Time Series Analysis of Milk Production (from Cows)

Daniel Mindlin

2019 M03 15

Contents

Abstract **2**

1 Introduction **2**

2 Dataset Breakdown **3**

2.1 Intital Analysis 3

2.2 Decomposition 4

3 Transformations **5**

3.1 Box-Cox Transform 5

3.2 Removing Seasonality and Trend 7

4 Detailed Model Analysis **8**

4.1 P, D, Q, p, d, q 8

4.2 Model Selection 9

4.3 Model Estimation 9

5 Diagnostics **9**

5.1 Normality 9

5.2 Independence 11

5.3 Constant Variance 12

6 Forecasting **12**

7 Conclusion **13**

8 References **14**

9 Appendix **14**

Abstract

Using detalied and comprehensive Time Series techniques, this project analyzes the per month, per cow production of milk, in pounds. Illustrated below are the results of said analysis, along with an attempt at forecasting future per month, per cow production of milk.The time series model constructed explains the behavior of per cow Milk production and endeavors to forecast it as well. Aspects of seasonality and other technical details of a Time Series are included.

1 Introduction

This project analyzes the trend and seasonal effect of Milk production. Milk is a staple of many diets. Providing a healthy source of protein and calcium, Milk is a delicious and nutritious drink which is clearly fascinating to study. The production of milk is also a very important commodity in financial markets worldwide, so analyzing its production could be incredibly useful. The project (and dataset) are in pounds of milk per cow (lbs per cow) and is adjusted to be on a monthly basis from January 1962 to December 1974 (155 entries).

Using R, I used BoxCox transformations, de-trended, and de-seasonalized the dataset.

I then selected a SARIMA model to describe the dataset and then used used AICc and BIC metrics to make sure it was was stationary and invertible.

I then used the Shapiro-Wilk test and Ljung-Box test as diagnostics to ensure the final model was satisfactory.

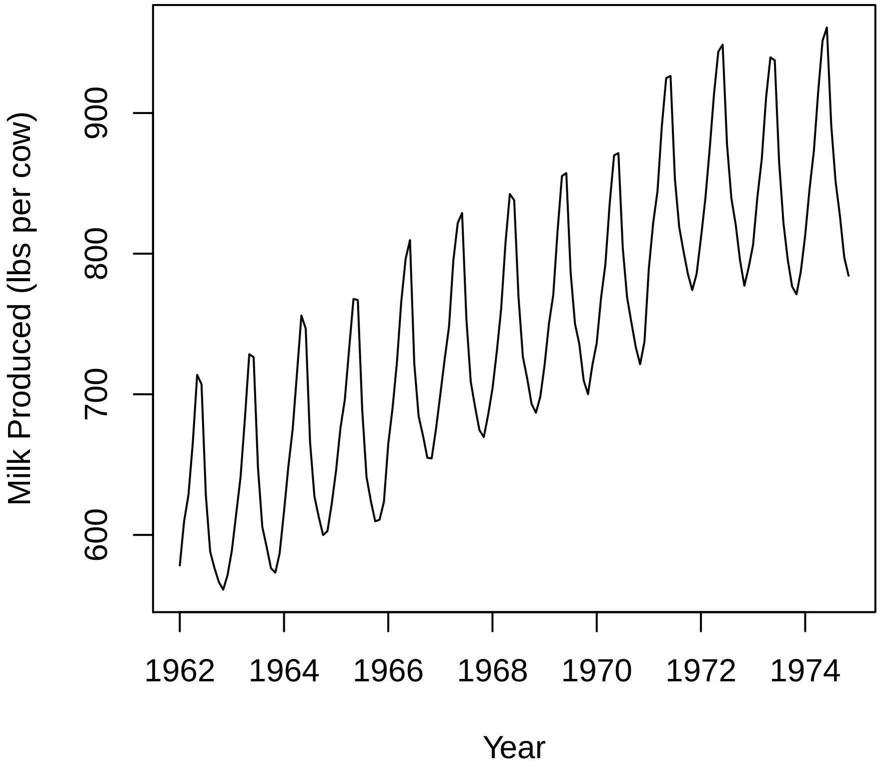
Since 2 of our models passed the necessary diagnostic checks, we chose the model based on the lowest BIC values with happened to be a $SARIMA(0,1,1)\mathbf{x}(1,1,0)_{12}$

Forecasting a year ahead showed that our forecasts coincided with the actual milk of those 12 periods.

2 Dataset Breakdown

2.1 Intital Analysis

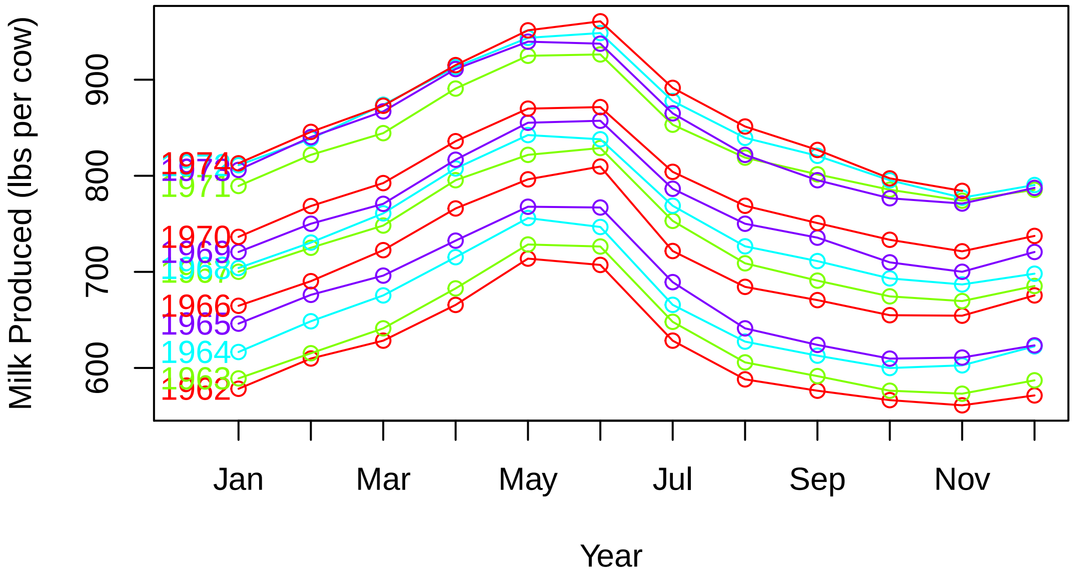
Figure 1: Monthly Milk Production, 1962−1974



The plot expresses a positive trend as well as annual seasonality. This seasonality is of note, showing that milk production must not have been constant over an annual period; cow output peaks in spring spring and reaches a minimum in autumn. This seasonality must be controlled for, otherwise the model will not be stationary. This cycle appears to be stable, never varying tremendously in range.

To illustrate this pattern, a monthly seasonal plot is edifying.

Figure 2: Seasonal Plot



This verifies what I thought from looking at the plot: Production peaks in Spring (May) and tumbling until Autumn (November). This cycle suggests seasonality in this dataset. Generally, milk production grows from year as time goes on as well.

2.2 Decomposition

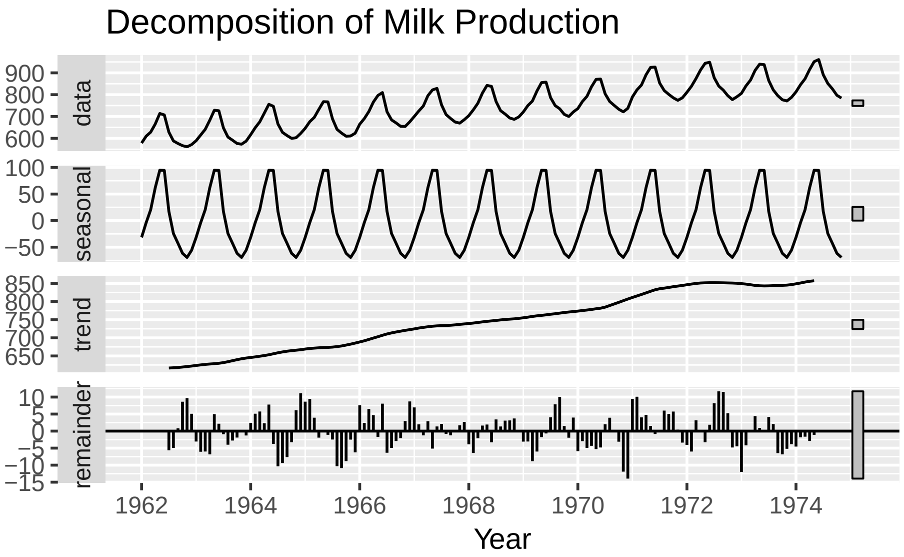
X_t is our milk production data. X_1 falls on January 1962, X_{155} falls on December 1974.

Trend: m_t

Seasonality: s_t

Error Term: Z_t

Decomposition Model: $X_t = m_t + s_t + Z_t$

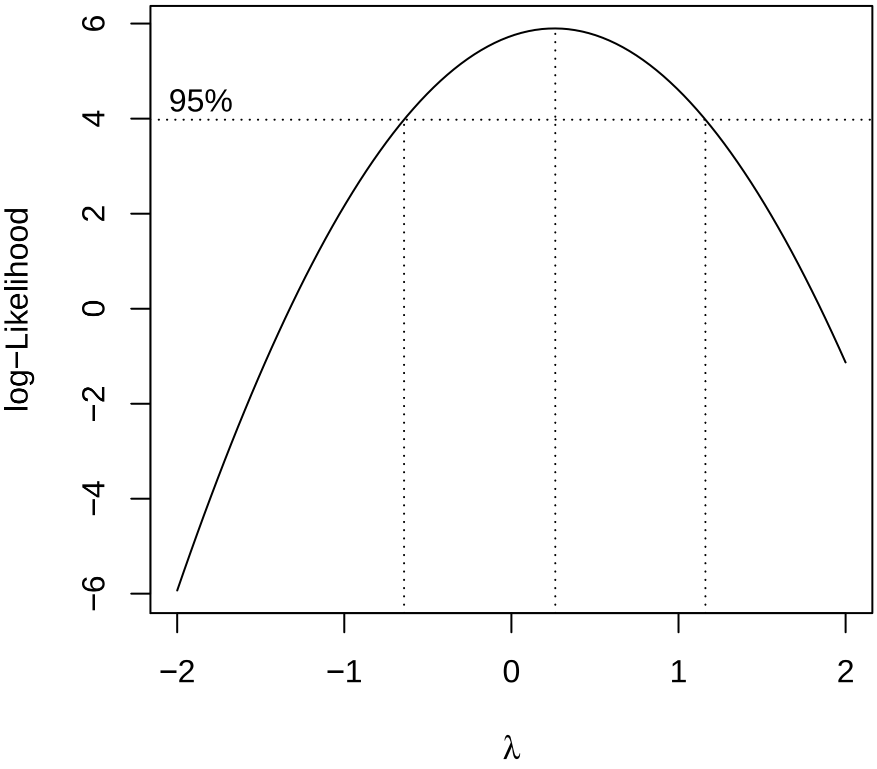


This verifies the same seasonal and upwards trends as noted above. The remainder *appears* stationary.

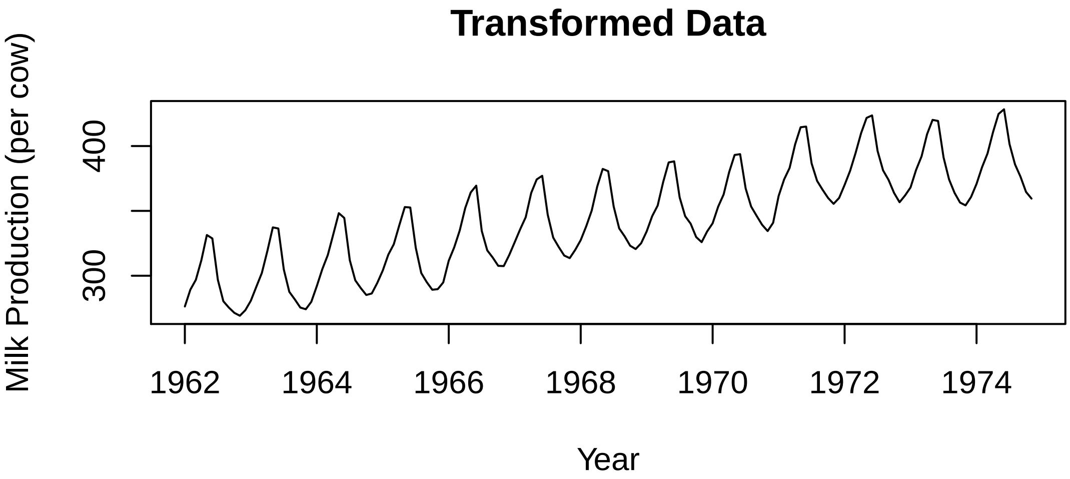
3 Transformations

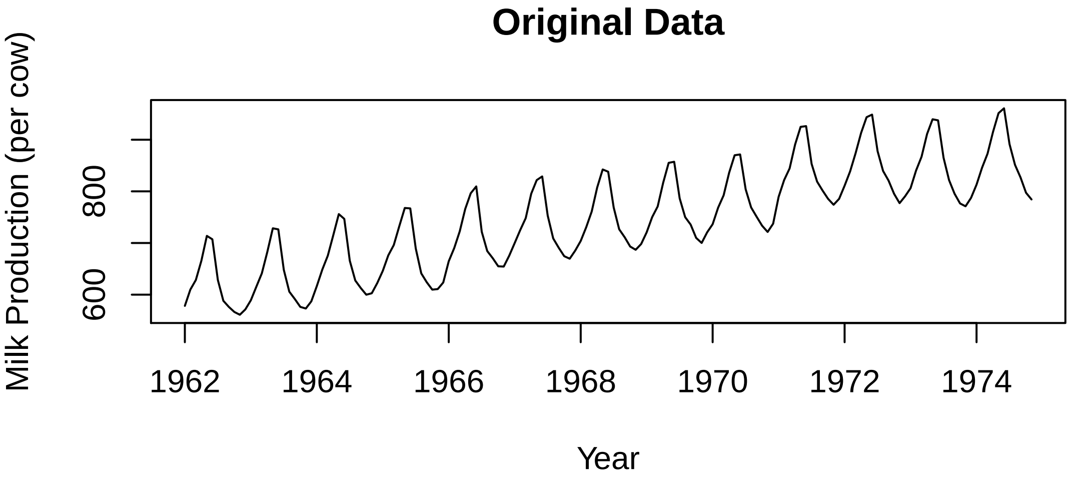
3.1 Box-Cox Transform

In order to stabilize the variance of milk production (per cow), I use Box-Cox to select a λ, selected to maximize log-likelihood.



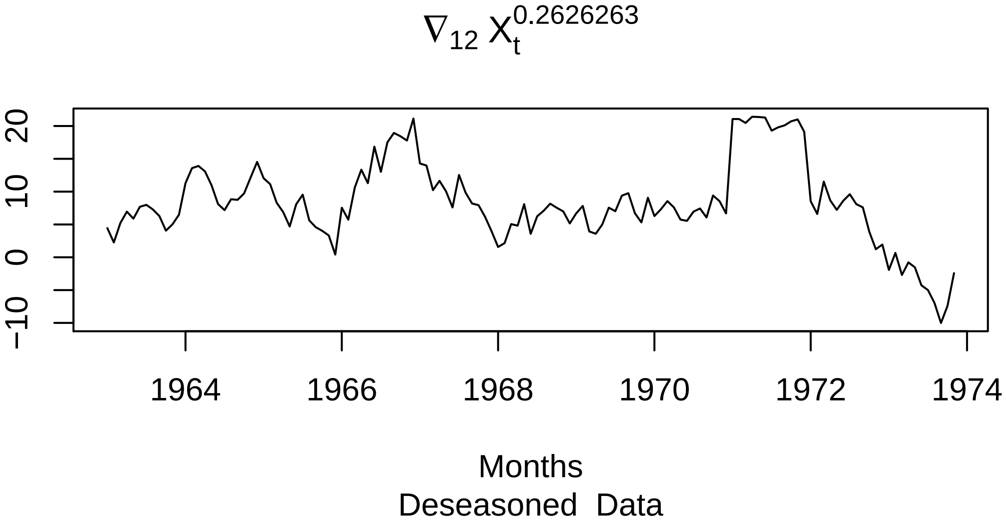
log-likelihood maximized at $\lambda = 0.2626263 \rightarrow Y_t^{0.2626263}$.

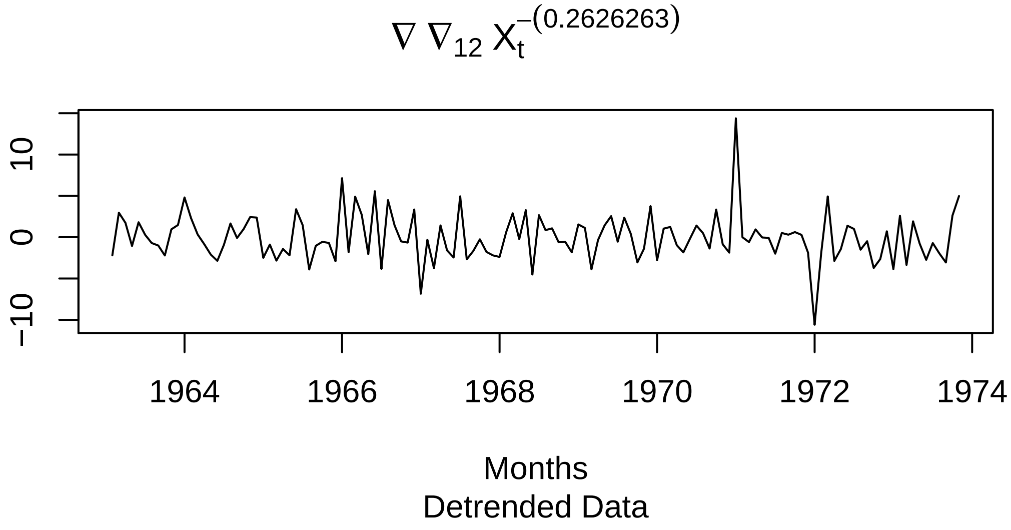




3.2 Removing Seasonality and Trend

To remove the seasonality and trend from the dataset, the differencing method is useful. Seasonality is the first to go.

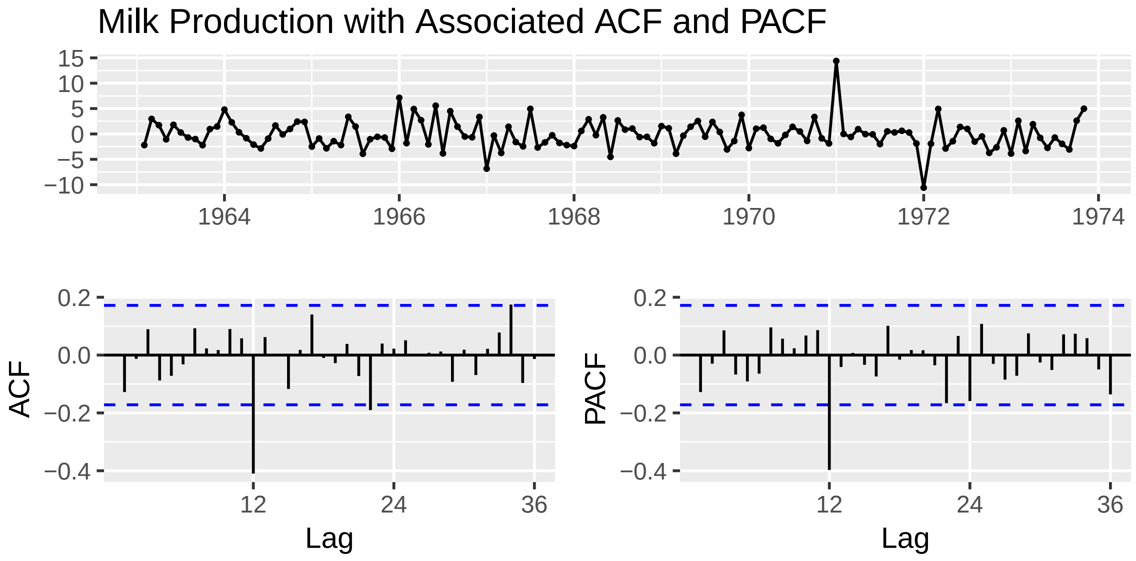




In the above plots, lag = 12 is used because I observe an annual cycle in the data. Utilizing an Augmented Dickey-Fuller Test with a null hypothesis that X_t is rejected yields a p-value of 0.01. This null hypothesis is therefore rejected, showing that the dataset is deseasonalized, detrended, and stationary.

4 Detailed Model Analysis

4.1 P, D, Q, p, d, q



ACF: Plot behavior at lag = 12 shows that AR, P = 1 is ideal. The strong upward trend suggests that ARIMA better, however.

PACF: Plot behavior shows that there is no seasonality, so Q = 0.

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

To identify p, d, and q, pertinent spikes appear at lag = 1,3. This suggests a MA(3) process. Furthermore, differencing the deseasonalized data at lag =2 shows an increase in variance, so d = 1.

4.2 Model Selection

2 models must be selected at this stage. Akaike's Corrected Information Criterion (AICc) and Bayesian Information Criterion (BIC) are useful tools for this. Both criterions measure a statistical model's quality in terms of Bias-Variance tradeoff. The lower a model's AICc and BIC values, the better.

Looking at our two tables of AICc and BIC values below, we see that the model with $p = 0, q = 3$ gives us the smallest AICc value and the model with $p = 0, q = 1$ gives us the smallest BIC value. This is expected as BIC has a larger penalty parameter for model complexity and thus favors a smaller model.

Therefore, our two models selected based on AICc and BIC are:

Model 1: $\text{SARIMA}(0, 1, 3) \times (1, 1, 0)_{12}$

Model 2: $\text{SARIMA}(0, 1, 1) \times (1, 1, 0)_{12}$

AICc values:

	MA(0)	MA(1)	MA(2)	MA(3)
AR(0)	2.789762	2.799876	2.814449	2.829731
AR(1)	2.800134	2.814555	2.814272	2.827993
AR(2)	2.814388	2.829477	2.814196	2.814156
AR(3)	2.829407	2.826799	2.841829	2.857292

BIC values:

	MA(0)	MA(1)	MA(2)	MA(3)
AR(0)	1.795896	1.826121	1.860593	1.895559
AR(1)	1.826379	1.860700	1.880100	1.913283
AR(2)	1.860533	1.895305	1.899487	1.918684
AR(3)	1.895235	1.912089	1.946357	1.980826

4.3 Model Estimation

The coefficients and parameters now need to be tested. Assume that data has zero mean, p and q are known.

Results from MLE:

	Model 1	Model 2
MA(1)	-0.1086	-0.1131
MA(2)	0.0475	NA
MA(3)	0.0653	NA
SAR(1)	-0.4276	-0.4258

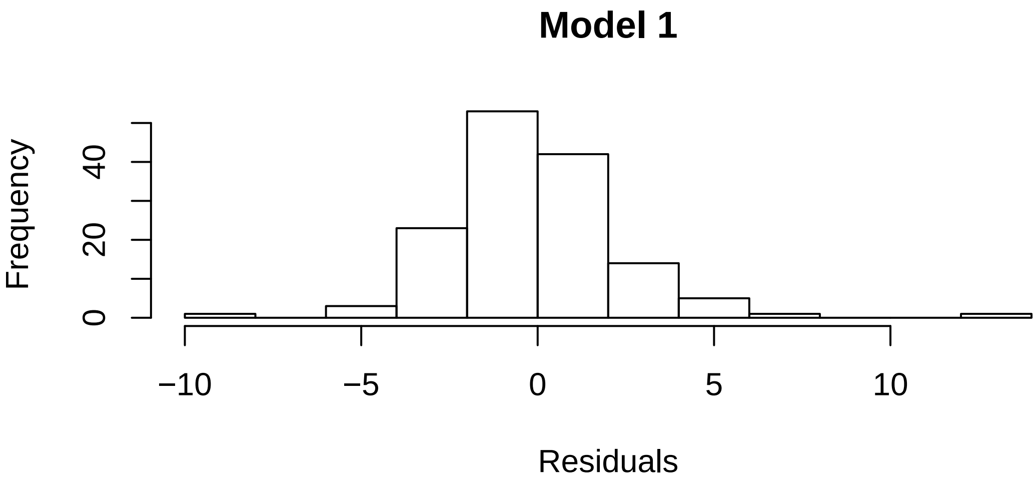
Model 1: SARIMA(0,1,3) $\times (1, 1, 0)_{12} = (1 + 0.4276B^{12})Y_t = (1 - 0.4276B)(1 + 0.4276B)(1 + 0.4276B)Z_t$.
Model 2: SARIMA(0,1,1) $\times (1, 1, 0)_{12} = (1 + 0.4258^{12})Y_t = (1 - 0.1131B)Z_t$.

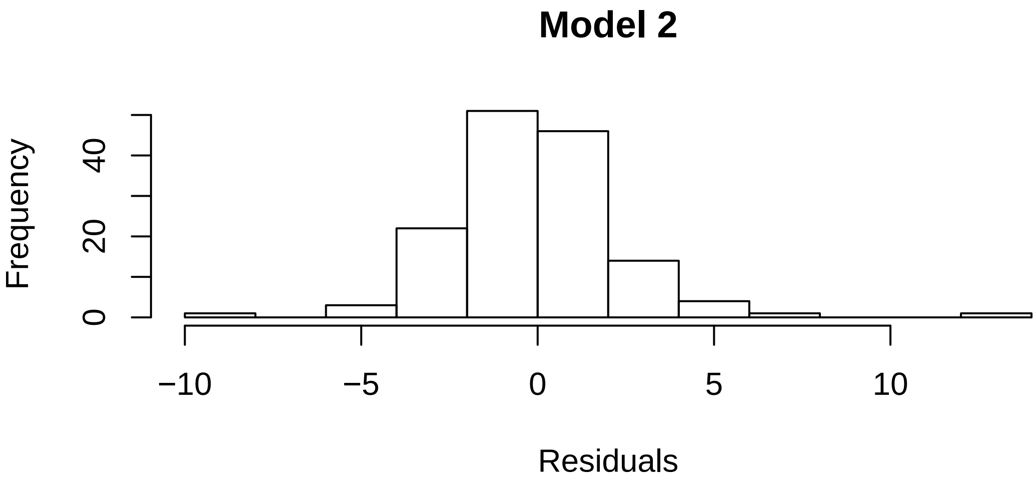
where $Y_t = \nabla\nabla^{12} X_t^{0.2626263}$.

5 Diagnostics

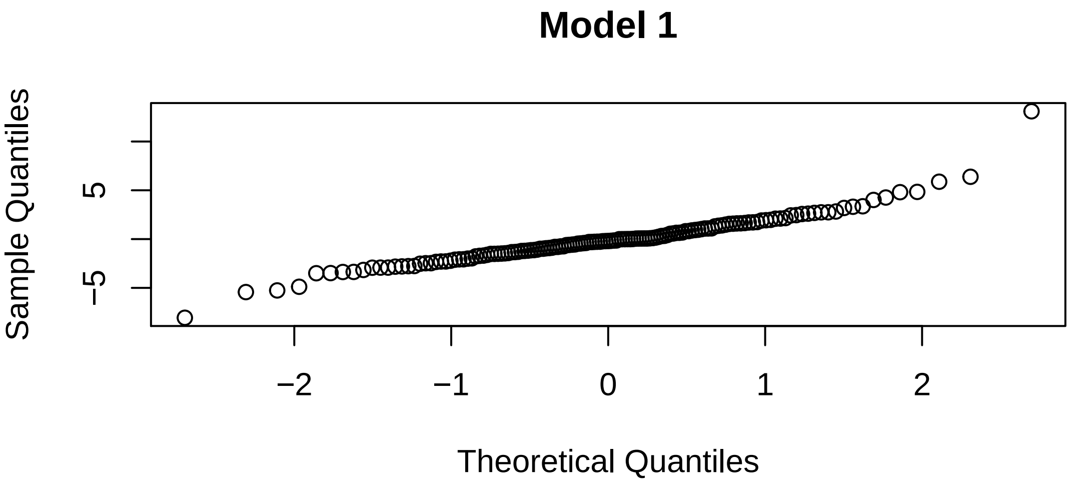
5.1 Normality

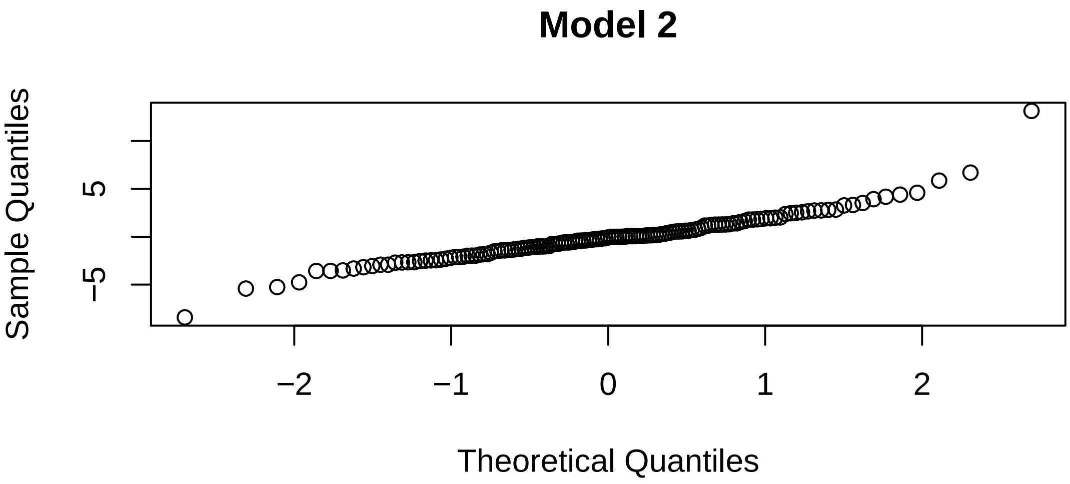
Normality is checked using two methods: residual histogram and Q-Q plot.





The residuals appear to be normal, which is desireable.





These mostly lie in a straight line, which is desireable.

Normality: Confirmed.

5.2 Independence

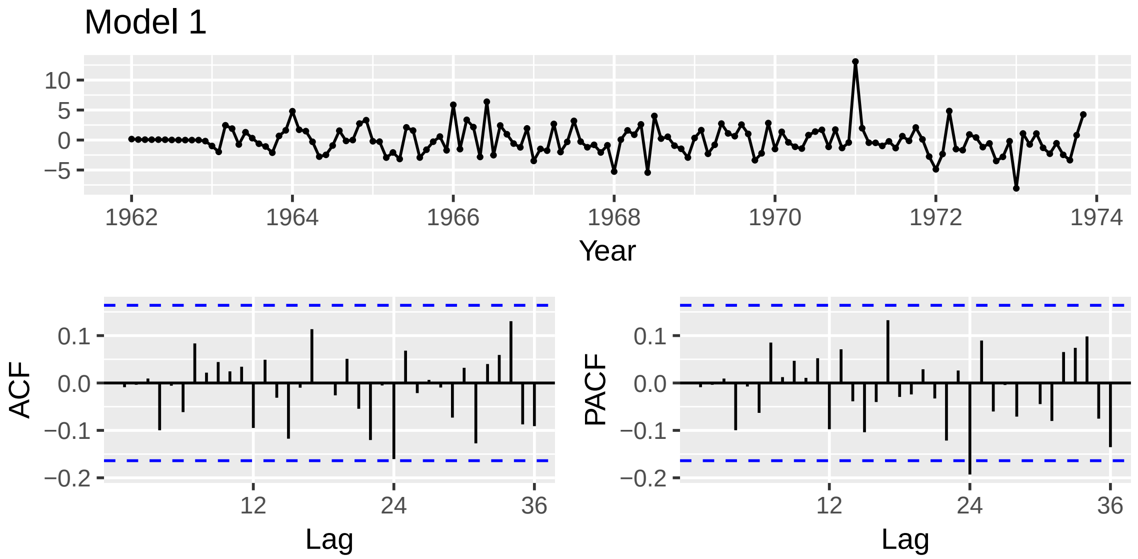
To ensure there is no serial correlation, Ljung-Box and Box-Pierce tests can be used.

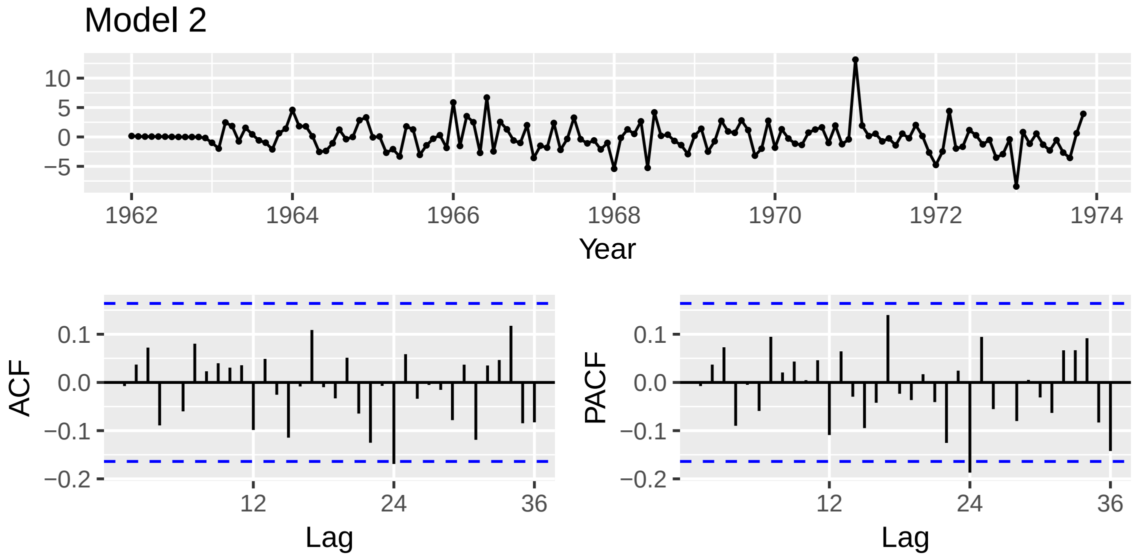
	Box-Pierce Test	Ljung-Box Test
Model 1	0.844796	0.814000
Model 2	0.902541	0.879298

Since the values here are very large, there is no reason to suspect serial correlation.

5.3 Constant Variance

Error terms in time series datasets ought not change over time. To ensure this, the ACF and PACF plots of the squared residuals must show that most of our error terms are within 95% white noise limits. This allows to assume constant variance.





Most of the error terms are within the blue lines, so constant variance can be assumed.

Since both models pass, Model 2 is preferable because it has lower AICc and BIC values.

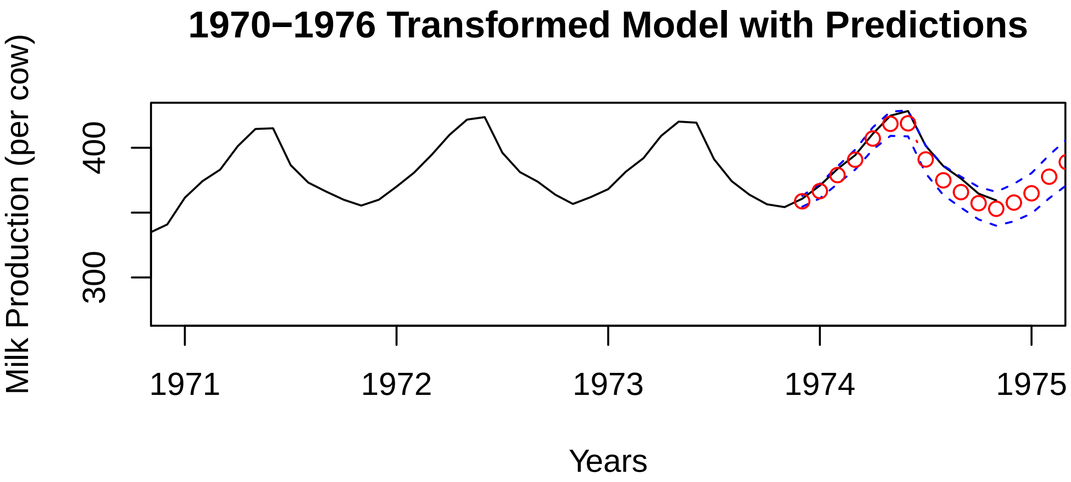
Final Model: $\text{SARIMA}(0, 1, 1) \times (1, 1, 0)_{12}$

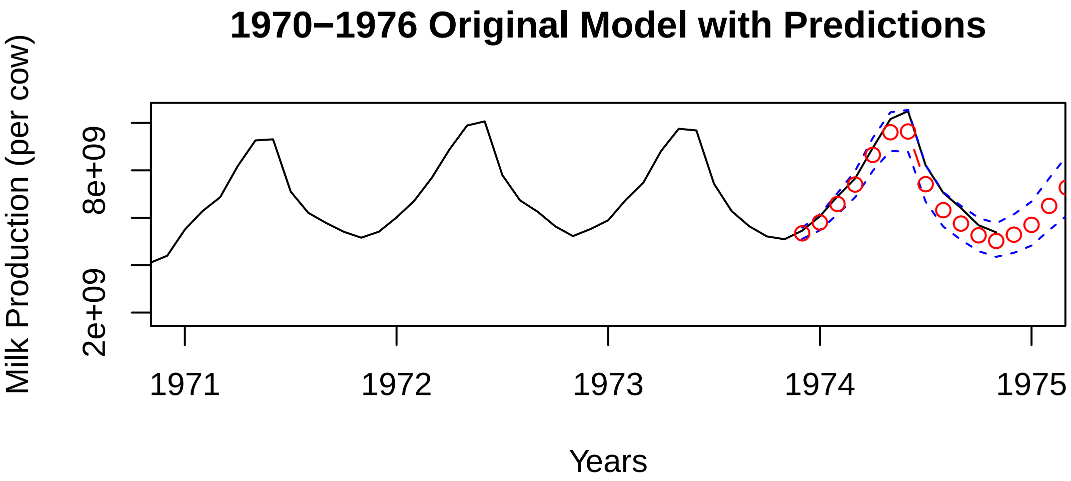
6 Forecasting

We now have a final model! For this dataset, if it follows the same pattern as in previous years, it should be possible to predict future performance.

The below plots shows both the transformed and observed data from 1971-74 along with the predicted results from 1974 for the next two years and a 90% confidence interval for this forecast. Forecasted points are in red.

While I attempted to find real-life data to confirm this, I was unfortunately unable to find an appropriate dataset to compare my forecast to.





7 Conclusion

To better understand the per cow production of Milk, I analyzed how per cow milk production has changed over the selected years and how it behaved on a seasonal basis. The time series model constructed explains the behavior of per cow Milk production and endeavors to forecast it as well.

My project shows a wonderful upward trend in per cow Milk production. From 1962-74, there was a noticeable growth in per cow Milk production. Warmer months and changing seasons likely caused the observed seasonality.

The final model: $\text{SARIMA}(0,1,1) \times (1,1,0)_{12}$.

This model satisfied all the assumptions that are expected of a satisfactory time series model.

8 References

Monthly milk production: Pounds per Cow. Jan 1962 - Dec 1974, Cryer (1986). https://datamarket.com/data/set/22ox/monthly-milk-production-pounds-per-cow-jan-62-dec-75#!ds=22ox&display=line

9 Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, fig.height = 3,
                      fig.width = 6, fig.align = "center")
library(tidyverse)
library(tseries)
library(qpcR)
library(tidyverse)
library(latex2exp)
library(fma)
library(astsa)
library(MASS)
library(TSA)
library(kableExtra)
library(GeneCycle)
library(forecast)
library(MuMIn)
library(robustbase)
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)
        }
        }
data <- read_csv("~/UCSB/PSTAT 174/Project/final/monthly-milk-production-pounds-p.csv")
data <- data[-nrow(data),2]
colnames(data) <- "Milk"
data <- ts(data, frequency = 12, start = 1962)
original.data <- read_csv("~/UCSB/PSTAT 174/Project/final/monthly-milk-production-pounds-p.csv")
original.data <- original.data[-nrow(original.data),2]
colnames(original.data) <- "Milk"
original.data <- ts(original.data, frequency = 12, start = 1962)
plot(data,
     main = 'Figure 1: Monthly Milk Production, 1962-1974',
     ylab = 'Milk Produced (lbs per cow)',
     xlab = 'Year')
# Plot seasonal data
seasonplot(data,12,col=rainbow(4),
           main="Figure 2: Seasonal Plot",
           ylab = 'Milk Produced (lbs per cow)',
           xlab = 'Year',
           year.labels.left = T)
# seasonal plot and decomposition plot
decomposed <- decompose(data)
autoplot(decomposed, main="Decomposition of Milk Production") + xlab("Year")
boxcox <- boxcox(data ~ as.numeric(1:length(data)))
lambda <- boxcox$x[which.max(boxcox$y)]
data.boxcox <- BoxCox(data, lambda = "auto")

plot(data.boxcox,
     main="Transformed Data",
     ylab="Milk Production (per cow)",
     xlab="Year",
     type="l")
plot(data, main="Original Data",
     ylab="Milk Production (per cow)",
     xlab="Year",
     type="l")
#data points to be modeled
data.modeled = ts(data.boxcox[1:143,], frequency = 12, start = 1962)

#data points to be tested
data.tested = ts(data.boxcox[144:155,], frequency = 12, start = 1974)

data.deseasoned <- diff(data.modeled, lag=12) # de-seasonalized data
```

```
# deseasonalized data plot
plot(data.deseasoned, ylab=NULL, xlab="Months", main=expression(nabla[12]~X[t]^0.2626263))
title(sub="Deseasoned  Data")

data.cleaned <- diff(data.deseasoned, lag=1) # remove trend


plot(data.cleaned, ylab=NULL, xlab="Months",
     main=expression(nabla~nabla[12]~X[t]^-(0.2626263)))
title(sub="Detrended Data")
ggtsdisplay(data.cleaned, main = "Milk Production with Associated ACF and PACF")
# AICc and BIC matrix:
AICc <- BIC.s <- matrix(NA, nrow=4, ncol=4)
for(p in 0:3){
  for(q in 0:3){
    AICc[p+1,q+1] <- sarima(data.modeled, p, 1, q, P = 0, D = 1, Q = 1,S =  12 , details = FALSE)$AICc
    BIC.s[p+1,q+1] <- sarima(data.modeled, p, 1, q, P = 0, D = 1, Q = 1,S =  12 , details = FALSE)$BIC
  }
}
AICc <- data.frame(AICc, row.names = c("AR(0)","AR(1)","AR(2)","AR(3)"))
BIC.s <- data.frame(BIC.s, row.names = c("AR(0)","AR(1)","AR(2)","AR(3)"))
#AICc==min(AICc)
#BIC.s==min(BIC.s)
kable(AICc, col.names = c("MA(0)","MA(1)","MA(2)","MA(3)"), format="markdown")
kable(BIC.s, col.names = c("MA(0)","MA(1)","MA(2)","MA(3)"), format="markdown")
fit1 <- arima(data.modeled, order=c(0,1,3), seasonal = list(order=c(1,1,0), period=12),
              method="ML")
fit2 <- arima(data.modeled, order=c(0,1,1), seasonal = list(order=c(1,1,0), period=12),
              method="ML")
coef1 <- unlist(fit1$coef)
coef2 <- unlist(fit2$coef)
coeffs <- cbind(coef1, c(coef2[1],NA,NA,coef2[2])) %>%
  data.frame(row.names = c("MA(1)","MA(2)","MA(3)","SAR(1)"))
kable(coeffs, format="markdown", digits = 4,
      col.names = c("Model 1", "Model 2"))
resid1 <- residuals(fit1)
resid2 <- residuals(fit2)
hist(resid1, main="Model 1", xlab = "Residuals")
hist(resid2, main="Model 2", xlab = "Residuals")
qqnorm(resid1, main="Model 1")
qqnorm(resid2, main="Model 2")
#box-pierce and ljung tests
bpA <- Box.test(resid1, lag=12, type="Box-Pierce", fitdf=3)$p.value
lbA <- Box.test(resid1, lag=12, type="Ljung", fitdf=3)$p.value

bpB <- Box.test(resid2, lag=12, type="Box-Pierce", fitdf=1)$p.value
lbB <- Box.test(resid2, lag=12, type="Ljung", fitdf=1)$p.value
boxes <- rbind(c(bpA,lbA),c(bpB,lbB)) %>% data.frame(row.names = c("Model 1","Model 2"))
kable(boxes, format="markdown", digits = 6, col.names = c("Box-Pierce Test","Ljung-Box Test"))
# Model 1
ggtsdisplay(resid1, main="Model 1", xlab="Year")
acfb=ggAcf(resid1^2, lag.max = 50,main="")
pacfb=ggPacf(resid1^2, lag.max = 50,main="")
```

```r
# Model 2
ggtsdisplay(resid2, main="Model 2", xlab="Year")
acfb=ggAcf(resid2^2, lag.max = 50,main="")
pacfb=ggPacf(resid2^2, lag.max = 50,main="")
#forecasting with model
pred <- predict(fit2, n.ahead = 24)
CI.u <- pred$pred + 1.64*pred$se
CI.l <- pred$pred - 1.64*pred$se
predpred <- pred$pred

ts.plot(data.boxcox,
        main="1970-1976 Transformed Model with Predictions",
        xlim=c(1971, 1975),
        ylab="Milk Production (per cow)",
        xlab="Years",
        type="l")
lines(CI.u, col="blue", lty="dashed")
lines(CI.l, col="blue", lty="dashed")
lines(predpred, col="red", type = 'b')

#Untransform data
original.data <- data.boxcox^(1/lambda)
#Plot original data

ts.plot(original.data,
        main="1970-1976 Original Model with Predictions",
        xlim=c(1971, 1975),
        ylab="Milk Production (per cow)",
        xlab="Years",
        type="l")
lines(CI.u^(1/lambda), col="blue", lty="dashed")
lines(CI.l^(1/lambda), col="blue", lty="dashed")
lines(predpred^(1/lambda), col="red", type = 'b')
```