STAT 153 Midterm 2 - ARIMA Modeling

Name

Date Submitted

I. Introduction

For this midterm project, I was able to work with three datasets and perform data analysis by fitting a particular ARIMA model on each of our datasets. Since I was only required to fit an ARIMA model to one particular dataset for the purposes of the report, I decided to work with the second dataset because I thought that it would be very interesting to dissect the trend and seasonal components of this particular data set; it looks like there may be multiple seasonal components – a high-frequency seasonal function as well as an underlying low-frequency seasonal function.

II. Exploratory Data Analysis

Violation of Stationarity & Variance-Stabilizing Consideration

Taking an initial glance at the second dataset **(Figure 1)**, there is an obvious periodic wavering of the univariate time series data, highlighting that seasonality will somehow play a factor in this dataset. In Figure 1, one can easily observe that the time series data is not stationary, as the mean function does not appear to be constant; for example, there seems to be a significant difference in the mean of the time series data depending on the time variable **(Figure 2)**. However, the variance seems to be roughly homoscedastic, so a variance-stabilizing transformation will not be entirely necessary for this data, as one can later see that using differencing will result in a stationary process. Thus, some form of differencing should be used to make the data appear stationary in order to forecast reasonable predictions using an appropriate (M)(S)ARIMA model.

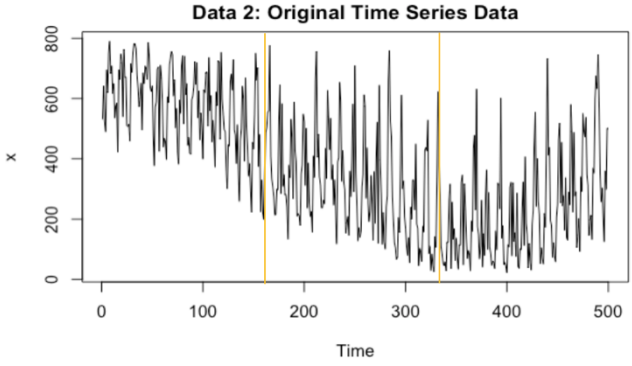

Figure 1: Graph of original (observed) time series data with yellow bars to separate into thirds to note non-stationarity.

Time <fctr>	Mean <dbl>	Variance <dbl>
1-167	569.2455	18413.85
168-333	310.3916	27789.55
334-500	231.2754	28111.67

Figure 2: Table of mean and variances of the original time series data grouped by their times into thirds to note non-stationarity due to the non-constant mean function. Variances are slightly different but not too significant to deem as heteroscedastic.

Transformation: (First) Differencing

To begin, an appropriate transformation to consider is the first difference of the data, as any patterns in the sample ACF and PACF plots of the resulting transformed (differenced) data can be used to identify if an ARMA, pure AR, or pure MA model would be appropriate for this differenced time series. However, due to seasonality, it is important to consider which type of

differencing to use (seasonal differencing, first differencing, multiple differences, etc.) by observing the sample ACF and PACF plots of the original data **(Figure 3)**.

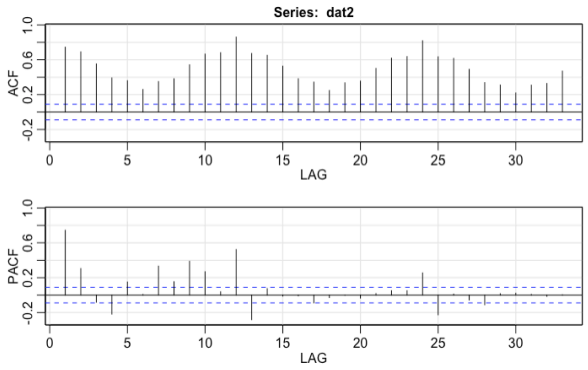

Figure 3: Graphs of the sample autocorrelation and sample partial autocorrelation functions of the original data. Note the periodic behavior of the sample ACF plot and the significant deviations in the sample PACF plot for the first 12 lags, specifically lags 12/13 as well as lags 24/25.

The sample ACF plot of the original data (Figure 3) exhibits strong seasonality with lag of multiples of 12, and the sample PACF plot exhibits significant behavior for the first 12 lags (and 13), but it also has a significant PACF value at lag h = 24 (a multiple of 12) as well as 25. Thus, a transformation to consider is the first difference of the data to make the data appear roughly stationary. After taking the first difference of the data, the resulting differenced data does appear roughly zero-mean stationary **(Figure 4)**. Thus, differencing the data removed the non-constant mean function that initially violated the assumption of stationarity and resulted in sample ACF/PACF plots that peak at lags of multiples of 12 **(Figure 5)**.

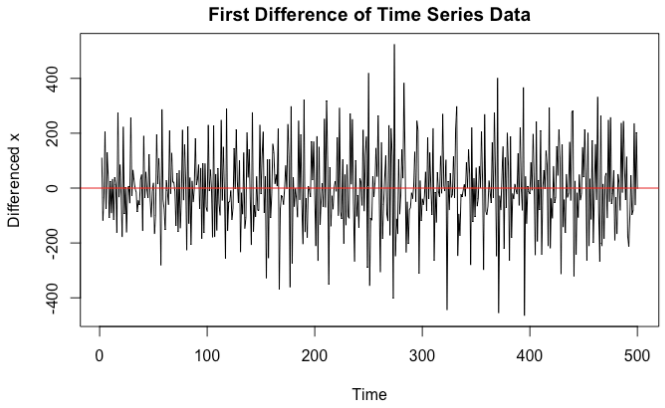

Figure 4: Graph of the first difference of the original (observed) time series data; appears to be roughly stationary with mean zero.

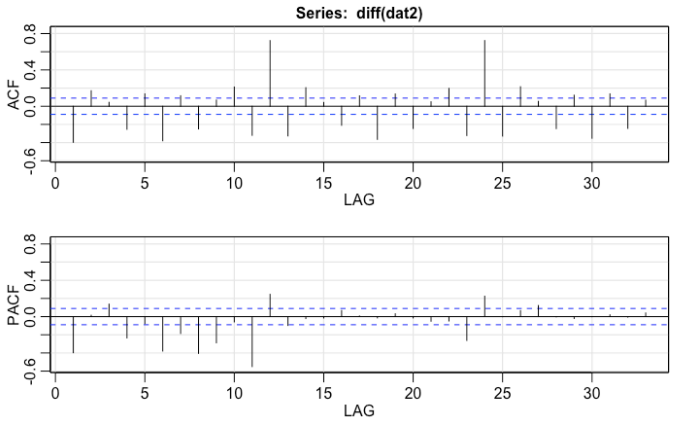

Figure 5: Graphs of the sample autocorrelation and sample partial autocorrelation functions of the differenced transformations. Note the peak lags of the sample ACF at multiples of 12 and the positive/negative alternating behavior of the sample PACF.

Seasonal Differencing of the (First) Differenced Data

Now that the original time series data has been transformed to a zero-mean stationary time series via first differencing (also meaning that we know that there will be a non-trivial, or non-zero, value of *d* in our ARIMA model of our original time series data), one can observe the sample ACF and PACF plots for the transformed data to try and estimate the *p* and *q* parameters in our ARIMA(*p*, *d*, *q*) model. When observing the sample ACF/PACF plots with just the first difference (Figure 5), one can observe that the sample ACF plot still has a peak at lags h = 12 ± 1, 24 ± 1, etc., and the sample PACF plot has reverse, consecutive peaks at lags h = 11/12, 23/24, etc. Thus, there must be some sort of multiplicative seasonal component that still remains at yearly lags (assuming that this dataset has monthly data points and is not a simulated dataset explains why the lags are at

multiples of 12). As a result, we should consider *both* first differencing with multiplicative seasonal differencing of lag 12 as the transformation. After differencing the data again using a lag of 12, the resulting process appears to be zero-mean stationary **(Figure 6)**, and the seasonal patterns of the sample ACF/PACF plots of this newly transformed data appear to be mostly eliminated despite one outlier at lag 12. **(Figure 7)**. With this particular transformation, it is now possible to identify which type of ARIMA model best fits on the original time series data.

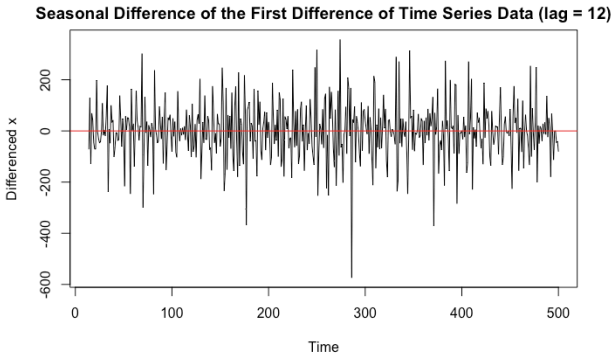

Figure 6: Graph of the seasonal difference of the first difference of the original (observed) time series data; appears to be roughly stationary with mean zero.

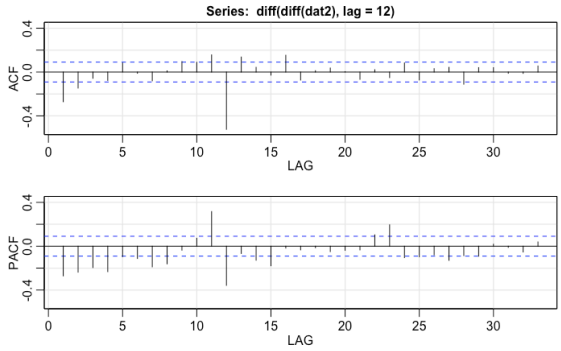

Figure 7: Graph of the seasonal difference of the first difference of the original (observed) time series data; appears to be roughly stationary with mean zero.

III. Identifying Suitable (M)(S)ARIMA Models

Which Model?

With the implications of a pure AR(p) model having a sample PACF that cuts off after lag p and a pure MA(q) model having a sample ACF that cuts off after lag q, the transformed data appears to be some combination of an ARMA(p, q) process with some form of multiplicative seasonality, as opposed to a pure MA(q) or pure AR(p) process. Thus, the observations of the sample ACF and sample PACF plots from the transformed data can be used to conclude that the p, d, and q values (number of parameters) in the ARIMA model of the original time series data are non-trivial with multiplicative seasonality of $S = 12$, and we can test this model compared to other ARIMA models by utilizing cross-validation analysis and comparing information criteria values.

Concerning the original data…*	**Reasoning:**
Why **not** AR(p) or MA(q)?	Sample ACF/PACF do not cut off abruptly.
Why **not** ARIMA($p, d, 0$) or ARIMA($0, d, q$)?	Sample ACF/PACF of the first differenced data do not cut off abruptly.
Why **not** ARIMA(p, d, q)?	Sample ACF/PACF still exhibit noticeable seasonality.
Why **not** ARIMA(p, d, q) × $(P, D, 0)_{12}$?	Sample ACF of the first differenced data is significant at lag of 11mod12, 0mod12, 1mod12.
Why ARIMA(p, d, q) × $(0, D, Q)_{12}$?	Sample ACF of the first differenced data is significant at lag of 11mod12, 0mod12, 1mod12, while its sample PACF exhibits a sharp turn-around at 11mod12 and 0mod12.

*assuming that p, d, q, P, D, Q, are not over-generalized/arbitrary and are significant/non-trivial (non-zero)

Choosing a Model Based on Information Criteria Comparison

After examining the sample ACF and PACF plots of the first differenced data as described earlier with Figure 4, there still seems to be an underlying seasonal pattern at lags of multiple 12. Since

the corresponding sample ACF plot (Figure 5) exhibits a very distinct pattern at lags 12 ± 1, 24 ± 1, this behavior can be attributed to both the non-zero parameters within the AR and MA processes after differencing. As a result, ARIMA(1, 1, 1) is the particular candidate to test along with a multiplicative MA seasonal component with period 12 and its counterparts, as necessarily shown in Figure 5 and Figure 6. Other forms in the ARIMA(1, 1, 1) family should also be considered with a seasonal differencing of S = 12, and information criteria such as AIC, AICc, and BIC (and later, cross validation) should be used to determine which exact order of the ARIMA process best fits the data. A summary of model comparison and their information criteria is shown in **Figure 8**.

Model <fctr>	AIC <dbl>	AICc <dbl>	BIC <dbl>
ARIMA(1,1,1)x(0,1,0) 12	10.077002	10.081099	9.093860
ARIMA(1,1,1)x(0,1,1) 12	9.552069	9.556231	8.577357
ARIMA(1,1,1)x(1,1,0) 12	9.790996	9.795158	8.816284
ARIMA(1,1,1)x(1,1,1) 12	9.556032	9.560275	8.589749
ARIMA(1,1,1)	10.890896	10.895058	9.916184

Figure 8: Table of values highlighting information criteria comparison between various SARIMA(1, 1, 1, P, D, Q, 12) models in an attempt to perform model selection.

The model with the lowest AIC, AICc, and BIC of the five potential candidates was the **ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$** model, as it was able to capture the seasonality of the moving-average process as observed in the sample ACF in Figure 7. In addition, the table below summarizes the strengths and weaknesses of the five candidates within the ARIMA(1, 1, 1) family chosen for information criteria comparison **(Figure 9)**:

Model	Advantages	Disadvantages	Rank
1: ARIMA(1, 1, 1) × (0, 1, 0)$_{12}$	Simple model	Fails to capture seasonality of MA	4
2: ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$	Low AIC, AICc, BIC	Ignores seasonal AR	1
3: ARIMA(1, 1, 1) × (1, 1, 0)$_{12}$	Low AIC, AICc, BIC	Fails to capture seasonality of MA	3
4: ARIMA(1, 1, 1) × (1, 1, 1)$_{12}$	Low AIC, AICc, BIC	Complex model (sAR not needed)	2
5: ARIMA(1, 1, 1)	Simple model	Fails to capture seasonality of MA	5

Figure 9: Table of appropriate models to fit the transformed data and their advantages/disadvantages. Ranks are based on AIC, AICc, and BIC values – Model 2 is the best based on information criteria.

Cross-Validation Analysis

Considering the different types of information criteria and comparing which SARIMA model to fit on the transformed (seasonal differencing of first differenced) data, cross-validation analysis can be applied to each the five SARIMA models to obtain CV scores as a method to determine which of the five models is the best. To conduct the cross-validation analysis, the 350 data points (70% of the data) will be used as the training set and the last 150 data points (30% of the data) will be the test set. For each model above, the training set can be used to predict the next ten observations through an iterative process and compared the mean squared errors of the 15 sets of ten estimates (150 values) to the test set to obtain a mean squared error term ("cross-validation

score"). The model with the lowest CV score should be chosen as the best fitting model, which also happens to be **ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$** for this particular data set **(Figure 10)**.

Model <fctr>	AIC <dbl>	AICc <dbl>	BIC <dbl>	CV <dbl>
ARIMA(1,1,1)x(0,1,0) 12	10.077002	10.081099	9.093860	106490.96
ARIMA(1,1,1)x(0,1,1) 12	9.552069	9.556231	8.577357	75471.12
ARIMA(1,1,1)x(1,1,0) 12	9.790996	9.795158	8.816284	92717.98
ARIMA(1,1,1)x(1,1,1) 12	9.556032	9.560275	8.589749	76488.62
ARIMA(1,1,1)	10.890896	10.895058	9.916184	677838.90

Figure 10: Table containing CV (cross-validation) scores for each of the five model candidates. Model 2 has the lowest CV score (and, thus, is the best model based on cross-validation procedure).

Why Split the Data Into 70% Training Set and 30% Test Set?

In cross-validation analysis, the specific division of the original data set into its training and test set should not impact the model selection assuming a robust model being fit on to the data. We could have utilized an 80%/20% training/test set, or a 60%/40% training/test set, and the results would also be the same, and we would want to choose the **ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$** model to represent the original time series data. However, if a smaller training set were to be used, another model may fit better, but, regardless, an **ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$** model is still very reasonable.

IV. Model Fitting and Forecasting

Analysis of Residuals & Ljung-Box Testing

Based on the results of both information criteria (AIC, AICc, BIC) comparison and cross-validation analysis between different suitable models, the **ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$** model seems to be the most reasonable model to fit on the original data. However, when fitting the model on to the data, the particular model fit should have roughly stationary standardized residuals (showing that there is no seasonal component), and the sample ACF of the residuals should behave similarly to a stationary process, and this is satisfied as shown in **Figure 11**. However, the Gaussian assumption may not be the best distribution to characterize the noise via the probability plot.

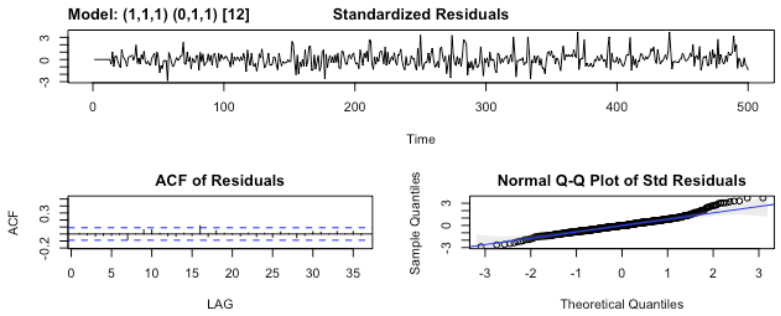

Figure 11: Graphs resulting from the *sarima()* function in R assuming an ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$, with the standardized residuals and its sample autocorrelation function and normal probability plot.

Lastly, a Ljung-Box test is conducted to determine whether the data seems like a good fit by observing whether or not the Ljung-Box statistics are insignificant (above the 5% threshold) to be

compared to white noise, as there would not be group-wise autocorrelations between residuals at certain lags **(Figure 12)**. Since all of these p-values for the test are insignificant, **ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$** does prove to be a very reasonable model for the original time series data.

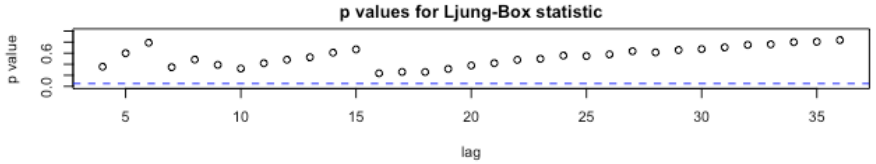

Figure 12: Graph of the Ljung-Box test statistics. Note that many of the p-values exceed the 5% significance level bands of group-wise autocorrelations, noting that the model is quite reasonable for the original data.

Parameter Estimates of ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$

Parameter	Estimate	Standard Error
$\hat{\phi}$ (MA)	0.3663	0.0476
$\hat{\theta}$ (AR)	-0.9280	0.0178
$\hat{\Phi}$ (seasonal AR)	N/A	N/A
$\hat{\Theta}$ (seasonal MA)	-0.8598	0.0301

```
Coefficients:
         ar1      ma1     sma1
      0.3663  -0.9280  -0.8598
s.e.  0.0476   0.0178   0.0301
```

Figure 13: Table of parameter estimates and their standard errors in the ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$ model for the original time series data. Obtained through the *sarima*() function in R.

Figure 13 displays the corresponding estimates and standard errors of the parameters of the ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$ model on the original data. By observing the roots of the AR and MA polynomials, one can easily show that this ARIMA process is both causal and invertible. In addition, each of these values is statistically significant from 1 using a t-test, indicating that these parameter estimates are all non-trivial, and the number of parameters in each very likely to be correct and necessary in the model fit (in terms of a lower bound minimum number of parameters).

Forecasting & Conclusions

This SARIMA model can be used to forecast future observations; **Figure 14** displays a graph of such predictions and **Figure 15** displays the actual predictions, which appear to be a reasonable fit to the actual data set (it even predicts the small upward trend between the troughs). Overall, the forecasts of this model look quite decent in relation to the previous behavior of this data set.

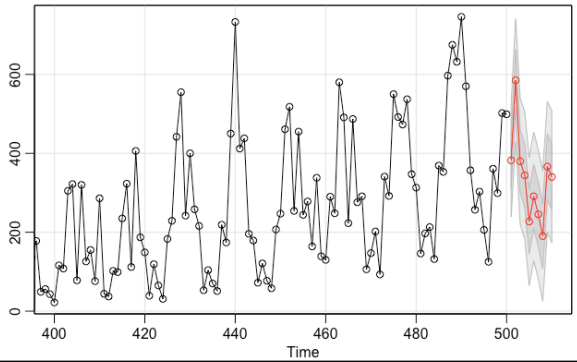

Figure 14: Graph of predictions and the corresponding confidence intervals associated with the forecasted values. From *sarima.for*() in R.

Time Index <int>	Prediction Value <dbl>
501	382.2626
502	585.2541
503	379.8301
504	344.3452
505	227.1662
506	291.0039
507	245.3819
508	190.3210
509	366.4625
510	340.1852

Figure 15: Table of predictions of the next ten points assuming an ARIMA(1, 1, 1) × (0, 1, 1)$_{12}$ model on the previous 500 data points.