¶ "Time Series Analysis of Monthly Milk Production from Jan. 1962 to Dec. 1975"

Ao Xu

7 Dec 2022

ABSTRACT

Based on monthly measurements gathered between the years 1962 and 1975, the time series analysis of milk production (pounds per cow) is performed in this project. The information was gathered in the R tsdl data library. The reference number is 203. There are 168 observations in the original dataset. The original dataset was split into two sets. The first 156 observations are in the training set, and the final 12 measurements are in the testing set. Applying the Box-Jenkins methodology, a suitable model is created from the training set to project future values and compare them to values from the testing set. The final model we selected is $SARIMA(0,1,1)(0,1,4)_{12}$. All predicted values fall within the 95% confidence interval according to this model.

1.0 INTRODUCTION

The dataset, which spans the period of January 1962 to December 1975, has 168 measures of monthly milk production (pounds per cow). It is a typical time series dataset with a enough number of observations for analysis and the creation of a prediction model. The original dataset is divided into a training set and a test set so that the prediction model's accuracy can be assessed. Monthly measurements from January 1962 to December 1974 make up the 156 values in the training set (denoted as c_train). Twelve values, representing the monthly measurements from January 1975 to December 1975, make up the testing set (denoted as c_test). We initially choose many models to estimate their coefficients using the Box-Jenkins approach. Then, by comparing AICcs and using diagnostic testing, we evaluate the models to determine which is best for forecasting. In order to verify the accuracy of our final model, we compare the forecast values to the data from the testing set.

1.1 Data Processing

The original data's plot displays recurring patterns and an upward trend, which point to seasonality and a non-constant mean. It appears that the data's variance is stable, nevertheless.

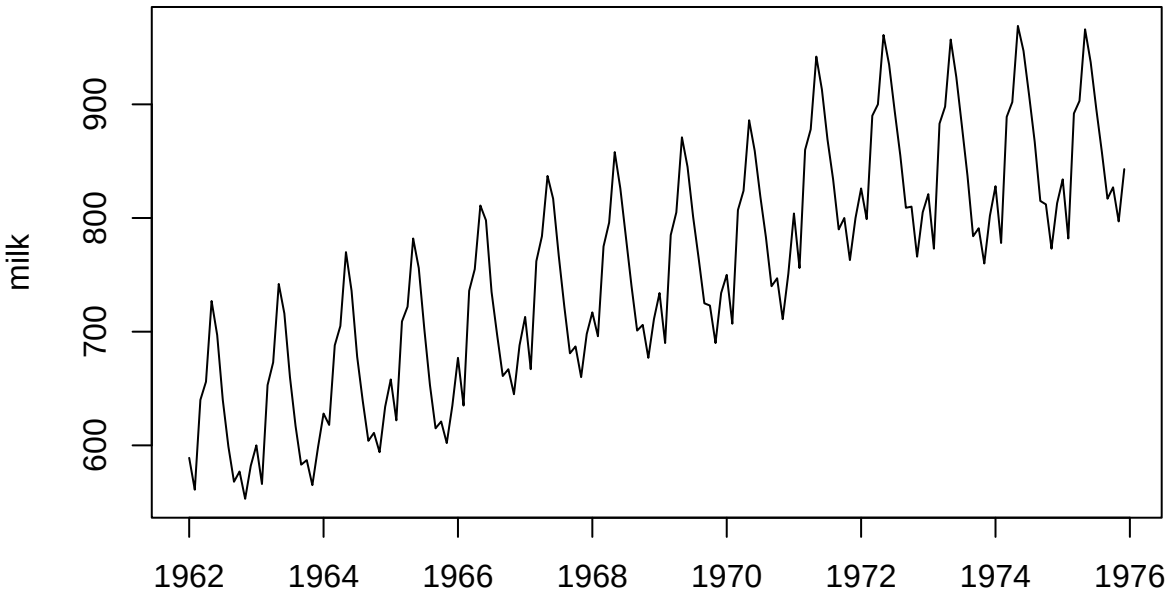

Figure 1: Monthly Measurements of Milk Prodcution Jan 1962 – Dec 1975.

We plot the histogram and ACF of the original dataset to further investigate its non-stationarity. The histogram has a right-skewed bias and is not symmetrical. Because the x axis is based on year, the ACF reveals a seasonality of 12 and a slow decay (shown in the graph as lag 1).

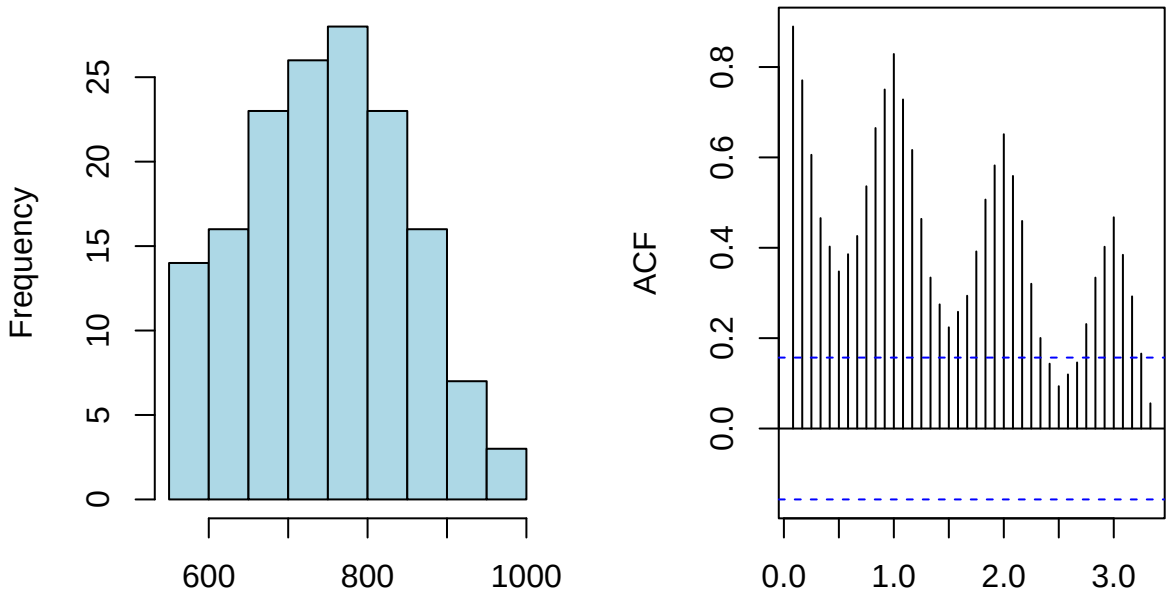

Figure 2: Histogram and ACF plot of c_train

Since the histogram does not exhibit a symmetric, normal pattern, we apply a box-cox transformation and refer to the resultant data as milk.log after log transformation.

```
## [1] 0.2626263
```

The box-cox plot gives us a result of $\lambda = 0.2626263$. The confidence interval of box-cox transformation includes 0 and 1. Therefore, we could experiment with non-transformation, log transformation, and box-cox transformation. We discovered that the log transformation was the most appropriate method for this data after completing the entire process and comparing the outcomes for each transformation. We set our log-transformed data set be milk.log.

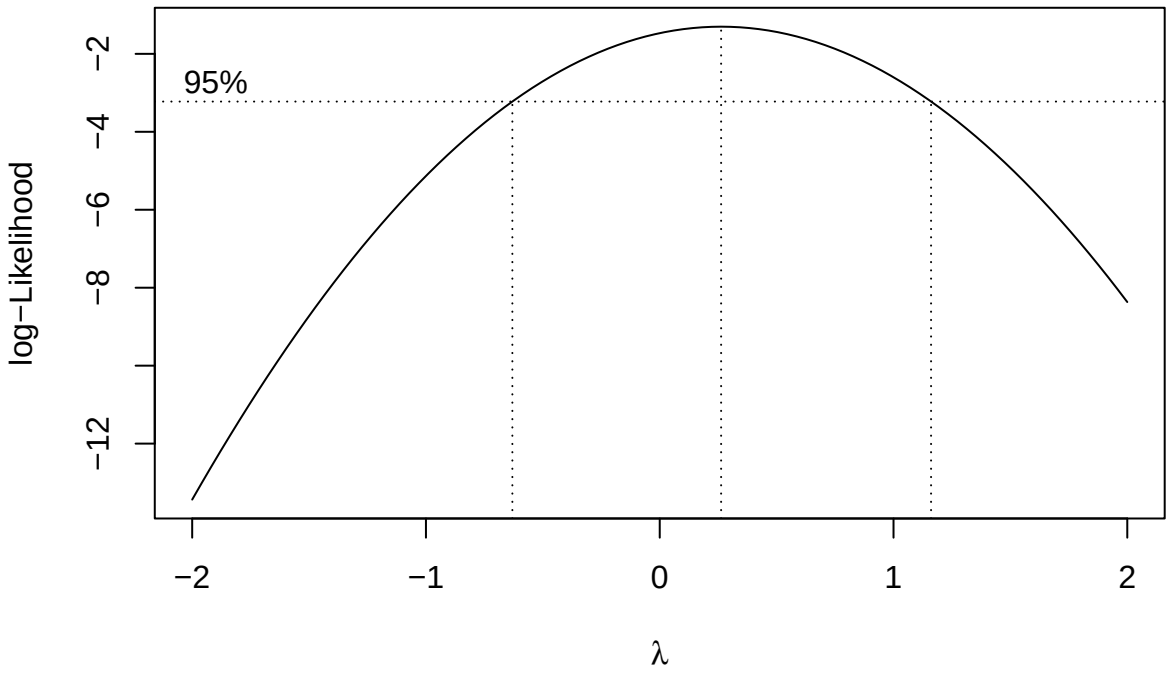

Figure 3: Box-cox transformation on c_train

We can further ensure that the log transformation is suitable for our data by contrasting the time series plot, histogram, and Q-Q plots of c_train and milk.log. We shall execute log transformation and continue working with milk.log. The next thing we'll do is to decompose the seasonality and trend of milk.log.

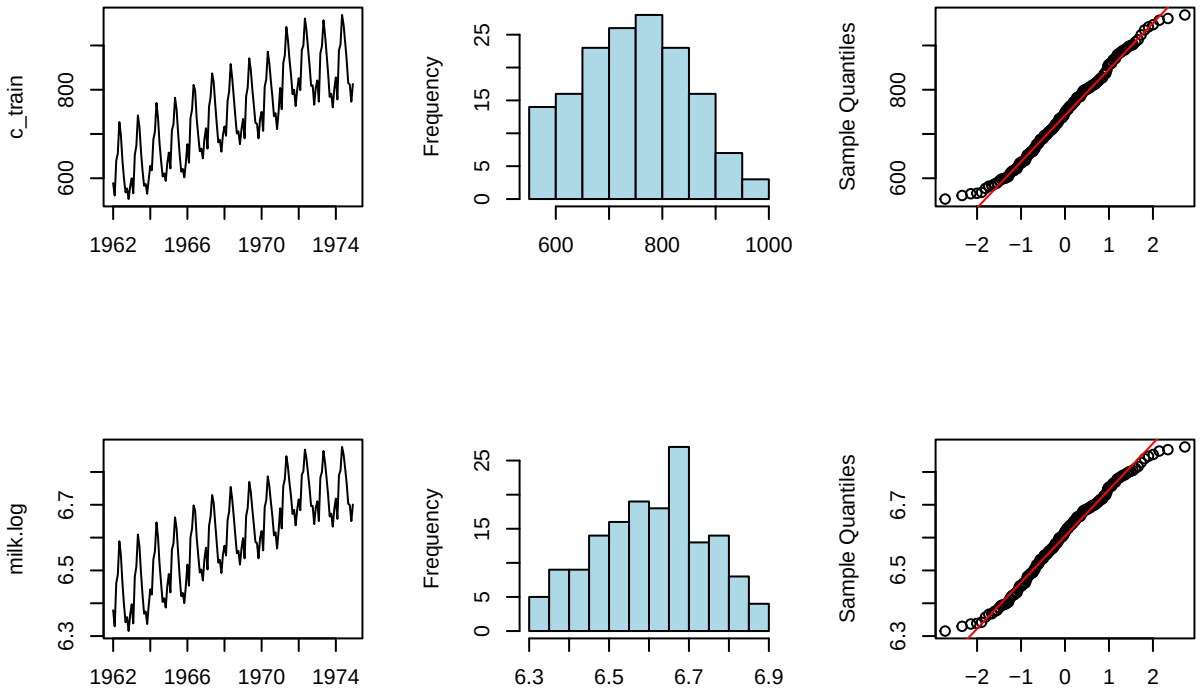

Figure 4: Comparison on c_train and milk.log: *Top Left*: Time Series Plot on c_train; *Top Middle*: Histogram of c_train; *Top Right*: Normal Q-Q Plot of c_train; *Bottom Left*: Histogram of milk.log; *Bottom Middle*: Histogram of milk.log; *Bottom Right*: Normal Q-Q Plot of milk.log

According on the decomposition plot, we predict that the difference between ∇_{12} and ∇_1 will cause our dataset to become stationary.

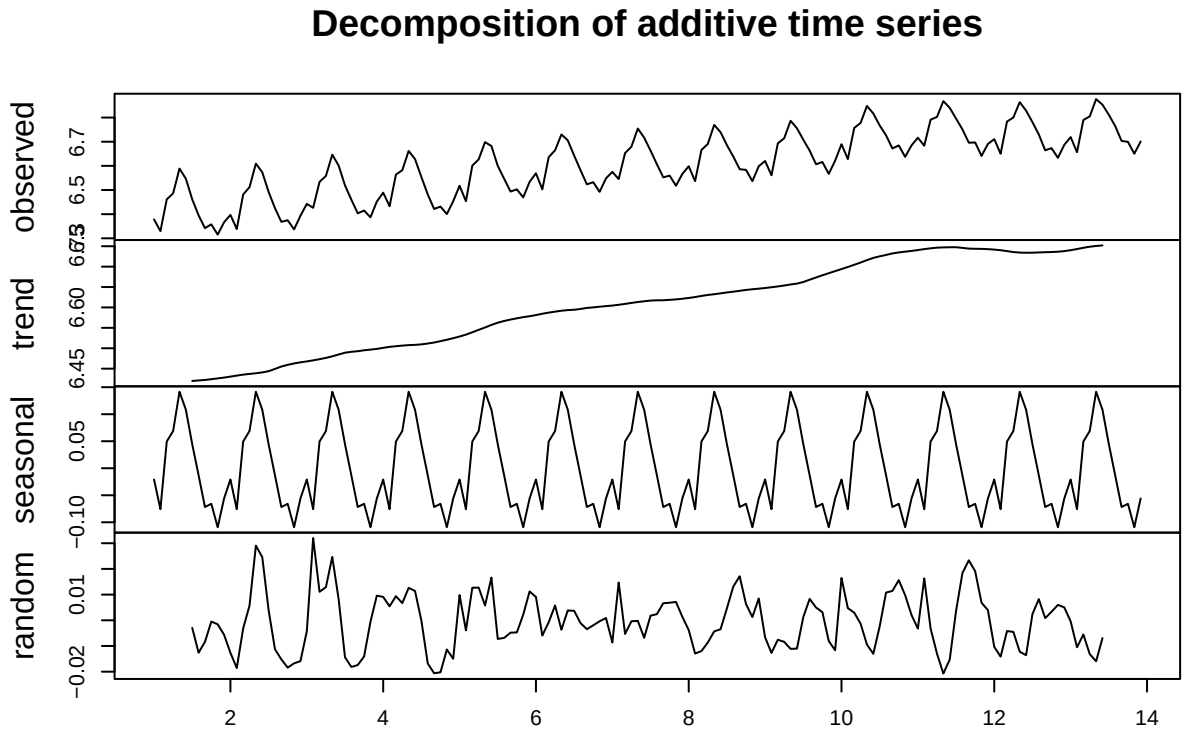

Figure 5: c_train: Decomposition of additive time series

The graph doesn't reveal any repeating patterns after we differentiate milk.log at ∇_{12}. However, a negative trend is still clearly visible. We therefore differenciate once more at ∇_1. Compared to the original, $\nabla_1\nabla_{12}$milk.log appears to be stationary. The graph's mean is also close to zero.

We keep making differences at lag 1. Our decision on the number of differences is finalized by comparing $\nabla_1\nabla_{12}$milk.log, and $\nabla_1\nabla_1\nabla_{12}$milk.log.

	$ln(U_t)$	$\nabla_{12}ln(U_t)$	$\nabla_1\nabla_{12}ln(U_t)$	$\nabla_1\nabla_1\nabla_{12}ln(U_t)$
Variance	0.018437	0.0004901	0.0001696	0.0004243

By comparing the variances above, we get the smallest variance is 0.0001696 when differenciating it at lag 12 and lag 1. However, it goes up to 0.0004243 with an additional difference at lag 1. As a result, we should stop the difference and choose $\nabla_1\nabla_{12}$milk.log.

The graphs below show a comparison of the ACF and PACF for milk.log, ∇_{12}milk.log, and $\nabla_1\nabla_{12}$milk.log. After being differentiated at lag 12, ∇_{12}milk.log does not exhibit spikes at multiples of lag 12. It still exhibits a declining pattern, indicating non-stationary. For $\nabla_1\nabla_{12}$milk.log, there are no decaying patterns after differenciating it at lag 1. A stationary process is indicated by the ACF and PACF plots of $\nabla_1\nabla_{12}$milk.log.

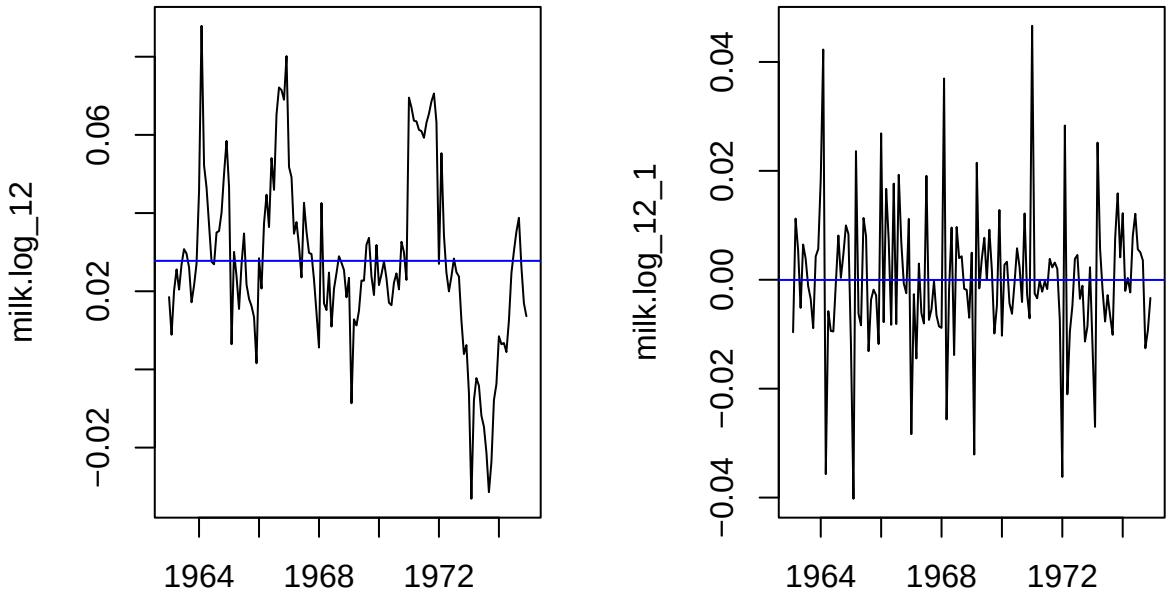

Figure 6: *Left*: Time Series Plot of $\nabla 12 ln(U_t)$; *Right*: Time Series Plot of $\nabla 1 \nabla 12 ln(U_t)$

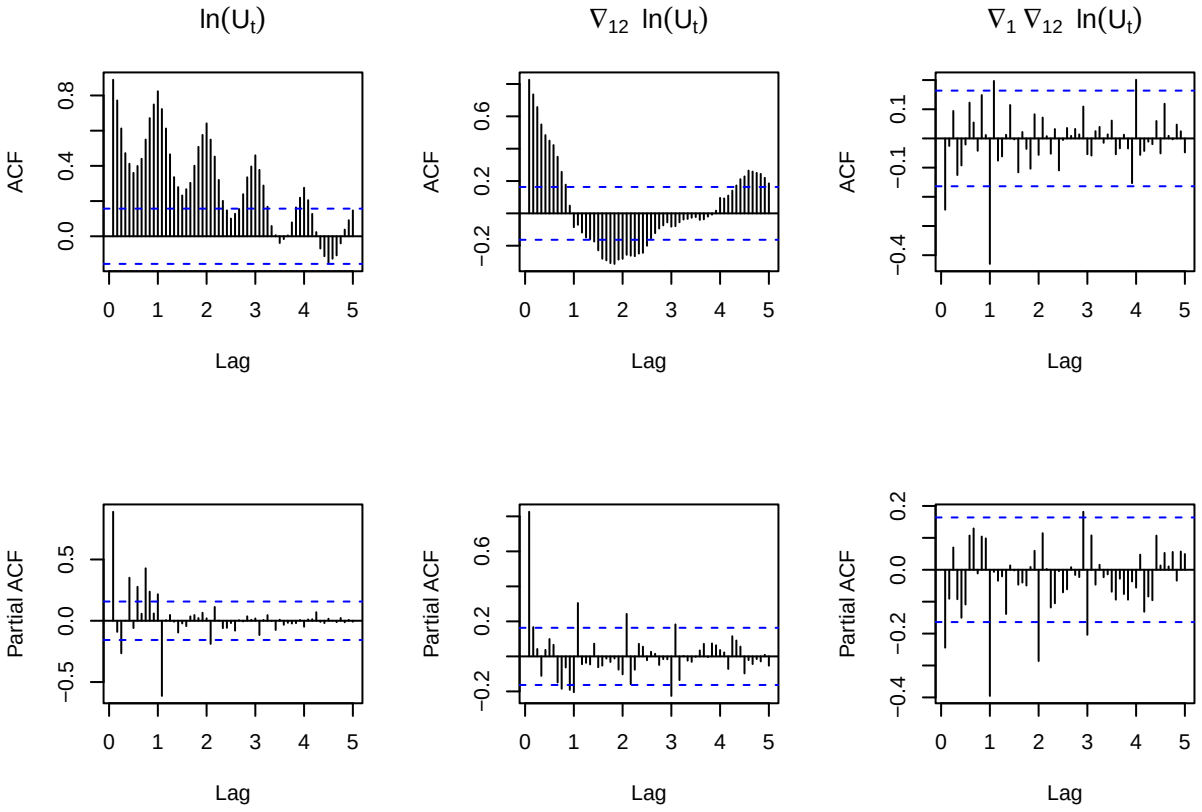

Figure 7: *Top Left*: ACF of $ln(U_t)$; *Top Middle*: ACF of $\nabla 12 ln(U_t)$; *Top Right*: ACF of $\nabla 1 \nabla 12 ln(U_t)$; *Bottom Left*: PACF of $ln(U_t)$; *Bottom Middle*: PACF of $\nabla 12 ln(U_t)$; *Bottom Right*: PACF of $\nabla 1 \nabla 12 ln(U_t)$

1.2 Model Identification

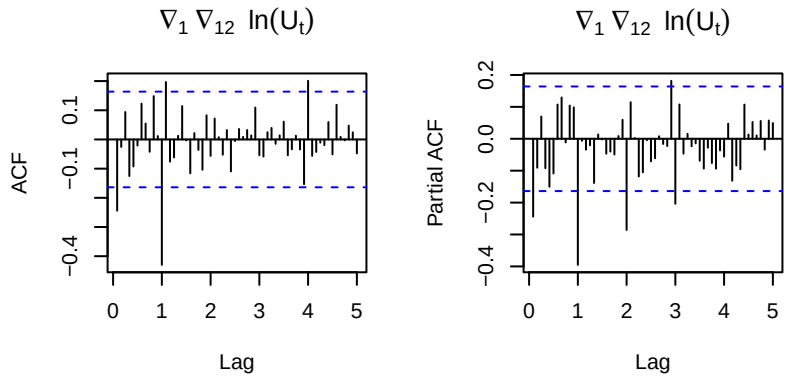

Figure 8: ACF and PACF of $\nabla 1 \nabla 12 ln(U_t)$

We choose to apply a SARIMA model for this dataset. Since we difference at lag 1 and lag 12, d = 1 and D = 1. The ACF plot of $\nabla_1 \nabla_{12}$milk.log shows spikes at lag 1, 12, 13 and 48. Then it cuts off at lag 48. Lag 13 have values outside of the confidence interval probably because of the influence from lag 1. Therefore, q = 1 and Q = 1 or 4. We see an alternating decaying pattern in PACF, which suggest P = 1 and the model might be a pure MA model. Within lag 12, PACF is significant at lag 1. Therefore, p = 1 and P = 1.

1.3 Model Estimation

The possible model suggested by ACF and PACF are $SARIMA(0,1,1)(0,1,1)_{12}$ and $SARIMA(0,1,1)(0,1,4)_{12}$. We compare the AICc of them.

	AICc
$SARIMA(0,1,1)(0,1,1)_{12}$	-882.5573
$SARIMA(0,1,1)(0,1,4)_{12}$	-885.7119

$SARIMA(0,1,1)(0,1,4)_{12}$ have the lowest AICc values. Therefore, we are going to pick $SARIMA(0,1,1)(0,1,4)_{12}$ for further analysis.

1.4 Model Checking

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

```
arima(milk.log, order = c(0,1,1), seasonal = list(order = c(0,1,4), period = 12), method = "ML")
```

```
##
## Call:
## arima(x = milk.log, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 4),
##     period = 12), method = "ML")
##
## Coefficients:
##           ma1     sma1     sma2    sma3    sma4
##       -0.1722  -0.6771  -0.0073  0.0132  0.2995
## s.e.   0.0869   0.1065   0.0983  0.1210  0.1361
##
## sigma^2 estimated as 9.925e-05:  log likelihood = 449.06,  aic = -888.11
```

The confidence intervals of sma2 and sma3 contain 0, so they are necessary to be fixed by zero.

```
arima(milk.log, order=c(0,1,1), seasonal = list(order = c(0,1,4), period = 12), fixed = c(NA,NA,0,0,NA)
```

```
##
## Call:
## arima(x = milk.log, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 4),
##     period = 12), fixed = c(NA, NA, 0, 0, NA), method = "ML")
##
## Coefficients:
##           ma1     sma1  sma2  sma3    sma4
##       -0.1729  -0.6771     0     0  0.3054
## s.e.   0.0866   0.0901     0     0  0.1200
##
## sigma^2 estimated as 9.928e-05:  log likelihood = 449.05,  aic = -892.1
```

```
polyroot(c(1, -0.1729))
```

```
## [1] 5.78369+0i
```

The root is outside unit circle, indicating the model is invertible.

The next step is to look at the Model residual. No trend, no seasonality, and no discernible variance change can be seen in the residual plot. Its mean is close to zero.

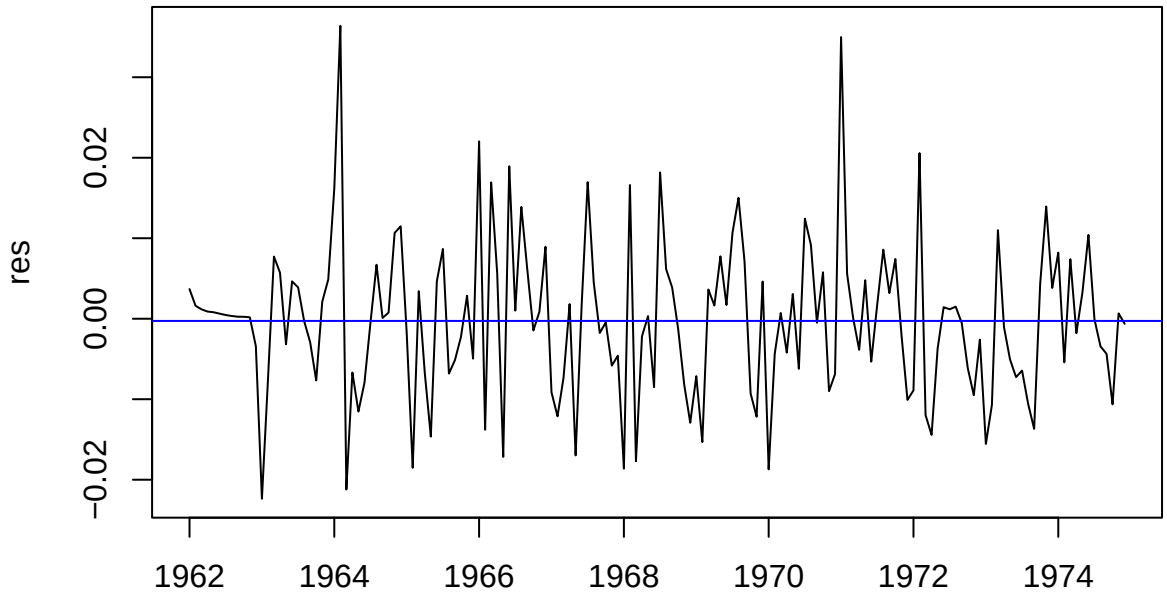

Figure 9: Time Series Plot of Model 1 $SARIMA(0,1,1)(0,1,4)_{12}$ residual

The histogram looks normal and the mean of it is nearly at zero. Also, the majority of the data lies on the qqline.

The ACF and PACF graphs show that Lag 4 is somewhat outside the 95% confidence interval. However, it can be counted as 0. All of other lags fall inside the confidence interval. We can infer that the residual of the model follows the white noise.

1.5 Model Diagnostics

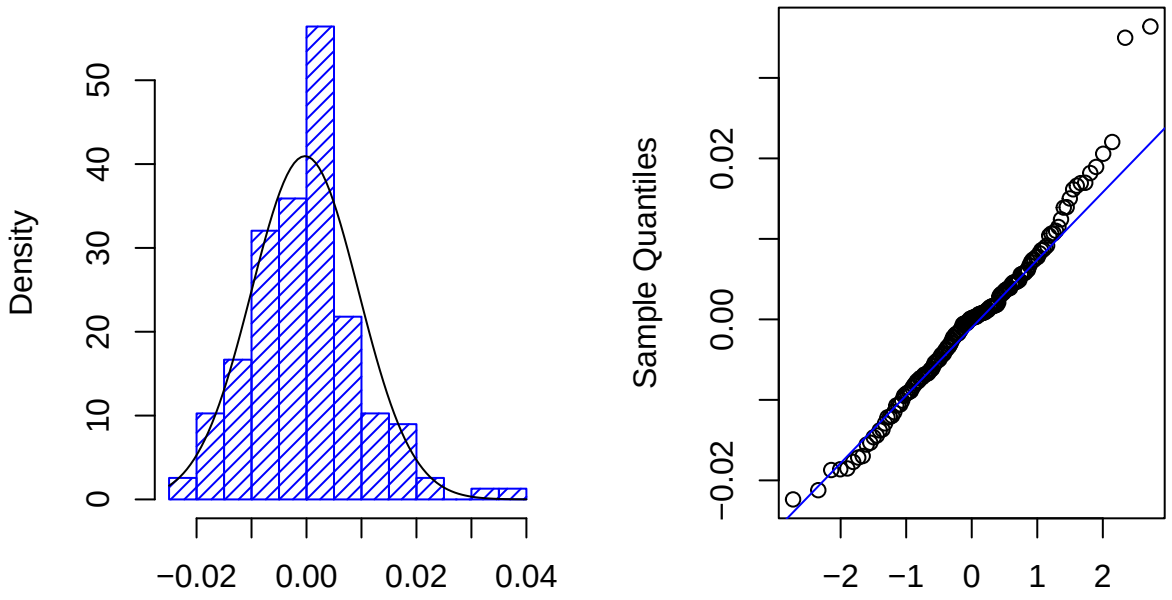

Figure 10: Model 1 $SARIMA(0,1,1)(0,1,4)_{12}$: *Left*: Histogram of residual; *Right*: Normal Q-Q Plot of residual

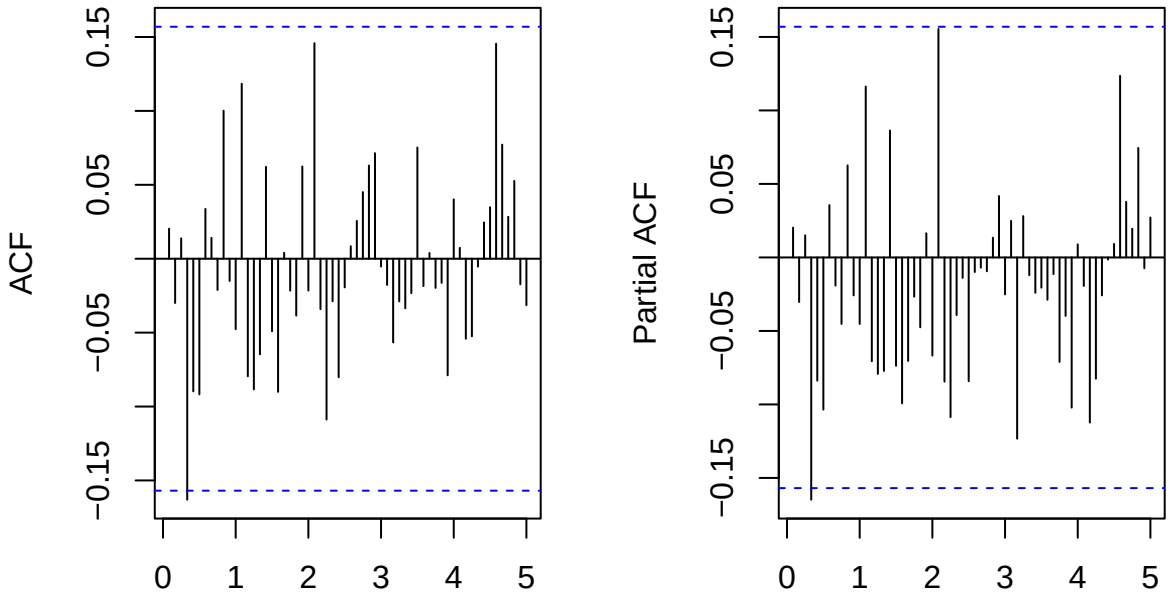

Figure 11: Model 1 $SARIMA(0,1,1)(0,1,4)_{12}$: *Left*: ACF of residual; *Right*: PACF of residual

Model: $SARIMA(0,1,1)(0,1,4)_{12}$

Since $\sqrt{156} \approx 12.49$, we set lag $= 12$ to perform model diagnostics. Also, since there are three parameters at the model functions, we set fitdf $= 3$ to do Box-Pierce test and Box-Ljung test .

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

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.9726, p-value = 0.003364
```

```
Box.test(res,lag = 12, type = c("Box-Pierce"),fitdf = 3) # 3 parameters
```

```
##
##  Box-Pierce test
##
## data:  res
## X-squared = 9.1806, df = 9, p-value = 0.4208
```

```
Box.test(res,lag = 12, type = c("Ljung-Box"),fitdf = 3) # 3 parameters
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 9.6626, df = 9, p-value = 0.3785
```

```
Box.test(res^2,lag = 12, type = c("Ljung-Box"),fitdf = 0)
```

```
##
##  Box-Ljung test
##
## data:  res^2
## X-squared = 16.63, df = 12, p-value = 0.1641
```

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

```
##
## Call:
## ar(x = res, aic = TRUE, order.max = NULL, mehod = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  9.492e-05
```

It passes all the diagnostic tests but not Shapiro-Wilk normality test. That's because the originial data is non-gaussian distribution, but the box-cox jenkins is already the best one. It cannot pass the Shapiro_Wilk normality test due to the influence of heavy tail. Also, there are two leverage points on the right side at qq-norm, which can affect the residual's normality.

1.6 Spectral Analysis

We investigate the periodicity of the Model residual before proceeding to our final forecasting step. The absence of strong spikes in the periodogram suggests that the residual lacks periodicity. We also fail to reject the null hypothesis that the residual is white noise in the Kolmogorov-Smirnov test.

```
fisher.g.test(res)
```

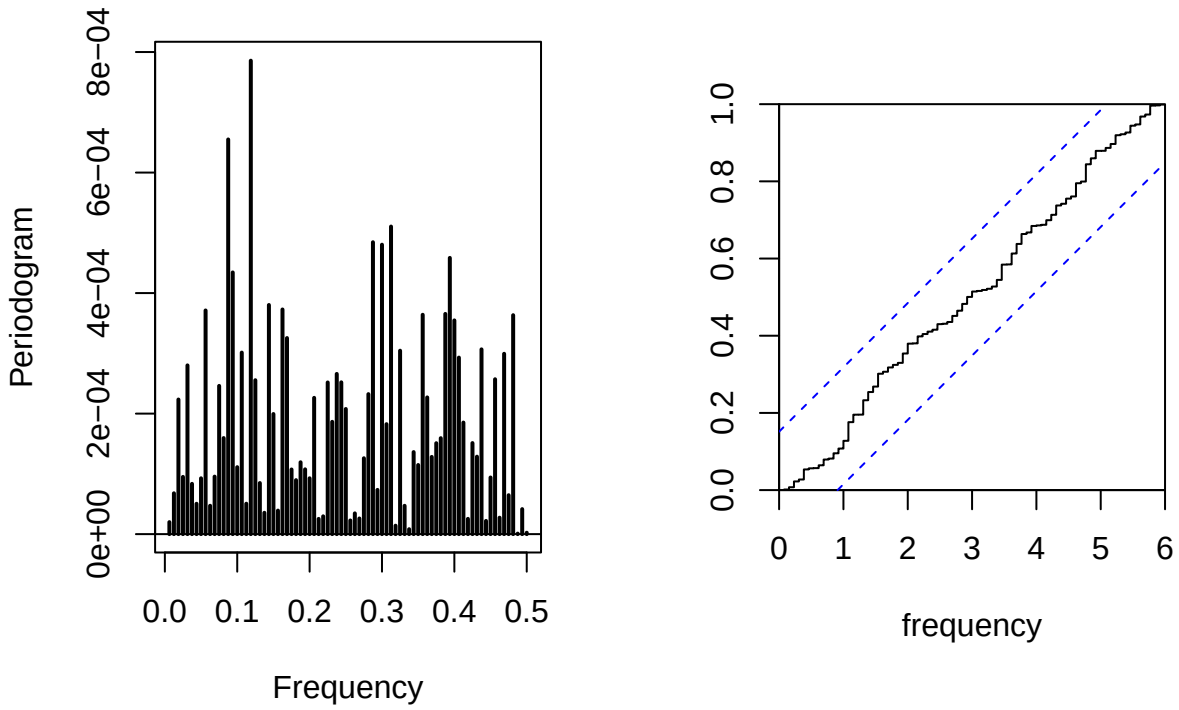

Figure 12: Model $SARIMA(0,1,1)(0,1,4)_{12}$: *Left*: Periodogram of residual; *Right*: Kolmogorov-Smirnov Test

```
## [1] 0.7995265
```

With a value of 0.7995265, the fisher test cannot reject the null hypothesis that the residual is Gaussian white noise. Then, we are confident to use our model for predicting after carefully examining the residual plots and passing the normality and periodicity tests.

Therefore, our final model is $SARIMA(0,1,1)(0,1,4)_{12}$

$$\nabla_1\nabla_{12}ln(U_t) = (1 - 0.1729_{(0.0866)}B)(1 - 0.6771_{(0.0901)}B^{12} + 0.3054B^{48}_{(0.1200)})Z_t$$

with $Zt - WN(0, \sigma_z^2)$ and $\sigma_z^2 = 9.928 * 10^{-5}$

1.7 Data Forecast

Firstly, we forecast on the transformed data by time series.

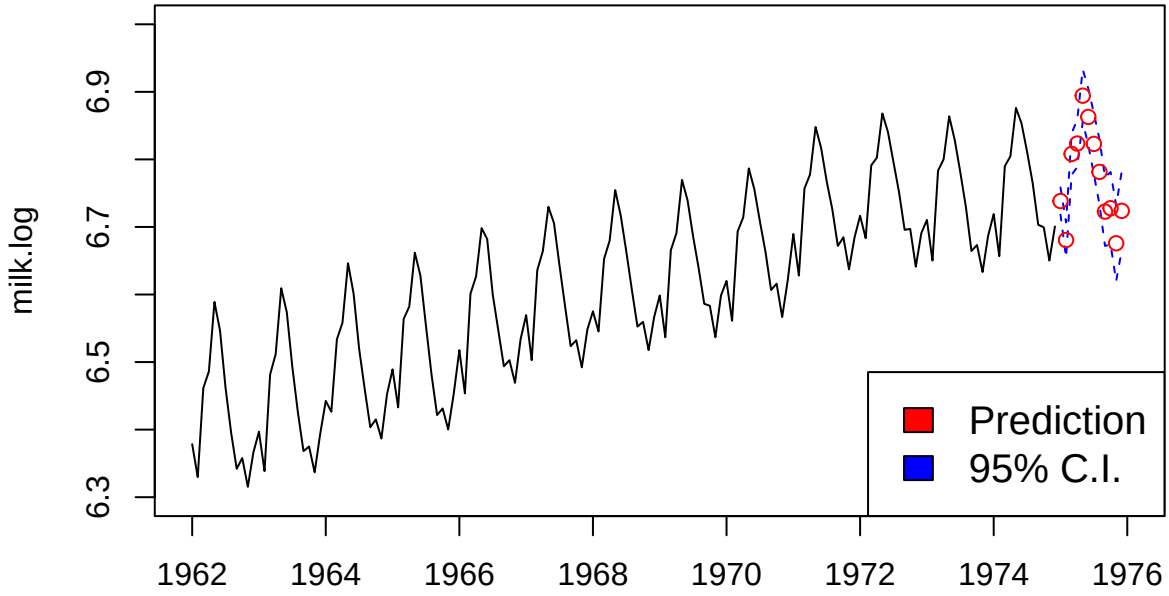

Figure 13: Forecast on the transformed data by time series

Next, we forecast on the original data by time series.

Then, we zoom in the forecast on the original data by time series.

Finally, we forecast on the original data set by time series with c_test data.

All testing set data fall inside the 95% confidence interval of prediction model. The numbers predicted by the chosen model are also fairly similar to the actual values, demonstrating the suitability and sufficiency of our model for forecasting milk production.

2.0 CONCLUSION

After comparing AICc values of two possible models - $SARIMA(0,1,1)(0,1,1)_{12}$ and $SARIMA(0,1,1)(0,1,4)_{12}$. The model $SARIMA(0,1,1)(0,1,4)_{12}$ has smaller AICc and provide credible forecasting results for our data set. It passes all diagnostic tests but not Shapiro-Wilk normality test. That's because the originial data is non-gaussian distribution, but the box-cox jenkins is already the best one. Also, it cannot pass the Shapiro_Wilk normality test due to the influence of heavy tail. There are two levergae opints on the right side at the qq norm plot, which can affect the residual's normality. However, we are still confident to claim that our $SARIMA(0,1,1)(0,1,4)_{12}$ model is appropriate to forecast the monthly measurements of milk production (pounds per cow).

The final model is:

$$\nabla_1\nabla_{12}ln(U_t) = (1 - 0.1729B)(1 - 0.6771B^{12} + 0.3054B^{48})Z_t$$

with

$$Z_t - WN(0, \sigma_z^2)$$

and

$$\sigma_z^2 = 9.928 * 10^{-5}$$

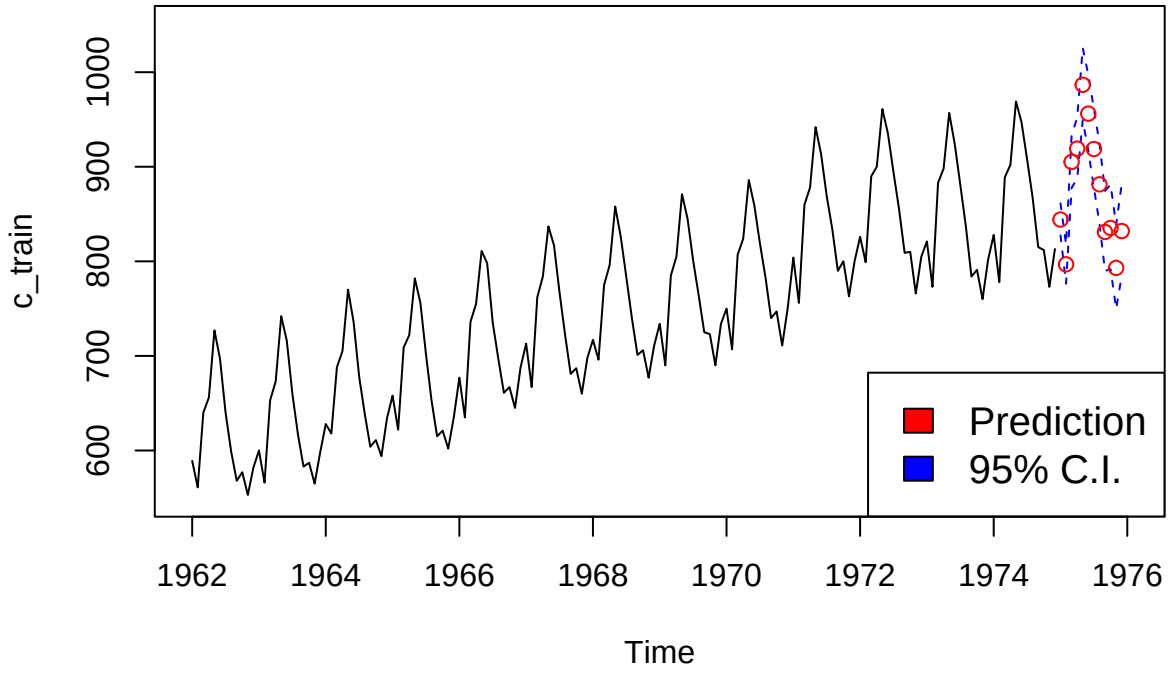

Figure 14: Forecast on the original data by time series

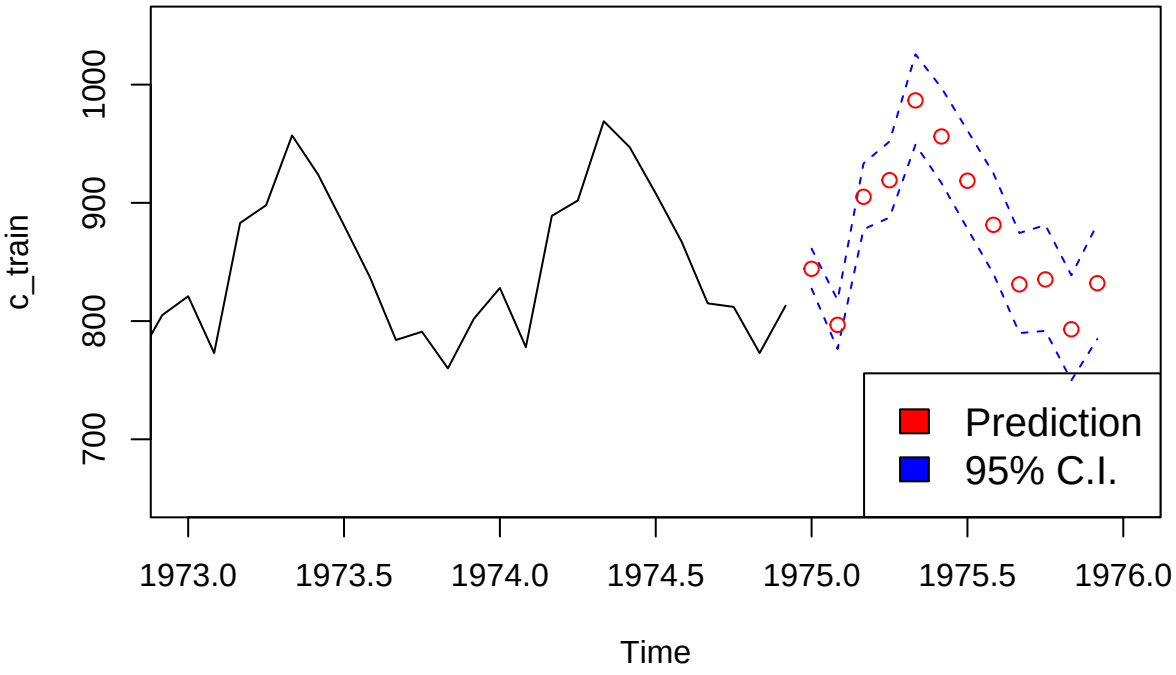

Figure 15: Zoomed in: Forecast on the original data set by time series

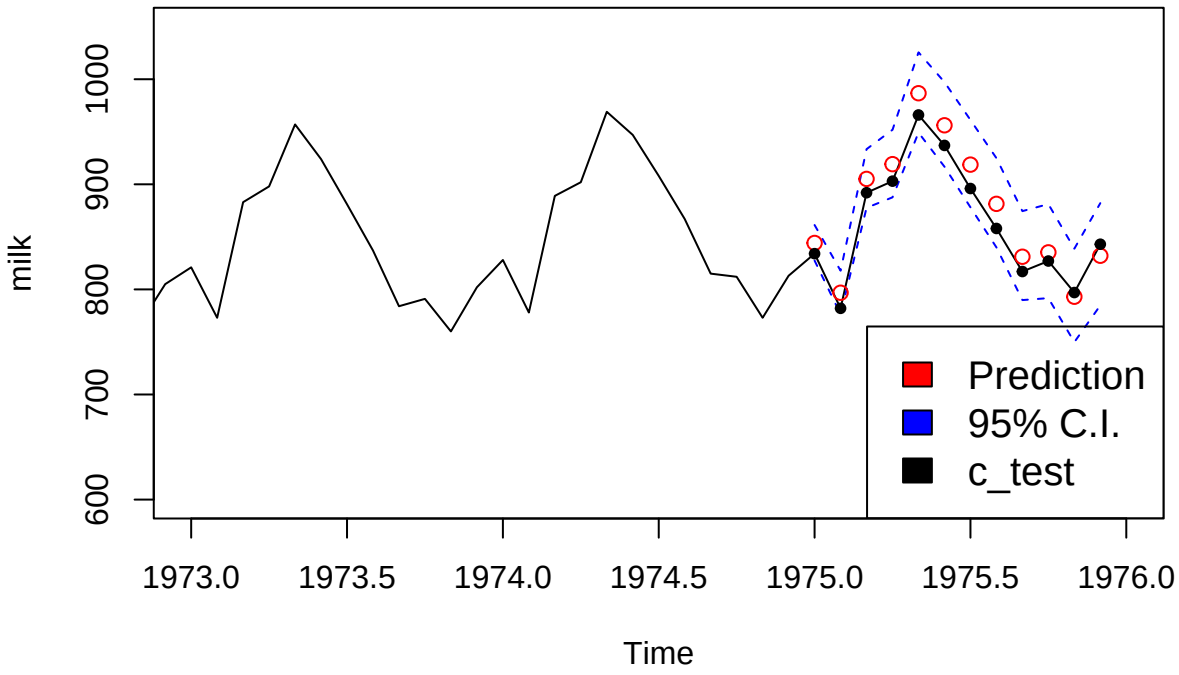

Figure 16: Zoomed in: Forecast on the original data set by time series with c_test data

3.0 REFERENCES

Feldman, R. (2021, December). *PSTAT 274 - Lecture 15. Fall 2022.* Santa Barbara; University of California, Santa Barbara.

Hyndman, R., & Yang, Y. (2018). tsdl: Time Series Data Library. v0.1.0.

4.0 APPENDIX

4.1 Library Used

```
library(tsdl)
library(forecast)
library(tidyverse)
library(MASS)
library(ggplot2)
library(ggfortify)
library(forecast)
library(GeneCycle)
library(qpcR)
require(TSA)
```

4.2 Data Processing

```
# View the data set NO.203 in tsdl
length(tsdl[[203]])
attr(tsdl[[203]],"subject")
attr(tsdl[[203]],"source")
attr(tsdl[[203]],"description")
```

```
# View the Original Data
milk <- tsdl[[203]]
plot.ts(milk,xlab = "",main = "")
# Set up training and testing group
# c_train totally 156 points, c_test totally 12 points
c_train <- ts(milk[1:156],start = c(1962,1),frequency = 12)
c_test <-ts(milk[157:168],start = c(1975,1),frequency = 12)
# Show histogram and acf plot of c_train
par(mfrow=c(1,2))
hist(c_train, col = "light blue",main = "")
acf(c_train, lag.max = 40, main = "")
# Perform box-cox transformation on c_train
bcTransform <- boxcox(c_train~as.numeric(1:length(c_train)))
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
c_train.bc = (1/lambda)*(c_train^lambda-1)
# Perform log transformation on c_train
milk.log <- log(c_train)
plot.ts(milk.log)
# Compare c_train and milk.log
par(mfrow=c(2,3))
plot.ts(c_train,xlab = "", main = "")
hist(c_train, col = "light blue", xlab = "", main = "")
qqnorm(c_train, main = "", xlab = "")
qqline(c_train, col = "red")
plot.ts(milk.log,xlab = "", main = "")
hist(milk.log, col = "light blue", xlab = "", main = "")
qqnorm(milk.log, main = "", xlab = "")
qqline(milk.log, col = "red")
# Decompose the trend and seasonality of c_train
x1 <- ts(as.ts(milk.log),frequency = 12)
decomp <- decompose(x1)
plot(decomp, xlab = "")
# Remove seasonality and trend
# ln(Ut) differenced at lag 12
var(milk.log)
milk.log_12 <- diff(milk.log, lag=12)
var(milk.log_12)
# ln(Ut) differenced at lag 12 and then lag 1
milk.log_12_1 <- diff(milk.log_12, lag=1)
var(milk.log_12_1)
# ln(Ut) differenced at lag 12, then lag 1 and then lag 1
milk.log_12_1_1 <- diff(milk.log_12_1, lag=1)
var(milk.log_12_1_1)
# Check differencing
par(mfrow=c(1,2))
plot.ts(milk.log_12, main = "")
abline(h=mean(milk.log_12), col = "blue")
plot.ts(milk.log_12_1, main = "")
abline(h=mean(milk.log_12_1), col = "blue")
# Create table to compare variance
v <- cbind(var(c_train), var(milk.log_12), var(milk.log_12_1), var(milk.log_12_1_1))
rownames(v) <- "Variance"
colnames(v) <- c("$U_t$","$\\nabla_{12}$$U_t$",
```

```
                "$\\nabla_{1}$$\\nabla_{12}$$U_t$",
                "$\\nabla_{1}$$\\nabla_{1}$$\\nabla_{12}$$U_t$")
knitr::kable(v)
# Compare the ACF and PACF of milk.log, milk.log_12, and milk.log_12_1
par(mfrow=c(2,3))
acf(milk.log, lag.max = 60, main = expression(ln(U[t])))
acf(milk.log_12, lag.max = 60, main = expression(nabla[12]~~ln(U[t])))
acf(milk.log_12_1, lag.max = 60, main = expression(nabla[1]~nabla[12]~~ln(U[t])))
pacf(milk.log, lag.max = 60, main = "")
pacf(milk.log_12, lag.max = 60, main = "")
pacf(milk.log_12_1, lag.max = 60, main = "")
# Decide the proper p,d,q,P,D,Q from the acf and pacf plots
acf(milk.log_12_1, lag.max = 60, main = expression(nabla[1]~nabla[12]~~ln(U[t])))
pacf(milk.log_12_1, lag.max = 60, main = "")
```

4.3 Model Estimation

```
# Calculate and compare the AICc for SARIMA (0,1,1)x(0,1,1) and SARIMA (0,1,1)x(0,1,4)
a <- AICc(arima(milk.log, order = c(0,1,1), seasonal = list(order = c(0,1,1),
                                                            period = 12), method = "ML"))
b <- AICc(arima(milk.log, order = c(0,1,1), seasonal = list(order = c(0,1,4),
                                                            period = 12), method = "ML"))
AICc_value <- rbind(a,b)
rownames(AICc_value) <- c("$SARIMA(0,1,1)(0,1,1)_{12}$", "$SARIMA(0,1,1)(0,1,4)_{12}$")
colnames(AICc_value) <- c("AICc")
knitr::kable(AICc_value)
```

4.4 Model Identification

$SARIMA(0,1,1)(0,1,4)_{12}$

```
# Estimate coefficients of the model
arima(milk.log, order = c(0,1,1), seasonal = list(order = c(0,1,4),
      period = 12), method = "ML")
arima(milk.log, order=c(0,1,1), seasonal = list(order = c(0,1,4),
      period = 12), fixed = c(NA,NA,0,0,NA), method="ML")
# Calculate the roots for check invertibility
polyroot(c(1, -0.1729))
# Plot the residuals
fit <- arima(milk.log, order = c(0,1,1), seasonal = list(order = c(0,1,4),
             period = 12), method = "ML")
res <- residuals(fit)
plot.ts(res, xlab = "")
abline(h=mean(res),col = "blue")
# Plot the histogram, Q-Q normal plot, ACF and PACF of the residual
par(mfrow=c(1,2))
hist(res, density = 20, breaks = 20, col = "blue", xlab = "",
     prob = TRUE, main = "")
mean <- mean(res)
std <- sqrt(var(res))
curve(dnorm(x,mean,std),add = TRUE)
qqnorm(res,main = "", xlab = "")
qqline(res, col = "blue")
```

```r
par(mfrow=c(1,2))
acf(res,lag.max=60, main = "", xlab = "")
pacf(res,lag.max=60, main = "", xlab = "")
```

4.6 Model Diagnostics

```r
# Perform test on Model's residual
shapiro.test(res1)
Box.test(res1,lag = 20, type = c("Box-Pierce"),fitdf = 2)
Box.test(res1,lag = 20, type = c("Ljung-Box"),fitdf = 2)
Box.test(res1^2,lag = 20, type = c("Ljung-Box"),fitdf = 0)
ar(res1,aic=TRUE,order.max=NULL, mehod = c("yule-walker"))
```

4.7 Spectral Analysis

```r
# install.packages("TSA")
require(TSA)
par(mfrow=c(1,2))
# Graph the periodogram of Model residual
periodogram(res, main = "")
abline(h = 0)
# Perform Kolmogorov-Smirnov Test on Model residual
cpgram(res, main = "")
# Perform fisher test on Model residual
fisher.g.test(res)
```

4.8 Data Forecast

```r
fit.1<-arima(milk.log, order = c(0,1,1),
             seasonal = list(order = c(0,1,4),
             period = 12),
             fixed = c(NA,NA,0,0,NA),
             method = "ML")
# Forecast on transformed data:
pred.tr <- predict(fit.1, n.ahead = 12)
U.tr = pred.tr$pred + 2*pred.tr$se
L.tr = pred.tr$pred - 2*pred.tr$se
ts.plot(milk.log,xlim = c(1962,1976), xlab = "", ylim= c(6.3,7.0))
lines(U.tr, col = "blue", lty = "dashed")
lines(L.tr, col = "blue", lty = "dashed")
points(pred.tr$pred,col = "red", pch = 1)
legend("bottomright", c("Prediction","95% C.I."),
       fill = c("red", "blue"), cex = 1.25)
# Forecast on original data
pred.orig <- exp(pred.tr$pred)
U = exp(U.tr)
L = exp(L.tr)
ts.plot(c_train, xlim= c(1962,1976), ylim= c(550,1050))
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points(pred.orig,col = "red", pch = 1)
legend("bottomright", c("Prediction","95% C.I."),
```

```
       fill = c("red", "blue"), cex = 1.25)
# Zoom in graph
pred.orig <- exp(pred.tr$pred)
ts.plot(c_train, xlim= c(1973,1976),ylim=c(650,1050))
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points(pred.orig,col = "red", pch = 1)
legend("bottomright", c("Prediction","95% C.I."),
       fill = c("red", "blue"), cex = 1.25)
# Forecast with test dataset
pred.orig <- exp(pred.tr$pred)
ts.plot(milk, xlim= c(1973,1976),ylim=c(600,1050))
lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
points(pred.orig,col = "red", pch = 1)
points(c_test,col = "black", pch = 20)
legend("bottomright", c("Prediction","95% C.I.","c_test"),
       fill = c("red", "blue","black"), cex = 1.25)
```