PSTAT 174 Final Project: Mauna Loa CO2 Time Series

Elias Parzen

11/19/2021

Executive Summary

A time series of CO2 measurements from a Hawaiian volcano was de-trended and de-seasonalized through differencing. Several candidate models were identified based on the ACF and PACF of the now-stationary data and compared via AICC. As a result, the time series was ultimately modeled as a SARIMA process of the form $(0,1,3) \times (0,1,1)$. This model passed all diagnostic checks and was used for forecasting. The predicted values was very close to the actual observations from the test data, indicating that the model accurately simulated the time series.

Full Report

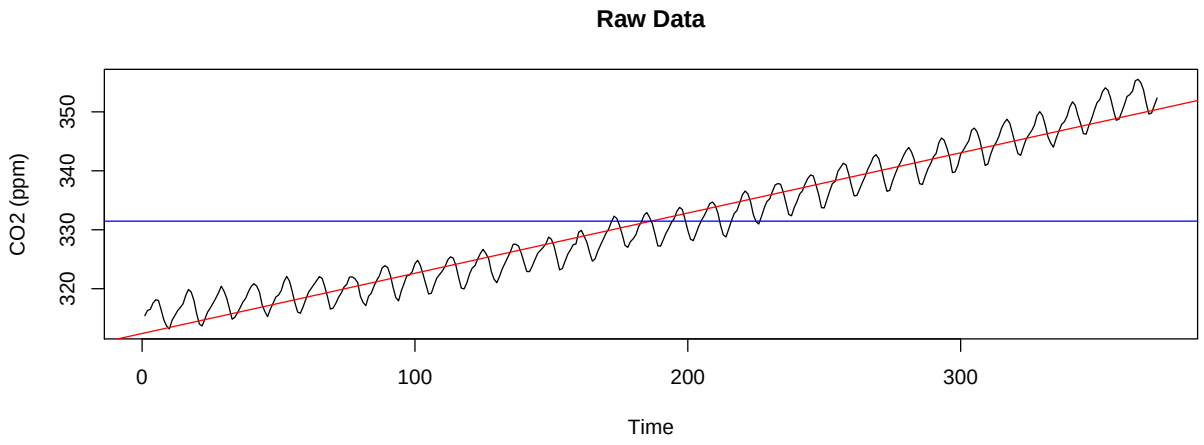
Introduction

This time series comes to us from the Oak Ridge National Laboratory climatology database, and is one of the data sets in the Time Series Data Library. It contains monthly measurements of CO2 in parts per million (ppm) from air samples above Mauna Loa, one of the largest active volcanoes on Earth. The data was collected by researchers from the Scripps Institute of Oceanography from January 1959 to December 1990, with a total of 384 observations. Missing values were filled by linear interpolation. The data was imported into R Studio, which was used to perform the analysis, model building, and forecasting.

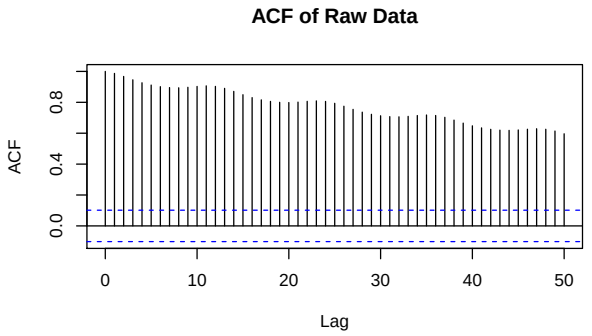
As someone who grew up in La Jolla, the home of the Scripps Institute, and who has had the pleasure of seeing Mauna Loa firsthand, this time series immediately caught my eye. The raw data itself intrigued me because the plot was indicative of a SARIMA process, one of the more complex models we covered in PSTAT 174, and something I wanted to improve my understanding of. My goal was to try to build a model of the data that would be able to forecast future measurements. I was able to de-trend and de-seasonalize the data through differencing at lags 1 and 12. At that point, I examined the ACF and PACF plots and identified a family of candidate models based on the significant lags. After narrowing the field down to three moving average SARIMA models by comparing AICCs, I put each candidate through diagnostic checks. Since they all passed, I chose the model with the lowest AICC for forecasting. My final model produced predictions quite close to the test data I had set aside before beginning my analysis, indicating that the $(0,1,3) \times (0,1,1)$ model accurately simulated the underlying process for this time series.

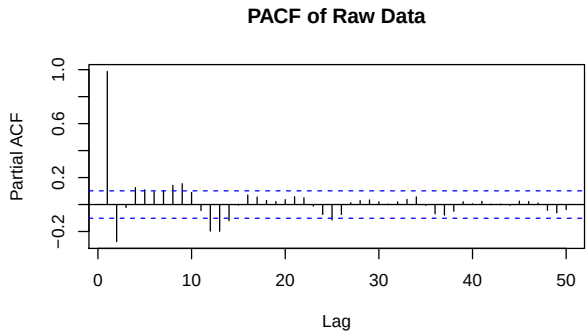
Time Series Plot and Analysis

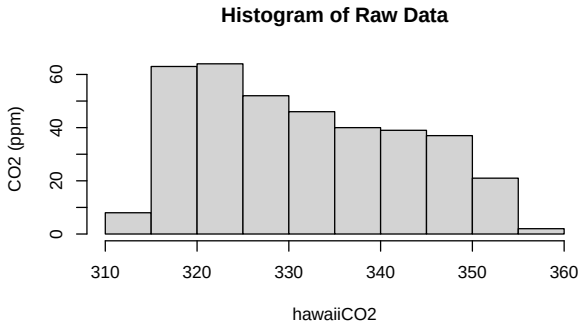
After splitting off the last 12 observations for testing purposes, plotting the raw data reveals several key features about the time series. The process is clearly non-stationary. It has an obvious, linear trend, as can be seen by comparing the blue line, representing the mean of the data, to the red line, which is the line of best fit. Furthermore, there is a clear seasonal component, likely with period 12. The trend at the beginning of the time series is somewhat less strong, and variance appears to remain fairly constant across the decades that the data was collected. There are no sudden changes in behavior that would complicate the modeling process



The histogram of the raw data is asymmetric, having a right skew, and the ACF plot, with its slow decrease and periodic bumps also points toward a trend and seasonality.



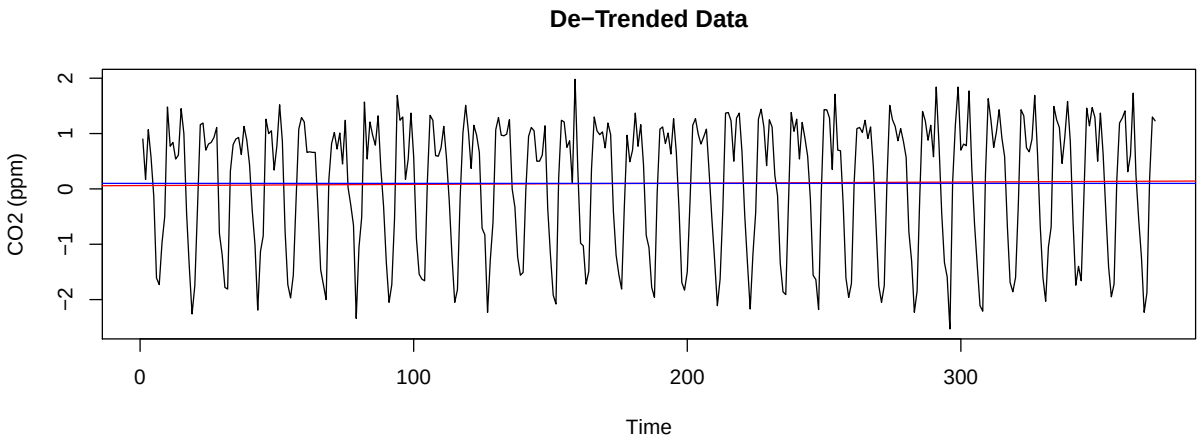




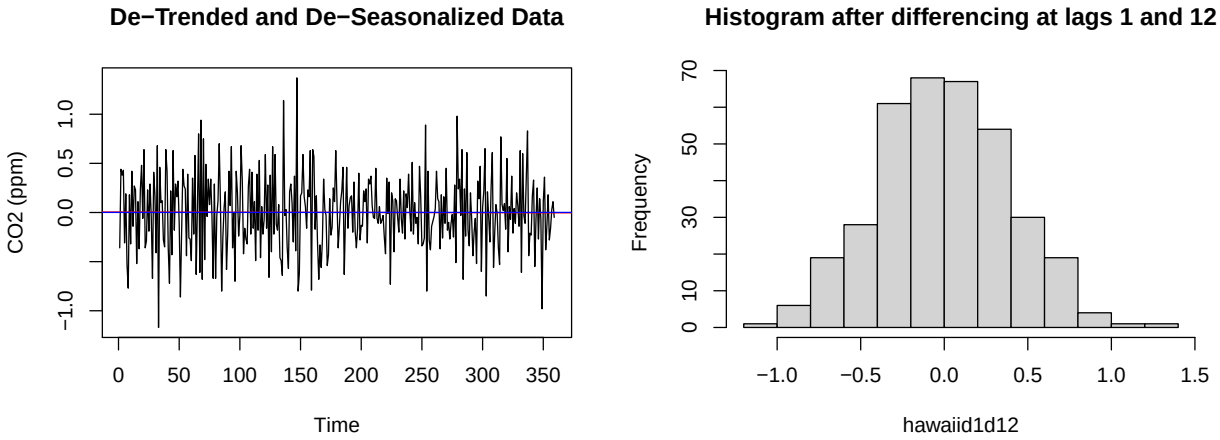
Based on these observations, it's apparent that in order to be made stationary, the time series will have to be differenced to remove trend and seasonality. Because there is no need to stabilize the variance, it will not be necessary to otherwise transform the data. If the variance was non-constant, the Box-Cox function could be used to find a suitable transformation.

Transformations and Stationarity

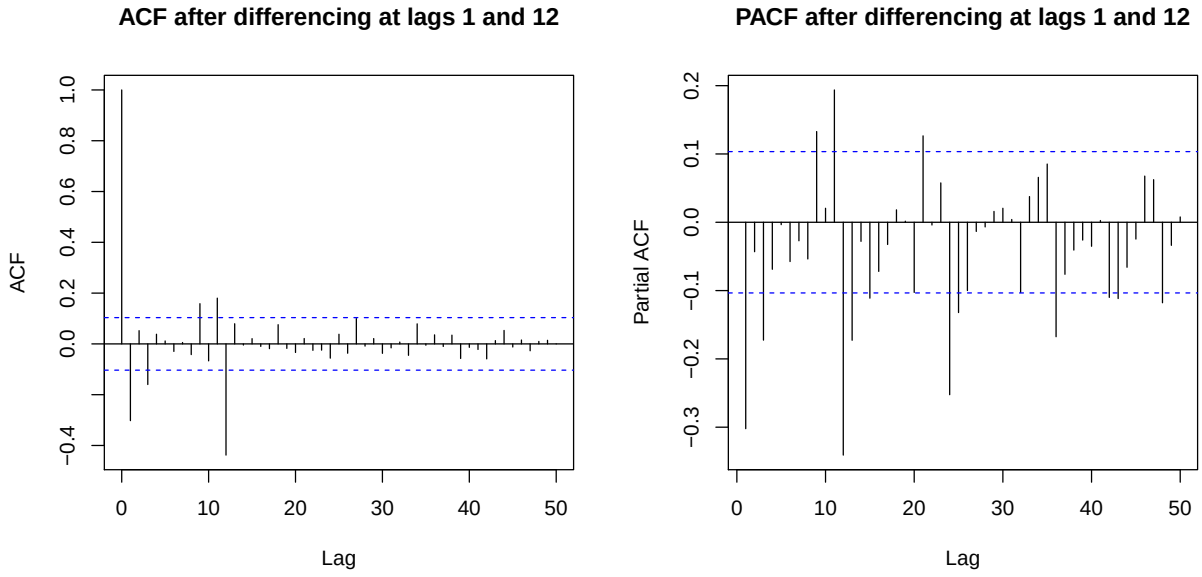
To remove the trend from the data, the time series was differenced once at lag 1. This form of differencing is effective at nullifying a linear trend, which is what the plot of the raw data suggested. The differencing resulted in a decrease in variance from 126.8622 to 1.3823. The line of best fit and the mean both became essentially 0, which shows that the trend was successfully eliminated.



To remove the seasonality, the data was differenced at the suspected period, 12. Differencing once at lag 12 decreased the variance again to 0.1535. The plot of the data looks to be of a stationary process, and the histogram shows that the differenced observations are approximately normally distributed.



The ACF plot no longer displays any sign of trend or seasonality, which had previously manifested as a steadily decreasing ACF with regular bumps.



With no remaining trend or seasonality, the data is now stationary and ready for modeling.

Candidate Models

When considering what types of models are candidates for this time series, the fact that the original data contained both a trend and a seasonal component tells us to look to the SARIMA family of models of the form $(p, d, q) \times (P, D, Q)_s$. Because differencing once at lags 1 and 12 caused the data to become stationary, we know that $d = 1, D = 1$ and $s = 12$. The ACF and PACF plots above hint at several possibilities for the remaining parameters. The most obvious feature of the ACF plot is that the autocorrelations cut off completely at lag 12. This tells us that it's safe to assume $Q = 1$. Some values at higher lags near the boundary of the confidence interval, but because we know the interval is calculated quite conservatively, we can confidently ignore them. The PACF plot shows exponential decay at lags that are a multiple of 12, indicating that it's likely $P = 0$.

In terms of the within-year parameters, we see ACF values outside the confidence intervals at lags 1, 3, 9, and 11. The value at lag 9 is close to the confidence interval, and the significant ACF at 11 is likely a result of $Q = 1$, so we should consider $q = 0, 1, 3$. In the PACF plot, we see that lags 1 and 3 have significant values beyond what we would expect from the pattern for $P = 0$. At lag 1, the PACF is always equal to the ACF, so the significant value there is likely just a consequence of having a significant ACF at that time index. Based on this, we should consider $p = 0, 3$ for our model.

To summarize, our candidate models are of the form $(p, 1, q) \times (0, 1, 1)$ where p could equal 0 or 3, and q could equal 0, 1, or 3.

Preliminary Model Selection

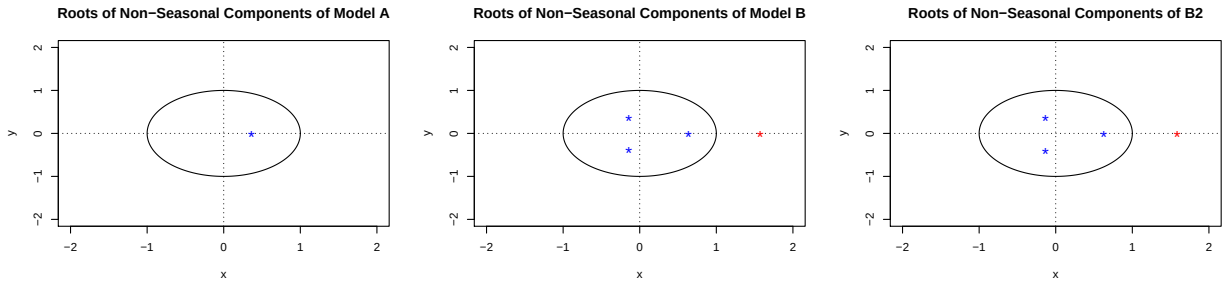
The above mentioned models were compared based on the Akaike information criterion (AICC) which calculates a score for each model balancing likelihood and complexity, with lower scores being better. The two lowest scoring models were Model A: $(0, 1, 1) \times (0, 1, 1)$ and Model B: $(0, 1, 3) \times (0, 1, 1)$ with scores of 136.45 and 135.8 respectively. These models also have a relatively low number of parameters, so the principle of parsimony suggests we should prefer them over more complex models. Based on these considerations, these two preliminary models were selected for estimation and diagnostic testing.

Model Diagnostics and Final Selection

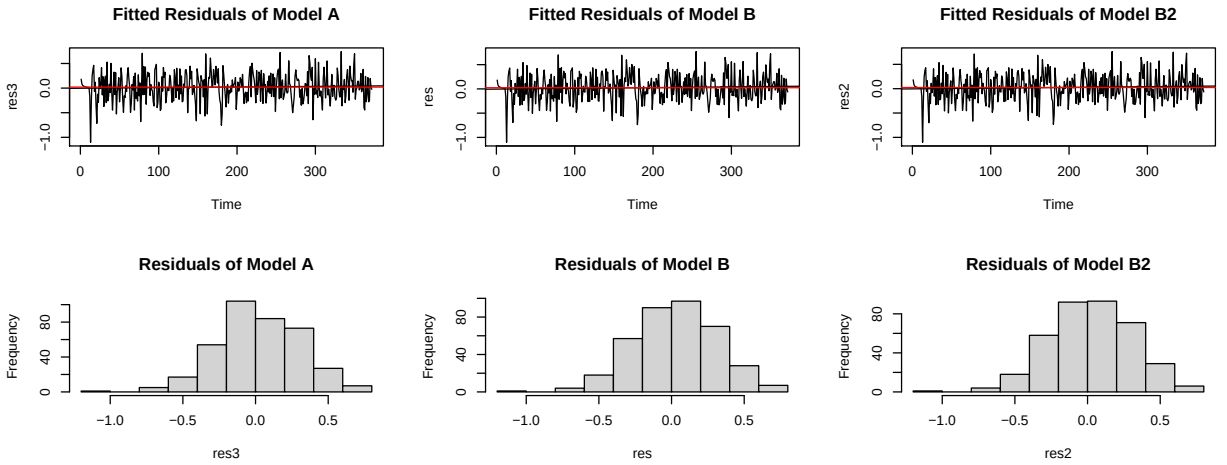
When the coefficients of Model B were estimated from the data, the coefficient of the second moving average term was -0.0296 with a standard error of 0.0562. Given that a 95% confidence interval would comfortably

include 0, I decided to test a new $(0, 1, 3) \times (0, 1, 1)$ model, Model B2, with the second coefficient fixed at zero. This lowered the AICC to 133.93, a significant improvement.

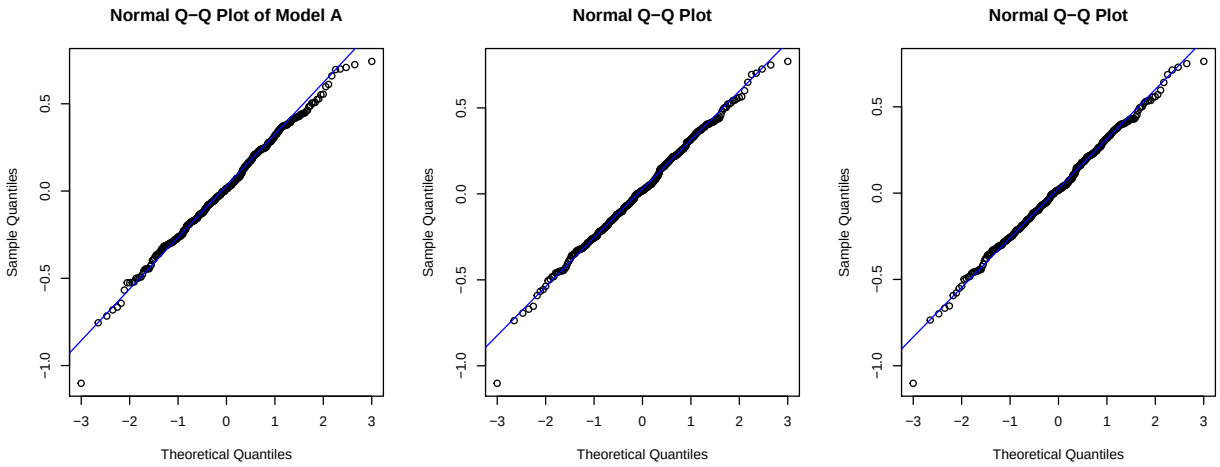
Each of the models was first checked for invertibility. Since they are all moving average models, and this type of model is always stationary, there was no need to test for that attribute. The roots (in red, blue are their inverses) of each model's characteristic polynomials were outside the unit circle, confirming invertibility.



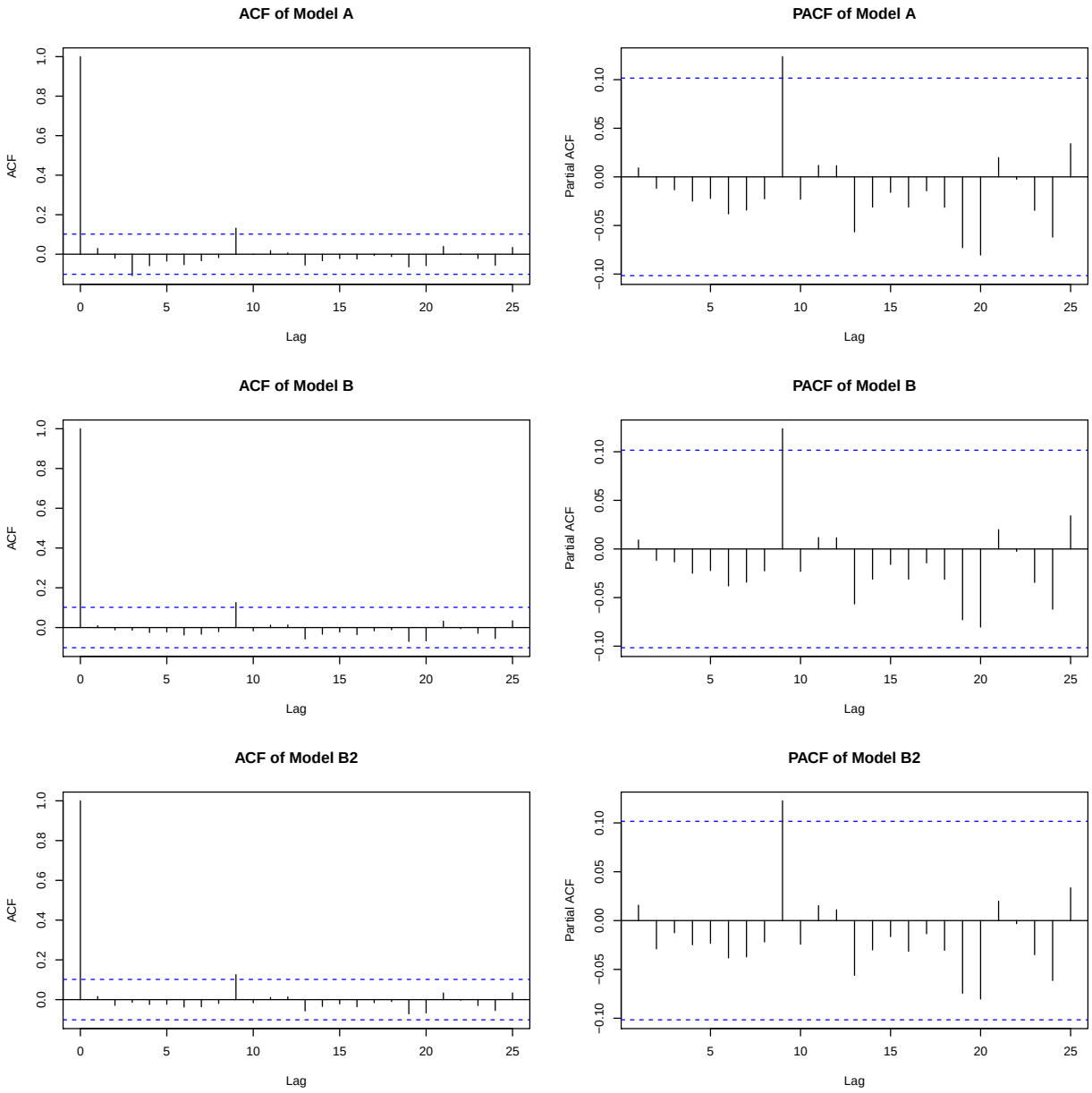
If a model fits the data well, its residuals, the difference between the models predictions and the actual data, should be white noise, uncorrelated and normally distributed with mean 0. To check this assumption, each model was subjected to a series of diagnostic tests. The histogram of each model's residuals visually seems like white noise with a roughly gaussian in distribution based on the histograms. The means were computed, and each was approximately 0.2



QQ-Plots of the residuals are close to the line representing normality for each model. Furthermore, Model A, Model B, and Model B2 each passed the Shapiro-Wilk test for normality with p-values of 0.291, 0.4299, and 0.3774 respectively. Based on these results, the residuals of each model seem to be normally distributed.



Next the ACF and PACF of the residuals were plotted. Since the residuals of a proper model are uncorrelated, there should be no significant correlations at any lag. For these three models, we see slight significance at lag 9 in the PACF plot for each model, but results of the remainder of the diagnostic tests, this was considered to be acceptable.



To further confirm the independence of the residuals, the Box-Pierce, Ljung-Box, and McLeod-Li tests were applied to test for linear and non-linear dependence. The maximum lag these tests can handle is approximately the square root of the number of observations, in this case, $\sqrt{n} = \sqrt{372} \approx 19$. The McLeod-Li test always uses 0 degrees of freedom, whereas the other tests use degrees of freedom equal to the number of parameters estimated from the data, so 2 for Model A, 4 for Model B, and 3 for model B2.

Each model passed every test by a wide margin, with the lowest p-value being 0.341 for the Model A Ljung-Box test.

Finally, to check the white noise assumption for the residuals even more thoroughly, the Yule-Walker equations were used on each model's residuals in an attempt to fit an autoregressive model to them. However, the algorithm returned a recommended model of AR(0), a.k.a white noise, each time.

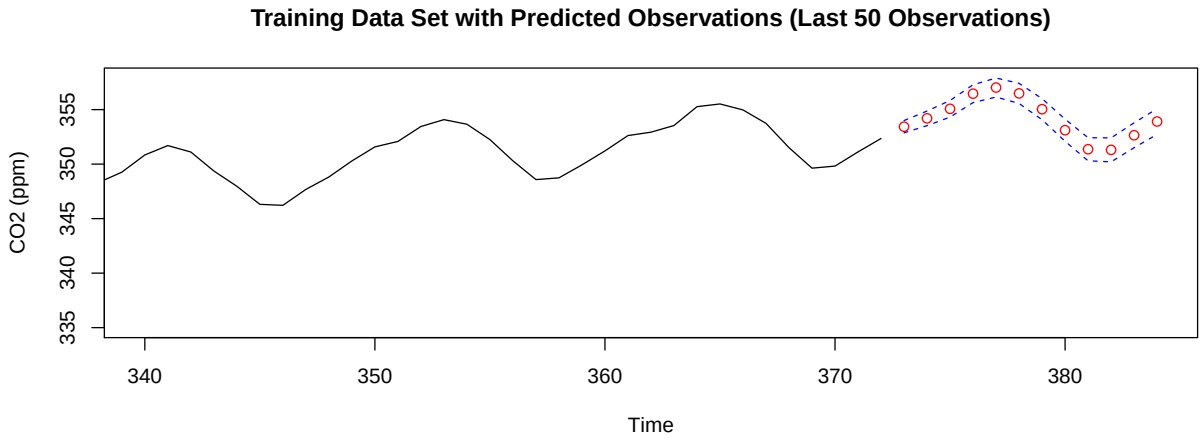
This result, in combination with the outcomes of the previous test, demonstrates that it was correct to ignore the slightly significant correlation at lag 9.

Each model passed all diagnostic tests, but only one would be used for forecasting. Since model B2, the $(0,1,3) \times (0,1,1)$ model with equation $(1-B)(1-B^{12})X_t = (1-0.3598B-0.1073B^3)(1-0.8559B^{12})Z_t$, had the lowest AICC (133.93 versus 136.45 for Model A and 135.8 for Model B), it was chosen as the final model over the other two candidates. B2 is also perfectly compatible with the ACF and PACF plots of the stationary data. It's worth noting that it's not uncommon for multiple different models to pass diagnostic testing and provide reasonable forecasts, so the fact that B2 was chosen does not mean that Models A and B were unworkable.

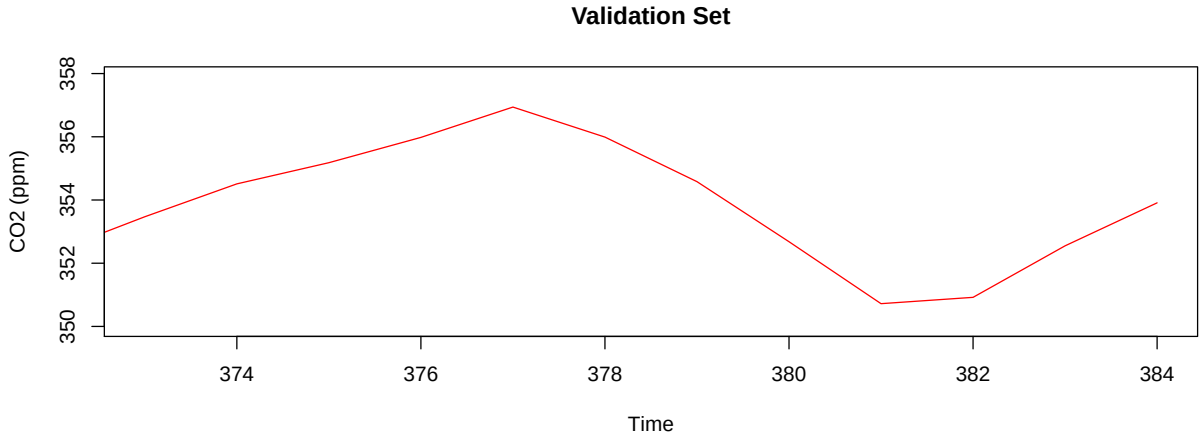
Forecasting

The purpose of analyzing this time series was to develop a model capable of predicting future values. To that end, the last 12 observations of this data set were split off at the beginning of the project and not used for training the models. These observations can be used as a validation to gauge the accuracy of the model.

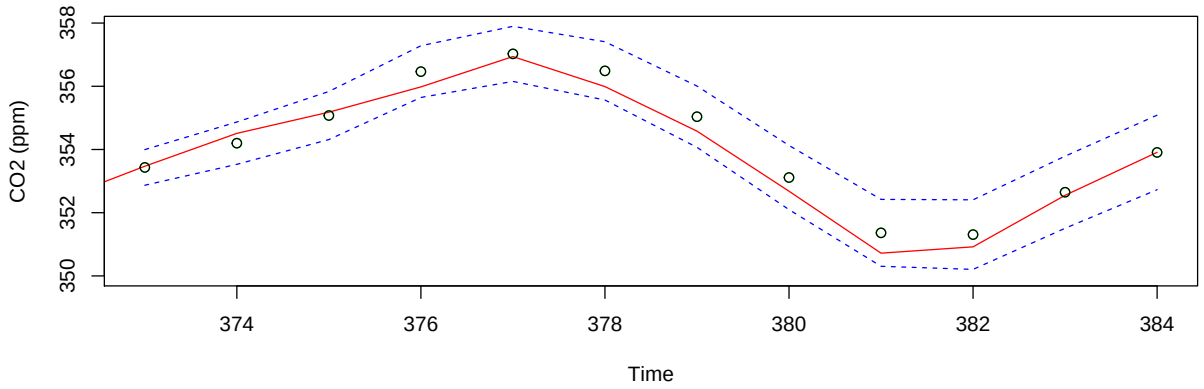
The predicted values for the last 12 observations are shown here as red circles. The dashed, blue lines represent the confidence intervals of the predictions



The true values of the final 12 observations are plotted here as the red line.



When we overlay the model's prediction upon the test data, we see that the prediction interval completely contains the true observations, and that the predicted values, shown here as black circles, fall quite close to the red line.



This result strongly suggests that the $(0,1,3)\times(0,1,1)$ Model B2 is capable of providing meaningful predictions of observations of CO2 levels on Mauna Loa. However, it should be noted that the data set only contains observations up to the year 1990, so although this model might work for predicting the CO2 levels of 1991, it does not necessarily follow that the model could estimate the CO2 levels of 2021.

Conclusion

The goal of this project was to identify a model capable of forecasting CO2 measurements on Mauna Loa based on the Scripps Institute data. The results outlined above are very encouraging, as the predicted measurements from the model, estimated to be $(1-B)(1-B^{12})X_t = (1-0.3598B-0.1073B^3)(1-0.8559B^{12})Z_t$, are very close to the actual observations from the test data. This indicates that the data analysis, model building, and diagnostic checking resulted in identifying a SARIMA process that closely matches the original time series data.

This satisfying result was only made possible through the instruction and support of my professor, Dr. Raisa Feldman, and my TA, Jasmine Li. They have my immense gratitude for helping me take my first steps into the world of time series.

References

Rob Hyndman and Yangzhuoran Yang (2018). tsdl: Time Series Data Library. v0.1.0.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
# Load libraries
library(tsdl)
library(MASS)
library(qpcR)
library(forecast)

# Set working directory
setwd("~/PSTAT 174/Final Project")

# Retrieve data
ag <- subset(tsdl,12,"Meteorology")
projData <- ag[[2]]

# Train/test split
hawaiiCO2 <- projData[c(1:372)]
hawaiiCO2Test <- projData[c(373:384)]
```

```r
allData <- projData[c(1:384)]

# Plot.roots function
plot.roots <- function(ar.roots=NULL, ma.roots=NULL, size=2, angles=FALSE, special=NULL, sqecial=NULL,my
{xylims <- c(-size,size)
      omegas <- seq(0,2*pi,pi/500)
      temp <- exp(complex(real=rep(0,length(omegas)),imag=omegas))
      plot(Re(temp),Im(temp),typ="l",xlab="x",ylab="y",xlim=xylims,ylim=xylims,main=main)
      abline(v=0,lty="dotted")
      abline(h=0,lty="dotted")
      if(!is.null(ar.roots))
        {
          points(Re(1/ar.roots),Im(1/ar.roots),col=first.col,pch=my.pch)
          points(Re(ar.roots),Im(ar.roots),col=second.col,pch=my.pch)
        }
      if(!is.null(ma.roots))
        {
          points(Re(1/ma.roots),Im(1/ma.roots),pch="*",cex=1.5,col=first.col)
          points(Re(ma.roots),Im(ma.roots),pch="*",cex=1.5,col=second.col)
        }
      if(angles)
        {
          if(!is.null(ar.roots))
            {
              abline(a=0,b=Im(ar.roots[1])/Re(ar.roots[1]),lty="dotted")
              abline(a=0,b=Im(ar.roots[2])/Re(ar.roots[2]),lty="dotted")
            }
          if(!is.null(ma.roots))
            {
              sapply(1:length(ma.roots), function(j) abline(a=0,b=Im(ma.roots[j])/Re(ma.roots[j]),lty="
            }
        }
      if(!is.null(special))
        {
          lines(Re(special),Im(special),lwd=2)
        }
      if(!is.null(sqecial))
        {
          lines(Re(sqecial),Im(sqecial),lwd=2)
        }
        }

# Plot raw data with mean and line of best fit
plot.ts(hawaiiCO2, main = 'Raw Data', ylab = 'CO2 (ppm)')
fit <- lm(hawaiiCO2~as.numeric(1:length(hawaiiCO2))); abline(fit, col = "RED")
abline(h=mean(hawaiiCO2), col = "BLUE")

par(mfrow = c(2,2))
# Plot ACF/PACF
acf(hawaiiCO2,lag.max = 50)
pacf(hawaiiCO2, lag.max = 50)
```

```r
# Plot histogram
hist(hawaiiCO2, main = "Histogram of Raw Data", ylab = 'CO2 (ppm)')

# Difference at lag 1 to remove trend
hawaiid1 <- diff(hawaiiCO2,1)
plot.ts(hawaiid1, main = "De-Trended Data", ylab = "CO2 (ppm)")
fit <- lm(hawaiid1~as.numeric(1:length(hawaiid1))); abline(fit, col = "RED")
abline(h=mean(hawaiid1), col = "BLUE")

par(mfrow = c(1,2))

# Difference at lag 12 to remove seasonality
hawaiid1d12 <- diff(hawaiid1,12)
plot.ts(hawaiid1d12, main = "De-Trended and De-Seasonalized Data", ylab = "CO2 (ppm)")
fit <- lm(hawaiid1d12~as.numeric(1:length(hawaiid1d12))); abline(fit, col = "RED")
abline(h=mean(hawaiid1d12), col = "BLUE")

# New histogram is symmetric/looks normally distributed
hist(hawaiid1d12, main = "Histogram after differencing at lags 1 and 12")

# New ACF/PACF no longer show signs of trend/seasonality
par(mfrow=c(1,2))
acf(hawaiid1d12, lag.max = 50, main = "ACF after differencing at lags 1 and 12")
pacf(hawaiid1d12, lag.max = 50, main = "PACF after differencing at lags 1 and 12")

# Check AICC for each candidate model
for(i in c(0,3)){
  for(j in c(0,1,3)){
    print(i)
    print(j)
    print(AICc(arima(hawaiiCO2, order = c(i,1,j), seasonal = list(order = c(0,1,1), period = 12), metho
  }
}

# Estimate models and AICCs
model <- arima(hawaiiCO2, order = c(0,1,3), seasonal = list(order = c(0,1,1), period = 12), method = "M
model2 <- arima(hawaiiCO2, order = c(0,1,3), seasonal = list(order = c(0,1,1), period = 12), fixed = c(
model3 <- arima(hawaiiCO2, order = c(0,1,1), seasonal = list(order = c(0,1,1), period = 12), method = "

AICc(model3)
AICc(model)
AICc(model2)

# Plot the roots of the candidate models
par(mfrow = c(1,3))
plot.roots(NULL,polyroot(c(1,-0.3642)), main="Roots of Non-Seasonal Components of Model A")
plot.roots(NULL,polyroot(c(1,-0.3532,  -0.0206,  -0.1010)), main="Roots of Non-Seasonal Components of M
plot.roots(NULL,polyroot(c(1,0,-0.3598)), main="Roots of Non-Seasonal Components of B2")

# Get residuals
res <- residuals(model)
res2 <- residuals(model2)
res3 <- residuals(model3)
```

```r
par(mfrow = c(2,3))

# Plot residuals with mean and line of best fit
ts.plot(res3,main = "Fitted Residuals of Model A")
t = 1:length(res3)
fit.res3 = lm(res3~t)
abline(fit.res3)
abline(h = mean(res3), col = "red")

ts.plot(res,main = "Fitted Residuals of Model B")
t = 1:length(res)
fit.res = lm(res~t)
abline(fit.res)
abline(h = mean(res), col = "red")

ts.plot(res2,main = "Fitted Residuals of Model B2")
t = 1:length(res2)
fit.res2 = lm(res2~t)
abline(fit.res2)
abline(h = mean(res2), col = "red")

# Plot residual histograms
hist(res3, main = "Residuals of Model A")
hist(res, main = "Residuals of Model B")
hist(res2, main = "Residuals of Model B2")

# QQ Plots
par(mfrow = c(1,3))

qqnorm(res3, main = "Normal Q-Q Plot of Model A")
qqline(res3,col ="blue")

qqnorm(res)
qqline(res,col ="blue", main = "Normal Q-Q Plot of Model B")

qqnorm(res2)
qqline(res2,col ="blue", main = "Normal Q-Q Plot of Model B2")

# Shapiro-Wilk tests
shapiro.test(res3)

shapiro.test(res)

shapiro.test(res2)

# Plot ACF and PACF
par(mfrow= c(3,2))
acf(res3,main = "ACF of Model A")
pacf(res,main = "PACF of Model A")

acf(res,main = "ACF of Model B")
pacf(res,main = "PACF of Model B")
```

```r
acf(res2,main = "ACF of Model B2")
pacf(res2,main = "PACF of Model B2")

# Independence tests
# Model A
Box.test(res3, lag = 19, type = c("Box-Pierce"), fitdf = 2)
Box.test(res3, lag = 19, type = c("Ljung-Box"), fitdf =2)
Box.test((res3)^2, lag = 19, type = c("Ljung-Box"), fitdf = 0)

# Model B
Box.test(res, lag = 19, type = c("Box-Pierce"), fitdf = 4)
Box.test(res, lag = 19, type = c("Ljung-Box"), fitdf = 4)
Box.test((res)^2, lag = 19, type = c("Ljung-Box"), fitdf = 0)

# Model B2
Box.test(res2, lag = 19, type = c("Box-Pierce"), fitdf = 3)
Box.test(res2, lag = 19, type = c("Ljung-Box"), fitdf = 3)
Box.test((res2)^2, lag = 19, type = c("Ljung-Box"), fitdf = 0)

# Yule-Walker AR(0) check
# Model A
ar(res3, aic = TRUE, order.max = NULL, method = c("yule-walker"))

# Model B
ar(res, aic = TRUE, order.max = NULL, method = c("yule-walker"))

# Model C
ar(res2, aic = TRUE, order.max = NULL, method = c("yule-walker"))

# Plot training data with prediction intervals and forecast values
pred.tr <- predict(model2, n.ahead = 12)
U= pred.tr$pred + 2*pred.tr$se
L= pred.tr$pred - 2*pred.tr$se
ts.plot(hawaiiCO2, xlim=c(340,length(hawaiiCO2)+12), ylim = c(335,max(U)), main = "Training Data Set wi
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points((length(hawaiiCO2)+1):(length(hawaiiCO2)+12), pred.tr$pred, col="red")

# Plot of test data
ts.plot(allData, xlim = c(373,length(hawaiiCO2)+12), ylim = c(350,max(U)), col="red", main = "Validatio

# Plot of test data with prediction intervals and forecast values
ts.plot(allData, xlim = c(373,length(hawaiiCO2)+12), ylim = c(350,max(U)), col="red", ylab = "CO2 (ppm)
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points((length(hawaiiCO2)+1):(length(hawaiiCO2)+12), pred.tr$pred, col="green")
points((length(hawaiiCO2)+1):(length(hawaiiCO2)+12), pred.tr$pred, col="black")
```