Forecasting Atmospheric Carbon Dioxide Concentration with SARIMA Models

Chris Meade

2/23/2017

Contents

1. Abstract 2

2. Introduction 2

    2.1 About the Data 2

    2.2 Software Used in Analysis 2

    2.3 Results 2

    2.4 Summary of Analysis 2

3. Analysis 3

    3.1 Exploratory Data Analysis 3

    3.2 Data Transformations 3

    3.3 ACF and PACF Analysis 7

    3.4 Model Diagnostics 8

    3.5 Forecasting 16

4. Conclusion 17

5. References 17

Appendix 18

1. Abstract

As a greenhouse gas, carbon dioxide (CO_2) is one of the driving forces behind global warming. Since the beginning of the industrial revolution, atmospheric CO_2 levels have risen more than 40% and give no idication of slowing down. This project utilizes seasonal autoregressive integrated moving average (SARIMA) models to accurately forecast atmospheric carbon dioxide concentrations 10 months into the future. It is our hope that these forecasts may be useful for anticipating other meteorological phenomena, such as catastrophic weather events and global temperature rises.

2. Introduction

Global warming is perhaps the greatest problem facing future generations. One of largest contributing factors to global warming is atmospheric carbon dioxide. Since the effects of global warming are widespread and catastrophic, it is important to accurately predict how atmospheric CO_2 levels will change into the future. The goal of this paper is to forecast monthly atmospheric carbon dioxide concentration by utilizing the Box-Jenkins methodology to fit an appropriate SARIMA model. Our efforts were successful and our results indicate that atmospheric CO_2 concentrations can be accurately forecasted 10 months into the future.

2.1 About the Data

The models in this project were trained from monthly mean atmospheric carbon dioxide concentrations measured in parts per million. The data were collected at the Mauna Loa Observatory in Hawaii, beginning in March 1958. These data are publicly available and freely distributed by the National Oceanic & Atmospheric Administration in accordance to their Global Greenhouse Gas Reference Network. The data table used in the project is available at the NOAA website.

2.2 Software Used in Analysis

All statistical analysis performed in this project was done with the `RStudio` integrated development environment. In addition to base `R`, the following software libraries were used: `MASS`, `stats`, `tseries`, and `forecast`. Please note that `forecast::Arima()` was used instead of `stats::arima()`. According to the creator of the `forecast` package, Rob J Hyndman, `Arima()` provides for more robust framework to forecast differenced data than that provided by `stats::arima()`.

2.3 Results

Using SARIMA models, we conclude that atmospheric CO_2 concentrations can be accurately forecasted.

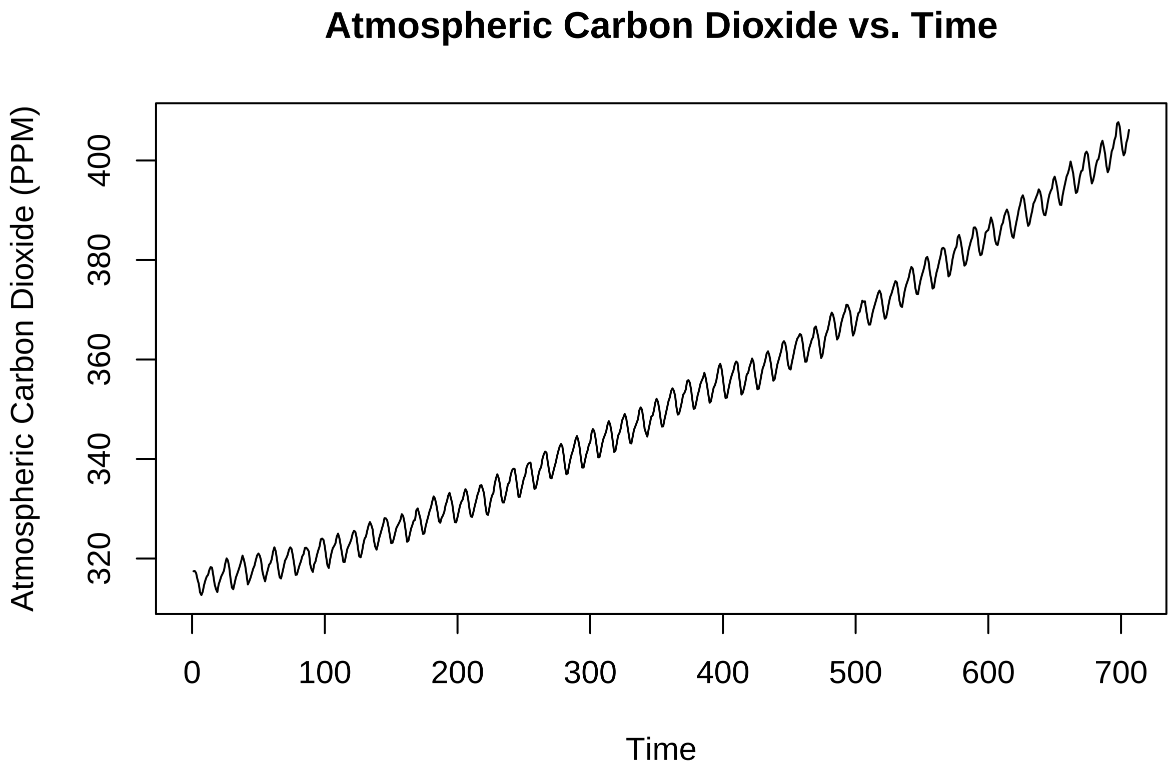
2.4 Summary of Analysis

We begin the project by loading our data in `R`. We first remove the last 10 observations for comparison against our forecasts. Next, we begin conducting exploratory analysis of the data and discover that it requires a transformation to statilize variance and differencing to remove trend and seasonality. We perform this transformation and differencing, yielding a stationary time series. Next, we examine the autocorrelations (ACF) and partial autocorrelations (PACF) of the stationary series to identify appropriate models. After constructing two potential models, we conduct model diagnostic checking to make sure each meets the assumptions of SARIMA. Finally, we choose a 'best' model and use it to forecast ten future observations. These forecasts are compared against the 10 observations we reserved in the beginning in order to evaluate model performance.

3. Analysis

3.1 Exploratory Data Analysis

We begin by loading the raw data into R and plotting the time series.

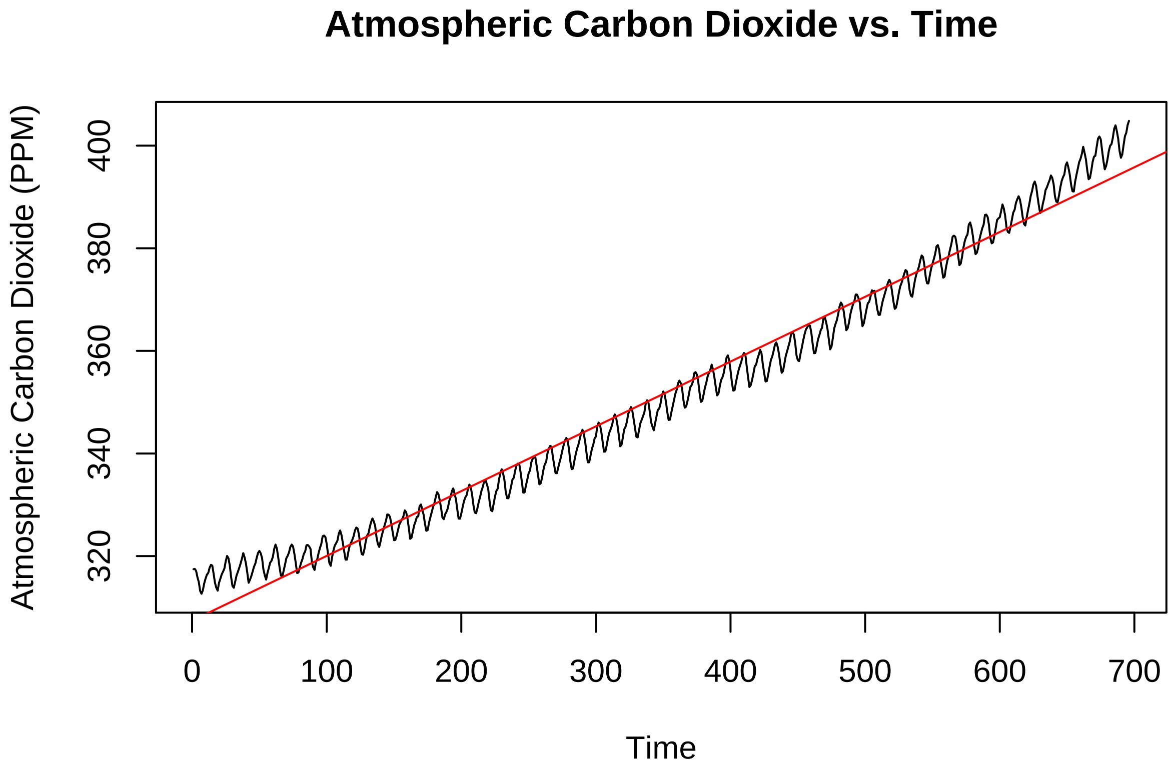


We make a few observations from this plot. First, there is a strong positive trend. The data is also yearly seasonal. Finally, variance slightly increases over time. As a result, we must transform the series to make it stationary.

We also remove the last 10 observations so that we may compare them against the forecasted data.

3.2 Data Transformations

To demonstrate the existence of a positive trend, we construct a linear model of the form $\hat{Y} = \beta_0 + \beta_1 X_t$ and impose it onto the time series.

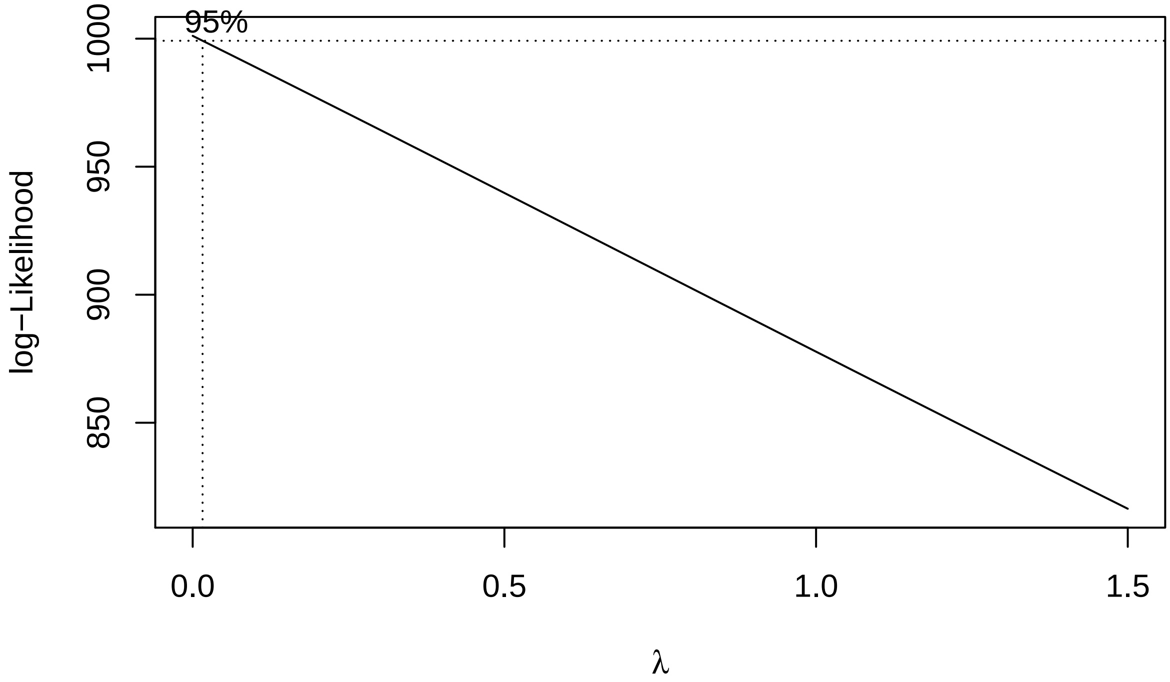


Apart from the tails, the regression line almost perfectly fits the data. This indicates that we will have to difference at Lag $= 1$ to remove the trend. We must also difference at Lag $= 12$ to remove seasonality.

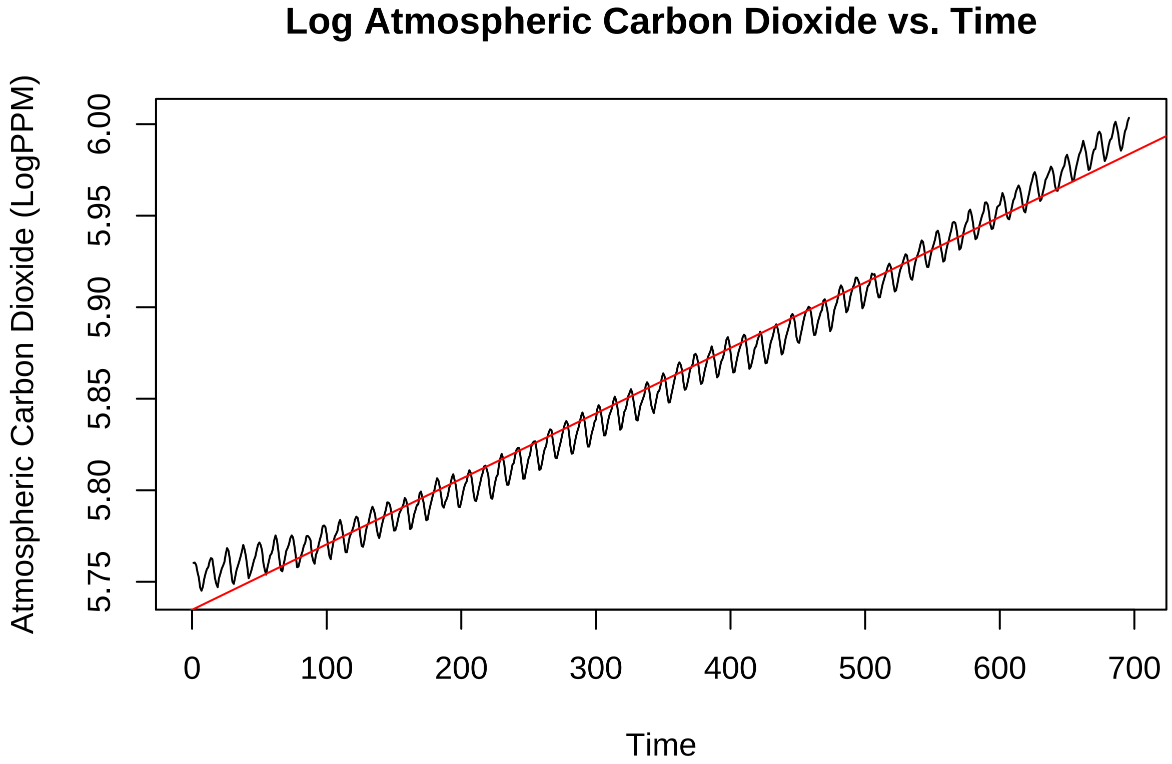
In order to stabilize variance, we employ the Box-Cox Power transformation. This method finds the best λ to apply to the following transformation:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log y_i, & \text{if } \lambda = 0 \end{cases}$$

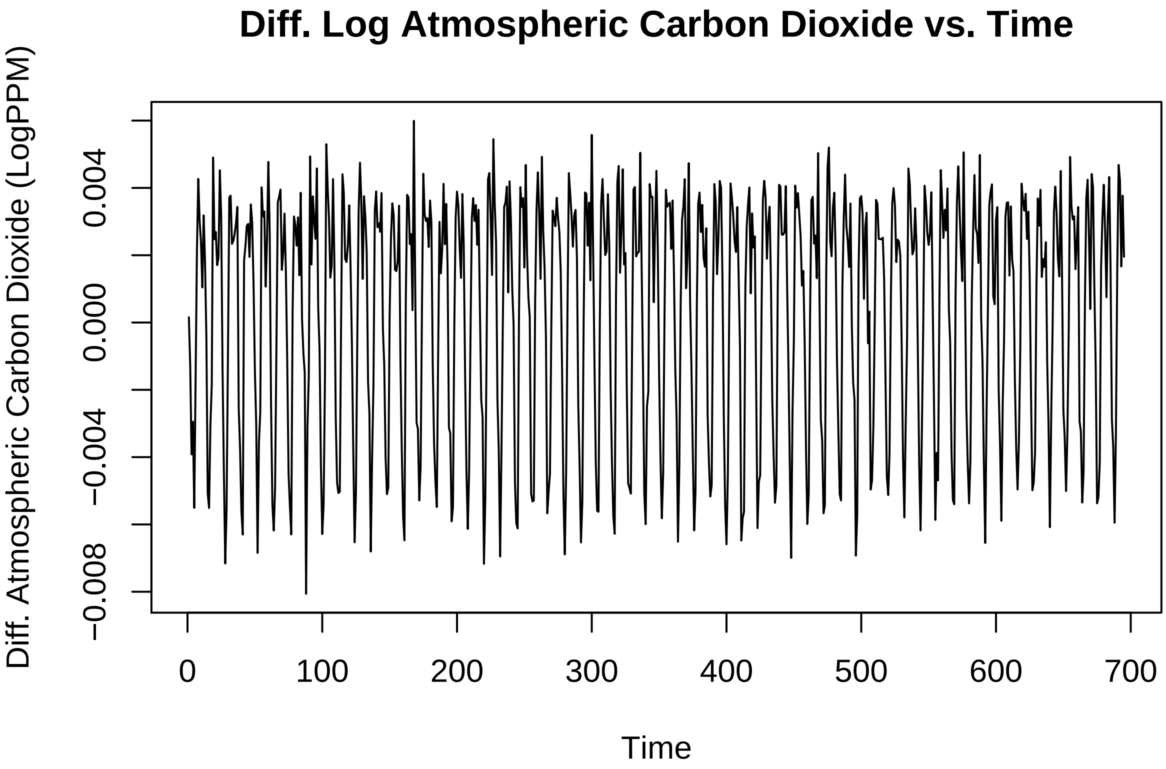
We use `R` to find the best λ. In accordance with the methodology in Brockwell & Davis, we restrict the parameter space so that $\lambda \in [0, 1.5]$.



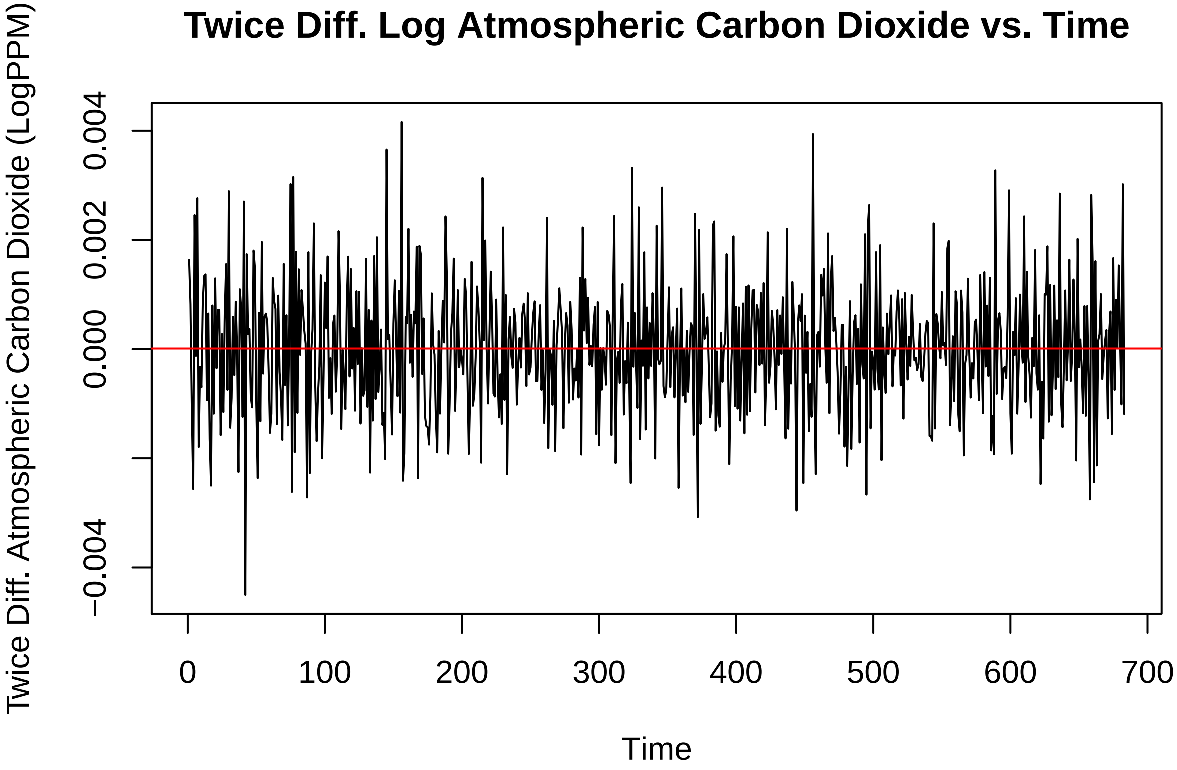
Clearly $\lambda = 0$ yields the best transformation, so we take the the log of the time series. We plot the transformed series and once again add the best-fitting regression line.



The transformation not only stabilizes variance, but also make the positive trend almost exactly linear. We now difference the data at lag = 1 to remove this trend and plot our results.



Model variance decreases by 0.005218006, so this differencing is justified. We now difference at lag = 12 to remove the seasonal component of the time series.



The variance again decreases, this time by `1.065186e-05`, so this differencing is again justified. The mean of the time series is added in red. The model also appears to be stationary, as neither mean nor variance appear to be dependent on time.

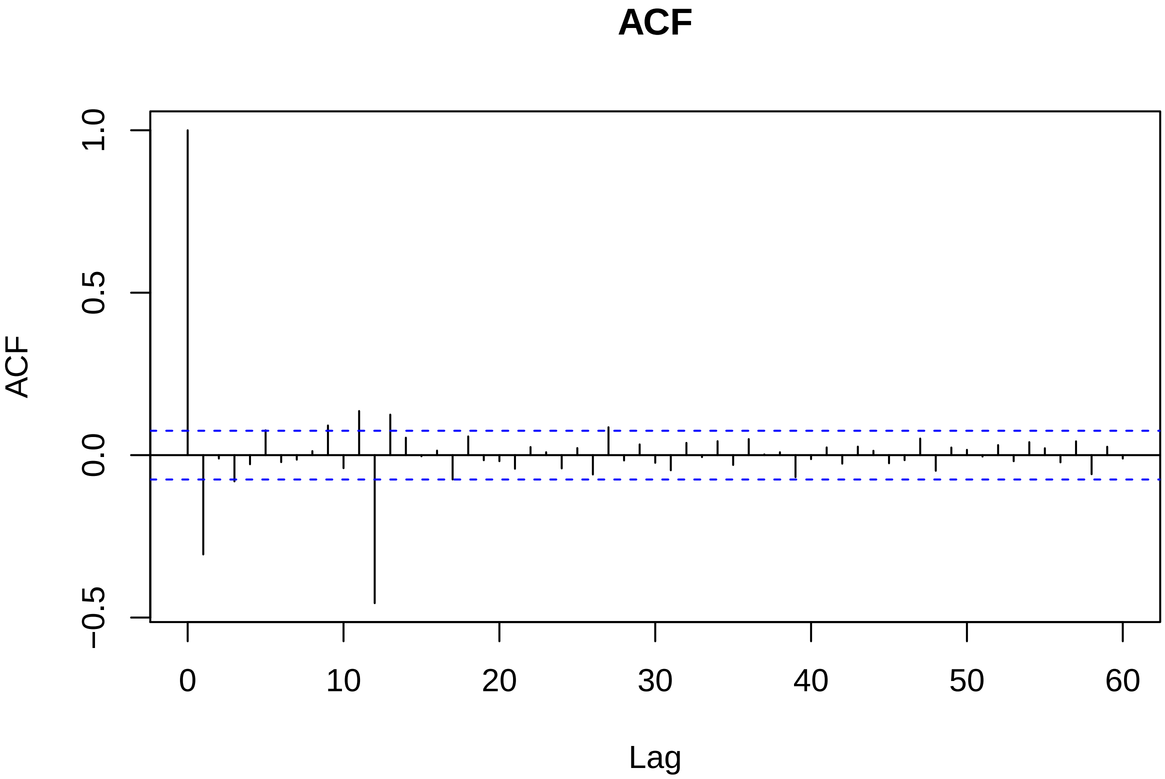
To confirm the stationarity of the series, we perform Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test in `R` under the null hypothesis that the series is stationary.

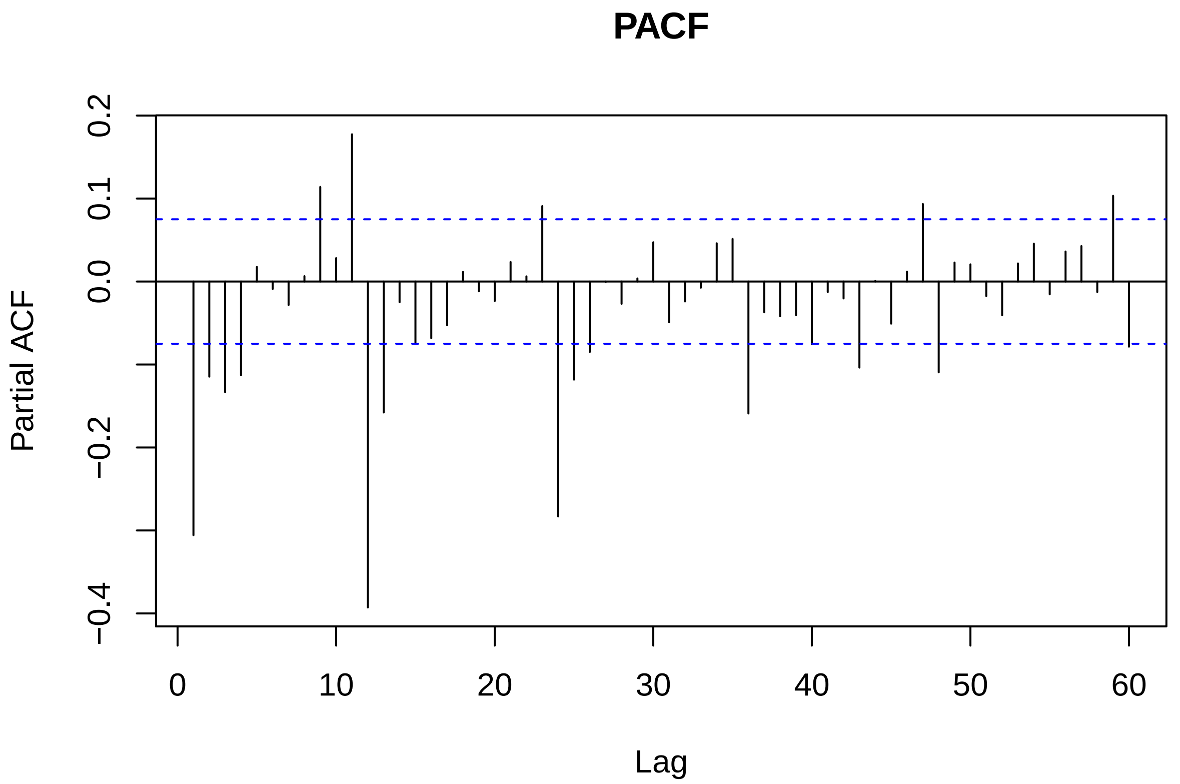
```
##
##  KPSS Test for Level Stationarity
##
## data:  ts.log.1.12
## KPSS Level = 0.010779, Truncation lag parameter = 6, p-value = 0.1
```

This test yields a p-value > 0.05, so we fail to reject the null hypothesis at the $\alpha = 0.05$ significance level and conclude that the model is indeed stationary.

3.3 ACF and PACF Analysis

Now that the time series is stationary, we analyze its ACF and PACF plots to identify the AR, MA, SAR, and SMA orders in the SARIMA model.





We begin by examing the seasonal compents, which are visble at lags $l = 12n$, $n \in \mathbb{N}$. We can clearly see that the ACF cuts off after lag 12, while the PACF exponentially decays at lags that are multiples of 12. This leads us to consider $SAR = 0$ and $SMA = 1$.

Next, we examine lags 1 through 11 to find the AR and MA orders of the model. The ACF cuts off after lag 1 and the PACF quickly decays, so we consider $MA = 1$ and $AR = 0$. It is also possible that the ACF is tailing off while the PACF cuts off after lag 4, which implies $AR = 4$ and $MA = 0$. With two possible model orders, we estimate their respective parameters using Maximum Likelihood Estimation.

We consider the following two models:

- 1. SARIMA $(0, 1, 1)$ x $(0, 1, 1)_{12}$
  - AICc $= -7673.12$ BIC $= -7659.57$
  - $\nabla_{12}\nabla Y_t = (1 - 3.9B)(1 - 0.89B^{12})Z_t$
- 2. SARIMA $(4, 1, 0)$ x $(0, 1, 1)_{12}$
  - AICc $= -7673.47$ BIC $= -7646.44$
  - $(1 + 0.37B + 0.17B^2 + 0.13B^3 + 0.1B^4)\nabla_{12}\nabla Y_t = (1 - 0.89B^{12})Z_t$

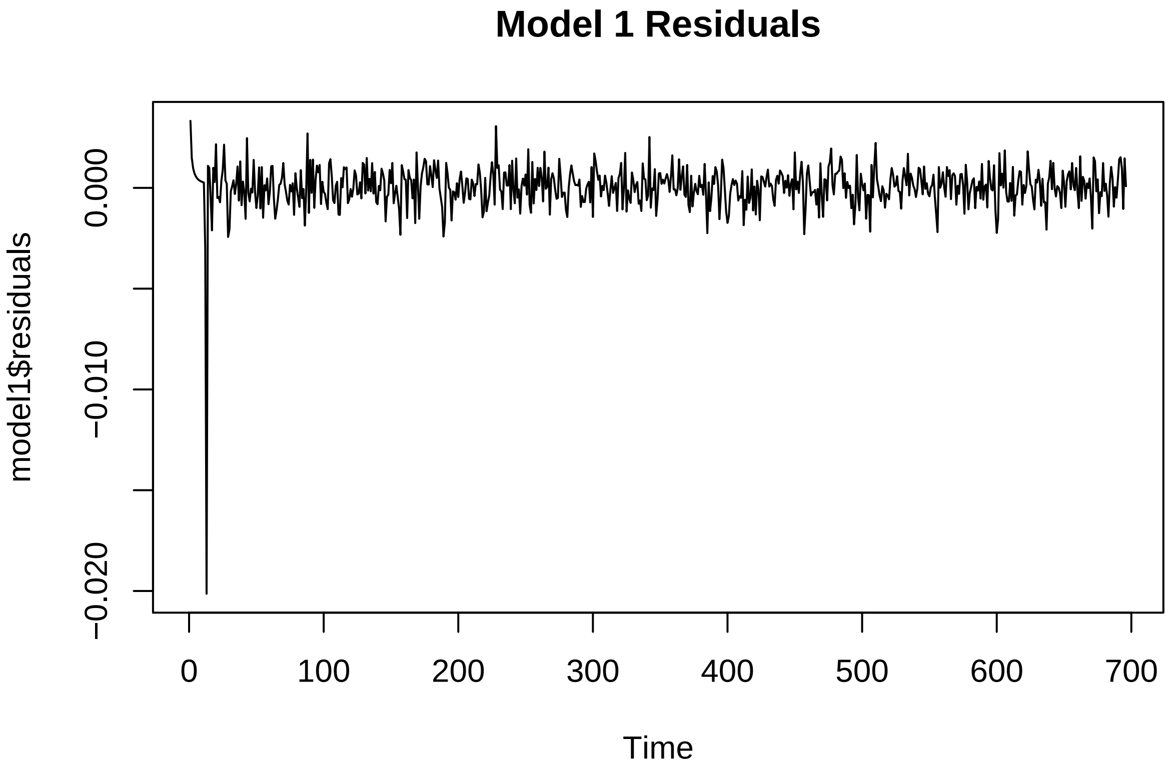
3.4 Model Diagnostics

3.4.1 Model 1

$$\nabla_{12}\nabla Y_t = (1 - 3.9B)(1 - 0.89B^{12})Z_t$$

We perform diagnostic checking for the SARIMA $(0, 1, 1)$ x $(0, 1, 1)_{12}$ model.

We begin by plotting the residuals of this model



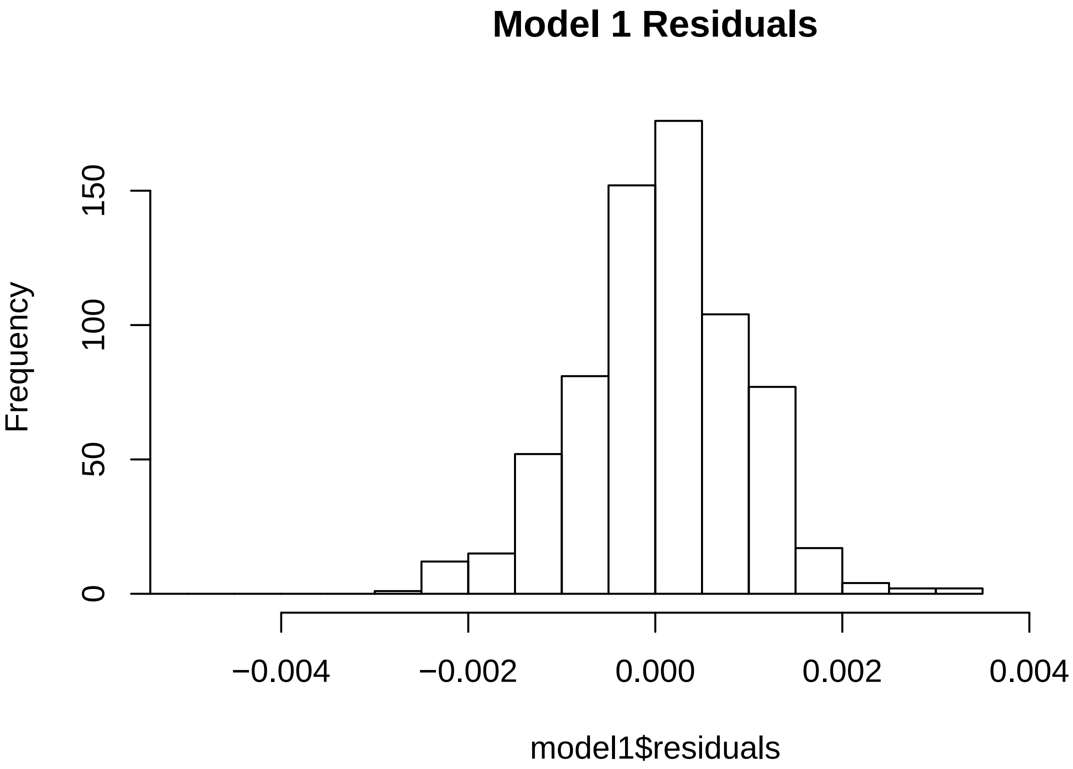
Apart from a single outlier at $t = 13$, the residuals appear to resemble white noise.

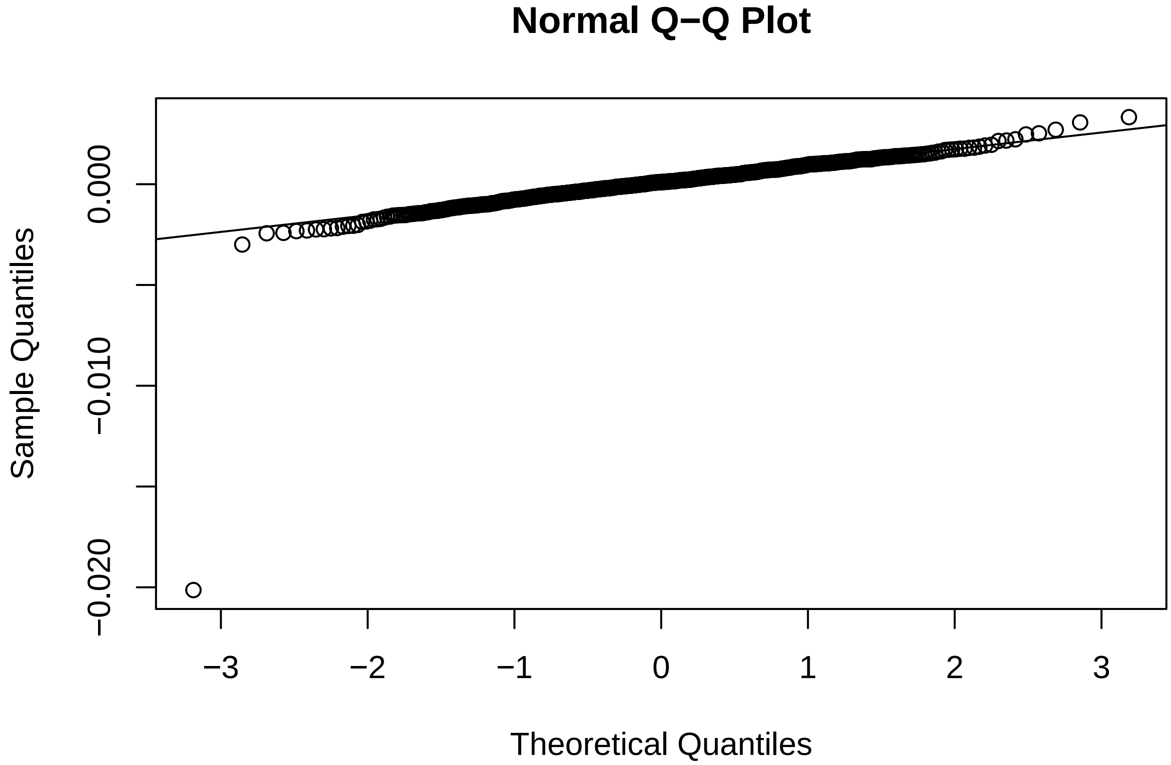
Interstingly, the residuals fail the Shapiro-Wilks normality test with this outlier present, put pass the test when it is removed.

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.68406, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals[14:706]
## W = 0.99634, p-value = 0.1168
```

We now plot the density and the QQ-Plot of the residuals.





From these plots, the residuals appears to be approximately normally distributed.

We now perform the Box-Pierce, Ljung-Box, and McLeod-Li tests on the residuals.

```
## [1] "Box-Pierce"
```

```
##
##  Box-Pierce test
##
## data:  model1$residuals
## X-squared = 19.126, df = 24, p-value = 0.7452
```

```
## [1] "Ljung-Box"
```

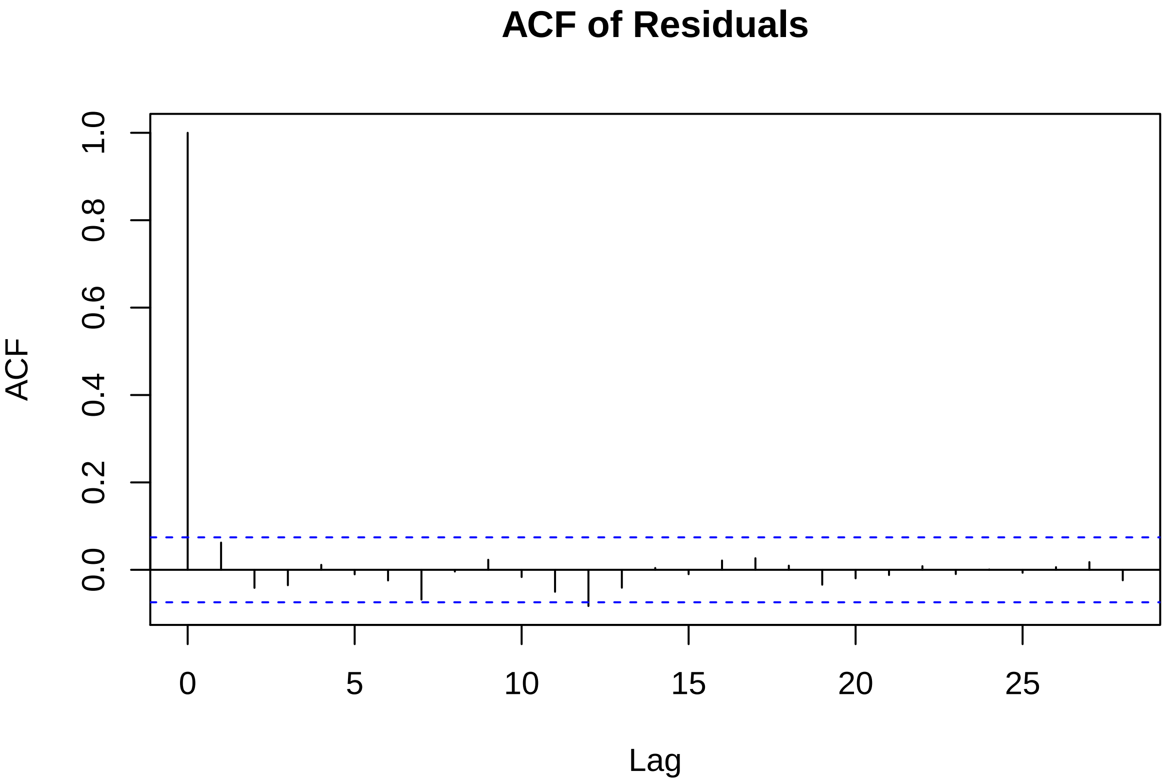
```
##
##  Box-Ljung test
##
## data:  model1$residuals
## X-squared = 19.435, df = 24, p-value = 0.7284
```

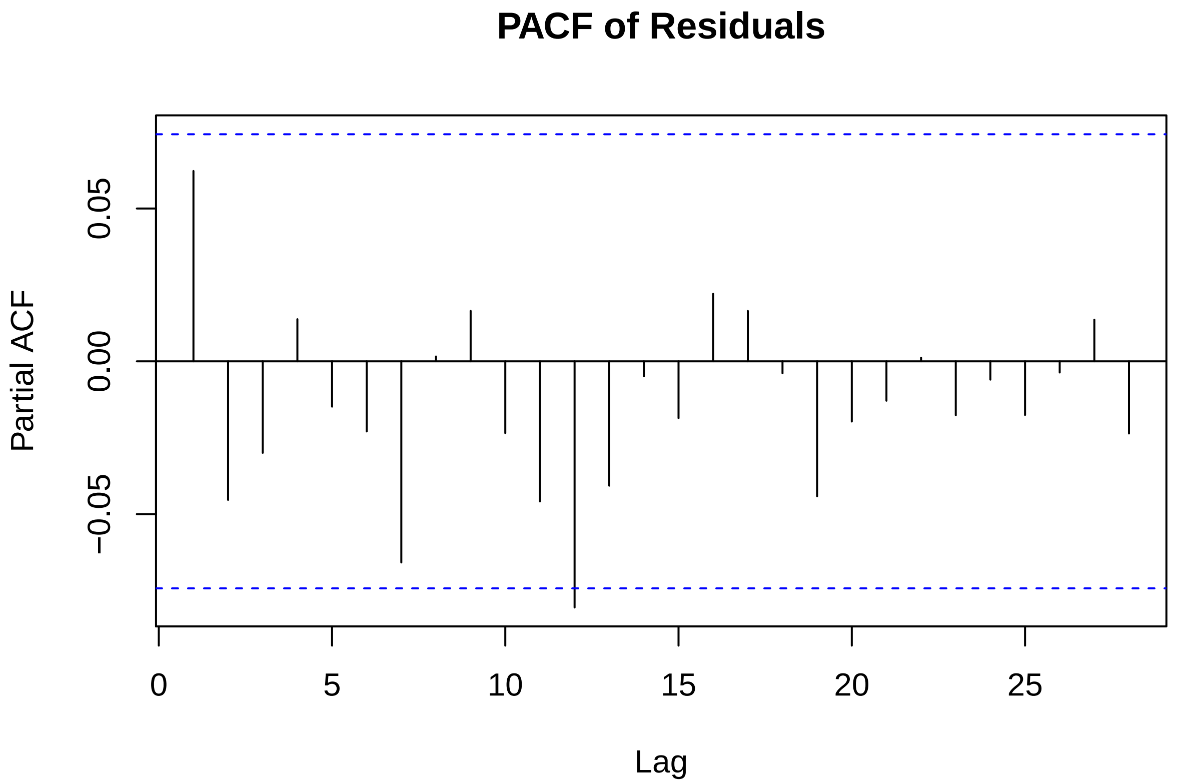
```
## [1] "McLeod-Li"
```

```
##
##  Box-Ljung test
##
## data:  (model1$residuals)^2
## X-squared = 1.0825, df = 26, p-value = 1
```

This model passes all adequacy tests at the $\alpha = 0.05$ significance level.

We now plot the ACF and PACF of the residuals to ensure they resemble white noise.





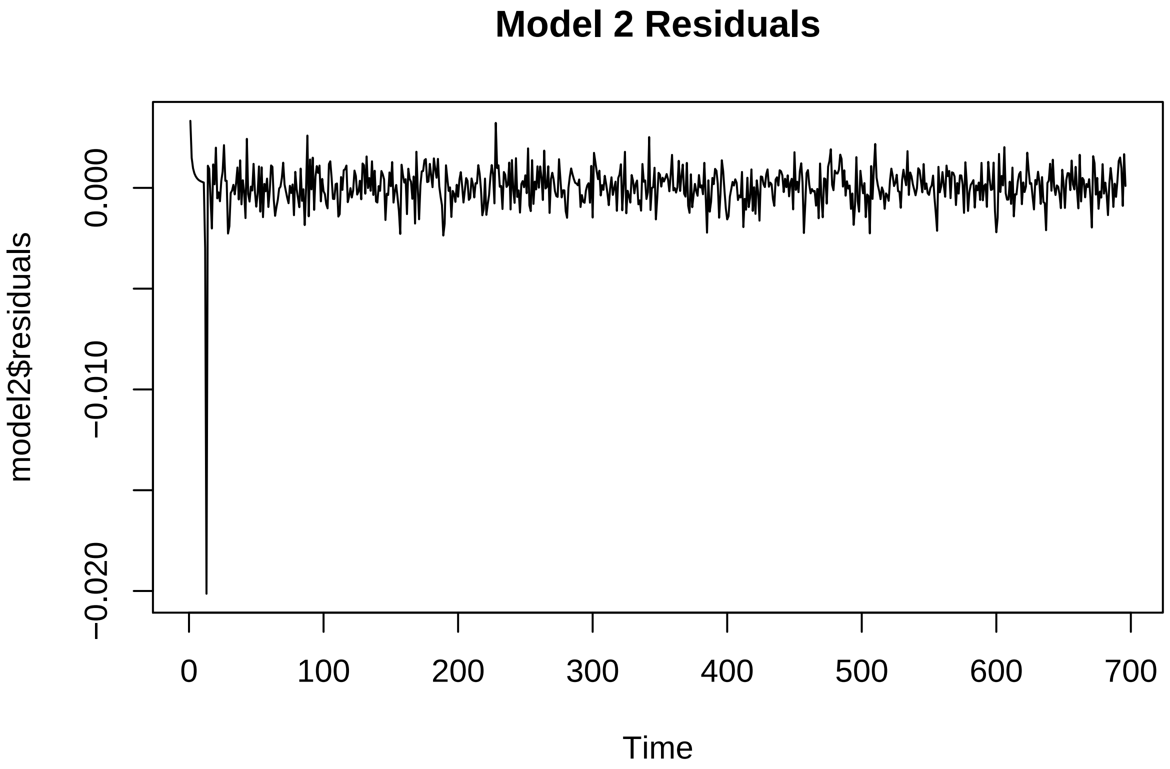
The ACFs and PACFs extend slightly beyond the confidence intervals at lag 12, but besides this resemble white noise.

Based on the results of this test, we deem this model adequate for forecasting.

3.4.2 Model 2

$$(1 + 0.37B + 0.17B^2 + 0.13B^3 + 0.1B^4)\nabla_{12}\nabla Y_t = (1 - 0.89B^{12})Z_t$$

We now check the adequacy of the second model. We begin by plotting the model residuals.

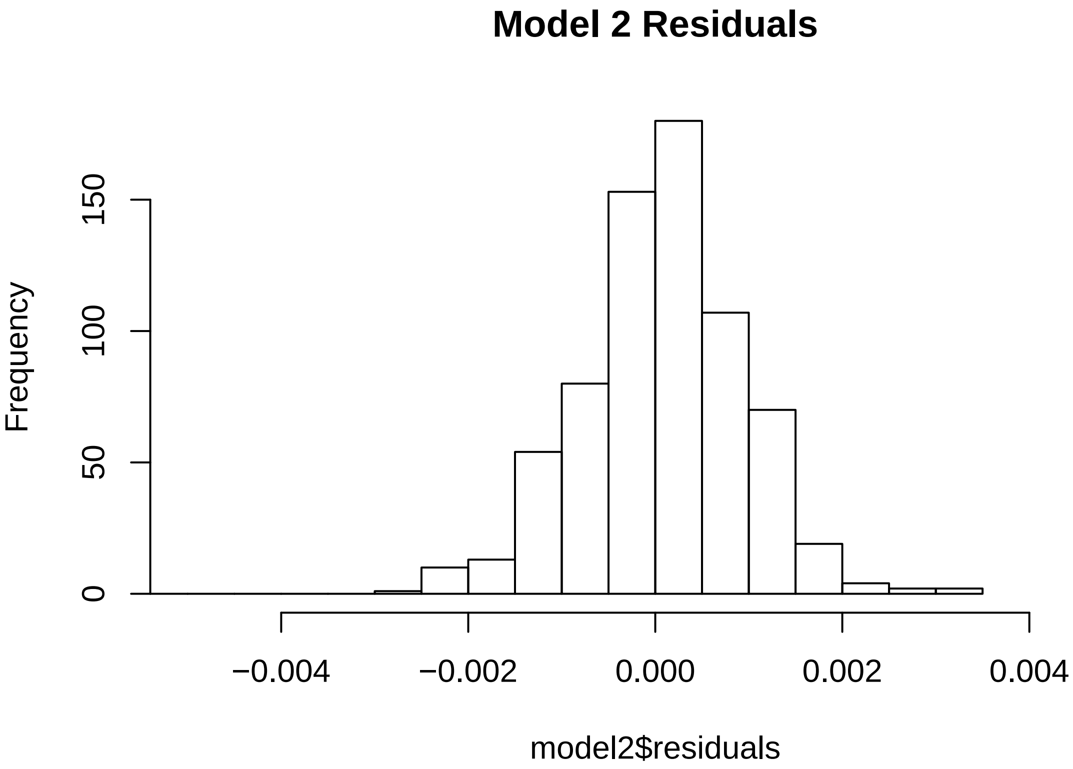


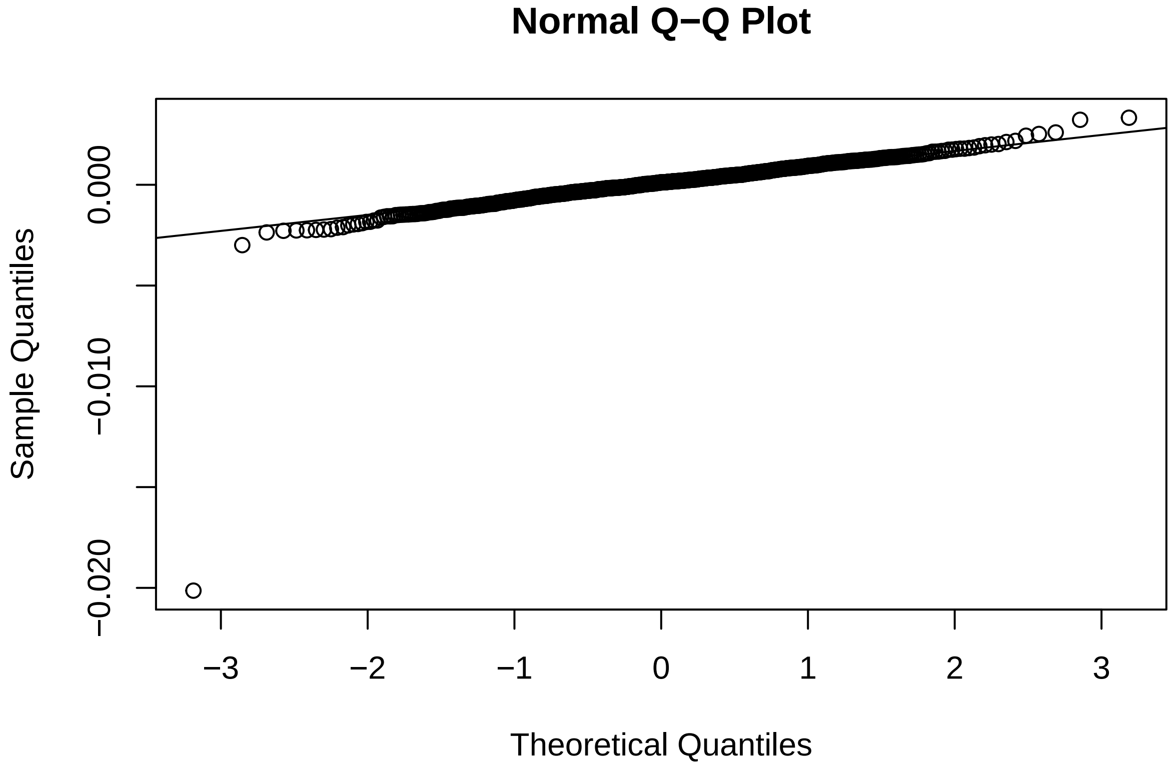
Just as before, there is a large outlier at $t = 13$. Besides this single point, the residuals resemble white noise. Just as before, the Shapiro-Wilks normality test fails for the unmodified residuals, but indicates normality when this point is removed.

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.68175, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals[14:706]
## W = 0.9964, p-value = 0.1245
```

We now plot the density of the residuals and construct a QQ-Plot.





The histogram and QQ-Plot also show that the residuals are distributed normally. We now perform the Box-Pierce, Ljung-Box, and McLeod-Li tests on the residuals.

```
## [1] "Box-Pierce"
```

```
##
```

```
##  Box-Pierce test
##
## data:  model2$residuals
## X-squared = 16.817, df = 21, p-value = 0.7221
```

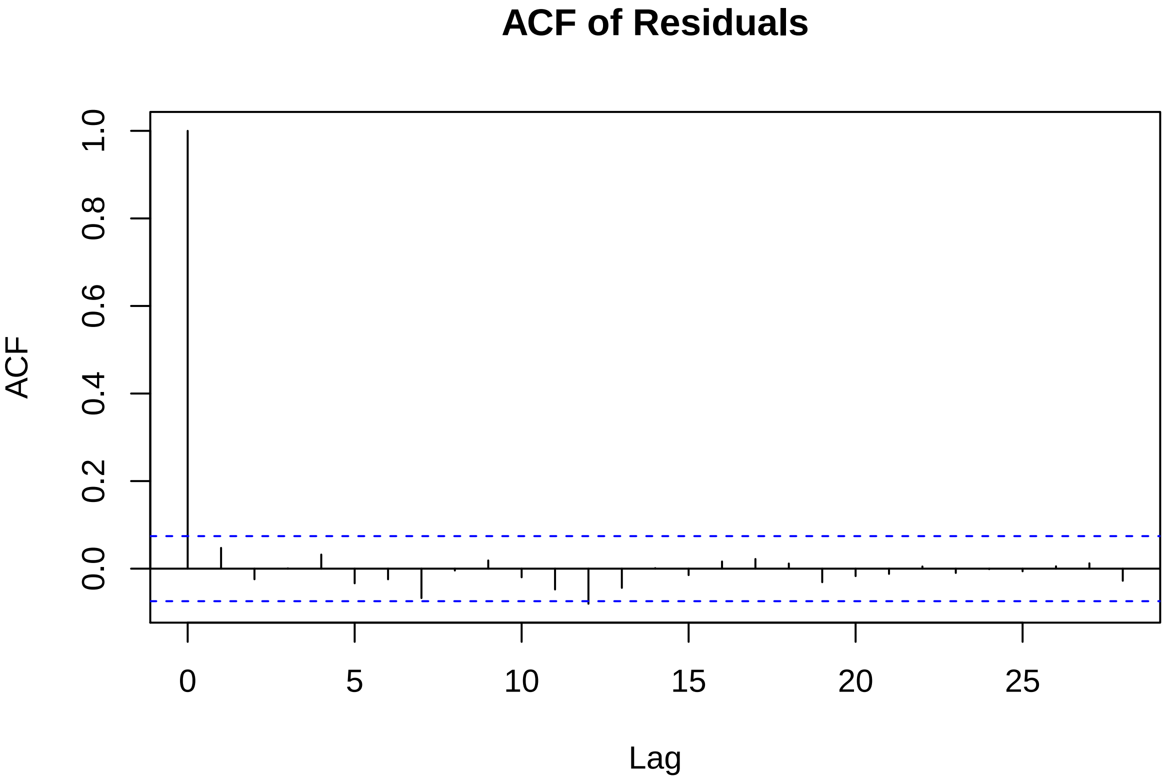
```
## [1] "Ljung-Box"
```

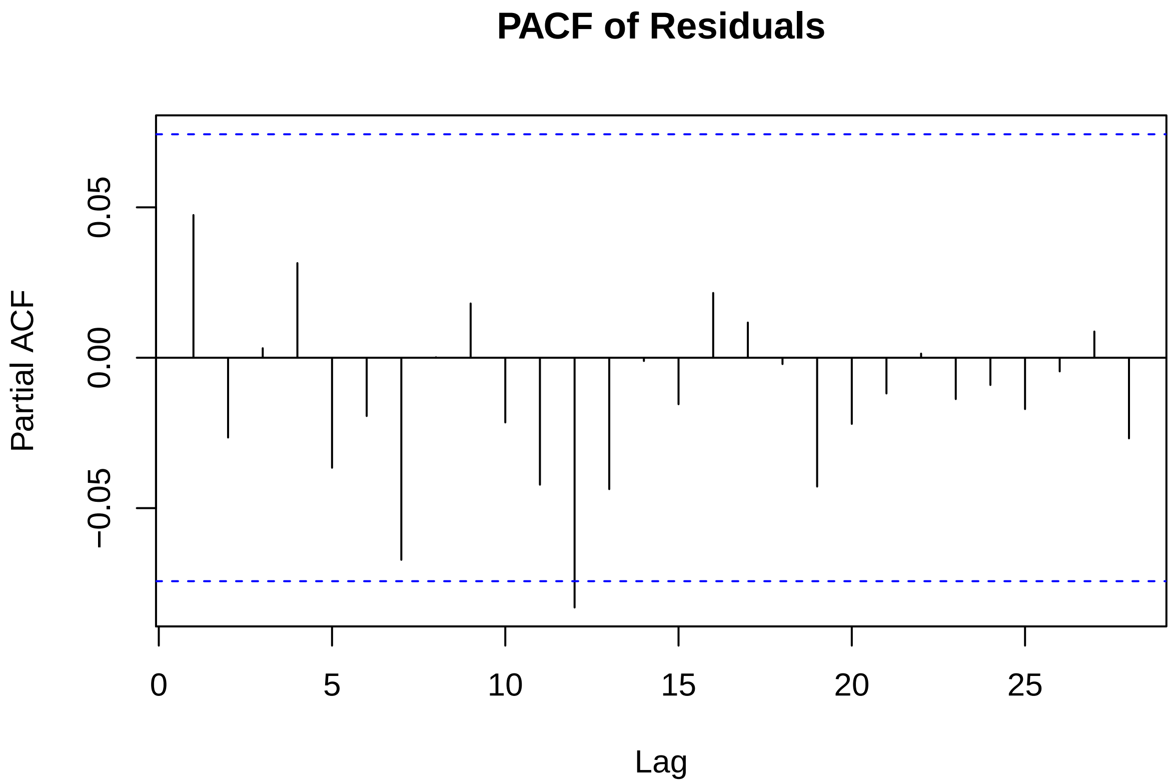
```
##
##  Box-Ljung test
##
## data:  model2$residuals
## X-squared = 17.103, df = 21, p-value = 0.7049
```

```
## [1] "McLeod-Li"
```

```
##
##  Box-Ljung test
##
## data:  (model2$residuals)^2
## X-squared = 1.0023, df = 26, p-value = 1
```

The model passes all adequacy tests at the $\alpha = 0.05$ significance level. Finally, we check that the ACF and PACF of the residuals resemble white noise.



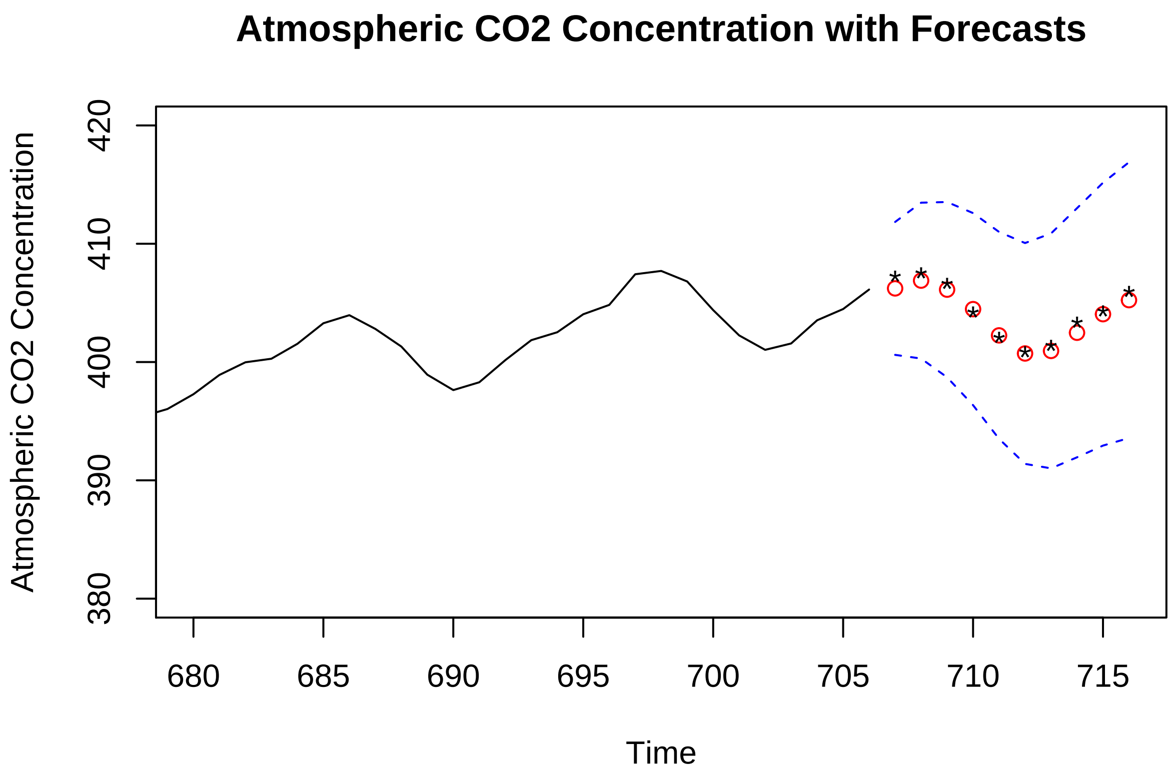


Just as before, there are signifcant spikes at lag 12 in both plots, but we deem this to be acceptable. Like the previous model, Model 2 is suitable for forecasting.

3.4.3 Choosing the Best Model Since both Model 1 and Model 2 pass our residuals tests and have similar AICc and BIC values, choosing a final 'best' model is difficult. We therefore must rely on the principle of parsimony. As such, we choose Model 1 as the final model because it is much simpler, having only two parameters against the five parameters of Model 2.

3.5 Forecasting

Now that we have a final model, we may begin forecasting. Forecasts are plotted below in red, and the actual observations appear as asterisks. The blue dotted lines represent the 95% confidence interval of the forecasts.



One can clearly see that the forecasts lie very close to the realized observations. Consequently, we conclude that the model provides great accuracy in forecasting this phenomenon.

4. Conclusion

Seasonal autoregressive integrated moving average time series models provide a viable framework for forecasting atmospheric carbon dioxide concentrations. Of the two SARIMA models considered, both meet the assumptions of the Box-Jenkins methodology for time series forecasting. Consequenty, we yield to Occam's razor and choose the model with the fewest paramerters as our final model. This model is given by:

$$\nabla_{12}\nabla Y_t = (1 - 3.9B)(1 - 0.89B^{12})Z_t$$

As one can see from the previous section, this model provides incredibly accurate forecasts of atmospheric CO_2 concentrations 10 months into the future.

5. References

- Warrick, Joby. 2014. The Washington Post - CO2 levels in atmosphere rising at dramatically faster rate, U.N. report warns. Link.
- Brockwell, PJ., Davis, RA. 2016. Introduction to Time Series and Forecasting.
- Hyndman, Rob J. 2012. Constants and ARIMA models in R. Link.
- Shumway, Robert H., Stoffer, David S. Time Series Analysis and Its Applications with R Examples. 3rd Edition. 2010.
- The National Oceanic and Atmospheric Administration
- The R Project for Statistical Computing

Appendix

The R code used to create the analysis in this report is provided in its entirety below.

```r
library(stats)
# Import Data
ts <- read.table("c02.txt", skip = 3, header = F, sep = "")

# Remove exraneous featurs
ts <- ts$V5

# Plot original series
plot.ts(ts, ylab = "Atmospheric Carbon Dioxide (PPM)",
        main = "Atmospheric Carbon Dioxide vs. Time")

# Copy series to new variable
orig <- ts

# Reserve last 10 observations
next10 <- orig[697:706]

# Convert vector to time series object
ts <- as.ts(ts, start=c(1958,3), frequency=12)
ts <- ts[1:696]

# Plot with regression line
linMod <- lm(ts~as.numeric(1:length(ts)))
plot.ts(ts, ylab = "Atmospheric Carbon Dioxide (PPM)",
        main = "Atmospheric Carbon Dioxide vs. Time")
abline(linMod, col = "red")

# Find lambda for Box-Cox transformation
library(MASS)
bcTransform <- boxcox(ts~as.numeric(1:length(ts)), lambda = seq(0,1.5,1/10))
trans <- bcTransform$x[which.max(bcTransform$y)]

# Transform data, plot with regression line
ts.log <- log(ts)
linMod2 <- lm(ts.log~as.numeric(1:length(ts.log)))
plot.ts(ts.log, ylab = "Atmospheric Carbon Dioxide (LogPPM)",
        main = "Log Atmospheric Carbon Dioxide vs. Time")
abline(linMod2, col = "red")

# Difference at lag1, compare variance
ts.log.1 <- diff(ts.log,1)
plot.ts(ts.log.1, ylab = "Diff. Atmospheric Carbon Dioxide (LogPPM)",
        main = "Diff. Log Atmospheric Carbon Dioxide vs. Time")
var.ts.log <- var(ts.log, na.rm = T)
var.ts.log.1 <- var(ts.log.1, na.rm =T)

# Difference at lag12, compare variance
ts.log.1.12 <- diff(ts.log.1,12)
plot.ts(ts.log.1.12, ylab = "Twice Diff. Atmospheric Carbon Dioxide (LogPPM)",
        main = "Twice Diff. Log Atmospheric Carbon Dioxide vs. Time")
```

```r
abline(h=mean(ts.log.1.12, na.rm = T), col = "red")
var.ts.log.1.12 <- var(ts.log.1.12, na.rm =T)

# KPSS Test for stationarity
library(tseries)
kpss.test(ts.log.1.12)

# Plot ACF, PACF
acf <- acf(ts.log.1.12, type = "correlation",
           plot = T, na.action = na.pass,
           lag.max=12*5, main ="ACF")
pacf <- acf(ts.log.1.12, type = "partial",
            plot = T, na.action = na.pass,
            lag.max=12*5, main = "PACF")

# Build two appropriate models
library(forecast)
model1 <- Arima(ts.log, order = c(0,1,1),
                seasonal = list(order = c(0,1,1), period = 12))
model2 <- Arima(ts.log, order = c(4,1,0),
                seasonal = list(order = c(0,1,1), period = 12))

# Model 1 Diagnostics
plot(model1$residuals, main="Model 1 Residuals")
shapiro.test(model1$residuals)
shapiro.test(model1$residuals[14:706])
hist(model1$residuals, xlim = c(-0.005,0.005),
     main = "Model 1 Residuals", breaks = 50)
qqnorm(model1$residuals)
qqline(model1$residuals)
paste("Box-Pierce")
Box.test(model1$residuals, lag = 26, type = "Box-Pierce", fitdf=2)
paste("Ljung-Box")
Box.test(model1$residuals, lag = 26, type = "Ljung-Box", fitdf=2)
paste("McLeod-Li")
Box.test((model1$residuals)^2, lag=26, type="Ljung-Box")
acf(model1$residuals, na.action=na.pass, main = "ACF of Residuals")
pacf(model1$residuals, na.action = na.pass, main = "PACF of Residuals")

# Model 2 Diagnostics
plot(model2$residuals, main="Model 1 Residuals")
shapiro.test(model2$residuals)
shapiro.test(model2$residuals[14:706])
hist(model1$residuals, xlim = c(-0.005,0.005),
     main = "Model 2 Residuals", breaks = 50)
qqnorm(model2$residuals)
qqline(model2$residuals)
paste("Box-Pierce")
Box.test(model2$residuals, lag = 26, type = "Box-Pierce", fitdf=5)
paste("Ljung-Box")
Box.test(model2$residuals, lag = 26, type = "Ljung-Box", fitdf=5)
paste("McLeod-Li")
Box.test((model2$residuals)^2, lag=26, type="Ljung-Box")
```

```r
acf(model2$residuals, na.action=na.pass,
    main = "ACF of Residuals")
pacf(model2$residuals, na.action = na.pass,
     main = "PACF of Residuals")

# Forecasting with Model 1
pred <- predict(model1, n.ahead = 10)
pred.orig <- exp(pred$pred)
pred.se <- exp(pred$pred)*pred$pred*pred$se

plot.ts(orig, xlim = c(680,length(orig)+10), ylim = c(380,420),
        ylab = "Atmospheric CO2 Concentration",
        main = "Atmospheric CO2 Concentration with Forecasts")
points((length(orig)+1):(length(orig)+10),pred.orig, col="red")
points((length(orig)+1):(length(orig)+10),next10, pch = "*")
lines((length(orig)+1):(length(orig)+10),pred.orig+1.96*pred.se,lty=2, col="blue")
lines((length(orig)+1):(length(orig)+10),pred.orig-1.96*pred.se,lty=2, col="blue")
```

This report was generated with `R Markdown` and LaTeX .