Car Sales in Quebec 1960-1968 Time Series Analysis

Austin Miles

6/1/2020

Abstract

The market for cars go through various increase and decreases in sales over time. The data set we use provides us monthly sales for cars in Quebec, Canada from 1960 to 1967. Using this data set, our goal is to forecast the amount of cars sold within the next 12 months and compare that to the actual number of cars sold in each of those 12 months. Forecasting these amounts will provide insight to fluctations in car sales which can benefit those who want to sell a car at an optimal time or buy a car at an optimal time.

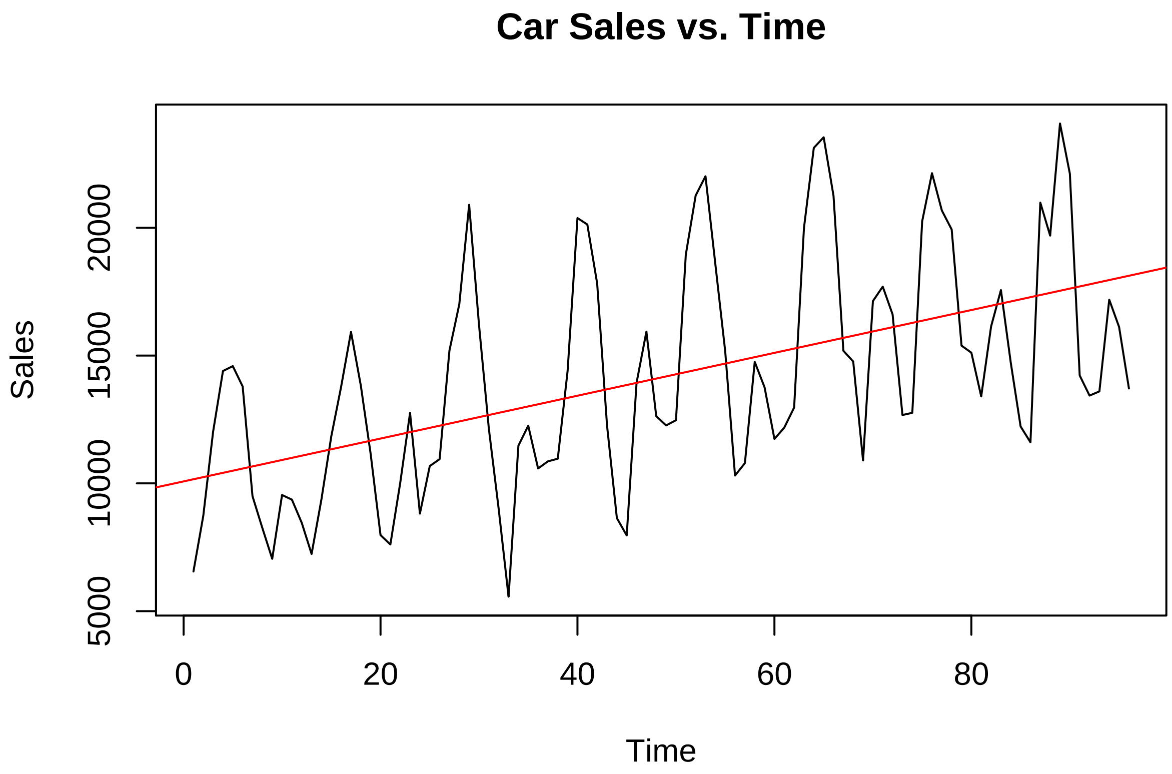
We are able to compare the actual results because we create a training set which does not include the last 12 months and perform our analysis on that set. For forecasting to be effective, the data must be stationary so we used a log transformation as well as differencing to get rid of the trend and seasonal component. After observing the ACF and PACF of the newly transformed data we get 3 potential SARIMA models which we compared AIC, examined residuals, performed diagonstics, calculated casuality and invertibility. We concluded that the model SARIMA (2,0,0) x $(0, 1, 1)_{12}$ was the most suitable model from these tests. The 95% confidence interval of the forcasted values were accuaratly able to contain the true results for the months in 1968.

Introduction

We want to forecast the future trends in car sales in Quebec to provide insights on when it is a good time to buy or sell a car. Basic economics teach us that when there are a lot of cars being sold, the car price is high; but when there is not a lot of cars being sold, the price is low. The data set contains 9 years worht of monthly car sales but we will only use the first 8 years in our study so that we can use the 9th year to compare our results and find our models accuracy. We load our Quebec car sales data into R and remove the last 12 observations as our test set for the forecasts. We then go into exploratory analysis and found that the data requires transformations to stabilize variance and differencing to remove trend and seasonality. Then we examine the ACF and PACF of the stationary series to identify appropriate models. We get three potential models which we conduct model diagnostics to make it a valid SARIMA. In the end, we use the model with the lowest AICc to forecast the next 12 observations which we compare to the test set, to evaluate the model's accuracy. The model we ended with was a good fit for our data, however we were unable to normalize the dataset perfectly which is common with real world datasets. We retrieved the data from https://data.world/perceptron/monthly-car-sales-quebec-1960 and used R for our analysis.

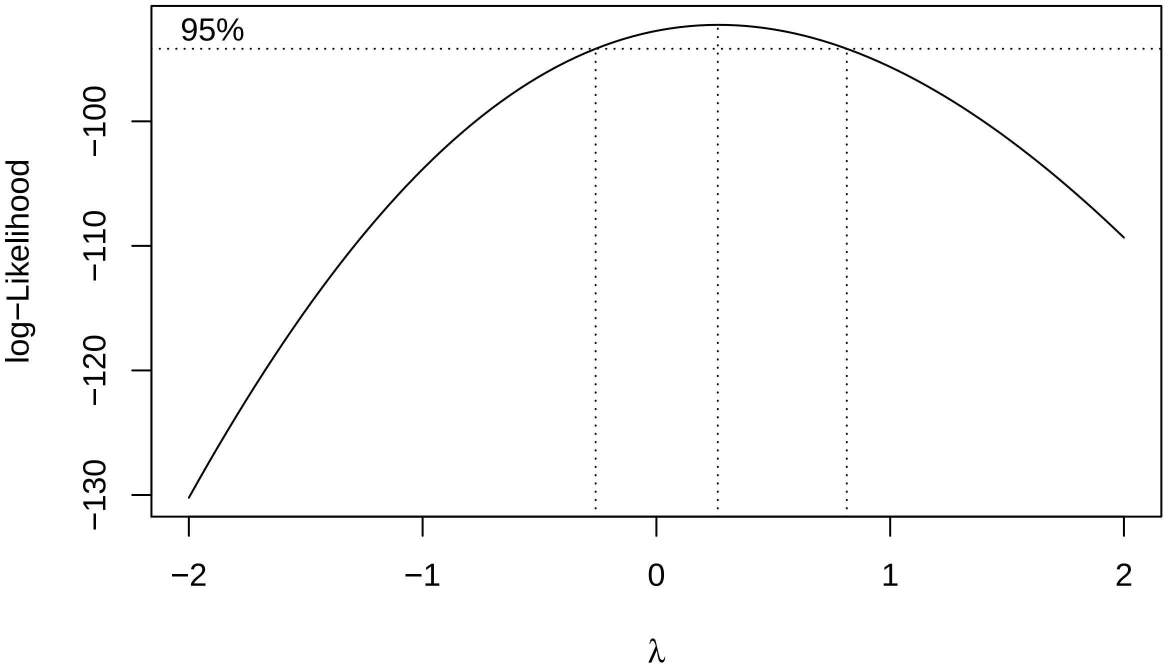
Plot and analysis of time series

The data we will use does not include the last 12 observations because we will compare them to the forecasted observations.

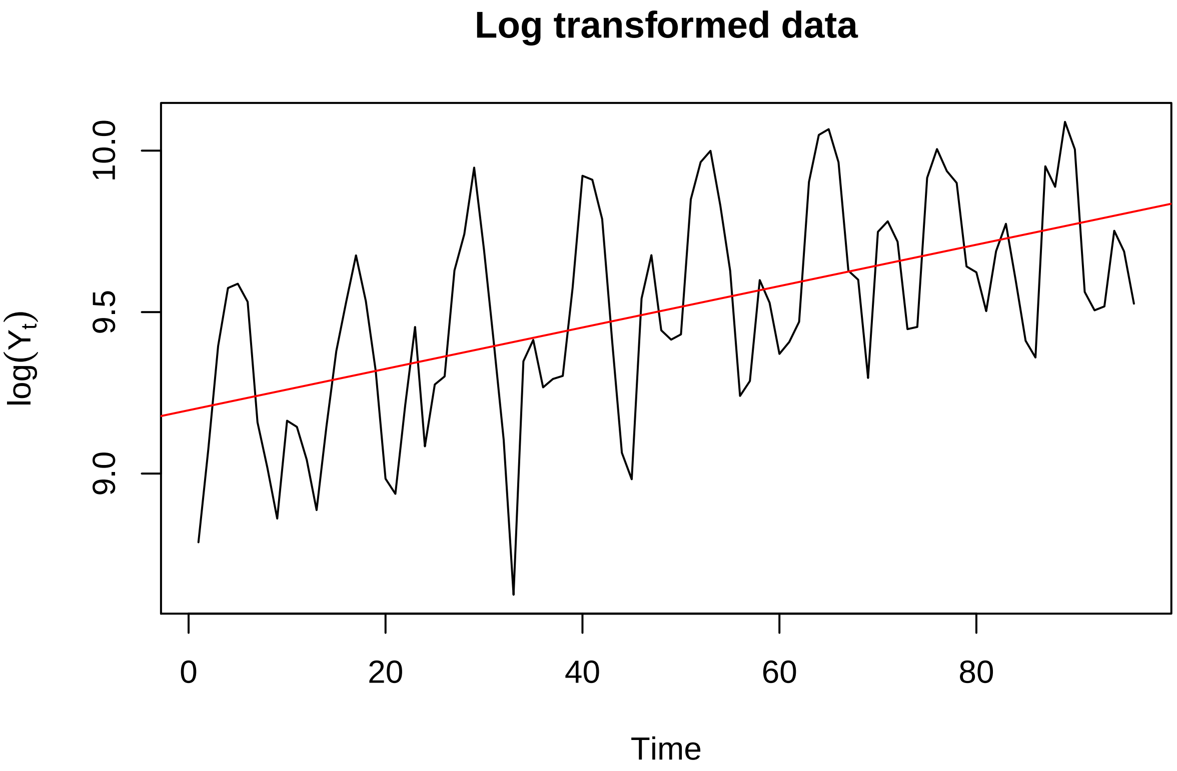


There is a positive trend and is seasonal. There are sharp changes in behavior around Time = 30 and variance seems to increase over time. We will transform the series to make it stationary. The trend line fits the data well so we will use difference at lag=1 to remove trend and difference at lag=12 to remove seasonality. However, the variance is not stable so we use a Box-Cox transformation.

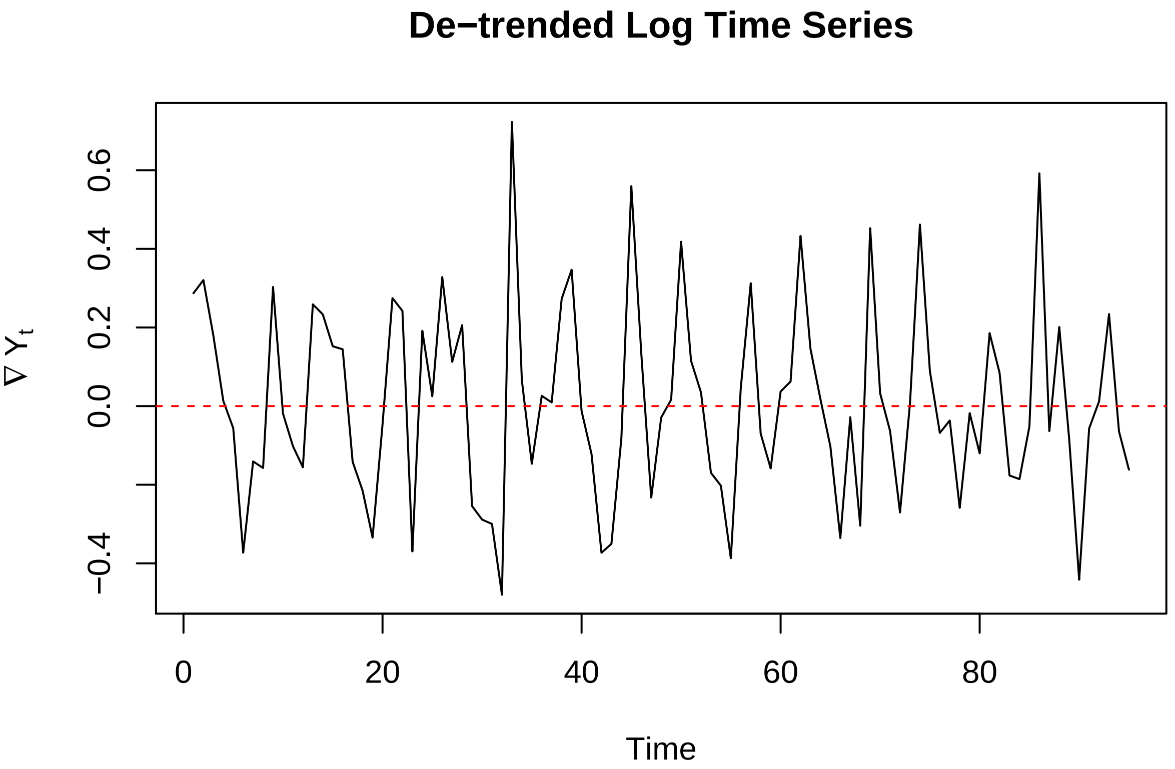
We use the Box-Cox transformation to stabilize the variance. To apply the transformation we find the best λ.



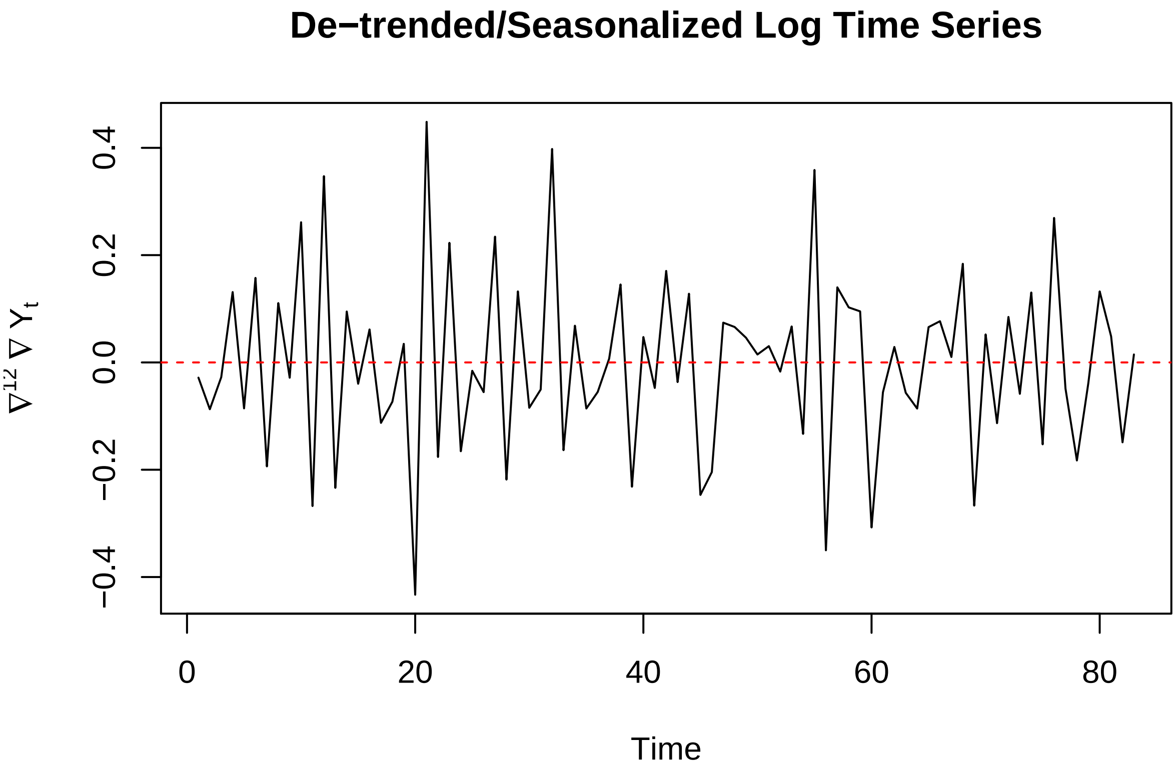
λ=0.2626263 is the optimal value and λ=0 is in the confidence interval, so we proceed with a log transformation instead.



The transformation clearly lowers the the variance from 19419113 to `0.1045338` and maintains the positive trend. So we then difference the data at lag=1 and lag=12.



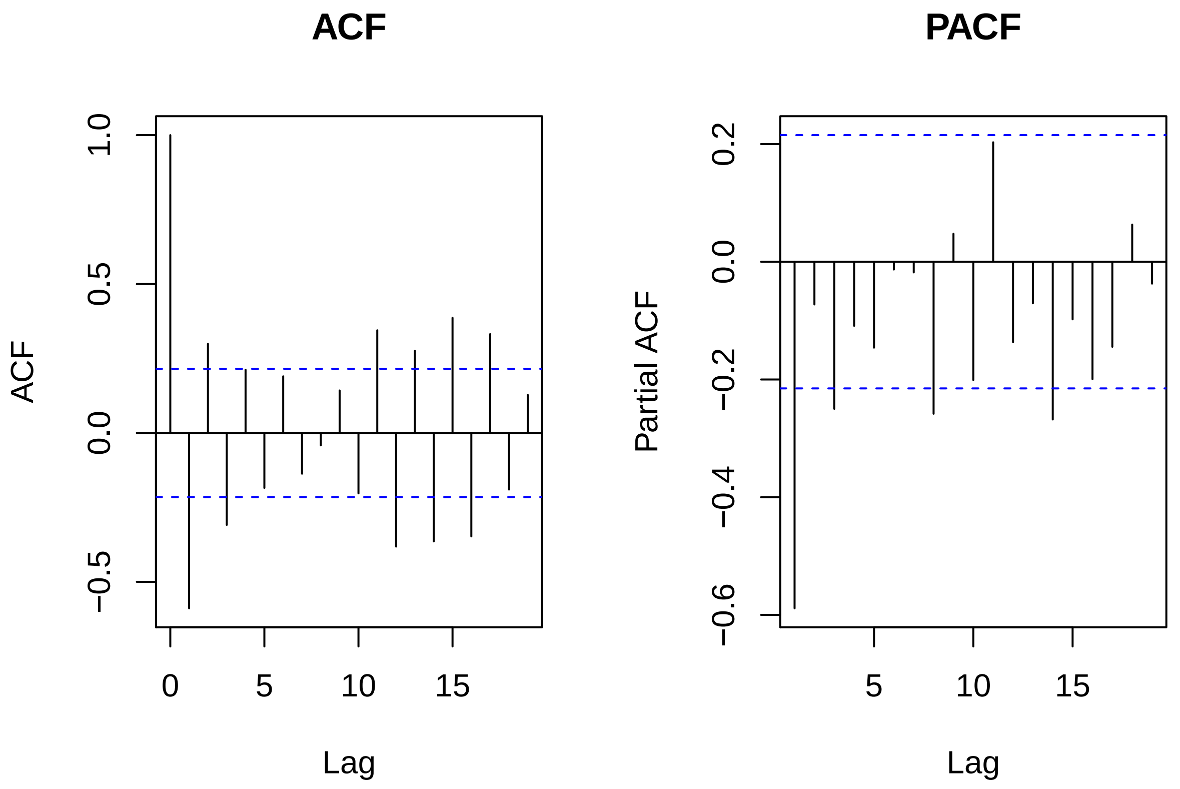
The variance of the model differenced at lag=1 decreases by `0.0451486`, so the differencing is justified. Now we difference the model at lag=12 to remove the seasonal component.



The variance of the model differenced at lag=12 decreases by `0.0315116`, so the differencing is justified. The data looks stationary so we will check ACF and PACF.

ACF and PACF

Since the differenced and transformed time series is stationary, we will look at the ACF and PACF plots to determine the variables in the SARIMA model.We will identify the possible AR, MA, SAR, SMA variables for our SARIMA model.



We observed that there is a seasonal component and potential values SMA(0) and SMA(1) since the ACF decreases after lag = 12. We also observed that potential values for SAR could be SAR(0) and SAR(1) since PACF cuts off after lag = 0 or 1. We chose these values since the lags slightly outside of the interval could be due to noise. We examine the first 11 lags to find the AR and MA of the model. ACF cuts off after lag 1,2,3 but is less significant for lag = 2 and 3. So we will consider AR(1) and AR(2). The PACF decays at lag=1 so we consider MA(1) and MA(0).

```
## Warning in arima(y0, order = c(1, 0, 1), seasonal = list(order = c(1, 1, :
## possible convergence problem: optim gave code = 1
```

Potential Models:

1. SARIMA (1,0,1) x $(0,1,1)_{12}$
   -AIC = -109.06
   -$(1-0.984B)X_t\nabla_{12}\nabla X_t = (1+0.782B)(1+0.4651B^{12})Z_t$

2. SARIMA (2,0,0) x $(0,1,1)_{12}$
   -AIC = -105.4

-$(1 - 0.341B)X_t\nabla_{12}\nabla X_t = (1 - 0.397B)(1 + 0.3B^{12})Z_t$

3. SARIMA (1,0,1) x $(1,1,2)_{12}$
   -AIC = -107.2
   -$(1 - 0.999B)(1 - 0.61B^{12})X_t\nabla_{12}\nabla X_t = (1 + 0.787B)(1 + 1.168B^{12})(1 - 0.191B^{24})Z_t$

```
## Warning: package 'glmulti' was built under R version 3.6.3
```

```
## Loading required package: rJava
```

```
## Warning: package 'rJava' was built under R version 3.6.3
```

```
## Loading required package: leaps
```

```
## Warning: package 'leaps' was built under R version 3.6.3
```

We compare the AICc values for each of the model.

```
## [1] -108.5566
```

```
## [1] -104.8926
```

```
## [1] -106.1092
```

The values are similar so we will check the normality of each model to decide which model to proceed with.

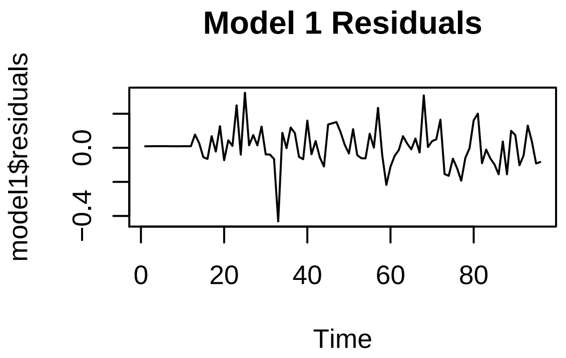
Model 1

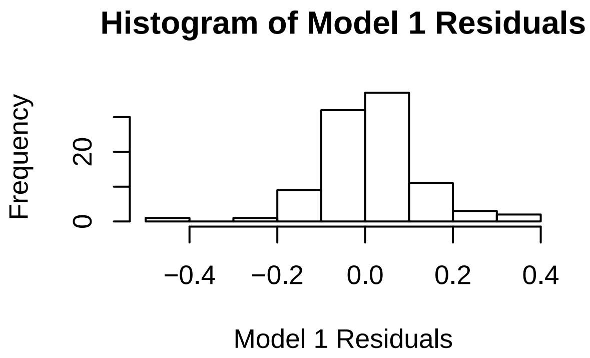
```
##
##  Box-Pierce test
##
## data:  model1$residuals
## X-squared = 16.515, df = 10, p-value = 0.0858
```

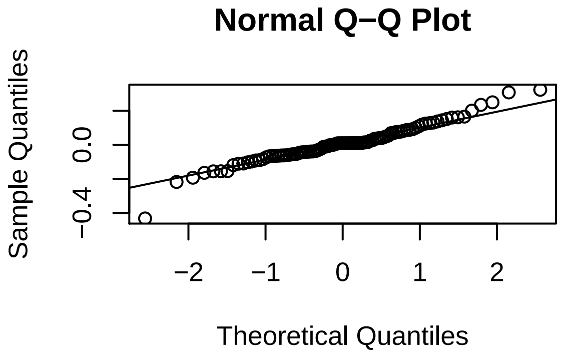
```
##
##  Box-Ljung test
##
## data:  model1$residuals
## X-squared = 18.212, df = 10, p-value = 0.0515
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.95978, p-value = 0.004925
```

Now we analyze the residuals and perform diagnostic checks for the models.





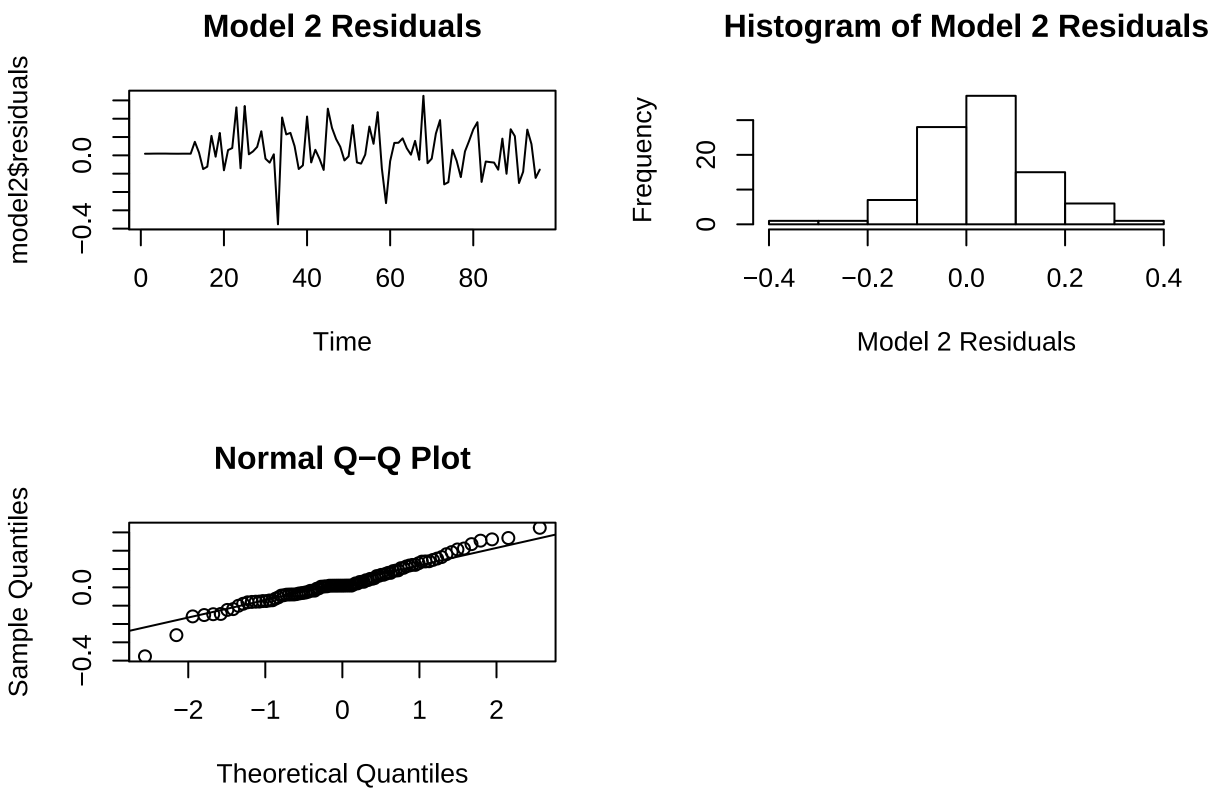


Model 2

```
##
##  Box-Pierce test
##
## data:  model2$residuals
## X-squared = 15.593, df = 10, p-value = 0.1119
```

```
##
##  Box-Ljung test
##
## data:  model2$residuals
## X-squared = 17.218, df = 10, p-value = 0.06968
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.97242, p-value = 0.04038
```

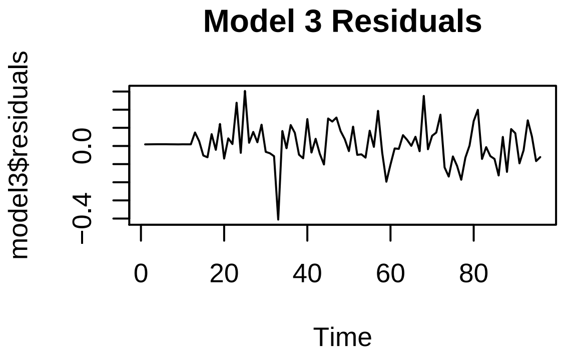


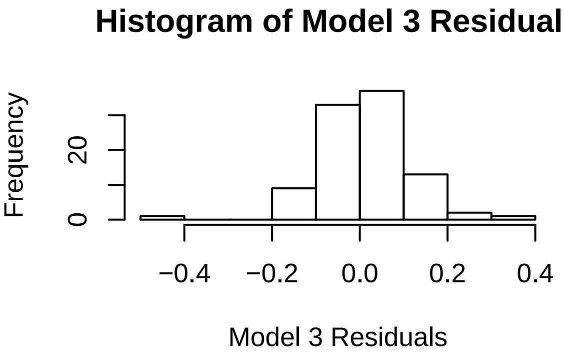
Model 3

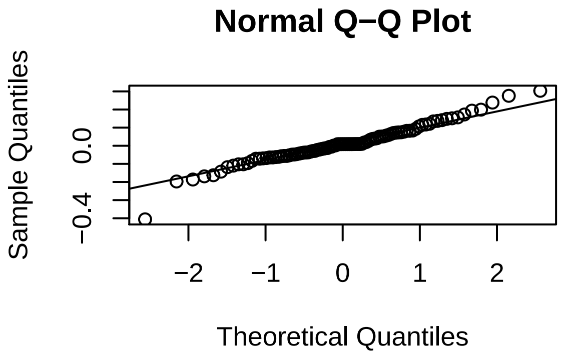
```
##
##  Box-Pierce test
##
## data:  model3$residuals
## X-squared = 15.738, df = 10, p-value = 0.1074
```

```
##
##  Box-Ljung test
##
## data:  model3$residuals
## X-squared = 17.284, df = 10, p-value = 0.06832
```

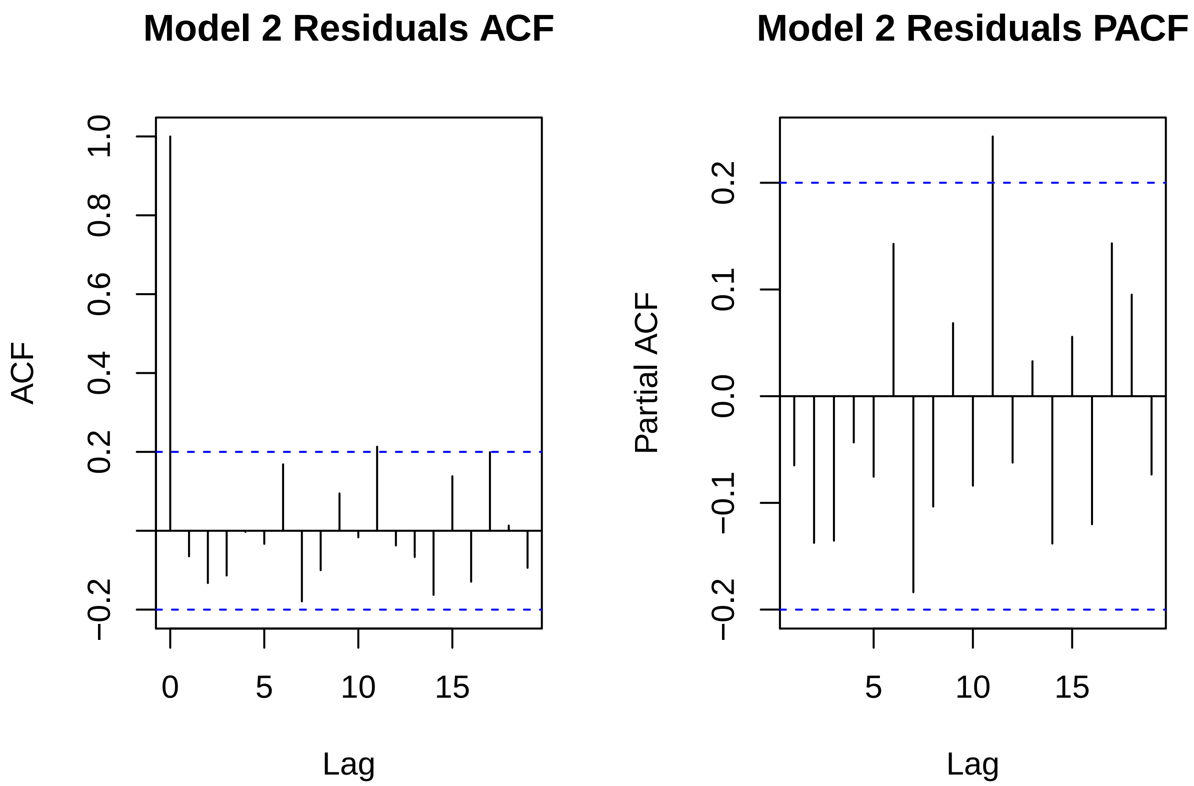
```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.96317, p-value = 0.008515
```



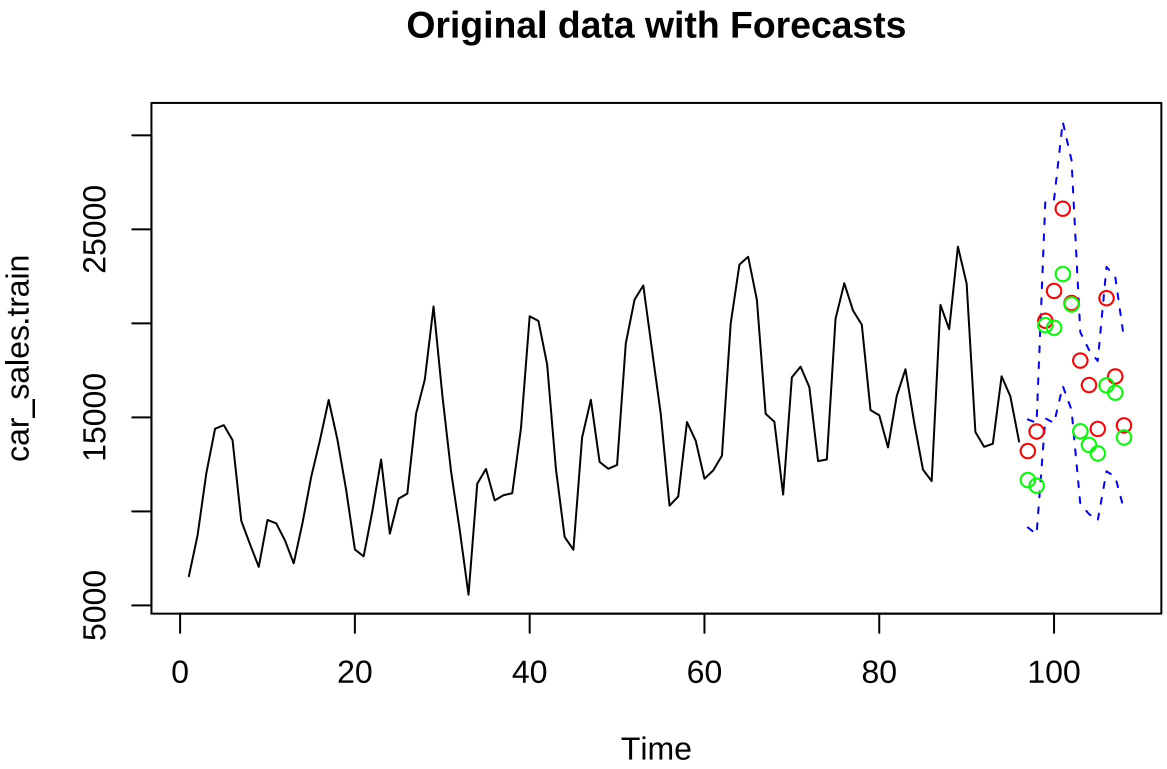




We decide to proceed with Model 2, SARIMA (2,0,0) x $(0,1,1)_{12}$, because it is the most normalized and the only model that passes the Shapiro-Wilk normality test at a 90% level. All 3 models have coefficents with absolute value less than 1 so they are all stationary and invertible.



The residuals for Model 2 show no trend, slight change in variance, and no seasonal component. The histogram and QQ plot show the best normality. The ACF plot stays within the confidence interval.



We forecast 12 points using model 2 (green points) with its 95% confidence interval (dashed blue line) and compare them to the actual points in our dataset (red). The true values are within the interval and the points are close to our model points, so our final model accurately forecasts the data.

Conclusion

Model 2. SARIMA (2,0,0) x $(0, 1, 1)_{12}$
$(1 - 0.341B)X_t\nabla_{12}\nabla X_t = (1 - 0.397B)(1 + 0.3B^{12})Z_t$
was used to forecast future car sales in Quebec since it had a low AIC and passed most of the diagnostic checks. The final graph shows that the model was able to accuaratly forecast the time-series.

References

https://data.world/perceptron/monthly-car-sales-quebec-1960

Appendix

```
# Read in the dataset
data <- read.csv("monthly-car-sales.csv")
```

```
# Create Time Series
car_sales <- ts(data$Sales, frequency=12, start=c(1960,1))
```

```
# seperate train and test set
car_sales.train <- car_sales[c(1:96)]
car_sales.test <- car_sales[c(97:108)]
```

```
# plot with trend
ts.plot(car_sales.train,main = "Car Sales vs. Time",ylab = 'Sales')
  abline(reg=lm(car_sales.train~time(car_sales.train)), col = 2)
```

```
library(MASS)
# Box-Cox Transformation
bcTransform = boxcox(car_sales.train~as.numeric(1:length(car_sales.train)),plotit = TRUE)
```

```
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
car_sales.bc = (1/lambda)*(car_sales.train^lambda-1)
```

```
# Log Transformation
y0 <- log(car_sales.train)

# Log
ts.plot(y0, main= "Log transformed data", ylab=expression(log(Y[t])))
  abline(reg=lm(y0~time(car_sales.train)), col = 2)
```

```
# variance of Log transformation
y0.var <- var(y0, na.rm =T)
```

```
# Diference at lag = 1 to remove trend component
y1 <- diff(y0, 1)
y1.var <- var(y1, na.rm = T)
y1.var.diff <- y0.var-y1.var
plot.ts(y1,main = "De-trended Log Time Series",ylab = expression(nabla~Y[t]))
  abline(h = 0,lty = 2, col = "red")
```

```
# Diference at lag = 12 to remove seasonal component
y12 <- diff(y1, 12)
y12.var <- var(y12,na.rm = T)
y12.var.diff <- y1.var - y12.var
plot.ts(y12,main = "De-trended/Seasonalized Log Time Series",ylab = expression(nabla^{12}~nabla~Y[t]))
  abline(h = 0,lty = 2, col = 'red')
```

```
# ACF
acf <- acf(y12, plot = TRUE, main = "ACF")
```

```
pacf <- pacf(y12, plot = TRUE, main = "PACF")
```

```
model1 <- arima(y0, order=c(1,0,1), seasonal = list(order = c(0,1,1), period = 12), method="ML")
model2 <- arima(y0, order=c(2,0,0), seasonal = list(order = c(0,1,1), period = 12), method="ML")
model3 <- arima(y0, order=c(1,0,1), seasonal = list(order = c(1,1,2), period = 12), method="ML")
```

```
## Warning in arima(y0, order = c(1, 0, 1), seasonal = list(order = c(1, 1, :
## possible convergence problem: optim gave code = 1
```

```
library(glmulti)
```

```
aicc(model1)
aicc(model2)
aicc(model3)
```

```
# Model 1 Tests
Box.test(model1$residuals, lag = 12, type = c("Box-Pierce"), fitdf = 2)
Box.test(model1$residuals, lag = 12, type = c("Ljung-Box"), fitdf = 2)
shapiro.test(model1$residuals)
```

```
# Model 1
plot(model1$residuals, main="Model 1 Residuals")
```

```
hist(model1$residuals, main = "Histogram of Model 1 Residuals")
```

```
qqnorm(model1$residuals)
qqline(model1$residuals)
```

```
# Model 2 Tests
Box.test(model2$residuals, lag = 12, type = c("Box-Pierce"), fitdf = 2)
Box.test(model2$residuals, lag = 12, type = c("Ljung-Box"), fitdf = 2)
shapiro.test(model2$residuals)
```

```
# Model 2
plot(model2$residuals, main="Model 2 Residuals")
```

```
hist(model2$residuals, main = "Histogram of Model 2 Residuals")
```

```
qqnorm(model2$residuals)
qqline(model2$residuals)
```

```
# Model 3 Tests
Box.test(model3$residuals, lag = 12, type = c("Box-Pierce"), fitdf = 2)
Box.test(model3$residuals, lag = 12, type = c("Ljung-Box"), fitdf = 2)
shapiro.test(model3$residuals)
```

```
# Model 3
plot(model3$residuals, main="Model 3 Residuals")
```

```
hist(model3$residuals, main = "Histogram of Model 3 Residuals"
)
```

```
qqnorm(model3$residuals)
qqline(model3$residuals)
```

```
acf(model2$residuals, main= "Model 2 Residuals ACF")
```

```
pacf(model2$residuals, main= "Model 2 Residuals PACF")
```

```
fit.A <- model2
pred.tr <- predict(fit.A, n.ahead = 12)
U.tr = exp(pred.tr$pred+2*pred.tr$se)
L.tr = exp(pred.tr$pred-2*pred.tr$se)
ts.plot(car_sales.train, xlim=c(1,length(y0)+12), ylim = c(min(car_sales.train),max(U.tr)), main="Origi
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
points((length(car_sales.train)+1):(length(car_sales.train)+12), car_sales.test, col="red")
points((length(car_sales.train)+1):(length(car_sales.train)+12), exp(pred.tr$pred), col="green")
```