Time Series Project

Chandler Crescentini

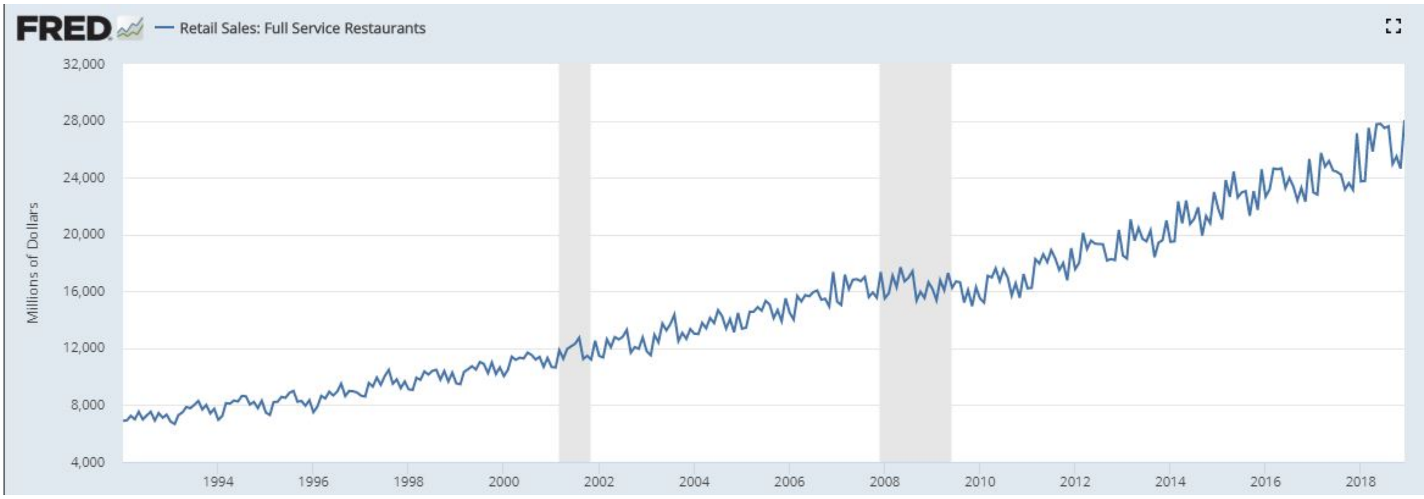
May 6, 2019

Introduction

The Monthly Retail Trade Survey is a voluntary, monthly survey of around 13,000 “brick-and-mortar” companies that sell merchandise and related services to the general public. These companies provide information on the dollar value of retail sales for that month, as well as end-of-month inventories. Of these 13,000 companies, around 2,500 with sales above a certain size cutoff are guaranteed to be included, and the rest are randomly selected. It is posited that this approach helps to simplify the calculations without sacrificing accuracy, since the 2,500 companies above the cutoff represent a vast majority of the total dollar amount collected. The data collected and compiled by this survey is then used by many business, academic, and governmental agencies for a variety of applications, mostly focused on economic trends and health. For example, the Bureau of Economic Analysis uses the estimates to calculate the Gross Domestic Product, the Council of Economic Advisors uses the data to analyze current economic activity, and financial and investment companies measure recent economic trends, just to name a few. Restaurant spending is an important predictor of overall economic heath because it can be considered an unnecessary expense and thus, we can draw the conclusions that people will spend more money going out to eat if they have more disposable income, and contrapositively, people will spend less money going out to eat if they have less disposable income.

The purpose of this project is to analyze the compiled data from the Federal Reserve Bank of St. Louis using the decomposition approach and the Box-Jenkins method to see if we can predict future consumer spending in full service restaurants.

Data Expoloration



We begin by looking at the time series plot of Retail Sales for Full Service Restaurants shown above. The plot shows monthly sales data, in millions of dollars, beginning January 1992 and ending December 2018. The data has a clear increasing trend, starting at approximately 7,000 (million) in 1992 and ending at approximately 28,000 (million) in 2018. The only deviation from this increasing trend occurs in the right hand shaded area. This shaded time frame, from December 2007 to June 2009, corresponds to the collapse of the United States real estate market, and a period that many consider to be the worst economic depression since World War II. As mentioned above, if going out to eat at restaurants is something that could be considered a unnecessary expense, it follows that for people struggling financially, this might be one of the first things that they cut out of their budget. At the beginning of 2010, we see this decreasing trend reverse and begin to increase again, as a result of the recovering economy.

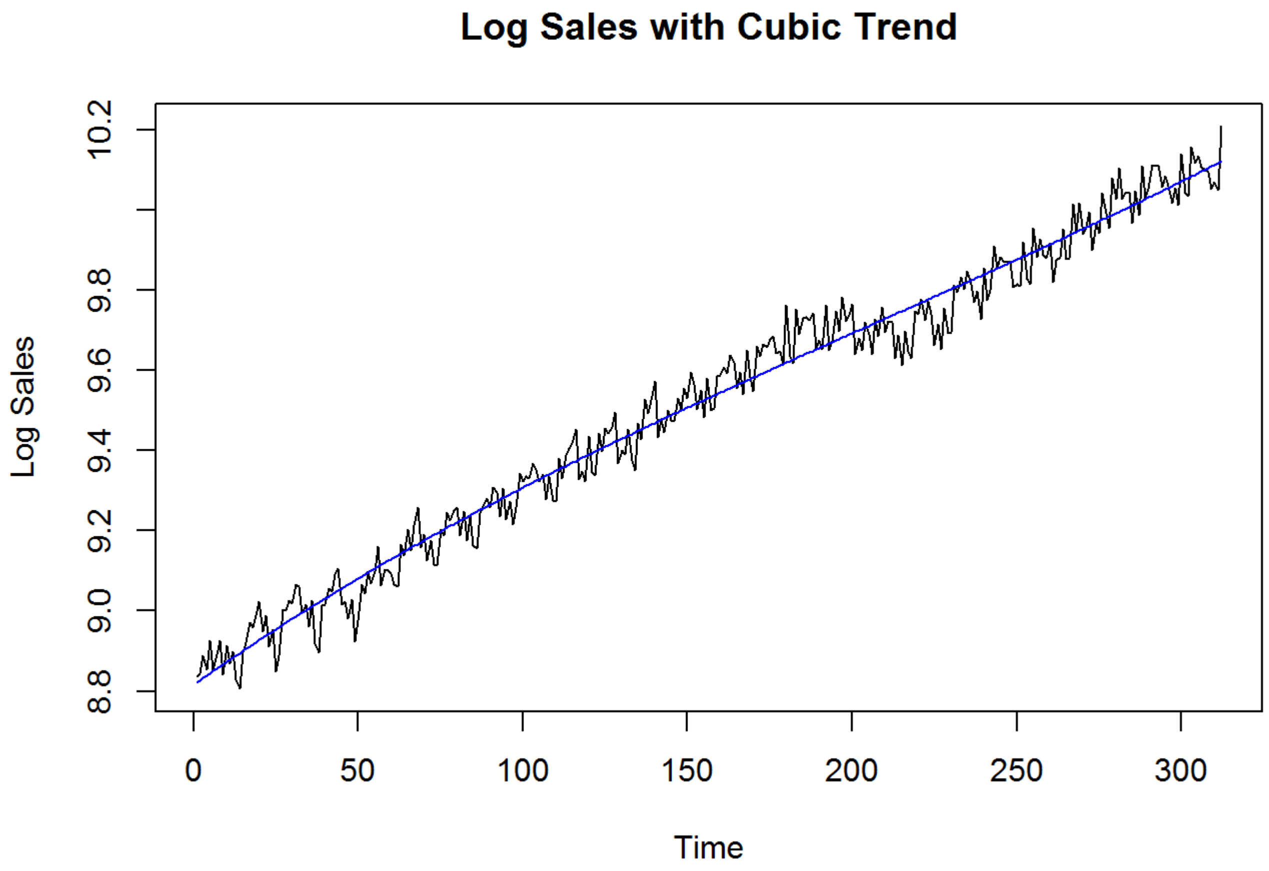
Within each year, a seasonal component also exists. Examining individual years at the beginning of the series, we see local maximums in the summer months and local minimums in the winter months. This seasonality continues, with increased variation throughout the series indicating that we have a multiplicative model. Something of interest is that as time increases, we see spending at restaurants in the month of December also increases. This may be due to end of year bonuses increasing people's amount of disposable income, or that it is simply a busy time of year with parties and holidays and people do not feel like cooking at home.

Analysis

Decomposition Approach

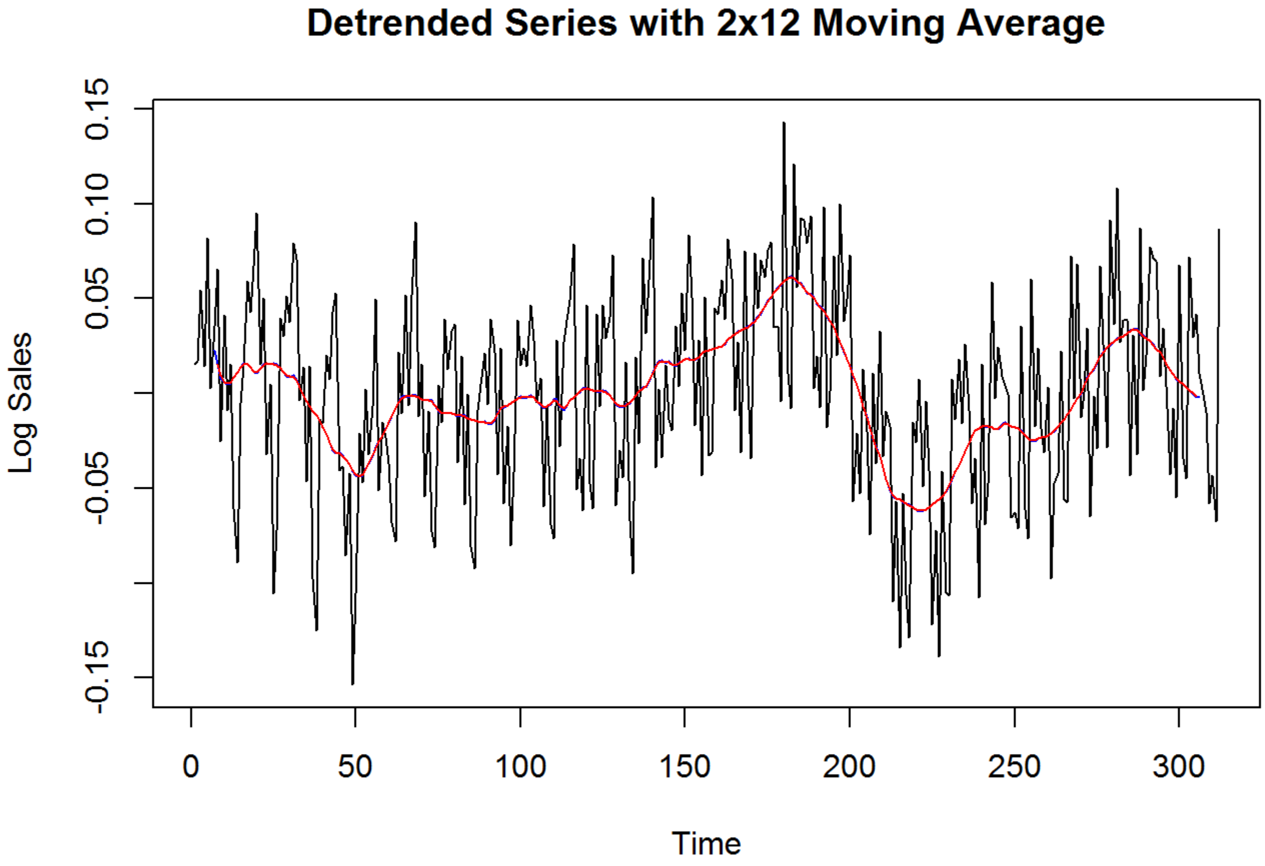
Our goal for the decomposition approach is to break down and estimate the trend and seasonal effects of the data and use these estimates to predict the final year of the time series. We hold out and test our prediction on the final year so we can see how accurate the prediction is. After all, it does us no good to make predictions if we don't know the accuracy of them.

Since we have a multiplicative model, we start by taking the log transform of the data in order to achieve constant variance throughout the data. In order to estimate the trend, we fit a linear, quadratic, and cubic regression line to see which fits the data best.

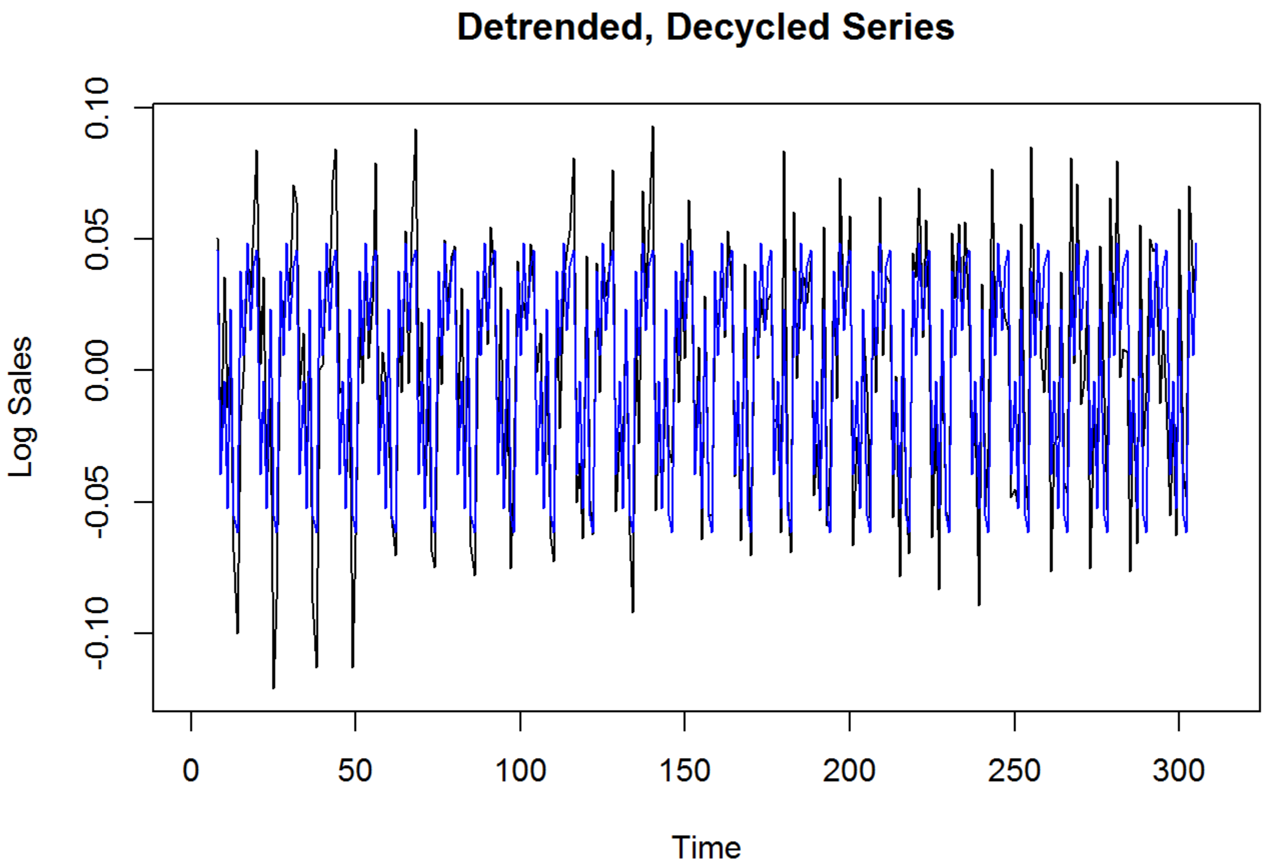

Log Sales with Cubic Trend

```
##
## Call:
## lm(formula = log ~ Time + I(Time^2) + I(Time^3), data = sales)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.153635 -0.038631  0.002672  0.038498  0.142915
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.816e+00  1.234e-02 714.507  < 2e-16 ***
## Time         5.784e-03  3.409e-04  16.971  < 2e-16 ***
## I(Time^2)   -1.037e-05  2.528e-06  -4.100 5.29e-05 ***
## I(Time^3)    1.677e-08  5.310e-09   3.158  0.00175 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05383 on 308 degrees of freedom
## Multiple R-squared:  0.9787, Adjusted R-squared:  0.9785
## F-statistic:  4712 on 3 and 308 DF,  p-value: < 2.2e-16
```

We see very similar R^2 values for all three trend models, with each being greater than 0.975. The overall F-test for all three are significant on the $\alpha = 0.05$ level, and the partial t-tests for each variable in each model are significant as well. Plotting each regression line on the log transformed plot shows that the hitch in the data, corresponding to the 2008 recession, is impacting the fit of the lines. This is causing the linear and quadratic models to overestimate values at the beginning of the time series and underestimate the values at the end. The hitch affects the cubic model as well, but it does a better job, compared to the other two, in following the overall trend of the data. Since we aren't in any real danger of running out of degrees of freedom, and we prefer a trend line that follows the heart of the data, we choose to proceed with the cubic model.



After finding the trend of the data, we plot the residuals to obtain the detrended series. From the detrended series, we examine the seasonal component. Before we can obtain a good seasonal model, we must first remove the cyclic component. In order to do this, we fit a 12 moving average followed by a 2 moving average, then subtract the fitted values of the moving average from the detrended values obtained above.


Detrended, Decycled Series
Log Sales
0.10
0.05
0.00
-0.05
-0.10
0
50
100
150
200
250
300
Time

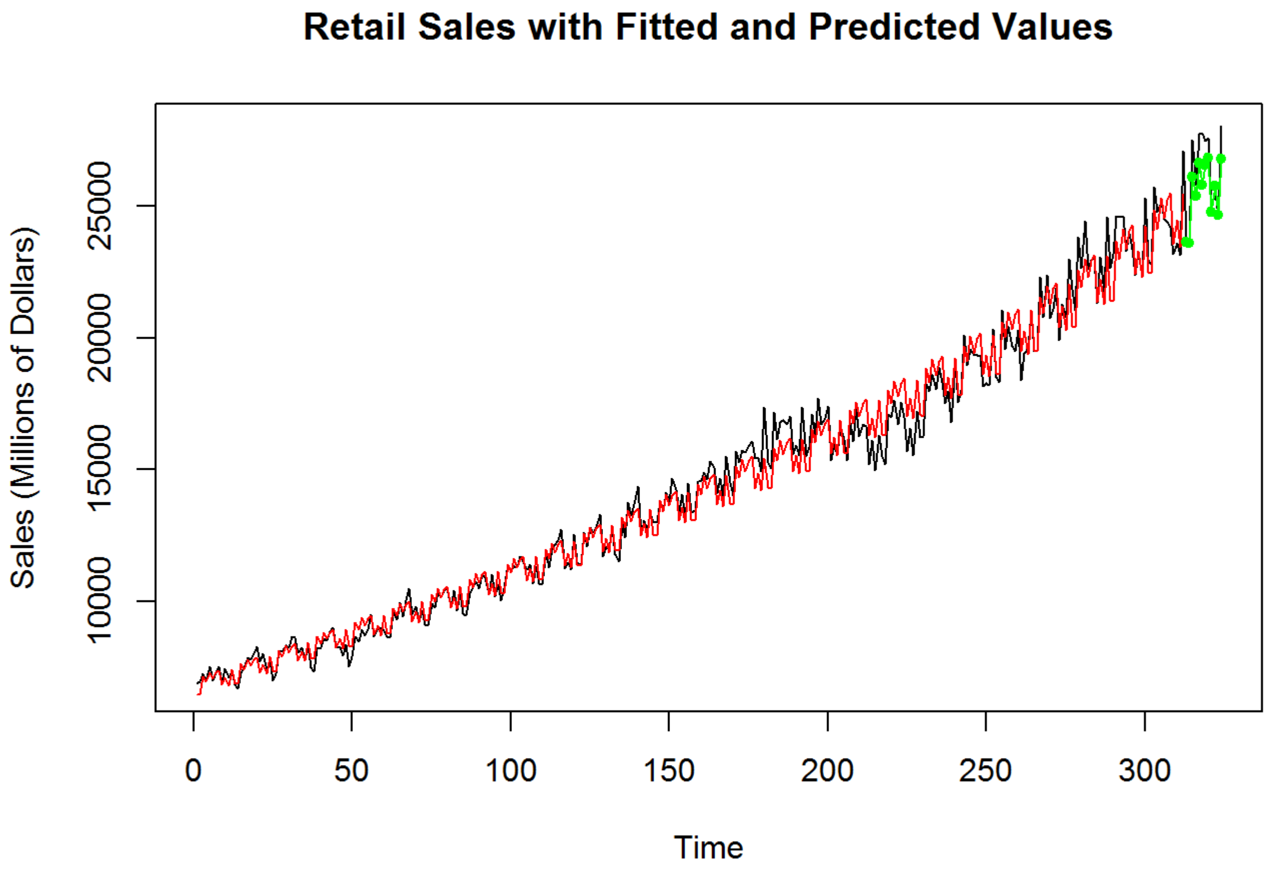
```
##
## Call:
## lm(formula = decycle$x ~ 0 + Month, data = decycle)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.065799 -0.013777 -0.000953  0.014064  0.060129
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## MonthApr  0.005731   0.004552   1.259  0.20902
## MonthAug  0.045707   0.004552  10.042  < 2e-16 ***
## MonthDec  0.023532   0.004552   5.170 4.40e-07 ***
## MonthFeb -0.061521   0.004552 -13.517  < 2e-16 ***
## MonthJan -0.055146   0.004552 -12.116  < 2e-16 ***
## MonthJul  0.039739   0.004645   8.554 7.31e-16 ***
## MonthJun  0.015227   0.004645   3.278  0.00118 **
## MonthMar  0.037806   0.004552   8.306 4.00e-15 ***
## MonthMay  0.048421   0.004552  10.638  < 2e-16 ***
## MonthNov -0.052297   0.004552 -11.490  < 2e-16 ***
## MonthOct -0.004018   0.004552  -0.883  0.37807
## MonthSep -0.039606   0.004552  -8.702 2.63e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02276 on 286 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.7654, Adjusted R-squared:  0.7556
## F-statistic: 77.77 on 12 and 286 DF,  p-value: < 2.2e-16
```

After removing the trend and cyclic components of the time series, we get a good look at the seasonal component. We first attempt to fit harmonic regressions to the detrended, decycled series. Fitting a two-component harmonic regression, only the cosine component is significant and reports an R^2 value of 0.3161. Increasing the number of components and fitting a four-parameter harmonic regression does not improve the model much, once we remove the insignificant sine components of the model results in an R^2 value of 0.3261.

Since neither of these regressions does a very good job of explaining the variance of the detrended, decycled data, we next attempt to utilize dummy regression to examine the contributions of individual months. Based on the dummy regression output above, we see that our analysis of the seasonal component from the Data Exploration section above is confirmed. Fall and winter months (September through February) with the exception of December have negative values indicating a decrease in spending, and spring and summer months (March through August) have positive estimate values indicating an increase in spending. The output also confirms our observations for the month of December, that despite winter months indicating a decrease in spending, spending at restaurants actually increases during this month. The dummy regression output reports an R^2 value of 0.7654, which tells us that over double the amount of variation in the reponse is explained by this model compared to the harmonic regression attempted previously. For this reason, we decide to proceed with the dummy regression model for the seasonal component of our time series.

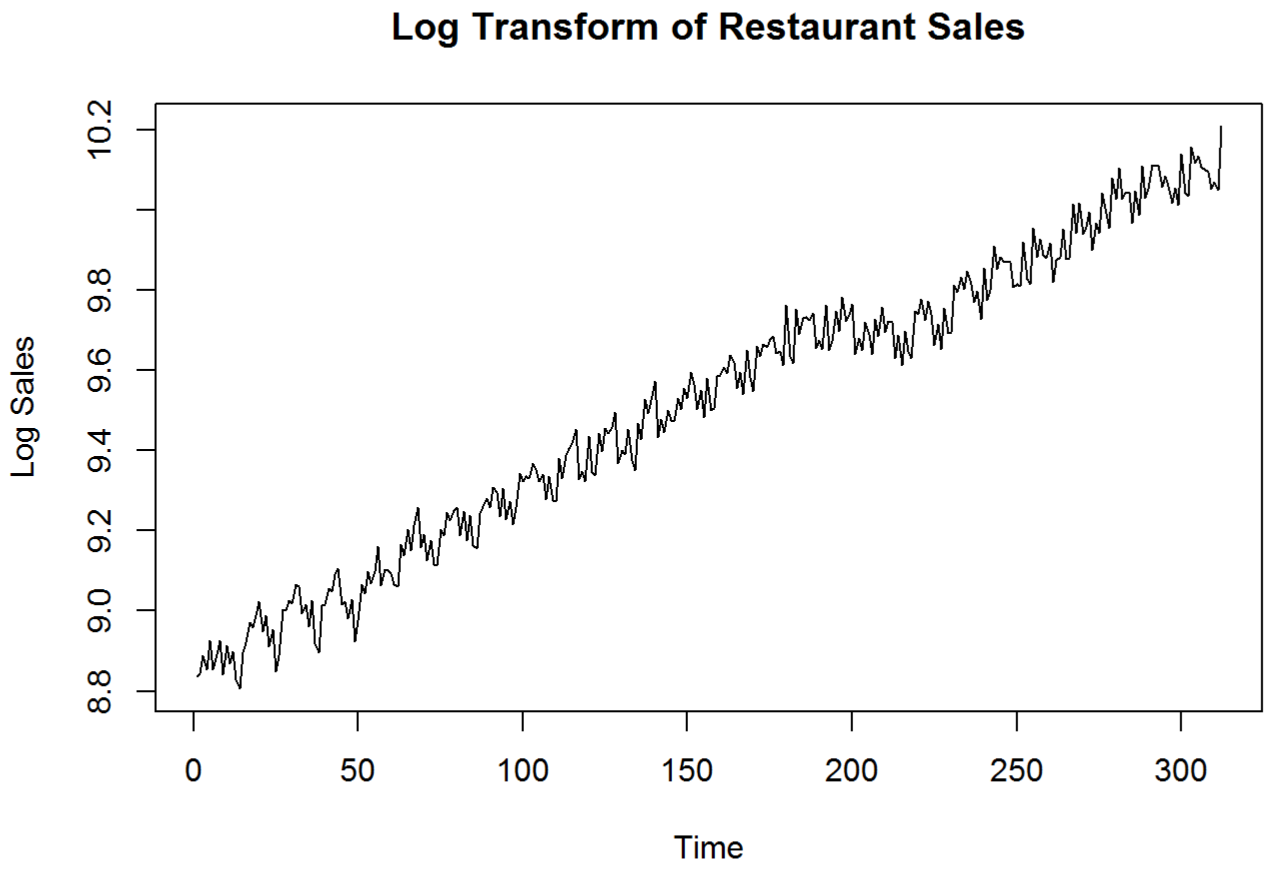
```
##
## Call:
## lm(formula = log ~ Time + I(Time^2) + I(Time^3) + Month, data = sales)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.100843 -0.024415 -0.001293  0.025284  0.117155
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.822e+00  1.090e-02 809.219  < 2e-16 ***
## Time         5.753e-03  2.345e-04  24.529  < 2e-16 ***
## I(Time^2)   -1.014e-05  1.740e-06  -5.826 1.48e-08 ***
## I(Time^3)    1.629e-08  3.654e-09   4.457 1.18e-05 ***
## MonthAug     3.826e-02  1.026e-02   3.730 0.000229 ***
## MonthDec     2.054e-02  1.026e-02   2.001 0.046264 *
## MonthFeb    -6.421e-02  1.026e-02  -6.261 1.34e-09 ***
## MonthJan    -5.801e-02  1.026e-02  -5.656 3.65e-08 ***
## MonthJul     3.209e-02  1.026e-02   3.129 0.001928 **
## MonthJun     8.419e-03  1.026e-02   0.821 0.412393
## MonthMar     3.257e-02  1.026e-02   3.176 0.001650 **
## MonthMay     4.350e-02  1.026e-02   4.241 2.97e-05 ***
## MonthNov    -5.827e-02  1.026e-02  -5.679 3.22e-08 ***
## MonthOct    -1.090e-02  1.026e-02  -1.063 0.288817
## MonthSep    -4.569e-02  1.026e-02  -4.454 1.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03698 on 297 degrees of freedom
## Multiple R-squared:  0.9903, Adjusted R-squared:  0.9898
## F-statistic:  2165 on 14 and 297 DF,  p-value: < 2.2e-16
```

The regression output for the full model is output above. We report an R^2 value of 0.9903 for the full model indicating that over 99% of the variation in the reponse can be explained by our model. Below, we plot the original data including the hold out, along with the fitted values from the full model above in red, and the predicted values in green. Overall, with the exception of the hitch corresponding to the 2008 recession, our fitted values follow the trend and seasonal fluctuations, and predict the hold-out values well. We will summarize the results and compare them to the Box-Jenkins analysis in the Summary section.

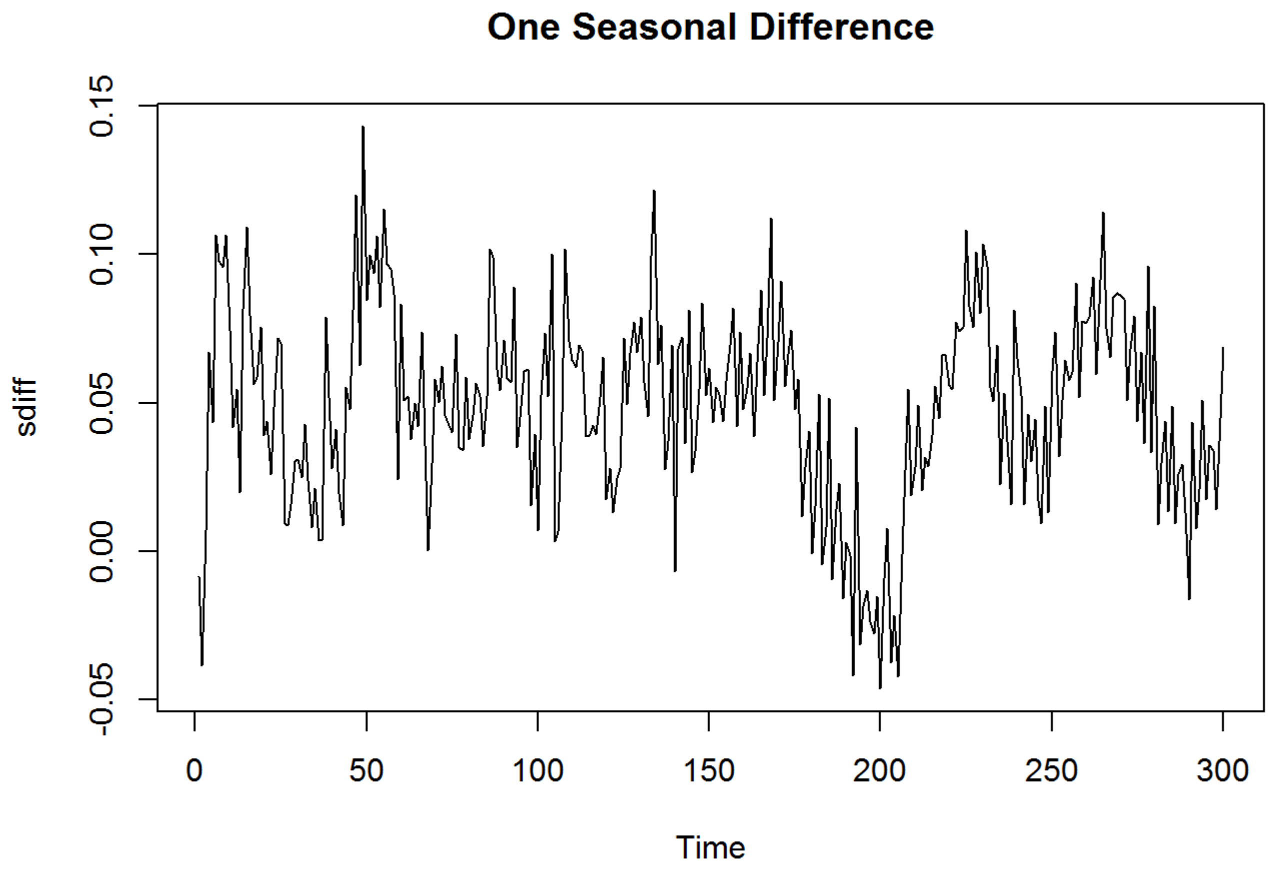


Box-Jenkins Analysis

Our goal for the Box-Jenkins analysis is to find a SARIMA model for our restaurant sales data, then use this model to predict the final year hold-out data, similarly to the decomposition method above. Since we have a multiplicative model, we begin by taking a log transform of the response in order to achieve constant variance. For convenience, we include a plot of the log transformed data again for this section.



To begin, we know that we have seasonal data, so we take a seasonal difference and run the Dickey-Fuller test in order to ensure that we have a stationary time series.



```
##
##  Augmented Dickey-Fuller Test
##
## data:  sdiff
## Dickey-Fuller = -3.4396, Lag order = 6, p-value = 0.04889
## alternative hypothesis: stationary
```

Examining the time series plot of the once seasonally differenced series, along with the Dickey-Fuller test output above, we conclude that we have a stationary time series and can continue with the analysis. In order to find the nonseasonal ARMA model, we propose six different models of varying parameters and choose the best one based on the lowest BIC value. The BIC values for each model are displayed in the table below. We choose to move forward with ARMA(1,1) and confirm our decision with the auto.arima and coeftest functions

Model	BIC
ARMA(0,1)	-1251.154
ARMA(1,0)	-1297.742
ARMA(1,1)	-1334.088
ARMA(1,2)	-1329.362
ARMA(2,1)	-1329.191
ARMA(2,2)	-1323.742

```
## Series: sdiff 
## ARIMA(1,0,1) with non-zero mean 
## 
## Coefficients:
##          ar1      ma1    mean
##       0.9121  -0.5607  0.0474
## s.e.  0.0311   0.0596  0.0071
## 
## sigma^2 estimated as 0.0006405:  log likelihood=678.45
## AIC=-1348.9   AICc=-1348.77   BIC=-1334.09
```

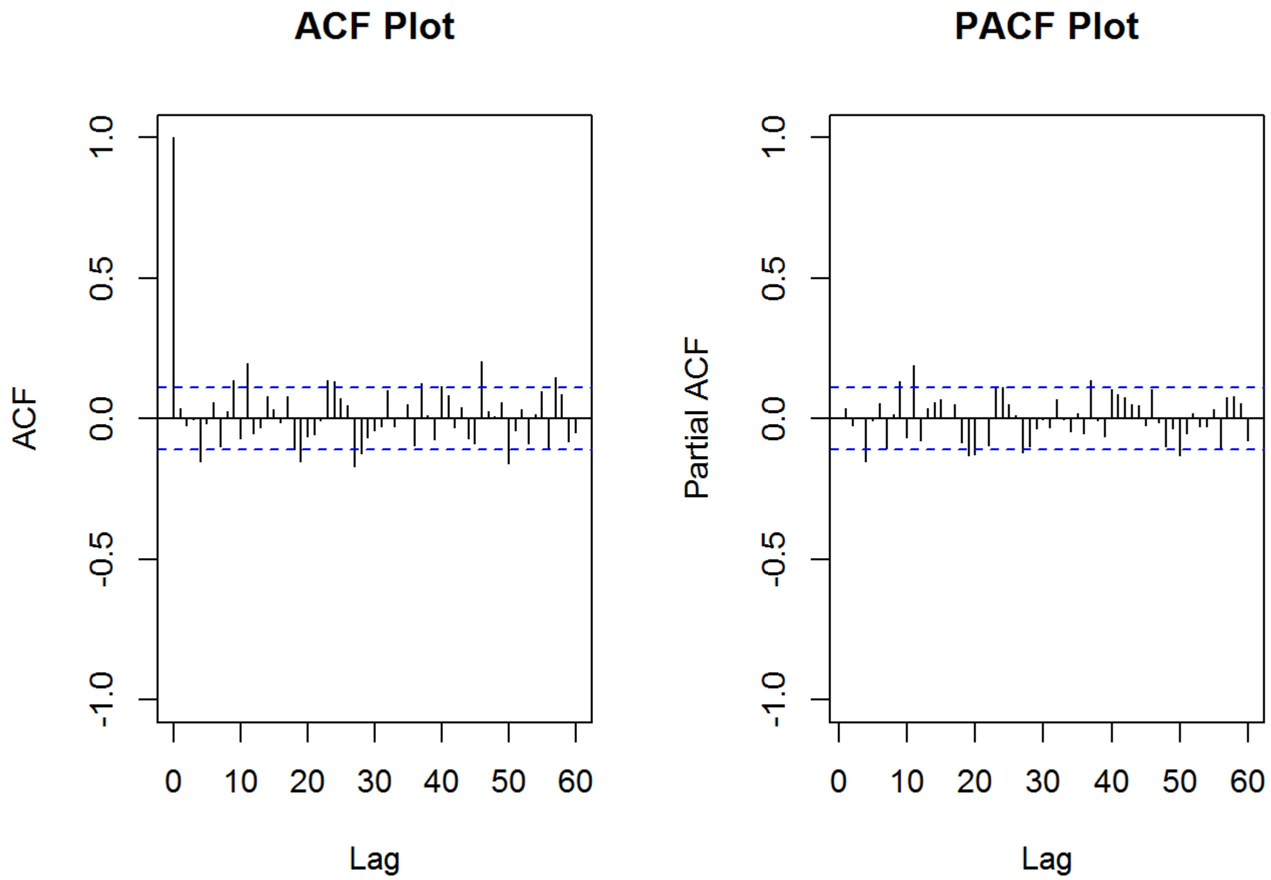
```
## 
## z test of coefficients:
## 
##             Estimate Std. Error z value  Pr(>|z|)    
## ar1        0.9121262  0.0310973 29.3314 < 2.2e-16 ***
## ma1       -0.5606700  0.0596169 -9.4045 < 2.2e-16 ***
## intercept  0.0474491  0.0070714  6.7100 1.947e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After determining the nonseasonal part of the model, we can proceed to find the seasonal ARMA model. We repeat the above process to iteratively find the best SARIMA model based on the lowest BIC value. For each model, we use the log value of the reponse along with the nonseasonal model found above when building the new models. The BIC values for each of these models is shown in the table below. We choose the SARIMA(1,0,1)(2,1,2)12 and examine the significance of each parameter using the coeftest function.

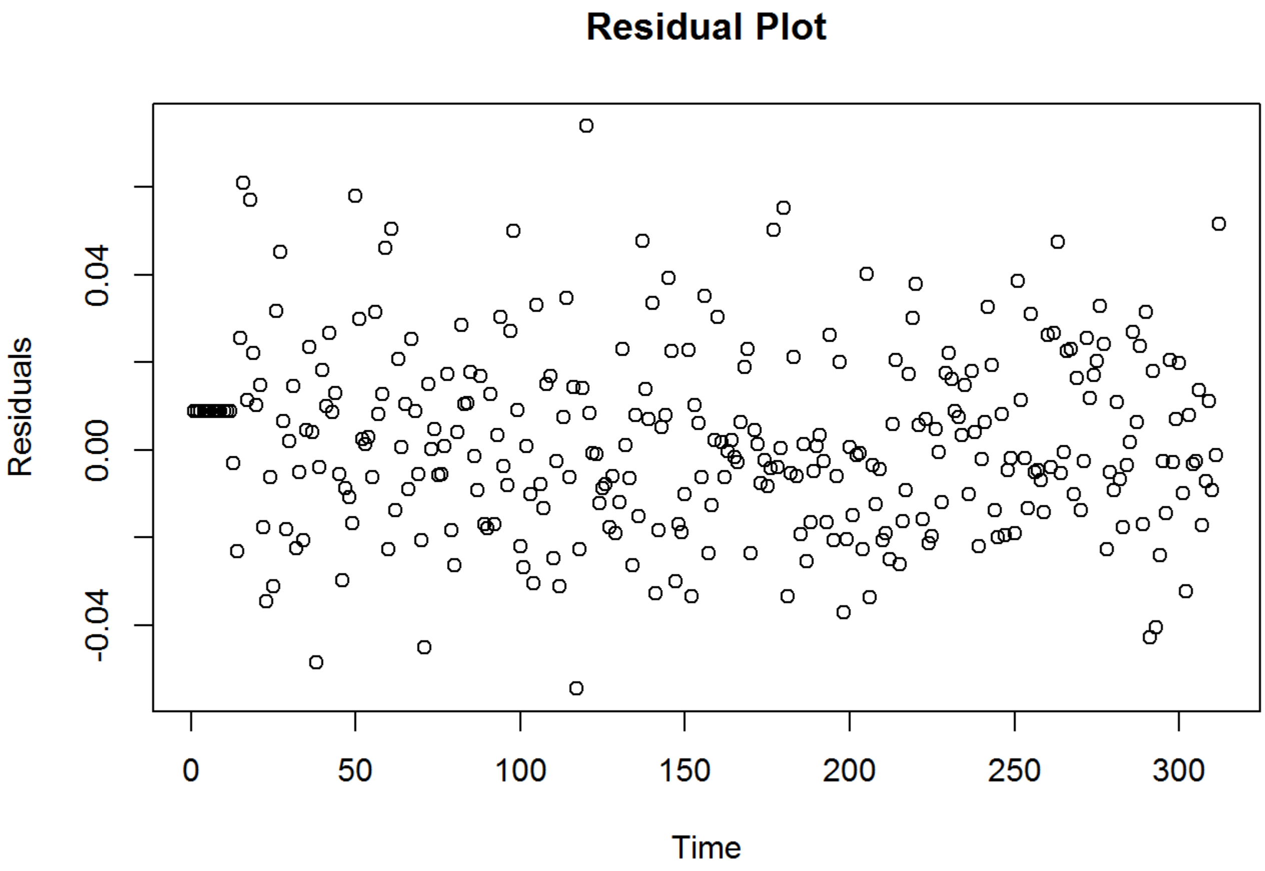
Model	BIC
SARIMA(1,0,1)(0,1,1)12	-1370.763
SARIMA(1,0,1)(1,1,0)12	-1337.831
SARIMA(1,0,1)(1,1,1)12	-1381.274
SARIMA(1,0,1)(0,1,2)12	-1379.293
SARIMA(1,0,1)(2,1,1)12	-1386.911
SARIMA(1,0,1)(2,1,2)12	-1407.199

```
##
## z test of coefficients:
##
##         Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.9963909  0.0037881 263.0326 < 2.2e-16 ***
## ma1  -0.4907071  0.0638353  -7.6871 1.505e-14 ***
## sar1  0.8154572  0.0780827  10.4435 < 2.2e-16 ***
## sar2 -0.6175055  0.0728866  -8.4721 < 2.2e-16 ***
## sma1 -1.3215610  0.0807942 -16.3571 < 2.2e-16 ***
## sma2  0.6426020  0.0807053   7.9623 1.688e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

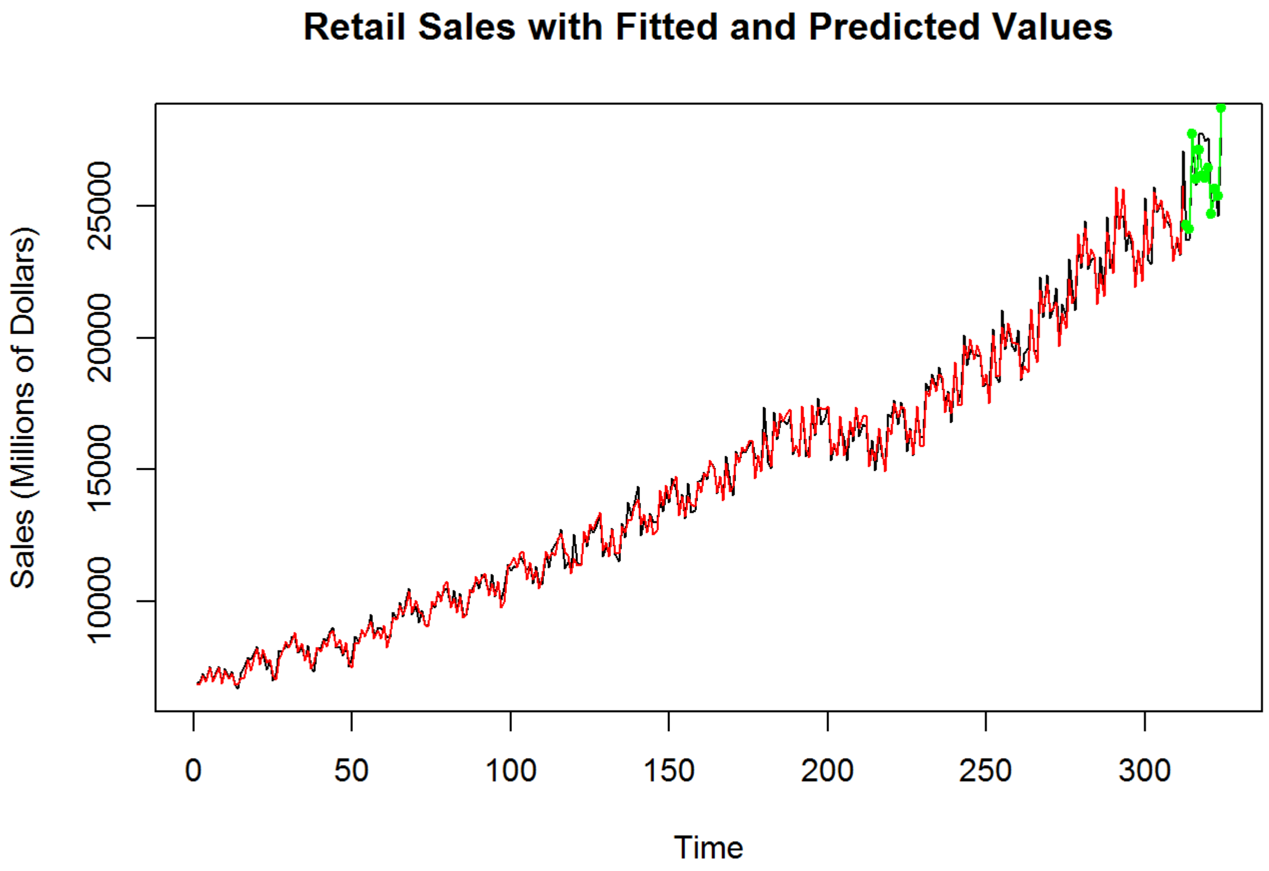
After confirming the significance of each parameter for the SARIMA model, we need to check the model's adequecy through analysis of the residuals. We want the ACF plot of the residuals to cut off after lag zero and the PACF plot of the residuals to stay around zero. We also check the plots of the residuals to ensure that they have mean zero and constant variance, as well as run the Box-Ljung test.



```
##
##  Box-Ljung test
##
## data:  residuals
## X-squared = 9.1372, df = 6, p-value = 0.166
```



After confirming the white noise of the residuals, we proceed to fit our SARIMA model to the data and predict the hold-out shown below. Below, we plot the time series data, the fitted values from our model in red, and the predcted values in green. We immediately notice that this model does a much better job of fitting the data, especially the hitch corresponding to the 2008 recession.



Summary

Measurement	Decomposition	Box-Jenkins
BIC	-1095.66	-1407.20
AIC	-1155.55	-1433.13
MSPE	859524	667681
MAPE	0.024039	0.005484

In summary, it is clear that the Box-Jenkins analysis is the superior method by all measurements, as well as through visual inspection. Through the use of this method, we achieved a mean absolute prediction error of a little over .5%, an improvement over the decomposition method by a factor of approximately 4.38. Since the ultimate goal of this project is to accurately predict the hold-out data, this is the most important statistic for determining the better method. The decomposition method did well in estimating the seasonal component but was unable to effectively deal with changes in the trend. With better estimation techniques, this can be improved, but we are able to bypass this with the use of the more robust Box-Jenkins method.

Sources

Data collected from: https://fred.stlouisfed.org/series/MRTSSM7221USN (https://fred.stlouisfed.org/series/MRTSSM7221USN)

Survey information: https://www.census.gov/retail/mrts/about_the_surveys.html (https://www.census.gov/retail/mrts/about_the_surveys.html)

Information on 2008 recession: https://en.wikipedia.org/wiki/Great_Recession#cite_note-IMF_WEO_2009-4 (https://en.wikipedia.org/wiki/Great_Recession#cite_note-IMF_WEO_2009-4)