Forecasting historical atmospheric carbon dioxide levels above Mauna Loa, Hawaii using the SARIMA model family.

John R. Inston *University of California Santa Barbara*

In this report we aim to select and fit an appropriate Seasonal Autoregressive Integrated Moving Average (SARIMA) model in order to produce accuruate five year ahead forecasts of atmospheric CO_2 levels above Mauna Loa, Hawaii. We investigate the underlying properties of the time series data, the suitability of SARIMA models, and whether the forecasts they produce accurately capture future observations. Throughout the project we implement various univariate time series analysis techniques including exploratory analysis, model selection procedures based on the adjusted Akaike Information Criteria, diagnostic checking, residual analysis and spectral analysis. Our first choice SARIMA$(0,1,3)(0,1,1)_{12}$ model performed well, producing approximately Gaussian WN residuals and accurate forecasts with confidence intervals capturing the true observed future values of the time series.

Keywords: Time Series Analysis, SARIMA Models, Atmospheric Science

Introduction

Several geophysical time series contain seasonal variation with one key example being carbon dioxide (CO_2) atmospheric concentration which typically is known to have annual seasonality - largely due to plant photosynthesis and respiration - superimposed upon an increasing trend caused primarily by human CO_2 production. In this project we examine a data set consisting of monthly measurements of atmospheric CO_2 above Mauna Lao, Hawaii in the 32 year period from between 1959 and 1990 with the aim of producing accurate forecasts - in our case a 5 year ahead forecast - by fitting a suitable model from the SARIMA model family. We use a variety of time series analysis including exploratory analysis (time series decomposition, stationarity transformations, inspecting series autocorrelations and partial autocorrelations, tests for normality, stationarity and autocorrelation), SARIMA fitting and model selection (checking for stationarity and invertibility, coefficient significance testing, performance criteria) and residual analysis (portmanteau tests, residual WN tests, model spectral decomposition). We find that the SARIMA$(0,1,3)(0,1,1)_{12}$ model is our optimal model candidate and our subsequent residual analysis, forecasting and spectral decomposition indicates that our model produces accurate forecasts which captured the future true observations of the time series. Our analysis was performed using the programming language R and a full breakdown of our code can be found in the Appendices.

Data

Data Description

Our data set contains monthly measurements of atmospheric carbon dioxide above Mauna Loa, Hawaii from January 1959 to December 1990, measured in parts per million (ppm). The data was originally collected from the Scripps Institute of Oceanography, La Jolla, California and made readily available through the `tsdl` package developed by Rob Hyndman, Professor of Statistics at Monash University, Australia. The data set initially contained missing values which were subsequently filled in by linear interpolation. Our raw data series, labelled by $x_t,\ t = 1, ..., 384$ contains 384 observations.

Exploratory Analysis

We split our data into disjoint training and testing sets, with the latter containing the last 5 years of observed CO_2 levels to be used to analyse the performance of the foreacsts produced by our chosen final model. Our

training data set contains 324 observations whilst our testing data set contains 60. Using the training data we produce a time series plot, shown below in Figure 1.

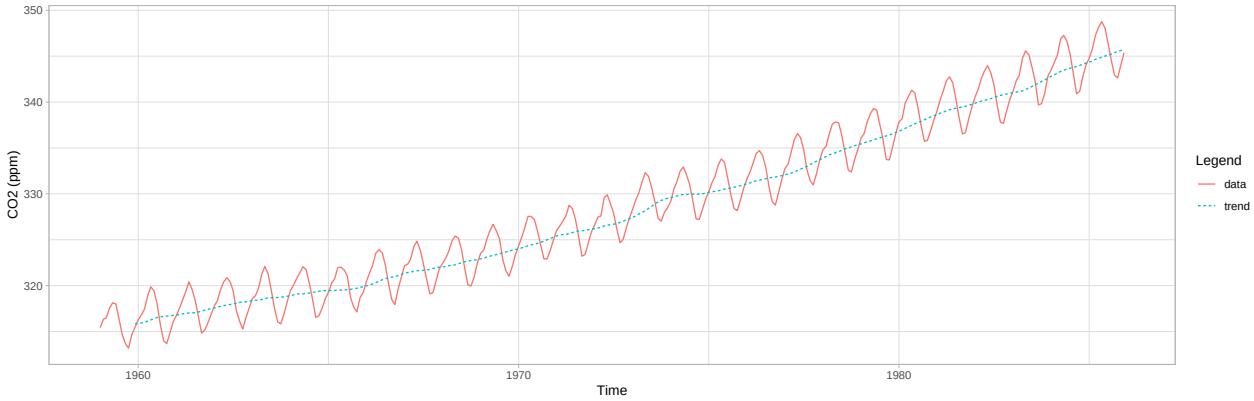


Figure 1: Atmospheric carbon dioxide above Mauna Loa, Hawaii.

We see clear evidence of both an increasing polynomial trend and a seasonal component (with a 12 month period) implying that our raw data is non-stationary. The variance of our data appears constant and there is no evidence of sharp changes in behaviour indicating that a SARIMA model will likely suit our data well. We decompose the time series in Figure 2 below to provide a clearer picture of the seasonal and trend components and since we observe no evidence of heteroscedastic behaviour in the residuals we opt not to transform our data.

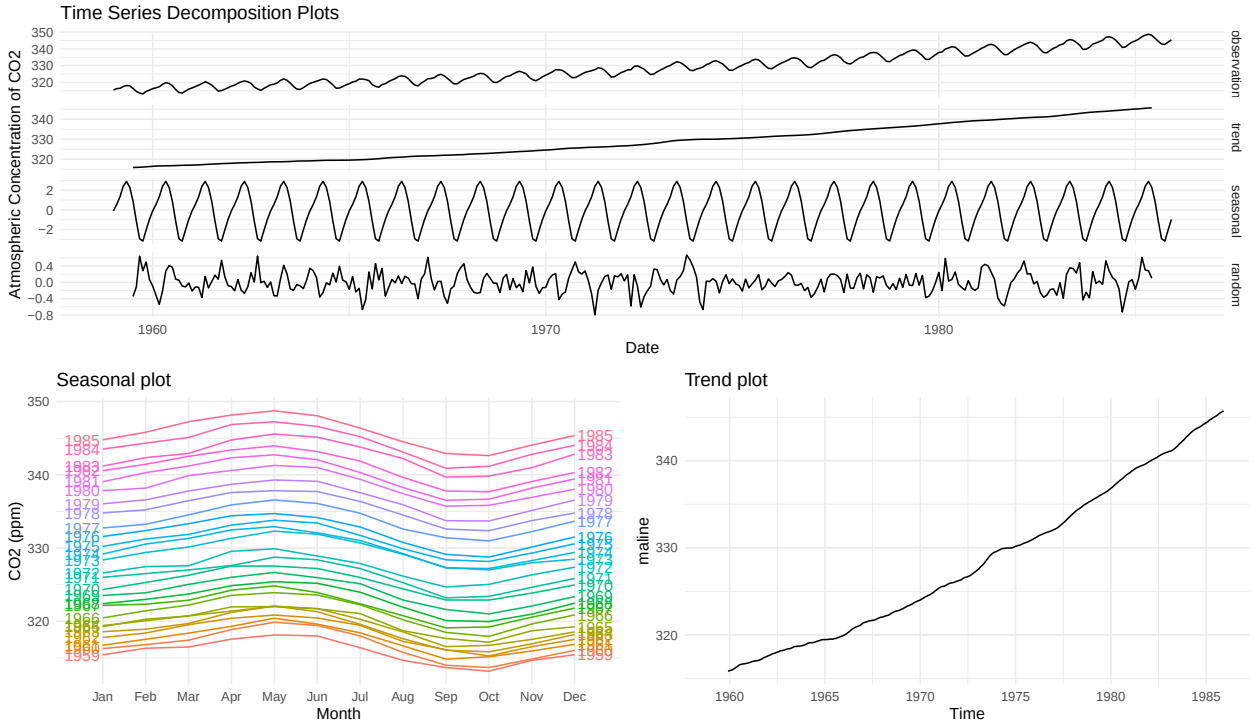


Figure 2: Time Series Decomposition.

We produce a plot of the autocorrelation function (ACF) of the data, shown below in Figure 3, from which we see highly significant autocorrelations for all lags up to 40 which slowly decay and show seasonality.

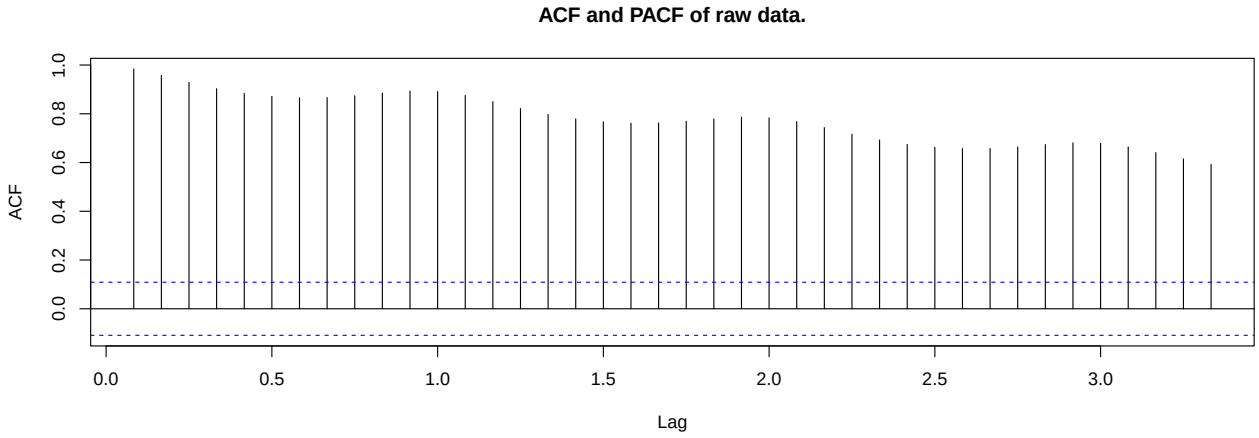


Figure 3: ACF and PACF of raw data.

We attempt to obtain stationarity by differencing, first with lag 12 to remove the seasonal trend. Mathematically this transformation is denoted:

$$y_t = x_t - x_{t-12} = (1 - B^{12})x_t = \nabla_{12} x_t$$

where y_t is our seasonally differenced data series and B is the back-shift operator whereby $B^k x_t = x_{t-k}$.

A time series plot of our differenced data is shown below in Figure 4 along with a trend line, from which we can see that all evidence of a seasonal trend has been removed but the data appears to still be non-stationary as a linear trend is still present. Alongside we include the ACF plot of our seasonally differenced data from which we see the decaying autocorrelation typical of data with a polynomial trend.

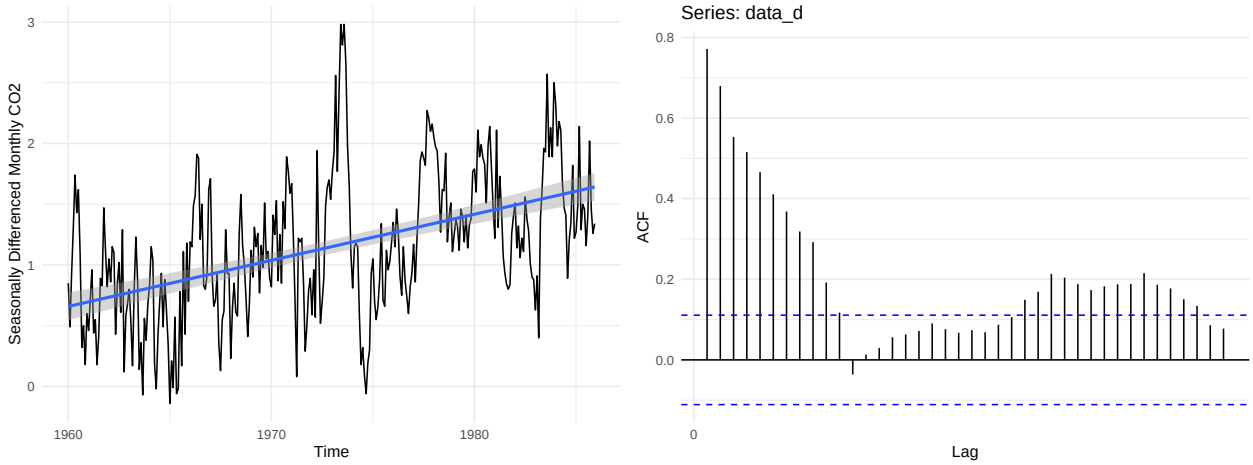


Figure 4: Time series and ACF plot of seasonally differenced data.

We therefore difference again with lag 1 to remove the linear trend from in the data, denoted mathematically by:

$$u_t = y_t - y_{t-1} = (1-B)y_t = (1-B)(1-B^{12})x_t = \nabla_1 \nabla_{12} x_t$$

where u_t is our twice differenced data. We produce another time series plot of our twice differenced data below in Figure 6.

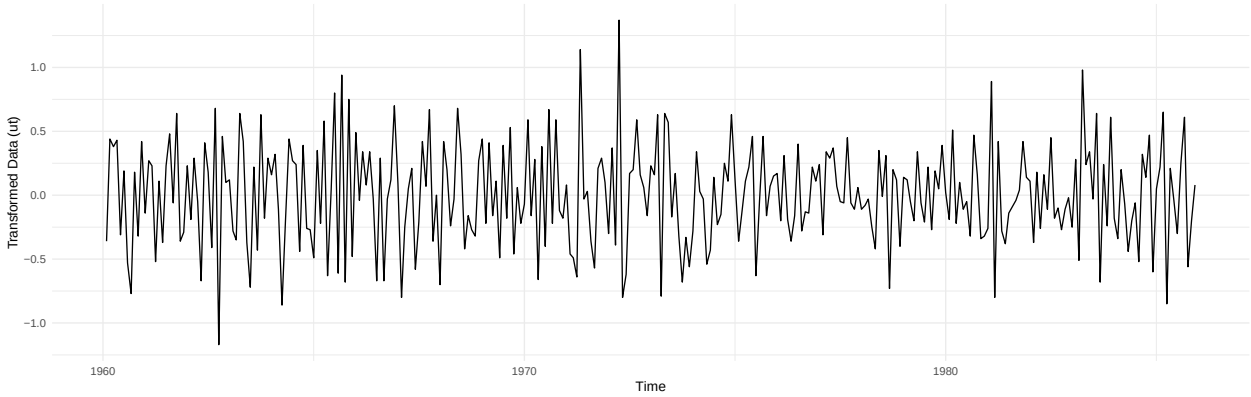


Figure 5: Time series plot of differenced and seasonally differenced data.

Our twice differenced data now appears stationary with no evidence of seasonality or polynomial trends with a stable variance in time. In Table 1 below we see that each transformation lowered the variance of our data.

Table 1: Stationarity Test Results

Transformation	**Variance**
None	86.0367259
Differenced lag 12	0.3449074
Differenced lag 12 and lag 1	0.1577327

In Figure 6 we see that the ACF of our twice differenced data shows no evidence of polynomial or seasonal trends. Further both the histogram and Q-Q plot shows that the empirical distribution of our data is symmetric and approximately Gaussian.

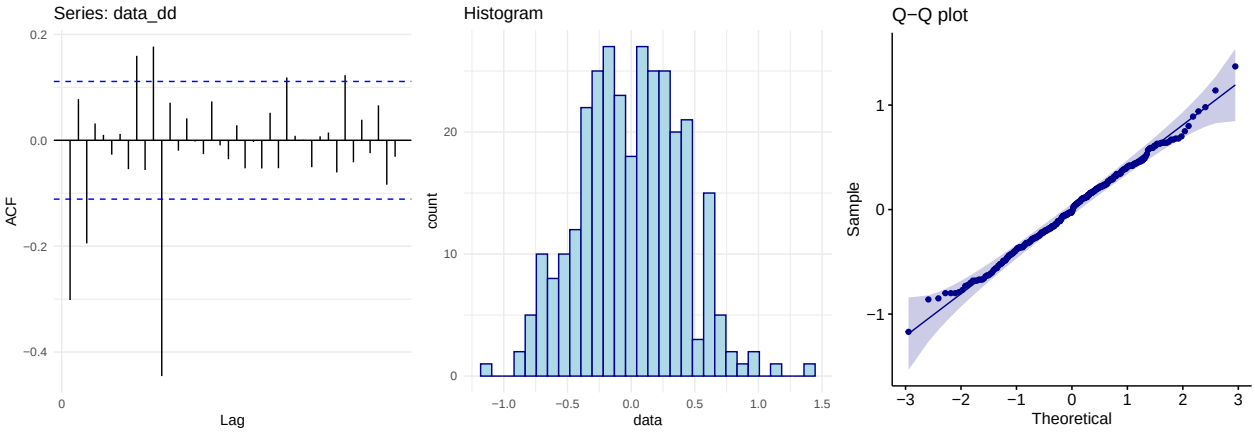


Figure 6: ACF and histogram plots of differenced and seasonally differenced data.

Finally, in Table 2 we see the p-values for both the Augmented Dickey–Fuller (ADF) t-statistic test (which examines whether the series has a unit root where a series with a trend line will have a unit root and result in a large p-value); and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (which tests the null hypothesis of trend stationarity where a low p-value will indicate a signal that is not trend stationary, has a unit root), both of which concluded that the twice differenced data is trend stationary.

Table 2: Trend stationarity test results

Test	**Null**	**pValue**	**Conclusion**
ADF Test	Process has a unit root and therefore a trend.	<0.01	Data is trend stationary
KPSS Test	Process is trend stationary	>0.1	Data is trend stationary

Model Fitting

Parameter Selection

When considering a potential SARIMA model we are required to select possible values for the 7 parameters p, d, q, P, D, Q and s.

Since we differenced our data with lag 1 once and with lag 12 once to obtain stationarity we have the parameters $d = 1$, $D = 1$ and $s = 12$. To determine a potential range of values of the remaining parameters (AR parameters p, P and MA parameters q, Q) we produce both the ACF and PACF plots of the differenced data, shown below in Figure 7.

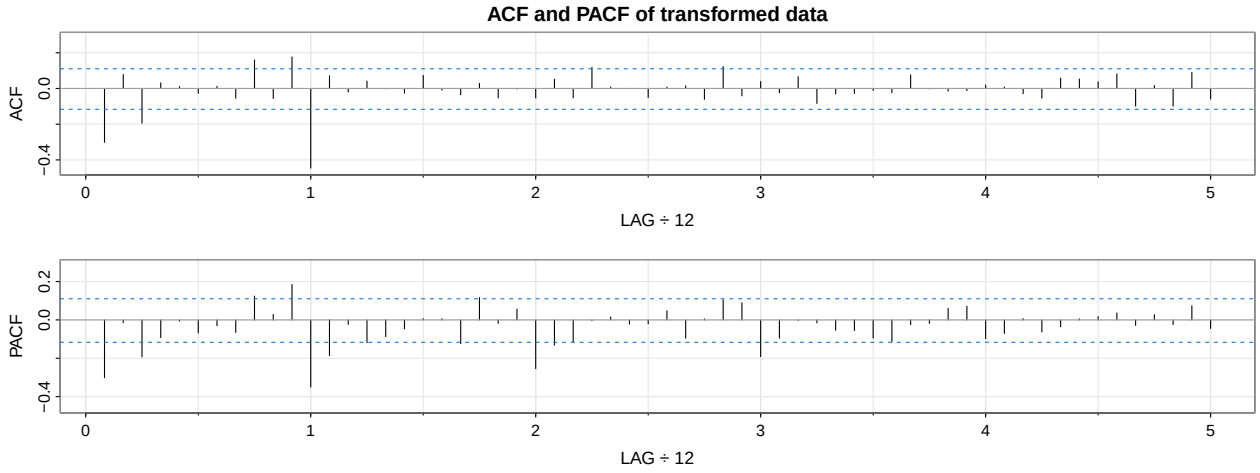


Figure 7: ACF and PACF of transformed data

To determine an upper bound for the AR term p range we look at the PACF and see that we have significant partial-autocorrelation spikes at lags 1 and 3 indicating we may require up to $p = 3$. We also note there are significant spikes at lags 9 and 11 although these are likely interactions with seasonal components to be considered later. To determine an upper bound for the MA term q range we look at the ACF and see that we have significant autocorrelation spikes at lags 1 and 3, indicating we may require up to $q = 3$. Similarly, we also see significant spikes at lags 9 and 11, likely due to interactions with seasonal components.

Now considering seasonal AR terms we again turn to the PACF, this time noting that we have significant partial-autocorrelation at lags 12,24 and 36 indicating that we may require a seasonal AR parameter up to $P = 3$. Finally, to determine an upper bound for the seasonal MA term Q range we note from the ACF plot that we have significant autocorrelation at lag 12 only, indicating the we likely require a seasonal MA parameter of $Q = 1$.

Model Selection

We proceed to select the two models using a procedure whereby we systematically fit, evaluate and compare models using the Akaike information criterion (adjusted for small sample sizes) - denoted AICc. Earlier we determined that at most our model with have up to $p = 3$, $q = 3$, $P = 3$ and $Q = 1$ however, since the AICc punishes models with large numbers of parameters our approach to selecting an optimal model will be to start with the simplest models and add parameters with the aim of minimizing the AICc.

For brevity we opt to exclude the model outputs for all models considered. A table summarizing all models considered, their parameter estimates with significance and their obtained AICc can be found in the Appendices.

Pure Moving Average Models

We begin by fixing our AR parameters $p = 0$ and $P = 0$ and determine the optimal pure Moving Average (MA) model. From our ACF and PACF plots we determined that our model should include up to $q = 3$ MA terms and $Q = 1$ seasonal MA terms. We proceed to fit various models with the results of the fitting procedure for our top performing models, shown below in Tables 3 and 4.

Model	SARIMA(0,1,3)(0,1,1)12			
	Est.	SE	z value	Pr(>|z|)
ma1	-0.3561760	0.0568682	-6.2632	3.772e-10***
ma2	0.0035662	0.0613929	0.0581	0.953678
ma3	-0.1570143	0.0550113	-2.8542	0.004314**
sma1	-0.8384782	0.0360002	-23.2909	< 2.2e-16***
sigma2	0.08458			
AIC				119.38
AICc				119.57
BIC				138.08

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3: SARIMA(0,1,3)(0,1,1)$_{12}$

Model	SARIMA(0,1,3)(0,1,1)12 (MA2=0)			
	Estimate	SE	z value	Pr(>|z|)
ma1	-0.354998	0.053140	-6.6804	2.382e-11***
ma3	-0.155974	0.052126	-2.9923	0.002769**
sma1	-0.838438	0.035989	-23.2972	< 2.2e-16***
sigma2	0.08431			
AIC				117.38
AICc				117.51
BIC				132.34

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4: SARIMA(0,1,3)(0,1,1)$_{12}$ (MA2=0)

From Table 3 we see that the SARIMA(0,1,3)(0,1,1)$_{12}$ model obtained an AICc of 119.57 and that of our 4 fitted coefficients, the MA(2) parameter is insignificant - since the confidence interval contains 0. Therefore, we fix this term to 0 and refit our model with results shown in Table 4. Our adjustment was successful with our new model achieving a lower AICc score of 117.51 and significant coefficients making this as our top performing pure MA model.

Considering Autoregressive Terms

We consider whether the introduction of autoregressive terms improves model performance per the AICc - namely whether the introduction of additional autoregressive terms improve performance enough that it outweighs the penalty from the AICc. We again consider a selection of models containing AR terms and show the fitting procedure results of our top two candidates in Tables 6 and 7 below.

Model	SARIMA(1,1,1)(0,1,1)12			
	Est.	SE	z value	Pr(>|z|)
ar1	0.344624	0.151353	2.2770	0.02279*
ma1	-0.675336	0.123186	-5.4822	4.2e-08***
sma1	-0.839105	0.035565	-23.5934	< 2e-16***
sigma2	0.08583			
AIC				123.24
AICc				123.37
BIC				138.2

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 5: SARIMA$(1,1,1)(0,1,1)_{12}$

Model	SARIMA(3,1,1)(0,1,1)12 (AR1=AR2=0)			
	Estimate	SE	z value	Pr(>|z|)
ar3	-0.138542	0.056595	-2.4480	0.01437*
ma1	-0.346842	0.055823	-6.2132	5.192e-10***
sma1	-1.196394	0.052177	-22.9296	< 2.2e-16***
sigma2	0.06065			
AIC				119.83
AICc				119.96
BIC				134.79

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6: SARIMA(3,1,1)(0,1,1)12 (AR1=AR2=0)

We found that the addition of seasonal AR terms typically decreased our models AICc and rarely had significant coefficients. From Table 5 we see that the SARIMA$(1,1,1)(0,1,1)_{12}$ model performed reasonably well, with all fitted coefficients being significant and achieving an AICc of 123.37. Furthermore, from Table 6 we also see that a potential candidate is the SARIMA$(3,1,1)(0,1,1)_{12}$ model with fixed zero AR(1) and AR(2) coefficients. This model again has all significant coefficients with an AICc of 119.96, making it the second best performing model overall per the AICc, however this model is not a viable candidate as the absolute value of the seasonal MA coefficient is greater than 1. Hence select the SARIMA$(1,1,1)(0,1,1)_{12}$ model as our second candidate and proceed to confirming stationarity and invertibility.

The equations for our selected models are therefore:

1. $\nabla_1\nabla_{12}X_t = (1 - 0.354998_{(0.053140)}B - 0.155974_{(0.052126)}B^3)(1 - 0.838438_{(0.035989)}B^{12})Z_t, \quad \hat{\sigma}_Z^2 = 0.08431$
2. $\nabla_1\nabla_{12}(1 - 0.344624_{(0.151353)}B)X_t = (1 - 0.675336_{(0.123186)}B)(1 - 0.839105_{(0.035565)}B^{12})Z_t, \quad \hat{\sigma}_Z^2 = 0.08583$

where Z_t is WN with $\mathbb{E}Z_t = 0$ and $\text{Var}(Z_t) = \sigma_Z^2$.

Model (A) is pure moving average so it is always stationary. Furthermore, from plot 1 in Figure 8 below we see that all ma roots lie outside the unit circle and since $|\Theta_1| < 1$ we have that the model is invertible.

Model (B) is invertible since $|\theta_1| < 1$ and $|\Theta_1| < 1$ and the model is stationary since $|\phi_1| < 1$.

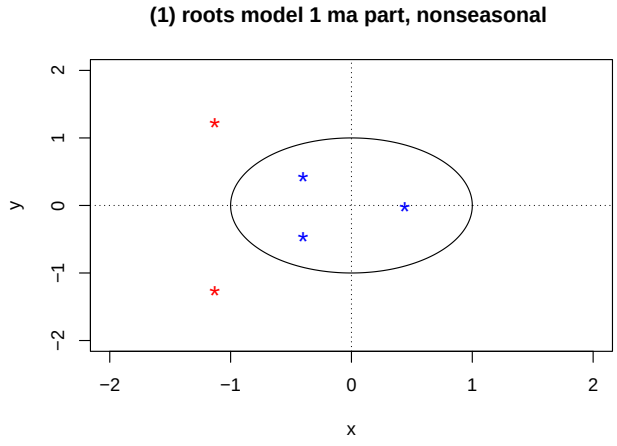


Figure 8: Model (A) Root Plot

Diagnostic Checking

Model (A)

We begin our diagnostic checking for model (A) by producing a time series plot of the residuals as well as their ACF plot, PACF plot, histogram, Q-Q plot and Ljung-Box plot which are all shown below in Figure 9.

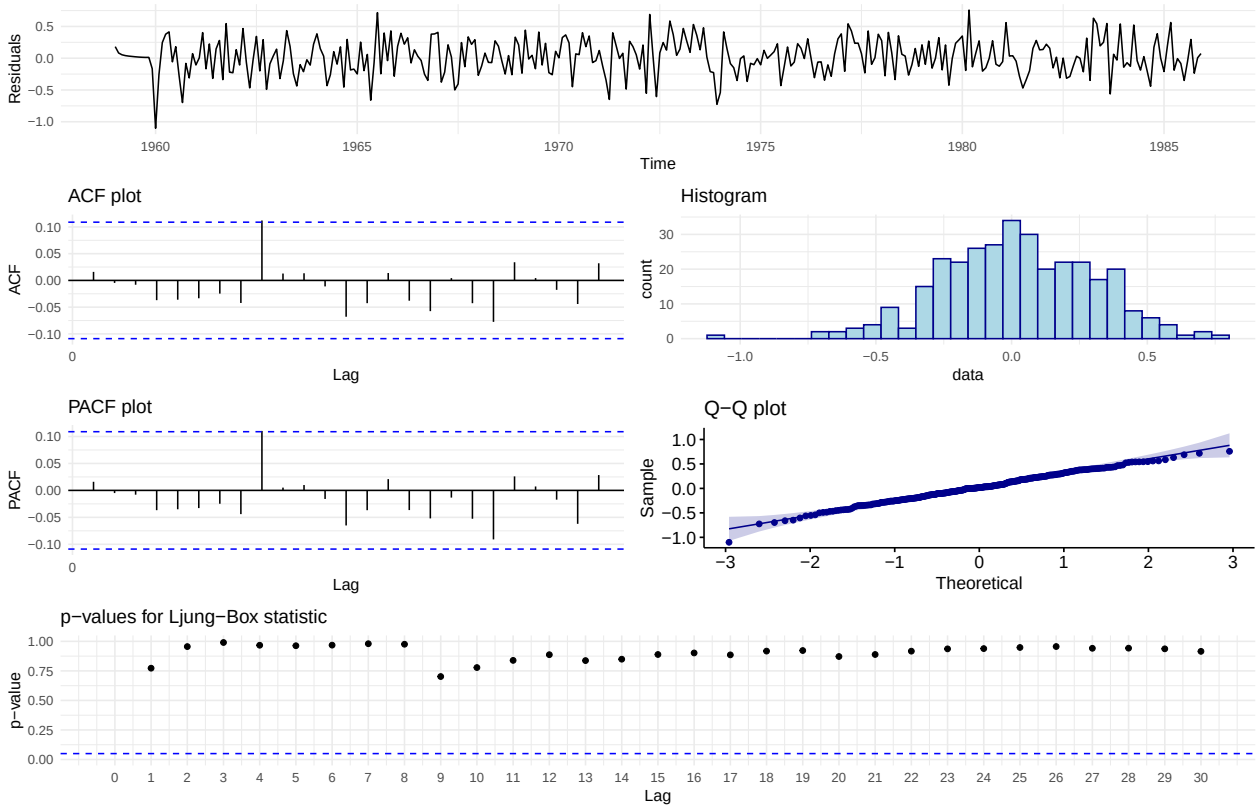


Figure 9: Model (A) Residual Analysis Plots

From the time series plot of the residuals show the characteristics of white noise with no trend, seasonality or conditional heteroscedasticity. From the ACF and PACF plots there is no evidence of significant autocorrelations or partial-autocorrelations at lags up to 25 but we do note a borderline spike in both at lag 9.

From the histogram we see that, bar a lower outlier, the empirical distribution of the residuals appears reasonably Gaussian (with a symmetric, unimodal distribution with non heavy tails), an observation which is further supported with our Q-Q plot which shows that, again bar the lower outlier, the points all lie closely to the diagonal line indicating that the distribution is close to Gaussian.

Finally, from the Ljung-Box statistic plot we see no significant p-values for all lags up to 30 indicating that there is no significant autocorrelations present in the residuals. We conduct a selection of statistical tests on the residuals, the results of which are shown below in Table 7.

Table 7: Model (A) - Residual Test Results

Test	p.value	Conclusion
Shapiro-Wilk	0.3162723	Normal
Box-Pierce	0.8187969	Independence
Ljung-Box	0.7928839	No autocorrelation in residuals
Ljung-Box (squared)	0.7703434	No autocorrelation in squared residuals

For each test we obtained an insignificant p-value hence our model passes all four tests.

Model (B)

We repeat our diagnostic checks for Model (B) with the same collection of residual analysis plots shown below Figure 10.

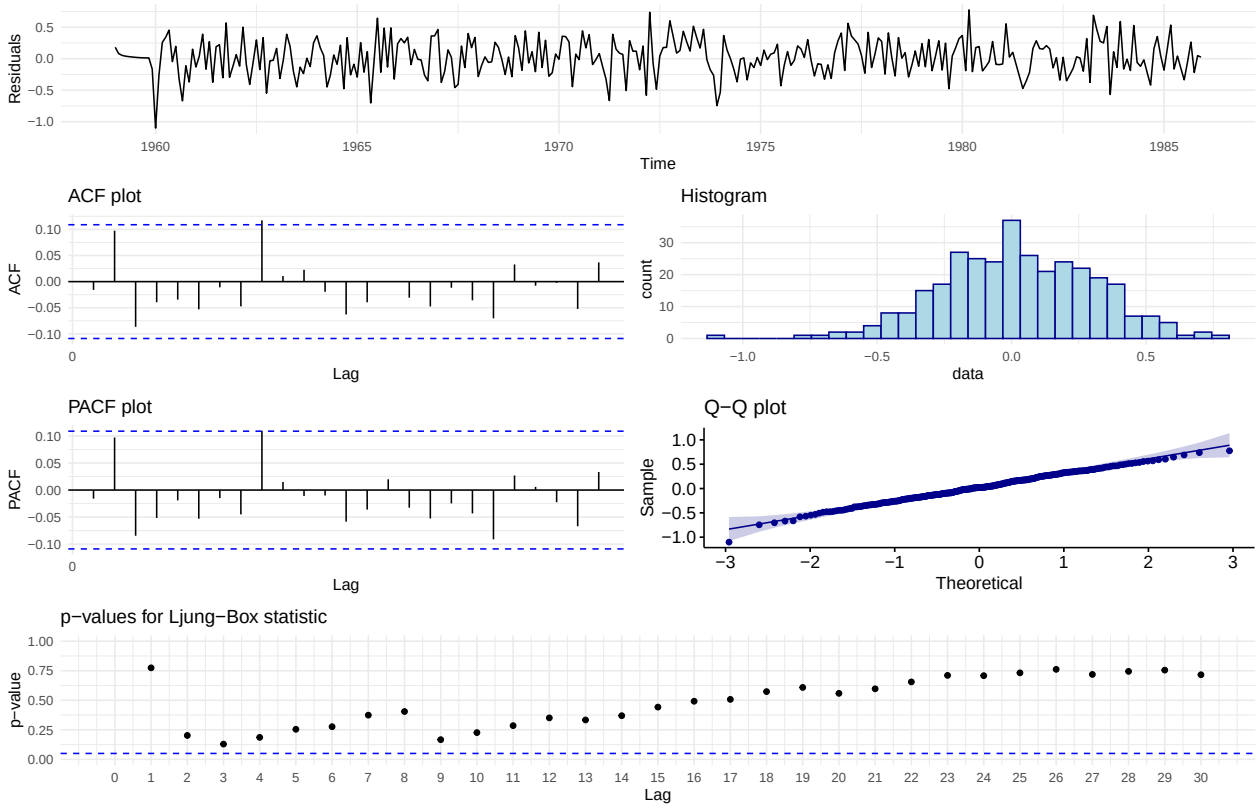


Figure 10: Model (B) - Residual Analysis Plots

From the time series plot of the residuals show the characteristics of white noise with no trend, seasonality or conditional heteroscedasticity. From the ACF and PACF plots there is no evidence of significant autocorrelations or partial-autocorrelations at lags up to 25 but we do note a borderline spike in both at lags 2 and 9.

From the histogram we see that, bar a lower outlier, the empirical distribution of the residuals appears reasonably Gaussian (with a symmetric, unimodal distribution with non heavy tails), an observation which is further supported with our Q-Q plot which shows that, again bar the lower outlier, the points all lie closely to the diagonal line indicating that the distribution is close to Gaussian. We do note that the distribution is perhaps a little heavier tailed that that for Model (A).

Finally, from the Ljung-Box statistic plot we see no significant p-values for all lags up to 30 indicating that there is no significant autocorrelations present in the residuals. Again, we note that although no p-values were significant they were much closer to the rejection region than Model (A). The equivalent test results are shown below in Table 8.

Table 8: Model (B) - Residual Test Results

Test	**p.value**	**Conclusion**
Shapiro-Wilk	0.5001295	Normal
Box-Pierce	0.3947198	Independence
Ljung-Box	0.3639211	No autocorrelation in residuals
Ljung-Box (squared)	0.7121532	No autocorrelation in squared residuals

For each test we obtained an insignificant p-value hence our model passes all four tests.

Both models pass diagnostic checking however based on the slight edge seen in the residual analysis of Model (A) as well as the lower AICc score we select Model (A) as our final model and proceed to forecasting. We note that our chosen model is simpler than our initial speculations from observing the ACF and PACF with the AICc selection process ensuring that our model complexity was justified by significant increased fit to maintain parsimony.

Forecasting

Our final model therefore is Model A which is a SARIMA(0,1,3)(0,1,1)$_{12}$ model with both the MA2 and MA3 coefficients set to 0 with model equation:

$$\nabla_1\nabla_{12}X_t = (1 - 0.354998_{(0.053140)}B - 0.155974_{(0.052126)}B^3)(1 - 0.838438_{(0.035989)}B^{12})Z_t, \quad \hat{\sigma}_Z^2 = 0.08431$$

Using our chosen model we produce a 5 year (60 month) ahead forecast, shown below in Figure 11.

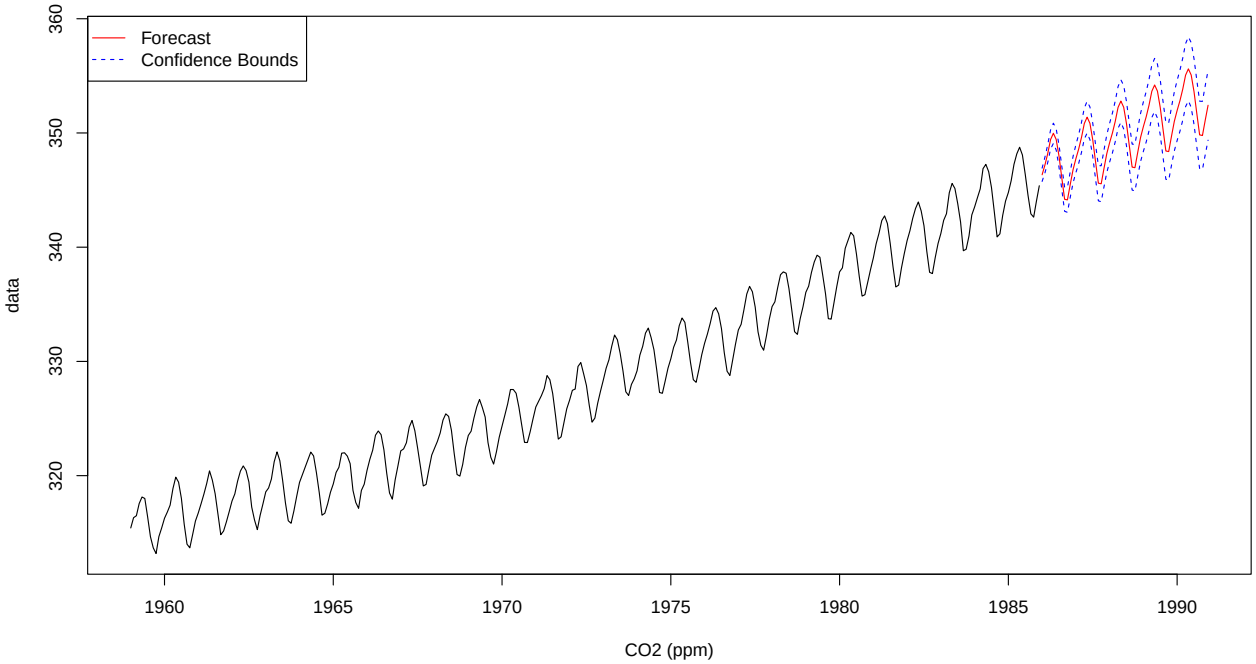


Figure 11: 60 month forecast of atmospheric carbon dioxide above Mauna Loa, Hawaii.

We produce a plot of our data since 1980 in order to focus on our forecasted values, shown below in Figure 12. Furthermore, we evaluate the performance of our fitted model by overlaying the true observed values of the series in our testing data set.

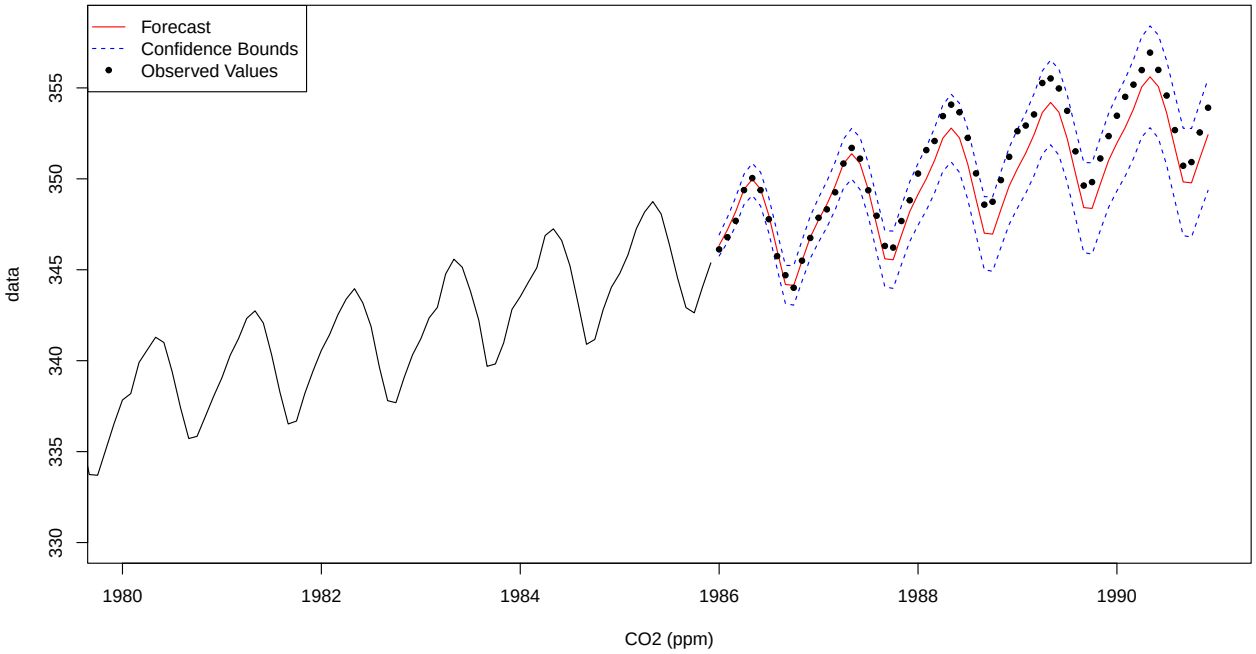


Figure 12: 60 month forecast pf atmospheric carbon dioxide above Mauna Loa, Hawaii.

From Figure 12 we see that the forecast produced by our model was successful. Throughout the five year period our observed values were well predicted by our forecast, lying within our confidence region throughout. We note a period from 1988 to 1990 where our model looked to be lagging below the observed values which began to approach the upper bound of our confidence region before returning back to lying close to our forecasted values by the end of 1990. Overall we conclude that despite the large (and somewhat impractical) forecasting window, our models forecast was accurate.

Spectral Analysis

We conduct spectral analysis on our chosen model by first producing two periodograms, one constructed using our un-differenced data and one constructed from the residuals of our final model, shown below

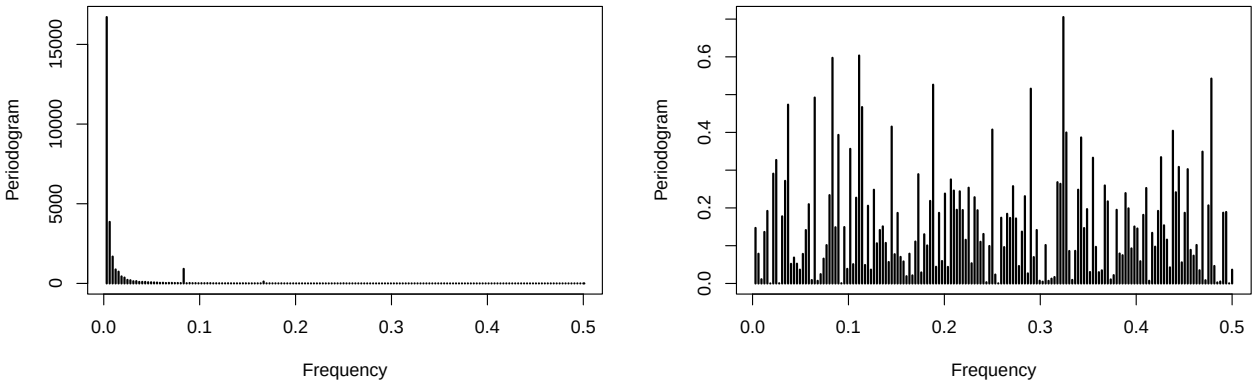


Figure 13: Periodograms for data and residuals

From plot 1 in Figure 13 we see that there is a spike around $0.08 \approx 1/12$ indicating a period of 12 months, an observation we detected initially with our seasonality plot. From plot 2 we see that there is no dominated frequency indicating that there is no significant seasonality remaining in our residuals.

Next we perform Fisher's test for periodicity detection on our final model residuals in order to test the data for the presence of hidden periodicities with unspecified frequency. The test results in a p-value of $0.8882769 > 0.05$, hence our model residuals passes Fisher's test for Gaussian WN and we conclude that there is insufficient evidence to suggest the model residuals are not Gaussian WN.

Another test for Gaussian WN is the Kolmogorov-Smirnov test for which we can construct the cumulative periodogram for the residuals of our final model, shown below in Figure 14. We see that our model has passed the Kolmogorov-Smirnov test since the values of the cumulative periodogram lie within the interval for all frequencies hence we conclude that our evidence suggests that our final model residuals are Gaussian WN.

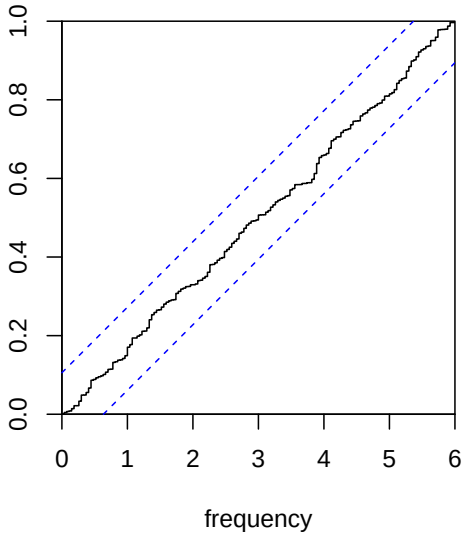


Figure 14: Cumulative Periodogram for Final Model Residuals

Conclusion

For this project we successfully achieved our goal of determining and fitting a model that provides accurate forecasts for our chosen time series data. We considered a data sample of atmospheric CO2 levels above Mauna Loa, Hawaii from January 1959 to December 1990. We conducted exploratory analysis and detected both a polynomial (linear) trend and a seasonal trend which we accounted for with differencing. We fitted various models from the SARIMA framework, selecting our top performing model first by using the AICc performance criterion and secondly by an in-depth residual analysis. Our final model equation is given as:

$$\nabla_1 \nabla_{12} X_t = (1 - 0.354998_{(0.053140)} B - 0.155974_{(0.052126)} B^3)(1 - 0.838438_{(0.035989)} B^{12}) Z_t, \quad \hat{\sigma}_Z^2 = 0.08431$$

Using our chosen model we produced a 5 year (60 month) ahead forecast which we compared to observed true future values and saw evidence that our model performs well for prediction with all future observations lying in our models confidence interval. We concluded by performing spectral decomposition, the results of which further support our selection of our final model with our evidence suggesting that our model produced residuals that were Gaussian WN and captured all significant seasonality present in the data.

References

1. *Introduction to Time Series and Forecasting*, by P. Brockwell and R. Davis, Springer (1996).
2. *Time Series Analysis with R Examples*, by R. H. Shumway and D. S. Stoffer, Springer (2011).
3. *UC Santa Barbara PSTAT274 Lecture Notes*, by Raya Feldman (2023).

Appendices

Summary of Fitted Models

Model	ar1	ar2	ar3	ma1	ma2	ma3	sar1	sma1	AICc
Pure MA Models									
SARIMA(0,1,1)(0,1,0)12				-0.344551***					279.55
SARIMA(0,1,0)(0,1,1)12								-0.889574***	154.53
SARIMA(0,1,2)(0,1,0)12				-0.345986***	-0.046864				280.87
SARIMA(0,1,1)(0,1,1)12				-0.353020***				-0.841135***	123.84
SARIMA(0,1,3)(0,1,0)12				-0.325178***	0.013010	-0.216530***			270.05
SARIMA(0,1,2)(0,1,1)12				-0.353873***	-0.046164			-0.838575***	125.30
SARIMA(0,1,3)(0,1,1)12				-0.356176***	0.0035662	-0.157014**		-0.838478***	119.57
SARIMA(0,1,3)(0,1,0)12 (MA2=0)				-0.321529***	0	-0.213898***			268.05
SARIMA(0,1,3)(0,1,1)12 (MA2=0)				-0.354998***	0	-0.155974**		-0.838438***	117.51
ARMA Models									
SARIMA(1,1,1)(0,1,1)12	0.344624*			-0.675336***				-0.839105***	123.37
SARIMA(0,1,1)(1,1,1)12				-0.356868***			-0.031260	-0.831075***	125.69
SARIMA(1,1,3)(0,1,1)12 (MA2=0)	0.025502			-0.376771**	0	-0.152454**		-0.838529***	119.55
SARIMA(3,1,3)(0,1,1)12 (MA2=0)	0.660242	0.234948	-0.269463	-1.007236*	0	0.193135		-0.846110***	123.1
SARIMA(3,1,1)(0,1,1)12	0.370575*	0.139399.	-0.103047	-0.722425***				-0.841411***	121.15
SARIMA(3,1,1)(0,1,1)12 (AR1=AR2=0)	0	0	-0.138542*	-0.346842***				-1.196394***	119.96

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 9: Model Comparison Summary

R Code

```
# Information -------

# Title: Time Series Project Analysis
# Author: John Inston
# Date: May 30th, 2023
# Course: PSTAT274
# Supervisor: Raya Feldman

# Setup ---------------

# load libraries
library(tsdl) # source for data
library(zoo) # rollmean
library(ggplot2) # plots and figures
library(tidyr) # updaed r syntax
library(gridExtra) # multiple ggplots
library(forecast) # seasonal plot
library(tseries) # adf test function
library(kableExtra) # tables
library(astsa) # acf2,
library(LSTS) #
library(MASS) # box-cox (potentially remove)
library(TSA)
library(ggpubr) # qqplot
library(lmtest)

# time series decomposition
#devtools:: install_github("brisneve/ggplottimeseries")
library(ggplottimeseries)

# set plot information
theme_set(theme_minimal())

# set seed
set.seed(8467)

# define plot roots function
plot.roots <- function(ar.roots=NULL,
                       ma.roots=NULL,
                       size=2, angles=FALSE,
                       special=NULL,
                       sqecial=NULL,
                       my.pch=1,
                       first.col="blue",
                       second.col="red",
                       main=NULL)
{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="dotted"))
  }
}
if(!is.null(special))
{
  lines(Re(special),Im(special),lwd=2)
}
if(!is.null(sqecial))
{
  lines(Re(sqecial),Im(sqecial),lwd=2)
}
}

# load data
tsdl_met <- subset(tsdl, 12, "Meteorology")
rawdata <- tsdl_met[[2]]

# select training data
data <- head(rawdata, n=length(rawdata)-(12*5))

# Exploratory Analysis ----------

## Undifferenced Data -----

## plot of training data (with trend line)
# trend line
maline <- rollmean(data, k=12, fill=NA, align='right')
# produce time series plot
plotdata <- data.frame(
  "Time" = seq(as.Date("1959/1/1"), as.Date("1985/12/1"), "months"),
  "data" = data,
  "trend" = maline
```

```
)

plotdata %>%
  tidyr::gather("id", "value", 2:3) %>%
  ggplot(., aes(Time, value, col=id, linetype=id))+
  geom_line() +
  theme_light() +
  labs(x = "Time", y = "CO2 (ppm)", color="Legend", linetype="Legend")

## time series decomposition
ts_decomp <- ggdecompose(dts2(data, type ="additive"))+
  xlab("Date")+
  ylab("Atmospheric Concentration of CO2")+
  ggtitle("Time Series Decomposition Plots")
# seasonal plot
seasplot <- ggseasonplot(data, year.labels=TRUE, year.labels.left=TRUE) +
  ylab("CO2 (ppm)") +
  ggtitle("Seasonal plot") +
  theme_minimal()
trendplot <- autoplot(maline, ts.color="red", main="Trend plot") +
  theme_minimal()
grid.arrange(ts_decomp, seasplot, trendplot,
             layout_matrix = rbind(c(1, 1),
                                   c(2, 3)))

## acf of undifferenced data
acf(data, 40, main="ACF and PACF of raw data.")

## conclusion
# - non-stationary, remove trend by differncing with lag 12

## Diff Lag 12 Data -------

# differencing data
data_d <- diff(data, 12)

# time series plot of differenced data
plot_data_d <- data.frame(
  "Time" = seq(as.Date("1960/1/1"), as.Date("1985/12/1"), "months"),
  "yt" = data_d
)
tsplot_data_d <- plot_data_d %>%
  tidyr::gather("id", "value", 2:2) %>%
  ggplot(., aes(Time, value))+
  geom_line() +
  theme_minimal() +
  labs(x = "Time", y = "Seasonally Differenced Monthly CO2") +
  geom_smooth(method=lm) #add linear trend line

# acf plot of differenced data
acfplot_data_d <- autoplot(acf(data_d, lag=40, plot = FALSE))

grid.arrange(tsplot_data_d,acfplot_data_d, ncol=2)
```

```r
## Diff Lag 12 & 1 Data ---------

# difference again with lag 1
data_dd <- diff(data_d, 1)

# plot of transformed data
# time series plot of differenced data
plot_data_dd <- data.frame(
  "Time" = seq(as.Date("1960/2/1"), as.Date("1985/12/1"), "months"),
  "yt" = data_dd
)
plot_data_dd %>%
  tidyr::gather("id", "value", 2:2) %>%
  ggplot(., aes(Time, value))+
  geom_line() +
  theme_minimal() +
  labs(x = "Time", y = "Transformed Data (zt)")

# table of variance of differenced data
var_table <- data.frame(
  "Transformation" = c("None", "Differenced lag 12", "Differenced lag 12 and lag 1"),
  "Variance" = c(var(data), var(data_d), var(data_dd))
)

# acf of differenced data
acf_data_dd <- autoplot(acf(data_dd, lag=40))

# histogram of differenced data
hist_data_dd <- ggplot(data.frame("data"=data_dd), aes(x=data)) +
  geom_histogram(color="darkblue", fill="lightblue") +
  ggtitle("Histogram")

# q-q plot of differenced data
qqplot_data_dd <- ggqqplot(data_dd, color="darkblue") + ggtitle("Q-Q plot")

grid.arrange(acf_data_dd, hist_data_dd, qqplot_data_dd,ncol=3)

# stationarity tests (adf and kpss)
adf_test <- adf.test(data_dd)
kpss_test <- kpss.test(data_dd, null="Trend")
table_data <- data.frame(
  Test = c("ADF Test", "KPSS Test"),
  Null = c("Process has a unit root and therefore a trend.",
           "Process is stationary"),
  pValue = c("<0.01", ">0.1"),
  Conclusion = c("Data is stationary", "Data is stationary")
)

# data appears stationary, proceed to model fitting

# Model Fitting ----------

# acf & pacf for parameter range selection
```

```r
acf2(data_dd, 60, main="ACF and PACF of transformed data")

## SIMA Model Fitting -----------

# model 1 - sarima(0,1,1)(0,1,0)12
model1 <- Arima(data,
                order=c(0,1,1),
                seasonal=list(order=c(0,1,0), period=12),
                method="ML")
model1
coeftest(model1)
# all coeffs significant
# AICc 279.55

# model 2 - sarima(0,1,0)(0,1,1)12
model2 <- Arima(data,
                order=c(0,1,0),
                seasonal=list(order=c(0,1,1), period=12),
                method="ML")
model2
coeftest(model2)
# all coeffs significant
# AICc 154.53

# model 3 - sarima(0,1,2)(0,1,0)12
model3 <- Arima(data,
                order=c(0,1,2),
                seasonal=list(order=c(0,1,0), period=12),
                method="ML")
model3
coeftest(model3)
# MA2 insignificant
# AICc 280.87

# model 4 - sarima(0,1,1)(0,1,1)12
model4 <- Arima(data,
                order=c(0,1,1),
                seasonal=list(order=c(0,1,1), period=12),
                method="ML")
model4
coeftest(model4)
# all coeffs signficant
# AICc 123.84 <- pretty bloody good!

# model 5 - sarima(0,1,3)(0,1,0)12
model5 <- Arima(data,
                order=c(0,1,3),
                seasonal=list(order=c(0,1,0), period=12),
                method="ML")
model5
coeftest(model5)
# MA2 insignificant
# AICc 270.05
```

```
# model 6 - sarima(0,1,2)(0,1,1)12
model6 <- Arima(data,
                order=c(0,1,2),
                seasonal=list(order=c(0,1,1), period=12),
                method="ML")
model6
coeftest(model6)
# MA2 insignificant
# AICc 125.3

# model 7 - sarima(0,1,3)(0,1,1)12
model7 <- Arima(data,
                order=c(0,1,3),
                seasonal=list(order=c(0,1,1), period=12),
                method="ML")
model7
coeftest(model7)
# MA2 insignificant
# AICc 119.57 <- second best SIMA model

# model 8 - sarima(0,1,3)(0,1,0)12 (MA2 removed)
model8 <- Arima(data,
                order=c(0,1,3),
                seasonal=list(order=c(0,1,0), period=12),
                method="ML",
                fixed=c(NA,0,NA))
model8
coeftest(model8)
# all coeffs signficant
# AICc 268.05

# model 9 - sarima(0,1,3)(0,1,1)12 (MA2 removed)
model9 <- Arima(data,
                order=c(0,1,3),
                seasonal=list(order=c(0,1,1), period=12),
                method="ML",
                fixed=c(NA,0,NA,NA))
model9
coeftest(model9)
# all coeffs signficant
# AICc 117.51 <- top SIMA model

## select final model 1
finalmodel1 <- model9

## AR Parameters -----

# model 10 - sarima(1,1,1)(0,1,1)12
model10 <- Arima(data,
                order=c(1,1,1),
                seasonal=list(order=c(0,1,1), period=12),
                method="ML")
model10
```

```
coeftest(model10)
# all coeffs signficant
# AICc 123.37 <- decent!

# model 11 - sarima(0,1,1)(1,1,1)12
model11 <- Arima(data,
                 order=c(0,1,1),
                 seasonal=list(order=c(1,1,1), period=12),
                 method="ML")
model11
coeftest(model11)
# SAR1 insignificant
# AICc 125.69

# model 12 - sarima(1,1,3)(0,1,1)12 MA2=0
model12 <- Arima(data,
                 order=c(1,1,3),
                 seasonal=list(order=c(0,1,1), period=12),
                 method="ML",
                 fixed=c(NA,NA,0,NA,NA))
model12
coeftest(model12)
# AR1 insignificant
# AICc 119.55

# model 13 - sarima(3,1,3)(0,1,1)12 MA2=0
model13 <- Arima(data,
                 order=c(3,1,3),
                 seasonal=list(order=c(0,1,1), period=12),
                 method="ML",
                 fixed=c(NA,NA,NA,NA,0,NA,NA))
model13
coeftest(model13)
# numerous insignificant coefficients
# AICc 123.1

# model 14 - sarima(3,1,1)(0,1,1)12 MA2
model14 <- Arima(data,
                 order=c(3,1,1),
                 seasonal=list(order=c(0,1,1), period=12),
                 method="ML",
                 fixed=c(NA,NA,NA,NA,NA))
model14
coeftest(model14)
# insignificant AR2 and AR3
# AICc 121.15

# model 15 - sarima(3,1,1)(0,1,1)12 AR1=AR2=0
model15 <- Arima(data,
                 order=c(3,1,1),
                 seasonal=list(order=c(0,1,1), period=12),
                 method="ML",
                 fixed=c(0,0,NA,NA,NA))
```

```
model15
coeftest(model15)
# all coeffs significant
# AICc 119.96

finalmodel2 <- model10

## final models
finalmodel1
coeftest(finalmodel1)
finalmodel2
coeftest(finalmodel2)

# Diagnostic Checking --------

## Model 1 ----

# residual analysis plots
resplot_model1 <- autoplot(finalmodel1$residuals, ylab="Residuals")
acf_model1_res <- autoplot(acf(finalmodel1$residuals)) + ggtitle("ACF plot")
pacf_model1_res <- autoplot(pacf(finalmodel1$residuals)) + ggtitle("PACF plot")
hist_model1_res <- ggplot(data.frame("data"=finalmodel1$residuals), aes(x=data)) +
  geom_histogram(color="darkblue", fill="lightblue") +
  ggtitle("Histogram")

qqplot_model1_res <- ggqqplot(finalmodel1$residuals, color="darkblue") +
  ggtitle("Q-Q plot")

model1_ljung <- Box.Ljung.Test(finalmodel1$residuals, lag=30)

grid.arrange(resplot_model1,
             acf_model1_res,hist_model1_res,
             pacf_model1_res,
             qqplot_model1_res,
             model1_ljung,
             layout_matrix = rbind(c(1, 1),
                                   c(2, 3),
                                   c(4, 5),
                                   c(6, 6)))

# portmanteau tests
model1_sw <- shapiro.test(finalmodel1$residuals)
model1_bp <- Box.test(finalmodel1$residuals,
                      lag =18, type = c("Box-Pierce"), fitdf = 3)
model1_lb <- Box.test(finalmodel1$residuals,
                      lag = 18, type = c("Ljung-Box"), fitdf = 3)
model1_lbs <- Box.test((finalmodel1$residuals)^2,
                       lag = 18, type = c("Ljung-Box"), fitdf = 0)

model1_tests <- data.frame("Test"=c("Shapiro-Wilk",
                                    "Box-Pierce",
                                    "Ljung-Box",
                                    "Ljung-Box (squared)"),
```

```
                          "p-value"=c(model1_sw$p.value,
                                      model1_bp$p.value,
                                      model1_lb$p.value,
                                      model1_lbs$p.value),
                          "Conclusion"=c("Normal",
                                         "Independence",
                                         "No autocorrelation in residuals",
                                         "No autocorrelation in squared residuals"))
model1_tests

## explore fitting ar models
#ar(model2$residuals, aic = TRUE, order.max = NULL, method = c("yule-walker"))

## Model 2 ------

# residual analysis plots
resplot_model2 <- autoplot(finalmodel2$residuals, ylab="Residuals")
acf_model2_res <- autoplot(acf(finalmodel2$residuals)) +
  ggtitle("ACF plot")
pacf_model2_res <- autoplot(pacf(finalmodel2$residuals)) +
  ggtitle("PACF plot")
hist_model2_res <- ggplot(data.frame("data"=finalmodel2$residuals), aes(x=data)) +
  geom_histogram(color="darkblue", fill="lightblue") +
  ggtitle("Histogram")

qqplot_model2_res <- ggqqplot(finalmodel2$residuals, color="darkblue") +
  ggtitle("Q-Q plot")

model2_ljung <- Box.Ljung.Test(finalmodel2$residuals, lag=30)

grid.arrange(resplot_model2,
             acf_model2_res,
             hist_model2_res,
             pacf_model2_res,
             qqplot_model2_res,
             model2_ljung,
             layout_matrix = rbind(c(1, 1),
                                   c(2, 3),
                                   c(4, 5),
                                   c(6, 6)))

# portmanteau tests
model2_sw <- shapiro.test(finalmodel2$residuals)
model2_bp <- Box.test(finalmodel2$residuals,
                      lag =18,
                      type = c("Box-Pierce"),
                      fitdf = 3)
model2_lb <- Box.test(finalmodel2$residuals,
                      lag = 18,
                      type = c("Ljung-Box"),
                      fitdf = 3)
model2_lbs <- Box.test((finalmodel2$residuals)^2,
                       lag = 18, type = c("Ljung-Box"), fitdf = 0)
```

```r
model2_tests <- data.frame("Test"=c("Shapiro-Wilk",
                                    "Box-Pierce",
                                    "Ljung-Box",
                                    "Ljung-Box (squared)"),
                           "p-value"=c(model2_sw$p.value,
                                       model2_bp$p.value,
                                       model2_lb$p.value,
                                       model2_lbs$p.value),
                           "Conclusion"=c("Not Normal",
                                          "Independence",
                                          "No autocorrelation in residuals",
                                          "No autocorrelation in squared residuals"))
model2_tests

# root plots
plot.roots(NULL,polyroot(c(1, 0.354998, 0, -0.155974)),
           main="(1) roots model 1 ma part, nonseasonal ")
plot.roots(polyroot(c(1, 0, 0, 0.138542)),NULL,
           main="(2) roots of model 2 ar part, nonseasonal")

# Forecasting --------

# forecast plot
pred.tr <- predict(finalmodel1, n.ahead = 60)
U.tr= pred.tr$pred + 2*pred.tr$se # upper bound of prediction interval
L.tr= pred.tr$pred - 2*pred.tr$se # lower bound
plot(data,xlim=c(time(data)[1],(time(rawdata)[length(rawdata)])),
     ylim = c(min(data),max(U.tr)), xlab="CO2 (ppm)")
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
lines(seq(time(rawdata)[length(rawdata)-59],
          time(rawdata)[length(rawdata)], by=(1/12)),
      pred.tr$pred, col="red")

# focussed forecast plot
plot(data,
     xlim=c(time(data)[length(rawdata)-130],(time(rawdata)[length(rawdata)])),
     ylim = c(330,max(U.tr)),
     xlab="CO2 (ppm)")
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
lines(seq(time(rawdata)[length(rawdata)-59], time(rawdata)[length(rawdata)], by=(1/12)),
      pred.tr$pred,
      col="red")
points(seq(time(rawdata)[length(rawdata)-59],
           time(rawdata)[length(rawdata)], by=(1/12)),
       rawdata[(length(rawdata)-59):length(rawdata)], pch=20)
legend("topleft", legend=c("Forecast", "Confidence Bounds", "Observed Values"),
       col=c("red", "blue", "black"), lty=c(1,2,NA), pch=c(NA,NA,20))

# Spectral Analysis ------

# periodogrms
```

```r
unloadNamespace("GeneCycle")
par(mfrow=c(1,2))
periodogram(data) #, main="Plot 1 - Periodogram from data")
periodogram(finalmodel1$residuals) #, main="Plot 2 - Periodogram from residuals")
par(mfrow=c(1,1))

# fisher g test
library(GeneCycle)
fisher.g.test(model2$residuals)

# cumulative periodogram
cpgram(finalmodel1$residuals, main="")
```