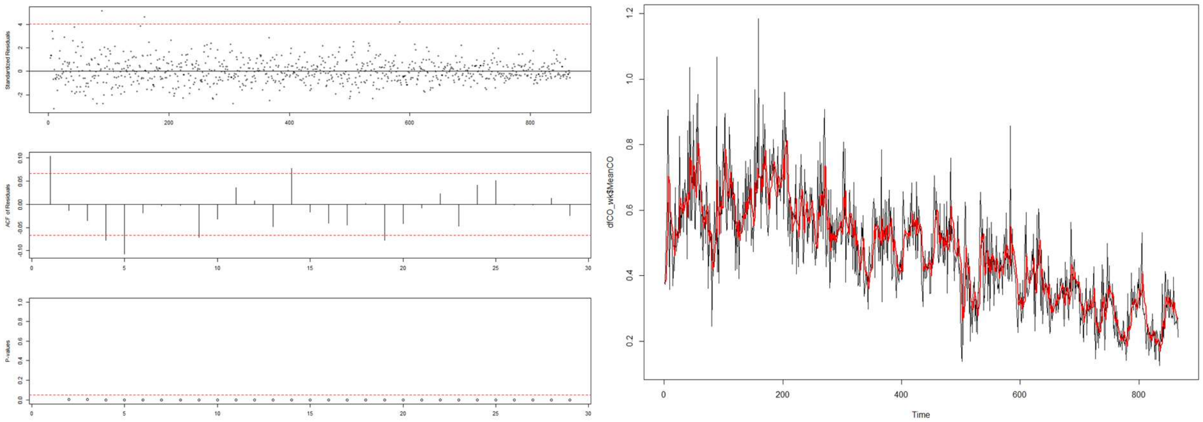


Figure 4.5.1

Based on the output of the residual plot of the intervention model containing step function, the residuals are not within the 95% tolerance. Therefore, this model was not adequate. We then plotted the impulse function with the ARIMA(0,1,1) model and found its residual plot also was not within the 95%. Therefore, this model was also not adequate.

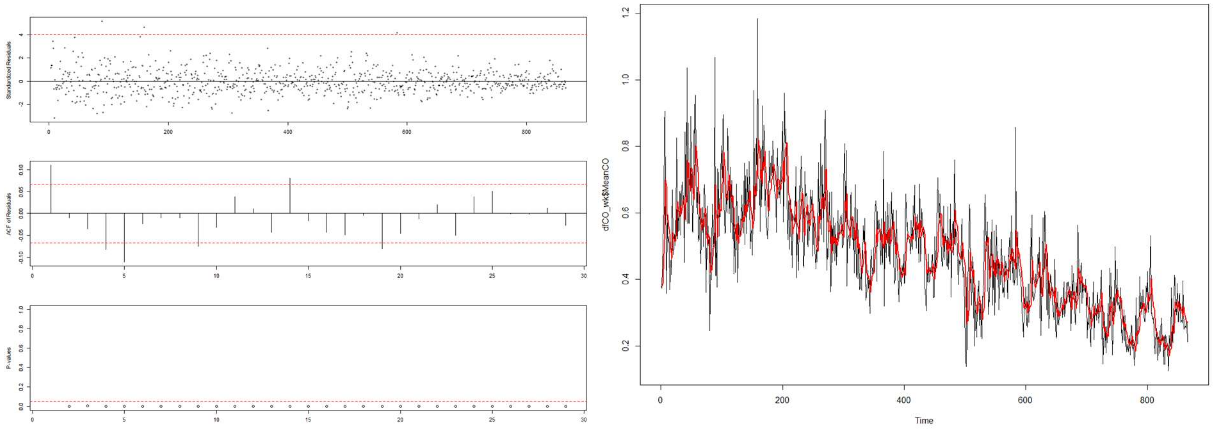


Figure 4.5.2

This means there was probably no intervention in the time series since we do not have the long-term data. May be the time series was continuously trending down. Therefore, we can only conclude there was no intervention in this dataset.

5. CONCLUSION

In this New York Pollutant time series analysis, we have analyzed several different models around the pollutant CO. We conducted ARIMA analysis, GARCH analysis, Cross Correlation, and Neural Network. They each have their strengths and weaknesses. For example, ARIMA model, GARCH model, and cross correlation provides models and are much faster. However, Neural Network requires lots of iterations and fine tuning of weights in order to figure our future outputs and it does not provide any mathematical model. From this reason, Neural Network would not be the best fit for time series. On the other hand, Cross-Correlation is based on dependence among two related series, in this case CO and NO. However, we could have also evaluated the correlation with Ozone (O3) or the Sulfur Dioxide (SO2) series. So, if there is any kind of correlation in those other correlated series, it may or may not reflected in the CO series and our prediction will be off. Therefore, it leaves us with ARIMA and GARCH analysis. These techniques are employed to find the best model. Thus, in this analysis, we narrowed down our best model to M2 (Difference Return) and M3a (Difference Log). The both are very parsimonious models structured as ARIMA (0,0,1) with GARCH(1,1) effects. Ultimately, M3a was chosen because it had a an excellently low MAPE (5.22%), but twice the log likelihood of the other identically sized model.

6. FUTURE WORK

Further work that could be conducted as a result of these findings would be to firstly to collect more daily data and aggregate it over weekly and monthly to train the time series. Second, it would be more interesting to add few more models calculated using Intervention Modeling and training Deep Neural Network. And then compare outputs of all models to see if backward testing from each technique are providing the same mean expected outcome. Once we know this review published article related to this specific topic to determine if our findings match publish results. However, the last part may be not be readily available.

7. REFERENCES

U.S. EPA, 2018, available at: https://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html; clean data downloaded from BrendaSo, U.S. Pollution Data, https://www.kaggle.com/sogun3/uspollution

Climate & Clean Air Coalition. (2019). *Tropospheric ozone*. [online] Available at: http://www.ccacoalition.org/ru/slcps/tropospheric-ozone [Accessed 17 Mar. 2019].

Canada.ca. (2018). *Smog and your health - Canada.ca*. [online] Available at: https://www.canada.ca/en/health-canada/services/air-quality/smog-your-health.html [Accessed 17 Mar. 2019].

Smedley, T. (2019). *Deadly air in our cities: the invisible killer*. [online] The Guardian. Available at: https://www.theguardian.com/environment/2019/mar/17/air-pollution-london-low-emission-zone-deadly-toxic-fumes [Accessed 17 Mar. 2019].

NYC Health. (2019). *Outdoor Air Quality*. [online] Available at: https://www1.nyc.gov/site/doh/health/health-topics/air-quality-air-pollution-protection.page [Accessed 17 Mar. 2019].

8. CODE

See separate folder containing R codes