

Midterm Two Project

Erica Wong (SID:24302634) and Bryana Gutierrez (SID:24504003)

April 18, 2017

In this project we received data and were tasked with coming up with the best predictions for the next 104 data points. Since our data are all time series, we took the time series approach. For the purposes of this report, we focused on the first data set Q1. Our first step was to load and format the data so that we may work with it. Then we checked to make sure that the data didn't have any missing values. Q1 did not contain any missing values so then we proceeded to plot the data.

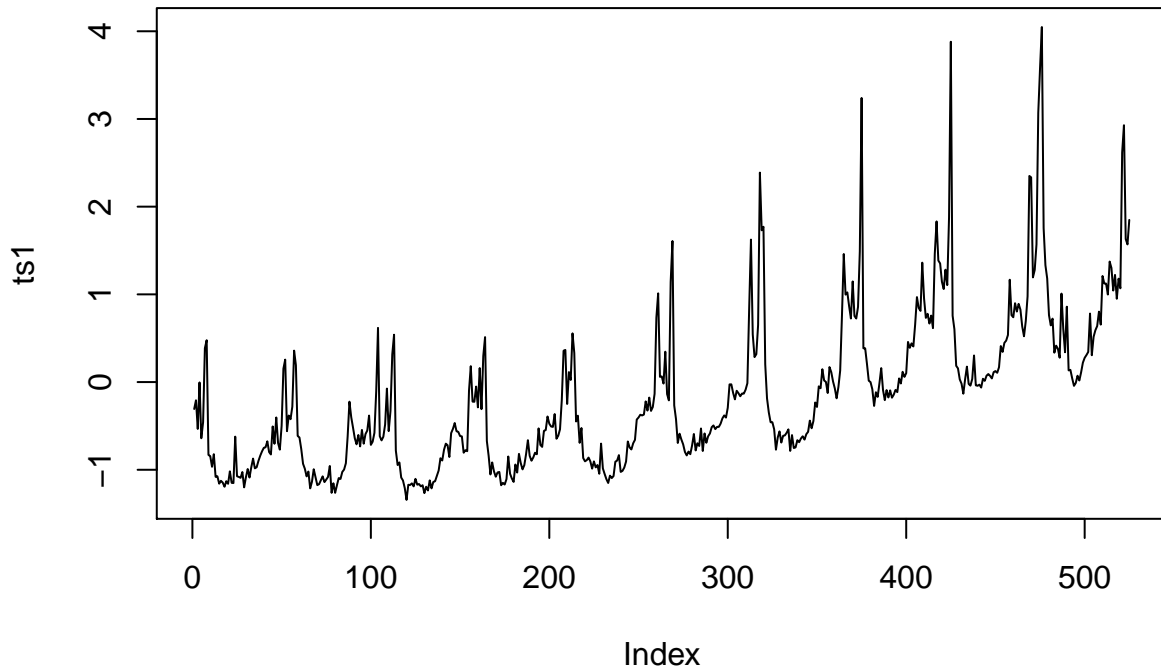


Figure 1: Plot of Original Data

From the plot of the original data, Figure 1, we can see that the variance of the data increases with time. Also, given the slight upward trend, we considered that the data might need to be differenced. To deal with the increase in variance we took the log of the data. Since there are negative values in the data, we added a constant of 1.5 which would later be subtracted when we calculated the predictions. The log transformed data looks as follows:

This helped take care of the variance that increases with time. Now it almost looks like the data decreases in variance with time. There is still a slight upward trend in the data, to take care of this we differenced the data. This is how the differenced data looked like.

```
#Doing Different Transformations to get a better idea of the wanted model
#took difference twice so data looked reasonably stationary
ts3 <- diff(ts2)
plot(ts3, type = "l")
```

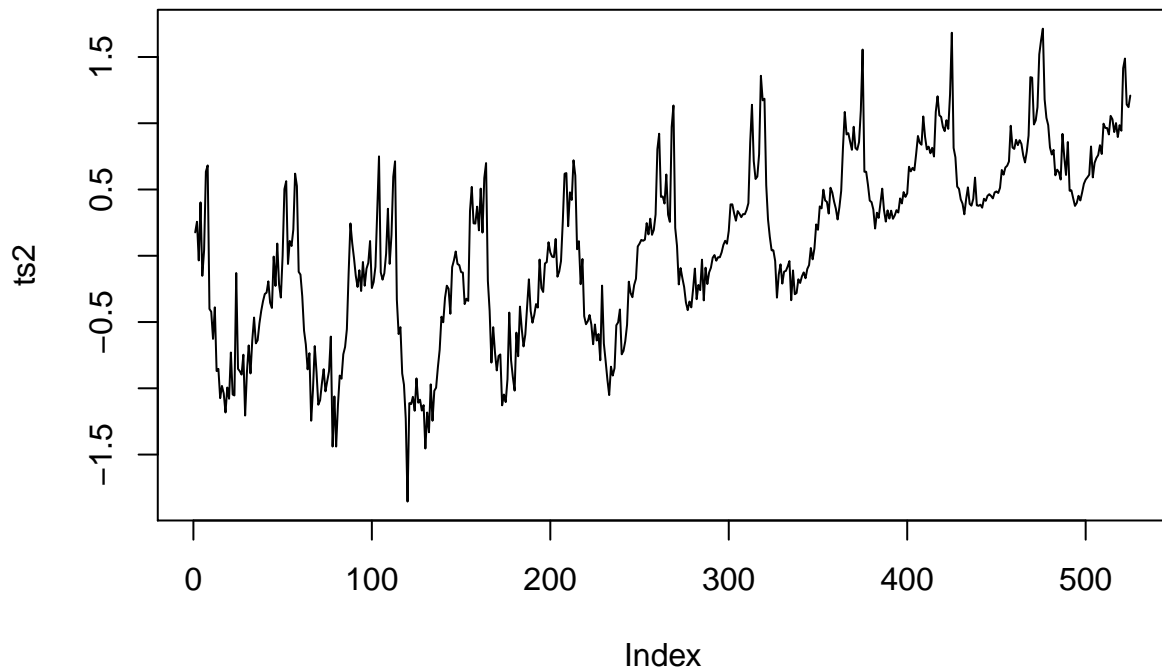
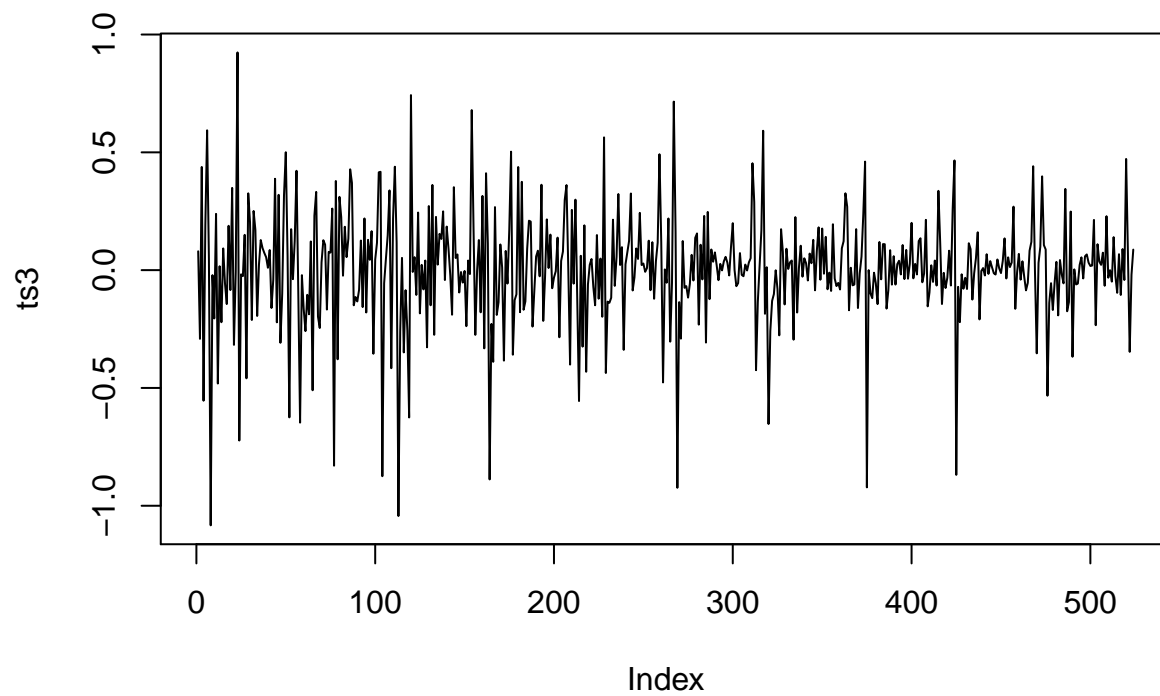


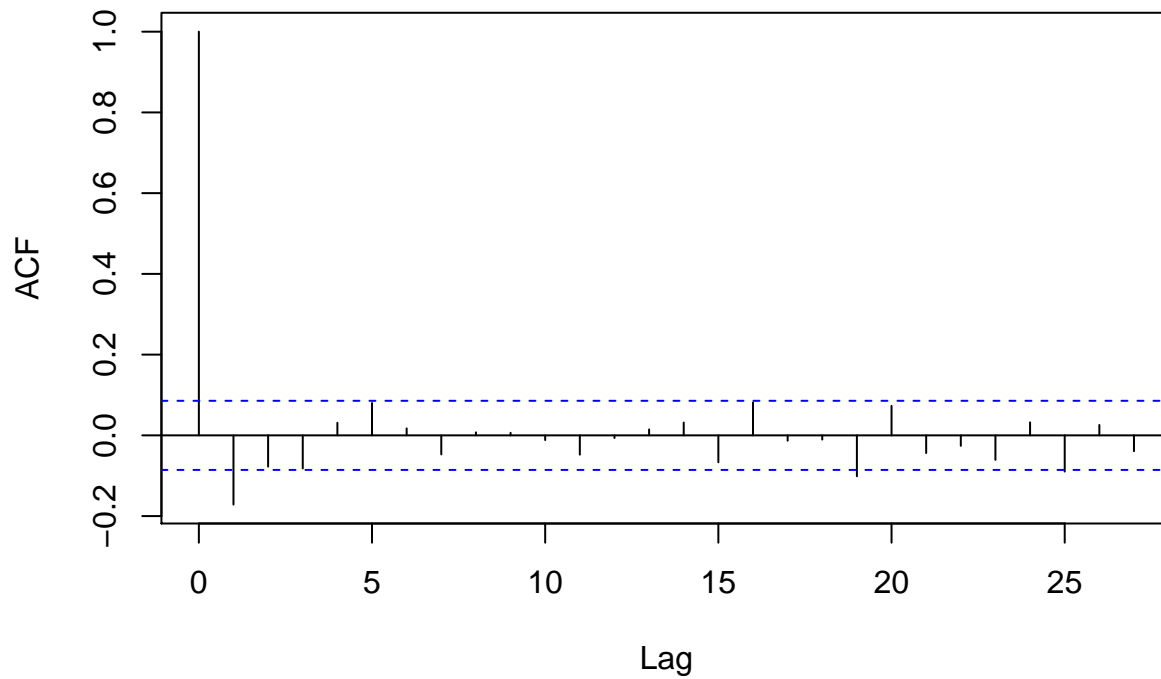
Figure 2: Plot of Data Shifted Up 1.5



This looked fairly stationary. So we moved on to look at what the ACF and PACF could tell us about the remaining data.

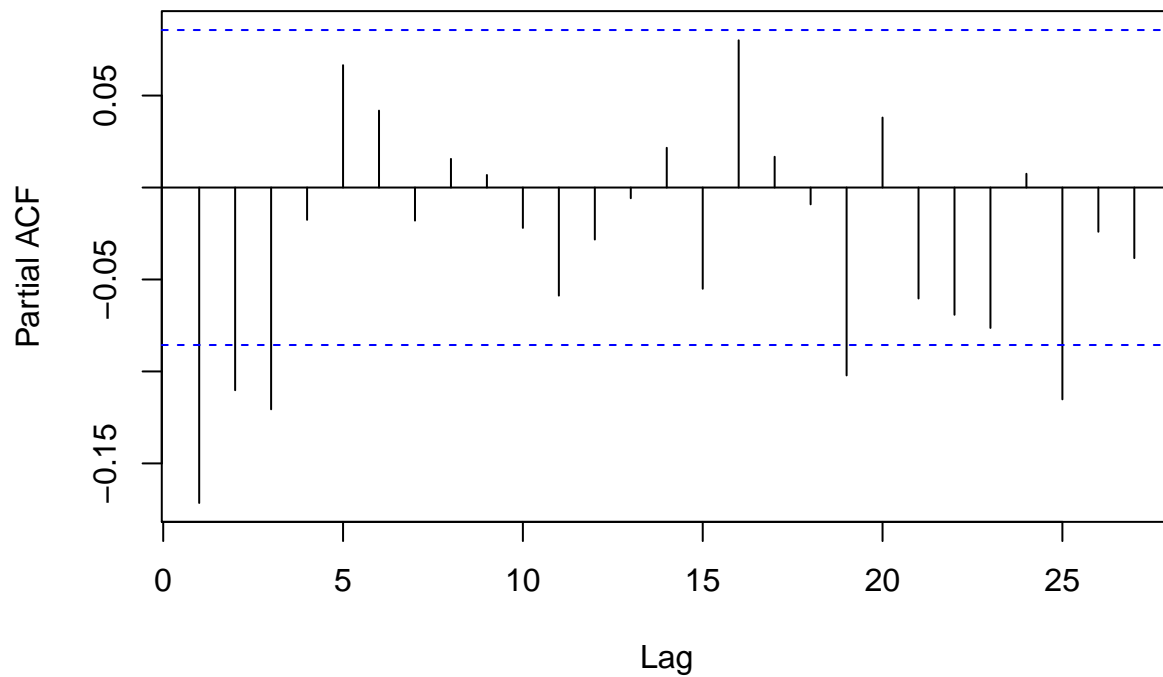
```
#plotted acf  
acf(ts3)
```

Series ts3



```
pacf(ts3)
```

Series ts3



We can see that there is some strong AR presence from the large PACF values and also some MA presence from some high values in the ACF plot. The ACF plot would imply that the model may be MA(2) because the lags die off after those points. The PACF plot shows that our model may also have some AR(4) component

because after those lags the lags become insignificant. To help give us an initial guess at the ARIMA distribution, we used `auto.arima`. Since an ARIMA model takes differencing into consideration, we used the data prior to differencing for the `auto.arima` function. R gave us the following output:

```
#running auto.arima to get a better idea of what the model is
auto.arima(ts2)
```

```
## Series: ts2
## ARIMA(4,1,2)
##
## Coefficients:
##          ar1          ar2          ar3          ar4          ma1          ma2
##        -0.4732   -1.0731   -0.2677   -0.1538    0.2798    0.9215
## s.e.    0.0778    0.0626    0.0477    0.0471    0.0685    0.0391
##
## sigma^2 estimated as 0.054:  log likelihood=23.99
## AIC=-33.98   AICc=-33.76   BIC=-4.15
```

We used this information to help give us a better idea about what our model is and then moved on to conduct some model selection.

One method of model selection is Cross Validation. First we wrote a function that will run the cross validation process for us. The output of the function will be the vector of mean-squared errors (MSE) from the different cross validation groupings from the same model. To compare the MSEs from each of the groups, we took the average MSEs from each of the models we tested.

Before actually conducting cross validation, we first had to come up with some models. We knew that in order for our data to look stationary, we must take 1 difference at one point. Additionally, from the PACF and ACF plot, we can see that there are AR(4) and MA(2) components in our model. So, we started off by thinking we should test an ARIMA(4,1,2) model. This model also happen to match what `auto.arima()` gave us. However, we also noticed from the original data (Figure 1) that there is some seasonality that exists. Since, this is data that is collected every week over years, we believed that the seasonality may be lag 52 because there are 52 weeks in a year. So, that is how we created our original model. We then conducted cross validation with models that were similar to what our original prediction was.

After coming up with various models and running cross validation, the best model in our case was from model 1 and the second best was model 6. In table 1, we can see a comparison of the MSEs.

	Average MSE
Model 1	2.5975
Model 2	2.8855
Model 3	4.2923
Model 4	2.7160
Model 5	2.6129
Model 6	2.6119

Table 1: MSE Table

After conducting cross validation, we still wanted to which model other model selection methods would tell us to pick. So, we conducted AIC and BIC. We knew that for AIC and BIC model selection, we wanted to pick the model with the lowest value. Tables 2 and 3 give us the values of AIC and BIC respectively. The best models for both are model 6, followed by model 4, and then model 1. However, when generating model 4, we did get some errors, so we believed that it would be best to eliminate that model from being considered.

From all three of our model selection processes, we get model 6 and model 1 as the best. So to help make our final decision, we decided to graph both of them to see which one we liked better or believed fit in our model better.

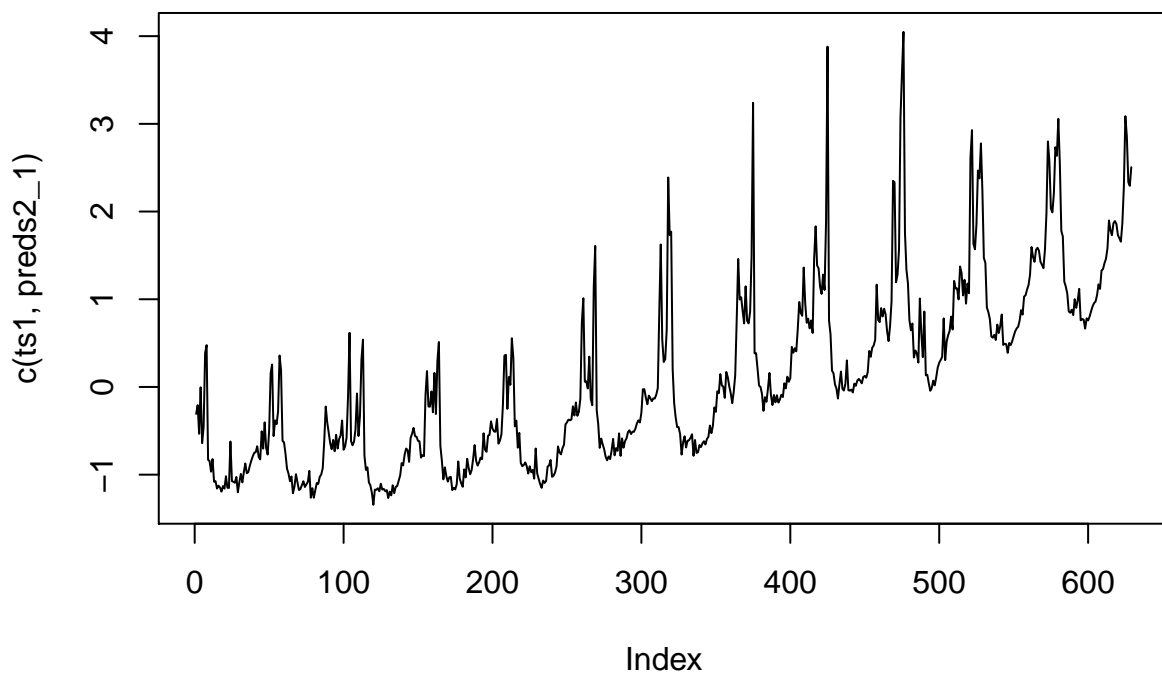


Figure 3: Model 1 Plot

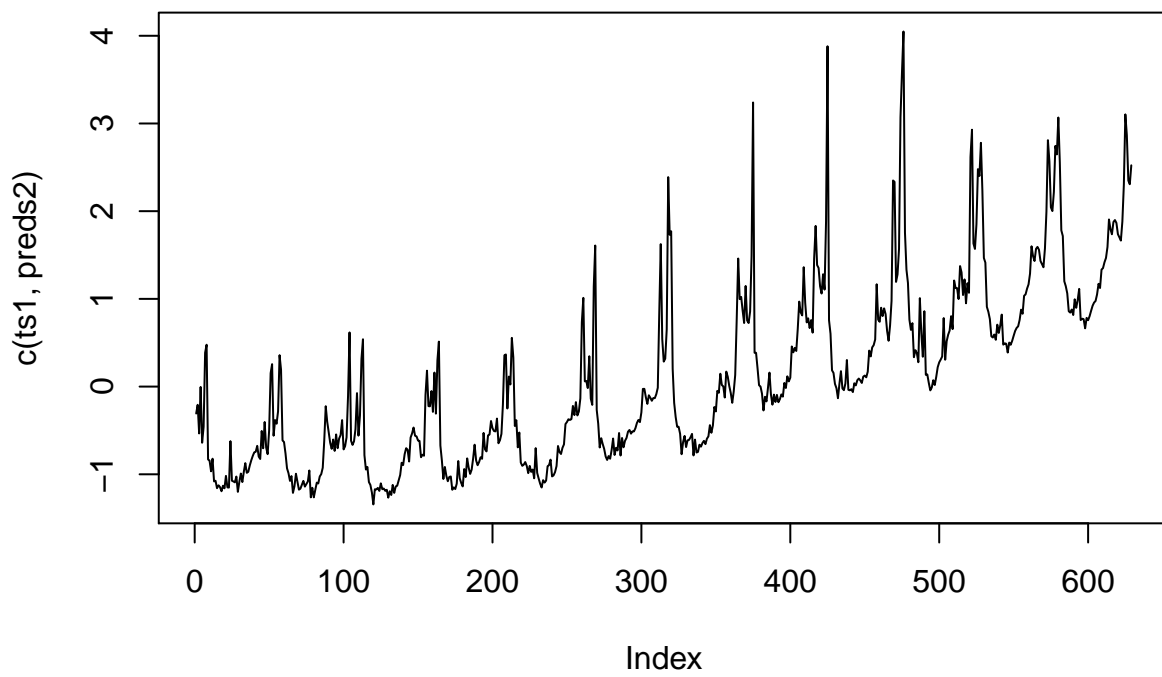


Figure 4: Model 6 Plot

	AIC Value
Model 1	-115.2144
Model 2	-107.8405
Model 3	-60.0934
Model 4	-121.1718
Model 5	-116.2298
Model 6	-117.2269

Table 2: AIC Table

	BIC Value
Model 1	-76.8609
Model 2	-69.5042
Model 3	-26.0015
Model 4	-78.5569
Model 5	-73.6149
Model 6	-83.1349

Table 3: BIC Table

The plots look the same to us, so we just went with model 6 because overall we believed that it performed better in model selection, ranking consistently in the top two best models. So, the final model that we went with was `arima(ts2, order = c(4,1,1), seasonal = list(order = c(1, 0, 1), period = 52))`.