PSTAT 174 Final Project

Philip Yoon

Abstract

This project will apply Box-Jenkins analysis techniques to forecast the monthly gasoline demand (in gallon millions) of Ontario, using time series data from 1960 to 1975. The goal is to create a (S)ARIMA model that can forecast the gasoline demand in the 12 months of 1975 using historical data from previous years. After a combination of transformations and differencing, the stationary data's ACF and PACF were analyzed to estimate models for forecasting. A stationary and invertible model were chosen using MLE estimation of parameters, AICc criterion, and the principle of parsimony, which was then used to forecast 1975 demand. The model did well; however, diagnostic checking indicated the results of said techniques were passable but not ideal.

Introduction

I will attempt to create an (S)ARIMA model to forecast Ontario's gasoline demand starting Janurary 1975 to December 1975 using time series data from 1960 up until 1975.

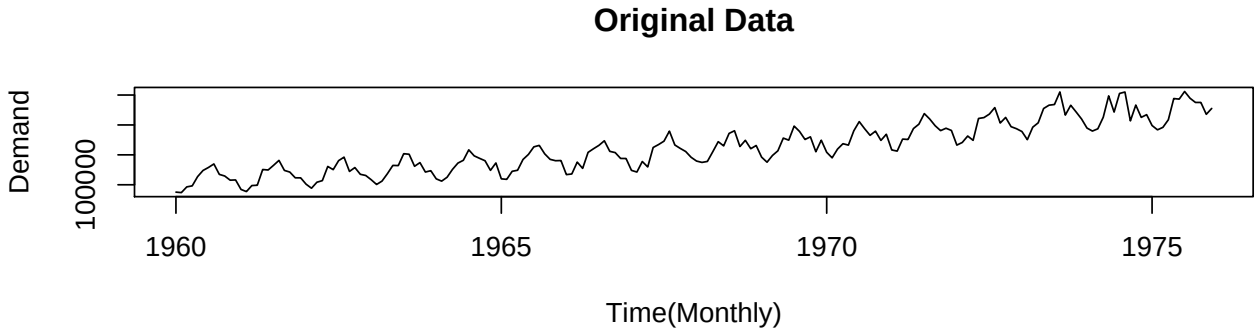
The demand of gasoline is important because it indicates whether the existing supply is enough for future gas requirements. Creating a model that can forecast the demand of gasoline from past data would be very useful. An example application is from the perspective of a gasoline distributor, where one would be able to determine if sufficient gas is stored for the change of gas demand in the future.

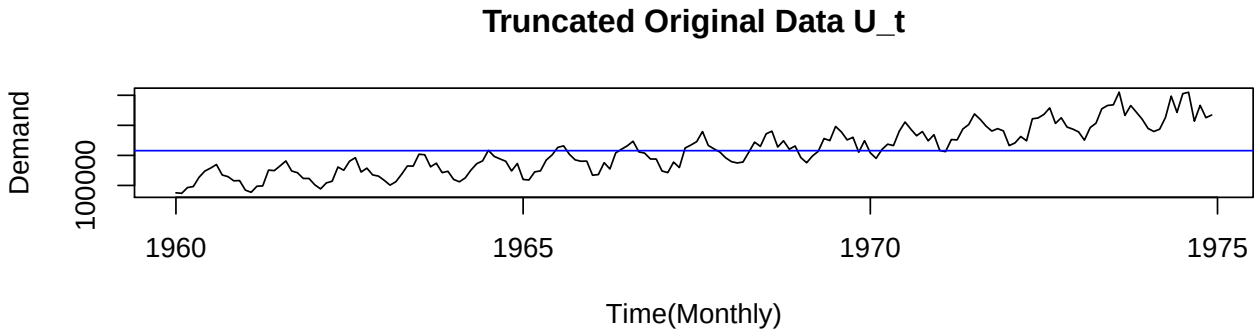
Using the Box Jenkins Method, I first wanted a set of stationary data with stable variance and symmetric distribution, which I achieved using a log transformation and differencing at lags 12, to remove seasonality, and lag 1, to remove linear trend. Viewing the ACF and PACF of this data, I then fit some SARIMA models which I compared using AICc criterion. Then I tested these models using Shapiro-Wilk, Box-Pierce, Ljung-Box, and McLeod-Li tests to obtain a final stationary and invertible model with residuals that imitate gaussian white noise. Finally, I used this model to forecast the demand of gasoline in Ontario during 1975, which I compared with the true values of demand. The forecasted values were not perfect but the true values were within the prediction interval of the model, so my final model did reasonably well.

Data is referenced from Abraham & Ledolter (1983). All code is run and compiled using RStudio.

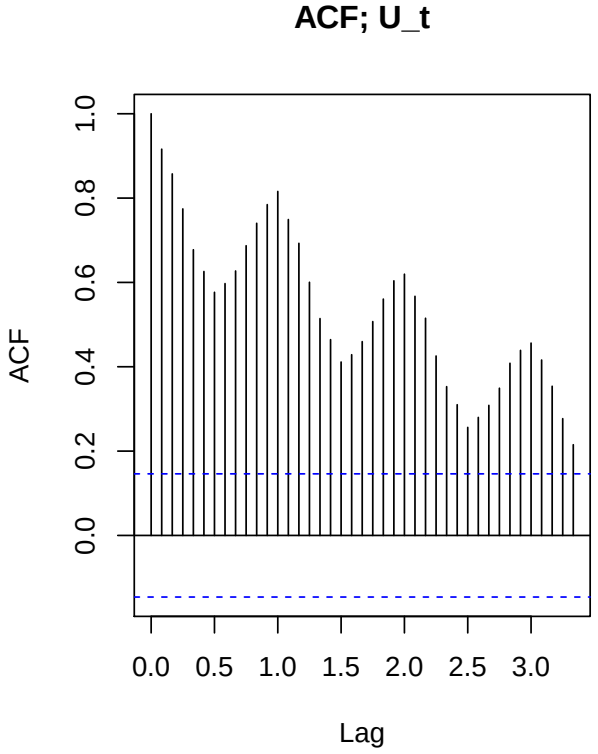
Since no new data is expected, I will leave the last 12 points for model validation, leaving the first 180 data points for model building. The truncated dataset will be called U_t.

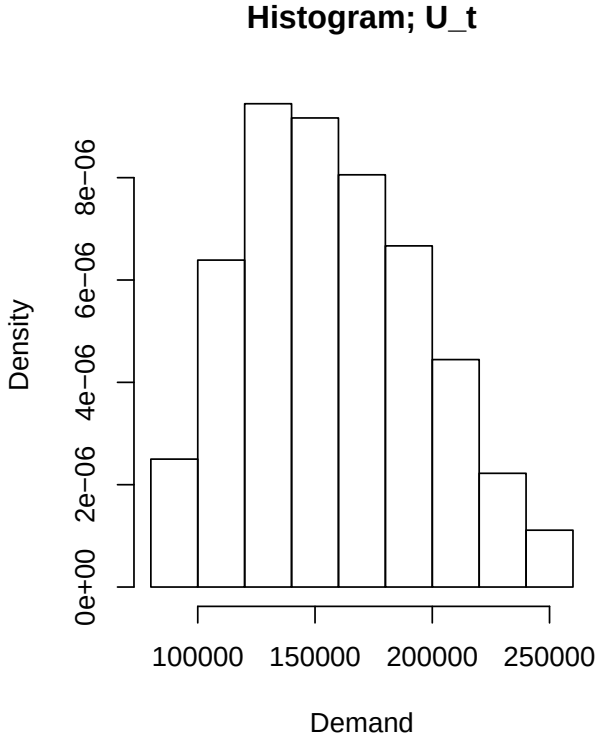
Exploratory Data Analysis





Immediately, I observe an upward linear trend as well as a clear seasonal component, which suggests highly non-stationary data.

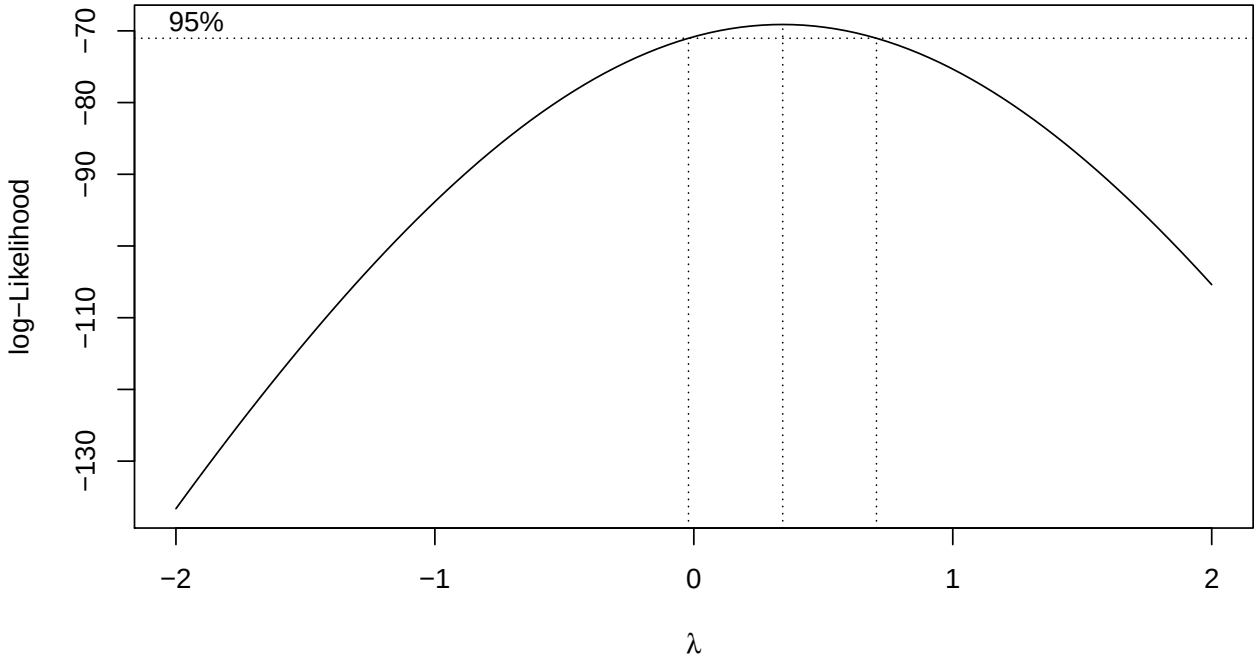




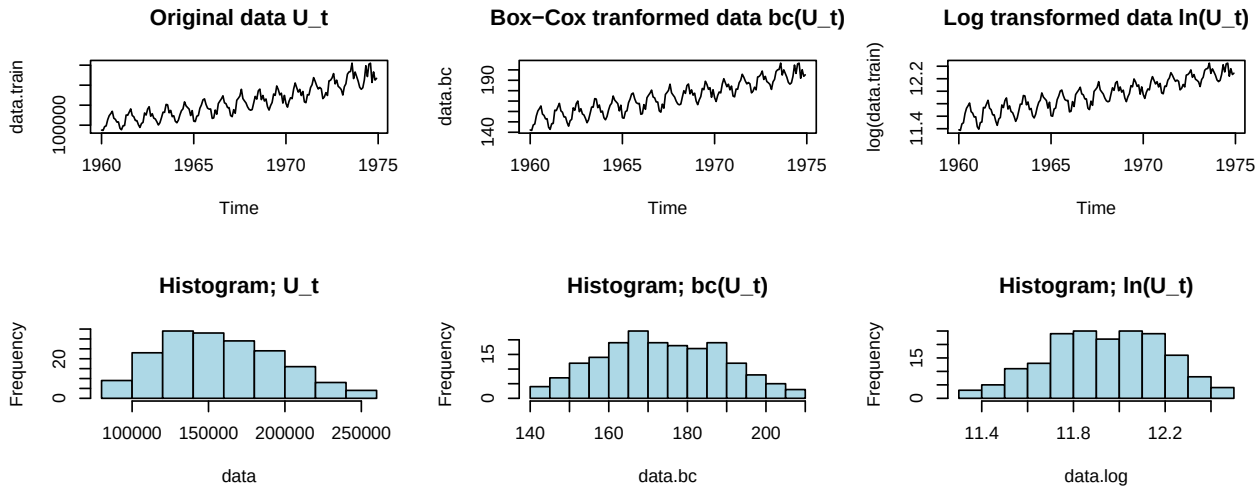
The skewed distribution from the histogram and the periodic and slow-decaying ACF confirm the non-stationarity of the original data.

In order to stabilize the variance and have a more symmetric distribution, I will attempt a transformation. To remove the seasonlity and trend, I will be differencing. The result will hopefully be a stationary time series.

First, I will try Box-Cox Transformation in order to make the data more normal.

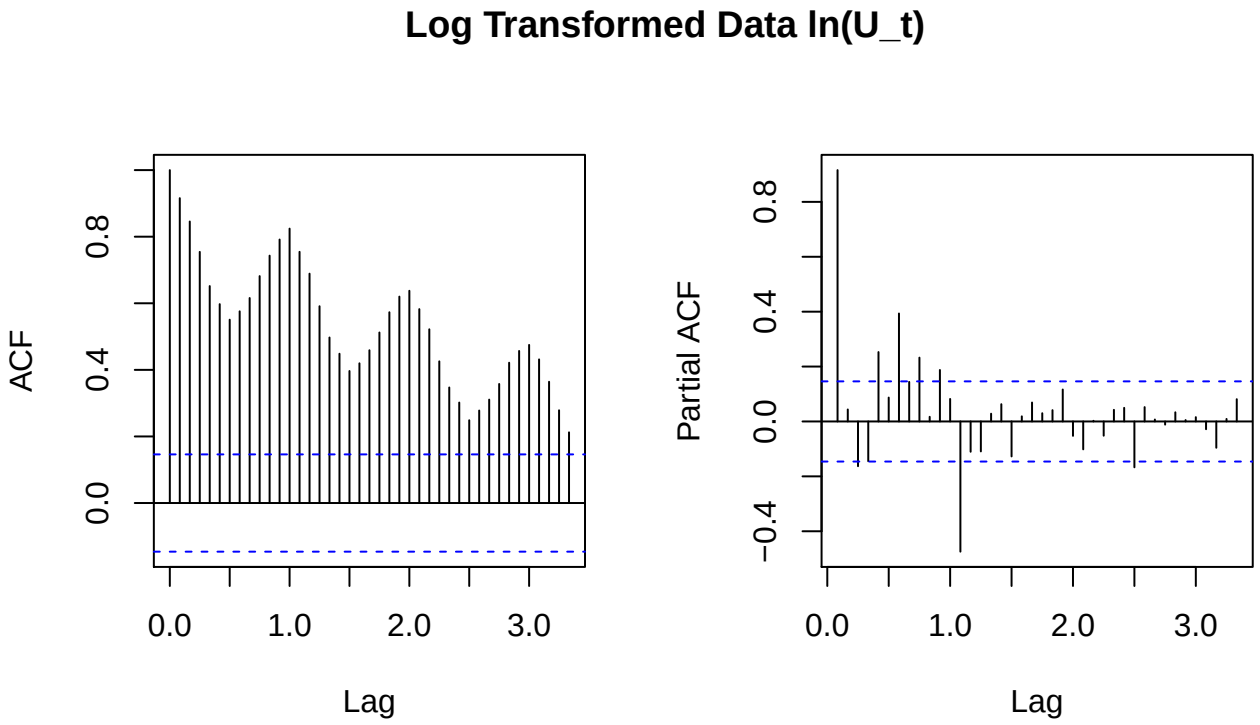


The BcTransform command gives a lambda of 0.3434, but because 0 is in the confidence interval, which is the same as a log transform, I will compare the results of both log and Box-Cox.



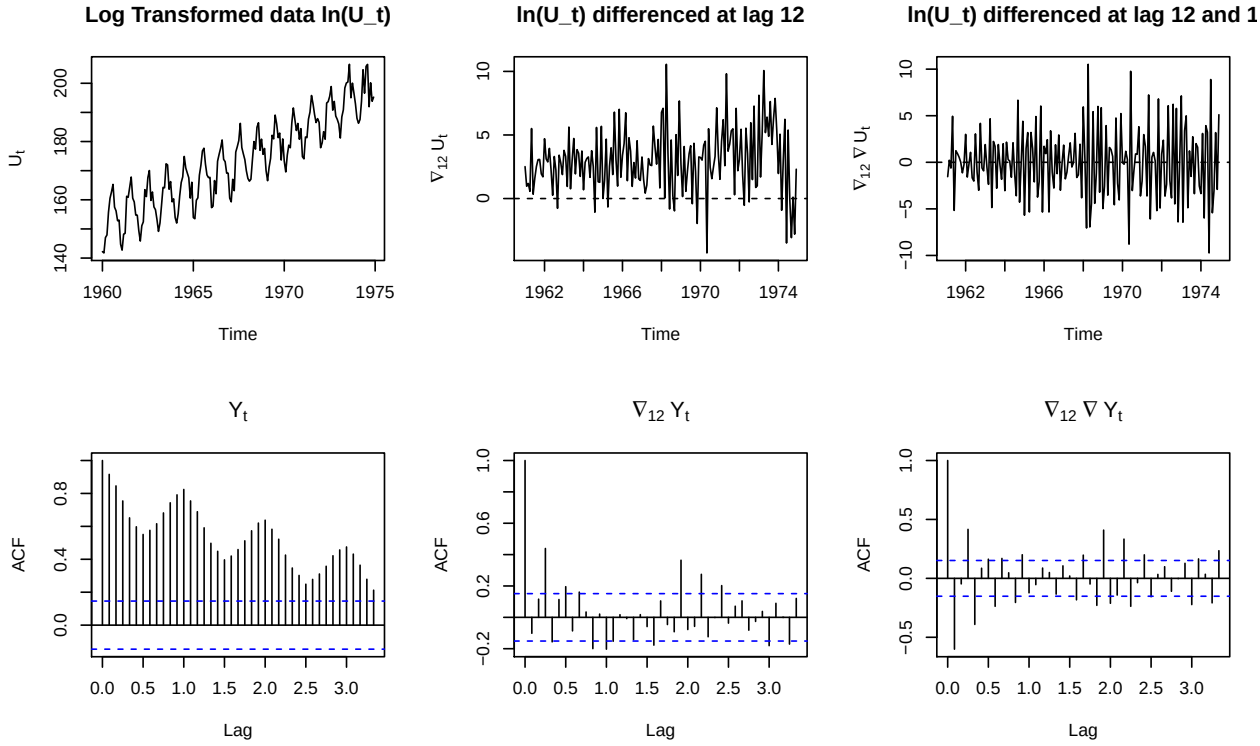
Both box-cox and log transformations successfully stabilize the variance and create better symmetry in the histograms. Viewing the plots, there is little difference between the two; however, the histogram of log-transformed data gives slightly more normal data, so I will choose the log transformation.

Next, I can address the seasonality and trend of the data.



The cyclical behaviour in the ACF of the transformed data suggests seasonality, and the jump in correlation magnitude every 12 lags suggests a period of 12.

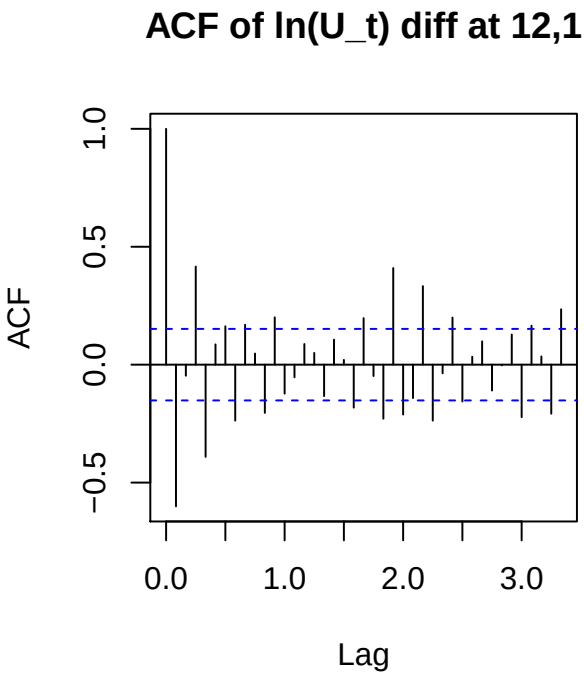
I will be difference at lag 12 to remove seasonality. Differencing at lag 1 will remove the upward linear trend of the data.

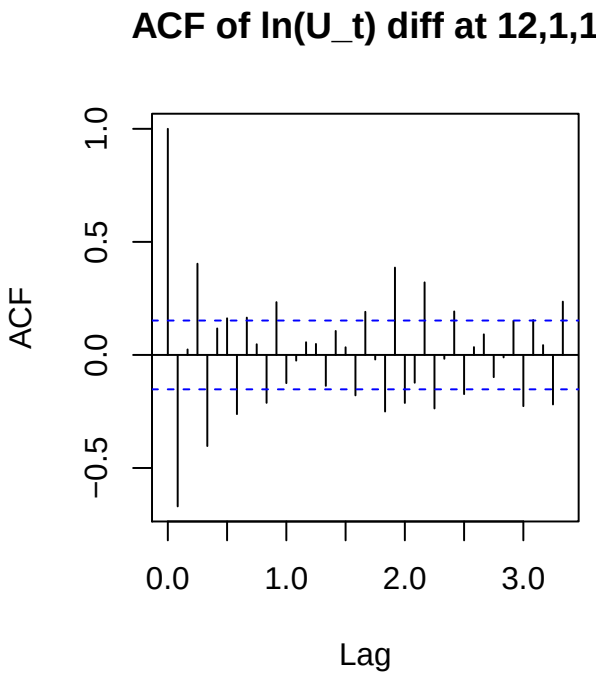


The difference at lag 12 successfully removes seasonality, but the plot still shows linear trend and the ACF

decays slowly which suggests non-stationarity. The additional difference at lag 1 removes the trend; however, the ACF still decays slowly so it may not be stationary.

I will attempt to difference once more at lag 1 to see if the lags die out quicker.





The differencing again at lag 1 did not help the ACF die out faster, so I will keep the original ln(U_t) differenced at lag 12 then 1.

My theory is that the data might visually look non-stationary, but might actually be stationary and sufficient for model building. I will use a unit root test, specifically the Augmented Dickey-Fuller test, to check for stationarity. In this test, the null hypothesis suggests the data is non-stationary.

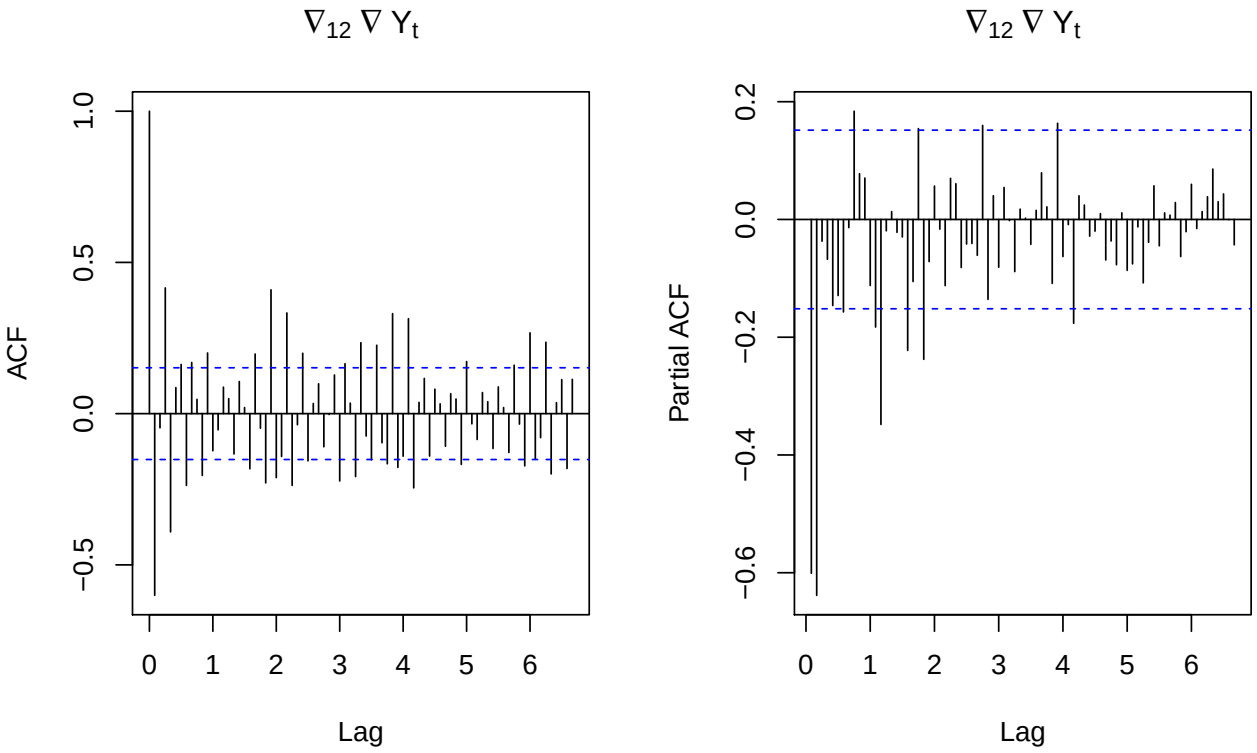
```
##
##  Augmented Dickey-Fuller Test
##
## data:  y1
## Dickey-Fuller = -7.9121, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

The test statistic is smaller than 0.05 so I can reject the null hypothesis. So I can conclude that the differenced data is stationary.

Model Identification

I will now look at the ACF and PACF plots find some candidate models.

```
# acf and pacf for differenced and transformed data
op <- par(mfrow = c(1,2))
acf(y1, lag.max=80, main="")
title(expression(nabla[12]~nabla~Y[t]))
pacf(y1, lag.max=80, main="")
title(expression(nabla[12]~nabla~Y[t]))
```



I applied one seasonal difference so D=1 at lag s=12. The PACF suggests P=4 and p=2. The ACF within periods seem to decay exponentially, a sign of AR for small p, q. Looking at the ACF, a Q=0,2,4,6 and q=0,1,3 seems like it could be a good fit. I will also try a non-seasonal AR(24). Some models not tested are because of R optimization errors.

```
##    P D Q , p d q  AICc
## A  4 1 2 , 2 1 0  -674.576
## B  4 1 2 , 2 1 1  -672.6677
## C  4 1 0 , 2 1 1  -662.1956
## D 24 0 0 , 0 0 0  -653.8955
```

Viewing the AICc of the 4 models, model A has the lowest AICc score and model B has the second lowest. I will consider both model A and B for forecasting.

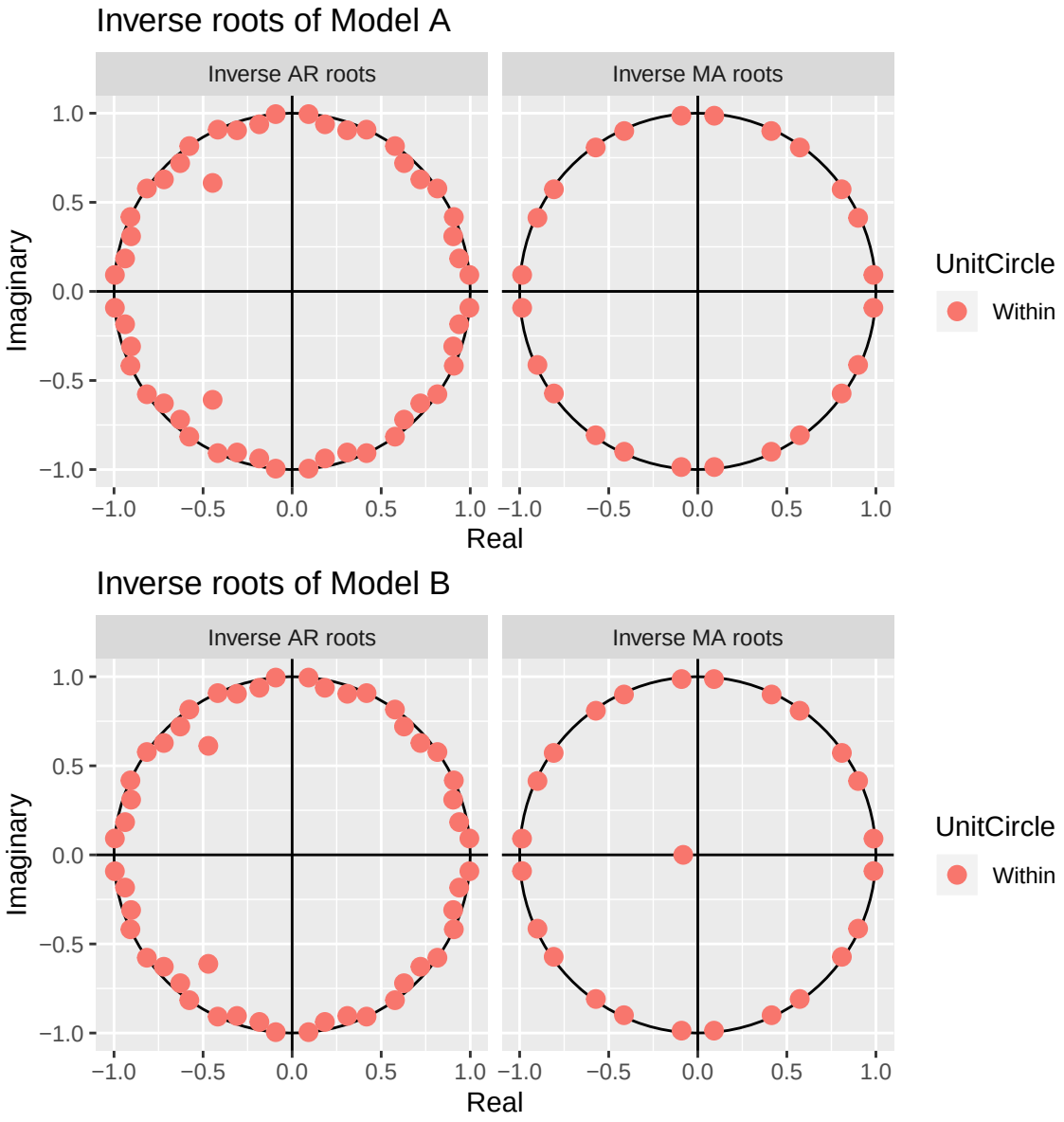
```
##
## Call:
## arima(x = data.log, order = c(2, 1, 0), seasonal = list(order = c(4, 1, 2),
##     period = 12), method = "ML")
##
## Coefficients:
##           ar1      ar2    sar1     sar2     sar3     sar4     sma1    sma2
##       -0.8928  -0.5698  0.0791  -0.6133  -0.4877  -0.3304  -0.7701  0.7897
## s.e.   0.0702   0.0696  0.2549   0.0878   0.0744   0.1502   0.3358  0.2575
##
## sigma^2 estimated as 0.0007451:  log likelihood = 346.71,  aic = -675.42
```

```
##
## Call:
## arima(x = data.log, order = c(2, 1, 1), seasonal = list(order = c(4, 1, 2),
##     period = 12), method = "ML")
```

```
##
## Coefficients:
##           ar1      ar2     ma1    sar1     sar2     sar3     sar4     sma1
##       -0.9410  -0.5956  0.0817  0.0958  -0.6175  -0.4789  -0.3309  -0.8048
## s.e.   0.1089   0.0794  0.1455  0.2713   0.0870   0.0781   0.1582   0.3713
##         sma2
##       0.8014
## s.e.  0.3009
##
## sigma^2 estimated as 0.0007356:  log likelihood = 346.86,  aic = -673.73
```

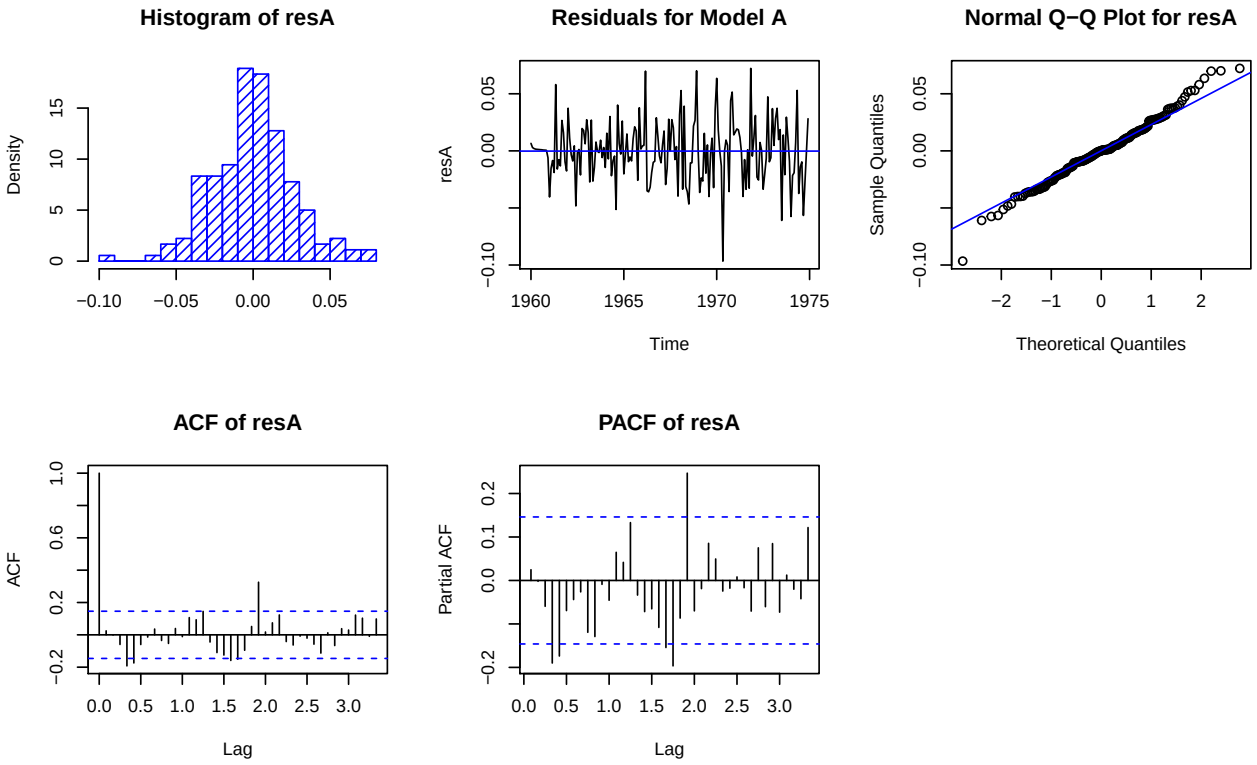
For model A, the sar1 coefficient is so small that it might not be significant. If 0 is in it's confidence interval, 0 would be a better coefficient, so I will make a new adjusted version of model A where that coefficient is 0. Same for model B, the ma1 and sar1 coefficients are small enough to justify replacing them with 0; however, R's optimization errors reject a new model where 0's replace small values so I will keep the original.

Now I will check for stationarity and invertibility by viewing the characteristic roots of Model A and B.



For a model to be stationary and invertible, all roots of the MA and AR part must exist outside the unit circle. Equivalently, all inverse of the roots must exist inside the unit circle; therefore, both models should be stationary and invertible.

Diagnostic checking for model A:



From the Residual Plot, there seems to be no trend and very little change of variance except for one spike after 1970. The QQ plot looks relatively normal except for one outlier in the beginning. The histogram also reflects this outlier but looks somewhat normal. The sample mean is close to 0 at -0.00009228181. The acf and pacfs look problematic as the acf has one spike at lag 23 and the pacf shows multiple values outside of the confidence interval, suggesting it does not be compared to a gaussian white noise process. I will try running Shapiro Wilk, Box Pierce, Ljung-Box, and Mc-Leod Li tests.

```
##
##  Shapiro-Wilk normality test
##
## data:  resA
## W = 0.98826, p-value = 0.1418
```

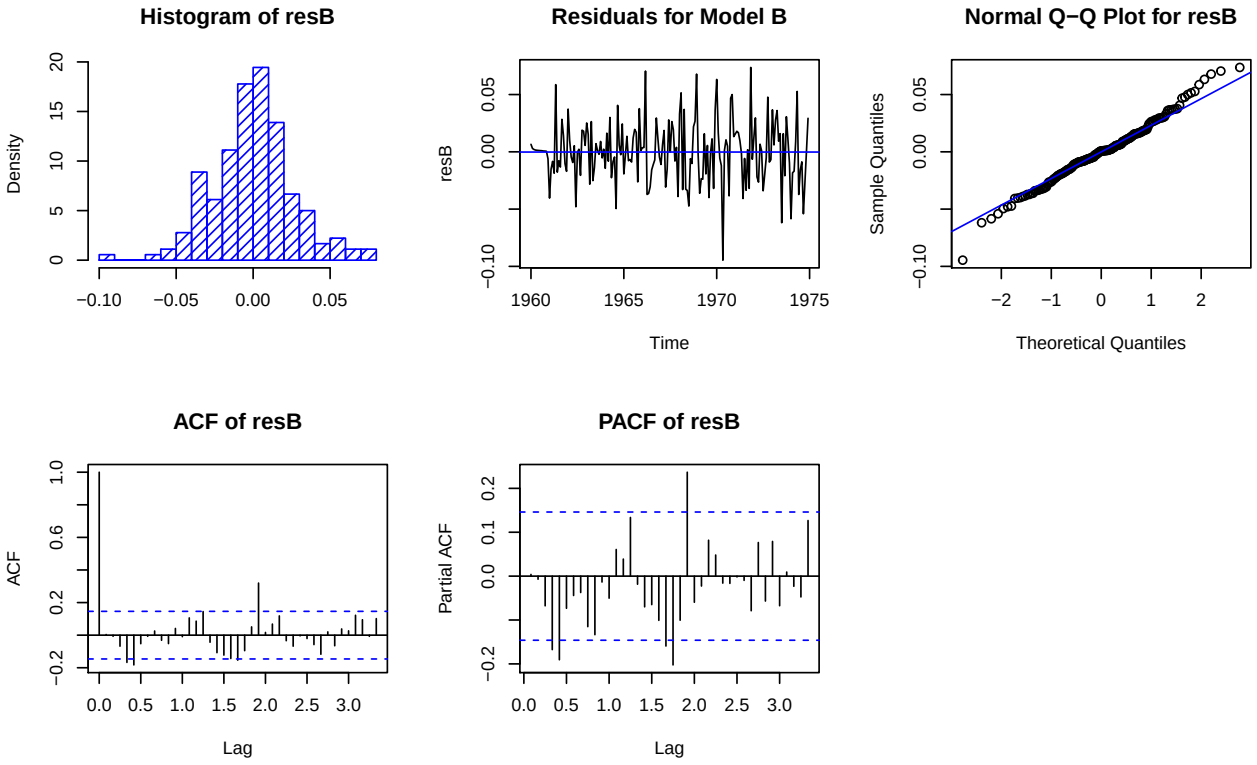
```
##
##  Box-Pierce test
##
## data:  resA
## X-squared = 14.772, df = 10, p-value = 0.1406
```

```
##
##  Box-Ljung test
##
## data:  resA
## X-squared = 15.354, df = 10, p-value = 0.1197
```

```
##
##  Box-Ljung test
##
## data:  resA^2
## X-squared = 14.462, df = 12, p-value = 0.2722
```

The Shapiro-Wilk tests for normality, the Box-Pierce and Ljung-Box tests if any ACF's are significantly different from zero, and the final McLeod Li tests for non-linear dependence of residuals. The model passed all tests by having p-values greater than 0.05 so I can consider it for forecasting.

Diagnostic checking for model B:



Like model A, model B residuals show no trend and very little if no change of variance. The QQ plot looks normal except for the one outlier and the histogram looks slightly more normal than model A. The sample mean is close to 0 at -0.0001817577 Also like model A, model B's ACF has one spike at lag 23 and it's PACF doesn't look much like white noise, so I will run a couple tests to determine if it can be used for forecasting.

```
##
##  Shapiro-Wilk normality test
##
## data:  resB
## W = 0.98855, p-value = 0.1548
```

```
##
##  Box-Pierce test
##
## data:  resB
## X-squared = 13.479, df = 10, p-value = 0.1981
```

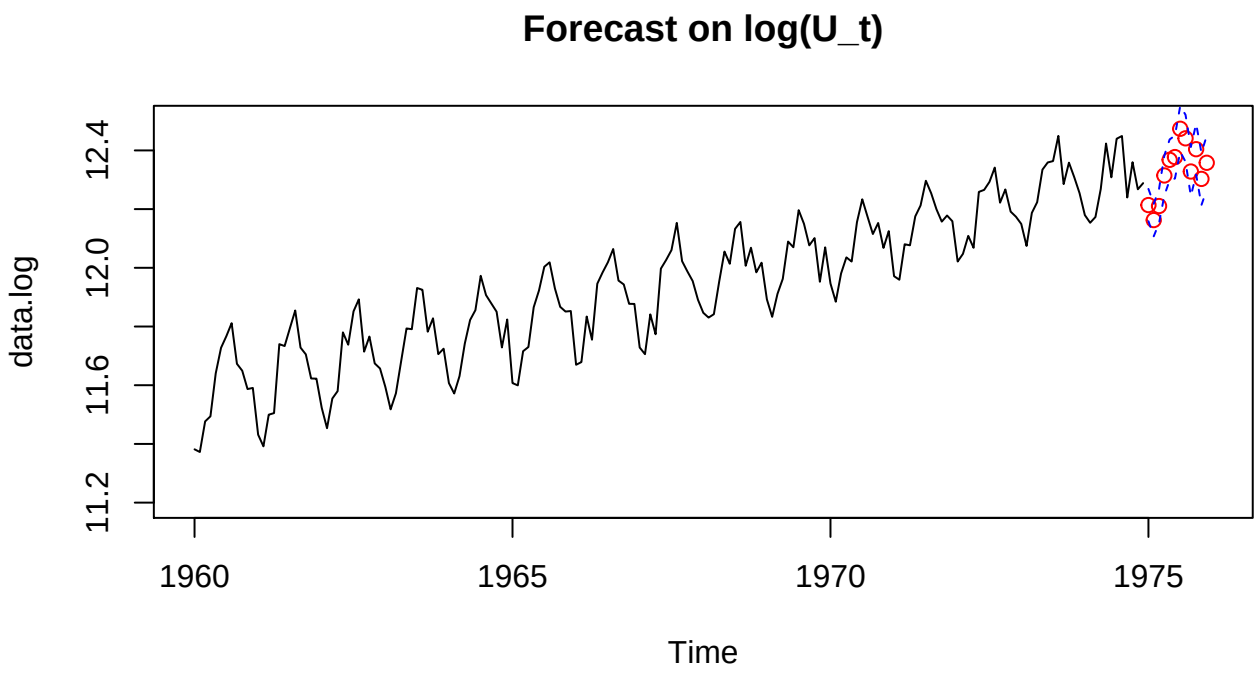
```
##
##  Box-Ljung test
```

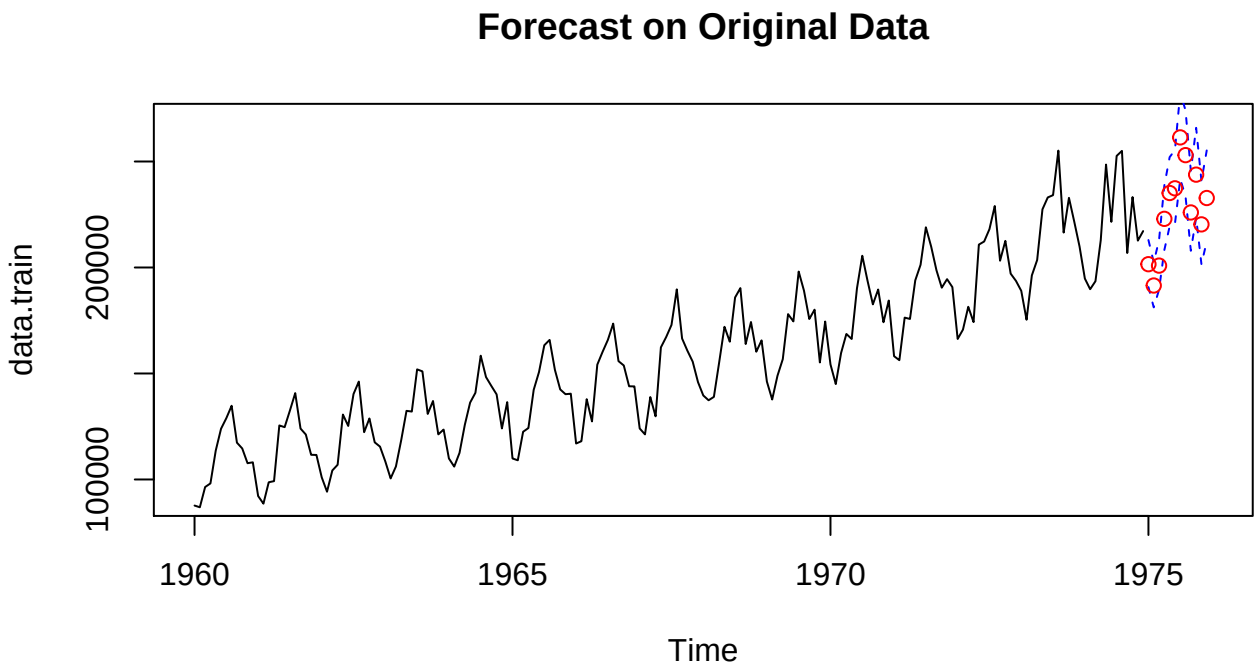
```
##
## data:  resB
## X-squared = 14.015, df = 10, p-value = 0.1723
```

```
##
##  Box-Ljung test
##
## data:  resB^2
## X-squared = 14.649, df = 12, p-value = 0.2612
```

The model passes all tests as well. Due to the principle of parsimony, model A being simpler than model B, and because model A had a slightly lower AICc score, I will choose model A for forecasting.

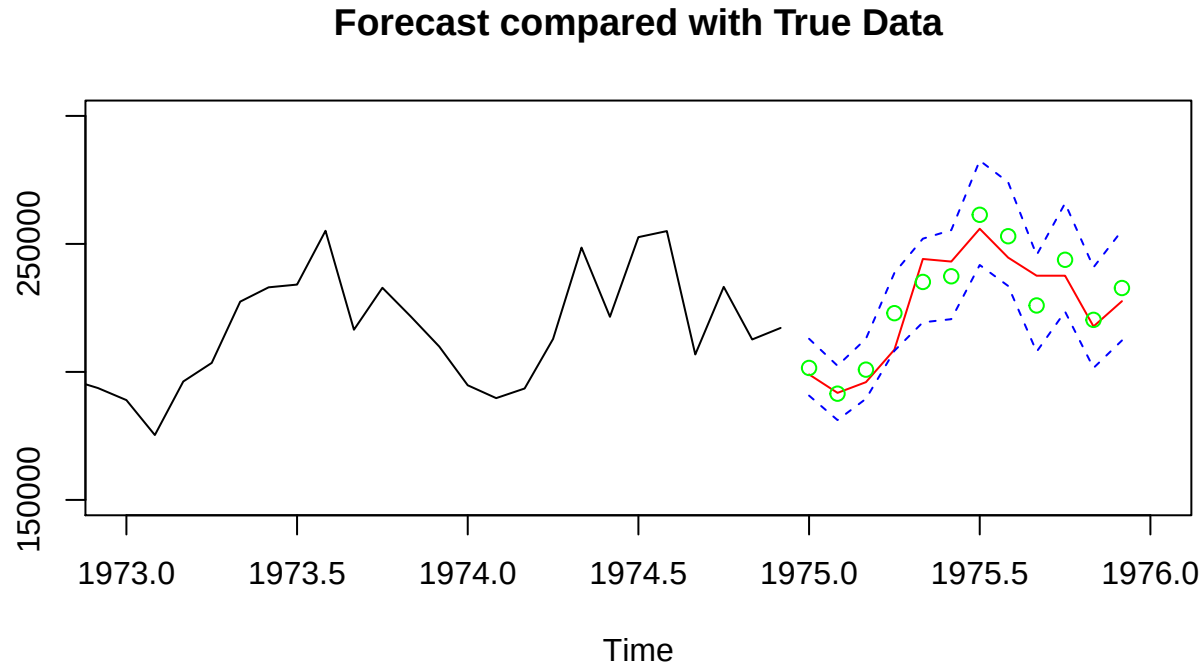
Forecasting





The red circles indicate forecasted demand of gasoline (in gallon millions) for the year of 1975 in Ontario, and the blue lines indicate the prediction interval.

Zooming in and adding in the test set:



The black line is the training set, and the red line is the test set, or true values for the year of 1975. The forecast was not perfect; however, the test set is still within the prediction intervals so model A is sufficient for finding future values of Ontario's gas demand.

Conclusion

The goal of this project was to create a (S)ARIMA model that sufficiently forecasts Ontario's gasoline demand starting Janurary 1975 to December 1975 using time series data from 1960 up until 1975.

Starting exploratory data analysis, the data was shown to have a significant linear trend and a seasonal component, as well as slight changes in variance. A log transformation successfully mitigated changes in variance. Differencing at lag 1 to removed linear trend and again at lag 12 removed a seasonal component with period 12. The remaining data looked reasonably like white noise; however, the ACF and PACF plots did not suggest stationary. The ACF did not decay quickly at all and was significant to a high lag. I tried differencing once more at lag 1, but it did very little change to the ACF. The data passed a Dickey-Fuller test for stationary, so I assume that it was passable for stationary data. This problem was likely due to the increase in change of variance near the end of the dataset.

The problematic ACF and PACF made model identification difficult. I made several estimates on the values of p, d, q and seasonal components but RStudio optimization errors restricted me from being able to extract the AICc score from every potential model. Therefore from a select 4 models I selected the 2 with the smallest AICc.

Checking unit roots, the models seemed to be stationary and invertible. The residual's plots, histogram, and qqplot seemed to look normal, but the ACFs did not. There were multiple values that looked significant. They both passed Shapiro Wilk, Box Pierce, Ljung-Box, and Mc-Leod Li tests, so I chose the model A according to the principle of parsimony and a slightly lower AICc score.

$FinalModel: (1-0.941_{(0.1089)}B-0.5956_{(0.0794)}B^2)(1+0.0958_{(0.2713)}-0.6175_{(0.087)}B^{12}-0.4789_{(0.0781)}B^{24}-0.3308_{(0.1582)})\ \Delta_{12}\Delta \ln(U_t) = (1+0.0817B_{(0.1455)})(1-0.8048_{(0.3713)}B^{12}-0.8014_{(0.3009)}B^{24})Z_t$

$\hat{\sigma}_z^2 = 0.0007356$

Finally, I used this model to forecast the demand of gasoline in Ontario during 1975, which I compared with the true values of demand. The forecasted values did not fit true data perfectly, but because the true values were within the prediction interval I believe it was satisfactory.

All in all, an adequate model was obtained based on past data using MLE estimation of parameters and AICc. It passed diagnostic checking, and the 1975 gasoline demand behaves as the past and therefore is described sufficiently by the model.

Appendix

```
# importing libraries and data source
#install.packages("devtools")
#devtools::install_github("FinYang/tsdl")
knitr::opts_chunk$set(fig.width=8, fig.height=5, fig.align="center")
options(warn=-1)
library(tsdl)
library(MASS)
library(qpcR)
library(forecast)
library(tseries)
library(urca)
# list view of all datasets with "Sales" attribute
tsdl_sales <- subset(tsdl, "Sales")

#for(i in 1:6)
#{
```

```r
#  if(length(tsdl_sales[[i]]) > 100)
#  {
#  cat("i =", i, attr(tsdl_sales[[i]], "description"), "  length = ", #length(tsdl_sales[[i]]), "\n")
#  }
#}
tsdl_sales <- subset(tsdl, "Sales")
data <- tsdl_sales[[1]]

# Leave 12 data points for model validation
train = data[c(1:180)]
test = data[c(181:192)]
data.train = ts(train, start=c(1960,1), frequency=12)
data.test = ts(test, start=c(1975,1), frequency=12)
op <- par(mfrow = c(2,1))
# Plot Original and Truncated Data
ts.plot(data, ylab="Demand", xlab="Time(Monthly)", main="Original Data");
ts.plot(data.train, ylab="Demand", xlab="Time(Monthly)", main="Truncated Original Data U_t")

# Add lines to identify mean, trend, and seasonality
abline(h=mean(data.train), col="blue")
fit <- lm(data.train ~ as.numeric(1:length(data.train)))
abline(fit, col="red")
abline(v=c(12,24,36), col="blue", lty=2)
# plot acf and histogram of truncated data
op <- par(mfrow = c(1,2))
acf(data.train,lag.max = 40, main = "ACF; U_t")
hist(data.train, xlab = "Demand", prob = TRUE, main = "Histogram; U_t")
# Box Cox Transformation
t <- 1:length(data.train)
fit <- lm(data.train ~ t)
bcTransform <- boxcox(data.train ~ t,plotit = TRUE)

lambda <- bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
data.bc <- (1/lambda)*(data.train^lambda-1)
# log transformed data
data.log = log(data.train)

# Plot original vs transformed
op <- par(mfrow = c(3,3))
ts.plot(data.train,main = "Original data U_t")
ts.plot(data.bc,main = "Box-Cox tranformed data bc(U_t)")
ts.plot(log(data.train), main="Log transformed data ln(U_t)")

# Plot transformed histogram
hist(data.train, breaks=10, col="light blue",xlab="data", main = "Histogram; U_t")
hist(data.bc, breaks=10, col="light blue",  main = "Histogram; bc(U_t)")
hist(data.log, breaks=10, col="light blue",  main = "Histogram; ln(U_t)", )
# acf and pacf of ln(U_t)
op <- par(mfrow = c(1,2))
acf(data.log,lag.max = 40,main = "")
pacf(data.log,lag.max = 40,main = "")
title("Log Transformed Data ln(U_t)", line = -1, outer=TRUE)
op <- par(mfrow = c(2,3))
```

```r
# ln(U_t) plot
ts.plot(data.bc, main = "Log Transformed data ln(U_t)", ylab=expression(U[t]))

# Difference at lag 12 to remove seasonal component
y12 <- diff(data.bc, 12)
ts.plot(y12,main = "ln(U_t) differenced at lag 12", ylab = expression(nabla[12]~U[t]))
abline(h = 0,lty = 2)

# Difference at lag 1 to remove trend component
y1 <- diff(y12, 1)
plot(y1, main = "ln(U_t) differenced at lag 12 and 1", type="l", ylab = expression(nabla[12]~nabla~U[t]))
abline(h = 0,lty = 2)

# ACF of ln(U_t)
acf(data.log,lag.max = 40,main=expression(Y[t]))

# ACF of ln(U_t) diff at lag 12
acf(y12,lag.max = 40, main=expression(nabla[12]~Y[t]))

# ACF of ln(U_t) diff at lag 12 then 1
acf(y1,lag.max = 40, main=expression(nabla[12]~nabla~Y[t]))

# acf comparison of diff at 12,1 and diff at 12,1,1
op <- par(mfrow = c(1,2))
y121 = diff(y1, 1)
acf(y1,lag.max = 40, main="ACF of ln(U_t) diff at 12,1")
acf(y121, lag.max=40, main="ACF of ln(U_t) diff at 12,1,1")
# dickey fuller test to check for non-stationarity
adf.test(y1, alternative = c("stationary", "explosive"),
         k = trunc((length(y1)-1)^(1/3)))
# acf and pacf for differenced and transformed data
op <- par(mfrow = c(1,2))
acf(y1, lag.max=80, main="")
title(expression(nabla[12]~nabla~Y[t]))
pacf(y1, lag.max=80, main="")
title(expression(nabla[12]~nabla~Y[t]))
# creating models from estimated p d q
modela <- arima(data.log, order=c(2,1,0), seasonal = list(order = c(4,1,2), period = 12), method="ML")
modelb <- arima(data.log, order=c(2,1,1), seasonal = list(order = c(4,1,2), period = 12), method="ML")
modelc <- arima(data.log, order=c(2,1,1), seasonal = list(order = c(4,1,0), period = 12), method="ML")
modeld <- arima(data.log, order=c(24,0,0), method="ML")
# list choice of p d q and aic score
cat("   P D Q , p d q  AICc\n")
cat("A  4 1 2 , 2 1 0 ", AICc(modela), "\n")
cat("B  4 1 2 , 2 1 1 ", AICc(modelb), "\n")
cat("C  4 1 0 , 2 1 1 ", AICc(modelc), "\n")
cat("D 24 0 0 , 0 0 0 ", AICc(modeld), "\n")

# View coefficients of model a and b
modela
modelb
# Check roots of model a and b
autoplot(modela, main="Inverse roots of Model A")
```

```r
autoplot(modelb, main="Inverse roots of Model B")
# Diagnostic checking for model A
resA <- residuals(modela)

op <- par(mfrow = c(2,3))
hist(resA, density=20, breaks=20, col="blue", xlab="", prob=TRUE, main="Histogram of resA")

plot.ts(resA, main="Residuals for Model A")
abline(h=mean(resA), col="blue")

qqnorm(resA, main="Normal Q-Q Plot for resA")
qqline(resA, col="blue")

acf(resA, lag.max=40, main="")
title("ACF of resA")
pacf(resA, lag.max=40, main="")
title("PACF of resA")
# Shapiro Wilk, Box Pierce, Ljung-Box, and Mc-Leod Li tests for model 3
shapiro.test(resA)
Box.test(resA, lag=12, type=c("Box-Pierce"), fitdf=2)
Box.test(resA, lag=12, type=c("Ljung-Box"), fitdf=2)
Box.test(resA^2, lag=12, type=c("Ljung-Box"), fitdf=0)
# Diagnostic checking for model B
resB <- residuals(modelb)

op <- par(mfrow = c(2,3))
hist(resB, density=20, breaks=20, col="blue", xlab="", prob=TRUE, main="Histogram of resB")

plot.ts(resB, main="Residuals for Model B")
abline(h=mean(resB), col="blue")

qqnorm(resB, main="Normal Q-Q Plot for resB")
qqline(resB, col="blue")

acf(resB, lag.max=40, main="")
title("ACF of resB")
pacf(resB, lag.max=40, main="")
title("PACF of resB")
# Shapiro Wilk, Box Pierce, Ljung-Box, and Mc-Leod Li tests for model B
shapiro.test(resB)
Box.test(resB, lag=12, type=c("Box-Pierce"), fitdf=2)
Box.test(resB, lag=12, type=c("Ljung-Box"), fitdf=2)
Box.test(resB^2, lag=12, type=c("Ljung-Box"), fitdf=0)

# Graph of forecast on transformed data
pred.tr <- predict(modela, n.ahead= 12)
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
ts.plot(data.log,  ylim=c(11.2,12.5), xlim=c(1960, 1976))
points(pred.tr$pred, col="red")
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
title("Forecast on log(U_t)")
```

```r
# Graph of forecast on original data
pred.orig <- exp(pred.tr$pred)
U= exp(U.tr)
L= exp(L.tr)
ts.plot(data.train, xlim=c(1960,1976), ylim=c(90000,270000))
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points(pred.orig, col="red")
title("Forecast on Original Data")
ts.plot(data.train, data.test, xlim= c(1973,1976), ylim= c(150000,300000), col=c("black", "red"))
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points(pred.orig, col="green")
title("Forecast compared with True Data")
```