University of California, Santa Barbara

The Rise and Fall of Carbon in Energy Production - The Gradual Reduction in Humankind's CO2 Footprint

PSTAT 174

Mu Pledge Class - Theta Apple Pie

Mclane Brown, Davis Messer, Diane Phan, Dennis Wang

Table of Contents

Abstract

The main interest of this project is to forecast the monthly emissions of carbon dioxide using the data collected from every month through years 1996 to 2015. To examine the monthly emissions of carbon dioxide, we performed analysis to create an optimal SARIMA model which best predicts the future of CO2 emissions. This includes transforming the data, removing seasonality and trend, estimating coefficients and ultimately testing for violations and forecasting monthly values within a 95% confidence interval. The forecasted values provide promise in solving real world problems regarding the effect carbon emission has on earth and its inhabitants.

Introduction

Carbon emissions are one of the primary greenhouse gases in earth's atmosphere, and as a result play an essential role in the lives of plants and animals. With the global demand for energy increasing each year, carbon dioxide is needed in order to support humans in their everyday necessities, activities, and technological advances of our generation. Although carbon emission provides major benefits to everyday development in humans, machinery, and wildlife, carbon emissions also have many negative effects on Earth and its inhabitants, including global warming. Thus, it is necessary to examine the frequency of carbon emission and propose solutions as to how it can be reduced and monitored to help preserve the planet for future generations.

The goal of the project is to forecast the monthly emissions of carbon dioxide based on data gathered each month from 1996 to 2015. Our data set, titled "Carbon Emission from Electricity Production" from Kaggle and includes monthly observations from 1973 to 2016 of eight different CO2 emission sources-- including Coal, Distillate Fuel, Petroleum Coke and Residual Fuel Oil-- as well as a Total CO2 Emissions (sum of all eight sources) for each month. One of the positives was dealing with the quadratic trends, which we were hesitant to interpret, but eventually, we were able to find appropriate models from the interpretations and challenges. On the other hand, one of difficulties we faced was interpreting the ACF and PACF plots to properly find a SARIMA model to use. Our preliminary model from our ACF and PACF analysis did not coincide with the Akaike Information Criterion (AIC) model.

Eventually, we were able to narrow our data down to 240 monthly observations of the total CO2 from all 8 sources from 1996 to 2015, we conducted data transformations, differencing, and ACF/PACF plot analysis and AIC comparison all in RStudio to produce our final model;

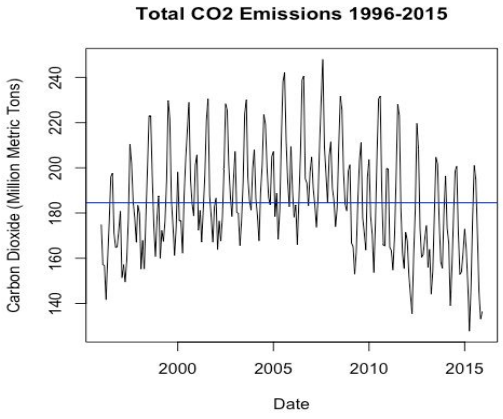
$$SARIMA(3,2,3) \times (0,1,1)_{12}$$

Our analysis included various tests to determine whether or not we had to transform our data, examine ACF and PACF plots, and remove trends or seasons to build a model that satisfies the conditions of diagnostic checks. That includes checking the normality of errors, detecting the serial correlation, and detecting heteroscedasticity. This process allowed us to explore research question such as what months or seasons tend to produce the most CO2 emissions? And our main research problem: Where will CO2 emissions go in the next 1-2 years? Lastly, we can compare our forecasted values to the first 7 months of 2016 which we omitted from our analysis, to see all of the actual values that lie within our forecasted confidence interval at 95%.

Source of data: https://www.eia.gov/electricity/data.php#elecenv
Our data set: https://www.kaggle.com/txtrouble/carbon-emissions
Special thanks to RStudio for helping us analyze the copious amounts of data in order to complete this project as well as our instructor Sudeep Bapat and teaching assistants for providing us feedbacks and codes during lab sections to utilize for this project.

Data Exploratory Analysis

Our data set contained monthly observations from 1973 to 2016 of eight different CO2 emission sources-- including Coal, Distillate Fuel, Petroleum Coke and Residual Fuel Oil-- as well as a Total CO2 Emissions (sum of all eight sources) for each month. We decided to use the Total CO2 Emissions and only include observations from 1996 to 2015, as it will give us the best forecast and answer our initial research questions.



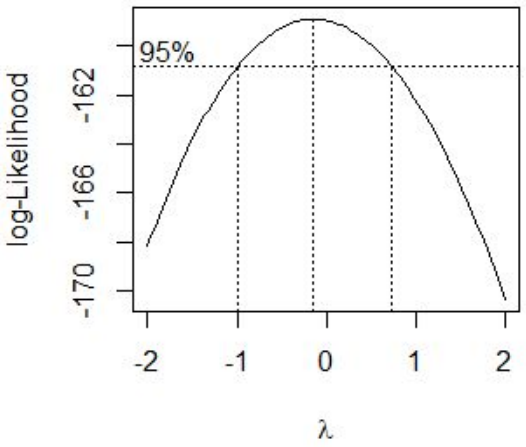
We began by plotting the 240 observations we had narrowed our research to, and look for 3 main

characteristics; trend, seasonal component, as well as any sharp changes in behavior. From the ensuing the time series plot, we used a blue line to denote the mean and observed a quadratic trend which is represented by the change in average CO2 emissions over time. Secondly, a clear seasonal component is apparent through the periodic peaks. Lastly, for our dataset of interest, there are no apparent sharp changes in behavior. Furthermore, variance also seems to be non-constant along the data timeline. Considering the trend, seasonality and non-constant variance based on observation, our time series data is non-stationary, resulting in the need for transformation and differencing. A more indepth break down of our data is provided in the analysis in future sections.

Data Transformation

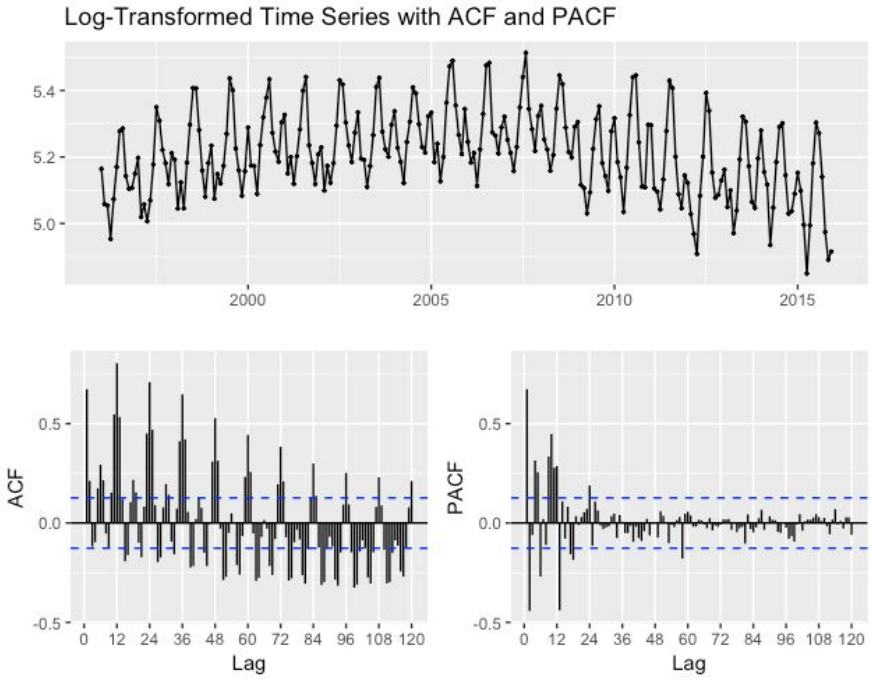
Stabilize Variance

Having done exploratory analysis of our time series data, the non-constant variance requires a transformation in order to stabilize the variance of Total CO2 Emissions. In order to do this, we used a Box-Cox transformation in order to find the optimal lambda value. The optimal lambda value returned by the Box-Cox transformation is -0.14. However, our Box-Cox plots 95% confidence interval includes 0. This means we can use a log transformation, which gives a more interpretable and easier to use transformation.

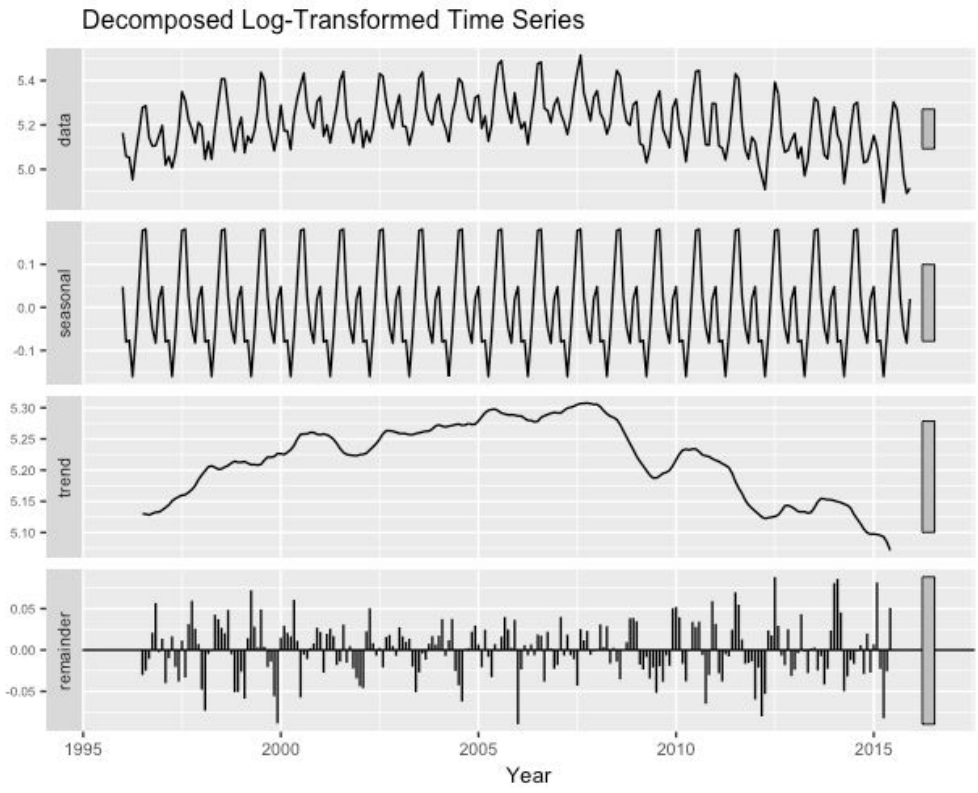


Differencing

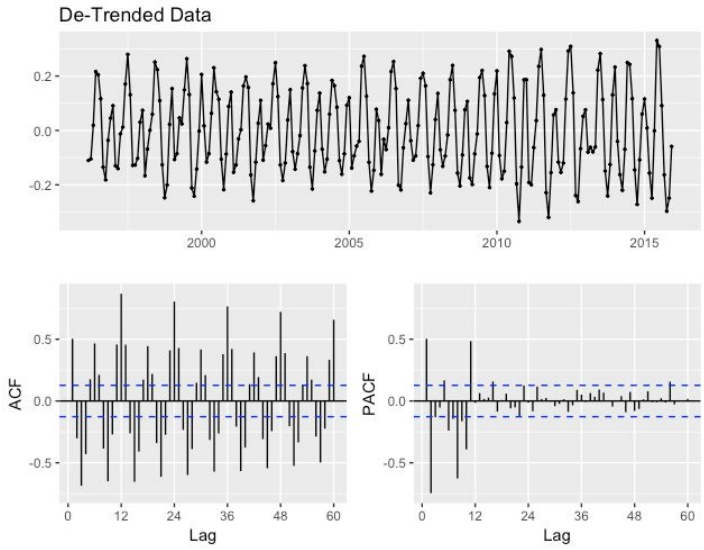
In order to remove seasonality and trend, we must difference the data. By removing seasonality and trend, we are able to view our models on the same level because there may be a case where a particular time period or season has significantly more carbon emissions compared to another season. De-seasonalizing and de-trending the models makes it easier to work with multivariate models and analyze them altogether.



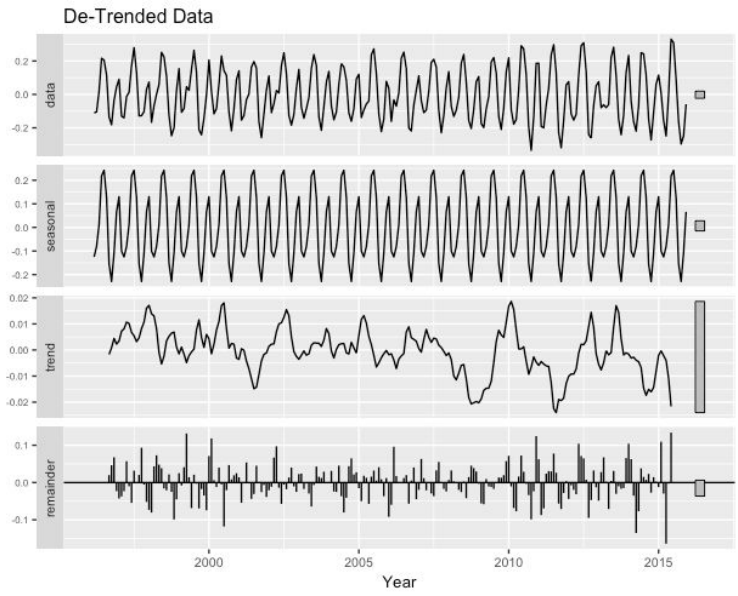
From the plots, it is clear that there is an increase in carbon emissions in the years between 2005 and 2010. There is a decay going downwards in the ACF plot, which implies non-stationary. The PACF shows a significant spike at lag 12. In addition, there are high spikes as we move towards lag 12 before the plot fits inside the interval, indicating within seasonal trends. Knowing that, we know that we want to difference the log-transformed data at lag 2 for quadratic trend and at lag 12 for seasonality. Upon examining the ACF and PACF plots, we can conclude that detrending and deseasonalizing is necessary.



The decomposition plot allows us to realize that the seasonal component shows consistent peaks and bottoms generated by carbon emission over the 12 month period. The seasonal component in this data set is 12, so it is necessary to difference the transformed data and remove the seasonal component at lag 12. The trending segment reveals a steady increase in carbon emissions, followed by a more dramatic decrease, suggesting it follows a quadratic trend. We first check the variance of the transformed data, then continue to difference the data until the variance increases instead of decreases. After differencing twice we see the variance increase, thus we choose to detrend at lag 2. The plots shown after decomposing, de-seasonalizing, and de-trending are shown below along with analysis of the ACF and PACF plots after applying log-transformation for the years 1996 to 2015.

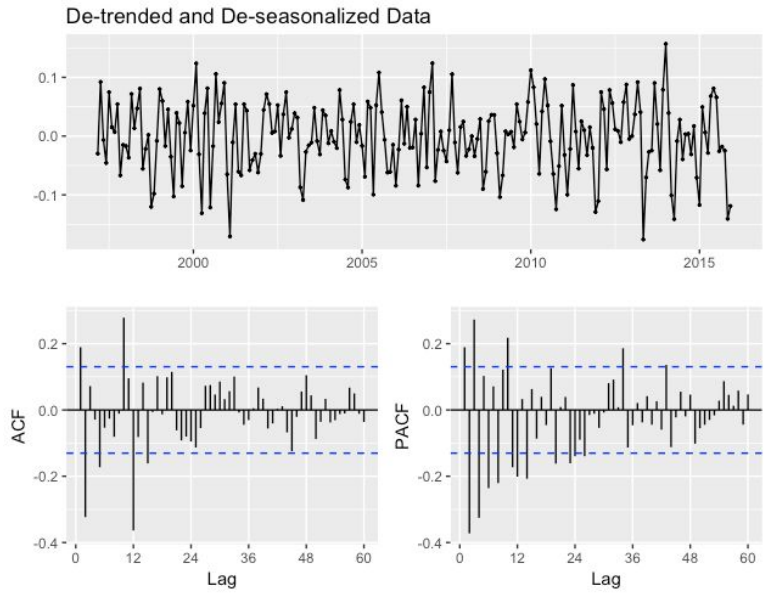


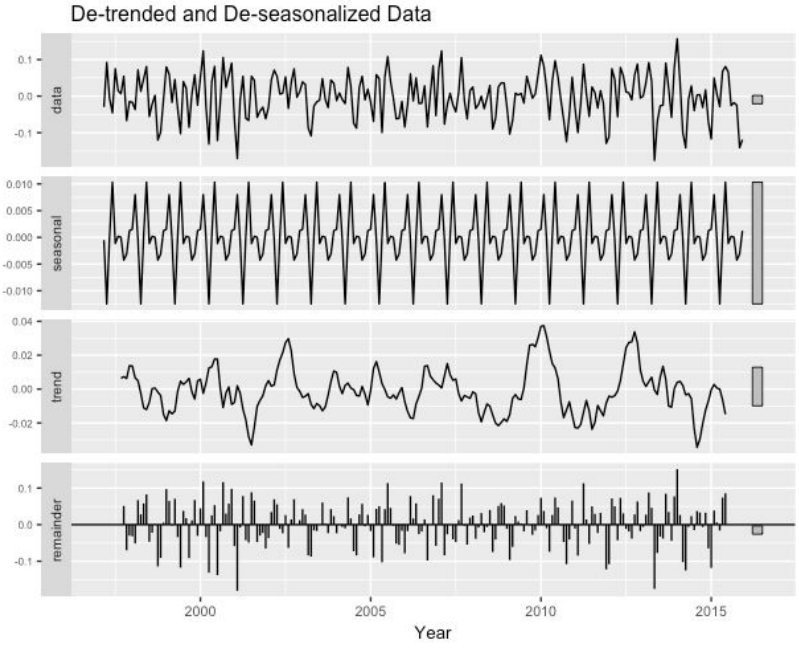
From this detrended plot, we can see that after differencing at lag 2, the time series of the data no longer shows obvious characteristics of a trend component suggesting that differencing the trend is indeed a correction step to fitting the model. Furthermore, the ACF also shows less spikes behaving within an interval of 12 and PACF also shows significantly less spikes as we approach lag 12, indicating successful removal of certain trends within the season. Below, we also composed a break down of the detrended plot and it does indeed align with our observations of stabilizing the data.



Compared to the original data, simply de-trending the data at lag 2 gives us better ACF and PACF plots. However, as seen from the decomposed plot of the detrended data, there are still trends

showing even after we removed the quadratic trend, suggesting a better solution could be possible and we want to see if we can get the best plots by breaking down the monthly data. We proceed by de-seasonalizing at lag 12.





Compared to the trended data shown earlier, we can see that the deseasonalized data shows a lower variance than just the purely detrended one as indicated by the lower range of possible y values in the graph, suggesting that a better solution of the model treatment does exist after

de-trending. Furthermore, we see that the ACF and PACF starts to tail off, and are stationary and within our confidence intervals after certain lags, indicating a more consistent plot.

Using the information gathered from the analysis and break down of the time series plots after detrending and deseasonalizing, it allows us to proceed to the next step - model building.

Model Building

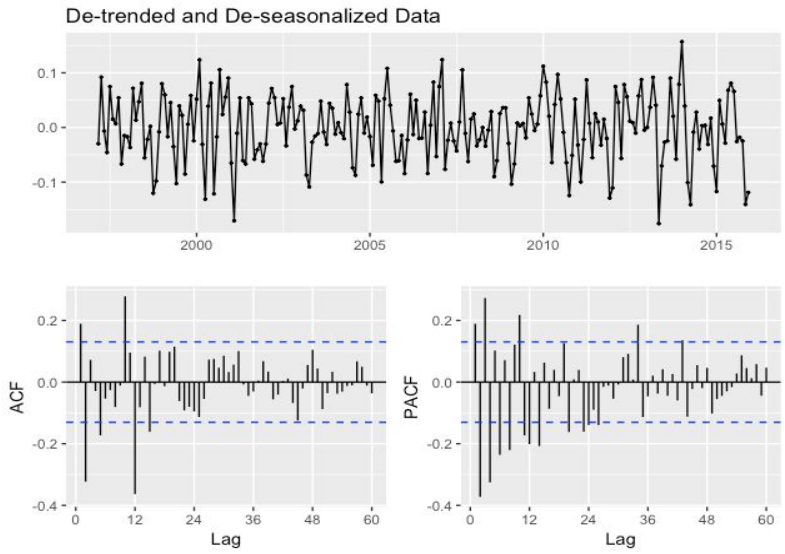
Preliminary Identification

To begin our model identification and build, we notice our data indicates a SARIMA model, which is simply a seasonal ARIMA model, due to the presence of both seasonal and non-seasonal terms. Once identifying SARIMA, our next task was to identify the order. SARIMA model notation is SARIMA(p,d,q)×(P,D,Q)s, where p, d and q represent non-seasonal AR order, differencing, and MA order respectively. Similarly, P, D and Q represent seasonal AR order, differencing and MA order, with s representing the length (in time) until the data repeats itself.

In order to stabilize our data, we differenced the log-transformed data at lag 12, to remove seasonality, and at lag 2 to remove the trend; thus, d = 2 and D = 1. We also determined s to equal 12 as our data is monthly. Now, we need to determine p, q, P and Q to produce an optimal model.

Analysis of ACF and PACF

To determine the order of our SARIMA model components, we take a look at the ACF and PACF plots for our de-trended and de-seasonalized data.



ACF plot

- Seasonal Terms (Q): Cuts off at lag 1. Q={1}
- Non-seasonal Terms (q): Cuts off at or tails off q={0,2}

PACF plot

- Seasonal Terms (P): Cuts off at lag 1. P={1}
- Non-seasonal Terms (p): Cuts of at lag 2 or at lag 3 or tails off. p={0,2,3}

This analysis results in a few different potential SARIMA models, where the model can contain an array of orders, ranging from p={0,2,3},q={0,2},P={1},Q={1} . Thus, more analysis is needed.

Model Selection

To identify models, we needed a way to compare many different models and the table below is the result of looping through four values for each parameter. Once parameters were established, tests were run and filtered resulting in the most optimal AICc, Shapiro and Ljung Box values to pick the two best models. All of the Shapiro and Ljung Box test P-values for models of interest are above 0.05 so we fail to reject the null hypothesis that the data is normally distributed, and that the data is independently distributed, meaning any correlation in the data is the result of randomness as we want.

	X.p.	X.q.	X.P.	X.Q.	X.AICc.	X.Shapiro.	X.LjungBox.
32	3	3	0	1	-5.603916	0.9580327	0.27682037
48	3	3	0	2	-5.596449	0.9445620	0.28330820
96	3	3	1	1	-5.596097	0.5973038	0.08947832
64	3	3	0	3	-5.595137	0.6515648	0.32822286
112	3	3	1	2	-5.589225	0.9474816	0.30549140
24	1	3	0	1	-5.585862	0.9404230	0.35519965

- Results after choosing d=2 and D=1 representing differencing in trend and seasonality and our screening criteria was sorted based on lowest AICc value

Model Fitting

After screening through the criterion, our two final models to be compared are:

$$SARIMA(3,2,3) \times (0,1,1)_{12} \quad SARIMA(3,2,3) \times (0,1,2)_{12}$$

based on the table above showing the AICc (Akaike Information Criterion), Shapiro and Ljung Box values. We prioritized choosing the lowest AICc since we have a relatively small dataset so the normal AIC will likely overfit the data. By using this, we assume the model has linear parameters and normally distributed residuals which we now check.

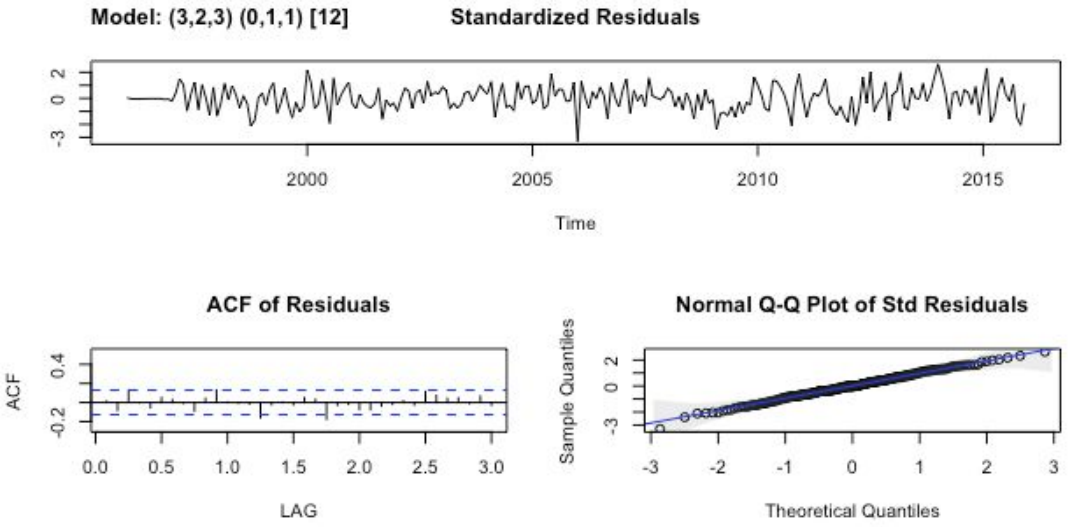
Diagnostics

After identifying and estimating a time series, model, we must confirm that the assumptions, residuals' normality, serial correlation, and heteroscedasticity of the SARIMA models are appropriate to proceed to the forecasting step. This is a crucial step, and for a time series model, we had to make sure that our fitted residuals of the model have similar properties as a White Noise distribution with $(0, \sigma^2)$.

We want to start off by drawing normal QQ-plots of our standardized residuals. The QQ plots should show points that are roughly on a straight line. For our first model:

$$SARIMA(3,2,3) \times (0,1,1)_{12}$$

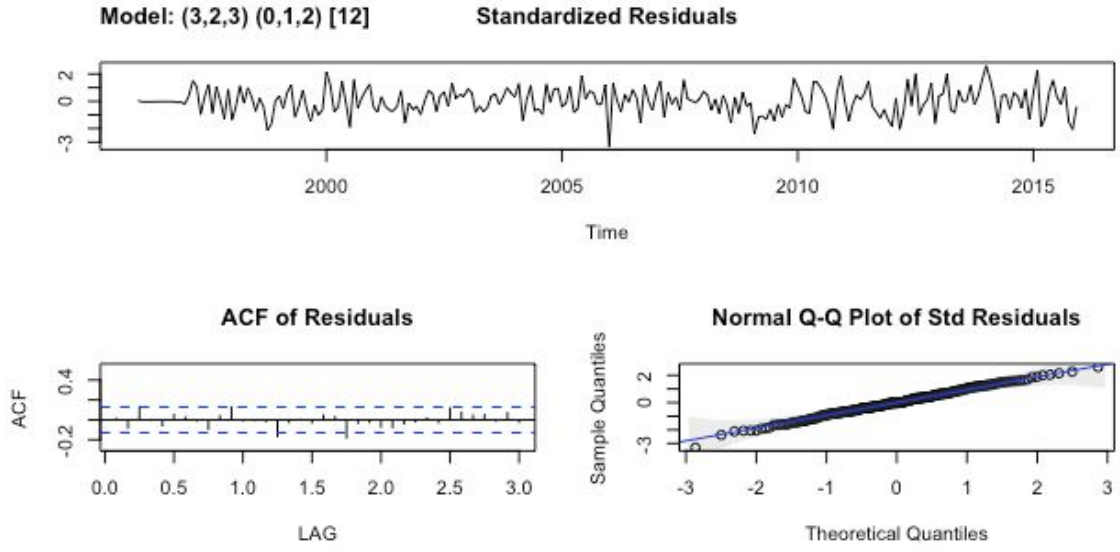
We see that the normal QQ plot reveal success in passing the normality test because the residuals are plotted in a shape of a very straight line.

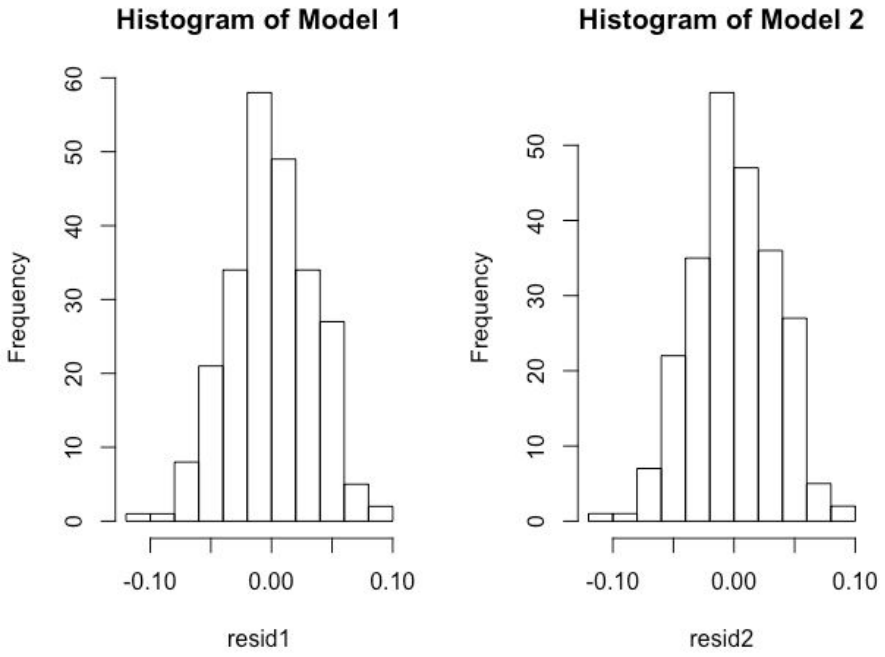


For our second model:

$$SARIMA(3,2,3) \times (0,1,2)_{12}$$

The residuals are plotted on a very straight line, similar to the first model.





The final plot we must examine before declaring success in the normality test is to analyze the histogram of the residuals. Both models have a fairly similar histogram that resemble a bell shaped curve. We can finally conclude that there is success for both models to pass the normality test.

Furthermore, we notice that the ACF values of squared residuals variance for both models 1 and 2 fall into the 95% confidence region. This a positive indication that implies an absence of heteroskedasticity, or violation of the constant error variance assumption, within either of the models. These models do not have heteroskedasticity so it passes the constant errors variance checking.

After conducting the series of tests, we can conclude that neither model 1 nor model 2 have a diagnostic problem. Thus, because of the similarities between the two models, model 1 is ultimately picked for its lowest AIC value and fewer parameters with the following coefficients.

```
> model1$fit$coef
        ar1         ar2         ar3         ma1         ma2         ma3        sma1    constant
-0.35243338  0.60426203 -0.01580566 -1.01963483 -0.92841671  0.94808696 -0.81018613  0.03235685
```

Thus, the final model is

$$SARIMA(3,2,3) \times (0,1,1)_{12}$$

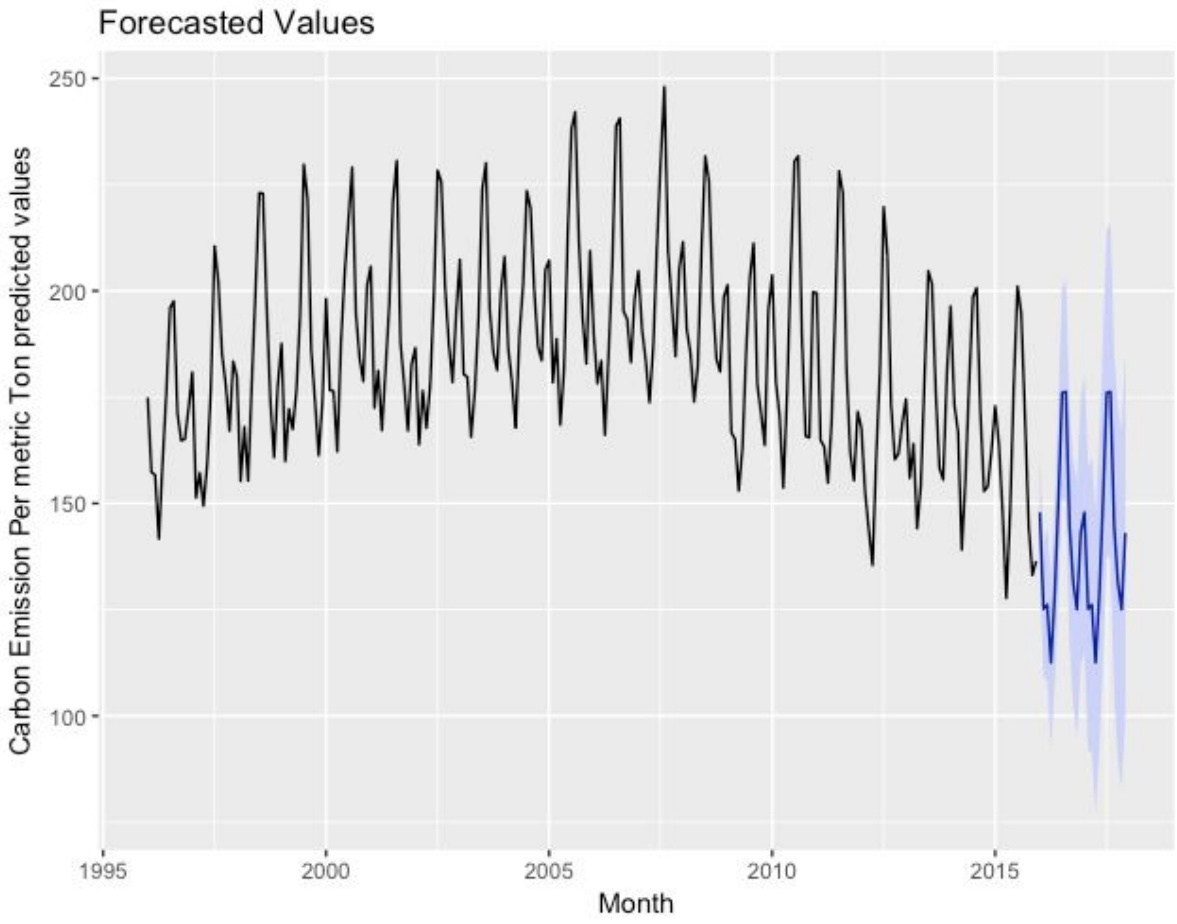
$$(1+0.3524B-0.6043B^2+0.0158B^3)\nabla_2\nabla_{12}Y_t = (1-1.0196B-0.9284B^2+0.9481B^3)(1-0.8102B^{12})Z_t+0.03236$$

Forecasting

Forecasting plays an important role in allowing us to apply our analysis to applicable, real world problems. We are also able to use our data to forecast and figure out which direction the carbon dioxide emissions will go towards in the next 1-2 years.

Now that we have a better understanding of the model we built using the training data, we will apply our chosen model to the remaining observations in 2016 that we intentionally left out at the beginning. We did this so that we can compare side by side how closely the model predicts the true values. In looking at the forecasted values, we see that they follow the seasonality and trend we accounted for in building the model which allows it to more accurately predict the future values. We notice that the confidence intervals for forecasted values increase further away from the actual data we get, so there is a limited range ahead in which it is useful to predict.

Based on the interpretation and conclusion we draw, the model would be beneficial in many different areas. Although the data studied is more broad since it is the total CO2 levels, the same procedure can be run on the components that make up that total as well. Oil companies could use forecasted data to see the effect their industry has on the total, or governments could use similar models to introduce regulations and insure the downward trend continues going forward.



	Point Forecast	Lo 95	Hi 95
Jan 2016	147.8895	134.48899	161.2900
Feb 2016	125.1367	109.27020	141.0033
Mar 2016	126.1625	108.16471	144.1603
Apr 2016	112.5592	92.65707	132.4613
May 2016	128.3768	106.73737	150.0163
Jun 2016	150.5462	127.29887	173.7936
Jul 2016	176.0384	151.28735	200.7894
Aug 2016	176.2669	150.09849	202.4354
Sep 2016	144.6511	117.13819	172.1640
Oct 2016	131.3373	102.54264	160.1320
Nov 2016	125.0628	95.04098	155.0845
Dec 2016	143.0313	111.83030	174.2323

Predicted Values of the Series

TXEIEUS	2016-01	159.03
TXEIEUS	2016-02	133.106
TXEIEUS	2016-03	115.997
TXEIEUS	2016-04	113.815
TXEIEUS	2016-05	129.44
TXEIEUS	2016-06	172.074
TXEIEUS	2016-07	201.958

Actual values omitted from data to test for accuracy

Conclusion

As with any typical project on time series, our overall goal was to understand the carbon emissions data well enough to create and pick a model to use in forecasting values for the future and address specific questions of interest.

We first looked at the overall structure of the data by plotting it, which clearly showed seasonality from the periodic drops and spikes in emissions, as well as a non-linear trend. It is important to note that this trend and seasonality can only be attributed to this timespan in which data was collected, and we should not extrapolate that carbon emissions always had or will always have these trends. For that we would need a much larger dataset and most likely a more complicated model.

Running a Box Cox test resulted in our decision to log transform the data to stabilize the variance, then differencing the data was our way of dealing with the seasonality and trend, once at lag 12, and once at lag 2. Lag 12 was chosen since the data follows monthly patterns, and lag 2 was used since it follows a quadratic trend and this produced the lowest variance.

The process of model building, selection and checking are next. We identify a model based on the ACF and PACF plots initially, then calculate values to compare different models and determine which we want to choose. To do this, we run through and compare many different parameter values based on the AICc numbers that give us an estimate of the quality of each model in relation to the others. Overall, it is balancing the goodness of fit and the simplicity of models, so the model chosen is not unnecessarily complex but still explains the data. Along with this, we insure the data is normal and the points are independently distributed via the Ljung Box and Shapiro tests. To check the models, we test that the residuals are normally distributed, then examine the serial correlation and heteroscedasticity. Ultimately, we find that both of the models do indeed work very well for our fitted data.

However,

$$SARIMA(3,2,3) \times (0,1,1)_{12}$$

$$(1+0.3524B-0.6043B^2+0.0158B^3)\nabla_2\nabla_{12}Y_t = (1-1.0196B-0.9284B^2+0.9481B^3)(1-0.8102B^{12})Z_t+0.03236$$

was ultimately selected due to a lower AIC value and fewer parameters.

Using the chosen model, we move on to forecasting future values. Since the model was built using the dataset with some of the last values removed, we forecast ahead and compare the graphs of the test dataset that includes all the values to the model and were able to successfully predict the values of the dataset within a 95% confidence interval.

Overall, we were proud of ourselves for challenging ourselves and applying our knowledge towards this project. For example, we were hesitant to interpret the quadratic trends, however, with time and patience, we were able to find an appropriate model to proceed with the analysis process. Furthermore, we ran into some difficulties in interpreting the ACF and PACF plots as accurately as possible. It was mostly difficult because the ACF and PACF analysis did not correlate with our AIC models completely, but we were eventually able to find a SARIMA model to use.

In the end, we were able to achieve our goals and understand and connect our time series project to the real world, specifically in the field of electricity production. We noticed that the summer season produced the most CO2 emissions, specifically around the months of July and August. This is reasonable because people are more inclined to turn on their air conditioning machines to keep cool in the heat. In fact, air conditioners use 6% of all electricity produced within the United States, resulting in 100 million tons of CO2 emissions each year. We noticed that the means are decreasing in the next 1-2 years. This might be due to the technological advances that have been emerging throughout the world today so that we are able to switch to more energy and gas efficient sources. People could be making more effort to save the Earth because they are more aware of global warming and the harmful uses of CO2 emissions.

References

1. "8.7 ARIMA Modelling in R." *4.4 Evaluating the Regression Model | OTexts*, www.otexts.org/fpp/8/7.
2. Bapat, Sudeep. "Take a Break! ." *SoundCloud*, soundcloud.com/sudeepbapat. 😎
3. Fuqua School of Business. "Identifying the Orders of AR and MA Terms in an ARIMA Model." *Seasonal Differencing in ARIMA Models*, people.duke.edu/~rnau/411arim3.htm
4. Grossman, Daniel. "Pros and Cons on 'Negative Emissions' Prospects » Yale Climate Connections." *Yale Climate Connections*, 24 May 2017, www.yaleclimateconnections.org/2017/05/pros-and-cons-on-negative-emissions-prospects/.
5. McNeill, Jason. "Carbon Emissions - Kaggle." *Countries of the World - Kaggle*, 6 Nov. 2016, www.kaggle.com/txtrouble/carbon-emissions.
6. Schlossberg, Tatiana. "How Bad Is Your Air-Conditioner for the Planet?" *The New York Times*, The New York Times, 20 Jan. 2018, www.nytimes.com/2016/08/10/science/air-conditioner-global-warming.html.

Appendix

```
# packages
library(MASS)
library(qpcR)
library(stats)
library(survMisc)
library(ggplot2)
library(ggpubr)
library(astsa)
library(forecast)

# read in data
getwd()
setwd("/Users/denniswang/Desktop")
emissiondata<-read.csv("/Users/denniswang/Desktop/MER_T12_06.csv",header=TRUE)
TXEIEUS<-read.csv("/Users/denniswang/Desktop/Pstat 174 files/TXEIEUS.csv",header=TRUE)
# time series data
Total_Energy_Electric_Power <- ts(TXEIEUS[,3],start=c(1996,1),frequency=12)

# time series
ts=ts(TXEIEUS$Value,start=c(1996,1),frequency=12)
ts=tsclean(ts)
t = 1:length(ts)
fit = lm(ts ~ t)

# transformation
bcTransform = boxcox(ts ~ t,plotit = TRUE)
lambda <- bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
# close to 0 so we'll use log of the data
ts.log<-log(Total_Energy_Electric_Power)
ts.log<-tsclean(ts.log)
plot(ts, main = "Original Time Series")
```

```
plot(ts.log, main = "Log-Transformed Time Series")

# decomposing
ggtsdisplay(ts,lag.max=120, main = "Original Time Series with ACF and PACF")
Decomp=decompose(ts) # survMisc package
autoplot(Decomp,main="Decomposed Time Series") +
theme(axis.text.y=element_text(size=6),text=element_text(size=10))+ xlab("Year") #forecast
package
 #log transformed decompose data
ggtsdisplay(ts.log,lag.max=120, main = "Log-Transformed Time Series with ACF and PACF")
Decomp2=decompose(ts.log) # survMisc package
autoplot(Decomp2,main="Decomposed Time Series") +
theme(axis.text.y=element_text(size=6),text=element_text(size=10))+ xlab("Year")

# detrending the data
dif2 = diff (ts.log , lag = 2)
ggtsdisplay ( dif2 , lag.max = 60, main = "De-Trended Data")
autoplot(decompose(dif2),main="De-Trended Data") +
 theme(axis.text.y=element_text(size=6),text=element_text(size=10))+
 xlab("Year")

# de-seasoning the data
dif12 = diff ( dif2 , lag = 12)
ggtsdisplay ( dif12 , lag.max = 60, main = "De-trended and De-seasonalized Data")
autoplot(decompose(dif12),main= "De-trended and De-seasonalized Data") +
 theme(axis.text.y=element_text(size=6),text=element_text(size=10))+
 xlab("Year")

# generate possible models to use with a matrix of the AICc values to compare
Fit = list ()
AICc=matrix (, nrow =200 , ncol =7)
colnames ( AICc ) =c (" p " ," q " ," P " ," Q " ," AICc " ," Shapiro " ," LjungBox ")
i =0
```

```
for (P in c (0:3) ) {
 for (Q in c (0:3) ) {
  for (p in c (0:3) ){
   for (q in c (0:3) ){
    Fit[[i+1]]=sarima (ts.log,p ,2 ,q ,P ,1 ,Q ,12 , Model = TRUE , details = FALSE )$fit
    plot.new()
    AICc[i+1,1]=p
    AICc[i+1,2]=q
    AICc[i+1,3]=P
    AICc[i+1,4]=Q
    AICc[i+1,5]=sarima(ts.log,p,1,q,P,1,Q,12,Model=TRUE,details=FALSE)$AICc
    plot.new()
    AICc[i+1,6]=shapiro.test(resid(Fit[[i+1]]))$p.value
    AICc[i+1,7]=Box.test(resid(Fit[[i+1]]),type=c("Ljung-Box"),lag=12)$p.value
    i=i+1
   }
  }
 }
}
AICc<-data.frame(AICc)
# sort by AICc ( Increasing order)
AICc_sorted<-AICc[order(AICc$X.AICc.),]
# Filter the models by Shapiro test
AICc_cmp<-subset(AICc_sorted,AICc_sorted$X.Shapiro.>0.05)

# Filter the models by LjungBox test (for checking independence)
AICc_cmp<-subset(AICc_cmp,AICc_cmp$X.LjungBox.>0.05)

# final models to consider
View(AICc_cmp)

model1 = sarima(ts.log ,3 ,2 ,3 ,0 ,1 ,1 ,12) #astsa package
model2 = sarima(ts.log ,3 ,2 ,3 ,0 ,1 ,2 ,12)
```

```
#Estimating coefficients
model1$fit$coef
 # diagnostic checking

# 1. normality
par ( mfrow = c (1 ,2) )
resid1 = residuals(model1$fit )
resid2 = residuals(model2$fit )
qqnorm (resid1,main ="Model 1 Normal Q - Q Plot")
qqline(resid1,col ="blue")
qqnorm (resid2,main ="Model 2 Normal Q - Q Plot")
qqline(resid2,col ="blue")
par(mfrow=c(1,2))
hist(resid1, main = "Histogram of Residuals from Model 1")
hist(resid2, main = "Histogram of Residuals from Model 2")

# 2. residual plots (when they fall within the confidence intervals
# then we know that they have constant variance. They almost all fall
# within the boundaries so we can assume this.)
ggtsdisplay ( resid1 , main ="Residual Plots for Model 1" , xlab =" Year ")
p1 = ggAcf ( resid1 ^2 , lag.max = 60 , main ="")
p2 = ggPacf ( resid1 ^2 , lag.max = 60 , main ="")
ggarrange ( p1 ,p2 , nrow = 2, ncol = 2) #ggpubr
ggtsdisplay ( resid2 , main ="Residual Plots for Model 2" , xlab =" Year ")
p11 = ggAcf ( resid2 ^2 , lag.max = 60 , main ="")
p22 = ggPacf ( resid2 ^2 , lag.max = 60 , main ="")
ggarrange ( p11 ,p22 , nrow = 2, ncol = 2) #ggpubr

# choosing our fit
fit=model1 #sarima (3,2,3)x(0,1,1)_12

# Forecast and transfer to original scale
```

```
library(forecast)

TXEIEUS2 <-read.csv("/Users/denniswang/Desktop/Pstat 174 files/TXEIEUS2.csv",header=TRUE)
ts2 <- ts(TXEIEUS2$Value,start=c(1996,1),frequency=12) # we include the actual values #for 2016
ts2.log<-log(ts2)

predict<-forecast(ts, level=c(95))
predict2<-forecast(ts2.log, level=c(95)) #log-transformed one
#Original forecasting transformation
autoplot(ts2,main="Original Data") + ylab("Carbon Emission Per metric Ton actual
values")+xlab("Month")+theme(legend.position="None")
autoplot(predict, main="Forecasted Values") + ylab("Carbon Emission Per metric Ton predicted
values")+xlab("Month")+theme(legend.position="None")

#Log transformed
autoplot(ts2.log,main="Log-transformed Original Data") + ylab("Carbon Emission Per metric Ton
actual values")+xlab("Month")+theme(legend.position="None")
autoplot(predict2, main="Log-transformed Forecasted Values") + ylab("Carbon Emission Per
metric Ton predicted values")+xlab("Month")+theme(legend.position="None")
```