Data Analysis Report

Luhuan Wu, 3032590198, Zihao Wang, 3032590289

3-1-2017

1 Introduction

The purpose of this project is to perform time series analysis on the given 5 data sets and make predictions on the next 104 observations. In this report, we specifically discuss about the data analysis and predictions of the 5th data set, which is $q5train.csv$, with the statistical tools from lectures, books and the Internet. We use R as our programming language and the relevant R-code is given in the Appendix.

2 Data Analysis

2.1 Exploratory data analysis

First, we read the file $q5train.csv$ into R workspace as $ts5$, and make the plot Figure 1.

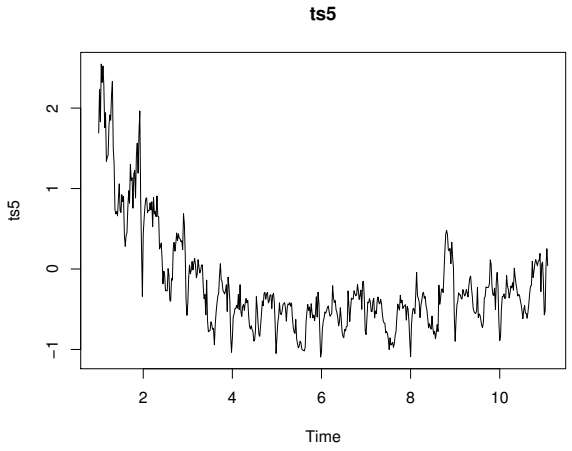


Figure 1: ts5

From 1, we have 4 main observations:

(i) **Seasonality:** There exists a periodicity of 52, which corresponds to the fact that it is a weekly dataset.

(ii) **Trend:** There is an obvious trend that goes downwards in observations interval [0, 400] with decreasing slope and tends to go upwards with small slope in interval [400, 500].

(iii) **Variance:** The variance is not stablized.

(iv) Data are all above -2.

According to these observations, we make adjustments as follows:

2.1.1 Set seasonality and shift to positive values

First we set the periocity of $ts5$ equal to 52, and shift the $ts5$ upwards by 2 to get $ts.positive$ so we could perform log and sqrt tranformation.

2.1.2 Variance stablization

We take the log and square-root transformation on $ts.positive$ to get $ts5.log$ and $ts5.sqrt$. From FIgure 2, we can see that the variance after transformation is more stablized.

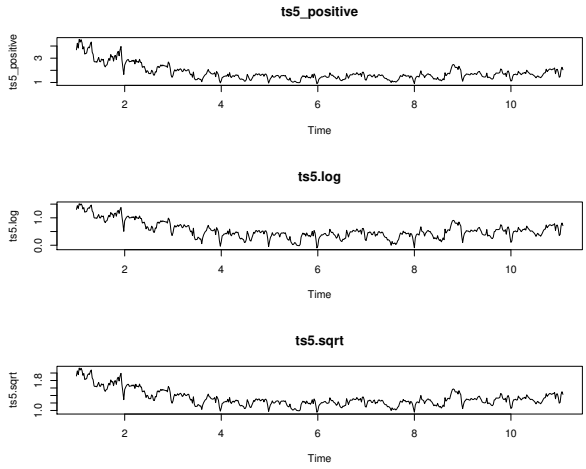


Figure 2: Comparison after transformation

2.1.3 Differencing

Because we see a trend and weekly seasonality in Figure 2, we decide to take a first order differencing at lag 1 and a first order differencing at lag 52, to get timeseries ts5.sqrt.dD. We get the time series plots Figure3 and acf, pacf plotsand Figure 4.

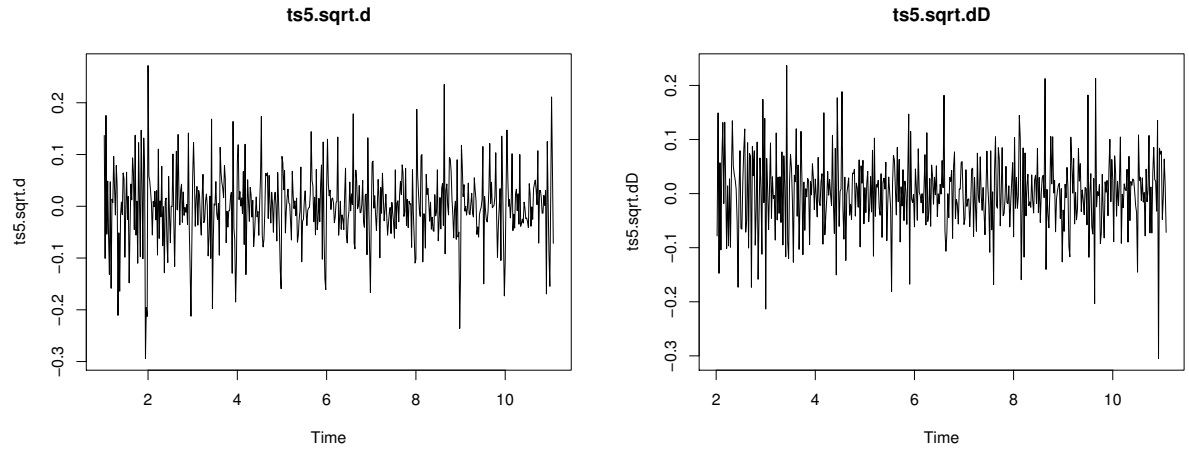


Figure 3: differenced ts5 plot

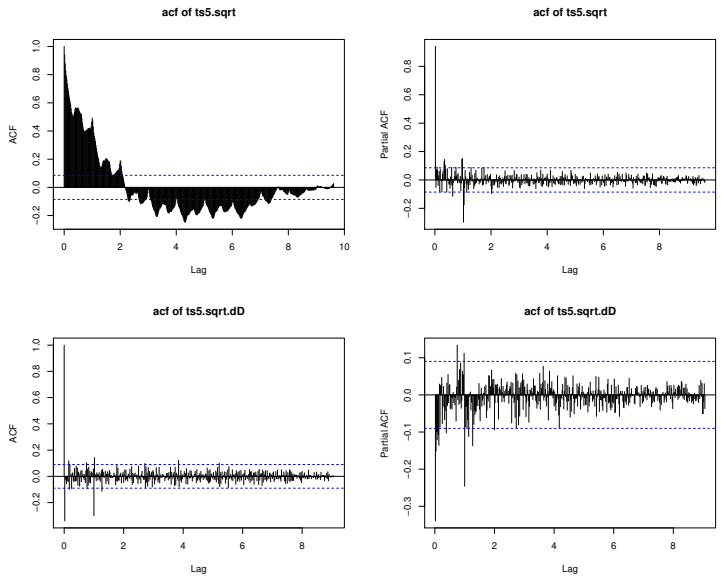


Figure 4: acf and pacf before and after differencing

We can see that two times differencings make the data more stationary-like. Beisides, the seasonal spikes in acf plot of $ts5.sqrt.dD$ significantly reduce, hence the seasonal difference is necessary.
In further analysis, we choose sqare-root-transfromed data $ts5.sqrt$ to fit the models. In addition, from the acf and pacf plot of $ts5.sqrt.dD$, we can find the acf cuts off after lag 1 and seasonal lag 52, and the pacf tails off. Based on the plots, we identify SARIMA model is $(0,1,1) \times (0,1,1)_{52}$ as a probability.

2.1.4 Another way to make assumptions

Using the *auto.arima* in *library* "$forecast$", we get the suggested SARIMA model $(0,1,2) \times (0,1,1)_{52}$.

2.2 Fit Multiplicative SARIMA Models I

In this part, we focus on finding the optimal estimations for seasonal coefficients in Multiplicative SARIMA Models.

Combining our intuitive model from analysis of plots which is $(0,1,1) \times (0,1,1)_{52}$ and the *auto.arima*'s suggestion $(1,1,2) \times (0,1,1)_{52}$, we try to get more multiplicative SARIMA models by "overfitting" a little bit on *ar* and *ma* coefficients. We first consider "overfitting" the seasonal coefficients. We "overfit" the model $(0,1,1) \times (0,1,1)_{52}$ to get models $(0,1,1) \times (0,1,2)_{52}$, $(0,1,1) \times (1,1,1)_{52}$, and overfit the model "$(0,1,2) \times (0,1,1)_{52}$ to get $(0,1,2) \times (0,1,2)_{52}$ and $(0,1,2) \times (1,1,1)_{52}$.

2.3 Diagnostics I

First, we check the detailed information of the 6 models. The R output follows:

```
1 arima(x = ts5.sqrt, order = c(0, 1, 1), seasonal = list(order = c(0, 1,
      1), period = 52))
2 Coefficients:
3           ma1      sma1
4       -0.4944   -0.4935
```

```
s.e.    0.0535    0.0476
sigma^2 estimated as 0.003322:  log likelihood = 669.77,  aic =
    -1333.54
Call:
arima(x = ts5.sqrt, order = c(0, 1, 1), seasonal = list(order = c(0, 1,
    2), period = 52))
Coefficients:
          ma1      sma1     sma2
      -0.4959   -0.5045   0.0224
s.e.    0.0535    0.0560   0.0572
sigma^2 estimated as 0.00332:  log likelihood = 669.85,  aic = -1331.69
Call:
arima(x = ts5.sqrt, order = c(0, 1, 1), seasonal = list(order = c(1, 1,
    1), period = 52))
Coefficients:
          ma1      sar1      sma1
      -0.4956   -0.0337   -0.4680
s.e.    0.0535    0.0995    0.0904
sigma^2 estimated as 0.003321:  log likelihood = 669.83,  aic =
    -1331.66
Call:
arima(x = ts5.sqrt, order = c(0, 1, 2), seasonal = list(order = c(0, 1,
    1), period = 52))
Coefficients:
          ma1      ma2      sma1
      -0.4641   -0.1573   -0.5147
s.e.    0.0442    0.0465    0.0487
sigma^2 estimated as 0.003237:  log likelihood = 675.1,  aic = -1342.2
Call:
arima(x = ts5.sqrt, order = c(0, 1, 2), seasonal = list(order = c(0, 1,
    2), period = 52))
Coefficients:
          ma1      ma2      sma1     sma2
      -0.4651   -0.1568   -0.5213   0.0140
s.e.    0.0444    0.0466    0.0560   0.0567
sigma^2 estimated as 0.003236:  log likelihood = 675.13,  aic =
    -1340.26
Call:
arima(x = ts5.sqrt, order = c(0, 1, 2), seasonal = list(order = c(1, 1,
    1), period = 52))
Coefficients:
          ma1      ma2      sar1      sma1
      -0.4649   -0.1569   -0.0219   -0.4981
s.e.    0.0443    0.0466    0.0984    0.0899
sigma^2 estimated as 0.003236:  log likelihood = 675.12,  aic =
    -1340.25
```

In addition, we can use $BIC()$ function in R to get the BIC values as follows :

```
> BIC(m5.sqrt.012011)
[1] -1325.569
> BIC(m5.sqrt.012012)
[1] -1319.473
> BIC(m5.sqrt.012111)
```

```
[1] -1319.461
> BIC(m5.sqrt.012112)
[1] -1313.268
> BIC(m5.sqrt.112011)
[1] -1326.578
> BIC(m5.sqrt.112012)
[1] -1320.574
```

From the R output, we can check the standard erors for the coefficients. From *Chapter 8: Model Diagnostics* in *Time Series Analysis with Applications in R, Second Edition (Cryer 2008)*, we know that if the s.e. is in the same or above the magnitude of the estimated coefficient, or the estimation is not statistically away from zero, this coefficient is very likely to be overfitting. For example, for model $(0,1,1)\times(0,1,2)_{52}$, $sma2 = 0.0224$ is less than half of $s.e. = 0.0572$. In addition, 0.0224 is not significantly different from 0. In addition, except for $sma2$ coefficient, the remaining coeffients are nearly the same. So model $(0,1,1)\times(0,1,2)_{52}$ is overfitted. Similarly, we can find $(0,1,1)\times(1,1,1)_{52}$, $(0,1,2)\times(0,1,2)_{t2}$ and $(0,1,2)\times(1,1,1)_{t2}$ to be overfitted as well. Now, we have 2 optimal models so far, which are $(0,1,1)\times(0,1,1)_{52}$ and $(0,1,2)\times(0,1,1)_{52}$.

2.4 Fit Multiplicative SARIMA Models II

So far, we have 2 optimal models: $(0,1,1)\times(0,1,1)_{52}$ and $(0,1,2)\times(0,1,1)_{52}$. Now we " overfit" ar and ma to get the optimal estimate for non-seasonal coefficients.
First, we fit $(0,1,1)\times(0,1,1)_{52}$ to get $(1,1,1)\times(0,1,1)_{52}$.

```
Call:
arima(x = ts5.sqrt, order = c(1, 1, 1), seasonal = list(order = c(0, 1,
    1), period = 52))

Coefficients:
          ar1      ma1     sma1
      0.3607  -0.7957  -0.5144
s.e.  0.0736   0.0493   0.0483

sigma^2 estimated as 0.003198:  log likelihood = 677.89,  aic =
   -1347.78
```

However, we do not get the expected overfitting result. The ar and ma coefficients behave statistically different from the original one. So we should take this model into consideration. We further overfit this model to get model $(1,1,2)\times(0,1,1)_{52}$:

```
Call:
arima(x = ts5.sqrt, order = c(1, 1, 2), seasonal = list(order = c(0, 1,
    1), period = 52))

Coefficients:
          ar1      ma1     ma2      sma1
      0.5523  -1.0018  0.1361  -0.5057
s.e.  0.1459   0.1539  0.1035   0.0484

sigma^2 estimated as 0.003192:  log likelihood = 678.68,  aic =
   -1347.36
```

We can see that $s.e.$ of $ma2$ is nearly equal to $ma2$, and other coefficients are no significantly different from those of$(1,1,1)\times(0,1,1)_{52}$. So this is overfitting. Also, overfitting to $(2,1,1)\times(0,1,1)_{52}$ yields the similar

result. So we stick with $(1,1,1) \times (0,1,1)_{52}$.

2.5 Diagnostics II

Now we run diagnostics on three models: $(0,1,1) \times (0,1,1)_{52}$, $(0,1,2) \times (0,1,1)_{52}$ and $(1,1,1) \times (0,1,1)_{52}$.

2.5.1 AIC, BIC

```
> AIC(m5.sqrt.011011)
[1] -1333.541
> AIC(m5.sqrt.012011)
[1] -1342.197
> AIC(m5.sqrt.111011)
[1] -1347.777
> BIC(m5.sqrt.011011)
[1] -1321.07
> BIC(m5.sqrt.012011)
[1] -1325.569
> BIC(m5.sqrt.111011)
[1] -1331.149
```

2.5.2 Residuals' correlation check

We use *tsdiag*() function in R which gives the output in Figure 5.

For the three models, there is basically no trend in standardized residuals, and the ACF of the residuals suggest white noise property. For model $(0,1,1) \times (0,1,1)_{52}$, corresponding p-value is around the blue band for all lags, and for model $(0,1,2) \times (0,1,1)_{52}$, the corresponding p-values for the first 3 lags are significantly above the blue band but the rest are aound. However, for model $(1,1,1) \times (0,1,1)_{52}$, in generall, most p-values except for the last three are significantly above the blue band. So we do not have statistically significant evidence to accept independence of the error terms in the first 2 models, while for the last model, we could accept the null hypothesis.

2.5.3 Cross Validation

Furthermore, we perform Cross Validation test on the 6 models, where k ranges from 1 to 5. We use MSE(Mean Sqruare Error) to estimate the errors. We get the following R output and the plot Figure 6:

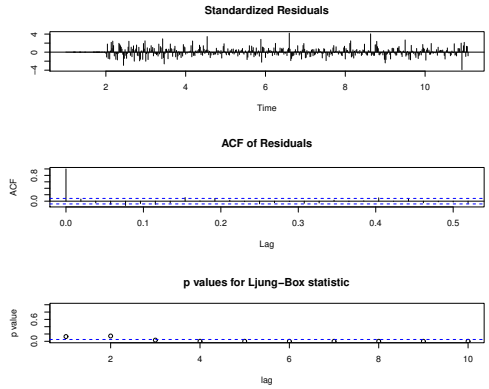
```

> MSE5.sqrt.011011
[1] 0.05026911 0.03702469 0.17117148 0.06571076 0.08994842
> MSE5.sqrt.012011
[1] 0.05660641 0.04139083 0.18246108 0.07509313 0.09226209
> MSE5.sqrt.111011
[1] 0.06157830 0.06194312 0.18914376 0.07625655 0.09664936
```

In General, model $(0,1,1) \times (0,1,2)_{52}$ has lowest MSE while model $(0,1,2) \times (0,1,1)_{52}$ has second lowest MSE.

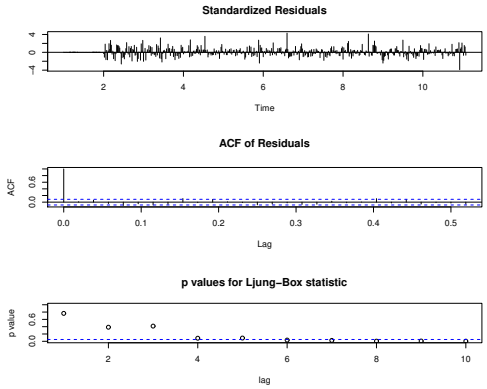
2.5.4 Summary

Overall, we can make a summary for these 3 models. See Table **??**.

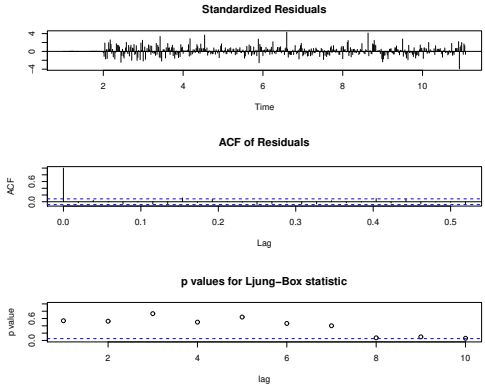
From Table **??**, we can see that the last two models outperfomance the first one in AIC, BIC and log likelihood, the first one has lowest MSE. However, we choose the first model $(0,1,1) \times (0,1,1)_{52}$ to be the



(a) $(0,1,1) \times (0,1,1)_{52}$



(b) $(0,1,2) \times (0,1,1)_{52}$



(c) $(1,1,1) \times (0,1,1)_{52}$

Figure 5: tsdiag

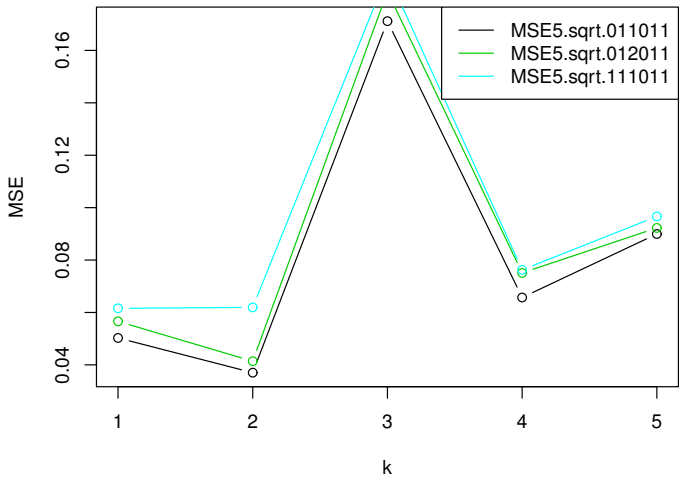


Figure 6: CV MSE

most optimal one since AIC and BIC are just heuristics while MSE is a more reliable criteria since it is an out-out-sample prediction validation.

3 Forecast

The prediction for the next 104 observations is show in Figure **??**. The black line stands for the real world data and the red line for predictions.

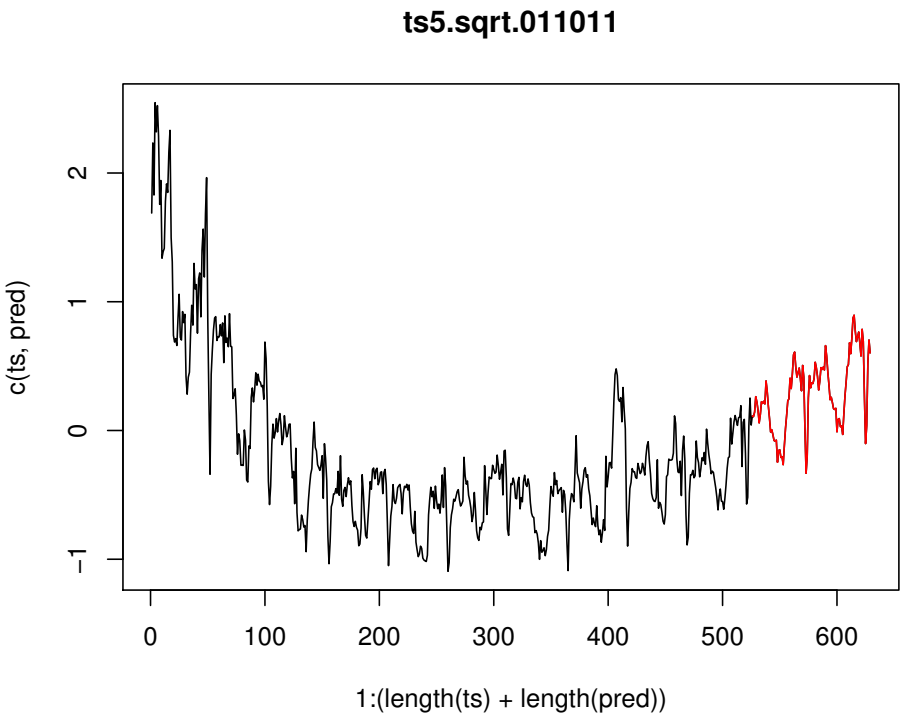


Figure 7: prediciton

4 Conclusions

We derive 2 models after exploratory data analysis and using *auto.arima*(). Based on that, we overfit the seasonal coefficients to check the selected model. We find seasonal coefficients to be (0,1,1), and then overfit the non-seasonal coefficients of the current-opimal models. However, we can not derive a satistisfying model which has best performance on all criteria including AIC, BIC, MSE, p-values, etc. In the end, we choose $(0,1,1) \times (0,1,1)_{52}$ which has lowest MSE for its simplicity.

SARIMA model	AIC	BIC	MSE	log likelihood
$(0,1,1) \times (0,1,1)_{52}$	-1333.541	-1321.07	0.10187308 0.07501398 0.34934075 0.12701968 0.17157842	669.77
$(0,1,2) \times (0,1,1)_{52}$	-1342.197	-1325.569	0.11445500 0.08433764 0.37143150 0.14543136 0.17596332	675.1
$(1,1,1) \times (0,1,1)_{52}$	1347.777	-1331.149	0.1243153 0.1276242 0.3844795 0.1477139 0.1842729	677.89

Table 1: models summary

5 Appendix

R code in Github page : `https://github.com/leahwu/stat153/blob/master/UCB/q5/q5_train.Rmd`.