Time Series Analysis of Monthly Measurements of Carbon Dioxide above Mauna Loa, Hawaii from 1959 -1990

Kavin Indirajith

2022-12-02

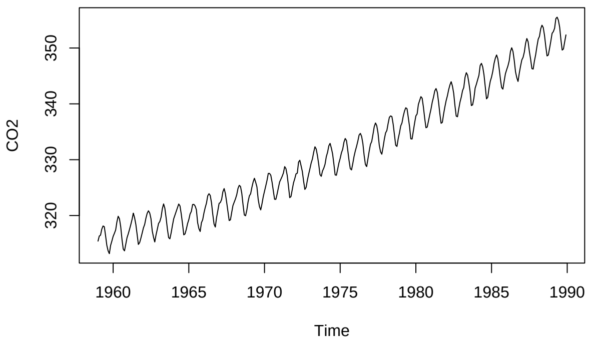
Abstract

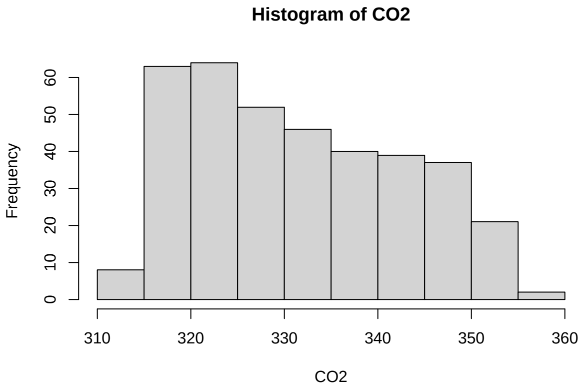
I analyze the time series data set that contains the monthly measurements of carbon dioxide above Mauna Loa, Hawaii from January 1959 to December 1990. I address whether this time series can be transformed and differenced into a stationary time series as wells whether it could be used to forecast future values after finding a suitable model. The data set was able to be turned into a stationary time series using differencing and a model was found using the stationary time series. The model was able to accurately forecast future values as the forecasted values lined up with the original data.

Introduction

In this report, I will be analyzing the monthly measurements of carbon dioxide above Mauna Loa, Hawaii over the course of 41 years, from January 1959 to December 1990. I find this dataset to be interesting as Mauna Loa is the largest volcano on earth, so being able to see how much carbon a volcano of that size puts out monthly and how that changes over time is very interesting. The source of this data set is the Time Series Data Library (tsdl) package in R and I will be using R to transform and forecast the data set. I will be using certain time series techniques to transform the data set into a stationary data set and find a suitable model to forecast future values. The techniques I will be using are transformations and differencing. After performing these techniques, I was able to turn the data set into a stationary data set, and, from that, found a model to use to forecast. Using the model to forecast future values, I found that the model was very suitable as the forecasted data was very close to the original data.

Section 1: Initial Plotting and Impressions

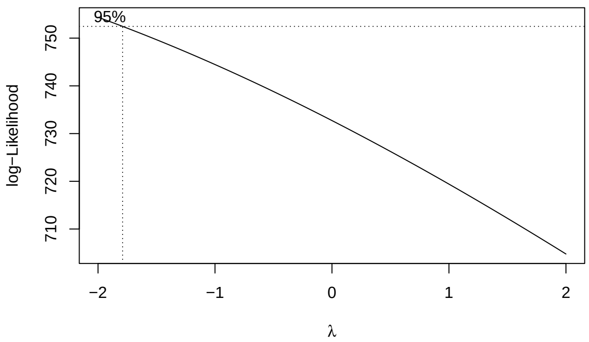


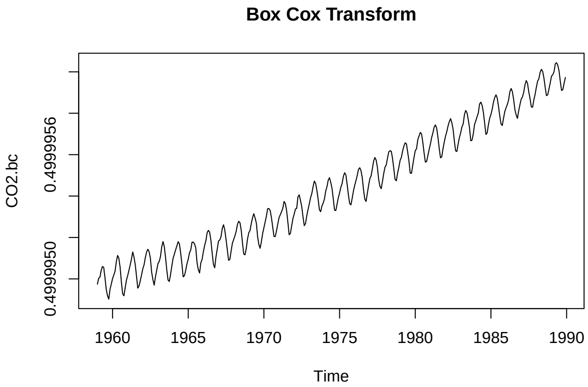


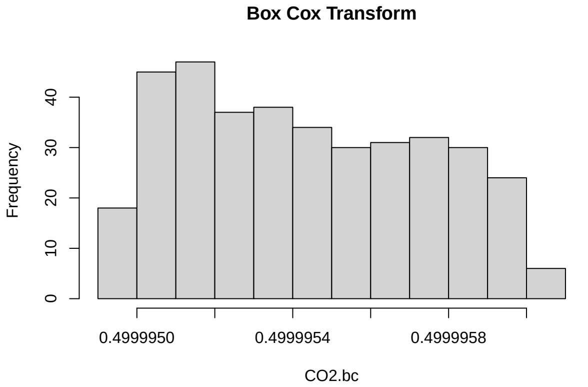
The graph does not have any apparent sharp changes in behavior. The graph exhibits the same behavior throughout, with a clear increasing trend. Additionally, the graph undergoes regular and predictable changes at fixed intervals. The CO2 emissions gradually increase from January to July, where it peaks. After July, the CO2 emissions gradually decrease until January, where the cycle starts once again. Based on this pattern, a seasonality component can be assumed to be present. The graph is non stationary due to these factors and it also has a non constant mean and variance.

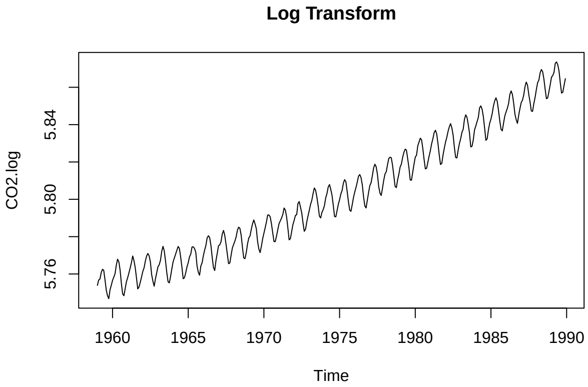
Section 2: Transformations and Differencing

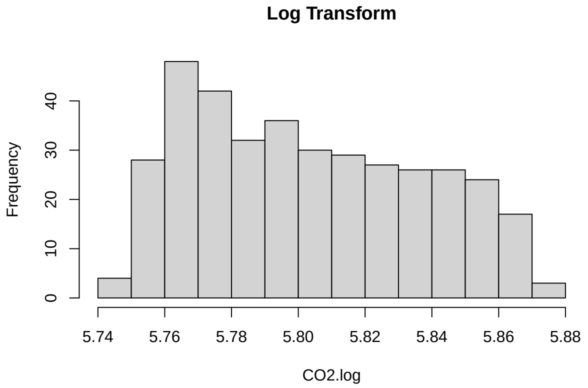
Transformations

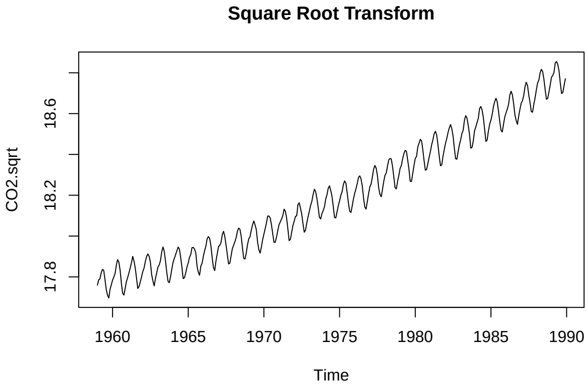


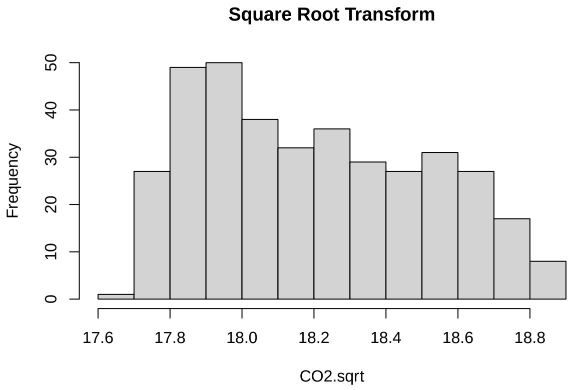






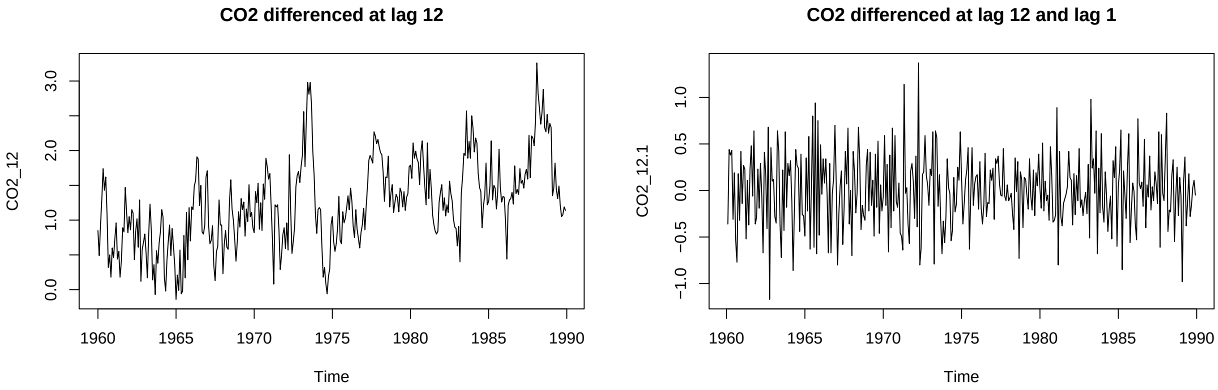


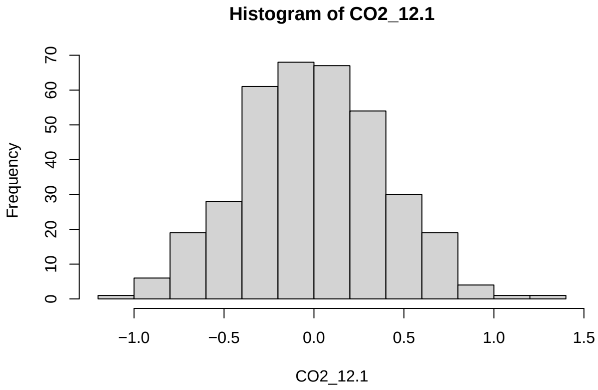




First, I decided to perform some transformations on the graph as the original graph had a non constant variance. This issue can be rectified by performing certain transformations on the time series. The three transformations I decided to perform were Box Cox, log, and square root. Looking at the three graphs, we can see that all three graphs look extremely similar. Thus, it is hard to choose a transformation from just the graphs. Moving on to the histograms, we can see that all the histograms display a right skewed distribution, which is extremely similar to the histogram of the original data. Next, we look at the log-likelihood vs λ graoh to see whether λ contains 1 in its confidence interval. 1 is within $\lambda's$ confidence interval so a transformation is not necessary. Considering all this information, I came to the conclusion that transformation is not required for this data set.

Differencing





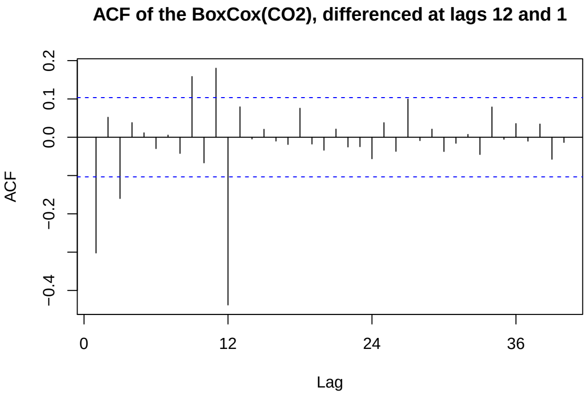
```
## [1] 126.8622
```

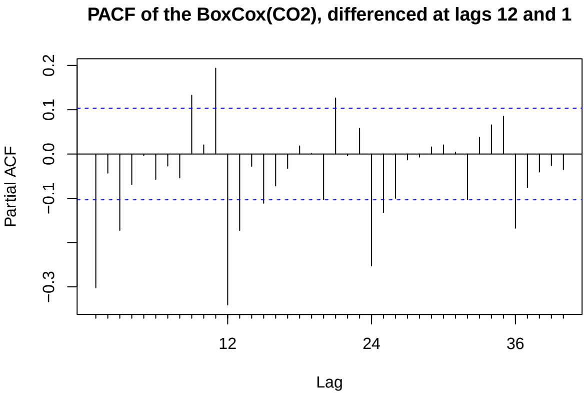
```
## [1] 0.387455
```

```
## [1] 0.153475
```

After deciding that no transformation was necessary for the time series, I moved on to differencing in order to make the time series stationary. First, I differenced at lag 12 as the graph showed a seasonality that was quite dependent on the months. Differencing at this lag made the seasonality, while still present, much less noticeable. This differencing reduced the original variance of 126.8622 by a large amount, with the new variance being 0.387455. With the seasonality mostly removed, I moved on to differencing once again as there was still a somewhat increasing trend to the graph. This time I differenced at lag 1 in order to remove the trend. This differencing removed the trend completely and gave me a stationary looking graph. Additionally, the variance was reduced once again, with it at 0.153475. The histogram of the data differenced at lags 12 and 1 is almost symmetrical and looks extremely similar to a normal distribution. Thus, the differencing was very effective in removing trend and seasonality.

Section 3: Plotting the ACF and PACF


ACF of the BoxCox(CO2), differenced at lags 12 and 1


PACF of the BoxCox(CO2), differenced at lags 12 and 1

The ACF is outside the confidence intervals at lags 1, 3, 9, 11, 12. The PACF is outside the confidence intervals at lags 1, 3, 9, 11, 12, 13, 22, 23, 24, 25, and 36. However, out of these lags, only lags 1, 3, 11, 12, 24, and 36 are significantly outside the interval.

Based on the ACF and PACF, the possible values for a SARIMA model are: Q = 1, P = 0, D = 1, q = 1 or 3, d = 1, p = 0, 1, 3, and s = 12.

Section 4: Choosing and Fitting the Model

Finding Suitable Models

```
##
## Call:
## arima(x = CO2, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 1), period = 12),
##     method = "ML")
##
## Coefficients:
##           ma1      ma2      ma3     sma1
##       -0.3532  -0.0206  -0.1010  -0.8556
## s.e.   0.0533   0.0562   0.0521   0.0328
##
## sigma^2 estimated as 0.07947:  log likelihood = -62.85,  aic = 135.69
```

```
## [1] 135.8648
```

```
##
## Call:
## arima(x = CO2, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 1), period = 12),
##     fixed = c(NA, 0, NA, NA), method = "ML")
##
## Coefficients:
##           ma1  ma2      ma3     sma1
##       -0.3598    0  -0.1073  -0.8559
## s.e.   0.0510    0   0.0492   0.0329
##
## sigma^2 estimated as 0.0795:  log likelihood = -62.91,  aic = 133.83
```

```
## [1] 133.9419
```

```
##
## Call:
## arima(x = CO2, order = c(1, 1, 3), seasonal = list(order = c(0, 1, 1), period = 12),
##     fixed = c(NA, NA, 0, NA, NA), method = "ML")
##
## Coefficients:
##          ar1      ma1  ma2      ma3     sma1
##       0.0602  -0.4125    0  -0.0974  -0.8557
## s.e.  0.1542   0.1423    0   0.0560   0.0328
##
## sigma^2 estimated as 0.07947:  log likelihood = -62.84,  aic = 135.67
```

```
## [1] 135.8412
```

```
##
## Call:
## arima(x = CO2, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12),
##     method = "ML")
##
## Coefficients:
##           ma1     sma1
##       -0.3642  -0.8573
## s.e.   0.0557   0.0327
##
## sigma^2 estimated as 0.08051:  log likelihood = -65.21,  aic = 136.42
```

```
## [1] 136.4851
```

I first tested two models: SARIMA(0,1,3)x$(0,1,1)_{12}$ and SARIMA(0,1,1)x$(0,1,1)_{12}$. They had an AICc of 135.8648 and 136.4851 respectively. However, I noticed that the ma2 coefficient of SARIMA(0,1,3)x$(0,1,1)_{12}$ was -0.0206, making it statistically insignificant. Thus, I decided to retest the model, this time with the ma2 coefficient removed. This updated model had an AICc of 133.9419, lower than before. Now, I introduced an AR component to this updated model to see whether it would give me a lower AICc than before. This new model of SARIMA(1,1,3)x$(0,1,1)_{12}$ had an AICc of 135.8412, marginally lower than the AICc of SARIMA(0,1,3)x$(0,1,1)_{12}$. I decided to choose SARIMA(0,1,3)x$(0,1,1)_{12}$ and SARIMA(0,1,3)x$(0,1,1)_{12}$ with the ma2 coefficient removed as my two models to diagnostic check.

Model A: $\nabla_1\nabla_{12}$ CO2 $= (1 - 0.3598_{0.0510}B - 0.1073_{0.0492}B^3)(1 - 0.8559_{0.0329}B^{12})Z_t$
$\hat{\sigma_z}^2$=0.0795
Model B: $\nabla_1\nabla_{12}$ CO2 $= (1 - 0.3532_{0.0533}B - 0.0206_{0.0562}B^2 - 0.1010_{0.0521}B^3)(1 - 0.8556_{0.0328}B^{12})$
$\hat{\sigma_z}^2$=0.07947

Checking for Stationarity and Invertibility

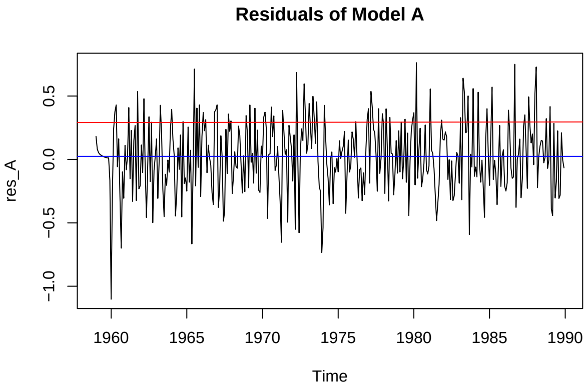
```
## [1]  1.587105+0.000000i -0.793552+2.289627i -0.793552-2.289627i
```

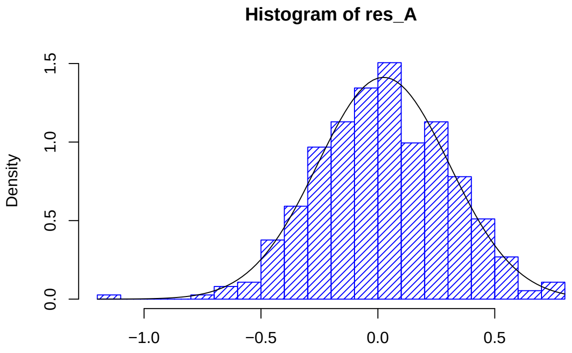
```
## [1]  1.573292-0.000000i -0.888626+2.345956i -0.888626-2.345956i
```

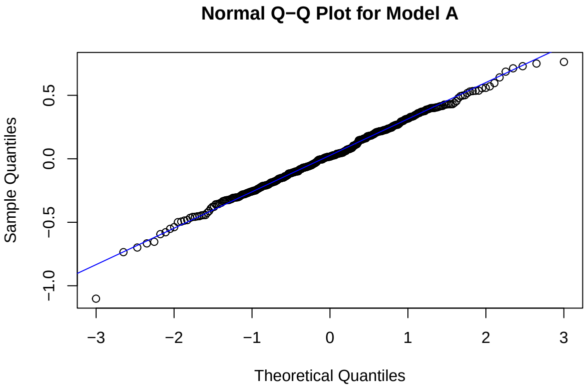
Both models are stationary as they are both pure MA models. Both models are both invertible as well as all their roots fall outside the unit circle.

Diagnostic Checking

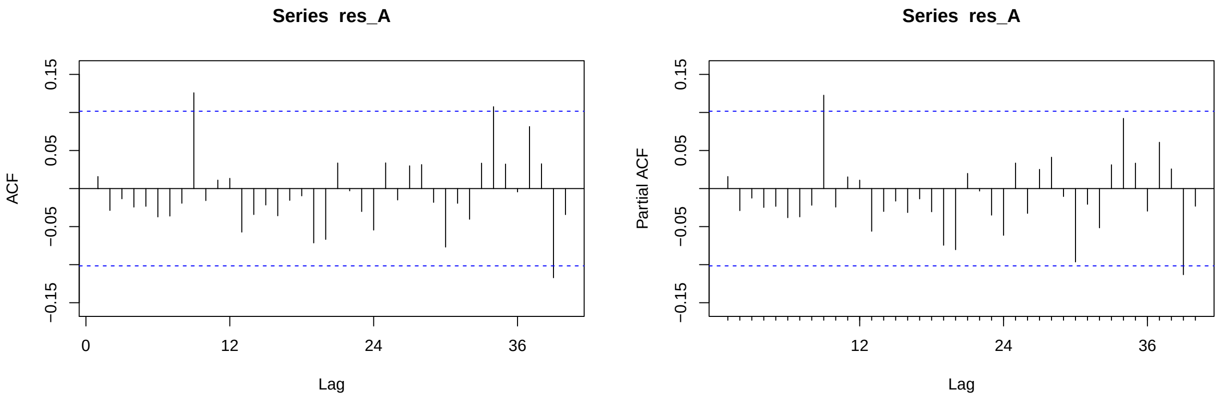
Model A







The plot of the residuals of Model A resemble WN. There is no trend, seasonality, and the mean is almost zero. The histogram resembles Gaussian and the QQ plot is almost a perfect straight line, with it only falling off towards the ends.


Series res_A

Series res_A

```
##
##  Shapiro-Wilk normality test
##
## data:  res_A
## W = 0.99558, p-value = 0.3774
```

```
##
##  Box-Pierce test
##
## data:  res_A
## X-squared = 8.12, df = 9, p-value = 0.5221
```

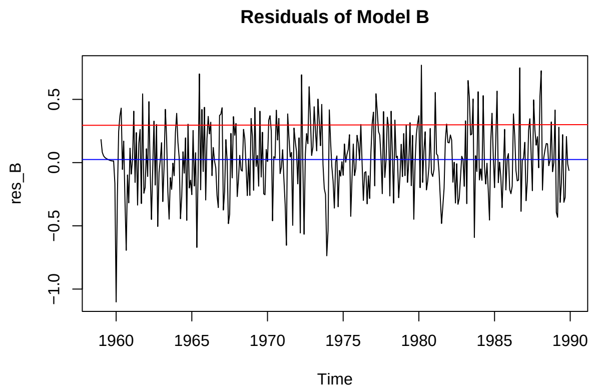
```
##
##  Box-Ljung test
##
## data:  res_A
## X-squared = 8.3451, df = 9, p-value = 0.4998
```

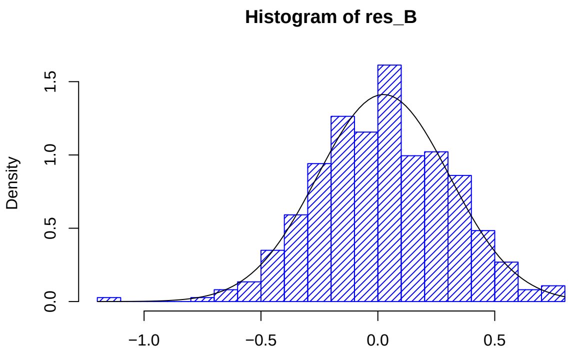
```
##
##  Box-Ljung test
##
## data:  (res_A)^2
## X-squared = 8.9936, df = 12, p-value = 0.7035
```

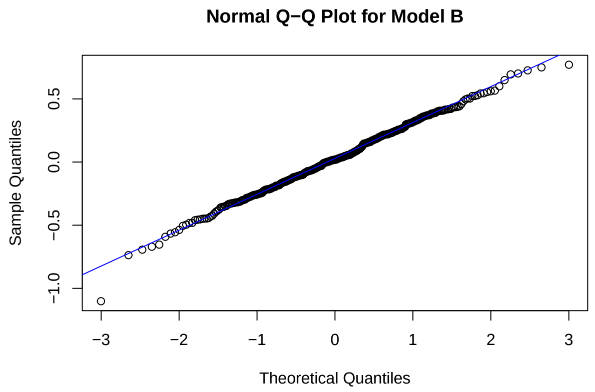
```
##
## Call:
## ar(x = res_A, aic = TRUE, order.max = NULL, method = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  0.07982
```

Since 95% of all lags in the ACF and PACF are within the confidence interval, by Bartlett's formula, the residuals can be considered WN. All p-values in the 4 tests are greater than 0.05 and the residuals were fitted to AR(0). Therefore, this model has passed diagnostic checking and can be used for forecasting.

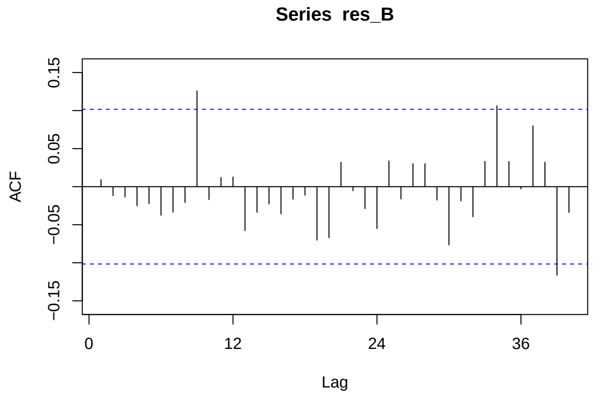
Model B

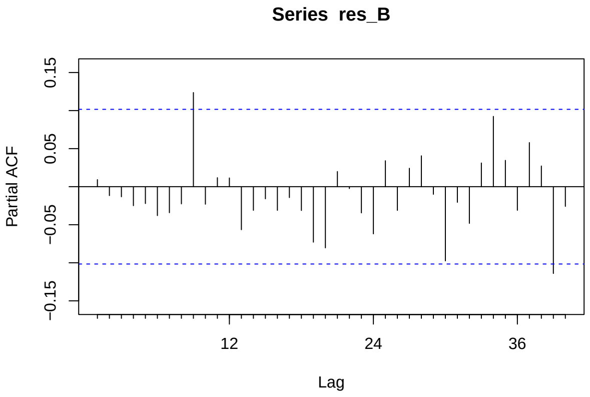






The graph of the residuals of Model B also follow WN, with no trend, seasonality, and a mean of near 0. The histogram resembles WN and the QQ plot is a near perfect straight line.





```
##
##  Shapiro-Wilk normality test
##
## data:  res_B
## W = 0.99582, p-value = 0.4299
```

```
##
##  Box-Pierce test
##
## data:  res_B
## X-squared = 7.7878, df = 9, p-value = 0.5557
```

```
##
##  Box-Ljung test
##
## data:  res_B
## X-squared = 8.0095, df = 9, p-value = 0.5332
```

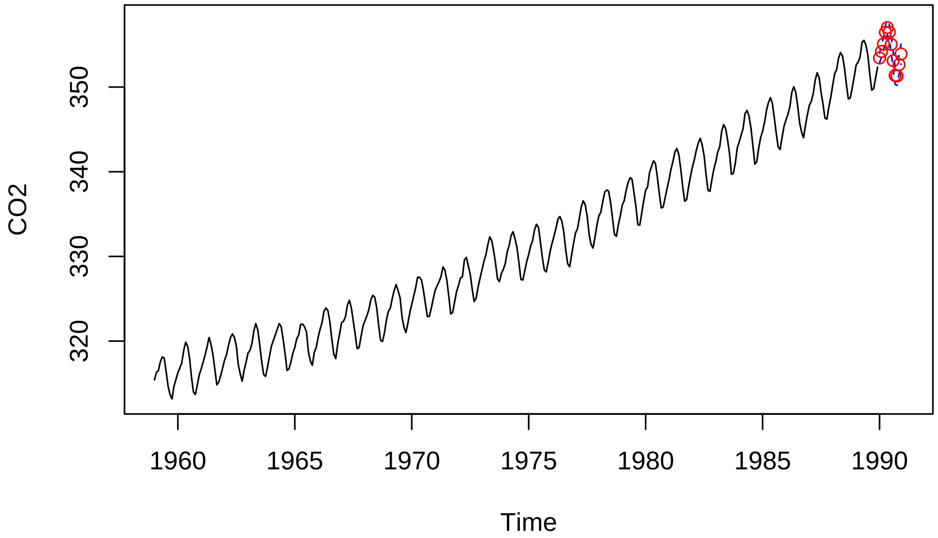
```
##
##  Box-Ljung test
##
## data:  (res_B)^2
## X-squared = 9.1637, df = 12, p-value = 0.6889
```

```
##
## Call:
## ar(x = res_B, aic = TRUE, order.max = NULL, method = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  0.07978
```

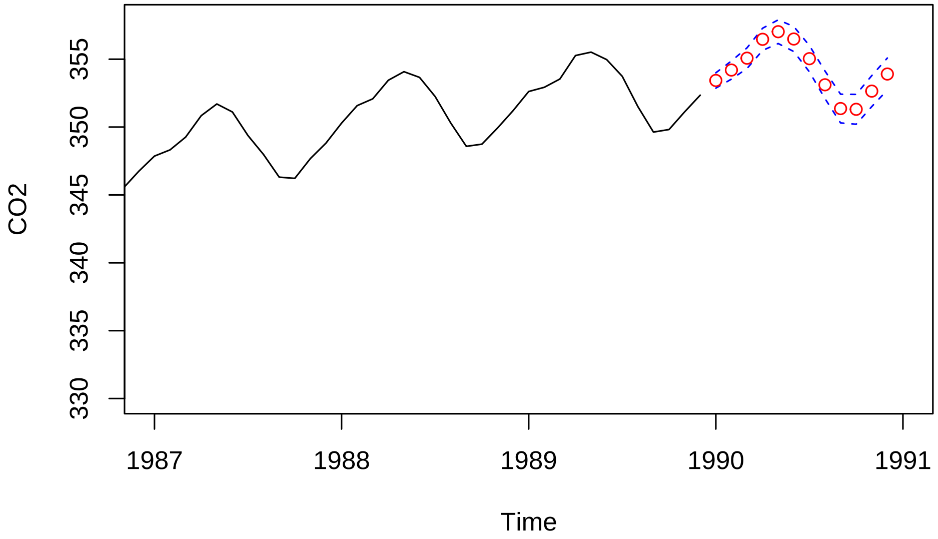
By Bartlett's Formula, the ACF and PACF of Model B follow WN. None of the p-values of the Portmanteau tests are less than 0.05 and the residuals were fitted to AR(0), so this model has passed diagnostic checking and thus ready to use for forecasting.

Since both models passed diagnostic checking, I will be using Model A as the model to use for forecasting as it is the simpler model.

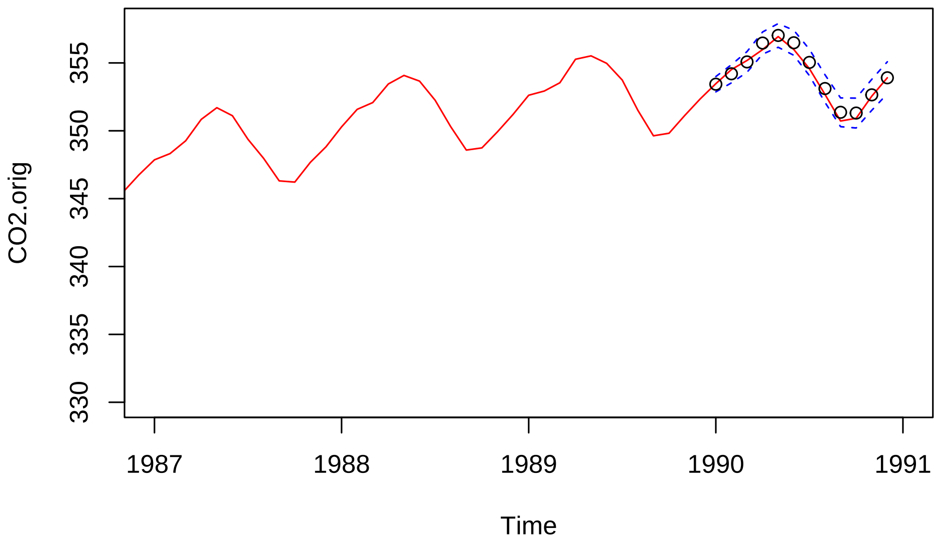
Section 5: Forecasting



Using Model A as the basis for forecast, I was able to forecast 12 months ahead of the dataset. However, due to the size of the data set, it is a bit hard to see in the full graph. As such, we will be looking at a zoomed graph that is limited to the time frame of 1987 to 1991.



Looking at the zoomed graph, we can see the forecasted points and their confidence intervals much more clearly. From the previous behavior of the graph and the shape of the confidence interval and forecasted points, it is highly likely that forecasting is correct and the data set will follow the pattern shown. We can confirm this by plotting the forecast on the original data set.



With the original data plotted onto the forecasted points, we can see that the original data did indeed follow the path forecasted by the model, with it completely inside the prediction intervals. So, the model I chose is good fit for the data set.

Conclusion

My goals for this project were to analyze a time series, transform it into a stationary time series, and find a model to forecast future values based off the transformed time series. I was able to remove the trend and seasonality from the original time series to make it stationary and I was able to find a very suitable model from that time series as the forecasted values were near accurate to the original data. The model I found was SARIMA(0,1,3)x$(0,1,1)_{12}$ without the ma2 coefficient. This model can algebraically written as $\nabla_1 \nabla_{12}$ CO2 $= (1 - 0.3598_{0.0510} B - 0.1073_{0.0492} B^3)(1 - 0.8559_{0.0329} B^{12}) Z_t$. So, all my goals were achieved to great success. I would like to thank Professor Feldman for looking over my project and pointing out any errors present within as well as helping me in the process of choosing a model.

References

Brockwell Peter J and Richard A Davis. 2002. Introduction to Time Series and Forecasting. 2nd ed. New York: Springer.

Shumway Robert H and David S Stoffer. 2006. Time Series Analysis and Its Applications : With R Examples 2Nd [updated] ed. New York: Springer.

Appendix

```
## Initial Plotting

x <- subset(tsdl,description = "Mauna") # Pulls data set from tsdl library
CO2.orig <- x[[1]] # Stores complete data set
```

```r
CO2 <- ts(CO2.orig[1:372], start = c(1959,1),
          frequency = 12) # Training data set with all data points except last 12
CO2_Test <- ts(CO2.orig[373:384], start = c(1990,1),
               frequency = 12) # Test data set with last 12 data points
plot.ts(CO2) # Plots training data set
hist(CO2) # Displays training data set

## Transformations

t <- 1:length(CO2)
bcTransform <- boxcox(CO2~t) # Plots log-likelihood against lambda to find optimal lambda
lambda <- bcTransform$x[which(bcTransform$y == max(bcTransform$y))] # Stores optimal lambda

CO2.bc <- (1/lambda)*(CO2^lambda-1) # Box Cox Transform
CO2.log <- log(CO2) # Log Transform
CO2.sqrt <- sqrt(CO2) # Square Root Transform

# Plots and Histograms of Transforms
plot.ts(CO2.bc, main = "Box Cox Transform")
hist(CO2.bc, main = "Box Cox Transform")
plot.ts(CO2.log, main = "Log Transform")
hist(CO2.log, main = "Log Transform")
plot.ts(CO2.sqrt, main = "Square Root Transform")
hist(CO2.sqrt, main = "Square Root Transform")

## Differencing

var(CO2) # Variance of original data
CO2_12 <- diff(CO2, lag = 12) # Differences data at lag 12 to remove seasonality
var(CO2_12) # Variance of data differenced at lag 12
plot(CO2_12,
     main = 'CO2 differenced at lag 12') # Plots data differenced at lag 12
CO2_12.1 <-
  diff(CO2_12, lag = 1) # Differences data differenced at lag 12 at lag 1 to remove trend
var(CO2_12.1) # Variance of data differenced at lag 12 and lag 1
plot(CO2_12.1,
     main = 'CO2 differenced at lag 12 and lag 1') # Plots data differenced at lag 12 and lag 1
hist(CO2_12.1) # Displays histogram of data differenced at lag 12 and lag 1


## Plotting ACF and PACF

Acf(CO2_12.1, main="ACF of the BoxCox(CO2), differenced at lags 12 and 1",lag.max = 40)
Pacf(CO2_12.1, main="PACF of the BoxCox(CO2), differenced at lags 12 and 1",lag.max = 40)


## Finding A Suitable Model

# Q = 1, P = 0, D = 1, q = 3, d = 1, p = 0, s = 12
arima(CO2, order=c(0,1,3),
      seasonal = list(order = c(0,1,1), period = 12), method="ML")
AICc(arima(CO2, order=c(0,1,3),
           seasonal = list(order = c(0,1,1), period = 12), method="ML"))
```

```
# Q = 1, P = 0, D = 1, q = 3, d = 1, p = 0, s = 12, with ma2 removed
arima(CO2, order=c(0,1,3),
      seasonal = list(order = c(0,1,1), period = 12), fixed = c(NA,0,NA,NA), method="ML")
AICc(arima(CO2, order=c(0,1,3),
           seasonal = list(order = c(0,1,1), period = 12), fixed = c(NA,0,NA,NA), method="ML"))

# Q = 1, P = 0, D = 1, q = 3, d = 1, p = 1, s = 12
arima(CO2, order=c(1,1,3),
      seasonal = list(order = c(0,1,1), period = 12), fixed = c(NA,NA,0,NA,NA), method="ML")
AICc(arima(CO2, order=c(1,1,3),
           seasonal = list(order = c(0,1,1), period = 12), fixed = c(NA,NA,0,NA,NA), method="ML"))

# Q = 1, P = 0, D = 1, q = 1, d = 1, p = 0, s = 12
arima(CO2, order=c(0,1,1),
      seasonal = list(order = c(0,1,1), period = 12), method="ML")
AICc(arima(CO2, order=c(0,1,1),
           seasonal = list(order = c(0,1,1), period = 12), method="ML"))

## Checking stationarity and invertibility

polyroot(c(1, -0.3598, 0, -0.1073))
polyroot(c(1,-0.3532,-0.0206,-0.1010))

## Diagnostic Checking

# Model A

ModelA <- arima(CO2, order=c(0,1,3), seasonal = list(order = c(0,1,1), period = 12),
                fixed = c(NA,0, NA, NA), method="ML") # Stores model in variable
res_A <- residuals(ModelA) # Stores residuals of model A
m_A <- mean(res_A) # Stores mean of residuals
std_A <- sqrt(var(res_A)) # Stores standard deviation of residuals
plot(res_A, main = "Residuals of Model A") # Plots residuals of model A
abline(h = m_A, col = 'blue') # adds mean line
fit <- lm(res_A ~ as.numeric(1:length(res_A))); abline(fit, col="red") # adds trend line

hist(res_A,density=20,breaks=20, col="blue", xlab="", prob=TRUE) # plots histogram of residuals
curve(dnorm(x,m_A,std_A), add=TRUE) # adds normal curve to histogram

qqnorm(res_A,main= "Normal Q-Q Plot for Model A") # plots Q-Q plot of residuals
qqline(res_A,col="blue") # adds line

Acf(res_A, lag.max = 40) # Plots ACF of residuals
Pacf(res_A, lag.max = 40) # Plots PACF of residuals

# Portmanteau Tests
shapiro.test(res_A)
Box.test(res_A, lag =12, type = c("Box-Pierce"), fitdf = 3)
Box.test(res_A, lag = 12, type = c("Ljung-Box"), fitdf = 3)
Box.test((res_A)^2, lag = 12, type = c("Ljung-Box"), fitdf = 0)

ar(res_A, aic = TRUE, order.max = NULL, method = c("yule-walker")) # Yule-walker estimation
```

```r
# Model B

ModelB <- arima(CO2, order=c(0,1,3),
                seasonal = list(order = c(0,1,1), period = 12),
                method="ML") # Stores model in variable
res_B <- residuals(ModelB) # Stores residuals of model B
m_B <- mean(res_B) # Stores mean of residuals
std_B  <- sqrt(var(res_B)) # Stores standard deviation of residuals
plot(res_B, main = "Residuals of Model B") # Plots residuals of model B
abline(h = m_B, col = 'blue') # adds mean line
fit <- lm(res_B ~ as.numeric(1:length(res_B))); abline(fit, col="red") # adds trend line

hist(res_B,density=20,breaks=20, col="blue", xlab="", prob=TRUE) # plots Q-Q plot of residuals
curve(dnorm(x,m_B,std_B), add=TRUE) # adds line

qqnorm(res_B,main= "Normal Q-Q Plot for Model B") # Plots ACF of residuals
qqline(res_B,col="blue") # Plots PACF of residuals

Acf(res_B, lag.max = 40) # Plots ACF of residuals
Pacf(res_B, lag.max = 40) # Plots PACF of residuals

# Portmanteau Tests
shapiro.test(res_B)
Box.test(res_B, lag =12, type = c("Box-Pierce"), fitdf = 3)
Box.test(res_B, lag = 12, type = c("Ljung-Box"), fitdf = 3)
Box.test((res_B)^2, lag = 12, type = c("Ljung-Box"), fitdf = 0)

ar(res_B, aic = TRUE, order.max = NULL, method = c("yule-walker")) # Yule-walker estimation


## Forecasting

pred <- predict(ModelA, n.ahead = 12) # predicts next 12 data points based on model
U = pred$pred + 2*pred$se # finds upper limit of confidence interval
L = pred$pred - 2*pred$se # finds lower limit of confidence interval
ts.plot(CO2, xlim = c(1959,1991),ylim = c(min(CO2),max(U))) # Plots data
lines(U , col="blue", lty="dashed") # adds upper confidence interval
lines(L , col="blue", lty="dashed") # adds lower confidence interval
points(pred$pred, col="red") # adds forecasted points

ts.plot(CO2, xlim = c(1987,1991),ylim = c(330,max(U))) # plots zoomed graph
lines(U , col="blue", lty="dashed") # adds upper confidence interval
lines(L , col="blue", lty="dashed") # adds lower confidence interval
points(pred$pred, col="red") # adds forecasted points

ts.plot(CO2.orig, xlim = c(1987,1991), ylim = c(330,max(U)), col="red") # plots original data
lines(U, col="blue", lty="dashed") # adds upper confidence interval
lines(L, col="blue", lty="dashed") # adds lower confidence interval
points(pred$pred, col="black") # adds forecasted points
```