Analysis of C-CPI-U of Public Transportation in U.S. Cities From 1990-2007 Using SARIMA Models

Evan Hu

December 2022

Abstract

The goal of this project is to analyze the relationship between Chained Consumer Price Index for all Urban Consumers (C-CPI-U) of public transportation in U.S. cities and time using a dataset from the U.S. Bureau of Labor Statistics. We are interested in finding possible seasonal patterns of C-CPI-U for public transportation and in forecasting the monthly C-CPI-U for public transportation. We utilize Box-Jenkins methodology to fit a SARIMA model to the dataset, to compare different time series models, and to perform diagnostics on the models. Our final model suggests that SARIMA models may not be suitable for this dataset.

Introduction

In this project, we use a dataset of monthly C-CPI-U for public transportation in U.S. Cities from the U.S. Bureau of Labor Statistics to study the relationship between C-CPI-U and time/seasonality. The dataset contains a total of 216 observations from January 1990 to December 2007. We will use the first 200 observations as the training dataset and the remaining 16 dataset as the testing dataset. Investopedia provides the following definition for C-CPI: "chain-weighted CPI, or chained CPI, is an alternative measurement for the Consumer Price Index (CPI) that considers changes to consumer spending patterns to provide a more accurate picture of the cost of living based on the goods that consumers actually buy". Though C-CPI is a less accurate measure of inflation than CPI, we can investigate monthly C-CPI-U of public transportation for those 17 years to learn more about cost of living and inflation in the United States. For forecasting, it's important to build a time series model that can help

predict future monthly C-CPI-U so we may get a better sense of the cost of living and inflation in the future.

We first perform data visualization to get a better sense of the data we are working with. The plots we obtain from the data visualization process suggest we difference the data at lag 12 to remove seasonality and at lag 1 to remove trend. We perform model identification with the sample ACF and sample PACF of the stationary data. Using SARIMA models, we perform model fitting and compare models. We use diagnostic checking tools for our candidate models such as residuals analysis and Portmanteau tests. However, the models do not pass some of these tests and the model we end up using to forecast values does not perform well.

We conclude that Box-Jenkins methodology may not be a suitable choice for this particular dataset. The analysis in this paper is made possible by the data provided by the U.S. Bureau of Labor Statistics. The analysis was completed using R software.

Section 1: Data Visualization

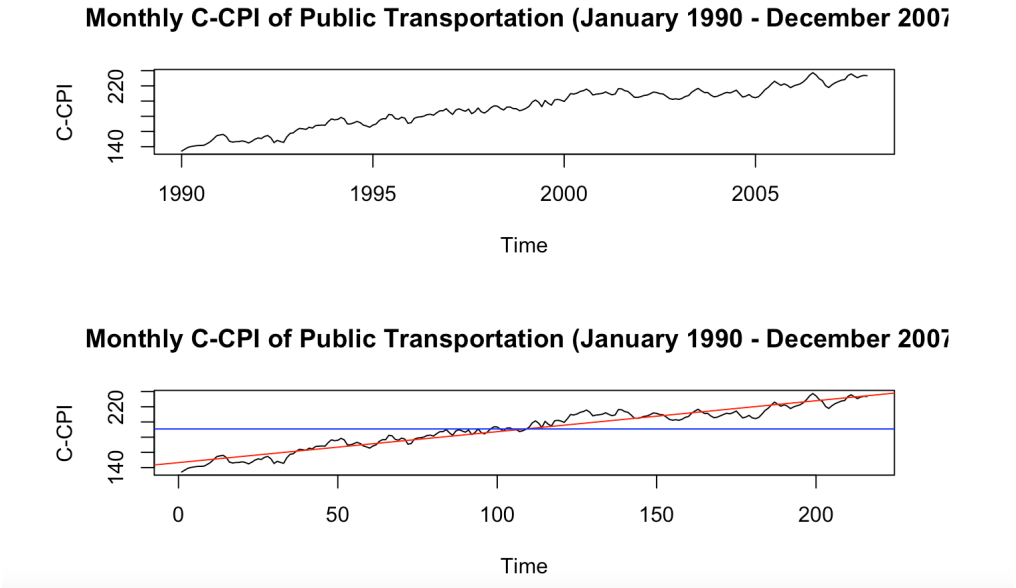

Figure 1: Monthly C-CPI-U for Public Transportation in U.S. Cities from January 1990 - December 2007 (red line: linear regression fit of the data, blue line: mean of data)

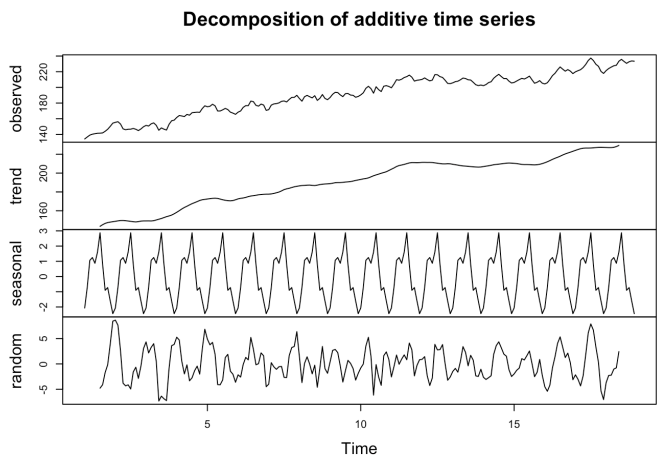

Figure 2: Decomposition of the Time Series in Additive Format (data = trend + seasonal + random components)

In Figure 1, the red line, which represents the regression fit of the time series, reveals that there is a positive linear trend between monthly C-CPI-U for public transportation and time. There also appears to be a seasonal pattern. We can also see that the variance seems to remain constant with time. Furthermore, the decomposition of the time series in additive format in Figure 2 clearly shows that our data is seasonal, has an upwards linear trend, and has rather constant variance. Thus, from these plots, it seems we do not need to transform our data, but we will need to difference our data to remove seasonality and to remove trend.

We first split our data into a training set and a testing set. We choose the first 200 observations (January 1990 - August 2006) as the training set and the remaining 45 observations (September 2006 - December 2007) as our testing set.

As noted above, the variance does not seem to change with time. Thus, we will forgo data transformation and continue to differencing the training data. After differencing the data at lag 12 to remove seasonality, we plot the differenced data in Figure 3 and can see that the seasonality is no longer apparent. Table 1 shows that the variance of the seasonally differenced training data is lower than the original training data. Then, we difference at lag 1 to remove trend. The variance decreases again. In Figure 3, the plot of the data differenced for both seasonality and trend does not show any signs of seasonality or trend. Additionally, the red regression fit line is very close to the mean and has a very small but negligible upwards trend. We conclude that the data differenced at lag 12 and at lag 1 is stationary.

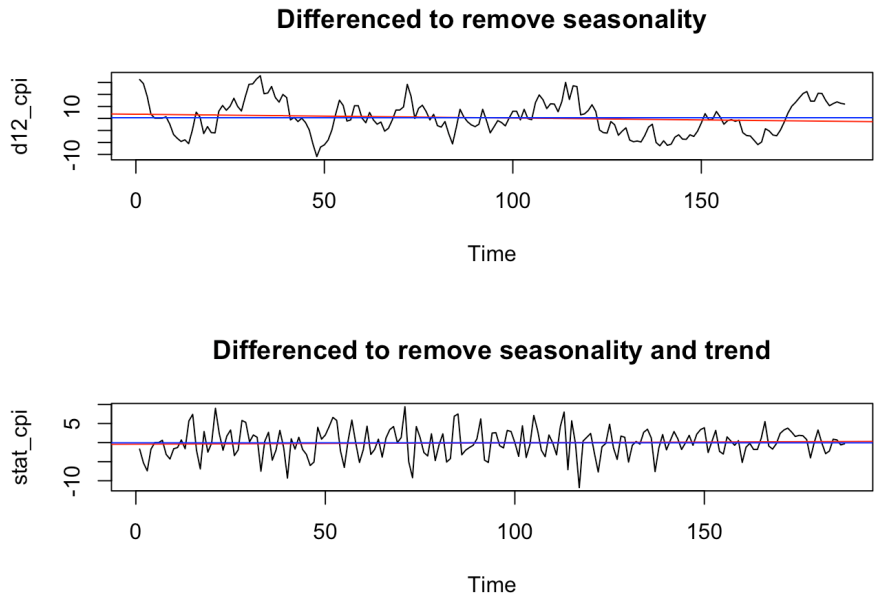

Figure 3: Plots of Differenced Data

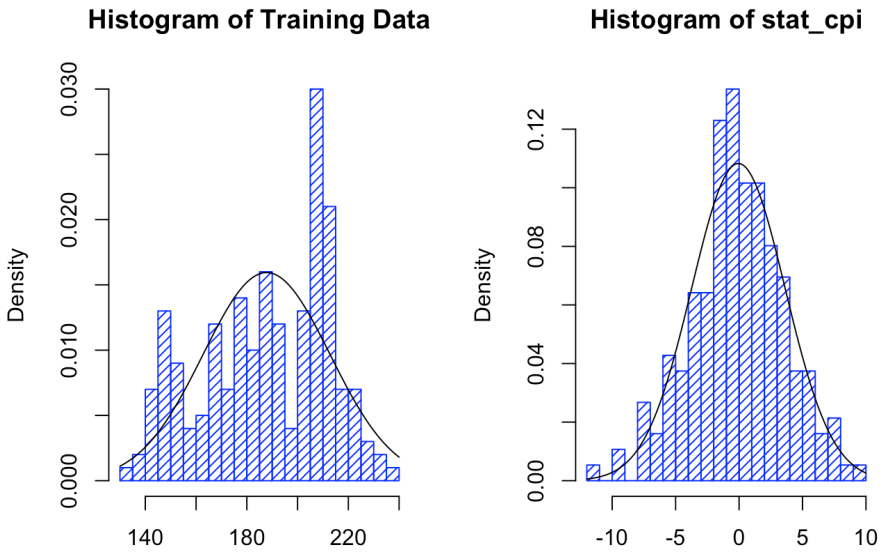

Figure 4: Left: Histogram of Training Data, Right: Histogram of Differenced Data

Data	Variance	Mean
Original training data	625.8865	187.775
Data after differencing at lag 12	47.93199	5.290426
Data after differencing at lag 1	13.56991	-0.0545454

Table 1: Variances and means of original and differenced data

In Figure 4, we compare the histograms of our original training data and our differenced data. We can see that after differencing for trend and seasonality, the data looks more normally distributed.

Section 2: Model Identification

In Section 2, we perform model identification based on the sample ACF and sample PACF. Since our data was differenced for both trend and seasonality, we will find an appropriate SARIMA model. From the plot of the sample ACF in Figure 5, ACF at lags 2, 8, 10, 12, and 14 are significant because the ACF at these lags are outside of the 95% confidence interval. Thus, we should consider q = 2, 8, 10 and Q = 1 for the moving average part of the model. Looking at the plot of the sample PACF, we have significant PACF at lags 2, 3, 11, 12, 24, and 39. We should consider p = 2, 3. There does not seem to be a seasonal AR part as the significant PACFs tail off.

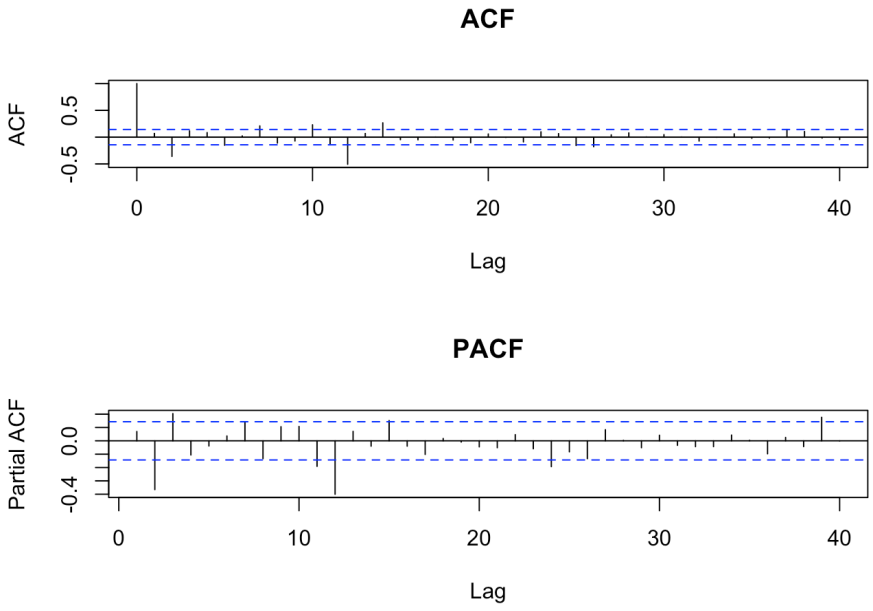

Figure 5: Top: Autocorrelation Function of Data, Bottom: Partial Autocorrelation Function of Data

Thus, we will try p = 2, 3 ; q = 2, 8, 10 ; Q = 1 ; P = 0. Since we differenced for seasonality and trend, D = 1 and d = 1 is a given.

Section 3: Model fitting

Section 3.1: Model A (SARIMA$(3,1,2)(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0)

After fitting different models from Table 2, the fourth model SARIMA$(3,1,2)(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0 is the model we choose based on the standard errors for the coefficient estimates of the models. The chosen model has the second lowest AICC of the 5 models. Despite the model not having the lowest AICC and ar3 having 0 in its confidence interval, the chosen model matches the plots of the sample ACF and PACF more.

Model	AICC
SARIMA$(2,1,2)(0,1,1)_{12}$	915.8224
SARIMA$(2,1,2)(0,1,1)_{12}$ with ar1 and ma1 fixed to 0	922.1931
SARIMA$(3,1,2)(0,1,1)_{12}$ with ar1 and ma1 fixed to 0	921.5869
SARIMA$(3,1,2)(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0	920.6108
SARIMA$(3,1,3)(0,1,1)_{12}$ with ar1 and ma1 fixed to 0	922.5992

Table 2: Fitted models and their AICC

SARIMA$(3,1,2)(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0						
	ar1	ar2	ar3	ma1	ma2	sma1
Coefficients	0	0	0.1266	0	-0.3469	-0.7442
Standard Error	0	0	0.0747	0	0.0798	0.0576
AICC: 920.6108						

Table 3: Summary of training data fitted to SARIMA$(3,1,2)(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0

To check if the residuals of Model A follow a white noise distribution, we use several diagnostic tools. We first check normality assumptions. Looking at Figure 6, the residuals seem to follow a normal distribution from the histogram and q-q plot. There is no seasonality or trend in the plot of the residuals. From Figure 7, the ACF of the residuals seems to be significant at lag 1 and the PACF of the residuals is significant at lags 1, 8, 48, and 51 because they are outside of the confidence interval. Thus, the residuals of Model A do not resemble white noise.

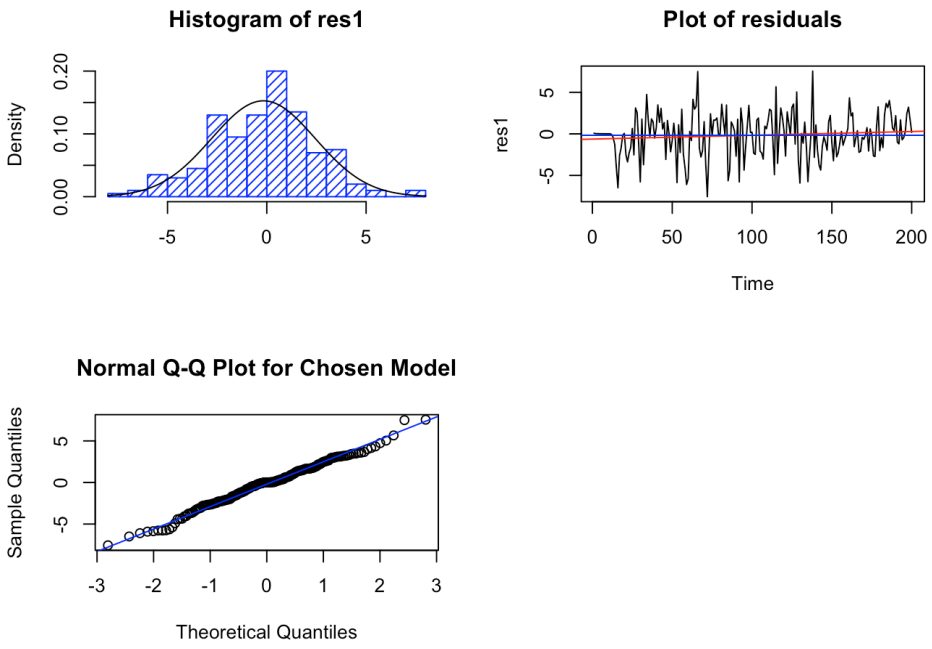

Figure 6: Diagnostic checking plots for normality of residuals of Model A

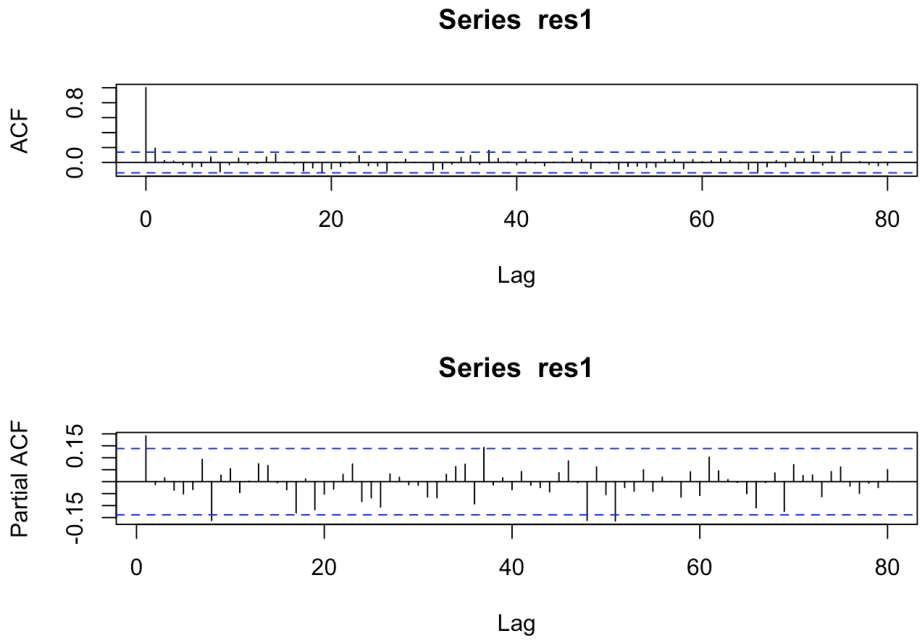

Figure 7: Top: ACF of residuals of Model A, Bottom: PACF of residuals of Model A

We also check for several independence assumptions in Table 4. The model only passes 2 of the four tests. Thus, since our model did not pass some diagnostic tests, we re-evaluate our model by changing our p and q.

Test	Statistics	P-value	Result
Shapiro-Wilk	0.99007	0.183	Pass
Box-Pierce	26	0.07447	Pass
Ljung-Box	27.68	0.04882	Did not pass
McLeod-Li	32.377	0.03945	Did not pass

Table 4: Testing results for Model A

Section 3.2: Model B (SARIMA$(3,1,3)(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0)

Upon trying to determine a new candidate model, fitting any model with q = 8 or 10 does not produce any fruitful results. With q = 10, the seasonal moving average part of the model with Q = 1 is not invertible because |sma| > 1. With q = 8 and different combinations of coefficients fixed to 0, ma8 consistently has 0 in its confidence interval.

From the plot of the ACF and PACF of the residuals of Model A, there is significant ACF and PACF at lag 1. Thus, we should consider increasing both or either p and q of Model A by 1. We first fit SARIMA$(4,1,2)(0,1,1)_{12}$ with ar1, ar2 and ma1 fixed to 0.

SARIMA$(4,1,2)(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0							
	ar1	ar2	ar3	ar4	ma1	ma2	sma1
Coefficients	0	0	0.1301	-0.0324	0	-0.3392	-0.7501
Standard Error	0	0	0.0751	0.0776	0	0.0797	0.0592
AICC: 922.5857							

Table 5: Summary of training data fitted to SARIMA$(4,1,2)(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0

Since ar4 has 0 in its confidence interval, we will now instead try increasing only q by 1. From Table 6, we can see that all the coefficients do not have 0 in their

confidence interval. Thus, we will consider SARIMA(3,1,3)$(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0 as our Model B and move onto diagnostic checking.

SARIMA(3,1,3)$(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0							
	ar1	ar2	ar3	ma1	ma2	ma3	sma1
Coefficients	0	0	0.6300	0	-0.2639	-0.6148	-0.7722
Standard Error	0	0	0.1629	0	0.0761	0.1903	0.0600
AICC: 921.8241							

Table 6: Summary of training data fitted to SARIMA(3,1,3)$(0,1,1)_{12}$ with ar1, ar2, and ma1 fixed to 0

To check if the residuals of Model B follow a white noise distribution, we use several diagnostic tools. We first check normality assumptions. Looking at Figure 8, the residuals seem to follow a normal distribution from the histogram and q-q plot. There is no seasonality or trend in the plot of the residuals. From Figure 9, the ACF of the residuals seems to be significant at lag 1 and the PACF of the residuals is significant at lags 1, 48, and 51 because they are outside of the confidence interval. Thus, the residuals of Model B do not resemble white noise.

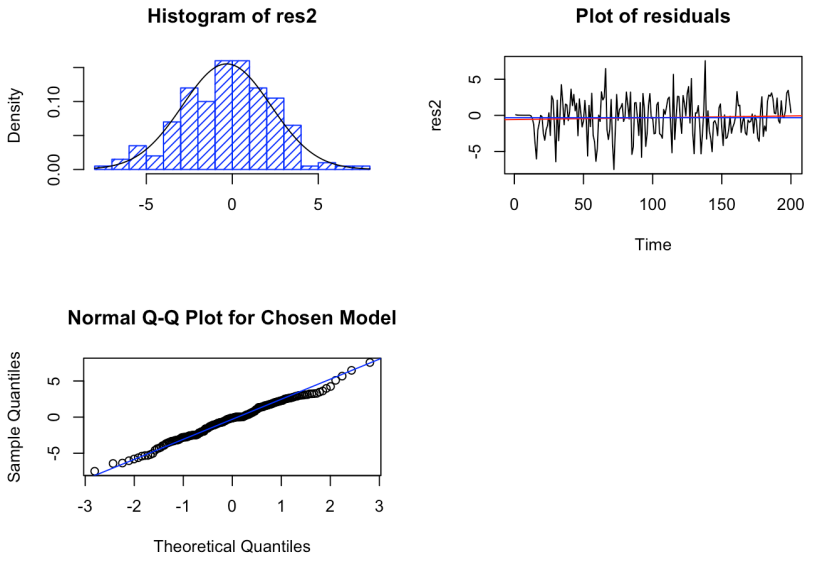

Figure 8: Diagnostic checking plots for normality of residuals of Model B

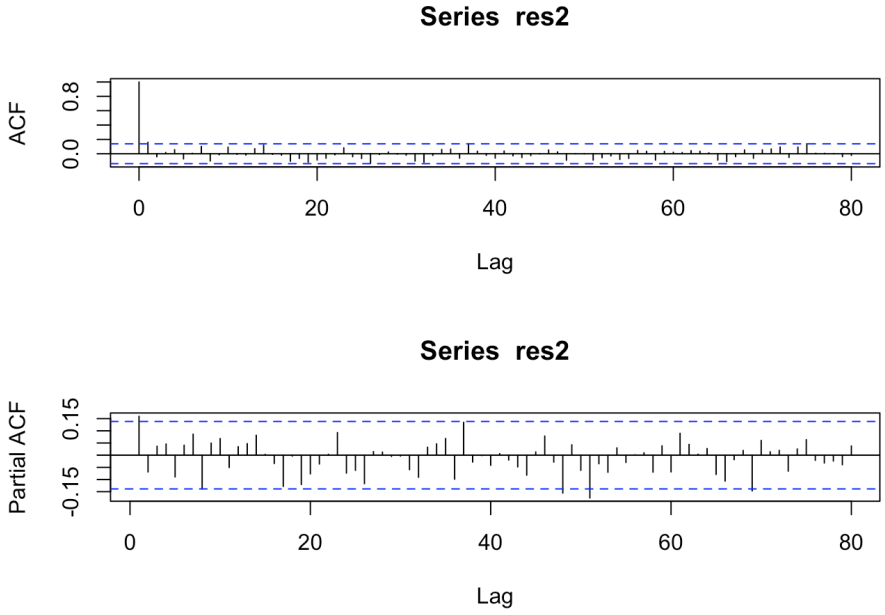

Figure 9: Top: ACF of residuals of Model B, Bottom: PACF of residuals of Model B

We check for several independence assumptions in Table 7. The model only passes 3 of the four tests. Thus, since our model did not pass some diagnostic tests, we re-evaluate our model by changing our p and q.

Test	Statistics	P-value	Result
Shapiro-Wilk	0.99195	0.3366	Pass
Box-Pierce	24.252	0.06094	Pass
Ljung-Box	25.864	0.03947	Did not pass
McLeod-Li	28.069	0.1078	Pass

Table 7: Testing results for Model B

Section 3.3: Model C

From the plot of the ACF and PACF of the residuals of Model B, there is significant ACF and PACF at lag 1. Thus, we should consider increasing both or either p and q of Model B by 1. We first fit SARIMA$(4,1,3)(0,1,1)_{12}$ with ar1, ar2 and ma1 fixed to 0.

SARIMA(4,1,3)(0,1,1)$_{12}$ with ar1, ar2, and ma1 fixed to 0								
	ar1	ar2	ar3	ar4	ma1	ma2	ma3	sma1
Coefficients	0	0	0.6481	0.0472	0	-0.2659	-0.6423	-0.762
Standard Error	0	0	0.1274	0.0609	0	0.0740	0.1417	0.0613
AICC: 922.66								

Table 8: Summary of training data fitted to SARIMA(4,1,3)(0,1,1)$_{12}$ with ar1, ar2, and ma1 fixed to 0

From Table 8, we notice that the AICC increased a bit compared to Model B, and ar4 has 0 in its confidence interval. Since ar4 is not significant, we will instead try q = 4. From Table 9, we can see that ar3 now has 0 in its confidence interval and ma4 is not significant.

SARIMA(3,1,4)(0,1,1)$_{12}$ with ar1, ar2, and ma1 fixed to 0								
	ar1	ar2	ar3	ma1	ma2	ma3	ma4	sma1
Coefficients	0	0	0.2059	0	-0.3360	-0.0667	-0.0925	-0.7684
Standard Error	0	0	0.2434	0	0.0737	0.2701	0.0871	0.0617
AICC: 923.5048								

Table 9: Summary of training data fitted to SARIMA(3,1,4)(0,1,1)$_{12}$ with ar1, ar2, and ma1 fixed to 0

To save time, I've gone ahead and tried fitting many different models. SARIMA(3,1,5)(0,1,1)$_{12}$ with ar1, ma1, and ma3 fixed to 0 seems to work the best with the training data. The AICC is the second lowest out of all the models, only falling behind in AICC to the first model in Table 2. All estimated coefficients are significant except ma5, but I've decided to leave the model as q = 5 because reducing the model to q = 4 would change coefficient estimates and standard errors and increase AICC.

SARIMA(3,1,5)(0,1,1)$_{12}$ with ar1, ma1, and ma3 fixed to 0									
	ar1	ar2	ar3	ma1	ma2	ma3	ma4	ma5	sma1
Coefficients	0	-0.89	0.1293	0	0.5961	0	-0.358	0.0869	-0.72
Standard Error	0	0.056	0.0509	0	0.0962	0	0.0791	0.0526	0.062
AICC: 919.96									

Table 10: Summary of training data fitted to SARIMA(3,1,5)(0,1,1)$_{12}$ with ar1, ma1, and ma3 fixed to 0

Having chosen our final model as SARIMA(3,1,5)(0,1,1)$_{12}$ with ar1, ma1, and ma3 fixed to 0, we proceed to diagnostic checking for the model.

To check if the residuals of our chosen model follow a white noise distribution, we use several diagnostic tools. We first check normality assumptions. Looking at Figure 10, the residuals seem to follow a normal distribution from the histogram and q-q plot. There is no seasonality or trend in the plot of the residuals. From Figure 11, the ACF of the residuals seems to be significant at lag 1. However, the ACF at lag 1 is barely above the confidence interval line, so we can use Bartlett's formula and count the ACF at lag 1 as within the confidence interval. Sadly, the PACF of the residuals is significant at lag 1. The residuals fail the PACF check.

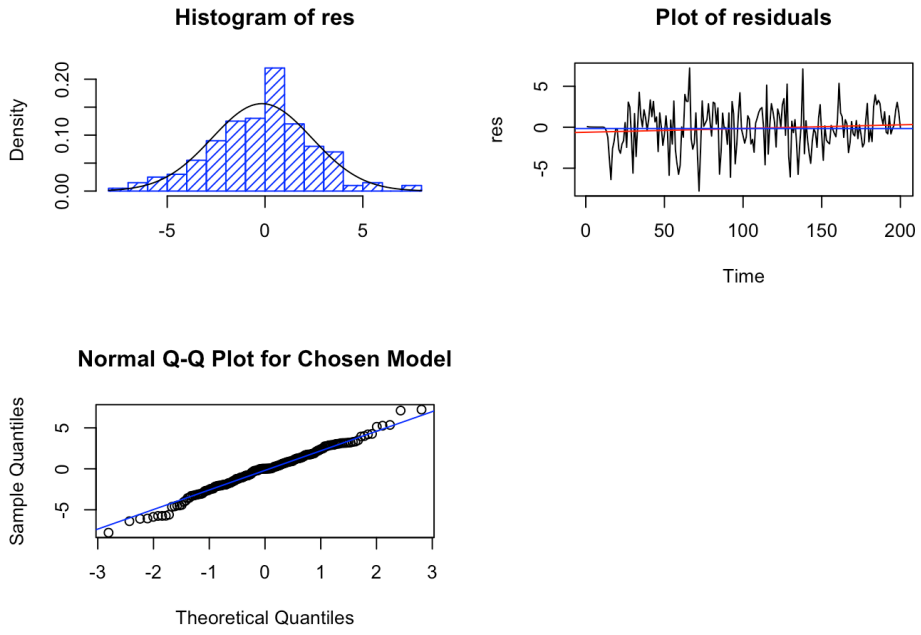

Figure 10: Diagnostic checking plots for normality of residuals of final model

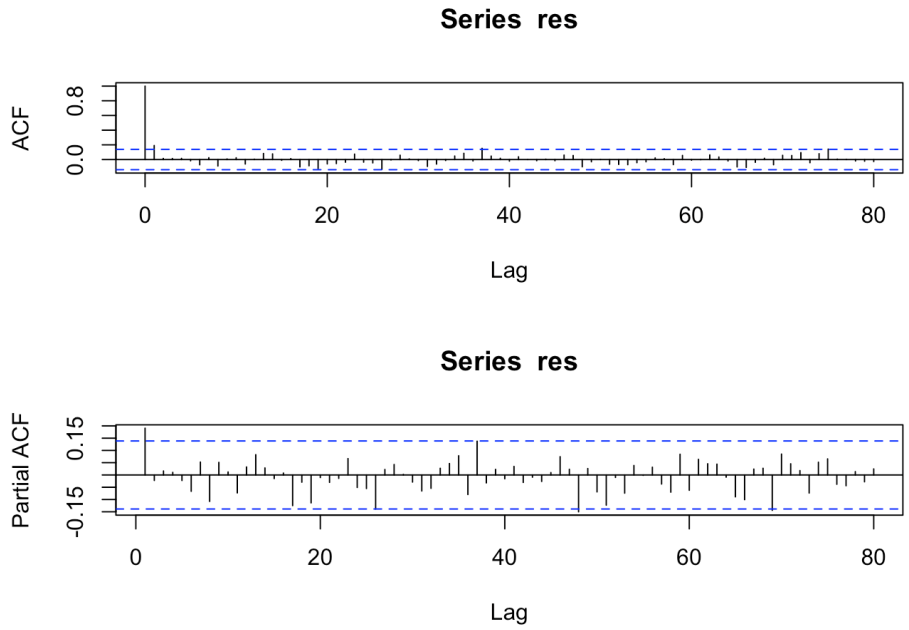

Figure 11: Top: ACF of residuals of final model, Bottom: PACF of residuals of final model

We check for several independence assumptions in Table 11. The model only passes 3 of the four tests. It does not pass the McLeod-Li test, which tests for nonlinear dependence. Thus, the squared residuals of the model are correlated. The residuals passed the Shapiro-Wilk test of normality. The residuals also passed the Box-Pierce and Ljung-Box test, which means the residuals should resemble white noise.

Test	Statistics	P-value	Result
Shapiro-Wilk	0.99013	0.1868	Pass
Box-Pierce	20.83	0.1061	Pass
Ljung-Box	22.156	0.07546	Pass
McLeod-Li	33.428	0.03026	Did not pass

Table 11: Testing results for chosen model

Though our model did not pass all the diagnostic checks, we will use this model for forecasting. I've tried many different models and all of them end up failing one of the tests and their residuals always seem to have significant ACF and

PACF at lag 1. Therefore, despite the original data looking rather suitable to be fitted into a SARIMA model, Box-Jenkins methodology may not work well for this data, and a different time series methodology may be more fruitful.

Finally, we want to determine if the model is invertible and stationary. In the seasonal moving average part, there is only one parameter, which has an estimated coefficient of -0.7201, so it is invertible. For the nonseasonal moving average part, we have $(1+0.5961B^2-0.3576B^4+0.0869B^5)$. Checking the roots of the MA polynomial in Figure 12, we can see that the roots all lie outside the unit circle. We should note that there are a couple of roots that are very close to the unit circle, which could indicate that the model is not invertible. For the nonseasonal autoregressive part of the model, we have $(1+0.8903B^2-0.1293B^3)$. Checking the roots of the AR polynomial in Figure 12, we can see that the roots all lie outside the unit circle. We should note that there are a couple of roots that are very close to the unit circle, which could indicate that the model is not stationary.

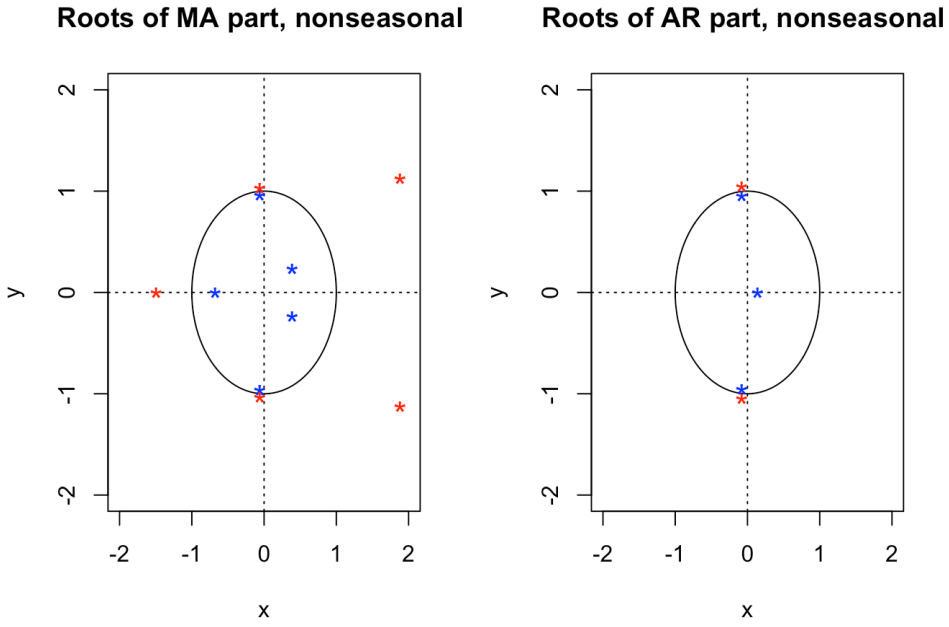

Figure 12: Red stars are the roots of the polynomials

Our final model:

$$(1-B)(1-B^{12})(1+0.8903B^2-0.1293B^3)X_t =$$
$$(1+0.5961B^2-0.3576B^4+0.0869B^5)(1-0.7201B^{12})Z_t$$

Section 4: Forecasting

For forecasting, we predict the C-CPI-U of public transportation in U.S. cities from September 2006 to December 2007 based on our final model and compare the predictions with the true values. Figure 13 has plots of the forecasted results and the predicted values are all a bit higher than the actual values. Thus, our model does not perform well.

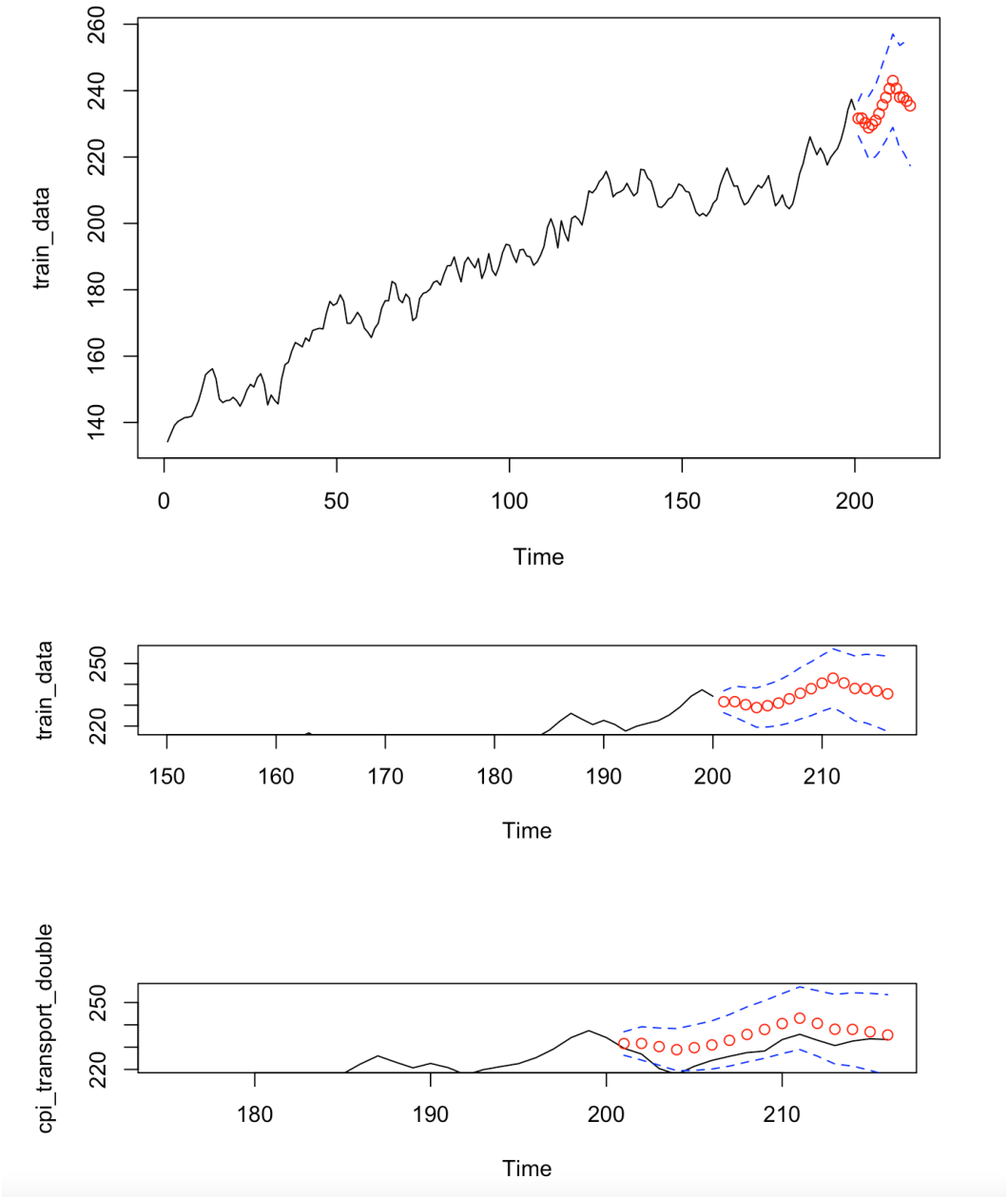

Figure 13: Black line represents the true time series data. The red dots are the predicted values and the blue lines are the confidence interval of the predictions.

Conclusion

In this project, we used a dataset of monthly C-CPI-U for public transportation in U.S. Cities from the U.S. Bureau of Labor Statistics to study the relationship between C-CPI-U and time/seasonality. We applied different time series and diagnostic techniques such as Box-Jenkins methodology, Portmanteau tests, checking residuals for normality in order to find possible seasonal patterns of C-CPI-U for public transportation and in forecasting the monthly C-CPI-U for public transportation. Our analysis suggests that SARIMA models may not be suitable for this dataset as all the models we fitted failed diagnostic checking. Although it's frustrating we were not able to meet our goal of fitting a SARIMA model to the dataset, I believe I was able to learn more about time series data analysis than if we were able to fit a SARIMA model.

The data may be better modeled by (S)ARIMA model if it was only differenced at lag 1. Differencing the original training data only at lag 1 for trend produced a lower variance than differencing at both lag 1 and 12 for trend and seasonality. However, the models fit to the data differenced only for trend were not able to pass the Box-Pierce and Ljung-Box tests.

References

[1] Data: U.S. Bureau of Labor Statistics. (n.d.). *BLS Data viewer*. U.S. Bureau of Labor Statistics. Retrieved December 8, 2022, from https://beta.bls.gov/dataViewer/view/timeseries/SUUR0000SAT

Appendix

```
#load packages
library(MASS)
library(tidyverse)
library(qpcR)
# load data
cpi_transport_list <- read.csv("/Users/evanhu/Desktop/PSTAT 174/final project/file.csv")
cpi_transport_list <- select(cpi_transport_list, "Value")
cpi_transport_double <- unlist(cpi_transport_list) # convert from 'list' to 'double' object
cpi_transport <- ts(cpi_transport_double, start = c(1990, 1), end=c(2007,12), frequency = 12) #
create time series object
```

Plot of the original series

```
par(mfrow=c(2, 1))
ts.plot(cpi_transport, main = "Monthly C-CPI of Public Transportation (January 1990 - December
2007)", ylab = "C-CPI")

ts.plot(cpi_transport_double, main = "Monthly C-CPI of Public Transportation (January 1990 -
December 2007)", ylab =  "C-CPI")
fit <- lm(cpi_transport_double ~ as.numeric(1:length(cpi_transport_double)))
abline(fit, col="red") # add trend to data plot
abline(h=mean(cpi_transport_double), col="blue") # add mean to data plot

x <- ts(as.ts(cpi_transport), frequency = 12)
decomp <- decompose(x)
plot(decomp)
```

Training/testing split, work with training set

```
train_data <- cpi_transport[c(1:200)] # training set
test_data <- cpi_transport[c(201:216)] # testing set

# plot training data
plot.ts(train_data)
fit <- lm(train_data ~ as.numeric(1:length(train_data)))
abline(fit, col="red")
abline(h=mean(train_data), col="blue")
```

Differencing data

```
var(train_data)
mean(train_data)

d12_cpi <- diff(train_data, lag = 12, 1)
```

```
var(d12_cpi)
mean(d12_cpi)

stat_cpi <- diff(d12_cpi, lag = 1, 1)
var(stat_cpi)
mean(stat_cpi)
par(mfrow=c(2, 1))
# plot of seasonally differenced data
ts.plot(d12_cpi, main = "Differenced to remove seasonality")
fit <- lm(d12_cpi ~ as.numeric(1:length(d12_cpi)))
abline(fit, col="red")
abline(h=mean(d12_cpi), col="blue")

# plot of stationary data
ts.plot(stat_cpi, main = "Differenced to remove seasonality and trend")
fit <- lm(stat_cpi ~ as.numeric(1:length(stat_cpi)))
abline(fit, col="red")
abline(h=mean(stat_cpi), col="blue")

par(mfrow=c(1, 2))
# Histograms of original and differenced data
hist(train_data, density = 20, breaks = 20, col= "blue", xlab="", main="Histogram of Training
Data", prob = TRUE)
m<-mean(train_data)
std<- sqrt(var(train_data))
curve(dnorm(x,m,std), add=TRUE )

hist(stat_cpi, density=20,breaks=20, col="blue", xlab="", prob=TRUE)
m<-mean(stat_cpi)
std<- sqrt(var(stat_cpi))
curve(dnorm(x,m,std), add=TRUE )
```

Sample ACF/PACF

```
par(mfrow=c(2, 1))
acf(train_data, lag.max = 80, main = "")
pacf(train_data, lag.max = 80, main = "")

acf(stat_cpi, lag.max = 200, main = "ACF")
pacf(stat_cpi, lag.max = 200, main = "PACF")

acf(stat_cpi, lag.max = 40, main = "ACF")
pacf(stat_cpi, lag.max = 40, main = "PACF")

fit1 <- arima(train_data, order = c(2, 1, 2), seasonal = list(order = c(0,1,1), period = 12), method =
"ML")
fit2 <- arima(train_data, order = c(2, 1, 2), seasonal = list(order = c(0,1,1), period = 12), method =
"ML", fixed =c(0,NA,0,NA,NA))
```

```
fit3 <- arima(train_data, order = c(3, 1, 2), seasonal = list(order = c(0,1,1), period = 12), method =
"ML",fixed =c(0,NA,NA,0,NA,NA))
fit4 <- arima(train_data, order = c(3, 1, 2), seasonal = list(order = c(0,1,1), period = 12), method =
"ML", fixed = c(0,0,NA,0,NA,NA))
fit5 <- arima(train_data, order = c(3, 1, 3), seasonal = list(order = c(0,1,1), period = 12), method =
"ML",fixed =c(0,NA,NA,0,NA,NA,NA))
AICc(fit1)
AICc(fit2)
AICc(fit3)
AICc(fit4)
AICc(fit5)

fit6 <- arima(train_data, order = c(3, 1, 8), seasonal = list(order = c(0,1,1), period = 12), method =
"ML",fixed =c(0,0,NA,0,NA,0,0,0,0,0,NA,NA))
fit6
fit7 <-arima(train_data, order = c(3, 1, 10), seasonal = list(order = c(0,1,1), period = 12), method =
"ML",fixed =c(0,0,NA,0,NA,0,0,0,0,0,0,NA,NA))
fit7
fit8 <-arima(train_data, order = c(3, 1, 10), seasonal = list(order = c(0,1,1), period = 12), method =
"ML",fixed =c(0,0,NA,0,NA,0,0,0,0,0,NA,0,NA,NA))
fit8
fit9 <- arima(train_data, order = c(4, 1, 2), seasonal = list(order = c(0,1,1), period = 12), method =
"ML", fixed = c(0,0,NA,NA,0,NA,NA))
fit9
fit10 <- arima(train_data, order = c(4, 1, 3), seasonal = list(order = c(0,1,1), period = 12), method
= "ML", fixed = c(0,0,NA,NA,0,NA,NA,NA))
fit10
fit11 <- arima(train_data, order = c(3, 1, 3), seasonal = list(order = c(0,1,1), period = 12), method
= "ML", fixed = c(0,0,NA,0,NA,NA,NA))
fit11
AICc(fit9)
AICc(fit10)
AICc(fit11)

# model to consider
model1 <- arima(train_data, order = c(3, 1, 2), seasonal = list(order = c(0,1,1), period = 12),
method = "ML",fixed =c(0,0,NA,0,NA,NA))
res1 <- residuals(model1)

par(mfrow=c(2, 2))
# diagnostic check for model 1
hist(res1, density=20, breaks=20, col="blue", xlab="", prob=TRUE)
m <- mean(res1)
std <- sqrt(var(res1))
curve(dnorm(x,m,std), add=TRUE )

plot.ts(res1, main = "Plot of residuals")
res_fit1 <- lm(res1 ~ as.numeric(1:length(res1)))
abline(res_fit1, col="red")
```

```
abline(h=mean(res1), col="blue")

qqnorm(res1,main= "Normal Q-Q Plot for Chosen Model")
qqline(res1,col="blue")

par(mfrow=c(2, 1))

acf(res1, lag.max=80)
pacf(res1, lag.max=80)

par(mfrow=c(1, 1))
shapiro.test(res1)
Box.test(res1, lag = 20, type = c("Box-Pierce"), fitdf = 3)
Box.test(res1, lag = 20, type = c("Ljung-Box"), fitdf = 3)
Box.test(res1^2, lag = 20, type = c("Ljung-Box"), fitdf = 0)
acf(res1^2, lag.max=40)

ar(res1, aic = TRUE, order.max = NULL, method = c("yule-walker"))

model2 <- arima(train_data, order = c(3, 1, 3), seasonal = list(order = c(0,1,1), period = 12),
method = "ML",fixed =c(0,NA,NA,0,NA,NA,NA))

res2 <- residuals(model2)

par(mfrow=c(2, 2))
# diagnostic check for model 2
hist(res2, density=20, breaks=20, col="blue", xlab="", prob=TRUE)
m <- mean(res2)
std <- sqrt(var(res2))
curve(dnorm(x,m,std), add=TRUE)

plot.ts(res2, main = "Plot of residuals")
res_fit2 <- lm(res2 ~ as.numeric(1:length(res2)))
abline(res_fit2, col="red")
abline(h=mean(res2), col="blue")

qqnorm(res2,main= "Normal Q-Q Plot for Chosen Model")
qqline(res2,col="blue")

par(mfrow=c(2, 1))

acf(res2, lag.max=80)
pacf(res2, lag.max=80)

par(mfrow=c(1, 1))
shapiro.test(res2)
Box.test(res2, lag = 20, type = c("Box-Pierce"), fitdf = 5)
Box.test(res2, lag = 20, type = c("Ljung-Box"), fitdf = 5)
Box.test(res2^2, lag = 20, type = c("Ljung-Box"), fitdf = 0)
acf(res2^2, lag.max=40)
```

```
ar(res2, aic = TRUE, order.max = NULL, method = c("yule-walker"))

model3 <- arima(train_data, order = c(3, 1, 4), seasonal = list(order = c(0,1,1), period = 12),
method = "ML",fixed =c(0,NA,NA,0,NA,0,NA,NA))

res3 <- residuals(model3)

par(mfrow=c(2, 2))
# diagnostic check for model 2
hist(res3, density=20, breaks=20, col="blue", xlab="", prob=TRUE)
m <- mean(res3)
std <- sqrt(var(res3))
curve(dnorm(x,m,std), add=TRUE)

plot.ts(res3, main = "Plot of residuals")
res_fit3 <- lm(res3 ~ as.numeric(1:length(res3)))
abline(res_fit3, col="red")
abline(h=mean(res3), col="blue")

qqnorm(res3,main= "Normal Q-Q Plot for Chosen Model")
qqline(res3,col="blue")

par(mfrow=c(2, 1))

acf(res3, lag.max=80)
pacf(res3, lag.max=80)

par(mfrow=c(1, 1))
shapiro.test(res3)
Box.test(res3, lag = 20, type = c("Box-Pierce"), fitdf = 5)
Box.test(res3, lag = 20, type = c("Ljung-Box"), fitdf = 5)
Box.test(res3^2, lag = 20, type = c("Ljung-Box"), fitdf = 0)
acf(res3^2, lag.max=40)

ar(res3, aic = TRUE, order.max = NULL, method = c("yule-walker"))
```

Check invertibility and stationarity of model

```
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)
} }
par(mfrow=c(1,2))
plot.roots(NULL,polyroot(c(1, 0, 0.6, 0, -0.28, 0.13)), main="Roots of MA part, nonseasonal ")
plot.roots(NULL,polyroot(c(1, 0, 0.8903, -0.1293)), main="Roots of AR part, nonseasonal ")
```

Diagnostic checking for final model

```
final_model <-arima(train_data, order = c(3, 1, 5), seasonal = list(order = c(0,1,1), period = 12),
method = "ML",fixed =c(0,NA,NA,0,NA,0,NA,NA,NA))

final_model

par(mfrow=c(2, 2))

res <- residuals(final_model)
hist(res, density=20, breaks=20, col="blue", xlab="", prob=TRUE)
m <- mean(res)
std <- sqrt(var(res))
curve( dnorm(x,m,std), add=TRUE )


plot.ts(res, main = "Plot of residuals")
res_fit <- lm(res ~ as.numeric(1:length(res)))
abline(res_fit, col="red")
abline(h=mean(res), col="blue")

qqnorm(res,main= "Normal Q-Q Plot for Chosen Model")
qqline(res,col="blue")
```

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

acf(res, lag.max=80)
pacf(res, lag.max=80)

par(mfrow=c(1, 1))
shapiro.test(res)
Box.test(res, lag = 20, type = c("Box-Pierce"), fitdf = 6)
Box.test(res, lag = 20, type = c("Ljung-Box"), fitdf = 6)
Box.test(res^2, lag = 20, type = c("Ljung-Box"), fitdf = 0)
acf(res^2, lag.max=40)

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

Forecasting

```
library(forecast)

forecast(final_model)
```

Plots for forecasting

```
pred.law <- predict(final_model, n.ahead = 16)
pred.law$pred
U.law= pred.law$pred + 2*pred.law$se #upper bound of prediction interval
L.law= pred.law$pred - 2*pred.law$se #lower bound
ts.plot(train_data, xlim=c(1,length(train_data)+16), ylim = c(min(train_data),max(U.law)))
lines(U.law, col="blue", lty="dashed")
lines(L.law, col="blue", lty="dashed")
points((length(train_data)+1):(length(train_data)+16), pred.law$pred, col="red")

par(mfrow=c(2,1))
ts.plot(train_data, xlim = c(150,length(train_data)+16), ylim = c(min(L.law),max(U.law)))
lines(U.law, col="blue", lty="dashed")
lines(L.law, col="blue", lty="dashed")
points((length(train_data)+1):(length(train_data)+16), pred.law$pred, col="red")

ts.plot(cpi_transport_double, xlim = c(175,length(train_data)+16), ylim = c(220,max(U.law)))
lines(U.law, col="blue", lty="dashed")
lines(L.law, col="blue", lty="dashed")
points((length(train_data)+1):(length(train_data)+16), pred.law$pred, col="red")
```