Monthly Employees Demand: Time series Analysis report

Zheyue Wang - PSTAT 174-final project defense: 2:00 – 2:30pm

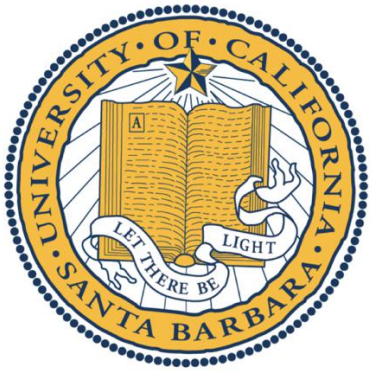
Mar 20 2010

e-mail: zheyue@ucsb.edu

2020-3-14

UCSB

[公司地址]



Contents

1 **Abstract** 3

2 **Introduction** 3

3 **Data Exploratory Analysis** 4

3.1 Preliminary Data Exploration 4

3.2 Decomposition Model 4

4 Data Transformation 5

4.1 Stabilize Variance of the Monthly employee 5

4.2 Remove Seasonality and Trend 7

5 Model Identification and Estimation 10

5.1 Preliminary Model Identification 10

5.2 Model Selection 12

5.3 Model Estimation 12

6. Diagnostics 19

6.1 Normality Checking…… 20

6.2 Independence Checking……21

6.3 Constant Variance Checking ……21

7.Forcasting 22

8 Conclusion 25

9 Reference 25

10 Appendix 26

1 Abstract

For this project, I built the SARIMA model for the monthly demand of employee from 1960 to 1975 to forecasting the amount of employee needed during the 12 month. In order to transform the original data to get more normality and best build the model with low variance, I use power of 1/5 for the lambda on transformation. Observed the lag with ACF and PACF, I selected a general SARIMA model. According to AIC score I selected the best p and q with the smallest AIC. After calculating the confident interval, I selected the best model to do the forecasting. Besides, I used diagnostic check to check whether the model is the best. In the end, I determine the best model is SARIMA (0,1,0) x (0,1,1).

2 Introduction

Number of employee demand per month for a company is an important data because it indicated how many more employee the company need to hire. Therefore, it's important to examine the trend and predict its future demand. The data I have found has provided its past demand of monthly employee in Winsock from 1960 until 1975. I choose this data because it shows a typical seasonal component and linear trend. Plotting the dataset, I can see the obviously upward linear trend indicating the increasing demand of employee, as well as a seasonal pattern. This is reasonable as

some month is highly demand for employee. After differencing at lag 12 and lag1, I find that the variance of this modified data gets smaller as compared to the original data. Time series techniques enable me to make this data stationary and allow me to fit it into a desirable model that can forecast the future demand of monthly employee. For example, I identify several possible SARIMA models based on ACF and PACF plot of the data, and further narrow the choice to use the AICc model selection criteria. Parameter estimation is also done using the MLE method. After running diagnostic checks such as the Ljung-Box Test, the Ljung-Box square Test, the Box-Pierce Test, the Shapiro Wilk Test and many more on the residuals of a model, and considering the AICc value, I decide the final model to be SARIMA$(0,1,0) \times (0,1,1)12$. Finally, I forecast the monthly demand of employee up to 12 months ahead and compare them with the true value. My result shows that all the predicted value falls within the 95% confidence interval, proving the feasibility of my final model

3 Data Exploratory Analysis

3.1 Preliminary Data Exploration

The dataset I used includes 2 variables: date on the monthly bases and monthly employee demand. There are 178 observation in total. I fist separate the training data by drop out the last 12 data, and those 12 data that I drop treat as my test dataset. I compare the forecasting result with the real data to obtain the accuracy of our prediction. Thus, there are total 166 training data and 12 test data.

As the preliminary data exploration, I save the training data in time series form and plot it with all 166 observation.

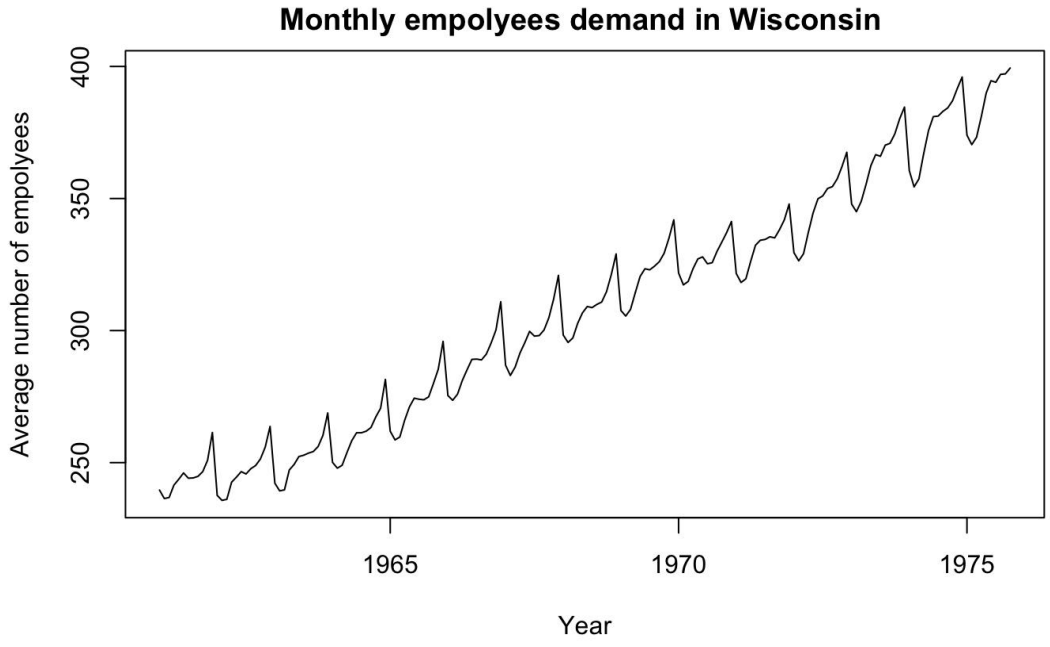


Figure 3.1

From the plot, I observe that there is upward trend of it. Besides, there is repeat pattern on a roughly fix time interval, looks like it repeats every 12 months, which is 1 year. Thus, I potentially conclude that there is also has seasonality in the plot.

3.2 Decomposition Model

I also decompose the original training data to obtain a classical decomposition model as $Y_t = m_t + s_t + S_t$, where Y_t is the original training data, m_t is the trend component, s_t is the seasonal component and S_t is a stationary process. The decompose graph is show below in figure 3.2:

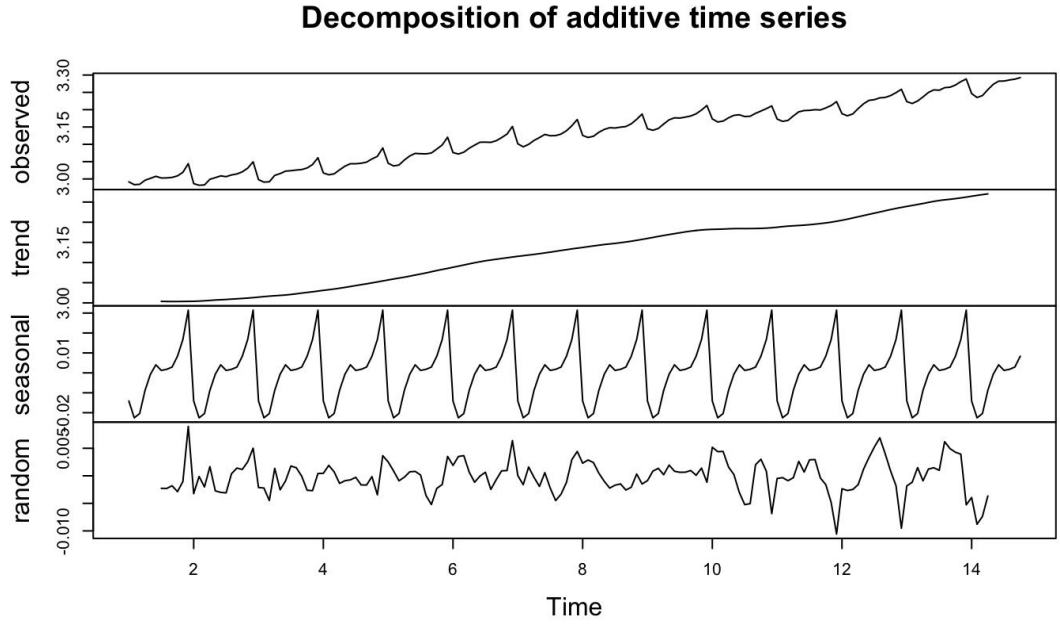


Figure 3.2

Figure 3.2 clearly shows seasonality composed with a high spike and a low spike on a yearly basic. Besides, it also has upward linear trend. Hence, Y_t (the original training dataset) is not stationary, and I have to transform and differentiate it to make

stationary. The reason why I cannot directly use Y_t (the original training data) is that the model I will build is based on stationary assumption.

4 Data Transformation

To make Y_t stationary, I decided to transform and difference it to do the future analysis. For transformation step, I need to stabilize the variance of Y_t and also remove the seasonality and linear trend.

4.1 Stabilize Variance of Y_t (the original training dataset)

I first use Box-cox transformation to decide the best transformation for Y_t. I plot the Box-Cox as in figure 4.1 in terms of lambda and observe that 0 is inside the 95% confident interval.

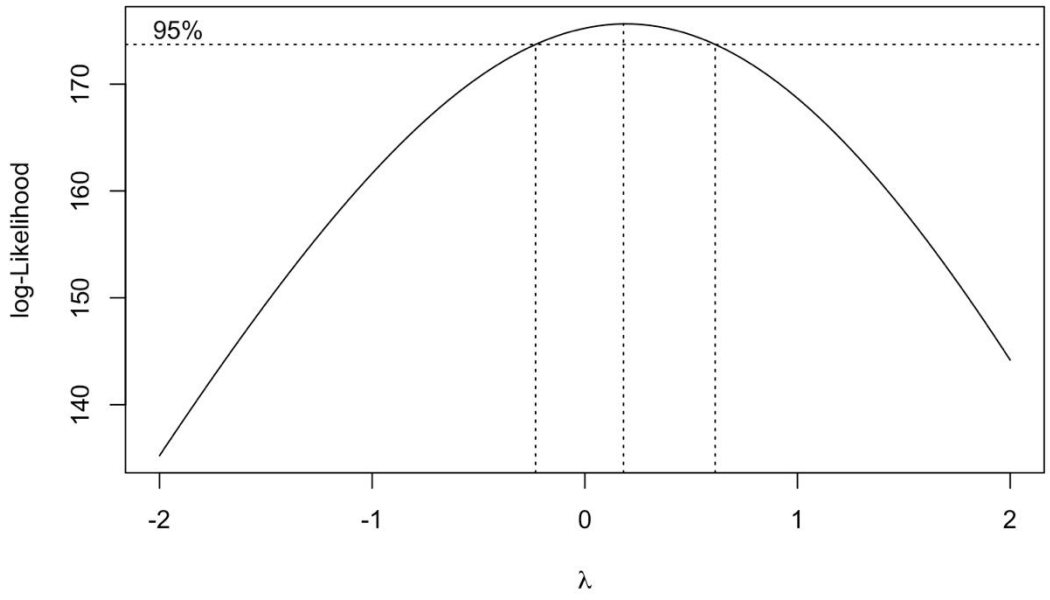


Figure 4.1

In this case, I decide to try other transformation such as log transformation and square root transformation. By computing the exacting lambda on R, I get lambda is 0.1818182. Since 1/5 = 0.20, and 0.1818182 is nearby the 0.20, so I use the 5^{th} root transformation for Y_t to make more accurate model. Therefore, I have the 5^{th} root transformed training data as:

$$V_t = Y_t^{0.2}$$ where Y_t is the original training data

The plot of ACF and PACF for V_t (the 5^{th} root transformed training data) indicated that my V_t is still non-stationary:

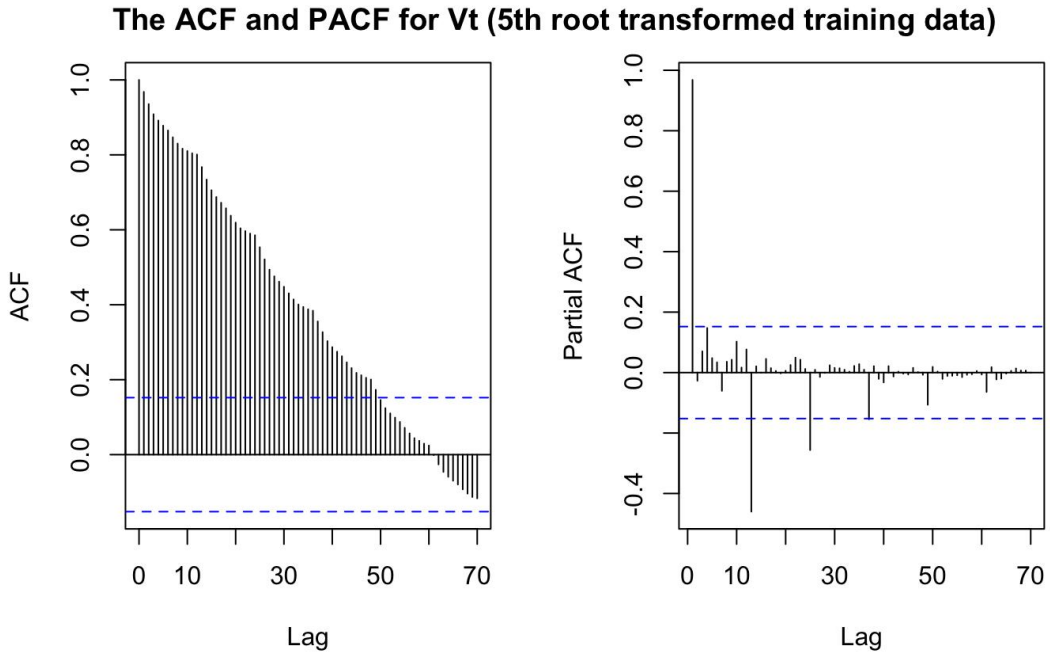


figure 4.1 (b)

figure 4.1 (b) shows that V_t (the 5^{th} root transformed training data) has stronger linear trend and seasonality

4.2 Remove seasonality and Trend

I use differencing method to remove the seasonality and trend. Figure 4.1(b) appear that V_t (my 5^{th} root transformed training data) is repeat a same patter for every 12 lags, so it is reasonable to assign the period of 12 towards the seasonal component to remove the seasonality of

period 12. I zoom up the graph and look closely:

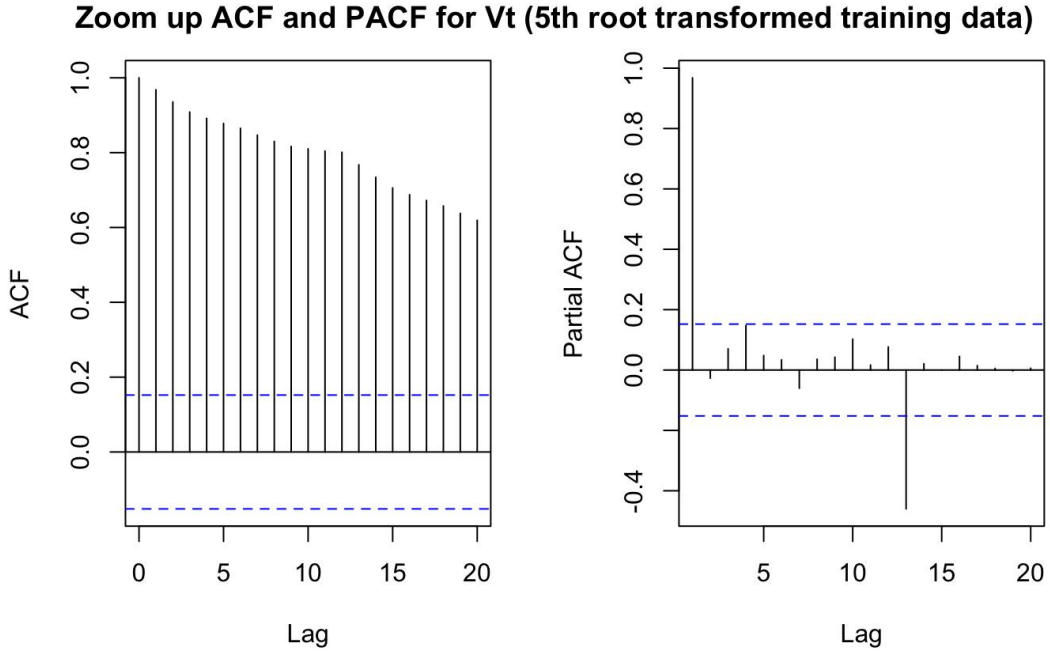


Figure 4.2(a)

I count on both ACF and PACF, and both repeat a pattern for every 12 lags. Thus, I apply differencing at lag s to Remove Seasonality:

$$W_t = \nabla_s V_t = V_t - V_{t-s} = (1 - B^s)V_t \quad and \ s = 12$$

$$so \quad we\ have \quad W_t = \nabla_{12} V_t = V_t - V_{t-12}$$

Figure 4.2 (b) is the ∇_{12} differenced data and variance is 9.90796e-05. I observe an obviously upward trend line on the graph.

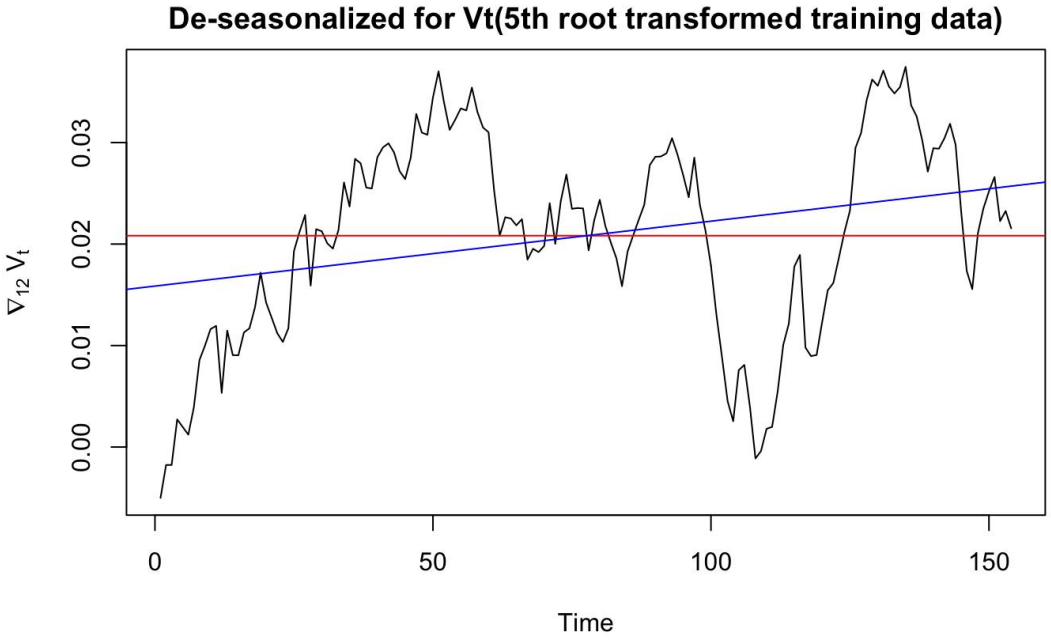


Figure 4.2 (b)

The red line is the mean and the blue line is the trend.

After differentiating at lag 12, the seasonal are removed but still have trend. To remove the trend, I elimination of trend by differencing at lag 1:

$$Therefore, \ W_t = \nabla(\nabla_{12} V_t)$$

Figure 4.2 (c) graph shows blow.

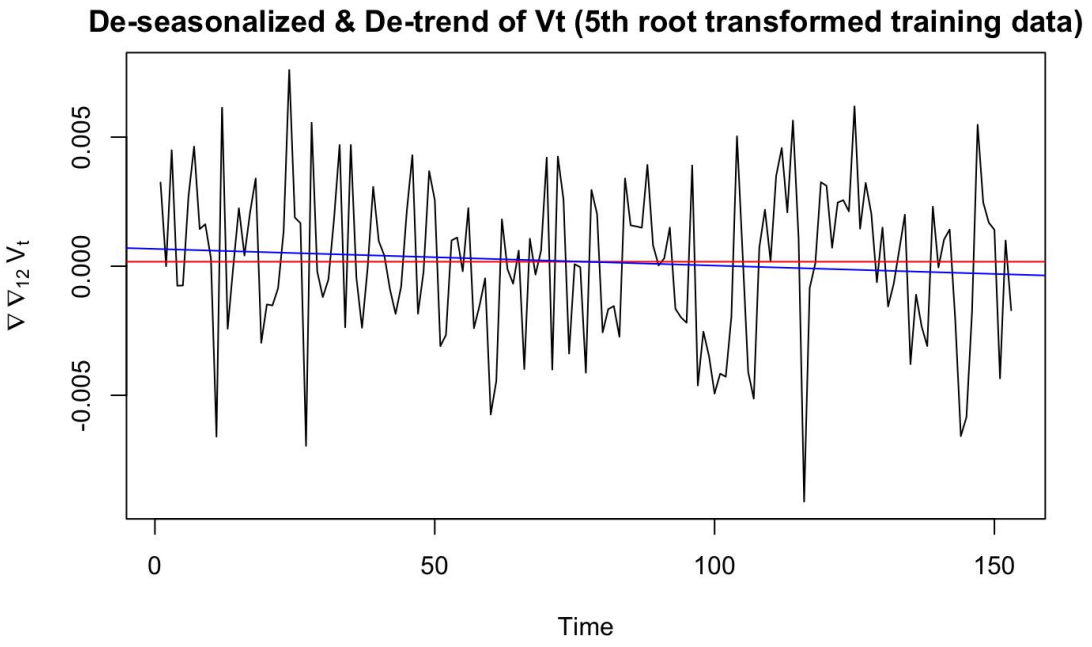


Figure 4.2 (c)

The red line is the mean and the blue line is the trend. The trend is slightly downward but is very close to the mean.

I obtain the variance of 8.967057e-06 and the trend become more horizontal after removing the trend at lag 1. If I remove at lag 1 once more, the variance increases to 1.588049e-05, which is larger than previous one. Therefore, difference twice at lag1 is over-differencing, so I only need to difference once to remove the trend as the final process below shows:

$$X_t = \nabla(\nabla_{12} V_t)$$

5 Model Identification and Estimation

According to the seasonal data of monthly employee demand, an appropriate SARIMA model will be adopted to analyze the data. The proper structure of SARIMA model is as below:

$$SARIMA(p, d, q) \times (P, D, Q)_s$$

where p = the order of non-seasonal AR process, d = non-seasonal differencing, q = the order of non- seasonal MA process, P = the order of seasonal AR process, D = seasonal differencing, P = the order of seasonal AR process, Q = the order of seasonal MA process and s = the period of the time lag.

In my specific case of the monthly demand of employee data, I have s = 12. Following the procedure to identify SARIMA and based on the differencing procedure in the previous step, I have D = 1 and d = 1 since I difference the V_t (5th root transformed training data) at lag 12 to remove seasonality and then differenced again at lag 1 to remove trend. Next, I need to find p, q and P, Q based on the ACF and PACF plots for the preliminary model identification

5.1 Preliminary Model Identification

I plot the ACF and PACF of stationary time series X_t as $X_t = \nabla(\nabla_{12} V_t)$ in order to identify seasonal terms P and Q. Since seasonal component s is 12 so we need to check value of ACF and PACF at seasonal lags 12, 24,36… Look back for my model,

I find that for the seasonal terms P and Q, ACF and PACF plot cuts off after lag 12, so Q = 1 and P = 1.

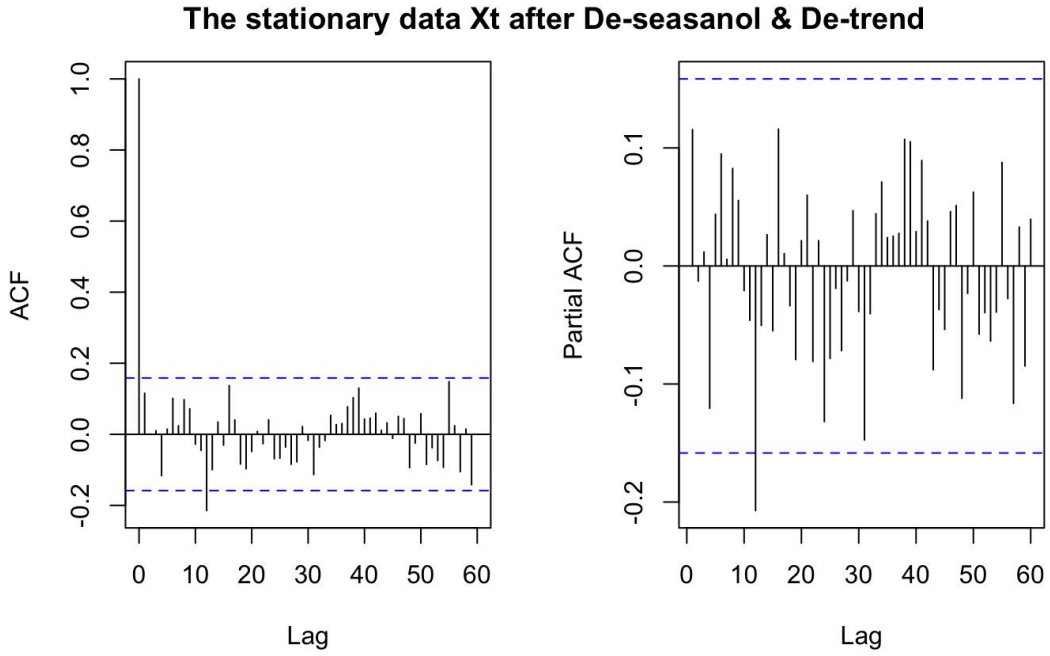


Figure 5.1(a)

Let's zoom up and observe more closely:

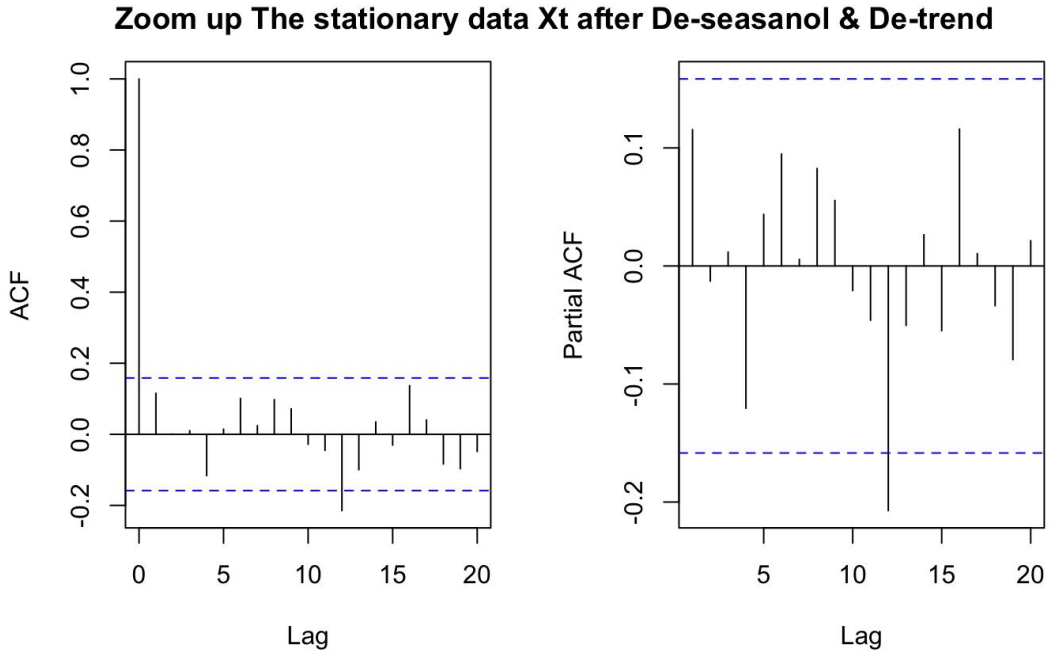


Figure 5.1(b)

Then based on figure 5.1(b) above I find p, q by looking at ACF and PACF plots of the zoomed plot at lag 1, 2, 3, 4…

I can see that q is 0 and p is also 0. Hence, I suggest that p = 0, which indicate AR (0) and q = 0, indicating MA (0) for ARMA (p, q) model. Thus, we can conclude that p, q both take value in 0. Fixed the seasonal terms P and Q, I test all the combinations of non-seasonal terms p, q between 0 to 1. The reason I change my range from 0 to 1 is because I want to know whether my AICc will decreasing if p and q increase. Thus, I have 4 models to consider.

5.2 Model Selection Through Criterion Model Fitting

I use Akaike's information criterion (AICc) to select the best model.

P	Q	AICc
AR (0)	MA (1)	-1350.208
AR (1)	MA (0)	-1347.908
AR (1)	MA (1)	-1347.120
AR (0)	MA (0)	-1342.450

From the AICc score table, I choose three possible model as my candidate models. There are:

1 MODEL I: $SARIMA(0,1,1) \times (1,1,1)_{12}$ (p = 0, q = 1)

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

3 MODEL III: $SARIMA(1,1,1) \times (1,1,1)_{12}$ (p = 1, q = 1)

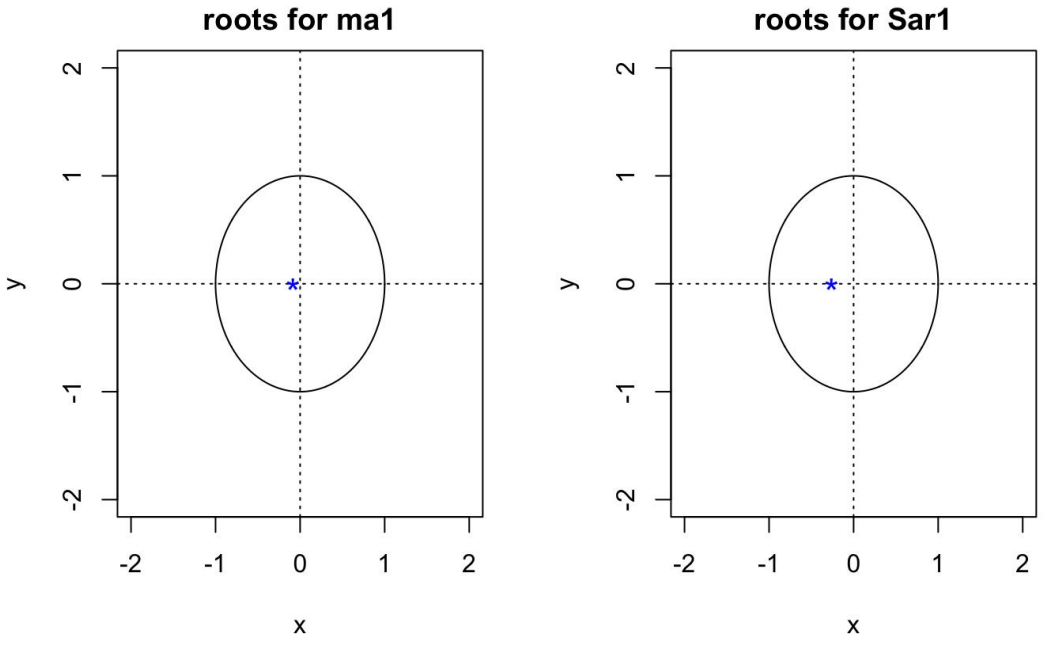
5.3 Model Estimation

Estimate MODEL I: $SARIMA(0,1,1) \times (1,1,1)_{12}$. aic = -1349.32

I use the MLE method to estimate the coefficients of our model, and the results are showed in the following table:

MA (1)	SAR (1)	SMA (1)
0.0815	0.2505	-05402

Plot the root to check for invertible and casual:



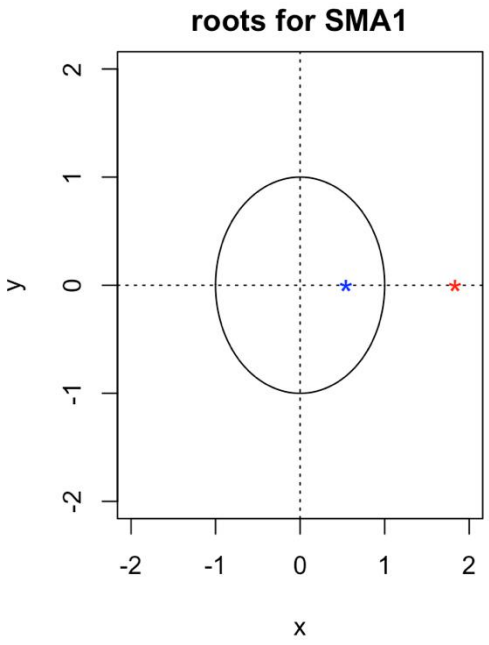


Figure 5.3(a) the root of Model I

The root of are below:

[1] –12.26994+0i

[1] –3.992016+0i

[1] 1.851166+0i

I notice that all integer root is outside of unit root, which means that my model is invertible and casual.

Calculate the confident interval to make batter model:

ma1: 0.0815-1.96*0.0810 = -0.07726, and 0.0815+1.96*0.0810 = 0.24026

sar1: 0.2505+1.96*0.2206 = 0.682876, and 0.2505-1.96*0.2206 = -0.181876

sma1: -0.5402-1.96*0.1913 = -0.915148, and -0.5402+1.96*0.1913 = -0.165252

for MA (1): since 0 is within the CI (-0. 07726, 0. 24026), so I set q = 0

for SAR (1): since 0 is within the CI (-0.181876, 0. 682876), so I set P = 0

for SMA (1): since 0 is not with the CI (-0.915148, -0.165252), so I keep this value

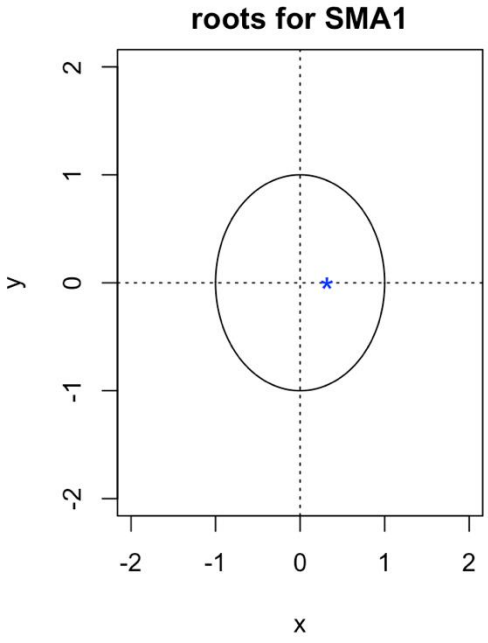
Therefore, I fixed Model I and find the coefficient with smallest AIC based on Model I:

MA (1)	SAR (1)	SMA (1)
0	0	-0.3259

aic = -1351.25 and sigma^2 estimated as 8.257e−06

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

The fixed model I is SARIMA (0,1, 0) x (0,1,1)s with s = 12. And aic is -1351.25. Since fixed model I has lower aic value, so I prefer fixed model I. Also, I need to check the invertible and casual for my fixed model I:



[1] 3.075031+0i

Obviously that the integer part is outside the unit circle, so my fix model I is invertible and casual.

Therefore, for my fixed model I *recall that* $X_t = \nabla(\nabla_{12} Y_t^{0.2})$:

Fixed Model 1: $SARIMA(0,1,0) \times (0,1,1)_{12}$

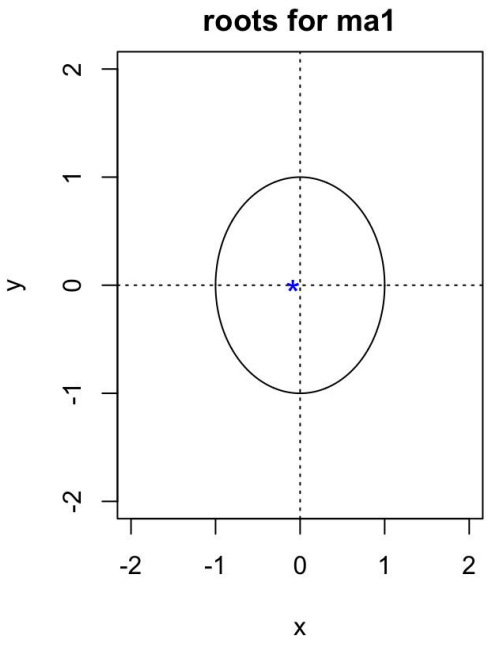
$$X_t = (1 - 0.3259B^{12})Z_t \quad \text{Where } Z_t \sim N(0, 8.141e\text{-}06)$$

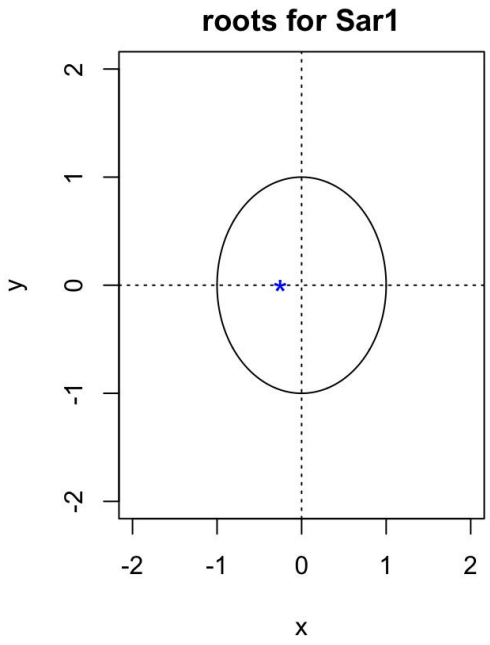
Estimate MODEL II: $SARIMA(1,1,0) \times (1,1,1)_{12}$. aic = –1349.35

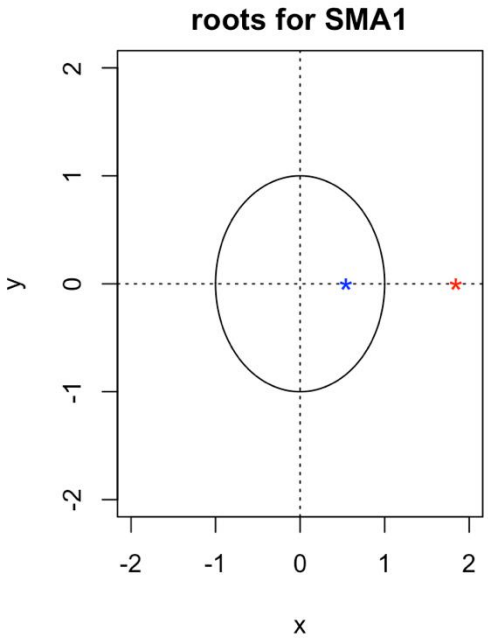
I use the MLE method to estimate the coefficients of my model, and the results are showed in the following table:

MA (1)	SAR (1)	SMA (1)
0.0834	0.252	-0.5419

Plot the root to check for stationary:







Below is the result of root

[1] –11.99041+0i

[1] –3.968254+0i

[1] 1.845359+0i

Calculate the confident interval to make batter model:

AR (1): 0.0834 + 1.96 * 0.0816 = 0.243336

0.0834 - 1.96 * 0.0816 = -0.076536 so 0 in within the CI set to 0

SAR (1): 0.252 +1.96 * 0.220 = 0.6832

0.252 -1.96 * 0.220 = -0.1792 so 0 is within the CI. set to 0

SMA (1): -0.5419 + 1.96 * 0.1903 = -0.168912

-0.5419 - 1.96 * 0.1903 = -0.914888 so 0 is not within the CI

Therefore, I fixed Model II and find the coefficient with smallest AIC based on Model II:

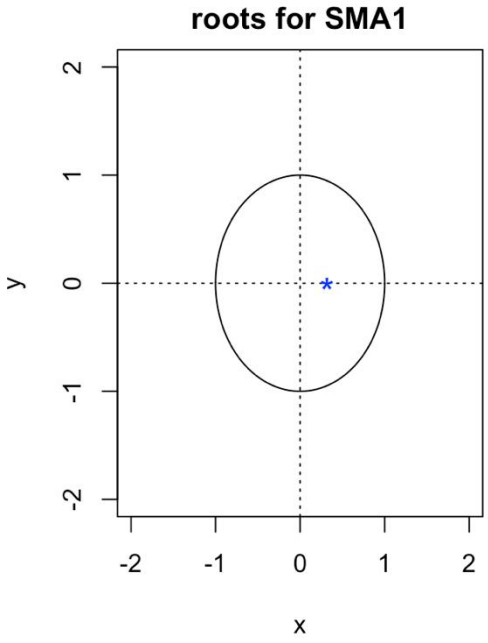
MA (1)	SAR (1)	SMA (1)
0	0	-0.3259

sigma^2 estimated as 8.257e-06 and aic = -1351.25

The fixed model II has aic = -1351.25. Since the fixed model II has lower aic value, so I prefer to choose the fixed model II.

Fixed MODEL II: $SARIMA(0,1,0) \times (0,1,1)_{12}$. aic = -1351.25

Also, I need to check the invertible and causal for my fixed model II:



Below is the root

[1] 3.068426+0i

Therefore, for my fixed model II recall that $X_t = \nabla(\nabla_{12} Y_t^{0.2})$

$$Fixed\ Model\ II: SARIMA(0,1,0) \times (0,1,1)_{12}$$
$$X_t = (1 - 0.3259B^{12})Z_t$$

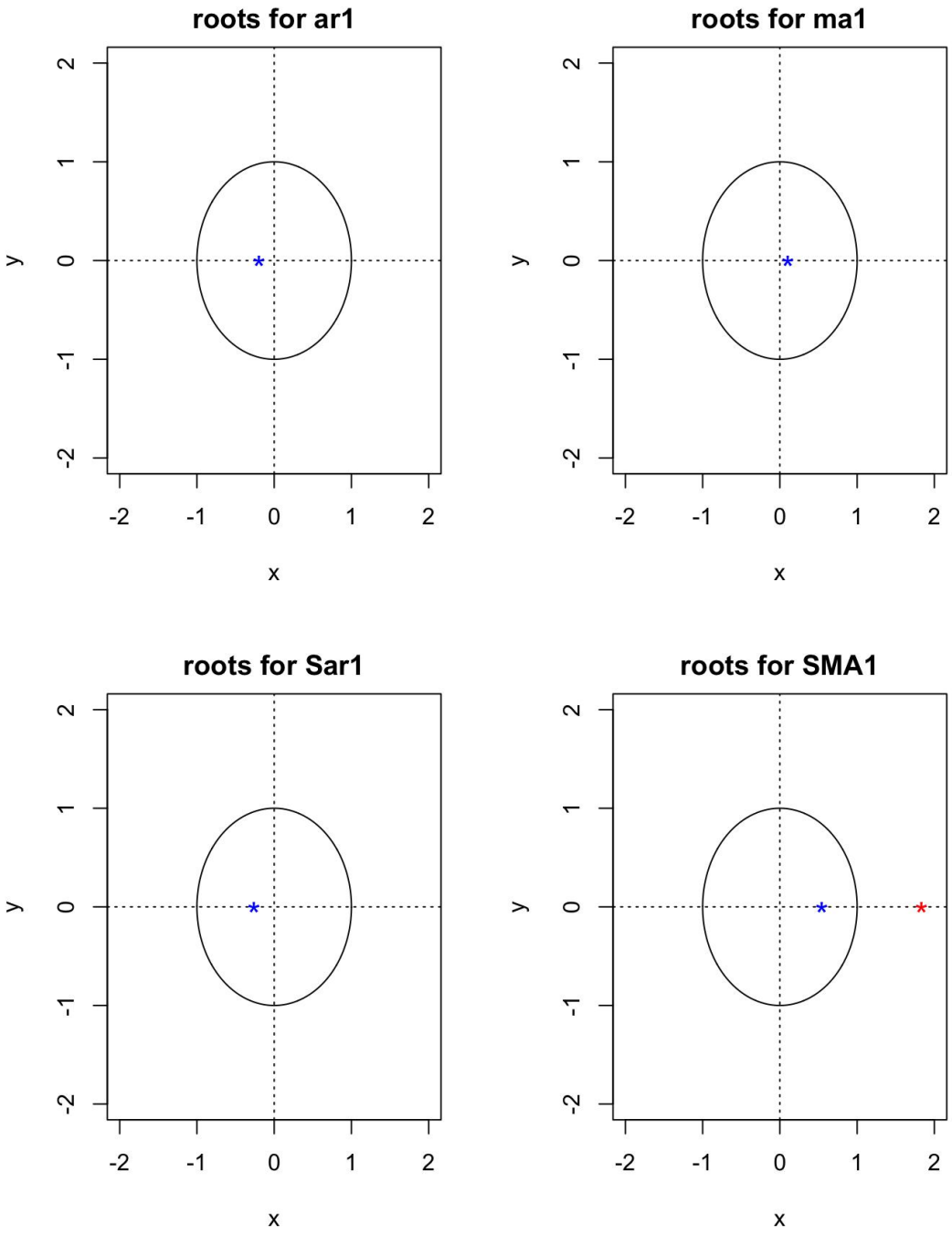
Where $Z_t \sim N(0, 8.257e\text{-}06)$

Estimate MODEL III: $SARIMA(1,1,1) \times (1,1,1)_{12}$. aic = -1347.37

We use the MLE method to estimate the coefficients of our model, and the results are showed in the following table:

AR (1)	MA (1)	SAR (1)	SMA (1)
0.1883	-0.1057	0. 2547	-0.5449

Plot the root to check for stationary:



Below is the result of root:

[1] −5.310674+0i

[1] 9.460738+0i

[1] −3.926188+0i

[1] 1.835199+0i

Check the CI for Model III:

For AR (1): 0.1883 + 1.96 * 0.8685 = 1.89056

0.1883 - 1.96 * 0.8685 = -1.51396 so 0 is within the CI and setting this with 0

MA (1): -0.1057 + 1.96 * 0.8778 = 1.614788

-0.1057 - 1.96 * 0.8778 = -1.826188 so 0 is within the CI and setting this with 0

SAR (1): 0.2547 + 1.96 * 0.2193 = 0.684528

0.2547 - 1.96 * 0.2193 = -0.175128 so 0 is within the CI so setting this with 0

SMA (1): -0.5449 + 1.96 * 0.1897 = -0.173088

-0.5449 - 1.96 * 0.1897 = -0.916712 so 0 is not with CI and keep this as NA

Therefore, I fixed Model III and find the coefficient with smallest AIC based on

Model III:

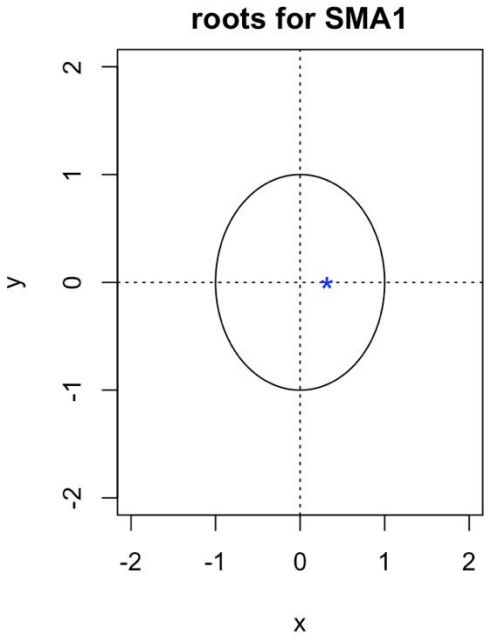
AR (1)	MA (1)	SAR (1)	SMA (1)
0	0	0	-0.3259

aic = -1351.25 and sigma^2 estimated as 8.257e-06

Fixed MODEL III: $SARIMA(0,1,0) \times (0,1,1)_{12}$ aic = -1351.25

Since fixed model III has smaller aic, so I chose fixed model III.

Also, we need to check the invertible and casual for my fixed model III:



Below is the result of root

[1] 3.068426+0i

Therefore, for my fixed model III, recall that $X_t = \nabla(\nabla_{12} Y_t^{0.2})$

$$\textit{Fixed Model III: } SARIMA(0,1,0) \times (0,1,1)_{12}$$
$$X_t = (1 - 0.3259B^{12})Z_t$$

Where $Z_t \sim N(0, 8.257e\text{-}06)$

From the all fixed model above, I found that all three fixed model suggest me $SARIMA(0,1,0) \times (0,1,1)_{12}$ is the only model, so I use this model as the model I selected.

Model Selected: $SARIMA(0,1,0) \times (0,1,1)_{12}$ Recall that $X_t = \nabla(\nabla_{12} Y_t^{0.2})$

$$X_t = (1 - 0.3259B^{12})Z_t$$

Where $Z_t \sim N(0, 8.257e\text{-}06)$

6. Diagnostic

Diagnostic checking on the behavior of residuals can demonstrate the validity of the fitted model as well as suggest reasonable modification of the model. Therefore, after identifying and estimating a time series model, diagnostic checking needs to be performed and three main assumptions should be verified including normality of residuals, independence (no serial correlation), and constant variance.

6.1 Normality Checking

I expect the residuals are normally distributed. In order to check this assumption, I draw the histogram and QQ plot of the residual series that are shown in Figure 6.1

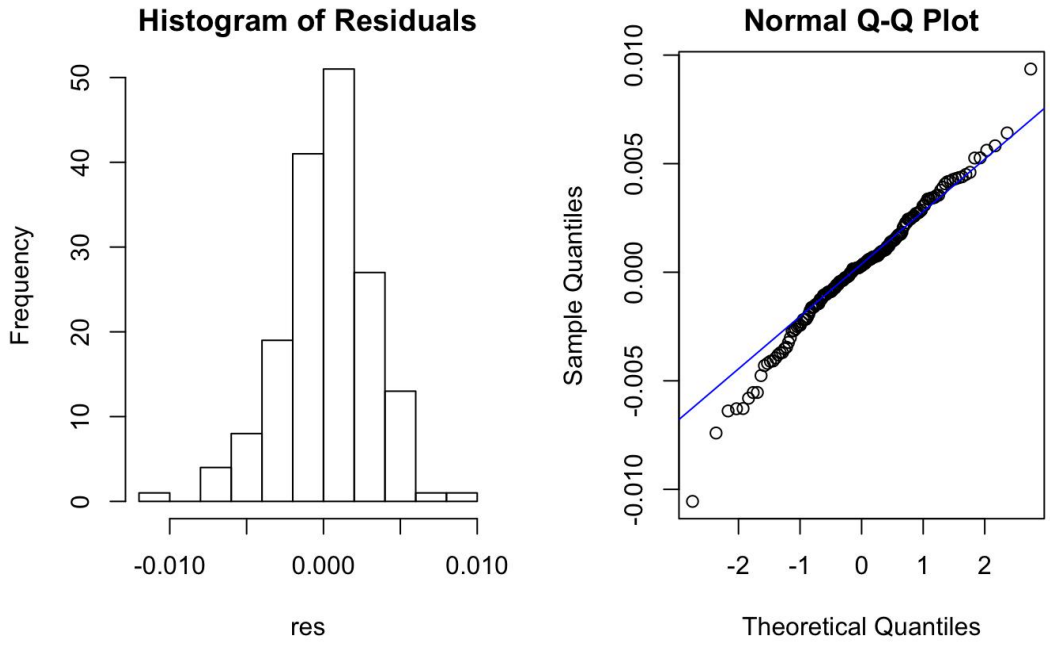


Figure 6.1

From figure 6.1, the histogram of the residuals of my model is symmetrical and has a bell shape which resembles the shape of normal distribution. Besides, the normal Q-Q plots shows that for my model, most of the points lie on the straight line, which indicate that the errors are normal.

Besides, I also perform the Shapiro Wilk Test as $\alpha = 0.05$ to further testify the assumption.

H_0 = residuals are normal

H_1 = residuals are not normal

	W. Statistics	**P-value**
Final Model	0.98406	0.05414

Table 6.1

As shown in Table 6.1, the p-value > 0.050, so we do not reject the assumption of normality. The residuals for my model are approximately Gaussian

6.2 Independence (Serial Correlation) Checking

The residuals are assumed to be uncorrelated with each other, and I perform three tests: Ljung-Box, Box-Pierce and Mcleod-Li (Ljung-box for squares) test at $\alpha = 0.05$ to check if the autocorrelation of residuals exists.

H_0 = Residuals are serially uncorrelated

H_1 = Residuals are not serially uncorrelated

Final model	Box-Pierce	Ljung-Box	McLeod-Li test: Ljung-Box for squares
P-value	0.8627	0.8341	0.5748

Table 6.2

The result in table 6.2 shows that for all three-test p-value > 0.05 for my model, hence I do not reject the assumption of serially uncorrelation between the residuals.

6.3 Constant Variance checking

I require my residuals to have constant variance for my model estimation and prediction to be efficient. The check for heteroskedasticity (the violation of constant variance of errors) can be done through the analysis of the ACF and PACF plots of the residuals in my model—the ACF and PACF values should lie within 95% of the White Noise limits. The result (Figure 6.3 below) shows that for my model, all of the values lie within the bound (the blue dash lines). Hence, I'm sure that the constant variance assumption is not violated and heteroskedasticity problem doesn't exist in my model.

Since my model pass the diagnostic test, so I use this model as my final model:

Recall that: $X_t = \nabla(\nabla_{12} Y_t^{0.2})$

Therefore,

Final Model: $SARIMA(0,1,0) \times (0,1,1)_{12}$

$$X_t = (1 - 0.3259B^{12})Z_t$$

Where $Z_t \sim N(0, 8.257e\text{-}06)$

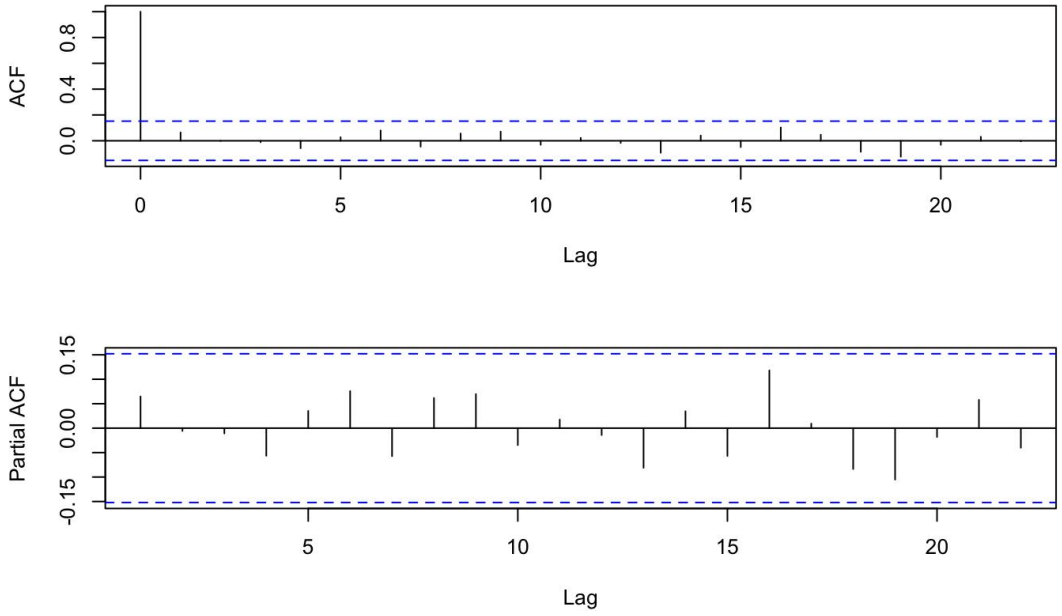


Figure 6.3

Besides, I also used the Yule-walker method find that the residual of my model is fitted AR (0), which means that my residuals are white noise. Thus, my model is ready to forecast. (see Appendix I)

7. Forecasting

Forecasting is our main goal of constructing a time series model. I now forecast 12 values ahead. Figure 7.1(a)(b) shows the V_t (the 5^{th} root transformed training data) and Figure 7.2(a)(b) shows the initial monthly employee demand data. The 12 red and orange dots represent the 12 forecasted value; the blue lines represent the boundaries of 95% confidence interval.

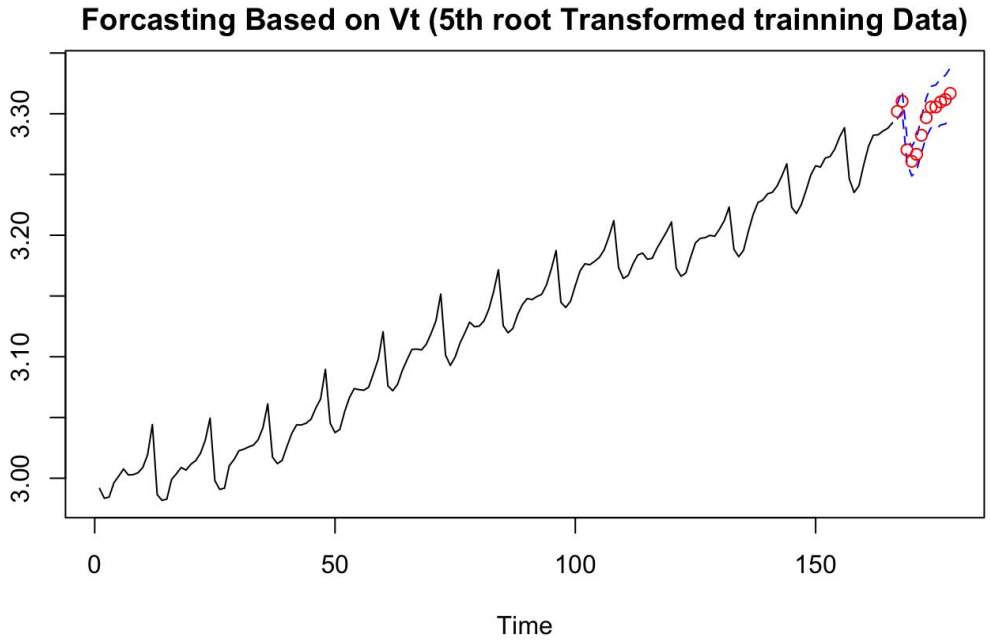


Figure 7.1 (a)

Zoom up for Figure 7.1(a):

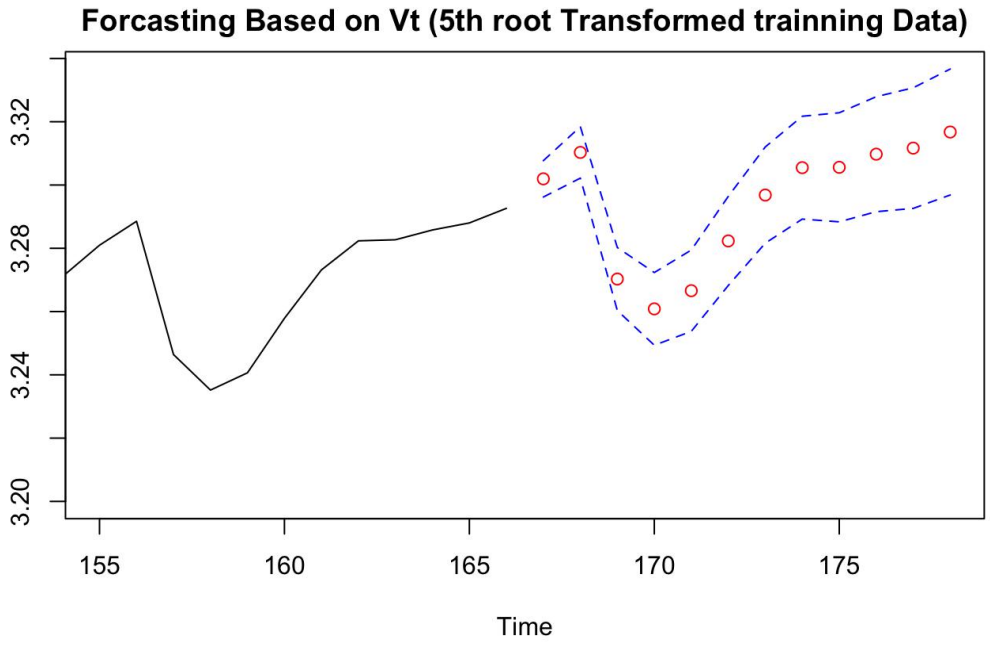


Figure 7.1(b)

Back to original data:

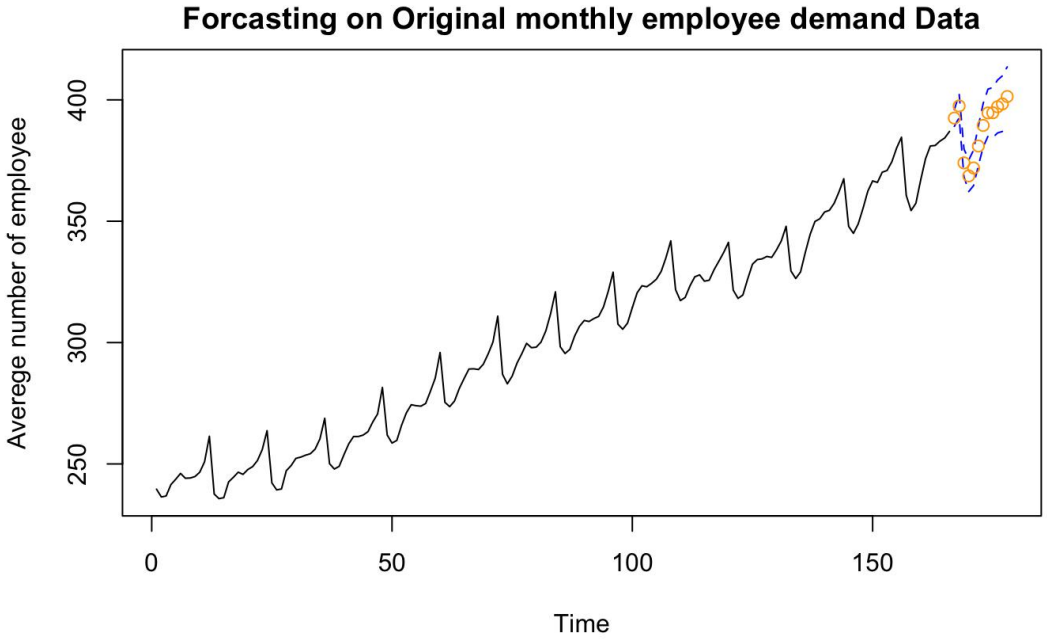


Figure 7.2(a)

Zoom up the confident interval part:

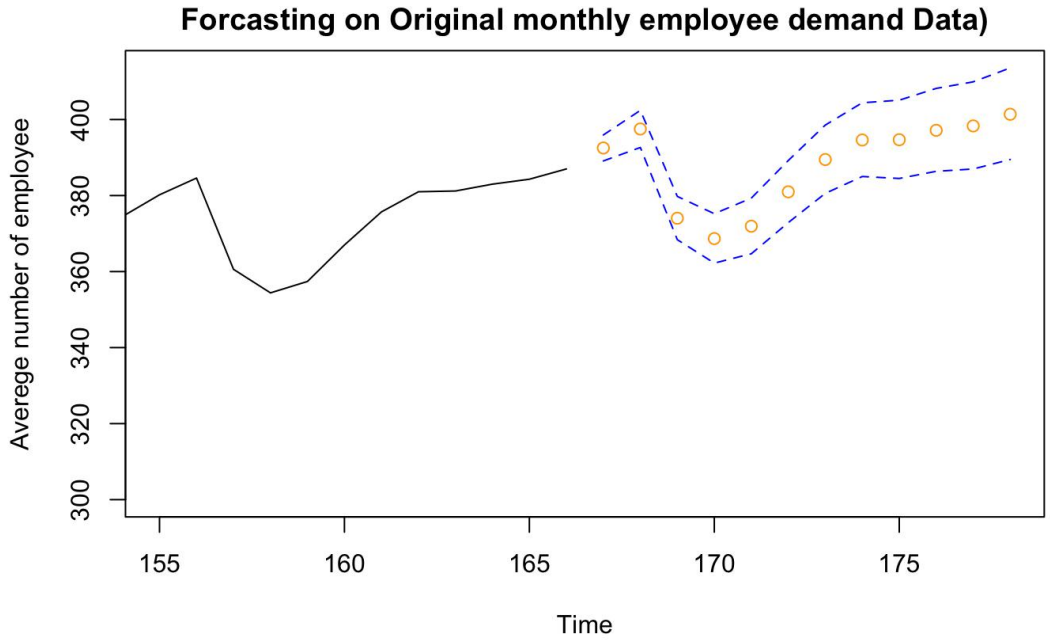


Figure 7.2(b)

A clearer look on the section, along with comparison between true observed values (black dots) and forecasted value (orange dots) can be seen in Figure 7.3:

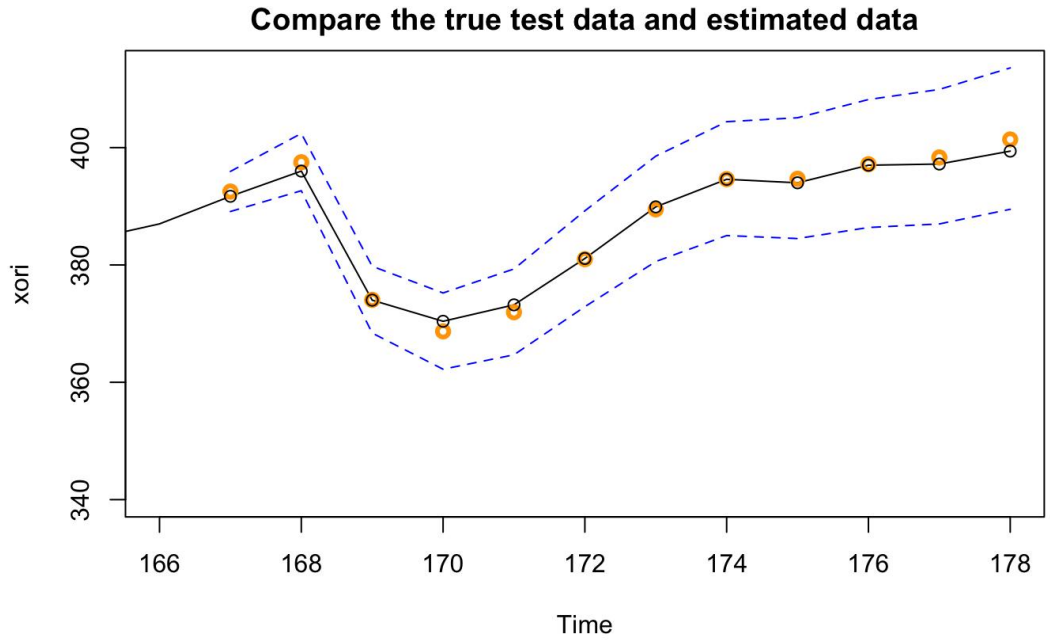


Figure 7.3

My final model clearly captures the overall trend and seasonality of the data. Staring from the second orange point, my forecasted value decrease indicating a downward trend and then goes up again, resembling the obvious pattern of my original data. Not only do they lie perfectly within the confidence interval, the observed and forecasted values are very close throughout the time period; some even overlaps. This proves the success of our final model.

8 Conclusion:

My goal is to construct a time series that can explain the monthly demand of employee in Wisconsin and predict its future need up to 12 months ahead. I found that there is an upward trend in the demand level which might be resulted from the

increase in size of company in the Wisconsin. Furthermore, I notice an obvious pattern and seasonality in my data—the demand level always reaches its peak between November and December every year, it may due to some holiday events such as Christmas day or thanksgiving day that demand more employee to work, and the cycle repeats every 12 months. After I make my data stationary, I fit it into several models, carry out the model selection process and run diagnostic check on them to select the best model. Recall that $X_t = \nabla(\nabla_{12} Y_t^{0.2})$. So final model is $SARIMA(0,1,0) \times (0,1,1)_{12}$. $X_t = (1 - 0.3259B^{12})Z_t$, Where $Z_t \sim$ N (0, 8.257e-06). Then I move on to forecast the monthly demand of employee from Nov, 1974 until Oct, 1975. My forecasted values not only just fall between the 95% confidence interval but also found to be very close to the true observed values. This proves the feasibility of my final model.

I would like to thank Professor Dr. Raya Feldman sincerely for this opportunity to apply all the Time Series knowledge she taught us into practice as well as offering helpful views on my project.

8. References

[1] Box Jenkins (1976), Time Series Data Library (TSDL) DataMarket http://datamarket.com/data/list/?q=provider:tsdl

[2]

https://gauchospace.ucsb.edu/courses/pluginfile.php/1996049/modresource/ content/1/plot.roots.R

9 Appendix: R code

```
Call:
ar(x = res, aic = TRUE, order.max = NULL, method = c("yule-walker"))

Order selected 0  sigma^2 estimated as  8.328e-06
```

Appendix I

get data from tsdl library.

```
library(tsdl) ## preparation, get data

tsdl

first_few = 5

for (i in 1:first_few)

cat("i= ", i, attr(tsdl[[i]], "subject"), "\n")

attr(tsdl[[544]],"subject")

attr(tsdl[[544]],"description")
```

choosing data and plot data

```
x <- tsdl[[544]]

x
```

```
length(x) # total we have 178 data

ts.plot(x,gpars=list(xlab="Year", ylab="Average number of employee", lty=c(1:
3)))

title("Monthly empolyees demand in Wisconsin ")
```

set up the training data

```
#drop last 12 data from transformed data for training

xtr<- x[c(1:166)] #this is the training data set

xtr

x.test <- x[c(167:178)]
```

stationary process

step 1: do the transformation

```
library(MASS)

t = 1:length(xtr)

fit = lm(xtr ~ t)

bcTransform = boxcox(xtr ~ t,plotit = TRUE)

lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]

xtr.bc = (1/lambda)*(xtr^lambda-1)

lambda    #[1] 0.1818182

x5tr <- xtr ^ (1/5) #the 5th root transformation
```

```
# compare the original training dataset plot and transformed training dataset
 plot
op=par(mfrow=c(1,2))
plot.ts(xtr)
plot.ts(x5tr)

#compare the variance of difference transformation

var(sqrt(xtr)) #[1] 1.51675
var(xtr.bc)    #[1] 0.1610108
var(log(xtr))  #[1] 0.02027031
var(x5tr)      #0.007805394 so we find that x5tr has the smallest      varia
nce
```

Preliminary Data Exploration

```
#check seasonal component: Decomposition
y <- ts(as.ts(x5tr),frequency = 12)
plot(decompose(y))
# the graph shows that my dataset have seasonal component
```

```
#plot ACF and PACF of the transformed training data Vt (5th root transform
ed data series)


op=par(mfrow=c(1,2))
acf(x5tr,lag.max=70,main="")
pacf(x5tr,lag.max=70,main="")
title("The ACF and PACF for Vt (5th root transformed training data)",line=-1,
outer=TRUE)
```

zoom up

```
#plot ACF and PACF of the transformed training data Vt (5th root transform
ed data series)
op=par(mfrow=c(1,2))
acf(x5tr,lag.max=20,main="")
pacf(x5tr,lag.max=20,main="")
title("Zoom up ACF and PACF for Vt (5th root transformed training data)",lin
e=-1,outer=TRUE)
```

step 2: remove seasonal and linear trend

```
#De-seasonal for Vt first
xd12 <- diff(x5tr,lag = 12)
ts.plot(xd12,main = "De-seasonalized for Vt(5th root transformed training dat
a)", ylab = expression(nabla[12]~V[t]))
```

```
var(xd12)

abline(h = mean(xd12), col = "red") #find the mean

fit <- lm(xd12 ~ as.numeric(1:length(xd12)))

abline( fit, col="blue")                # find the trend



#remove the linear trend after de-seasonal

xd22 <- diff(xd12,1)

ts.plot(xd22,main = "De-seasonalized & De-trend of Vt (5th root transformed

 training data)", ylab = expression(nabla~nabla[12]~V[t]))

abline(h = mean(xd22), col = "red")

fit <- lm(xd22 ~ as.numeric(1:length(xd22)))

abline( fit, col="blue")                #check the trend

var(xd22)                                  #8.967057e-06


#test if remove trend once more

xd32 <- diff(xd22,1)                       #over difference

var(xd32)


#check the acf and pacf after de-seasonal and de-trend

op=par(mfrow=c(1,2))

acf(xd22,lag.max=60,main="")
```

```
pacf(xd22,lag.max=60,main="")     #So xd22 is the stationary data Xt in repo
rt
title("The stationary data Xt after De-seasonal & De-trend",line=-1,outer=TR
UE)



#compare the original training data Yt and stationary dataset Xt
op=par(mfrow=c(1,2))
ts.plot(xtr)
ts.plot(xd22)
title("compare the original training data Yt and stationary data Xt ",line=-1,ou
ter=TRUE)
```

Therefore, I notice that for seasonal part P = 1 and Q = 1 s=12

zoom up to identified the non-seasonal part p,q

```
xd12 <- diff(x5tr,lag = 12)
xd22 <- diff(xd12,1)
op=par(mfrow=c(1,2))
acf(xd22,lag.max=20,main="")
pacf(xd22,lag.max=20,main="")
title("Zoom up The stationary data Xt after De-seasanol & De-trend",line=-1,
outer=TRUE)
```

Augmented Dickey—Fuller Test to verify whether Xt is stationary or not

```
library(tseries)

adf.test(xd22)



#Dickey-Fuller = -4.5822, Lag order = 5, p-value = 0.01

#alternative hypothesis: stationary
```

Model selecting:

step 1：Identify the model by selecting AIC

```
library(qpcR)

#model estimate

aiccs = matrix(NA, nr = 36, nc = 3)

colnames(aiccs) = c("p", "q", "AICc")

i=0

for(p in 0:1){

  for(q in 0:1){

    aiccs[i+1, 1] = p

    aiccs[i+1, 2] = q

    aiccs[i+1, 3] = AICc(arima(x5tr, order = c(p,1,q), method="ML",seasonal =

 list(order = c(p,1, q), period = 12)))

    i = i+1
```

```
  }
}


aiccs[order(aiccs[,3])[0 :4],]
```

Model I: SARIMA (0,1,1) x (1,1,1) aic =-1349.32

```
arima(x5tr, order=c(0,1,1), seasonal = list(order = c(1,1,1), period = 12), metho
d="ML")
```

step 2: check the invertible and casual for model I

```
source('/Users/zheyue/174/plot.roots.R')
par(mfrow = c(1,2))
plot.roots(NULL, polyroot(c(1,0.0815)), main = "roots for ma1")
plot.roots(NULL, polyroot(c(1,0.2505)), main = "roots for Sar1")
plot.roots(NULL, polyroot(c(1,-0.5402)), main = "roots for SMA1")


polyroot(c(1,0.0815))
polyroot(c(1,0.2505))
polyroot(c(1,-0.5402))


##Roots outside the unit circle, invertible and casual
```

```
#The blue star means the imaginary part and the red star means the real p
art
```

step 3：Fixed Model I

```
arima(x5tr, order=c(0,1,1), seasonal = list(order = c(1,1,1), period = 12), fixed =
 c(0, 0, NA), method="ML")


#Model I fixed : SARIMA(0,1,0) x (0,1,1) aic = -1351.25
```

step 4：check the invertible and casual for MODEL I fixed

```
source('/Users/zheyue/174/plot.roots.R')
par(mfrow = c(1,2))
plot.roots(NULL, polyroot(c(1,-0.3252)), main = "roots for SMA1")
polyroot(c(1,-0.3252))


##Roots outside the unit circle, invertible and casual
```

MODEL II：SARIMA $(1,1,0)\times(1,1,1)12$

```
arima(x5tr, order=c(1,1,0), seasonal = list(order = c(1,1,1), period = 12), metho
d="ML")
```

repeat step2：check the invertible and casual for model II

```
source('/Users/zheyue/174/plot.roots.R')

par(mfrow = c(1,2))

plot.roots(NULL, polyroot(c(1,0.0834)), main = "roots for ma1")

plot.roots(NULL, polyroot(c(1,0.252)), main = "roots for Sar1")

plot.roots(NULL, polyroot(c(1,-0.5419)), main = "roots for SMA1")



polyroot(c(1,0.0834))

polyroot(c(1,0.252))

polyroot(c(1,-0.5419))



##Roots outside the unit circle, invertible and casual

#The blue star means the imaginary part and the red star means the real p
art
```

repeat step 3: fixed Model II

```
arima(x5tr, order=c(1,1,0), seasonal = list(order = c(1,1,1), period = 12), fixed =
 c(0, 0, NA), method="ML")



#fixed Model II: SARIMA(0,1,0)x(0,1,1)  aic = -1351.25
```

repeat step 4: check the invertible and casul for MODEL II fixed

```
source('/Users/zheyue/174/plot.roots.R')

par(mfrow = c(1,2))
```

```
plot.roots(NULL, polyroot(c(1,-0.3259)), main = "roots for SMA1")

polyroot(c(1,-0.3259))


##Roots outside the unit circle, invertible and casual
```

Model III: SARIMA (1,1,1) x (1,1,1)12

```
arima(x5tr, order=c(1,1,1), seasonal = list(order = c(1,1,1), period = 12), method
="ML")
```

repeat step2: check the invertible and casual for model III

```
source('/Users/zheyue/174/plot.roots.R')

par(mfrow = c(1,2))

plot.roots(NULL, polyroot(c(1,0.1883)), main = "roots for ar1")

plot.roots(NULL, polyroot(c(1,-0.1057)), main = "roots for ma1")

plot.roots(NULL, polyroot(c(1,0.2547)), main = "roots for Sar1")

plot.roots(NULL, polyroot(c(1,-0.5449)), main = "roots for SMA1")


polyroot(c(1,0.1883))

polyroot(c(1,-0.1057))

polyroot(c(1,0.2547))

polyroot(c(1,-0.5449))
```

```
##Roots outside the unit circle, invertible and casual

#The blue star means the imaginary part and the red star means the real p
art
```

repeat step 3: fixted Model III

```
arima(x5tr, order=c(1,1,1), seasonal = list(order = c(1,1,1), period = 12), fixed =
 c(0, 0, 0,NA), method="ML")

#fixed Model III: (0,1,0)x(0,1,1) aic = -1353.25
```

repeat step 4: check the invertible and casual for MODEL III fixed

```
source('/Users/zheyue/174/plot.roots.R')

par(mfrow = c(1,2))

plot.roots(NULL, polyroot(c(1,-0.3259)), main = "roots for SMA1")

polyroot(c(1,-0.3259))



##Roots outside the unit circle, invertible and casual
```

Model Diagnostics: analyze the residual

```
fit1=arima(x5tr, order=c(0,1,0),seasonal=list(order=c(0,1,1),period=12),include.mea
n=F)

res = residuals(fit1)

mean(res)

var(res)
```

#Normality Checking

```
#Test for normality of residuals

# plot the Histogram of residuals

hist (res,main = "Histogram of Residuals")

# q-q plot

qqnorm(res)

qqline(res,col ="blue")
```

Shapiro Test

```
shapiro.test(res)
```

Independence (Serial Correlation) Checking

```
#n = √178 =14, and p+d+q+P+D+Q = 0+1+0+0+1+1 = 3

Box.test(res, lag = 14, type = c("Box-Pierce"), fitdf = 3)  #Box-Pierce test

Box.test(res, lag = 14, type = c("Ljung-Box"), fitdf = 3)   #Box-Ljung test


Box.test(res^2, lag = 14, type = c("Ljung-Box"), fitdf = 0) #McLeod-Li test:

Ljung-Box for squares


#plot the Mcleod-Li test

library("TSA")

McLeod.Li.test(fit1)
```

Test for constant variance of residuals

```
layout(matrix(c(1,1,2,3),2,2,byrow=T))
ts.plot(res,main = "Fitted Residuals")
t = 1:length(res)
fit.res = lm(res~t)
abline(fit.res)              # the black line indecated that my residuals are alm
ost overlap the line 0, indicated that my model fit well.
abline(h = mean(res), col = "red")



# acf
acf(res,main = "")
# pacf
pacf(res,main = "")
```

Fitted residuals to AR(0), i.e. WN!

```
ar(res, aic = TRUE, order.max = NULL, method = c("yule-walker"))
```

Forecasting

```
library(forecast)
fit1=Arima(x5tr, order=c(0,1,0),seasonal=list(order=c(0,1,1),period=12),include.me
```

```
an=F)

forecast(fit1)

length(fit1)
```

Forecast next 12 observations of Vt (5th root Transformed trainning Data)

```
pred.xtr<-predict(fit1,n.ahead = 12)

Utr=pred.xtr$pred + 2*pred.xtr$se          # construct the upper bond confiden

t interval

Ltr=pred.xtr$pred - 2*pred.xtr$se          # construct the lower bond confiden

t interval

ts.plot(x5tr, xlim=c(1,length(x5tr)+12),ylim = c(min(x5tr),max(Utr)), main="Foreca

sting Based on Vt (5th root Transformed training Data)",ylab="")

lines(Utr,col="blue",lty="dashed")

lines(Ltr,col="blue",lty="dashed")

points((length(x5tr)+1):(length(x5tr)+12),pred.xtr$pred,col="red")
```

zoom up

```
#Forecast next 12 observations of Vt (5th root Transformed trainning Data) f

rom 167th to 178th

pred.xtr<-predict(fit1,n.ahead = 12)

Utr=pred.xtr$pred + 2*pred.xtr$se

Ltr=pred.xtr$pred - 2*pred.xtr$se

ts.plot(x5tr, xlim=c(155,length(x5tr)+12),ylim = c(3.20,max(Utr)), main="Forecast
```

```
ing Based on Vt (5th root Transformed training Data)",ylab="")
lines(Utr,col="blue",lty="dashed")
lines(Ltr,col="blue",lty="dashed")
points((length(x5tr)+1):(length(x5tr)+12),pred.xtr$pred,col="red")
```

Forecasting the initial monthly employee demand Data

```
pred.orig = pred.xtr$pred ^5
U = Utr^5          #previous CI upper bound
L = Ltr^5          #previous CI lower bound
ts.plot(xtr, xlim=c(1,length(xtr)+12), ylim = c(min(xtr),max(U)), main="Forecastin
g on Original monthly employee demand Data",ylab="Averege number of emp
loyee")
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points((length(xtr)+1):(length(xtr)+12), pred.orig, col="orange")
```

rescale to zoom up

```
pred.orig = pred.xtr$pred ^5
U = Utr^5
L = Ltr^5
#xori <- ts(x)
ts.plot(xtr, xlim=c(155,length(xtr)+12), ylim = c(300,max(U)),main="Forecasting
on Original monthly employee demand Data)",ylab="Average number of emplo
```

```
yee")
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points((length(xtr)+1):(length(xtr)+12), pred.orig, col="orange")
```

Compare the true data and theoretic value

```
pred.orig = pred.xtr$pred ^5
U = Utr^5
L = Ltr^5
xori <- ts(x)
ts.plot(xori, xlim=c(166,length(xtr)+12), ylim = c(340,max(U)), main = "Compare
 the true test data and estimated data")
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")


points((length(xtr)+1):(length(xtr)+12), pred.orig, col="orange",lwd = 3)  # the
model predict data point
points((length(xtr)+1):(length(xtr)+12),x.test,col="black") # the real data to test
```