Atmostpheric CO2 Concentration in Hawaii: Time Series Analysis

Roupen Khanjian, Drew Pritchard, Michael Wang

Supervised by Professor Sudeep Bapat

Contents

Abstract

A controversial topic of today is the effect of polution on the biosphere. From climate change to the death of the great barrier reef it seems that a key argument is that industrialized societies are manufacturing and irresponsibly ignoring their CO_2 emissions. Throughout this project we answer the question of just how much the CO_2 concentration our atmosphere has been changing in recent decades. We also dive into how - if at all - the concentration of CO_2 behaves on a micro-scale, aka monthly basis, rather than just its behavior between a several decade period. In order to address these questions, we use time series analysis consisting of exploratory data analysis, data transformation, differencing, seasonal autoregressive integrated moving average model selection, residual diagnostic checking and spectral analysis. Finally, to check any predictive power we may have, we forecasted a year ahead of our analyzed data.

```
## Parsed with column specification:
## cols(
##   Month = col_character(),
##   `CO2 (ppm) mauna loa, 1965-1980` = col_character()
## )
```

1 Introduction

In 2013 NASA scientists detected C02 levels of above 400ppm from the Mauna Loa Observatory in Hawaii.[1]. For scientists, this was noted as an unfortunate milestone.The rapid rise in C02 levels in our atmosphere for the past century has contributed to many catastophic climate disastors. Understanding in what ways the C02 levels have been incresing throguhout time is an important aspect to solving the global crisis of climate change.

We decided to analyze the C02 levels recorded monthly from the Mauna Loa Observatory in Hawaii during the years 1965-1980 [2]. This data set consists of 192 data points where the C02 levels were measured monthly. Taking out the last year of data points (12 data points), our training data set consisted of 180 points.

Using Rstudio, we Box-cox transformed, de-trended, and de-seasonalized the time series values. After making sure our data was stationary we went ahead with selecting an appropraite SARIMA model to describe our data. While verifying our models are invertible and stationary, we used AICc and BIC values to narrow down our search for an ideal model.

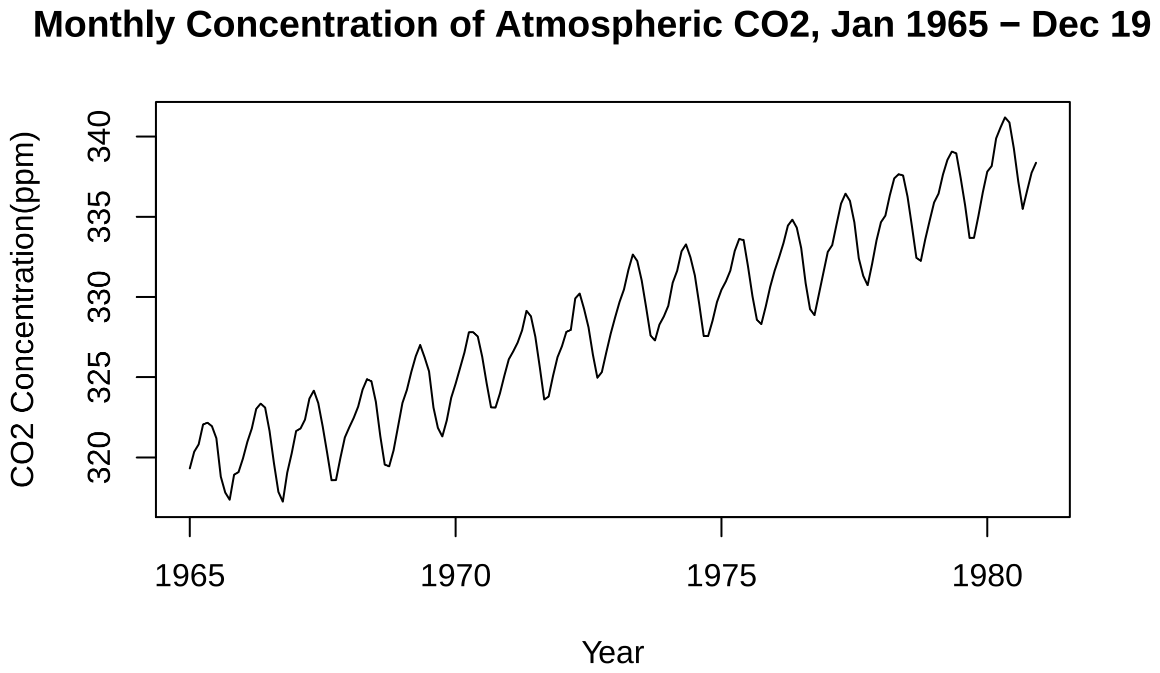
Next we tested the models with different diagnostic checks including Shapiro-Wilk test for normality, Ljung-box test for serial correlation and other tests to ensure our final model would be valid. Since 2 of our models passed the necessary diagnostic checks, we chose the model based on the lowest BIC values with happened to be a $SARIMA(0, 1, 1)$x$(1, 1, 0)_{12}$

Afterwords we forecasted a year ahead and made sure our forecasts coincices with the obsereved C02 levels that year. Finally we conducted spectral analysis on the stationary data and concluded that the data can be modeled using a linear combinations of 12 sinisuoids.

2 Data Exploratory Analysis

2.1 Preliminary Exporation

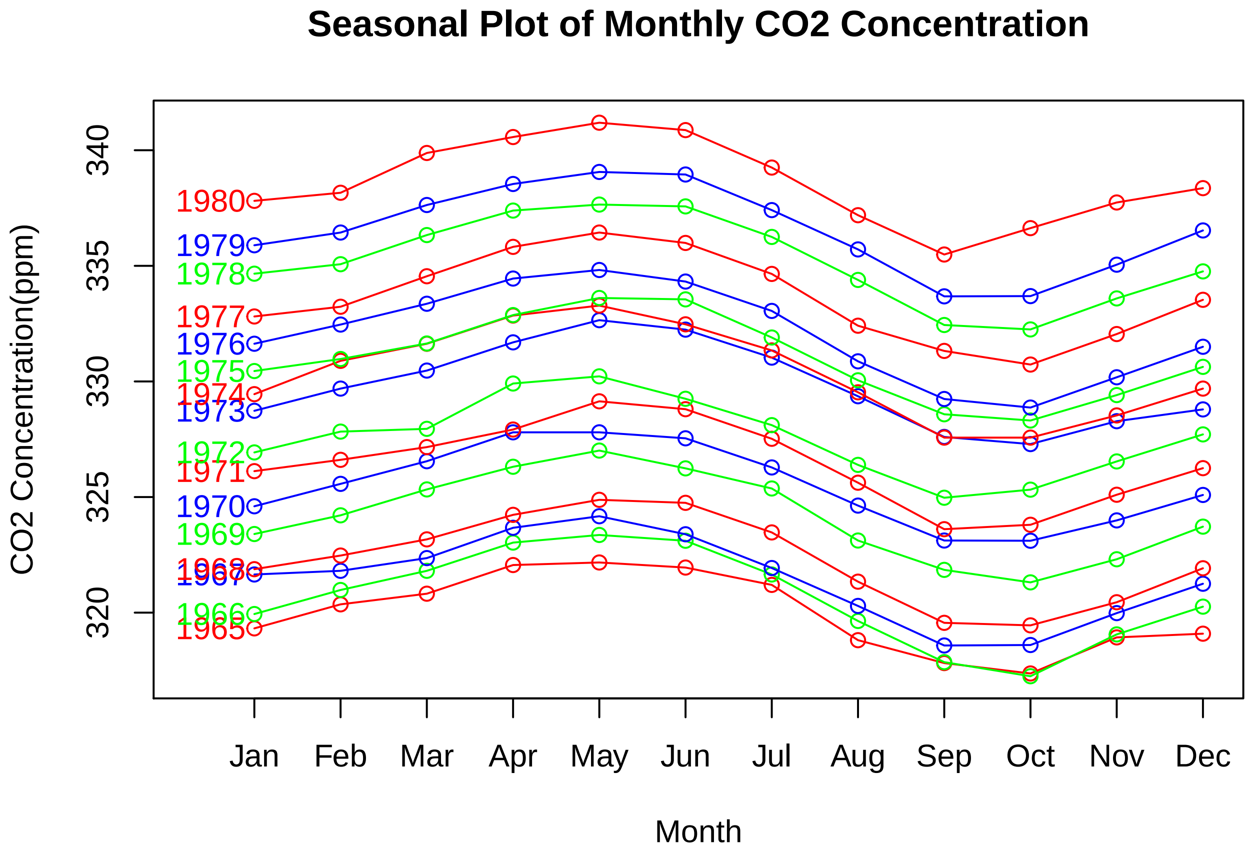
The data we have contains two variables: the Month which also includes the year in which it was recorded and quarter as well as production of clay data in million units. We have 155 observations in total and will reserve the last 15 observations as our testing data for forecasting.



The plot clearly shows an upward trend and yearly seasonality. The presence of seasonality is interesting because it displays evidence that concentration of CO_2 is not constant throughout the year; for some reason it spikes in the spring and reaches a minimum in the fall. We will need to control for this seasonality soon in order to reach a stationary process model with which we can use to forecast.

Fortunately, we don't see a clear display of non-constant variance. Every year's dips and spikes are close to if not the same length from each other over the course of all of our data.

To more closely examine the seasonal pattern we have observed, we create a seasonal plot by month for each year we have data for.



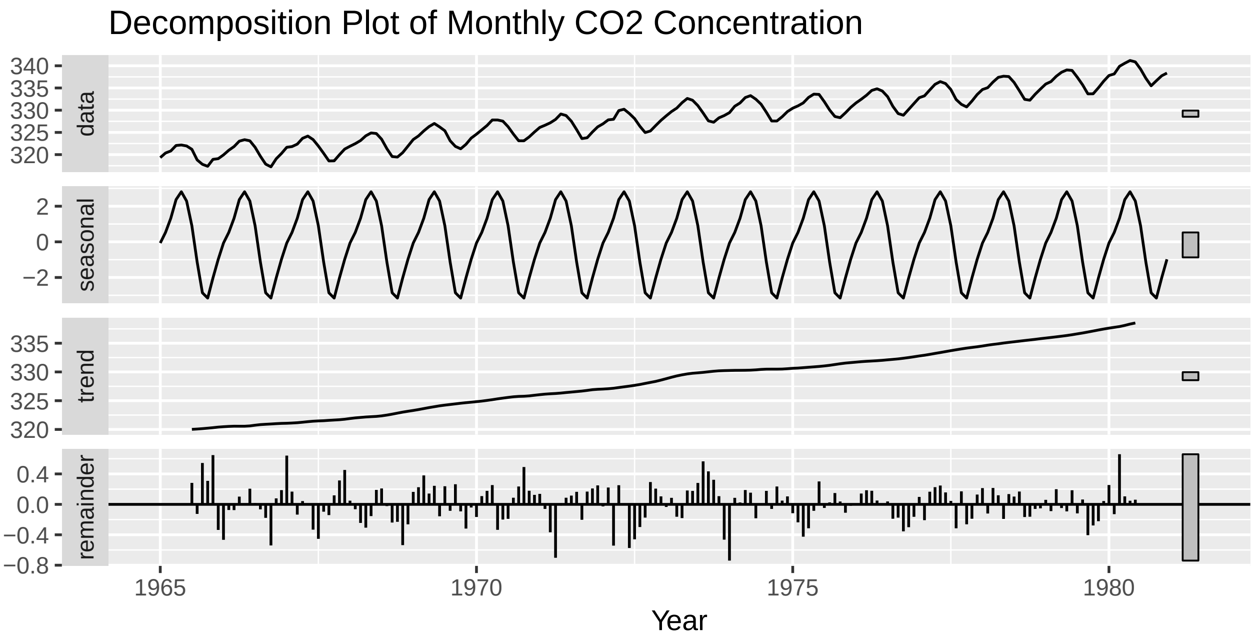
By splitting each year's data and plotting them in parallel we again see what was aforementioned: a gradual increase in ppm from October to May, then a gradual decrease from May to October. The pattern is repeated every year, leading to a conclusion of a strong seasonality component in our data.

It is also apparent that the ppm for each month in each year is directly higher than it's predecessor - a result of the upward trend in our data.

2.2 Decomposition

By splitting our data into three distinct parts - trend, seasonality, and cyclical components - we can further explore the nuances of the data.

If we let Y_t be our CO2 data where $Y_1 =$ Jan 1965, $Y_2 =$ Feb 1965, etc. Then we can write a decomposition model of our data as $Y_t = m_t + s_t + S_t$.

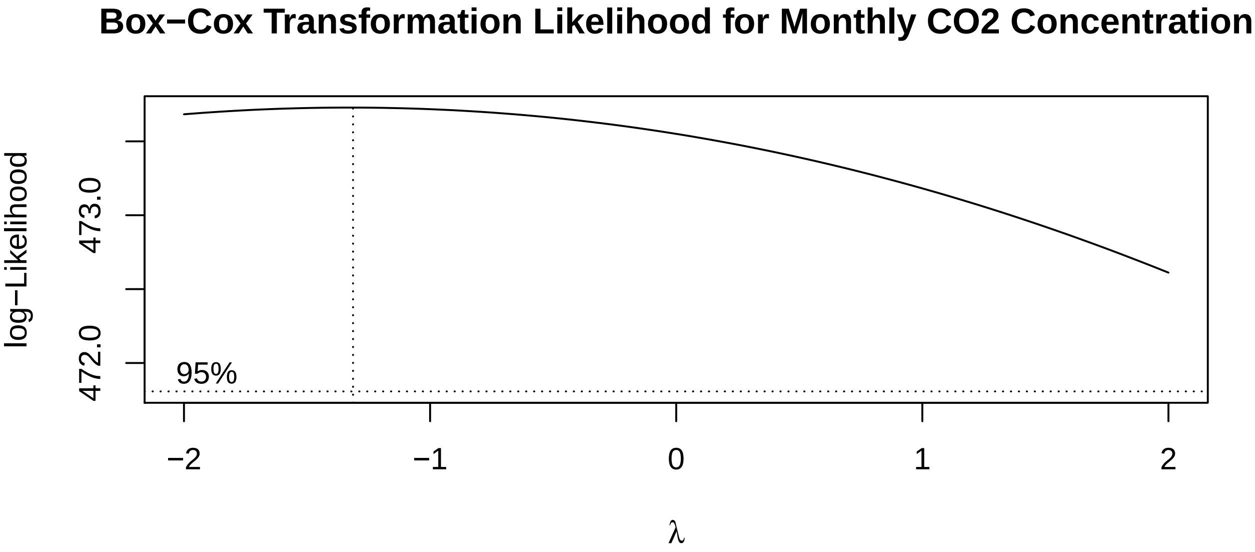


3 Transformations

3.1 Variance Stabilization via Box-Cox

Although there was no obvious heteroskedacity(non-constant variance) when we originally plotted our data, we still want to stabelize and reduce our variance as much as possible.

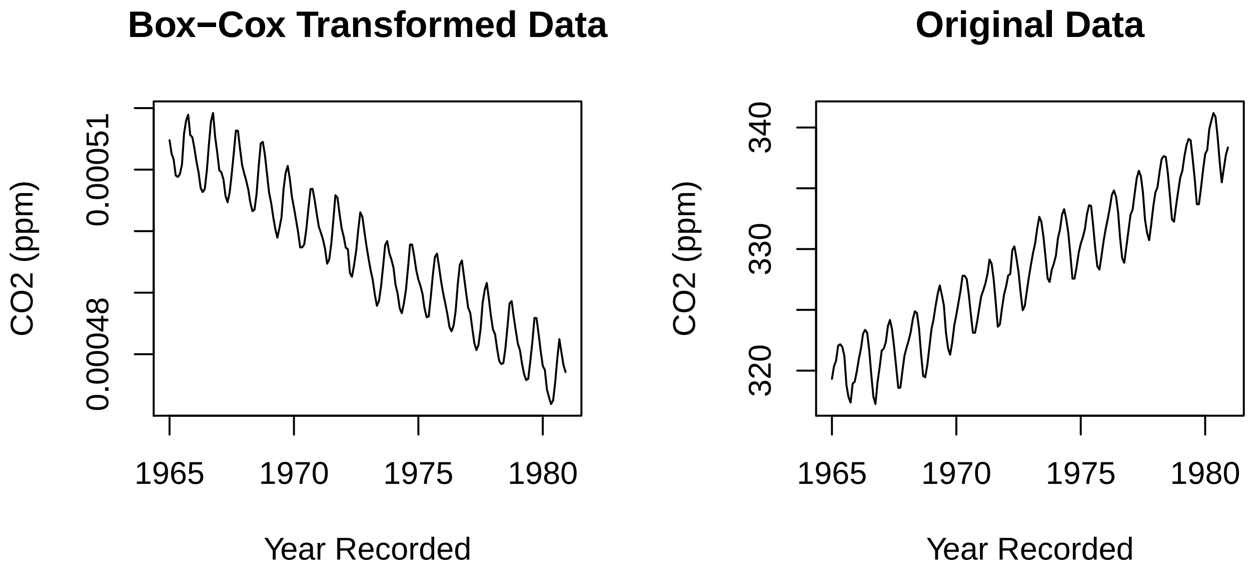
A simple and and often effective way to stabilize variance across time is to apply a transformation on the data. We used the Box-Cox method to find a value of λ that determines the power transformation to perform on our data. This value is chosen based on the maximized profile log-likelihood. For seasonal data, a linear time trend with seasonal dummy variables is used.



Our log-likelihood is maximized when λ = -1.3131313. However, the confidence interval of our MLE includes the value of 0 so we could decide to instead go with a log transformation. Still, we stick with -1.3131313 and transform our data into $Y_t^{.9090}$.

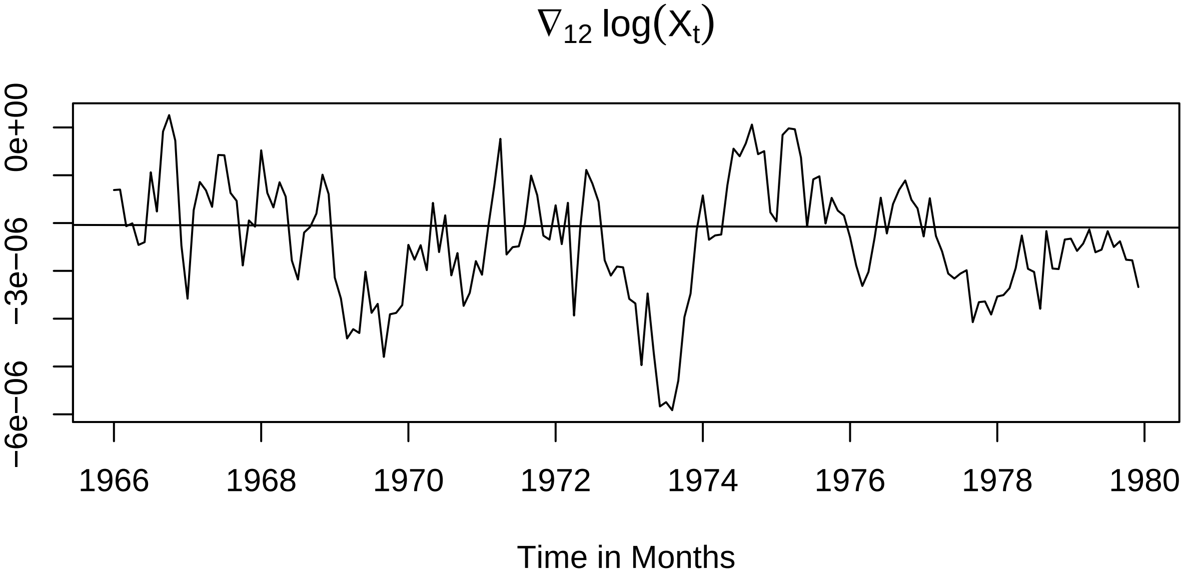
Table 1: Log of Variance of Co2 Data

	Before Box-Cox	After Box-Cox
CO2	-22.83756	-25.9675



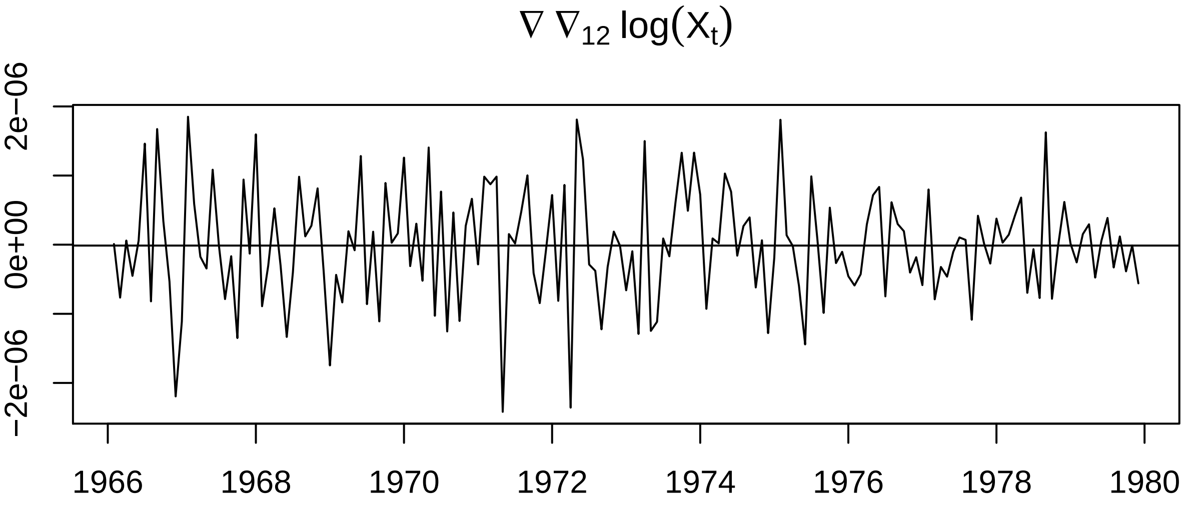
In Table 1 the difference between the variances of before and after the transformation can be seen. We have to take the log of the variances because they are so small and R naturally rounds them both to 0

3.2 Removing Seasonality



(Figure 6: Plot of CO2 Concentration Differenced at Lag 12)

Figure 6 depicts a plot of the differenced data along with a linear trend line overlayed on it. We no longer have a seasonality component but we must further difference to remove the trend.



Time in Months
(Plot of CO2 Concentration Differenced at Lag 12 and 1)

This looks much better. There

continue

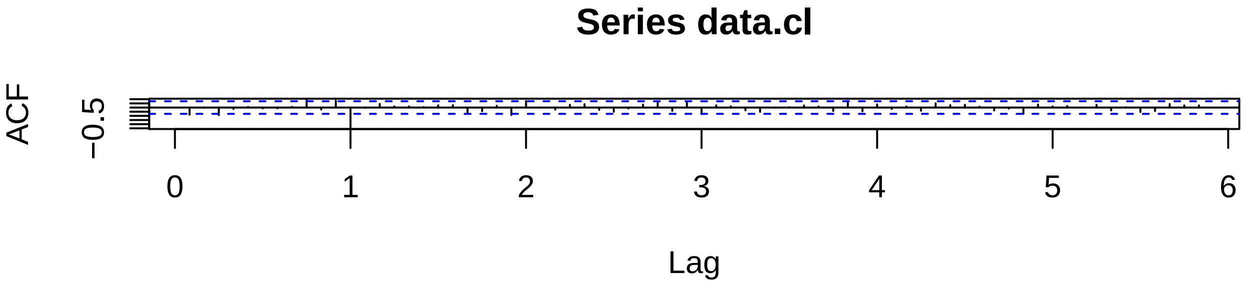
We still want to check what the effects of differencing more times would be. A common, quick way is to check the variance before and after differencing.

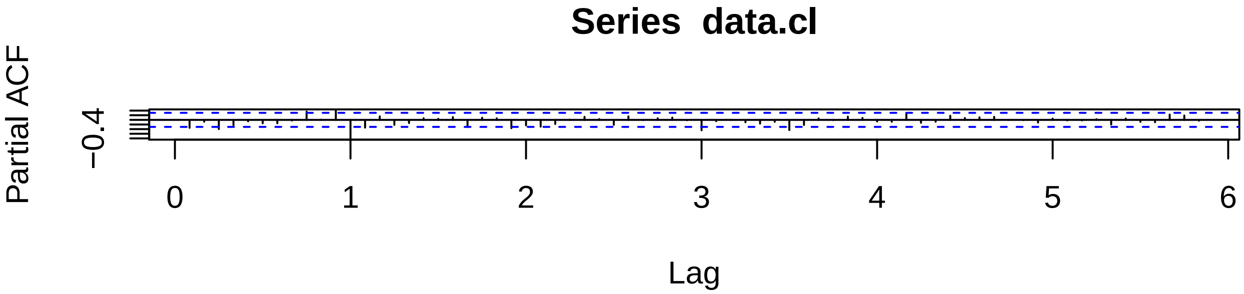
Differenced at 12,1	Differenced at 12,1,1
0	0

The variance more than doubles! It doesn't look like we should difference any further. Otherwise we would be at risk of over-differencing.

4 Model Identification, Selection, and Estimation

4.1.1 Identify Seasonal Order: P, D, Q





By looking at lags 12, 24, 36, . . . , on the ACF plot, we can determine that it tails off from lag 12 onward. Hence we have an AR part with P = 1.

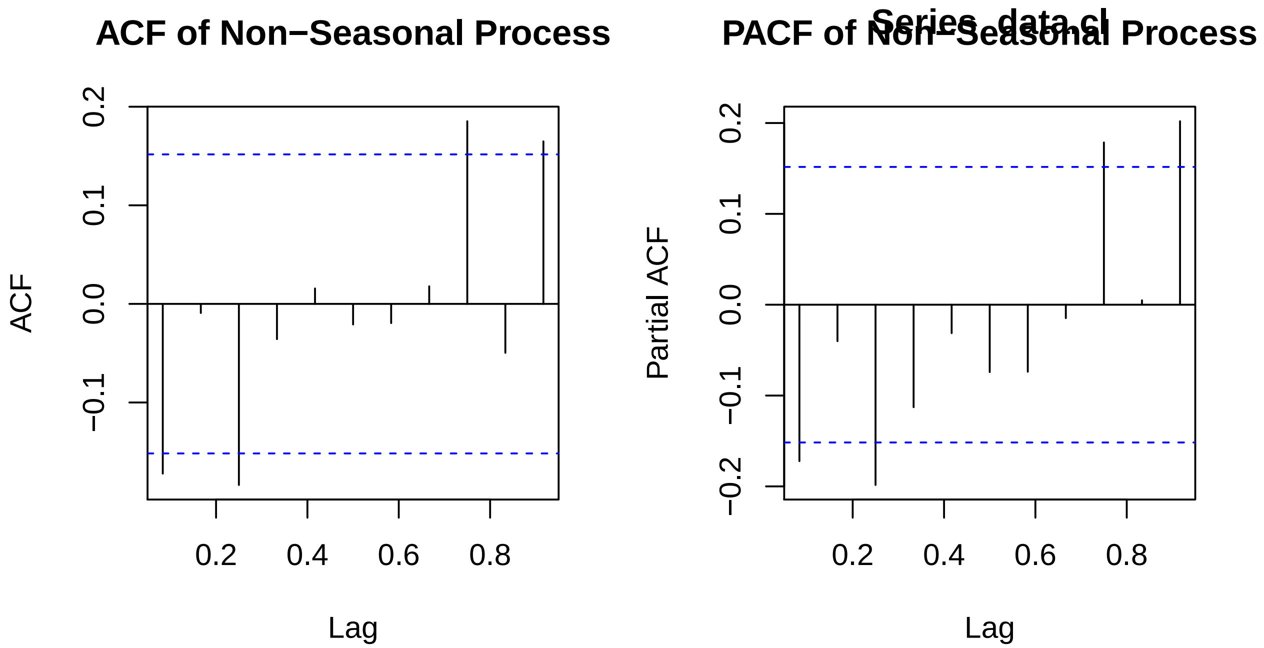
For the PACF plot it appears to just completely cut off at lag 1. Again indicating that we have a an SAR(1) component in our model. This also lets us know that there is **no** MA part, i.e. Q = 0.

While this is somewhat subjective, it could make all the difference and is quite an important piece for our model. It is quite possible that someone else looking at these plots could have concluded that the PACF actually tails off somewhat from lag 12 onward. However we stick to our original assumption that we have:

$$\text{Seasonal Part} \Rightarrow (1, 1, 0)_{12}$$

4.1.2 Identify Inter-Seasonal Order: p, d, q

The next step is for us to identify the between-season order of our model. We already know that d = 1 because of the second difference at lag 1 we did in section 3.3. To find p and q we zoom in on the ACF and PACF plots.



From our zoomed ACF and PACF plots we can see that we have spikes at lags 1 and 3 as well as 9 and 11. We will assume that the spikes at lags 9 and 11 are outliers in our dataset. PACF tails off and ACF cuts off after lag 3, which can be interpreted as an $MA(3)$ process ($p = 0, q = 3$). We should also consider the possibility where we have an $ARMA(p, q)$ process where both ACF and PACF tail off after lag 3 i.e. $max(p, q) = 3$. Because we're working with sample data, the zoomed ACF and PACF plots may not accurately depict the theoretical model. Thus we test all combinations of p and q for $p, q \in \{0, 1, 2, 3\}$ resulting in 16 preliminary models.

4.2 Model Selection

From our 16 models, we select the best 2 based off Akaike's corrected information criterion (AICc) and Bayesian information criterion (BIC). Both information criterion measure the "quality"" of a statistical model for a given dataset relative to other models by using goodness of fit and penalizing for model complexity. We wish to select the two models with the lowest AICc and BIC values.

Looking at our two tables of AICc and BIC values below, we see that the model with $p = 0, q = 3$ gives us the smallest AICc value and the model with $p = 0, q = 1$ gives us the smallest BIC value. This is expected as BIC has a larger penalty parameter for model complexity and thus favors a smaller model.

Therefore, our two models selected based on AICc and BIC are:

$$\text{Model A: } SARIMA(0, 1, 3) \times (1, 1, 0)_{12}$$

$$\text{Model B: } SARIMA(0, 1, 1) \times (1, 1, 0)_{12}$$

	MA(0)	MA(1)	MA(2)	MA(3)
AR(0)	-27.58629	-27.72067	-27.71123	-27.72657
AR(1)	-27.71424	-27.71829	-27.71207	-27.71634
AR(2)	-27.70359	-27.71658	-27.70863	-27.72174
AR(3)	-27.70846	-27.71527	-27.70571	-27.69779

	MA(0)	MA(1)	MA(2)	MA(3)
AR(0)	-28.58004	-28.69706	-28.67039	-28.66864
AR(1)	-28.69063	-28.67746	-28.65415	-28.64145
AR(2)	-28.66275	-28.65865	-28.63374	-28.63004
AR(3)	-28.65053	-28.64038	-28.61401	-28.58941

4.2 Model Estimation

Now that we have chosen two models with the best information criterion values, we need to estimate the coefficients and parameters. We are under the assumption that the data has zero mean, and that we know p and q. We will estimate the coefficients in our models using the maximum likelihood method. The results are shown below:

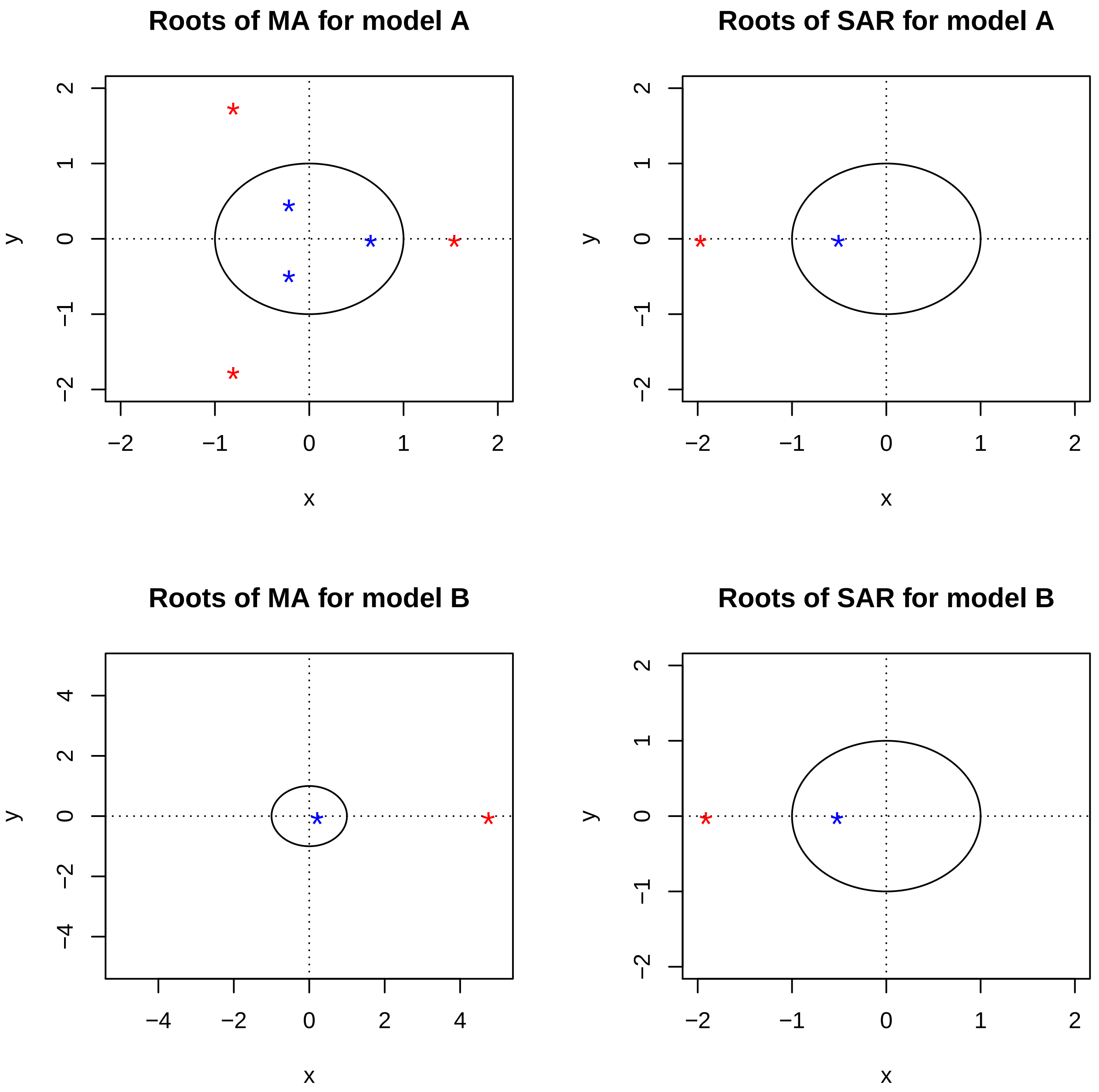
	Model A	Model B
MA(1)	-0.2177	-0.2105
MA(2)	-0.0133	NA
MA(3)	-0.1744	NA
SAR(1)	-0.5072	-0.5225

Thus our fitted models have the algebraic form:

$$\begin{aligned}
&\text{Model A: } SARIMA(0,1,3)\times(1,1,0)_{12}\\
&\quad(1+.5072B^{12})Y_t = (1-0.2177B)(1-0.0133B)(1-0.1744B)Z_t\\
&\qquad\qquad Z_t \sim N(0, 4.435e-13)\\
&\text{Model B: } SARIMA(0,1,1)\times(1,1,0)_{12}\\
&\quad(1+.5225B^{12})Y_t = (1-0.2105B)Z_t\\
&\qquad\qquad Z_t \sim N(0, 4.589e-13)
\end{aligned}$$

where $Y_t = \nabla\nabla^{12}X_t^{-1.313}$.

Next we check for causality and invertibility by examining the roots of our polynomial. Causality and invertibility are implied when the roots of our polynomial in the AR and MA, respectively, lie outside the unit circle. From the plots below we can see that all the polynomial roots (red) lie outside the unit. Also, note that the absolute value of the polynomial coefficients are all less than 1. Thus we can conclude that both model A and model B are causal and invertible.



Roots of MA for model A

Roots of SAR for model A

Roots of MA for model B

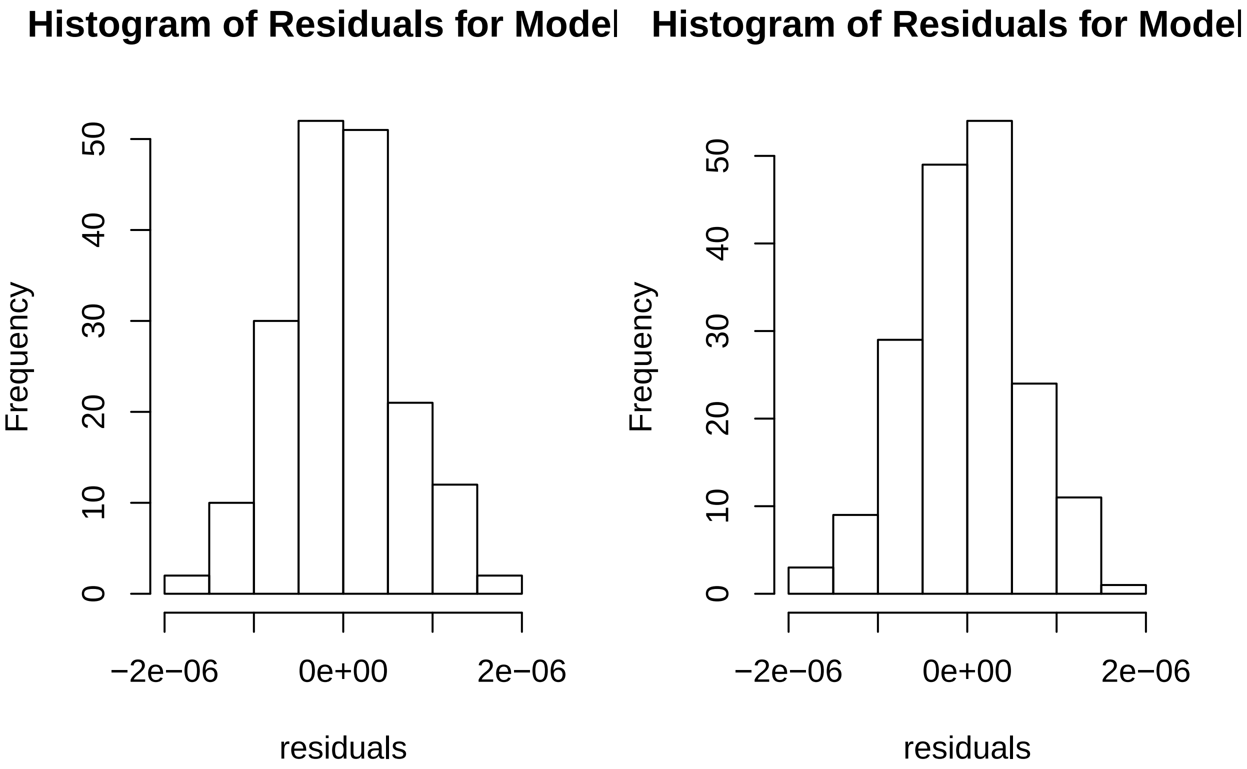
Roots of SAR for model B

5 Diagnostic Checks

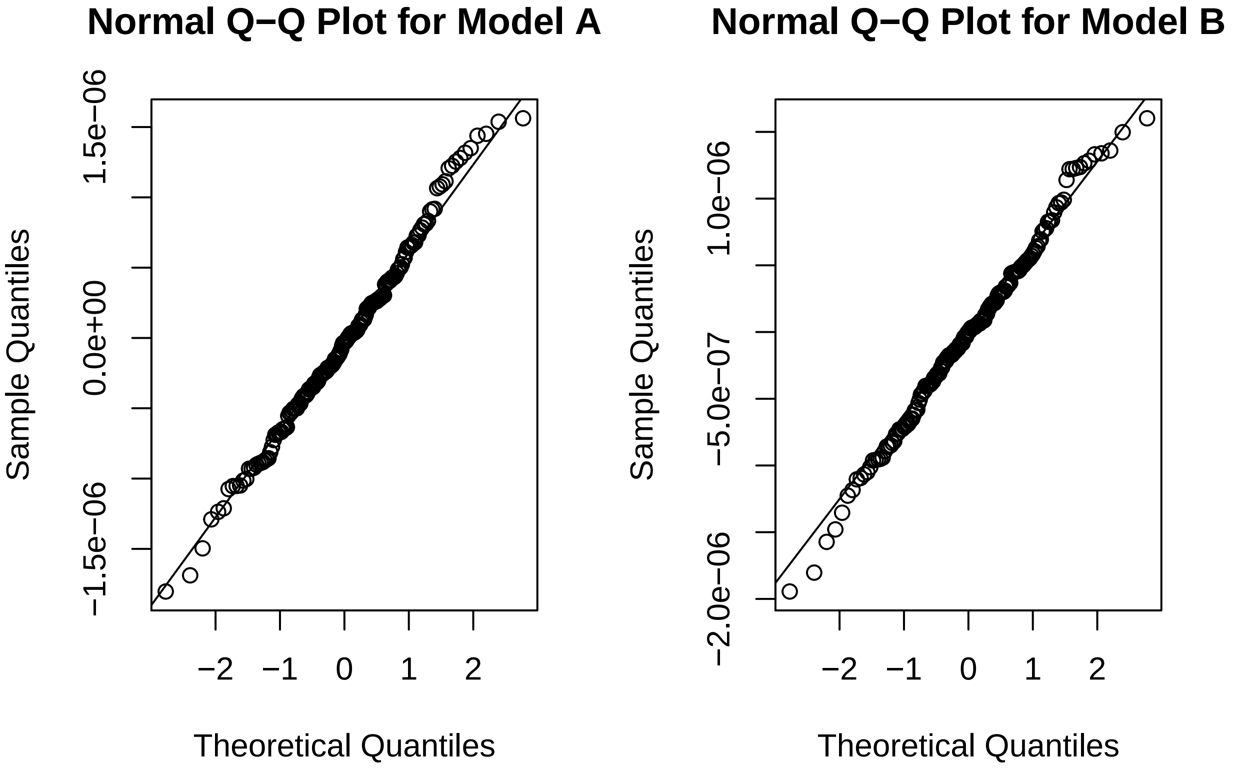
After identifing the two models with the smallest AICc and BIC scores, we ran diagnostic tests to make sure our models are reliable, valid and accurte. We tested for normality of the errors terms, independence via lack of serial correlation, and detecting heteroscedasticity.

5.1 Normality of Residuals

We decided to see if the error terms were normally distrubuted using 3 different checks. First we made a histogram of the residuals of both models and noticed that they evenly distrubed and symmetrical similar to a gaussian distribution.



Next we graphed a Normal Q-Q plot of the standardized residuals and noticed they roughly lie on a straight line. This indiciated to us that since the sample quantiles conicde with the theoretical quantlies, the residuals were evenly distributed.



In order to confirm our results from the our histograms and Q-Q plots we performed the Sapiro Wilk test. The null hypothesis for the Shapiro Wilk test is that the error terms are normally distributed. The p-values for both of our models are > 0.05, thus we would fail to reject the null hypothesis, thus confirming that both of our models have error terms that are normally distributed.

	Model A	Model B
P-values	0.588	0.7069

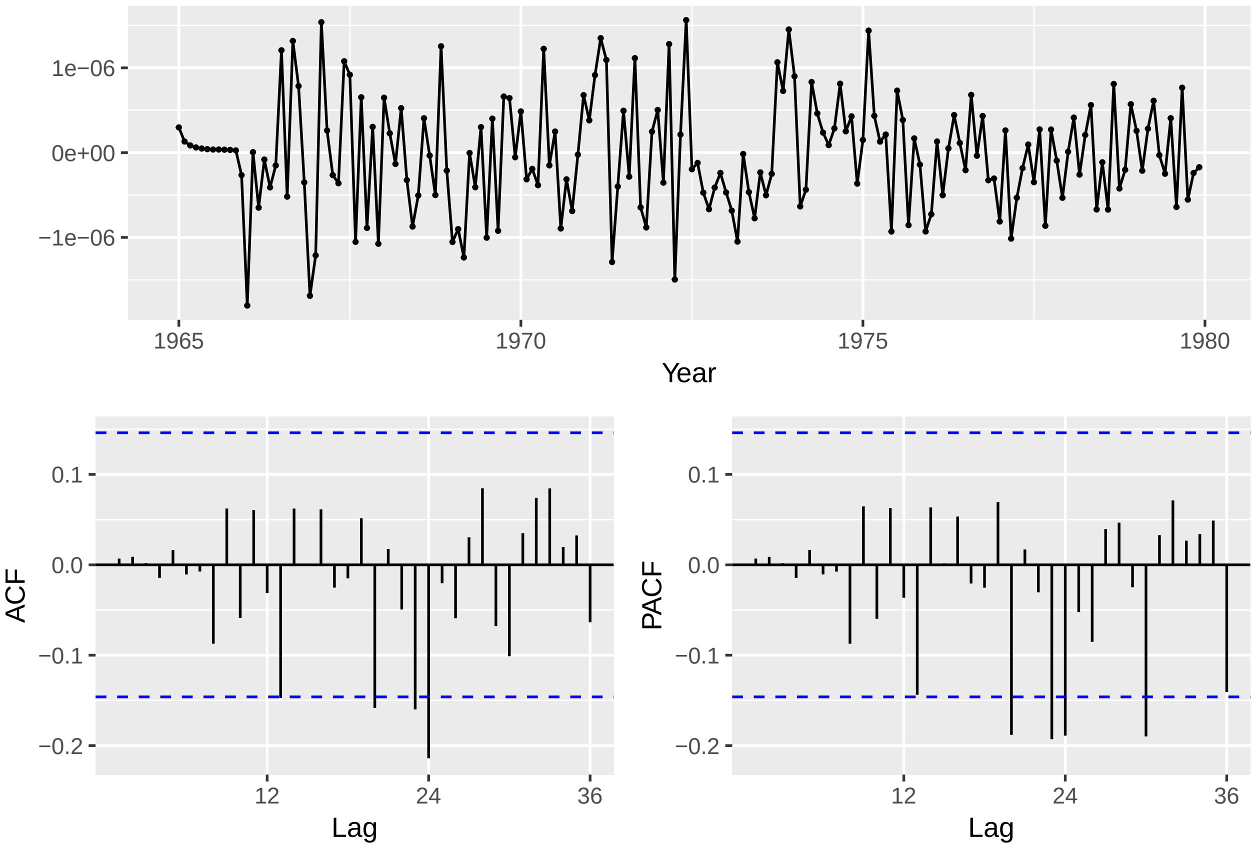
5.2 Serial Uncorrelation of Residuals (Independence checking)

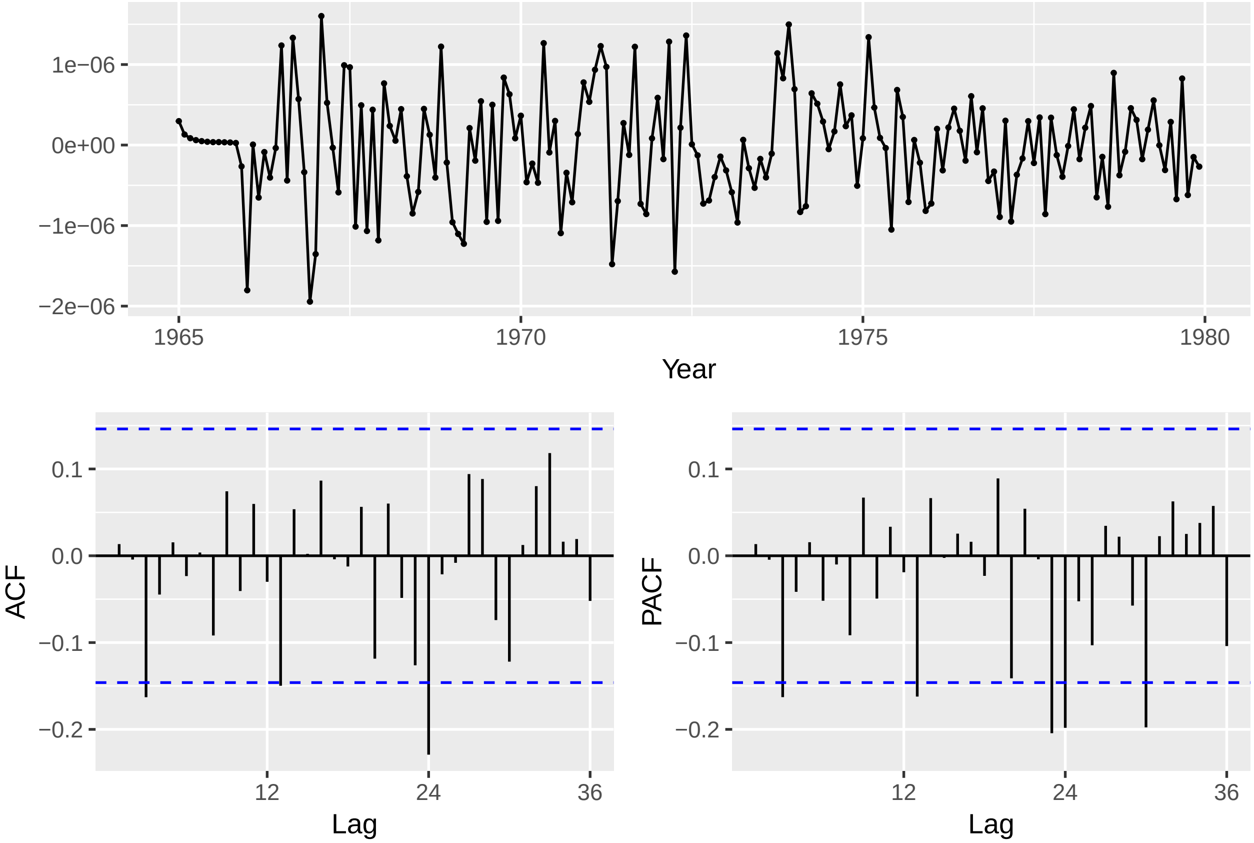
When analyzing time series data it is important that the error terms are independent of one another and there is no correlation between observations over different lags. To ensure we do not have this problem known as serial correlation we use the Ljung-Box and Box-Pierce test. The null hypothesis for these tests of is that serial correlation does not exist. Looking at our table x, we do not see any p-values less than 0.05, thus we would fail to detect serial correlation in either model and can conclude the error terms are indeed independent.

	Box-Pierce	Ljung-Box
Model A	0.9321	0.9177
Model B	0.6287	0.5927

5.3 Detection of Heteroskedasticity (Checking for nonconstant variance)

One of the diagnostic checks that a researcher must make regarding a time series data set is the error terms must not change over time. In order to make sure our model does not have heteroscedasticity we need to analyze the ACF and PACF plots of the squared residuals. If we see that most of our error terms are within the 95% White Noise limits then we can assume the error terms experience constant variance. Looking at figure x we notice that most of our error terms are indeed within the limits denoted by the blue dotted lines. Thus we can conclude that we do not detect heteroscedasticity in either one of our models.

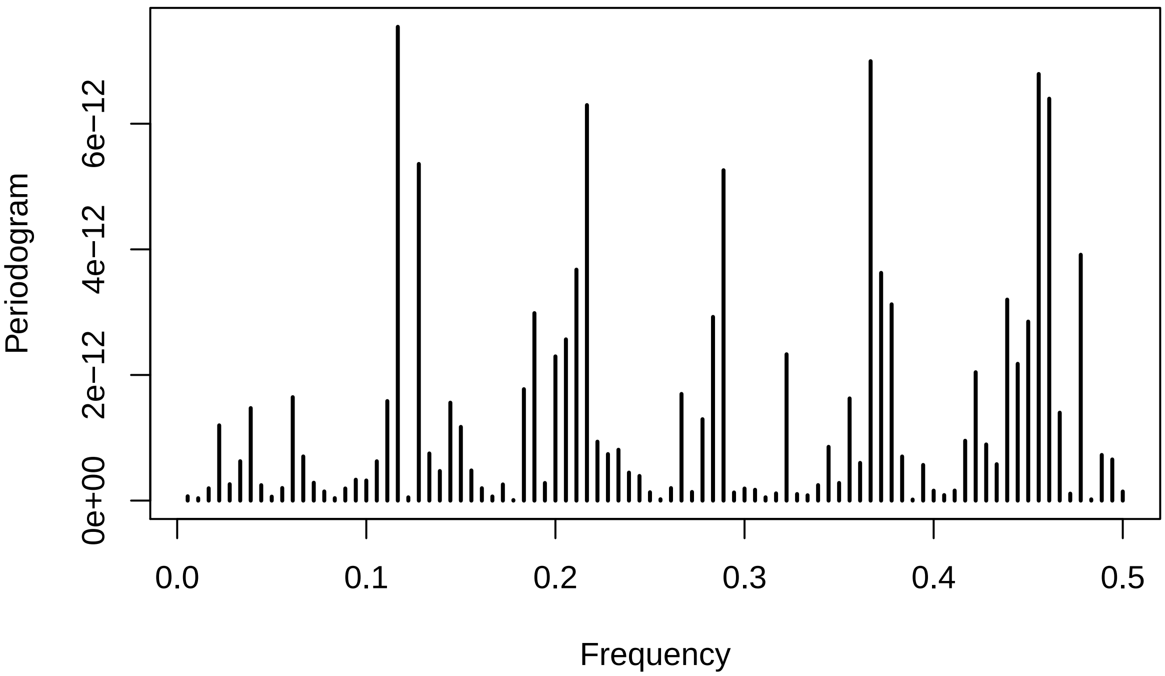

1e−06
0e+00
−1e−06
1965
1970
1975
1980
Year
ACF
0.1
0.0
−0.1
−0.2
12
24
36
Lag
PACF
0.1
0.0
−0.1
−0.2
12
24
36
Lag


1e−06
0e+00
−1e−06
−2e−06
1965
1970
1975
1980
Year
ACF
0.1
0.0
−0.1
−0.2
12
24
36
Lag
PACF
0.1
0.0
−0.1
−0.2
12
24
36
Lag

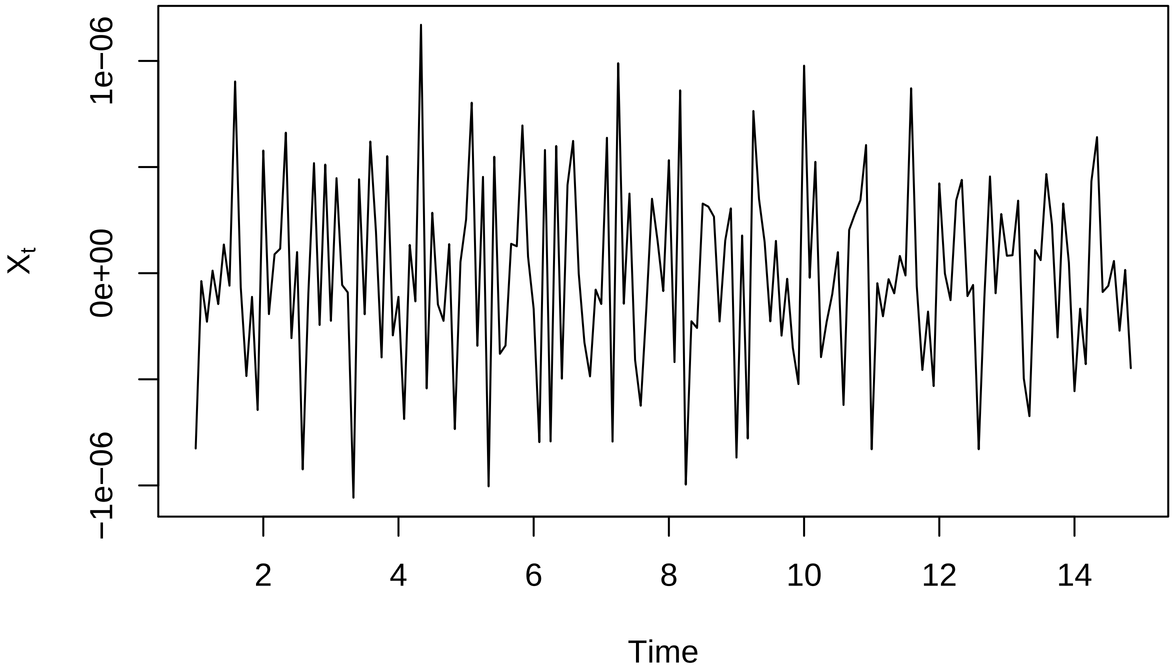
6 Spectral Analysis

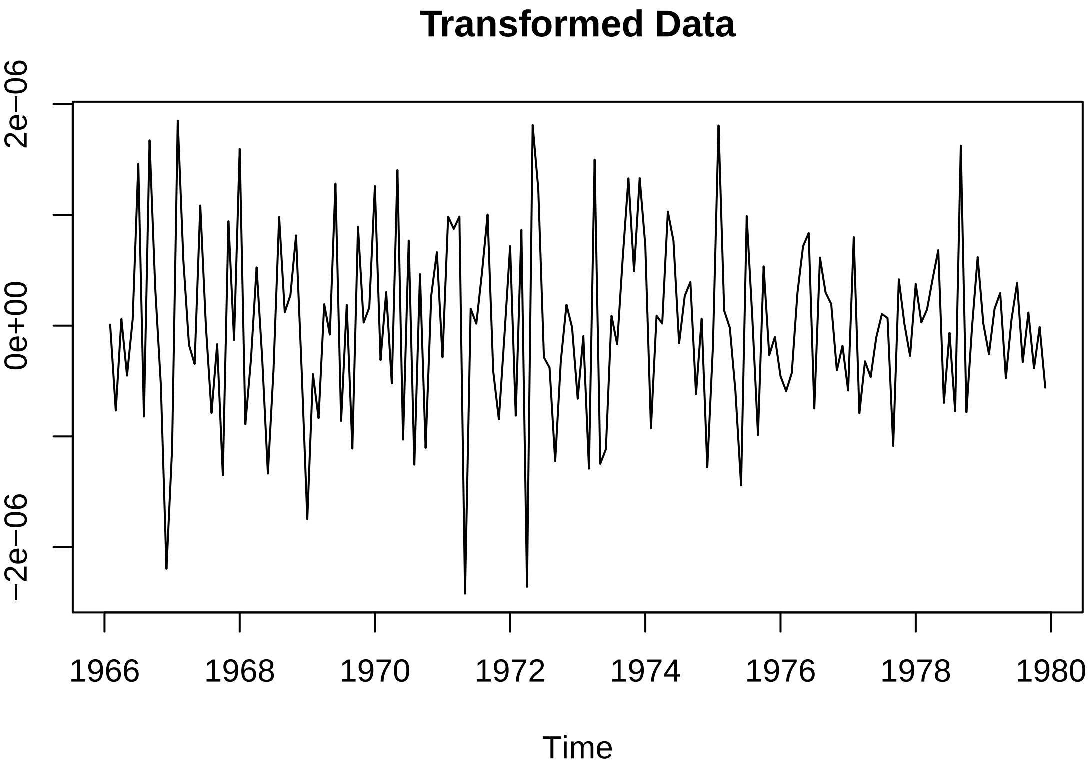
6.1 Peridogram

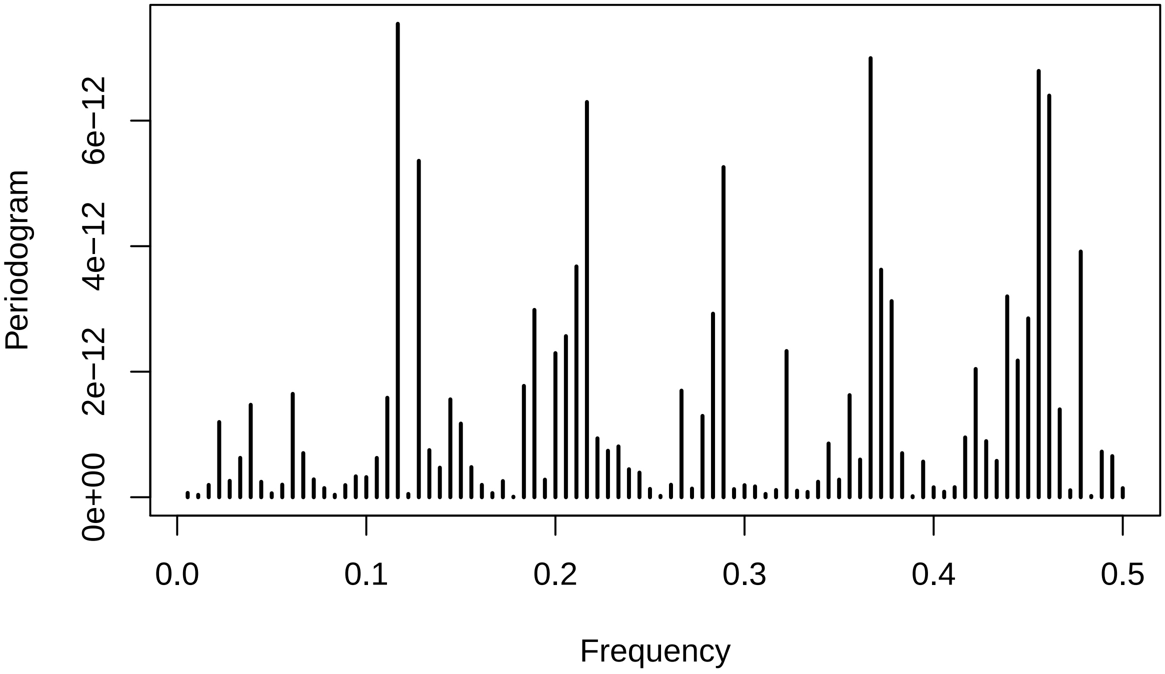
Peridogram of Stationary CO2 Data



Sinusoidal Expression of Data

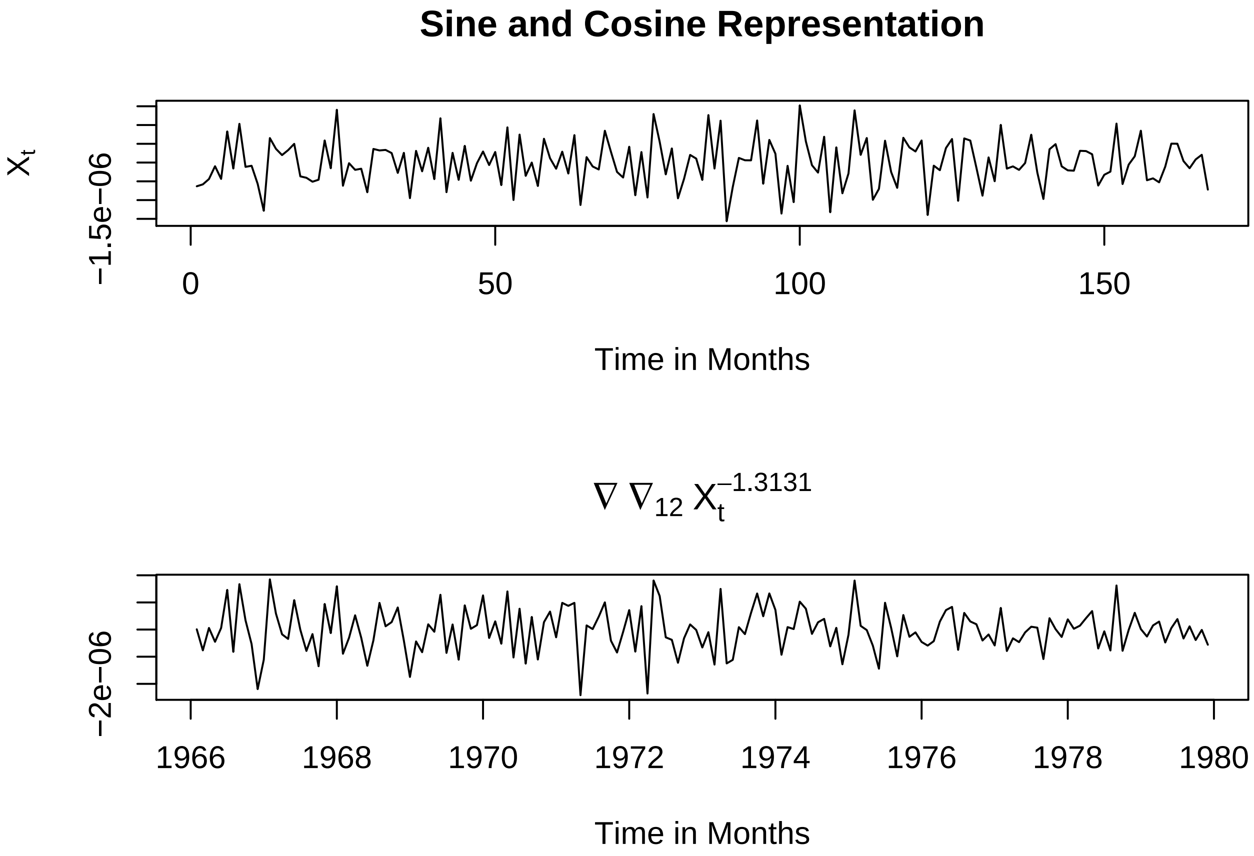






```
##
## Call:
## lm(formula = data.cl ~ ., data = m)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.627e-06 -3.494e-07  2.191e-08  4.025e-07  1.198e-06
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.293e-09  4.447e-08  -0.074 0.941083
## V1           1.707e-07  6.316e-08   2.703 0.007715 **
## V2          -2.686e-07  6.302e-08  -4.263 3.66e-05 ***
## V3           2.686e-07  6.361e-08   4.222 4.30e-05 ***
## V4          -1.118e-07  6.302e-08  -1.774 0.078275 .
## V5          -1.024e-07  6.390e-08  -1.603 0.111116
## V6          -2.746e-07  6.307e-08  -4.353 2.55e-05 ***
## V7           2.492e-07  6.357e-08   3.920 0.000137 ***
## V8           2.887e-08  6.342e-08   0.455 0.649665
## V9           4.072e-08  6.326e-08   0.644 0.520815
## V10         -2.751e-07  6.290e-08  -4.374 2.35e-05 ***
## V11          2.318e-07  6.334e-08   3.660 0.000355 ***
## V12         -1.531e-07  6.281e-08  -2.438 0.016021 *
## V13          1.795e-07  6.310e-08   2.844 0.005110 **
## V14         -1.802e-07  6.268e-08  -2.875 0.004662 **
## V15         -1.740e-07  6.328e-08  -2.750 0.006738 **
## V16         -1.790e-07  6.359e-08  -2.814 0.005580 **
## V17          2.050e-07  6.332e-08   3.237 0.001505 **
## V18          7.242e-08  6.284e-08   1.153 0.251014
## V19         -1.619e-07  6.363e-08  -2.545 0.012009 *
## V20          2.656e-08  6.301e-08   0.421 0.674036
## V21         -8.917e-08  6.344e-08  -1.405 0.162091
## V22         -1.695e-07  6.331e-08  -2.677 0.008298 **
## V23          1.920e-07  6.353e-08   3.022 0.002984 **
## V24          6.779e-08  6.309e-08   1.074 0.284453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.743e-07 on 142 degrees of freedom
## Multiple R-squared:  0.5726, Adjusted R-squared:  0.5003
## F-statistic: 7.925 on 24 and 142 DF,  p-value: < 2.2e-16
```



## 6.2 Fisher Test

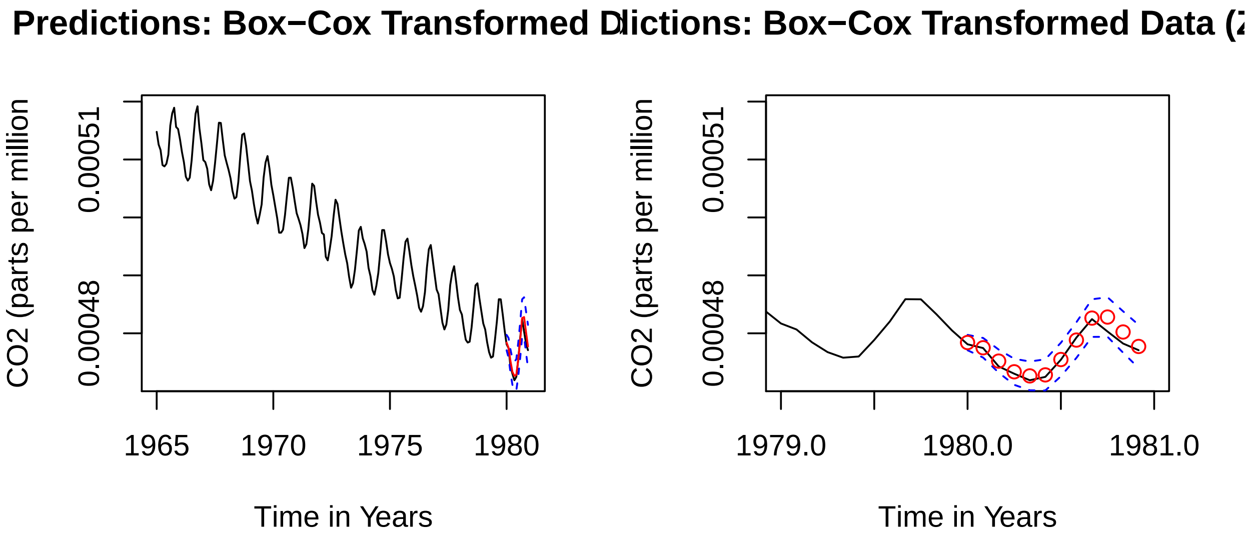
6.3 Kolmogorov-Smirnov Test

7 Forecasting

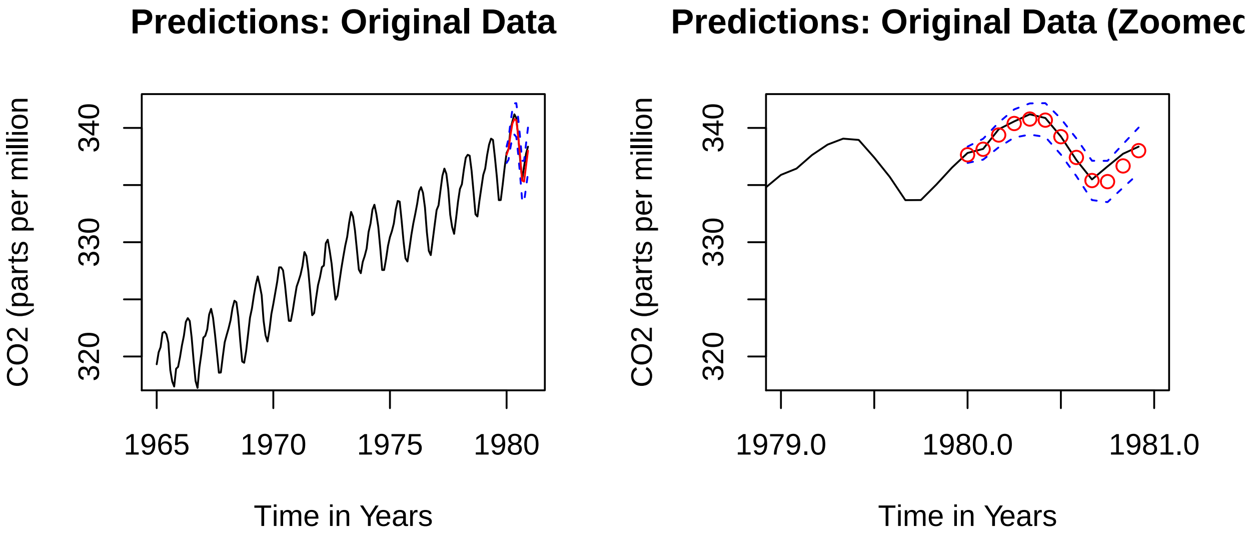
Now that we have a final model that accurately estimates our data we can begin forecasting. The two main objectives in time series analysis is to model and then harness that model's predictive power to get an accurate idea of what the future would entail. Specific to our case, the predictions we make should provide insight on how the concentration of CO_2 in our atmosphere will change in the coming years if it continues the pattern it's been following for the last 30+ years.

Because we have a complete model with all parameters estimated, we can predict the next twelve months and use the data we set aside near the beginning of the analysis to compare to our predictions.

The first set of plots depict our predictions on the Box-Cox transformed data: On the left we see the transformed data in black and overlayed atop the last years line is our prediction in red along with both the upper and lower confidence interval limit in blue dashed lines. On the right we can see a zoomed-in version of the same plot to more clearly identify how accurate we were with our predictions.



Now we show to similar plot with the same schema as before but now we are using our original, untransformed data.



These values can be interpreted as real-world predictions. For instance we predicted that on January of 1980 the atmospheric CO_2 ppm would be 337.6487023 when the true, observed ppm is 337.81. If we take the MSE of our predictions we get 0.3007988 which is very small relative to our data.

Our model clearly captures the true trend and seasonality of the data. All of our predicted points follow the true points very closely and our 95% confidence interval easily encompasses all of the observed points. This is evidence that our model is successful in estimating the process of CO_2 ppm in the atmosphere.

8 Conclusion

In order to better understand the progression of carbon dioxide emissions into our atmosphere we made it our goal to analyze how the levels of CO_2 have changed over the last few decades and how the levels behave on a seasonal basis. To do this, we constructed a time series model that both explains the behavior and can be used to forecast its behavior further into the future.

Our analysis showed an undeniable upward trend in CO_2 levels. From 1965 until 1980, every year's measurements were greater than the ones before it. We also found that the concentration is seasonal with a distinct spike and trough in the same spot every year. This seasonality component could be explained by shifting tropical winds as well as other factors.

After a tedious investigation using an array of methods, we reached our model:

Using the results of our forecasting, we conclude that our model is more than satisfactory. Our 95% confidence interval encapsulates all the true data and our predicted points are very close to what was actually observed.