Sean Pili & Michael McDonald
STAT 4534: Time Series
Final Paper

# Background/Data Cleaning

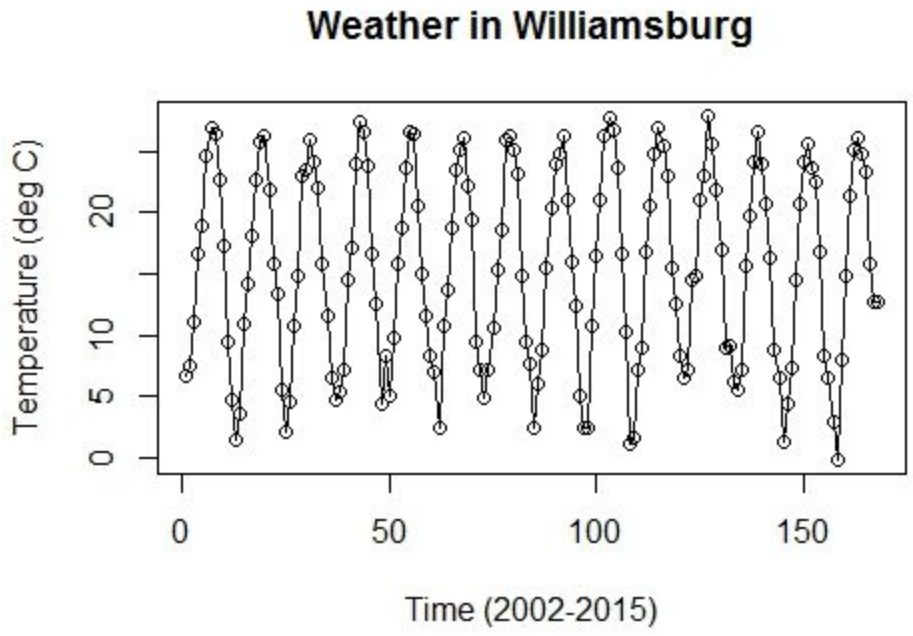
We collected our data from the Williamsburg 2 N weather station, (37.30N, 76.7W) ID:425004491510, from Williamsburg, Virginia. We found the dataset on NASA's website and we have included a link to the dataset in the reference page of our report  The dataset contains the monthly average surface temperatures measured at the station (as well as three month averages) from January 1901-February 2017 in degrees Celsius.

Unfortunately, the data had a separate column for each month and stored missing values as 999.9 degrees Celsius (which never happens in real life). To clean our dataset we selected only the columns containing the monthly average temperature (dropping the year variable and the three month average temperature variable) and took the transpose of those values so that each column would contain a year's worth of observations so that we could convert the dataframe into a vector whose first observation is Williamsburg's average surface temperature in January, 1901 and whose last observation is Williamsburg's average surface temperature in February, 2017 via the as.vector() function in R. We then constructed a new Year variable that spanned the dataset using the rep function to repeat a span of 1901:2017 12 times and then using the order() function to make sure the years were in ascending order and appended it to our vector containing our monthly average temperatures. We then used a similar tactic to create a month variable that spanned our data and appended it to our new dataframe. After discussing a number of possibilities on how to deal with missing data values, we decided to subset the most recent, large section of the data that contained no NA's. We settled for 2002-2016 which contains no missing values  and eclipses the most recent full year of data recorded. We wanted recent data for two reasons; it would be easier to make future predictions with recent data and we were taught that the technology used to record weather data has changed over the years (and has remained fairly constant since the 1970's), so we decided to choose a range of data that hopefully would not contain a major change in the way our data was recorded.

The specific research question we will try to answer is "Can we predict the average monthly temperature in Williamsburg, VA?"

# Decomposition

**Figure 7**




After considering the time series plot of our data, we believe that it did not contain a trend as it's mean appears to remain constant but the dataset does exhibit additive seasonality. Taking this into account we decided to fit three different types of models, each of which were created with and without a time term to be sure that the data did not contain a high order polynomial trend. The first models we created were two-parameter harmonic regression models, one with a time feature and one without a time feature. As we suspected, the coefficient for our time variable was insignificant. We then created four-parameter harmonic regression models but did not include them in this report because none of the extra coefficients were significant for either model. After that we created our seasonal adjustment models with L =12. We initially chose L to be 12 because our data is recorded over 12 different months and did not think that 4 months (one for summer fall winter and spring) would capture the variation in the data because each yearly peak and trough spanned roughly 5-6 datapoints. Again, as we suspected, the coefficient for our time variable in our seasonal adjustment model that contained time as a feature was insignificant. Interestingly, all months in both seasonal average models were significant except for February (which was denoted as M2 in our code.) The model summaries of our decomposition models can be found in our appendix listed under **figures 9-12.** We will show the model validation plot for our best decomposition model under our comparison section (listed as **figure 18)**

# SARIMA

The time series plot in **Figure 7** shows a section of the original data (**Figure 13**) that does not have any missing values and is the most recent (Background Section). The data itself

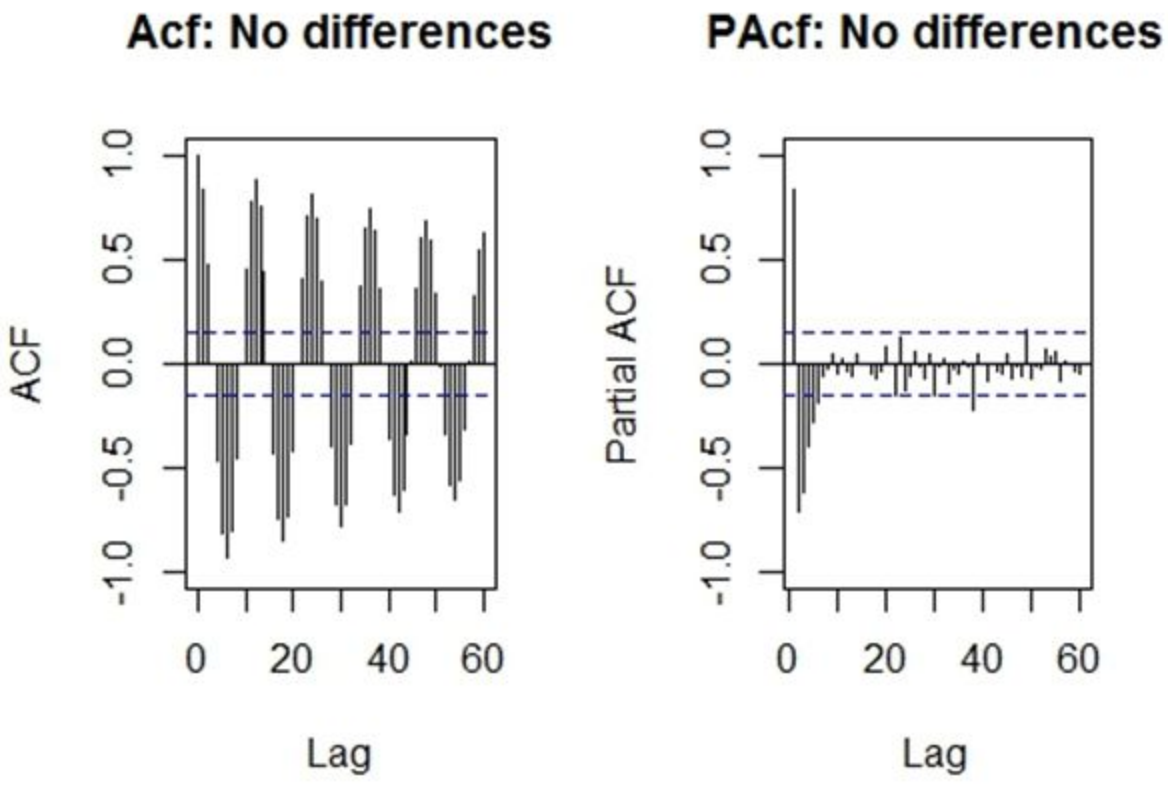
already shows constant mean and constant variance so we can proceed with looking at the Acf/PAcf plots.

There were two approaches in determining a suitable SARIMA model: the first approach used the method of differences with interpreting the Acf/PAcf graphs, the second approach primarily used brute force. Then, the models' predictive and descriptive capabilities were compared (in Comparison Section) by using the MSPE and BIC metrics.

## A First Approach: Informative

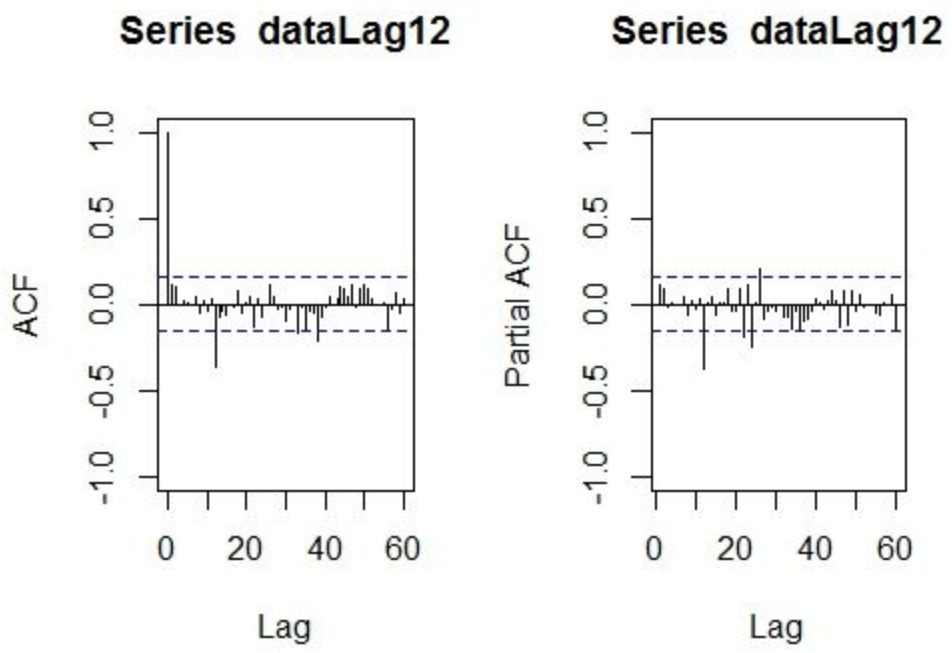
The Acf and PAcf indicate that a SARIMA model must be used. The non-seasonal parts look to be both decaying. Therefore, the nonseasonal parts will be tested as ARIMA(1,0,1), ARIMA(1,0,2). ARIMA(2,0,1) and ARIMA(2,0,1).

**Figure 14**




Now, the seasonal differences were taken to obtain the following Acf/PAcf plots:

**Figure 8**




The differenced Acf and PAcf plots show significance at a P=1 (Seasonal AR part) and Q= 2 (Seasonal MA part). Also, the model looked stationary after taking the differences shown in **Figure 6** (show in the Appendix Section) so D=0. Therefore the proposed SARIMA models are as follows: SARIMA(1,0,1)(1,0,2)12, SARIMA(1,0,2)(1,0,2)12, SARIMA(2,0,1)(1,0,2)12, SARIMA(2,0,2)(1,0,2)12. Eventually, the coefficient test showed that SARIMA(1,0,2)(1,0,2)12 had the most significant parameters and had residuals displaying white noise shown in **Figure 2** and F**igure 3.**

**Figure 1**

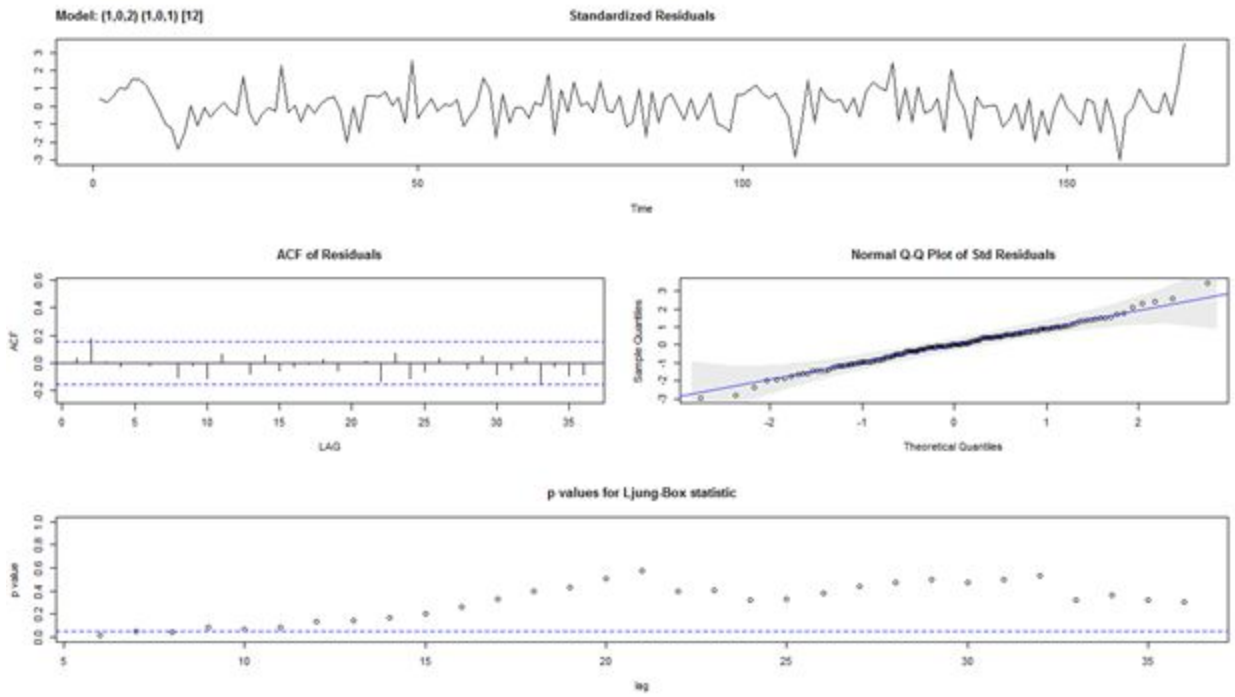
```
> coeftest(modelSarimaRework2$fit) # test the significance of the model

z test of coefficients:

        Estimate  Std. Error   z value  Pr(>|z|)
ar1   0.99941134  0.00082317 1214.0989 < 2.2e-16 ***
ma1  -0.81162952  0.06977272  -11.6325 < 2.2e-16 ***
ma2  -0.18391874  0.06964871   -2.6407  0.008274 **
sar1  0.99995627  0.00019486 5131.5894 < 2.2e-16 ***
sma1 -0.90108624  0.11817688   -7.6249 2.442e-14 ***
sma2 -0.06032907  0.09741534   -0.6193  0.535720
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Figure 3**

```
> coeftest(modelSarimaRework2$fit) # test the significance of the r

z test of coefficients:

       Estimate  Std. Error   z value  Pr(>|z|)
ar1   0.99940808  0.00083709 1193.9068 < 2.2e-16 ***
ma1  -0.81042747  0.06958282  -11.6469 < 2.2e-16 ***
ma2  -0.18511114  0.06945803   -2.6651  0.007697 **
sar1  0.99991997  0.00037633 2657.0045 < 2.2e-16 ***
sma1 -0.94793509  0.11989423   -7.9064 2.649e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
```

**Figure 1** shows that all the parameters are significant except the seasonal MA part at lag 24. the model should be further reduced to SARIMA(1,0,2)(1,0,1)12 shown in **Figure 2** and **Figure 3**, whereby **Figure 3** show white noise.

**Figure 2**



# A Second Approach: Brute Force

A second approach to finding a SARIMA model was through brute force eventually resulting in a SARIMA model SARIMA(1,0,0)(0,1,1)12. **Figure 4** shows the residuals that exhibit white noise and **figure 5** shows the coefficients.

**Figure 4**

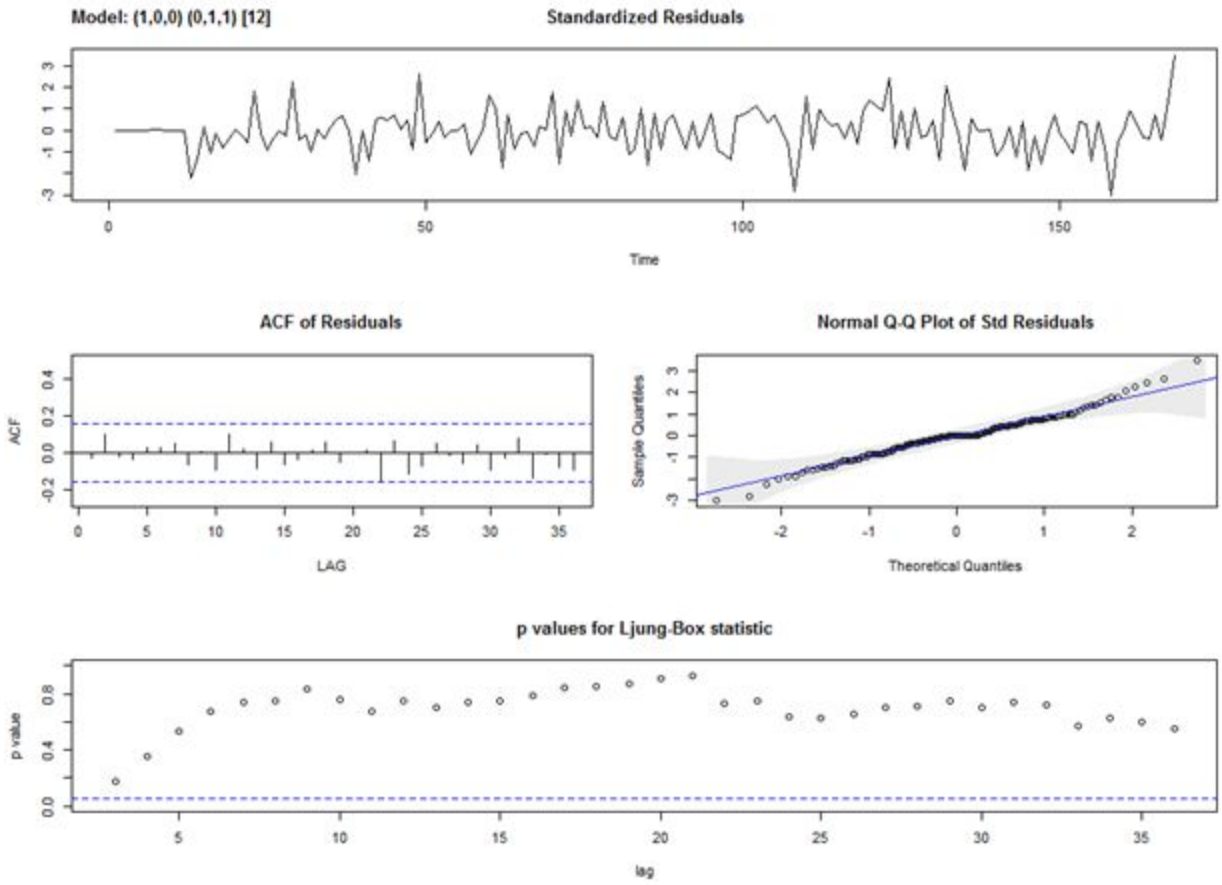



**Figure 5**

```
> coeftest(modelSarimaSean$fit) # test the significance of the model

z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
ar1   0.172221   0.082894  2.0776   0.03775 *
sma1 -0.928612   0.144766 -6.4146 1.412e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This model eventually resulted in the lowest BIC out of all the models with residuals that satisfied the white noise assumptions.

# Comparisons

Figure 15

| Model | BIC/AIC | MSPE (Mean Square Prediction Error) |
|---|---|---|
| SARIMA (1,0,0)(0,1,1)12 | 654.5588 | 2.441992 |
| SARIMA(1,0,2)(1,0,1)12 | 675.8393 | 2.168959 |
| Decomposition: 2 parameter Harmonic Regression NO TREND | 658.6687 | 2.643479 |
| Decomposition: 2 Parameter harmonic regression TIME | 663.8039 | 2.608179 |
| Decomposition: Seasonal Average + time feature | 705.2272 | 2.192646 |
| Decomposition: Seasonal Average NO TIME | 700.1069 | 2.21267 |
| ARIMA(1,0,0) | 701.891 | 10.84727 |
| ARIMA(1,0,1) | 705.835 | 10.11572 |
| ARIMA(1,0,2) | 710.8849 | 3.410056 |
| ARIMA(2,0,1) | 710.3143 | 3.220333 |
| ARIMA(2,0,2) | 715.3606 | 10.16177 |
| ARMA(0,1) => MA(1) | 705.3504 | 10.94372 |
| ARMA(0,2) => MA(2) | 702.3125 | 10.28911 |

**Comparison of decomposition models**

The BIC of our harmonic regression model that did not contain time ( 658.6687) was noticeably lower than the BIC of our harmonic regression model that did include time 663.8039 and there was only a slight difference in their MSPE's, 2.608179 vs. 2.643479, in favor of the model that included time. This further affirmed our belief that our dataset did not contain a trend. As expected, the BIC of our time-less model (700.1069) was noticeably lower than the BIC of the model that included time (705.2272) and their difference in MSPE is negligible (2.192646 v. 2.21267 in favor of the model that included time.) Taking all of this into consideration, we decided that our two parameter harmonic regression (without a time feature) was the best decomposition model we tested because it had the best BIC and had a competitive MSPE. It is

important to note that since we are measuring surface temperature in degrees celsius, we must take the square root of our MSPE for it to make sense because we do not think in degrees Celsius squared. Although the difference in our harmonic regression model that contained a time feature and our chosen model was approximately .5 degrees celsius squared, the difference in their mean prediction errors is approximately .14 degrees celsius whereas the difference in their BIC's is 46.5472 units. **Figure 18** below shows the residuals v. fitted value plot for our final decomposition model. The points in the graph do not appear to form a pattern so we believe that the linearity assumption is validated for our two parameter harmonic regression model. Also, harmonic regression without a linear trend will have the same fitted values for each month, thus the verticality of the residuals are not important. Since the points in **figure 18** appear to move somewhat closer together from left to right, there is some concern that the constant variance assumption of our model is violated, but since it is not a drastic decline in variance, we feel that the assumption is still satisfied for our model. **Figure 19** below shows a graph of our 2 parameter harmonic regression model's predicted points and its fitted values. We were impressed with both of them. The plot also includes the prediction interval for our predicted points. We were pleased to find that it was a rather tight interval and matches the seasonal pattern in the data well.

Figure 18

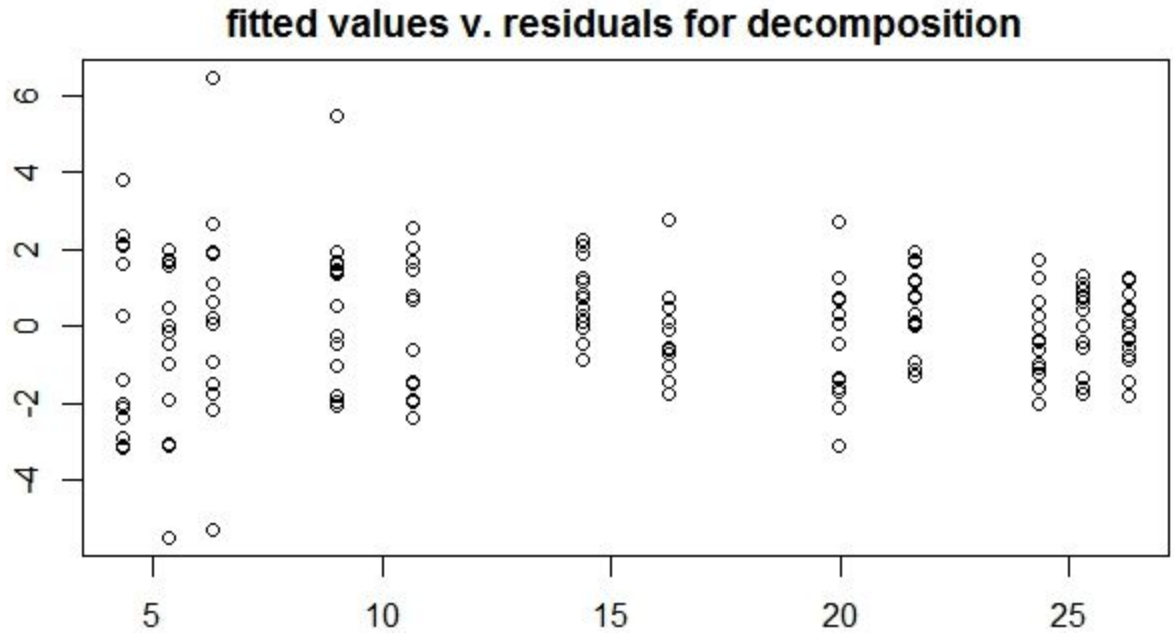
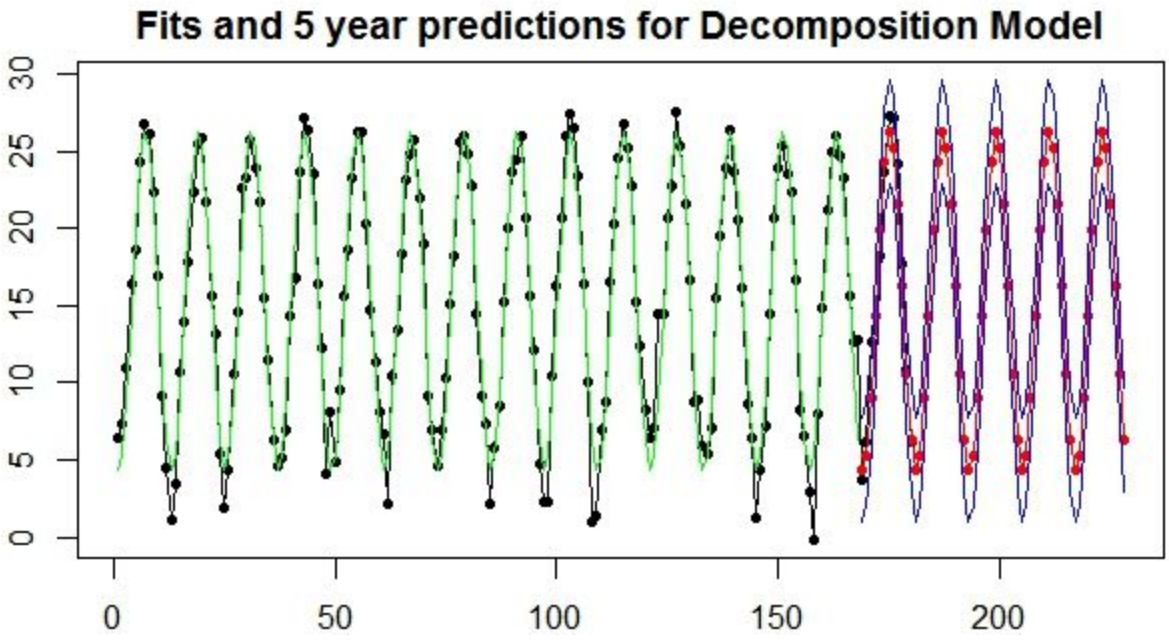



Figure 19




**Comparison of SARIMA models**

Although it was derived by brute force, we deemed our SARIMA(1,0,0)(0,1,1)12 model as our best SARIMA model because it fit all of the white noise assumptions, had a better BIC than the SARIMA (1,0,2)(1,0,1)12 model (654.5588 v. 675.8393) and it had a reasonable MSPE (2.441992 v. 2.168959 in favor of the model we did not select.) Therefore, with a much better BIC and a close MSPE, the SARIMA(1,0,0)(0,1,1)12 model was better.

**Overall comparison**

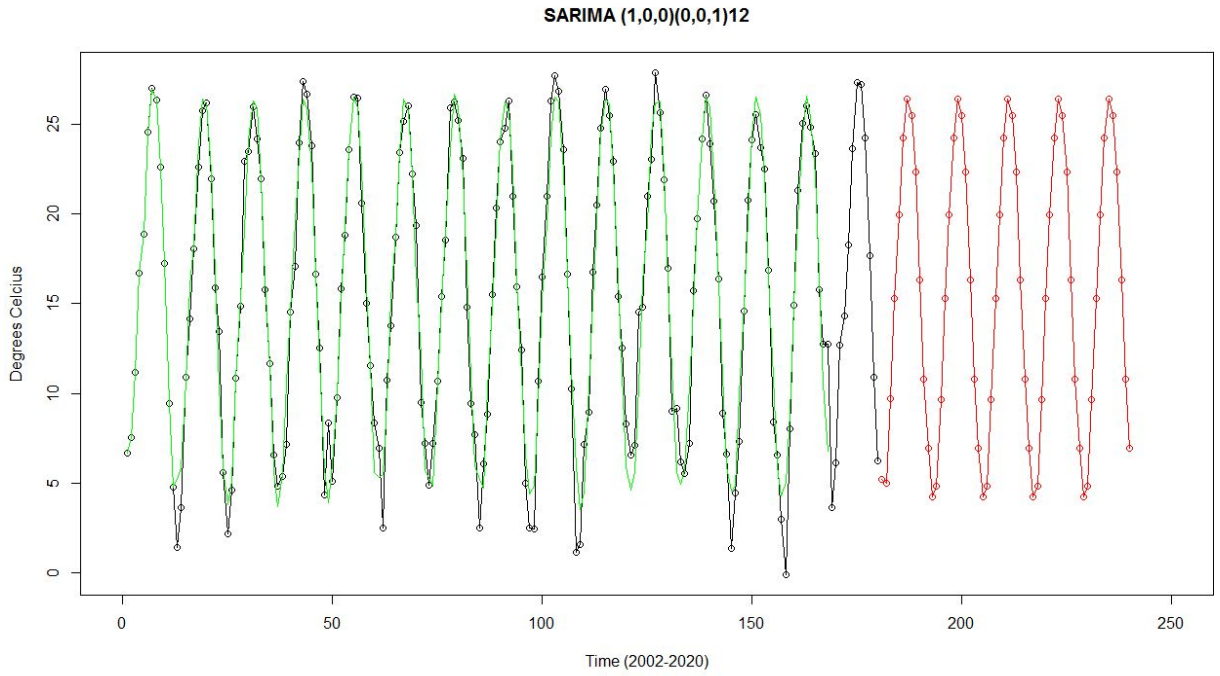
As we did previously, we used BIC and MSPE to determine our final model. Visually, both models fitted our data very well and had believable predicted values, so we had to resort to other measures to decide which model was the best. Fortunately, the decision was relatively easy because our final SARIMA model edged out our final decomposition model in both BIC (654.5588 to 658.6678) and MSPE (2.441992 to 2.643479.) So, based on BIC and MSPE, we believe that our SARIMA (1,0,1)(0,1,1)12 model best describes our data and will provide accurate predictions.

**Conclusions**

The SARIMA model found through brute force SARIMA(1,0,0)(0,1,1)12 had the lowest BIC and had a reasonable MSPE level. Therefore, this model suggests that the seasonal average temperatures, where seasons are months in this analysis, are slowly rising. With a fair MSPE, this model can be used to answer the original research question to predict the average temperatures within a couple of years. **Figure 16** shows the fitted plot with 5 years of

predictions and **figure 20** is the same plot without the fitted lines but has the prediction intervals (found in the Appendix Section)..

**Figure 15**



# References

https://data.giss.nasa.gov/cgi-bin/gistemp/stdata_show.cgi?id=425004491510&dt=1&ds=5

http://handbook.cochrane.org/chapter_16/16_1_2_general_principles_for_dealing_with_missing_data.htm

# Appendix

Figure 1

```
> coeftest(modelSarimaRework2$fit) # test the significance of the model

z test of coefficients:

        Estimate  Std. Error    z value  Pr(>|z|)
ar1   0.99941134  0.00082317 1214.0989 < 2.2e-16 ***
ma1  -0.81162952  0.06977272  -11.6325 < 2.2e-16 ***
ma2  -0.18391874  0.06964871   -2.6407  0.008274 **
sar1  0.99995627  0.00019486 5131.5894 < 2.2e-16 ***
sma1 -0.90108624  0.11817688   -7.6249 2.442e-14 ***
sma2 -0.06032907  0.09741534   -0.6193  0.535720
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2

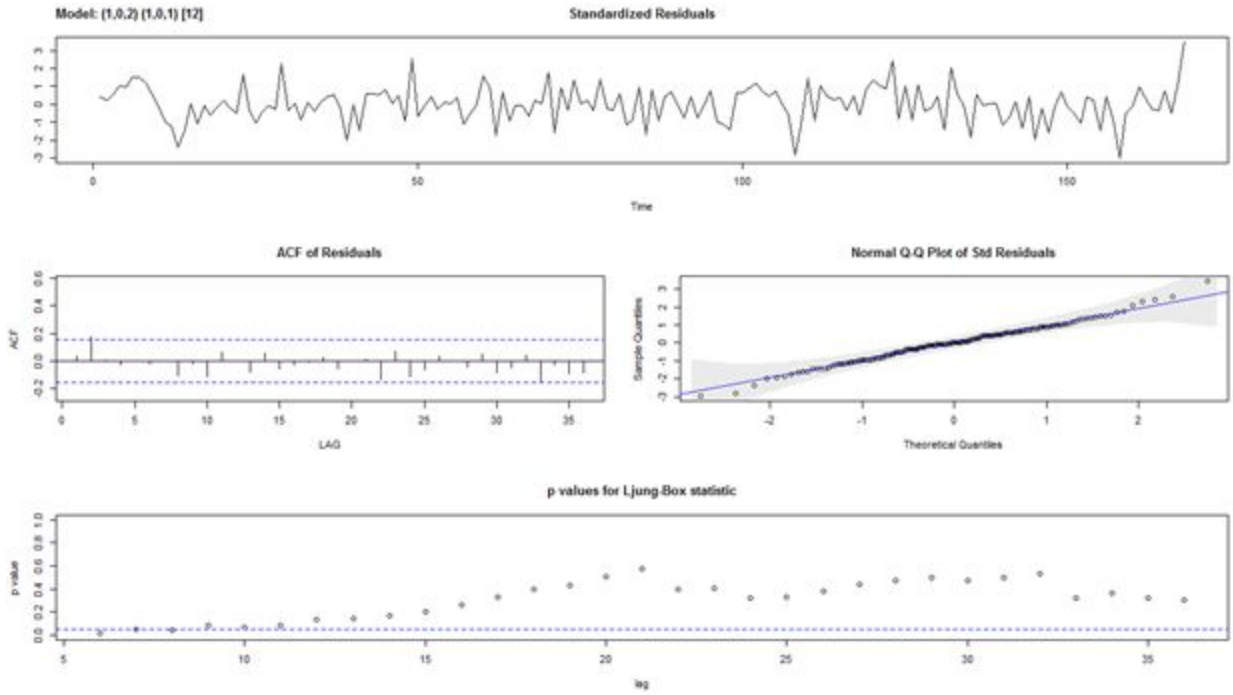



Figure 3

```
> coeftest(modelSarimaRework2$fit) # test the significance of the r

z test of coefficients:

        Estimate  Std. Error    z value  Pr(>|z|)
ar1   0.99940808  0.00083709 1193.9068 < 2.2e-16 ***
ma1  -0.81042747  0.06958282  -11.6469 < 2.2e-16 ***
ma2  -0.18511114  0.06945803   -2.6651  0.007697 **
sar1  0.99991997  0.00037633 2657.0045 < 2.2e-16 ***
sma1 -0.94793509  0.11989423   -7.9064 2.649e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
```

Figure 4

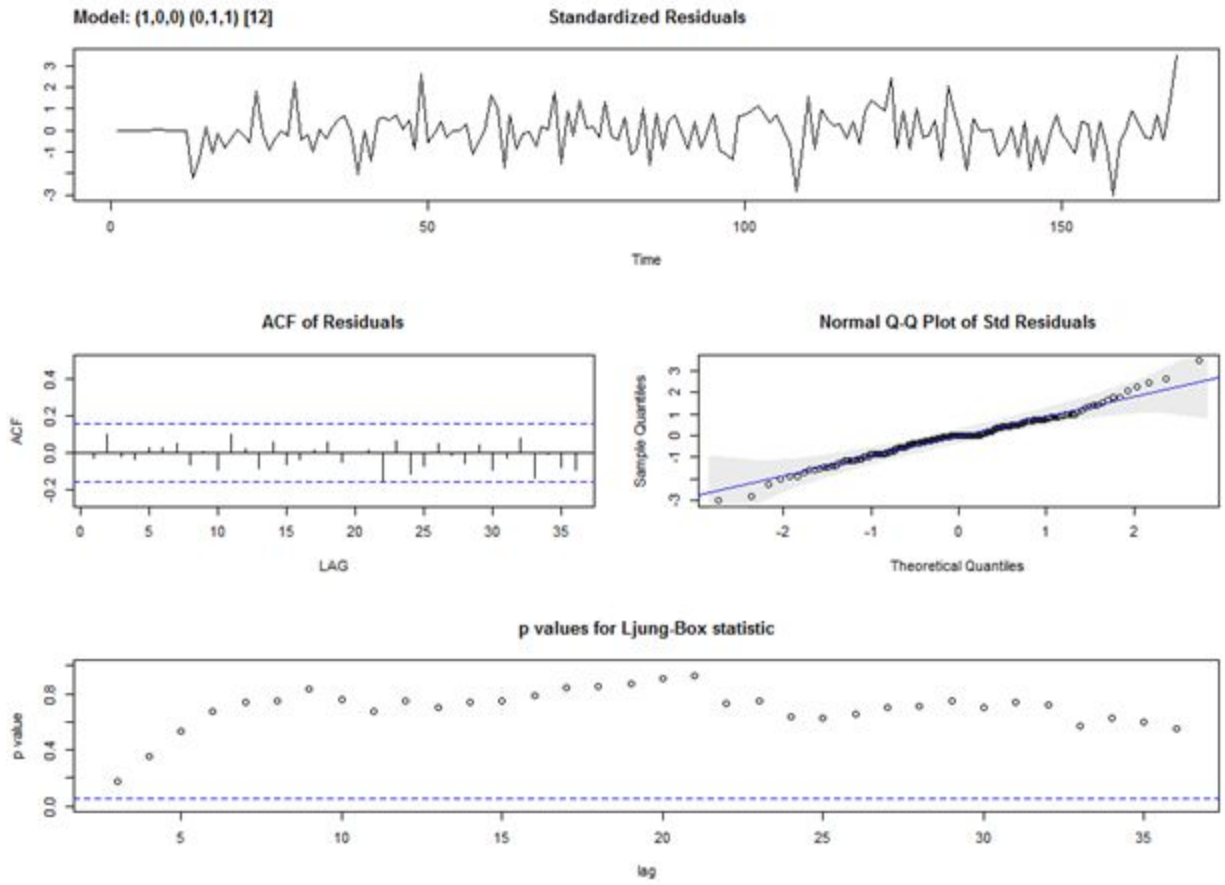



Figure 5

```
> coeftest(modelSarimaSean$fit) # test the significance of the model

z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1   0.172221   0.082894  2.0776   0.03775 *
sma1 -0.928612   0.144766 -6.4146 1.412e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6

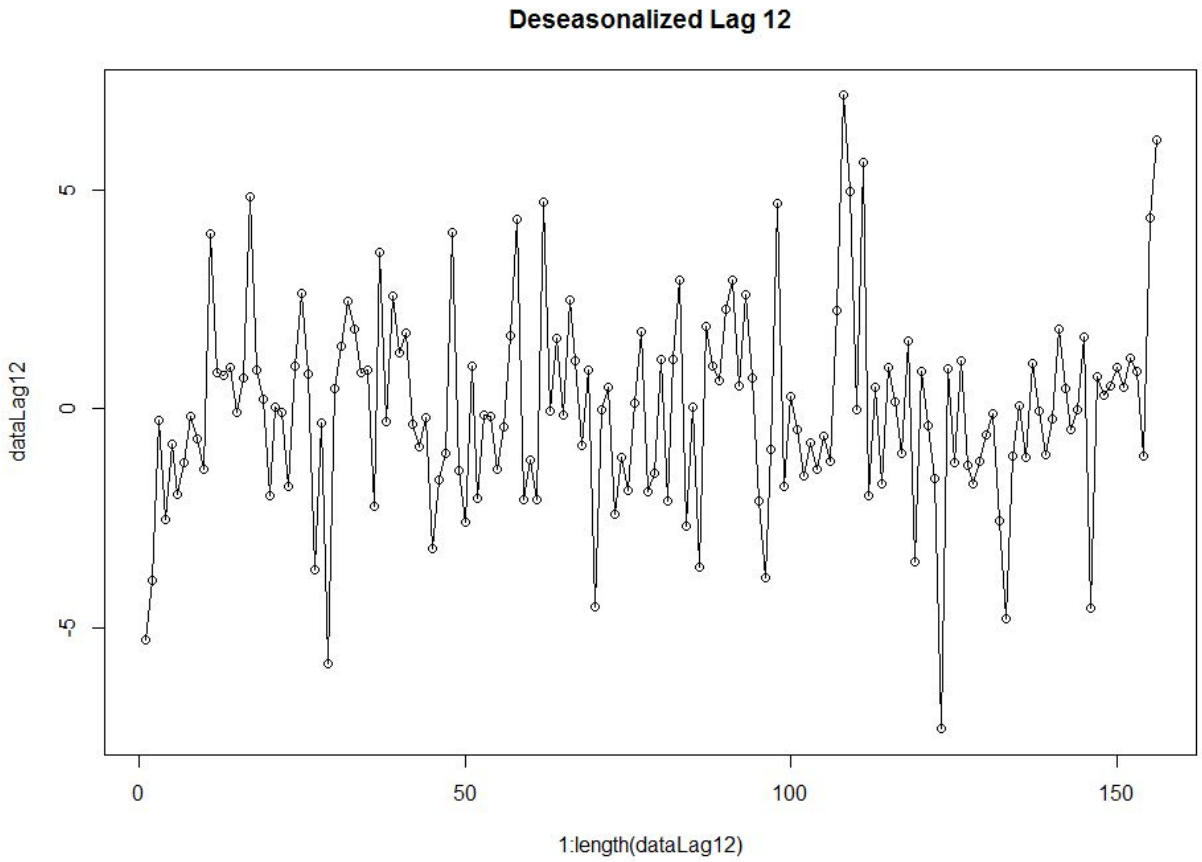



Figure 7

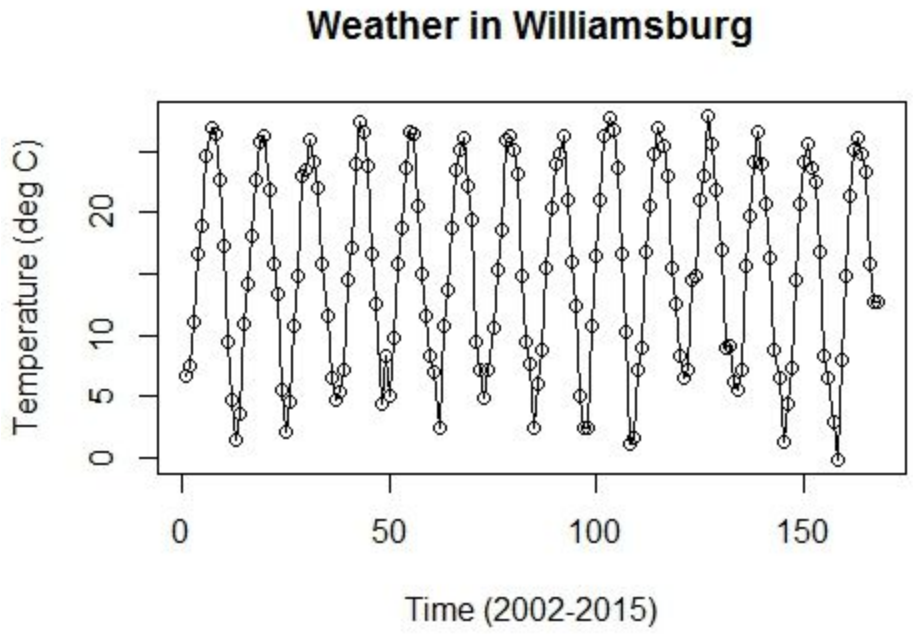

Weather in Williamsburg
Temperature (deg C)
0
5
10
20
0
50
100
150
Time (2002-2015)

Figure 8

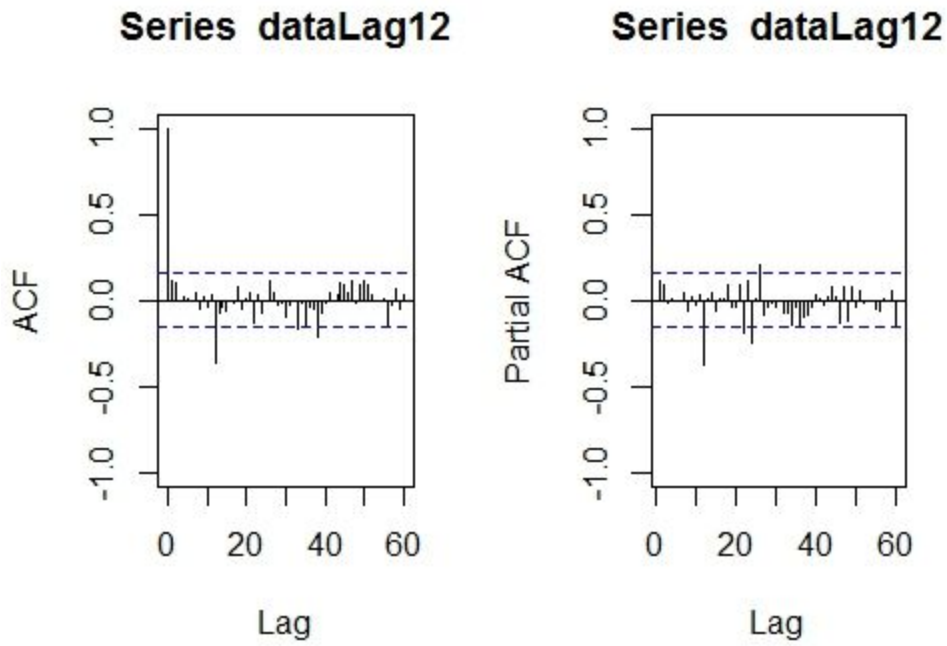

Series dataLag12
ACF
1.0
0.5
0.0
-0.5
-1.0
0
20
40
60
Lag
Series dataLag12
Partial ACF
1.0
0.5
0.0
-0.5
-1.0
0
20
40
60
Lag

Figure 9

```
Call:
lm(formula = trainnumz ~ time + I(sin(2 * pi * time/L)) + I(cos(2 *
    pi * time/L)), data = train.set)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5230 -1.0950  0.0774  1.1385  6.4339

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             15.2917837  0.2617844  58.414   <2e-16 ***
time                     0.0002825  0.0026880   0.105    0.916
I(sin(2 * pi * time/L)) -6.3182744  0.1843367 -34.276   <2e-16 ***
I(cos(2 * pi * time/L)) -9.0431695  0.1840831 -49.125   <2e-16 ***
```

Figure 10

```
Call:
lm(formula = trainnumz ~ I(sin(2 * pi * time/L)) + I(cos(2 *
    pi * time/L)), data = train.set)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5015 -1.1027  0.0792  1.1430  6.4572

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              15.3157     0.1298  118.03   <2e-16 ***
I(sin(2 * pi * time/L))  -6.3193     0.1835  -34.44   <2e-16 ***
I(cos(2 * pi * time/L))  -9.0429     0.1835  -49.28   <2e-16 ***
```

Figure 11

```
Call:
lm(formula = trainnumz ~ time + Month, data = train.set)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5478 -1.1360 -0.0261  0.9855  6.1927

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.985e+00  4.980e-01   8.003 2.66e-13 ***
time        1.577e-04  2.688e-03   0.059 0.953303
MonthM10    1.212e+01  6.375e-01  19.007  < 2e-16 ***
MonthM11    6.621e+00  6.376e-01  10.385  < 2e-16 ***
MonthM12    2.525e+00  6.377e-01   3.960 0.000114 ***
MonthM2     6.977e-01  6.370e-01   1.095 0.275095
MonthM3     5.560e+00  6.370e-01   8.728 3.87e-15 ***
MonthM4     1.107e+01  6.371e-01  17.369  < 2e-16 ***
MonthM5     1.561e+01  6.371e-01  24.503  < 2e-16 ***
MonthM6     2.000e+01  6.371e-01  31.387  < 2e-16 ***
MonthM7     2.218e+01  6.372e-01  34.808  < 2e-16 ***
MonthM8     2.133e+01  6.373e-01  33.476  < 2e-16 ***
MonthM9     1.810e+01  6.374e-01  28.391  < 2e-16 ***
```

Figure 12

```
Call:
lm(formula = trainnumz ~ Month, data = train.set)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5450 -1.1464 -0.0157  0.9920  6.2050

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9979     0.4490   8.904 1.30e-15 ***
MonthM10     12.1179     0.6350  19.084  < 2e-16 ***
MonthM11      6.6229     0.6350  10.430  < 2e-16 ***
MonthM12      2.5271     0.6350   3.980 0.000105 ***
MonthM2       0.6979     0.6350   1.099 0.273440
MonthM3       5.5600     0.6350   8.756 3.14e-15 ***
MonthM4      11.0657     0.6350  17.427  < 2e-16 ***
MonthM5      15.6114     0.6350  24.586  < 2e-16 ***
MonthM6      19.9986     0.6350  31.496  < 2e-16 ***
MonthM7      22.1807     0.6350  34.932  < 2e-16 ***
MonthM8      21.3350     0.6350  33.600  < 2e-16 ***
MonthM9      18.0964     0.6350  28.500  < 2e-16 ***
```

Figure 13

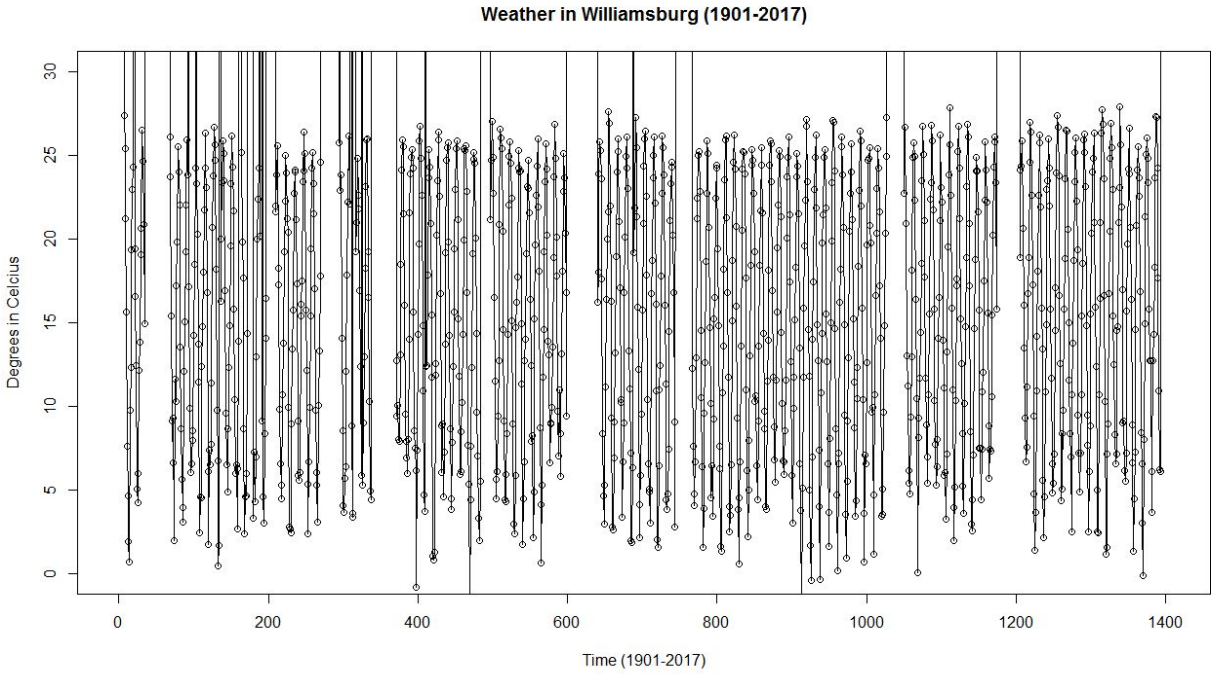

Weather in Williamsburg (1901-2017)
Degrees in Celcius
0
5
10
15
20
25
30
0
200
400
600
800
1000
1200
1400
Time (1901-2017)

Figure 14

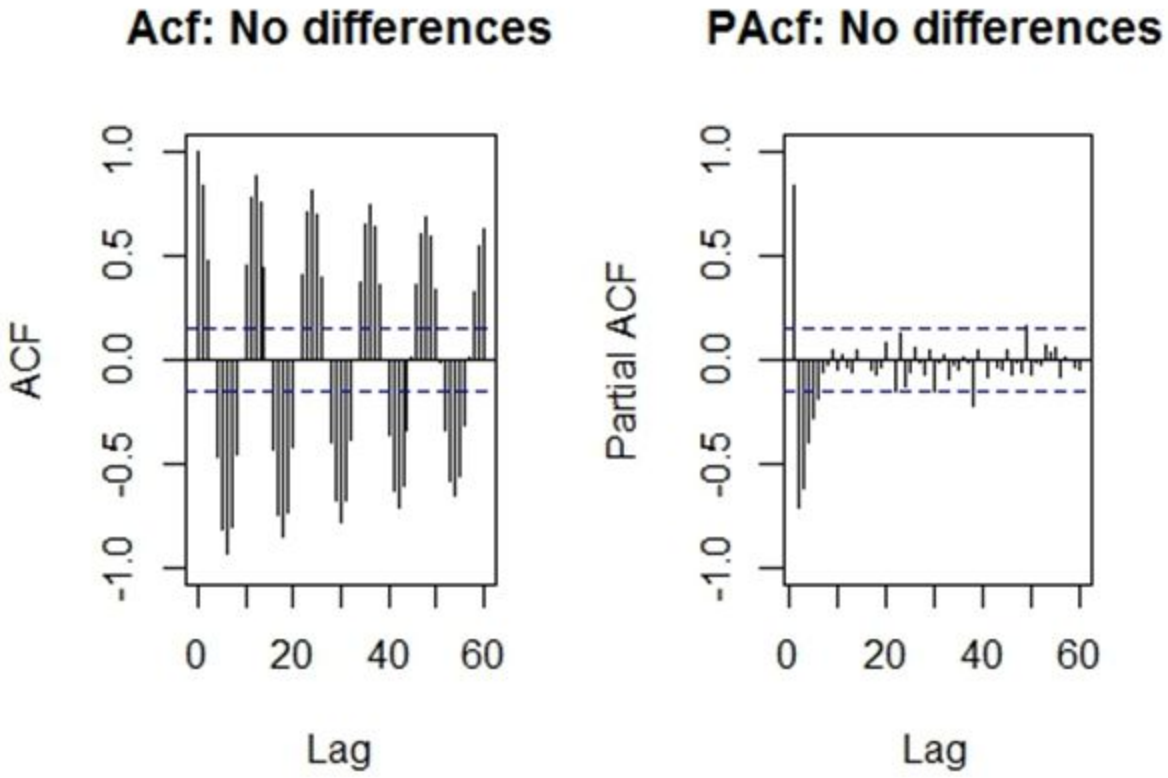



Figure 15

| Model | BIC/AIC | MSPE (Mean Square Prediction Error) |
|---|---|---|
| SARIMA (1,0,0)(0,1,1)12 | 654.5588 | 2.441992 |
| SARIMA(1,0,2)(1,0,1)12 | 675.8393 | 2.168959 |
| Decomposition: 2 parameter Harmonic Regression NO TREND | 658.6687 | 2.643479 |
| Decomposition: 2 Parameter harmonic regression TIME | 663.8039 | 2.608179 |
| Decomposition: Seasonal Average + time feature | 705.2272 | 2.192646 |
| Decomposition: Seasonal Average NO TIME | 700.1069 | 2.21267 |
| ARIMA(1,0,0) | 701.891 | 10.84727 |
| ARIMA(1,0,1) | 705.835 | 10.11572 |
| ARIMA(1,0,2) | 710.8849 | 3.410056 |
| ARIMA(2,0,1) | 710.3143 | 3.220333 |
| ARIMA(2,0,2) | 715.3606 | 10.16177 |
| ARMA(0,1) => MA(1) | 705.3504 | 10.94372 |
| ARMA(0,2) => MA(2) | 702.3125 | 10.28911 |

Figure 16:

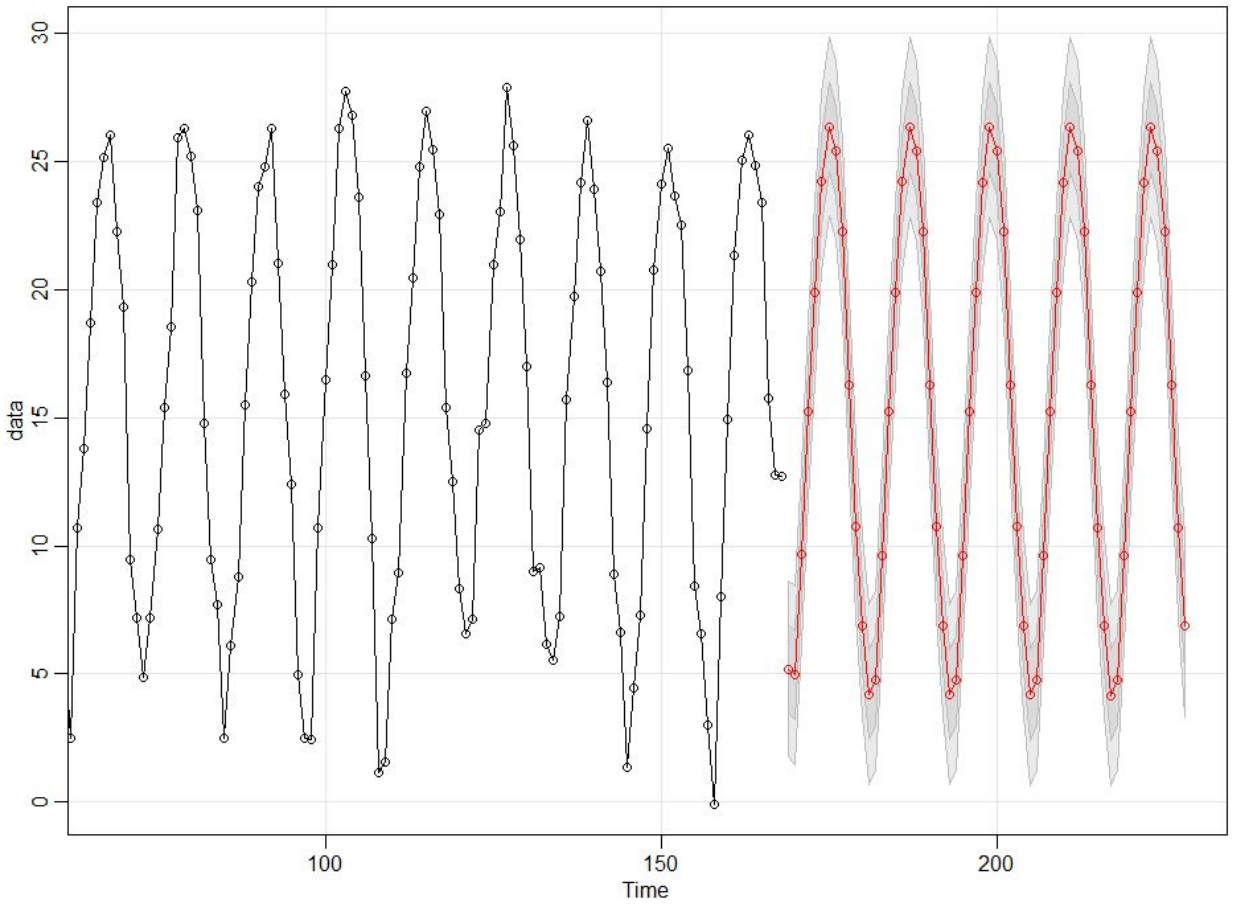



Figure 17

```
Console ~/Virginia Tech/STAT 4534 Time Series/
> predict(modelSarimaRework2, 12)
Error in ets(object, lambda = lambda, allow.multiplicative.trend = allow.multiplicative.tren
d,  :
  y should be a univariate time series
> predict(modelSarimaRework2.reduced, 12)
Error in ets(object, lambda = lambda, allow.multiplicative.trend = allow.multiplicative.tren
d,  :
  y should be a univariate time series
> BICSarimaRework2.reduced <- n*log(sum(resid(modelSarimaRework2.reduced$fit)^2/n)) + 4*log(
n) + n + n*log(2*pi)
Error: object 'n' not found
> n <- length(data)
> BICSarimaRework2.reduced <- n*log(sum(resid(modelSarimaRework2.reduced$fit)^2/n)) + 4*log(
n) + n + n*log(2*pi)
> BICSarimaRework2.reduced
[1] 675.8393
> predict(modelSarimaRework2.reduced$fit, 12)
Error in eval(expr, envir, enclos) : object 'xmean' not found
> predict(modelSarimaRework2.reduced$fit, 12)
Error in eval(expr, envir, enclos) : object 'xmean' not found
>
```

Figure 18

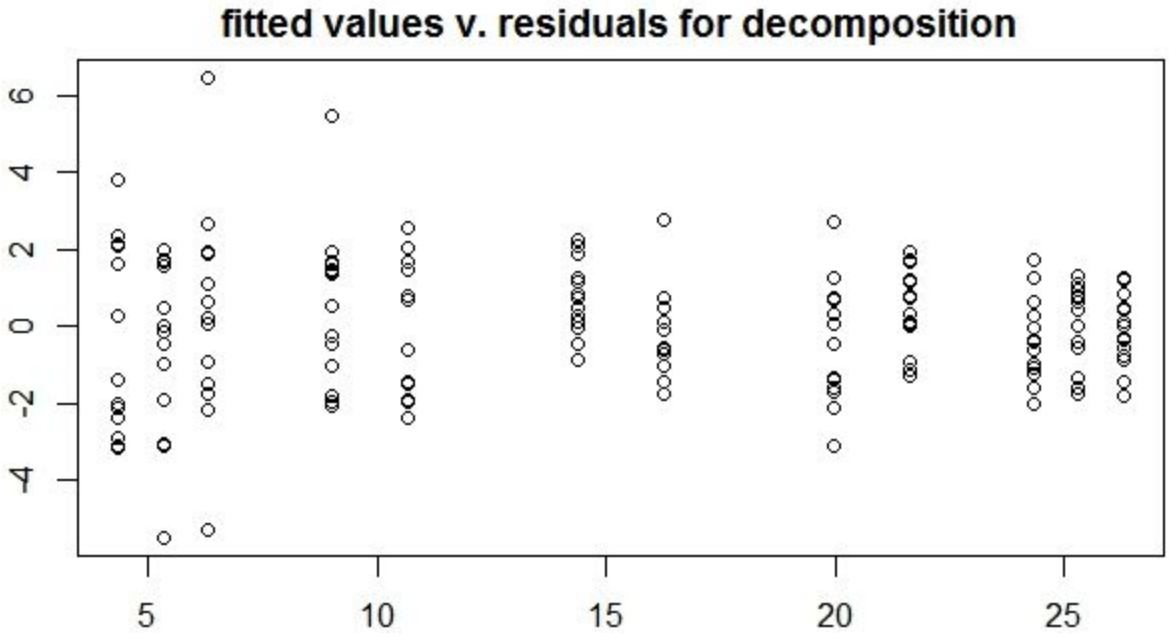



Figure 19

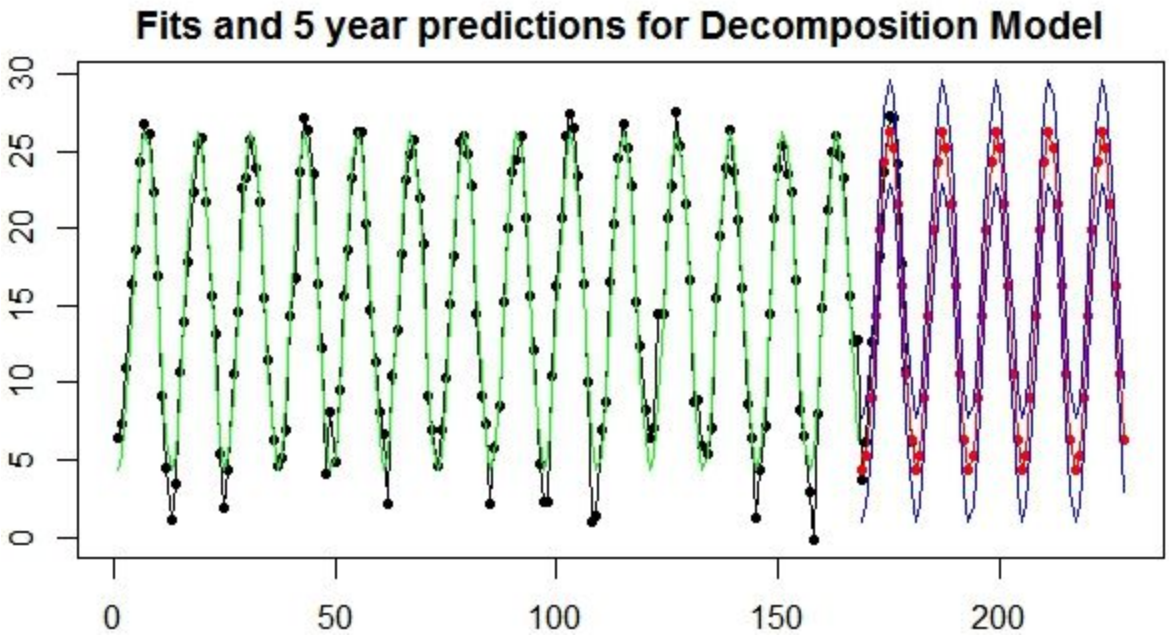
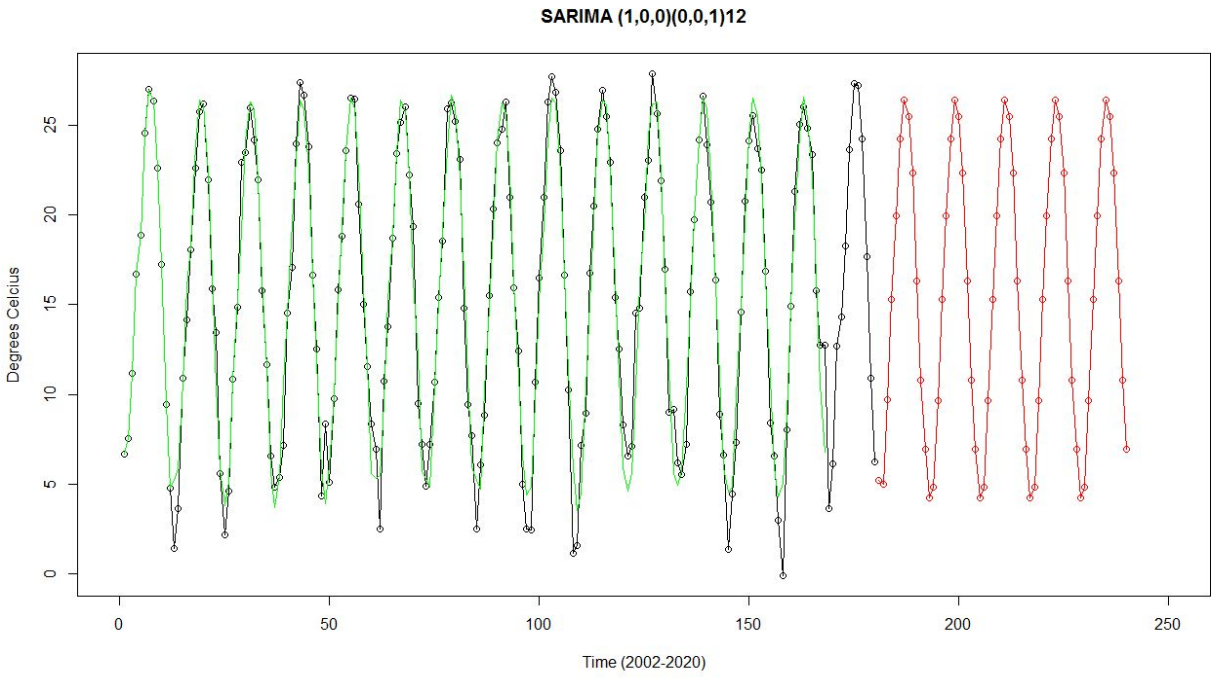

Fits and 5 year predictions for Decomposition Model
0
5
10
15
20
25
30
0
50
100
150
200

Figure 20


SARIMA (1,0,0)(0,0,1)12
Degrees Celcius
0
5
10
15
20
25
0
50
100
150
200
250
Time (2002-2020)