Analysis and Forecasting of Atmospheric CO2 Levels

Rutvij Kortikar

3/8/2020

Introduction

I remember watching Al Gore's "An Inconvenient Truth" many years ago and always thinking what I could do. When I bought my first car, I opted for diesel as opposed to gasoline due to the higher energy density of diesel and the possibility to convert said vehicle to vegetable oil to reduce my reliance on fossil fuels. Although I haven't been able to afford an electric car, my second Job was at Tesla. I care a lot about the environment, so when it came to this final project, my first thought was trying to see how we could use the techniques we learned in PSTAT 174 to explore atmospheric trends and better understand the trends behind our precarious climate situation.

Choose a dataset that you will be interested to analyze for your class final project. URLs of time series libraries are posted on Gauchospace. Provide information about the project.

(a) Data set description: briefly describe the data set you plan to use in your project.

This data set is the CO2 concentration in the air over time. These data, known as the keeling curve. are collected at the Muana Loa Observatory on Hawaii from 1958 until today. The curve measures the mole fraction of CO2 in the air over time with units of micromol/mol.

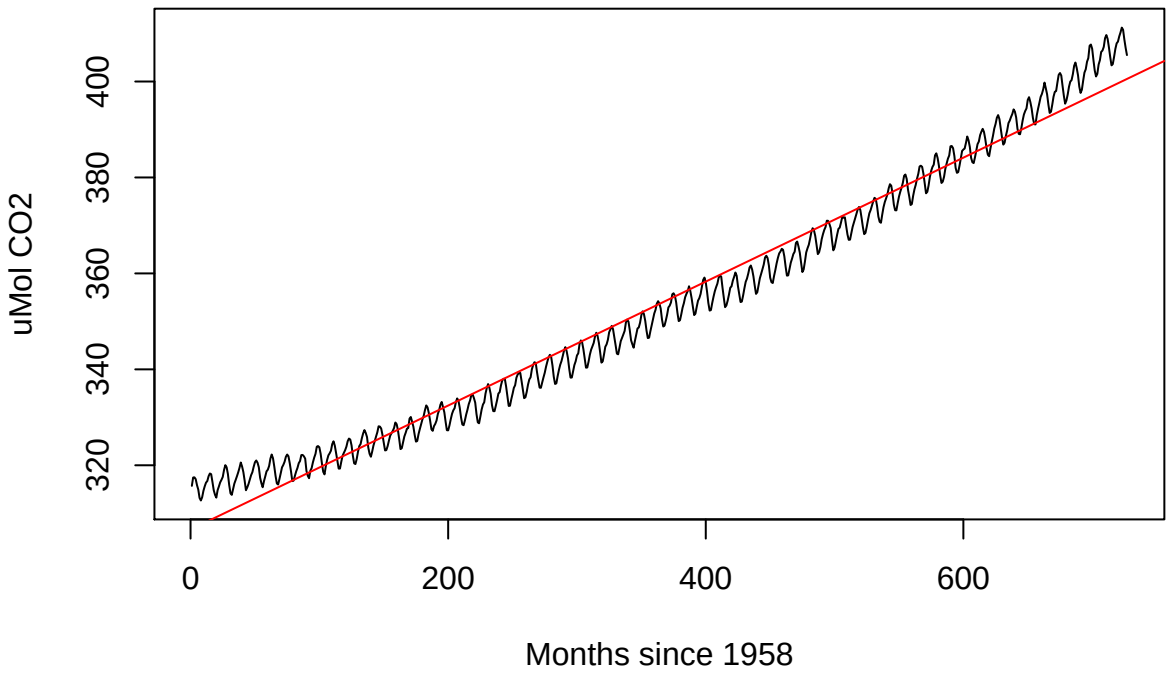
I plan to use this data set in my project to explore seasonality, trend, and other properties of time series. The science behind climate change is clear, and this is a great opportunity for me to learn more about how we can make mathematical extrapolations based on these data. Furthermore, forecasting behind significant climate changes could allow one to consider appropriate policy proposals to counteract or at the very least mitigate these trends.

(b) Motivation and objectives: briefly explain why this data set is important or interesting. Provide a clear description of the problem you plan to address using this dataset (for example to forecast)

This data set is very important because if CO2 concentrations rise above a certain point, the polar ice caps could melt, which would further accelerate the increase in temperature due to the loss of the cooling white ice that reflects sunlight.

Furthermore, this dataset has clear seasonality since oceans absorb and retain different amounts of gas depending on the season in the Northern Hemisphere (Since it contains the greatest biomass and hence absorbs most CO2 during the Summer), hence it is a good oppportunity to explore concepts we have learned in PSTAT174. It would be good to forecast how much time we have left to address climate change if we know how much longer we have before the CO2 levels reach catastropic proportions and cause widespread flooding following the melting of the polar ice caps which would further increase the global temperature.

(c) Plot and examine the main features of the graph, checking in particular whether there is (i) a trend; (ii) a seasonal component, (iii) any apparent sharp changes in behaviour. Explain in detail.

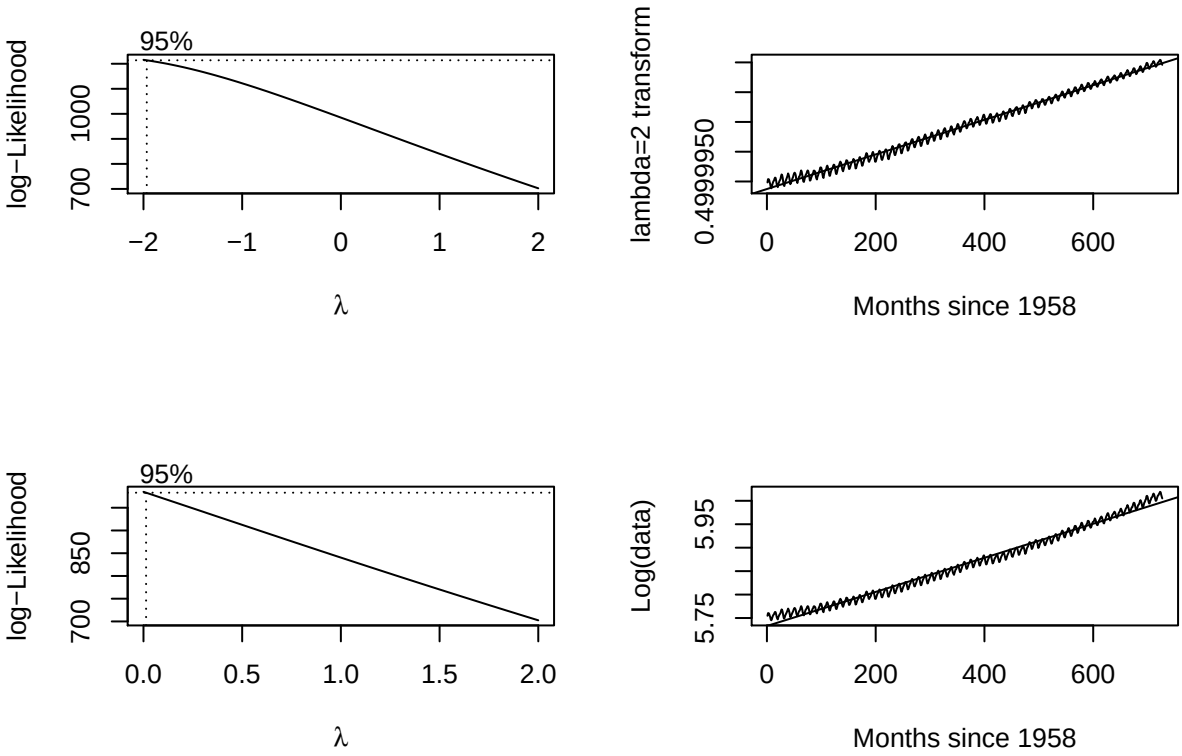


(i) There is a clear increasing trend in this time series. From when the data begins in 1958 to when it ends in 2018, there is a clear increase in CO2 concentrations. In fact, levels are higher now than they have ever been in history.

(ii) There is a definite seasonality component in the time series. Concentration increases in summer and decreases in winter. This means that differencing by lag 12 would be a good idea, since the data is collected on a monthly basis.

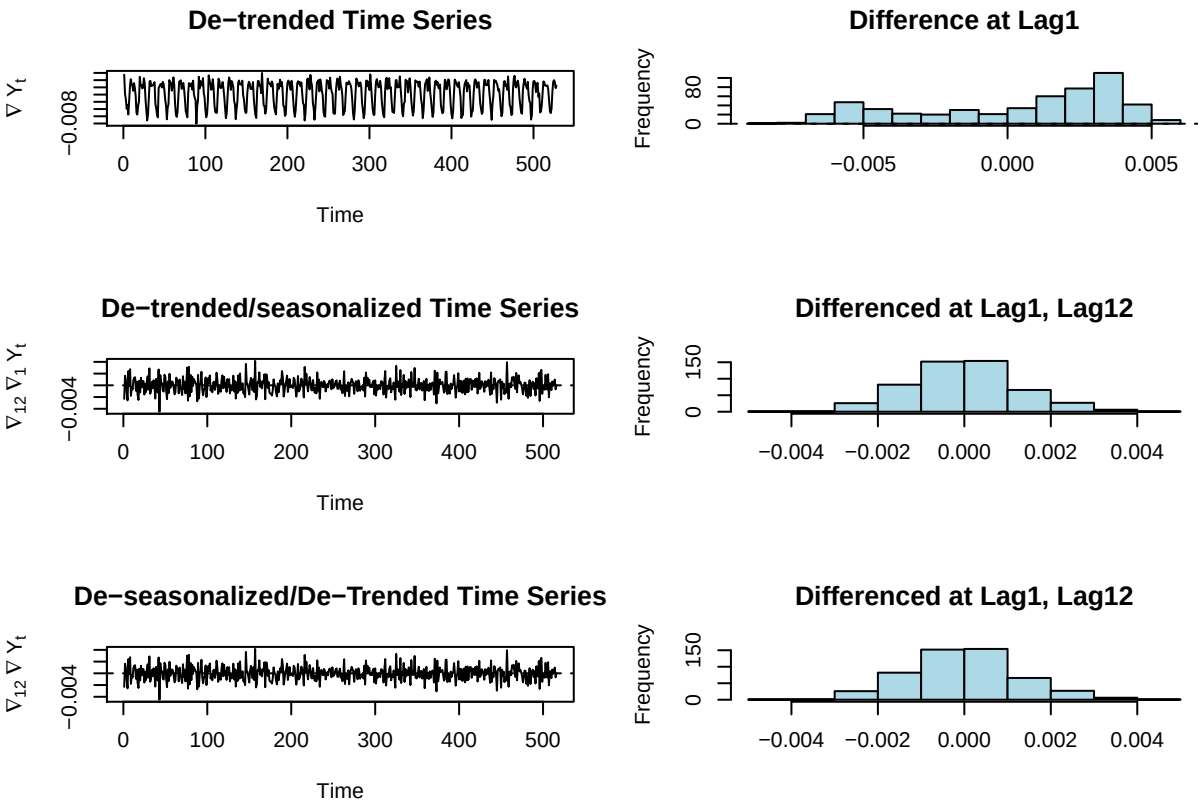
(iii) There are no apparent abberations in behaviour. This makes this a good candidate for time series analysis. If abberations were present, this could imply the presence of a nonlinear time series, and would require the utilisation of tools beyond the scope of this class.

As we can see, these data demand a Box-Cox transformation since the 'tails' of the data diverge from the linear trend despite the middle of the data being approximately linear. Firstly, we do a Train/Test split upon the data so that we can avoid overfitting in the model and see if we have a good margin of prediction when we finally select the model.



Without the constraint of lambda > 0, Box-Cox yields a lambda of -2.56 to maximize the likelihood. This initial graph gives a *much* better linear fit than using a log transform, however I notice from visual inspection that the variances are *not* constant as they are much higher for the first 200 values and much lower for values with an index greater than 600. We cannot proceed if we have unequal variances. Hence, we see the justification for the criteria in Brockwell/David for why lambda must have a lower bound of zero.

Based on these results, we take log transform and proceed with differencing. Log transform marginally improves the linear fit of the model as seen via decreased variance calculated and smoothes the slight tendency towards greater variance in CO2 concentrations for more recent values.



We decide to use lag 1 differencing for our model due to the analysis of the histograms. We can further improve upon this model by differencing at lag 12 to account for seasonality due to the periodic trends observed in the data. Furthermore, I observe that the order in which we difference at lag 1 and lag 12 does not matter, as both lead to a stationary distribution with residuals approximately gaussian as seen via histogram. This implies stationarity as demonstrated in an earlier HW problem.

```
## [1] 1.243124e-05

## [1] 1.570856e-06

## Warning in adf.test(data, k = 23): p-value greater than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  data
## Dickey-Fuller = -0.10269, Lag order = 23, p-value = 0.99
## alternative hypothesis: stationary

## Warning in adf.test(l1, k = 23): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  l1
## Dickey-Fuller = -6.0916, Lag order = 23, p-value = 0.01
## alternative hypothesis: stationary
```

```
## Warning in adf.test(l_1_12, k = 23): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  l_1_12
## Dickey-Fuller = -7.3594, Lag order = 23, p-value = 0.01
## alternative hypothesis: stationary
```

The Variances progressively go down as I difference, justifying the transformation. Interestingly, the Dickey-Fuller test for stationarity returns a statistically significant P value if we only difference for trend. This would imply that it is sufficient to difference with lag1 only and treat CO2 concentrations as a nonseasonal ARIMA model. However, I want to find the time series model that will most accurately map to the data, so I will difference further with lag 12 to account for seasonality and subsequently find the relevant parameters to do predictive forecasting.

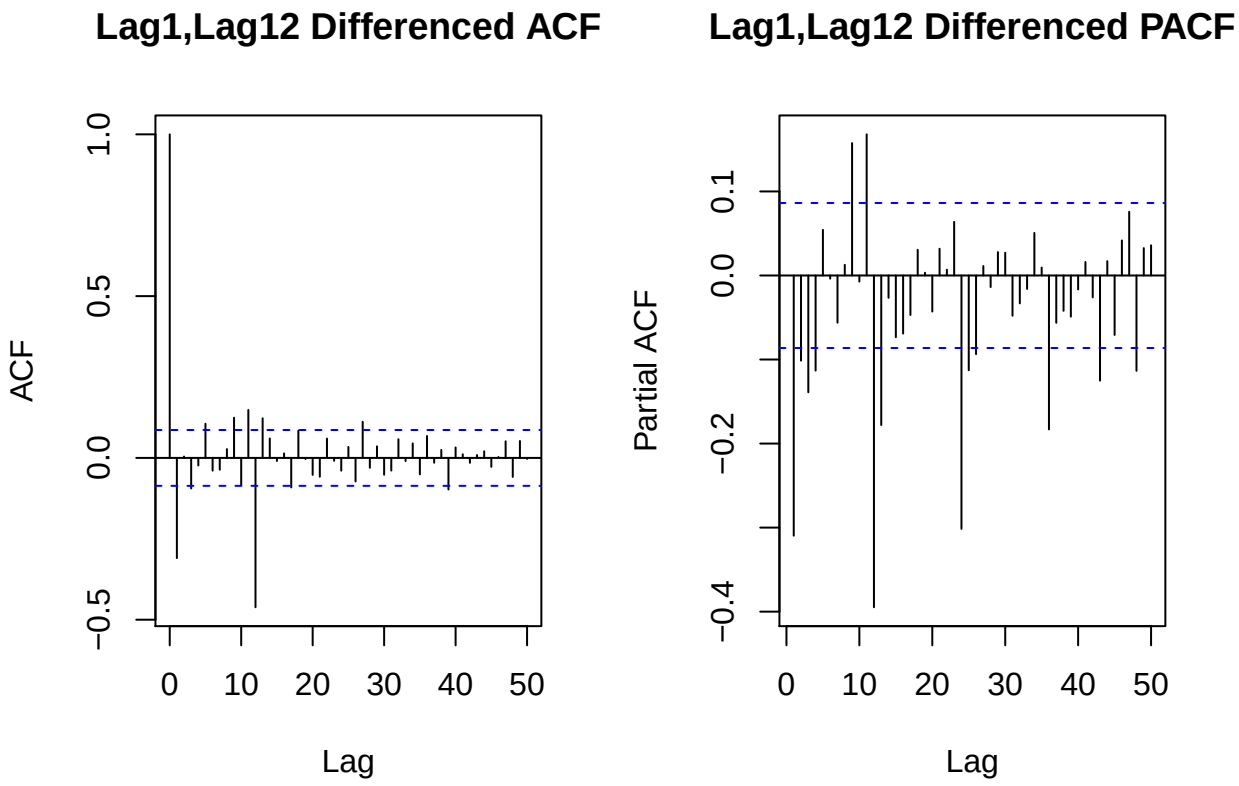
(d) Use any neccesary transformations to get a stationary series. Give a detailed explanation to justify your choice of a particular procedure. If you have used transformation, justify why. If you have used differencing, what lag did you use? Why? Is your series stationary now?

I ended up using a log transform along with differencing at lags 1 and 12 to obtain a stationary series. A log transform was used because the Box-Cox procedure yielded a lambda equal to zero which implies a log transform would maximize likelihood. Technically, the liklihood is maximized at a lambda of -2.5, however this lambda results in non-constant variances and would ultimately result in a non-stationary series. Hence I used the log transform which reduced the variances, indicating that it is an appropriate transform.

These data clearly indicate increasing CO2 concentration in the atmosphere as well as a seasonal trend when the CO2 concentration changes with an annual periodicity. We will need to use differencing with lag 12 to account for seasonality. The series also benefits from differencing at lag 1 to account for trend.

Both these operations respectively lower the variance and are commutative and yield the same underlying series, which imples that the final series is stationary. We can also perform PACF/ACF analysis as well as examine the properties of residuals to determine stationarity and find the appropriate models.

(e) Plot and analyze the ACF and PACF to preliminary identify your model(s): Plot ACF/PACF. What model(s) do they suggest? Explain your choice of p and q here.



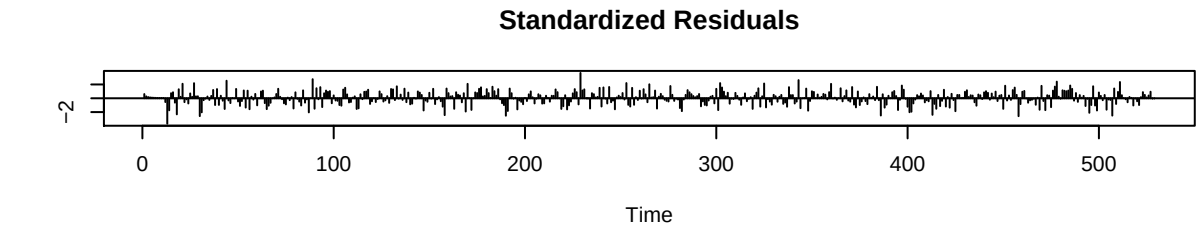
Based on these graphs I can postulate parameters for the relevant SARIMA models. Firstly I note that we have s=12 due to the periodicity in the exponential decay for the PACF. Furthermore we can determine d=1 since due to the statistically significant spike at lag1.We assume MA(q) since the values of sample acf can be taken as 0 after some lag q, which implies invertibility. Since the PACF cuts off after the first 4 values (notwithstanding correlated spikes due to the effect of seasonality) while the ACF tails off in a pattern of exponential decay, we can consider three models.

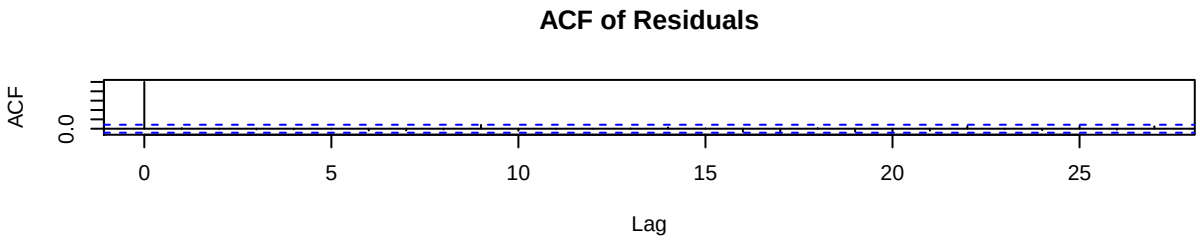
The first takes into account the first four statistically significant PACF values. The second is the simplest model we could use that explains these data. The third accounts for second order differencing, essentially postulating a geometric trend for CO2 growth.

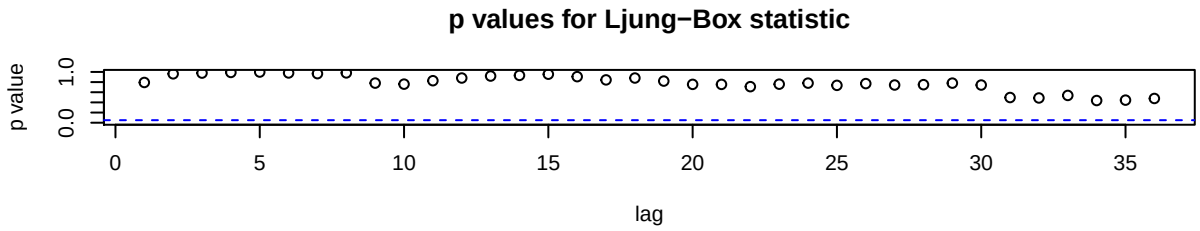
MODEL 1: SARIMA $(5,1,0) \times (0,1,1)_{12}$

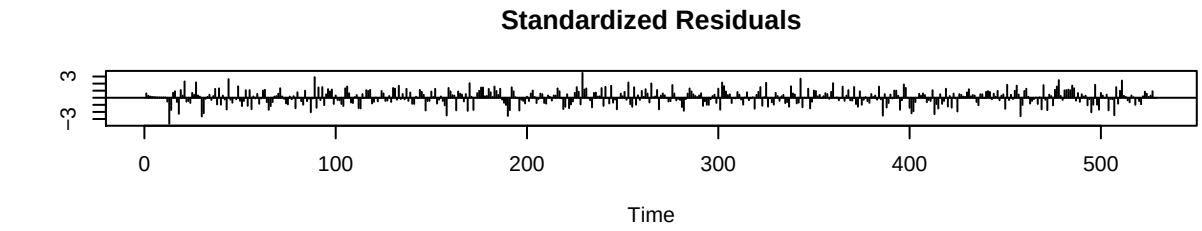
MODEL 2: SARIMA $(0,1,1) \times (0,1,1)_{12}$

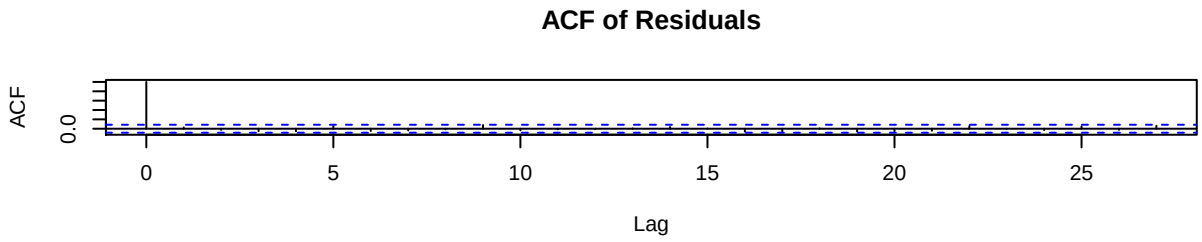
MODEL 3: SARIMA $(0,2,1) \times (0,2,1)_{12}$

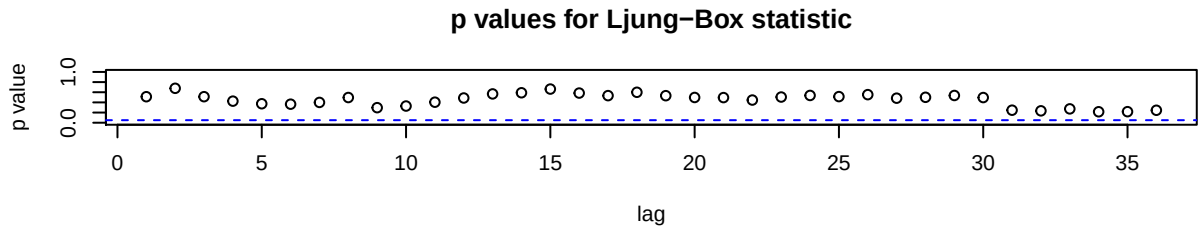

Standardized Residuals
−2
0
100
200
300
400
500
Time


ACF of Residuals
ACF
0.0
0
5
10
15
20
25
Lag

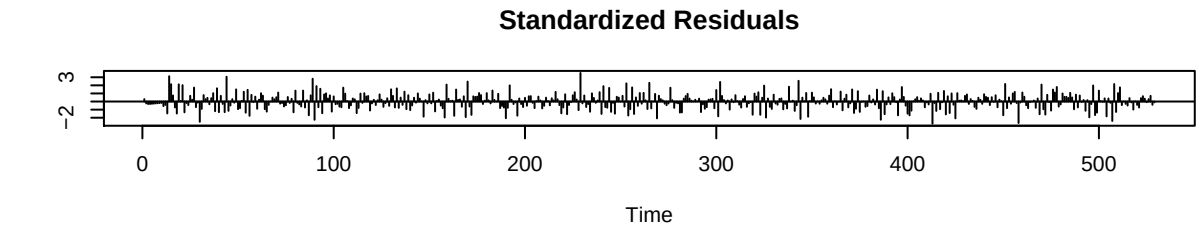

p values for Ljung−Box statistic
p value
0.0
1.0
0
5
10
15
20
25
30
35
lag

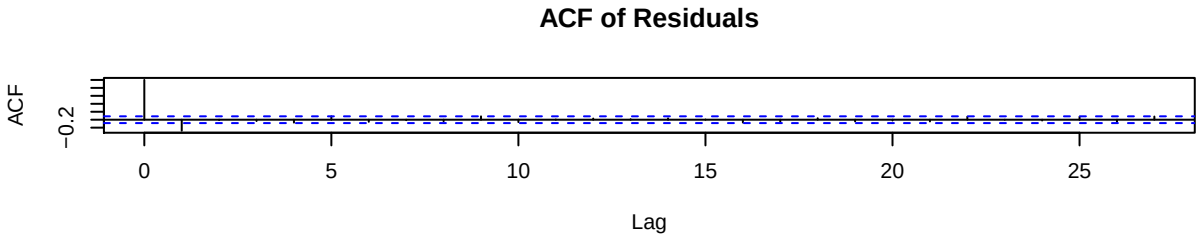


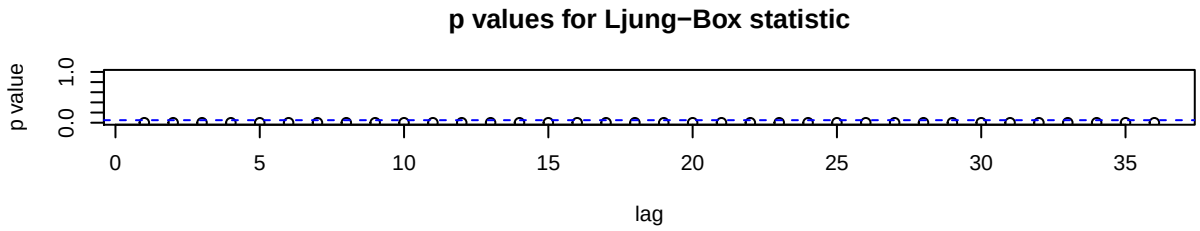




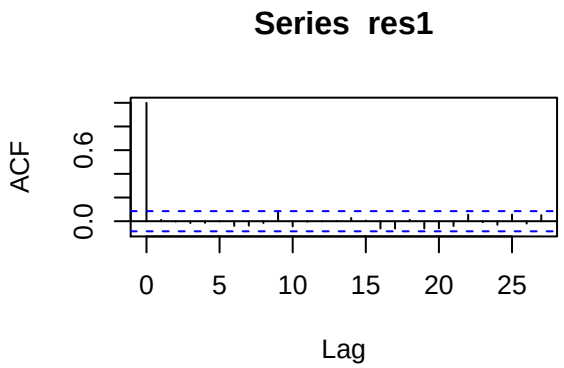

```
## Warning in arima(data, order = c(p, d, q), seasonal = list(order = c(P, :
## possible convergence problem: optim gave code = 1
```

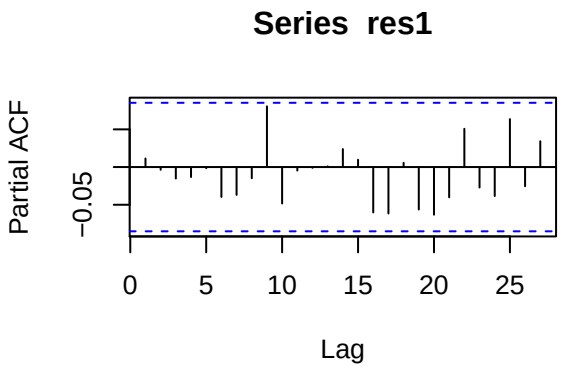


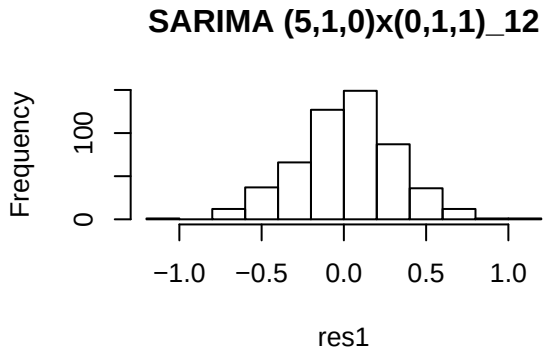


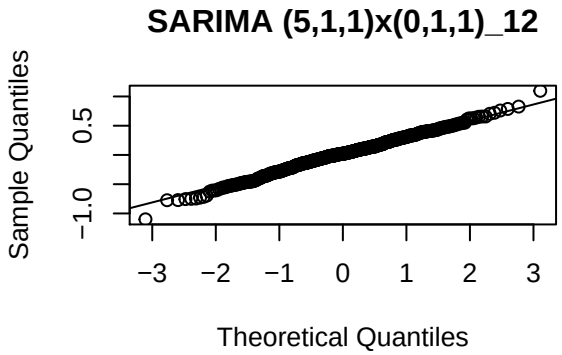


After examining the plots for the three different models, I observe that the plots for residuals and ACF have no anomalies. However, the Ljung-Box statistic is anomalous for the third model, indicative of autocorelation in the residuals. Hence this is a sign of overdifferencing. We can eliminate this model and generate qq plots and run the Shapiro-Wilk Test to ascertain normality.





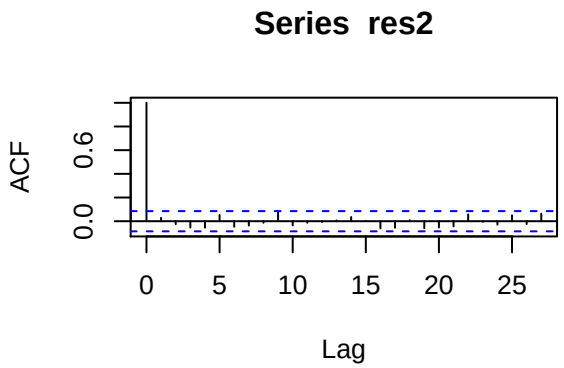


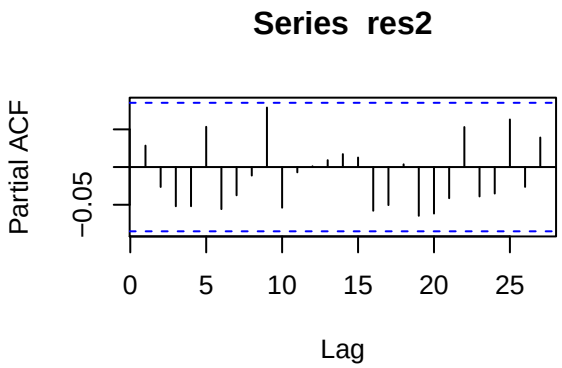


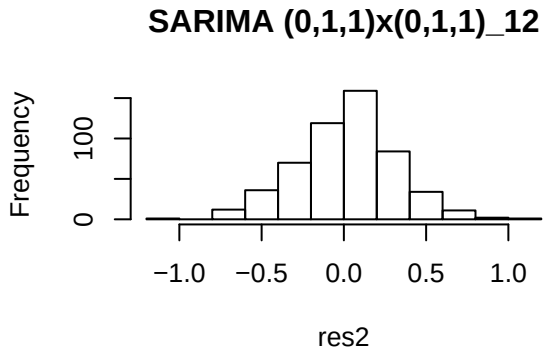

```
##
##  Shapiro-Wilk normality test
##
## data:  res1
## W = 0.99575, p-value = 0.1628
```

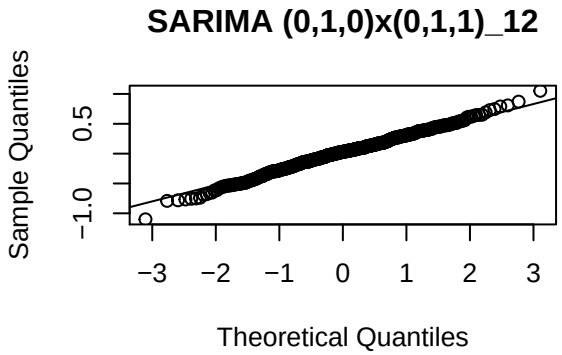
```
##
## Call:
## ar(x = residuals(mod1$fit), aic = TRUE, order.max = NULL, method = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  0.08947
```

Histogram shows pattern of Gaussian white noise and Quantiles are approximately linear. ACF values are within bounds of confidence interval and no anomalous behaviour detected for PACF. The Shapiro-Wilk p value is not statistically significant, indicative of normally distributed residuals. Since the autoregressive term for this model is 5, I can attempt yule-walker estimation for variance and obtain a value of 0.08947.










```
##
##  Shapiro-Wilk normality test
##
## data:  res2
## W = 0.99601, p-value = 0.2023
```

Again ACF/PACF look fine, and Shapiro value is in line with normally distributed residuals. We were able to eliminate one model The Shapiro-Wilk test fails for both models, however when we graph the log transformed residuals, a single value sticks out. We can provide justification for the removal of this value, namely, that it occurs near the beginning when values can deviate more from the trend. Hence we can simply take our training data to be between 15 and 460. The Shapiro-Wilk test now passes and we can see that our models reasonably extrapolate the test data. Yule-walker estimation is not appropriate here since we have no autoregressive terms.

AICc / Residual Analysis

```
## [1] "AICc for 3 models respectively"
```

```
## [1] -1.388756
```

```
## [1] -1.395485
```

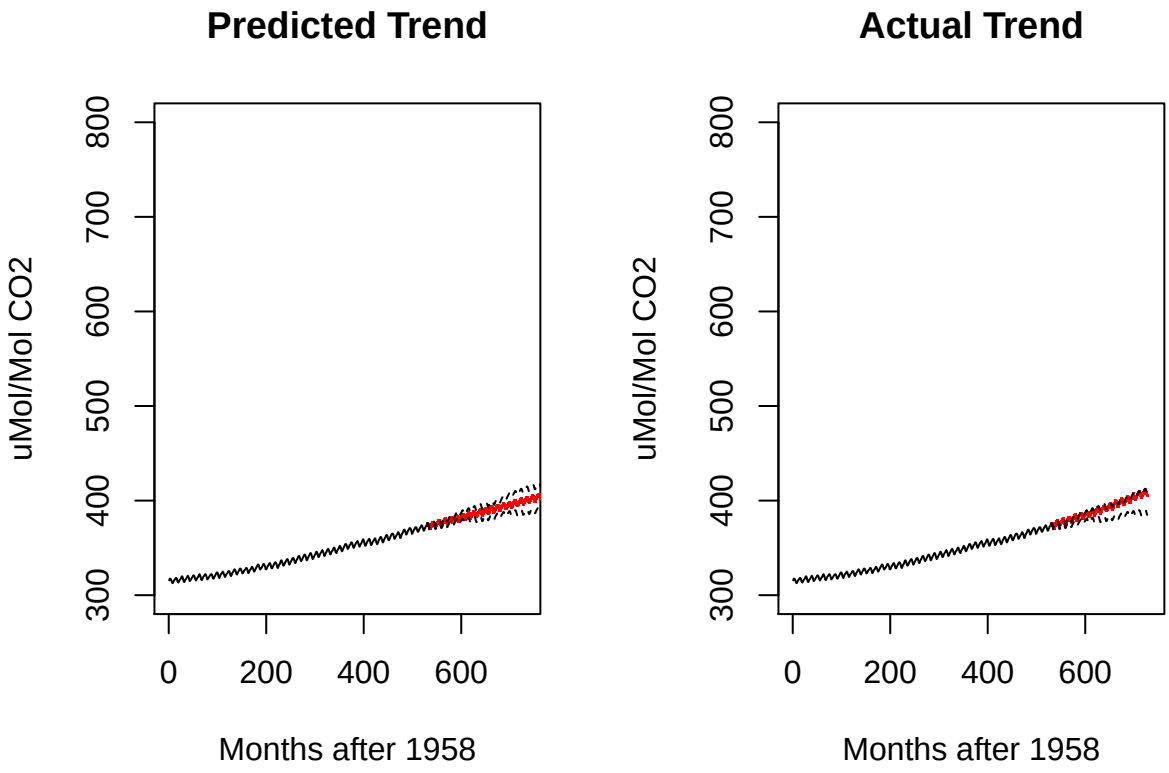
```
## [1] -1.270888
```

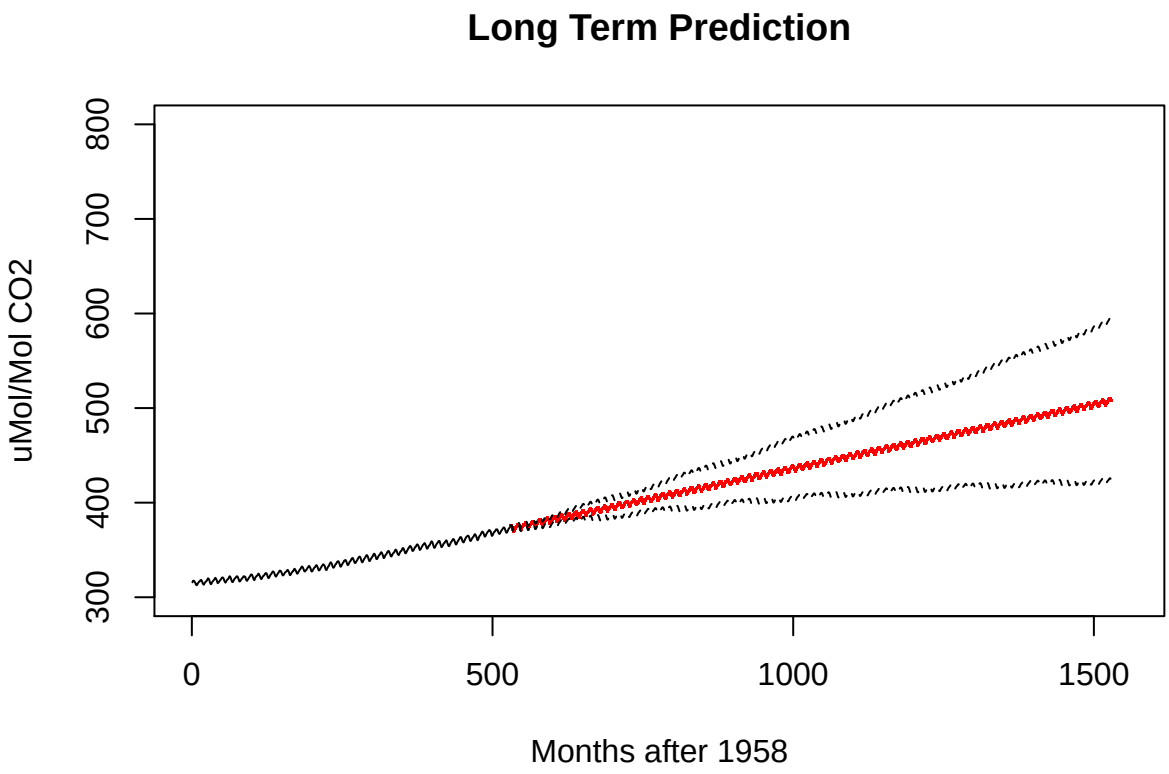
```
##
##  Box-Pierce test
##
## data:  res1
## X-squared = 15.772, df = 16, p-value = 0.469

##
##  Box-Pierce test
##
## data:  res1
## X-squared = 15.772, df = 16, p-value = 0.469

##
##  Box-Ljung test
##
## data:  res1^2
## X-squared = 18.664, df = 21, p-value = 0.6067

##
##  Box-Pierce test
##
## data:  res2
## X-squared = 19.913, df = 20, p-value = 0.4634

##
##  Box-Pierce test
##
## data:  res2
## X-squared = 19.913, df = 20, p-value = 0.4634

##
##  Box-Ljung test
##
## data:  res2^2
## X-squared = 18.754, df = 21, p-value = 0.6009

##
##  Box-Pierce test
##
## data:  res3
## X-squared = 64.444, df = 20, p-value = 1.431e-06

##
##  Box-Pierce test
##
## data:  res3
## X-squared = 64.444, df = 20, p-value = 1.431e-06

##
##  Box-Ljung test
##
## data:  res3^2
## X-squared = 35.574, df = 21, p-value = 0.0244
```

Here we set lag equal to the square root of the number of values, in this case sqrt(441)=21. Box-Ljung and Box-Pierce tests pass for Model 1 and Model 2. However, Box-Ljung, Box-Piece, and McLeod-Li (Box-Ljung of squared residuals) test fails for Model 3 indicating that differencing at lag 2 would be a case of overdifferencing, despite the reduction in variance of residuals and would **not** be appropriate. We then calculate the AICc for the remaining models.

Hence we find that the second model has the lowest AICc. (-1.4 < -1.39) As per the principle of parsimony, opt for this model. Now, we are able to extrapolate the results towards our test data and determine our predictive efficacy and appropriateness of the model and see how well we can extrapolate into the future.

Forecasting





The model shows that the predicted values are within the range of our testing data, hence this model is a good idea to forecast into the future. Of course, with any time series model, the variance increases as we look further into the future. Hence, any predictions must be taken with a grain of salt. Although we have a geometric growth rate for CO2, it is likely that factors such as decrease in human population growth as well as increasing use of green technologies and a more environmentally conscious attitude amongst people could slow or reverse this trend. Nevertheless, assuming that current trends continue, we can establish a resonable bound for when the uMol CO2 will increase by 50% to a level of 600 uMol/Mol causing the irreversible melting of the polar ice caps.

Conclusions/Policy Suggestions

Based on the chosen time series model we can propose that the sea levels will reach cataclysmic levels in around 30 to 80 years if current trends continue. Furthermore, when we look at our Forecast, I notice that due to our train test split, the more recent data encapsulating the geometric growth of the CO2 concentration during the 1990s is not included. This biases our model to *undershoot* the predicted CO2 concentration. Hence we should assume an earlier date for the predicted climate apocalypse.

When the ppm reaches 600, Antarctica melts. Once the Antarctic shelf melts to a significant degree, the resultant cascading chain reaction from the loss of the reflective polar ice cap will compound the effects of global warming and cause further temperature increase. Were both polar ice caps to melt, the sea levels would rise by approximately 300 metres. Granted, the planet has already demonstrated a surprising ability to maintain homeostasis despite the adverse consequences inflicted upon the environment by our species. However, it is unlikely that this will last. Already we are seeing the impact of historic CO2 levels through the increased incidence of fires and hurricanes as well as the destruction of 95% of the Great Barrier Reef. Although such climate change cannot be stopped quickly without annihilating industrial society as we know

it, it may perhaps be curbed in a reasonable amount of time so that we can avoid irreversible effects from the displacement of billions of people and numerous capital cities rendered inhabitable by the rising tide.

Such policy proposals should be taken with a grain of salt, shutting down all coal plants and imposing severe taxes upon the use of gasoline may curb the rapidity of current climate change trends. However, severe policy changes could displace the livelihood of billions, especially in developing countries currently reliant upon coal, logging, and other environmentally harmful means of generating electricity, managing transportation, cooking food, etc etc... Therefore, an ideal solution must be implementable within a decade, hence we cannot seriously consider techniques such as reforesting the Sahara/Australian Desert as the forests will take too much time to achieve significant rates of carbon capture.

IMMEDIATE PLAN ::

The solution to the climate crisis is clearly nuclear energy. However the type of plant I am proposing does not burn Uranium/Plutonium, but Thorium. Such reactors not only generate far less nuclear waste, but creates waste with fewer long lived superactinides and becomes negligibly radioactive within 100 years. Thorium salt reactors are also resistant to nuclear proliferation as well as nuclear meltdown due to inherently different design. Such reactors were sucessfully explored in the past, however such designs were ultimately discarded due to the military's needs for a reactor to produce Plutonium as opposed to energy due to the demands of the cold war.

Ultimately, the UN, combined with a country that has significant thorium reserves such as India or the US could spearhead such a "Manhattan Project" style effort to create an efficient Thorium salt reactor and put the patents in the public domain. Simple market forces would then displace the coal and oil industries and petroleum would only be used for niche chemical uses such as the manufacture of plastics and pharmaceuticals. Provided that such changes are made in a conservative timeline of around 20 years, less than 500 million people will be displaced and the inevitable global catastrophe will be mitigated to the level of a worldwide crisis, albeit a severe one.

Bibliography/References

SOURCE FOR DATA :: https://datahub.io/core

TEXTBOOK :: Brockwell & Davis, PA. 2011. Introduction to Time Series and Forecasting.

SOURCE FOR MELTING ESTIMATES :: https://www.livescience.com/64507-antarctica-ice-melt-earth-tilt.html

KEELING CURVE OFFICIAL SITE :: https://web.archive.org/web/20170426045237/https://scripps.ucsd.edu/programs/keelingcurve/

R DOCUMENTATION :: https://www.rdocumentation.org/

LFTR THORIUM REACTOR CITATION :: https://www.world-nuclear.org/information-library/current-and-future-generation/molten-salt-reactors.aspx

Appendix

```
knitr::opts_chunk$set(echo = TRUE)

## LOAD FILES

co2.csv = read.table("co2-mm-mlo_csv.csv",
sep=",", header=FALSE, skip=1, nrows=727)
#head(co2.csv) ## COMMENTED OUT FOR EASE OF READABILITY
#tail(co2.csv)

#here we use the interpolated values since they contain properly imputed values
#V5 would be trend, but we do not wish to apply a smoothing function yet
data <- co2.csv$V4
dataTrain <- data[1:529]
dataTest <- data[530:727]

mod1 <- lm(data~as.numeric(1:length(data)))
plot.ts(data,xlab="Months since 1958",ylab="uMol CO2")
abline(mod1, col="red")

## BOXCOX TRANSFORM
#boxcox transformation
require(MASS)
par(mfrow=c(2,2))
bct_co2 <-boxcox(data~ as.numeric(1:length(data))) #plots the graph
bct_co2$x[which(bct_co2$y== max(bct_co2$y))] #yields lambda

dataprime <-(data^-2 -1)/-2
mod2 <- lm(dataprime~as.numeric(1:length(dataprime)))
plot(dataprime, type="l",xlab="Months since 1958", ylab="lambda=2 transform") #UNEQUAL VARIANCES
abline(mod2)

bct_co2 <-boxcox(data~ as.numeric(1:length(data)), lambda=seq(0,2,1/42)) #plots the graph
bct_co2$x[which(bct_co2$y== max(bct_co2$y))] #yields lambda

mod3<- lm(log(data)~as.numeric(1:length(data)))
plot(log(data), type="l", xlab="Months since 1958", ylab="Log(data)") #UNEQUAL VARIANCES
abline(mod3)

## DIFFERENCING AT LAGS 1 AND 12
ldata <- log(dataTrain); #data has been log transformed

# Diference at lag = 1 to remove trend component
par(mfrow = c(3,2))
l1 = diff(ldata, 1)
ts.plot(l1,main = "De-trended Time Series",ylab = expression(nabla~Y[t]))
hist(l1, col="light blue", xlab="", main="Difference at Lag1")
abline(h = 0,lty = 2)
l_1_12 = diff(l1, 12)
ts.plot(l_1_12,main = "De-trended/seasonalized Time Series",
ylab = expression(nabla[12]~nabla[1]~Y[t]))
abline(h = 0,lty = 2)
```

```r
hist(l_1_12, col="light blue", xlab="", main="Differenced at Lag1, Lag12")
#If we switch the order of differencing we end up with the same time series:
l12 = diff(ldata, 12)
l_12_1 = diff(l12, 1)
ts.plot(l_12_1, main = "De-seasonalized/De-Trended Time Series",
ylab = expression(nabla[12]~nabla~Y[t]))
abline(h = 0,lty = 2)
hist(l_1_12, col="light blue", xlab="", main="Differenced at Lag1, Lag12")

## COMPARISON OF VARIANCES
#Var Lag1
var(l1)
#Var Lag2,Lag12
var(l_1_12)

## DICKEY-FULLER TESTS
library(tseries)
adf.test(data,k=23) #original series
adf.test(l1,k=23)  #differenced for trend
adf.test(l_1_12,k=23) #lag 1, and lag 12 difference.529 Values so lag=23

## PRELIMINARY RESIDUAL ANALYSIS
par(mfrow=c(1,2))
acf(l_1_12,lag=50,main="Lag1,Lag12 Differenced ACF")
pacf(l_1_12,lag=50,main="Lag1,Lag12 Differenced PACF")

## SARIMA MODEL + RESIDUALS
sarima=function(data,p,d,q,P=0,D=0,Q=0,S=-1){
  n=length(data)
  constant=1:n
  xmean=matrix(1,n,1)
  if (d>0)
  fitit=arima(data, order=c(p,d,q), seasonal=list(order=c(P,D,Q), period=S),xreg=constant,include.mean=
  if (d<.00001)
  fitit=arima(data, order=c(p,d,q), seasonal=list(order=c(P,D,Q), period=S),xreg=xmean,include.mean=F)
  if (d+D>1)
  fitit=arima(data, order=c(p,d,q), seasonal=list(order=c(P,D,Q), period=S))
  if (S < 0) goof=20 else goof=3*S
  tsdiag(fitit,gof.lag=goof)
  k=length(fitit$coef)
  BIC=log(fitit$sigma2)+(k*log(n)/n)
  AICc=log(fitit$sigma2)+((n+k)/(n-k-2))
  AIC=log(fitit$sigma2)+((n+2*k)/n)
  list(fit=fitit, AIC=AIC, AICc=AICc, BIC=BIC)
}

mod1 = sarima(dataTrain, 5, 1, 0, 0, 1, 1, 12)
mod2 = sarima(dataTrain, 0, 1, 1, 0, 1, 1, 12)
mod3 = sarima(dataTrain, 0,2,1,0,1,1,12)

res1 = residuals(mod1$fit)
res2 = residuals(mod2$fit)
res3 = residuals(mod3$fit)
```

```r
## SHAPIRO TEST/QQ-PLOT
par(mfrow=c(2,2))
acf(res1)
pacf(res1)
hist(res1,main="SARIMA (5,1,0)x(0,1,1)_12")
qqnorm(res1,main="SARIMA (5,1,1)x(0,1,1)_12")
qqline(res1)

shapiro.test(res1)
ar(residuals(mod1$fit), aic=TRUE, order.max=NULL, method=c("yule-walker"))

par(mfrow=c(2,2))
acf(res2)
pacf(res2)
hist(res2,main="SARIMA (0,1,1)x(0,1,1)_12")
qqnorm(res2,main="SARIMA (0,1,0)x(0,1,1)_12")
qqline(res2)

shapiro.test(res2)

# #AICC/BOX-LJUNG/MCLEOD-LI TESTS
par(mfrow=c(2,2))
library(qpcR)
print("AICc for 3 models respectively")
mod1$AICc
mod2$AICc
mod3$AICc

Box.test(res1, lag = 21, type = c("Box-Pierce"), fitdf= 5)
Box.test(res1, lag = 21, type = c("Box-Pierce"), fitdf= 5)
Box.test(res1^2, lag = 21, type = c("Ljung-Box"), fitdf= 0) #McLeod-Li

Box.test(res2, lag = 21, type = c("Box-Pierce"), fitdf= 1)
Box.test(res2, lag = 21, type = c("Box-Pierce"), fitdf= 1)
Box.test(res2^2, lag = 21, type = c("Ljung-Box"), fitdf= 0) #McLeod-Li

Box.test(res3, lag = 21, type = c("Box-Pierce"), fitdf= 1)
Box.test(res3, lag = 21, type = c("Box-Pierce"), fitdf= 1)
Box.test(res3^2, lag = 21, type = c("Ljung-Box"), fitdf= 0) #McLeod-Li

## FORECASTING PLOTS
par(mfrow=c(1,2))
mypred = predict(mod2$fit, n.ahead=1000)
ts.plot(dataTrain, xlim=c(0,733),ylim=c(300,800),xlab="Months after 1958",ylab="uMol/Mol CO2", main="Pr
points(530:1529,mypred$pred,pch=46,col="red")
lines(530:1529,mypred$pred+1.96*mypred$se,lty=2)
lines(530:1529,mypred$pred-1.96*mypred$se,lty=2)

ts.plot(dataTrain, xlim=c(0,733),ylim=c(300,800),xlab="Months after 1958",ylab="uMol/Mol CO2", main="Ac
points(530:727,dataTest,pch=46,col="red")
lines(530:727,mypred$pred[1:198]+1.96*mypred$se[1:198],lty=2)
lines(530:727,mypred$pred[1:198]-1.96*mypred$se[1:198],lty=2)
```

```r
## LONG TERM PLOT
par(mfrow=c(1,1))
mypred = predict(mod2$fit, n.ahead=1000)
ts.plot(dataTrain, xlim=c(0,1555),ylim=c(300,800),xlab="Months after 1958",ylab="uMol/Mol CO2", main="L
points(530:1529,mypred$pred,pch=46,col="red")
lines(530:1529,mypred$pred+1.96*mypred$se,lty=2)
lines(530:1529,mypred$pred-1.96*mypred$se,lty=2)
```