# Time Series Project

# Hungarian Chickenpox Cases 2005-2014

Andrea Marchi

Ben Taczy

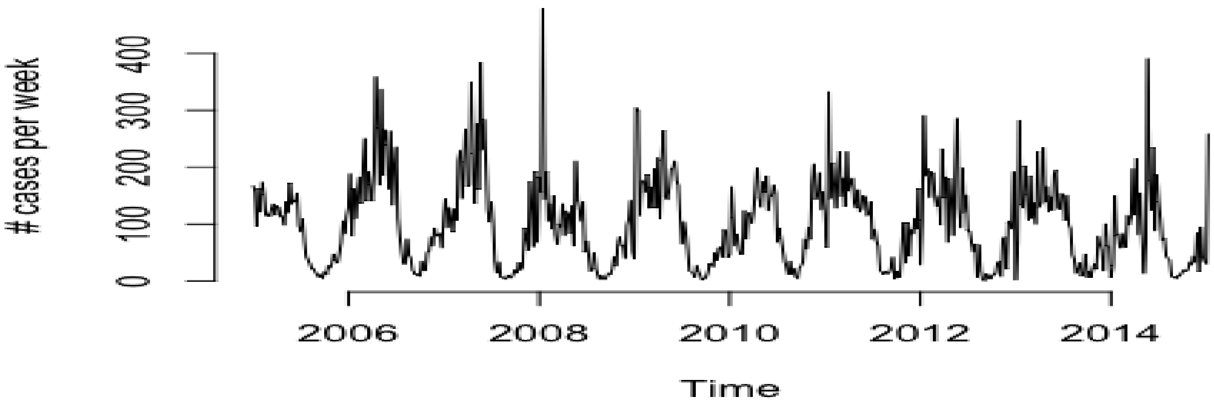
Dustin Sparks

# Table of Contents

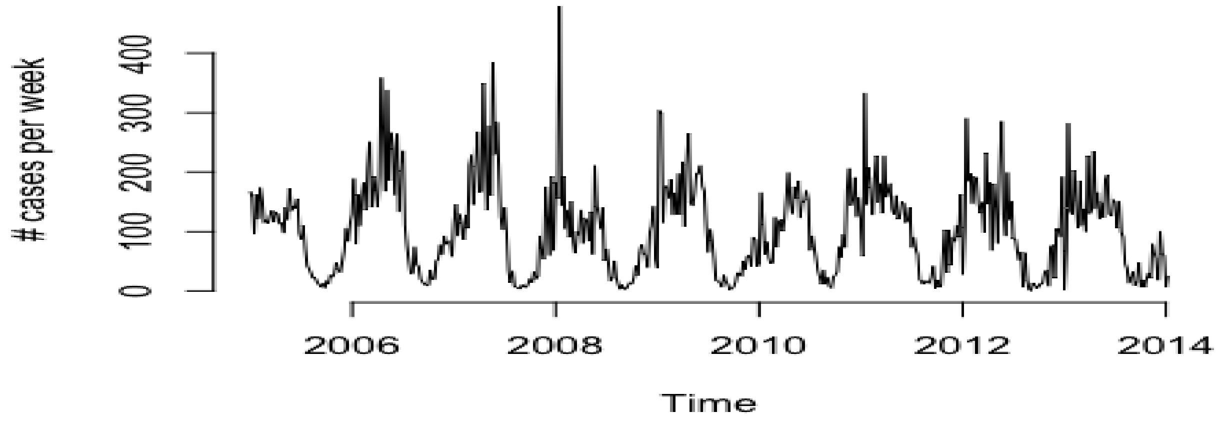
## Dataset Background

These data were found on the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Hungarian+Chickenpox+Cases#) and originally sourced from Benedek Rozemberczki at the University of Edinburgh. This data is a time series data set for chickenpox cases including most cities Hungary and we elected to focus on the city of Budapest. Budapest was selected as our variable of focus since it is the largest city in Hungary. Data was collected weekly from 2005-2014, which implies that there are 52 data points per year, for a total of 520 points, and so we decided to build a model for future forecasting using the first 9 years of the collected data for a total of 468 data points spanning from 2005 to 2013. Then we used the last 52 data points of this time series to verify the accuracy of the model by comparison with actual collected data.

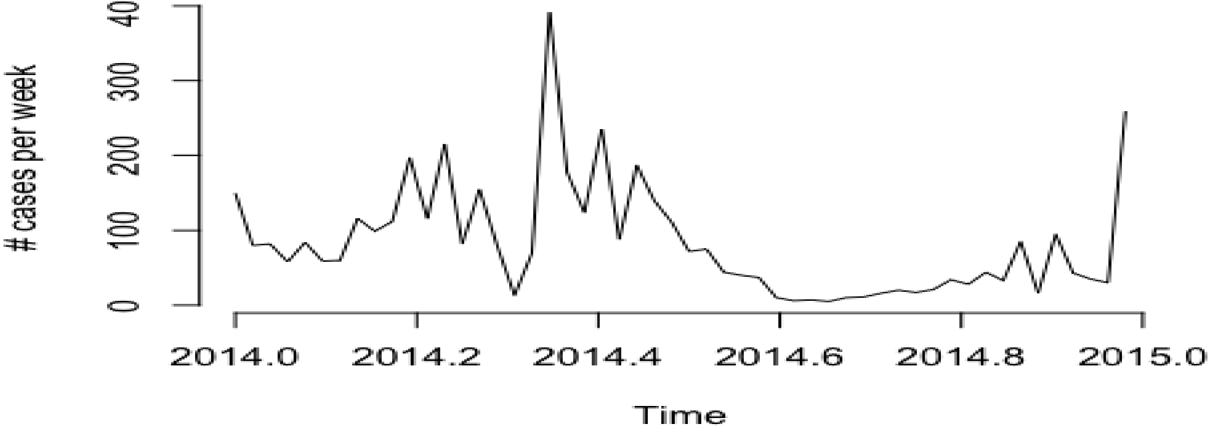
```
## t   DATE   #Cases
## 1 03/01/05 168
## 2 10/01/05 157
## 3 17/01/05  96
## 4 24/01/05 163
## 5 31/01/05 122
## 6 07/02/05 174
## .   .  .     .
## .   .  .     .
## .   .  .     .
## .   .  .     .
## .   .  .     .
```



Complete Time Series (2005-2014)



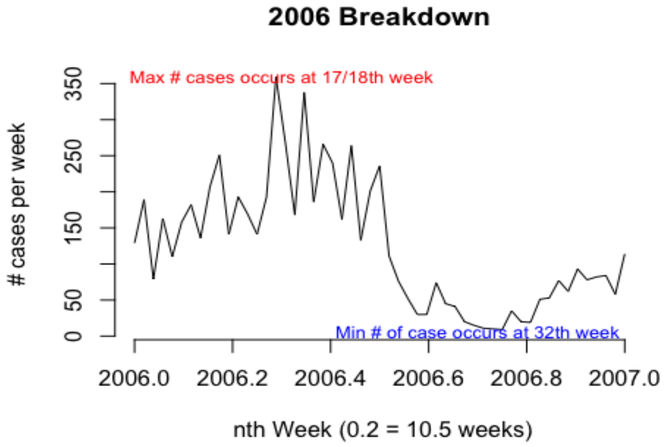
Train Time Series (2005-2013)

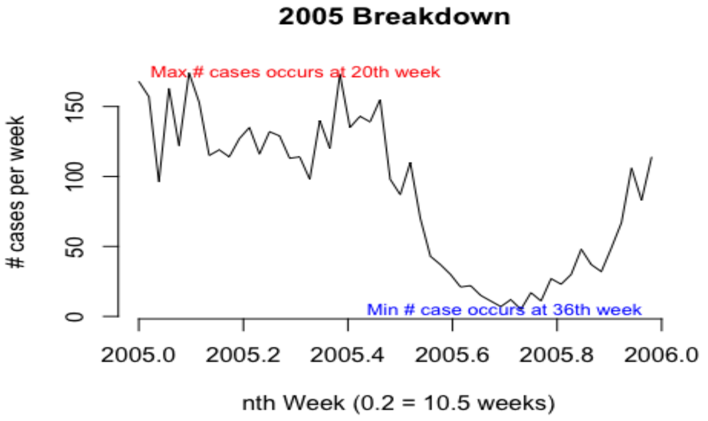


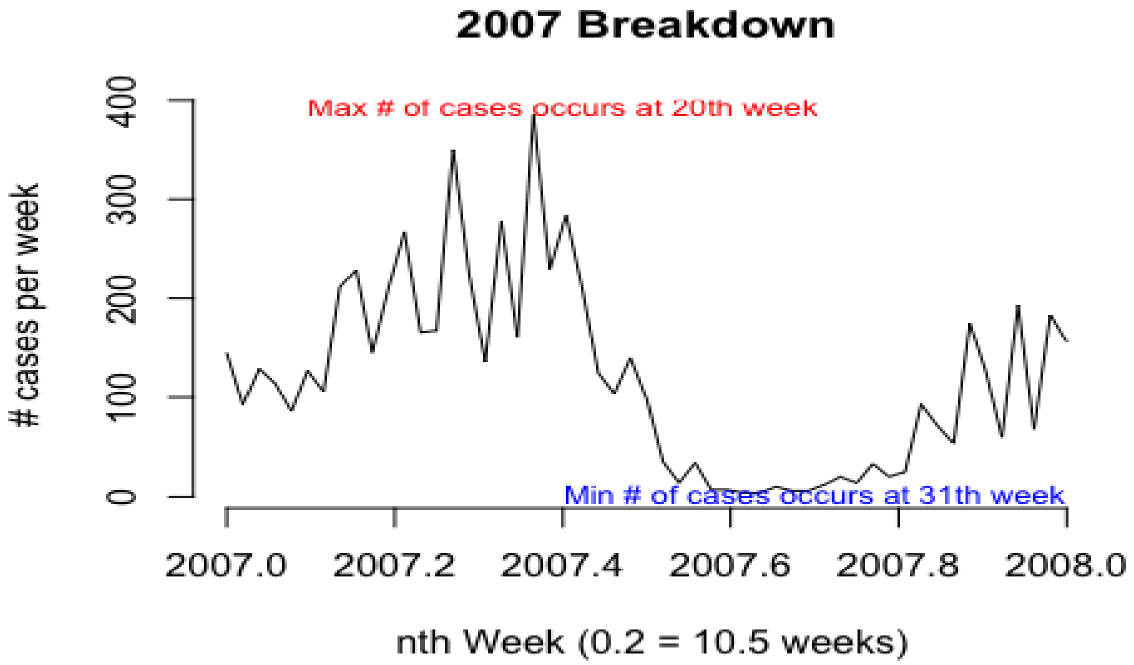
Test Time Series (2014)

## Observation from Complete Time-Series, Year-to-Year Breakdown and Decomposition:

From the plot of the complete time series (2005-2014), we can observe both a cyclical effect and a seasonal variation as well. However, we want to look at trends per year by considering and isolating the first three years to check for patterns and finally using decomposition of components.



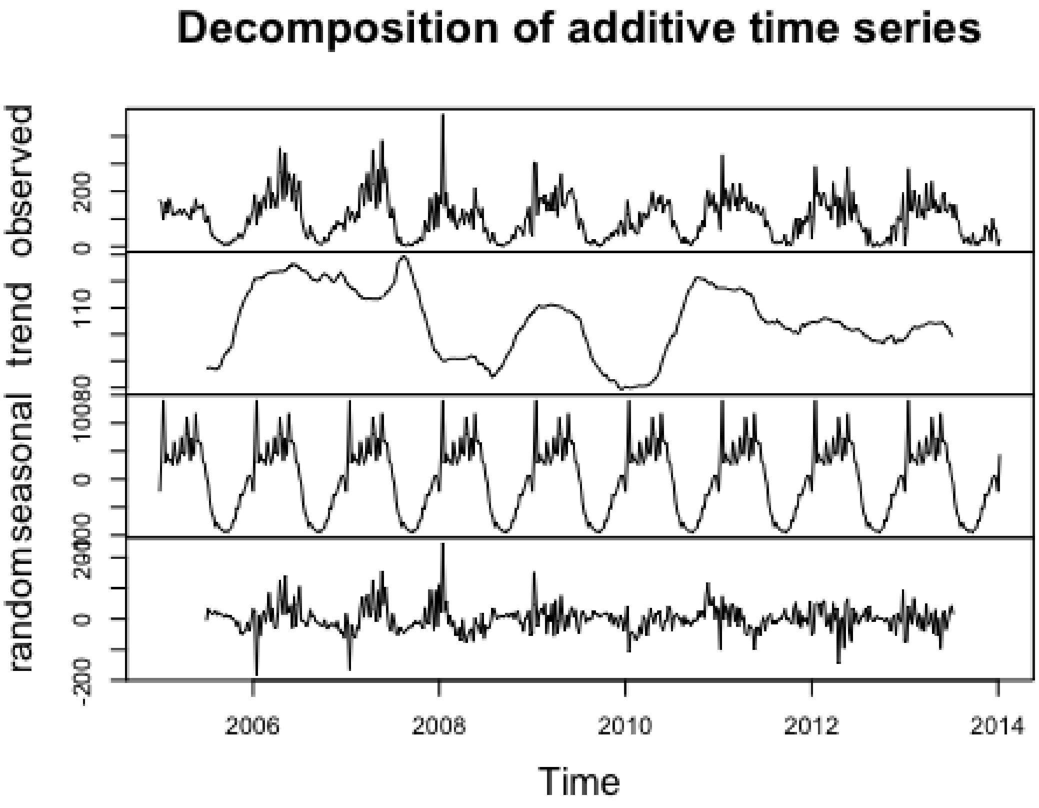




### Observations

We can notice that between years we have repeated patterns between years with surges of cases around the 20th week of the year (April), followed by a sudden negative rate of change where the number of cases hit the lowest point of the year around the 32nd week (late summer). We performed the entire analysis year by year, and we noticed that this behavior repeats for all the ten years without an apparent secular trend. Therefore, overall, the number of chickenpox cases in Budapest from 2005 to 2014 oscillates but remains constant. To confirm our thesis, we wanted to plot the individual components of the time series.

## Decomposition of Time Series



## Observations

The plot of the individual time series components confirms our suspects. We cannot say that there is a recognizable line in the trend graph, so we can infer that the secular trend effect is negligible. On the other hand, because we can observe repeated patterns in the seasonal component graph, we can conclude that there is indeed an evident seasonal variation affecting the number of chickenpox cases in Budapest.

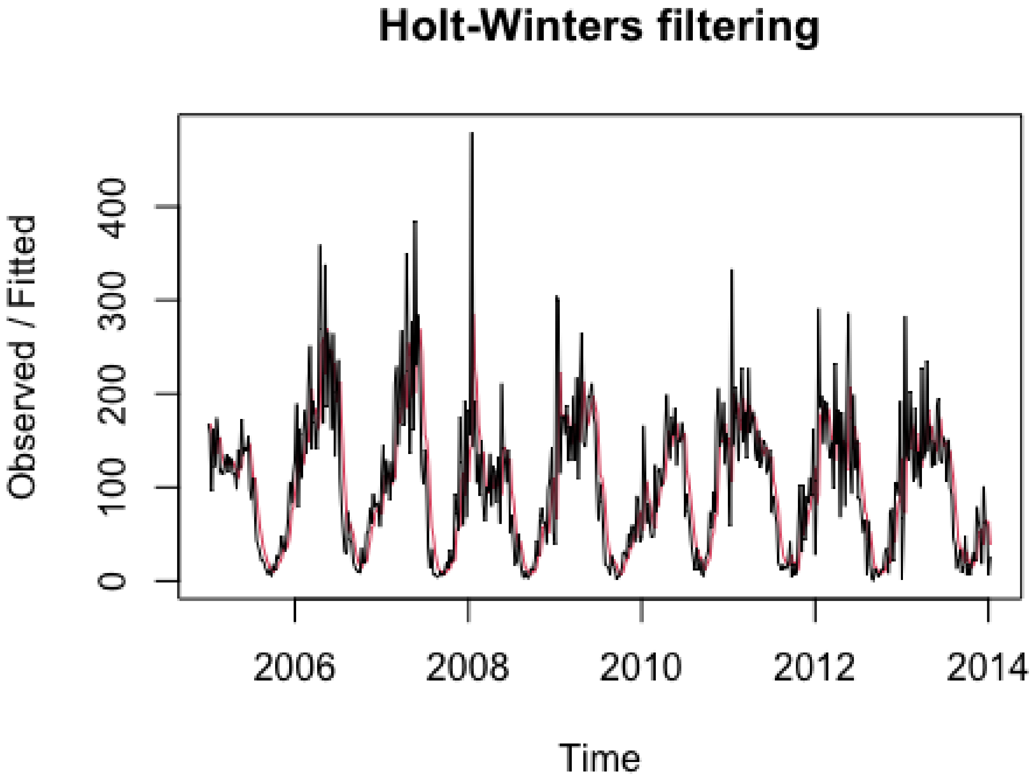
We can notice the negative value from the seasonal component graph for every half of a year. This behavior corroborates the evidence of our preliminary investigation of the time series. It looks like a spike of chickenpox cases in the 20th week of any given year, is followed by a fast decrease, and this pattern repeats itself from year to year from 2005 to 2014.

## Building Models for Forecasting

### *Model 1: exponential smoothing*

In exponential smoothing we wanted to find an alpha parameter that can reduce all the ‘noise’ attributed by the residual effect in this time series. In this case the best suited parameter is an alpha = 0.419.

```
## Holt-Winters exponential smoothing without trend and without seasonal component.
## Smoothing parameters:
##  alpha: 0.4193598
##  beta : FALSE
##  gamma: FALSE
##
## Coefficients:
##      [,1]
## a 33.26326
```

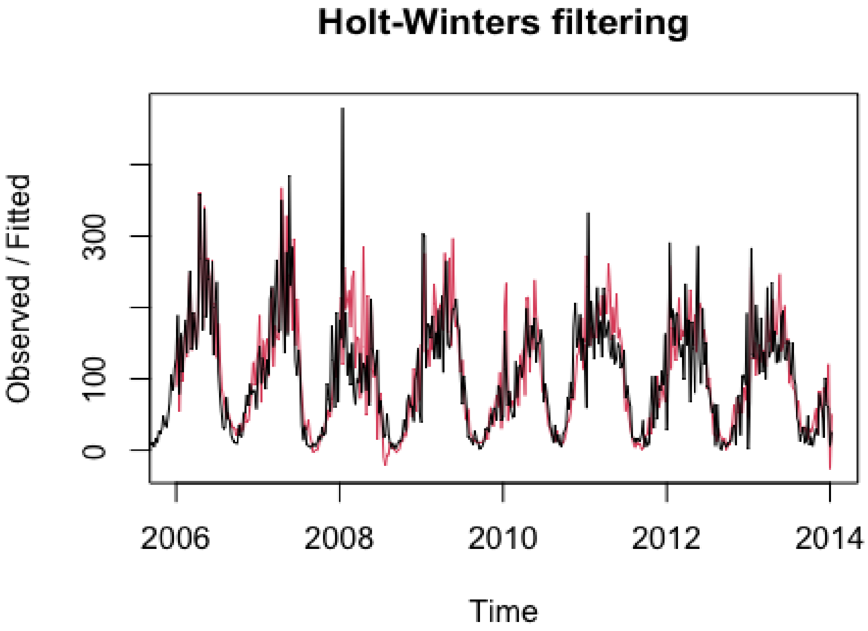


By looking at the graph, this smoothing technique gives a parameter of 128.9041. When it comes to finding best-fitted values, the smoothed time series shows a reasonable degree of accuracy, as we can notice by looking at the graph. However, when it comes to forecasting, the smoothed time series with a = 0.41 remain constants to 33.263 chickenpox cases for all the future values of $t + n \cdot weeks$ after week 1 of 2014.

### Model 2: Holt-Winters

Because of the very high seasonal variation we observed in the time series, we wanted to apply smoothing on seasonal variation as well. In doing so we are going to use the following parameters for Holt-Winters, alpha = 0.214, beta = 0.002, gamma = 0.561. Because of the negligible secular trend and the high seasonal variation, we wanted to keep beta very low and gamma high. Here are shown coefficients for first 10 seasons (weeks) only.

```
## Holt-Winters exponential smoothing with trend and additive seasonal component.
## Smoothing parameters:
##  alpha: 0.2138998
##  beta : 0.002204984
##  gamma: 0.5611843
##
## Coefficients:
##            [,1]
## a    73.4751230
## b     0.3081583
## s1  160.6071252
## s2   17.2133983
## s3   25.9618062
## s4   36.7737947
## s5    8.9646005
## s6   -7.2432362
## s7   30.6786292
## s8  -21.6410398
## s9    2.9556759
## s10 -26.7635856
## .    .    .   .

## .    .    .   .

## .    .    .   .
## s52 -39.0066177
```



Although, smoothing with Holt-Winters gives fitted values for the time series a little bit off with respect of exponential smoothing, because it takes seasonal variation in consideration, we predict that it may gives moreprecise forecast values once compared with the test time series.

### Model 3: Regression

We wanted to construct a regression model for fitted values and forecasting as well. Since it is now known by different analyses that the number of chickenpox cases is affected by seasonality, in our regression model, we want to include seasonal variation in our regression model. Again, for simplicity, we are omitting categoricals for weeks over the 10th.

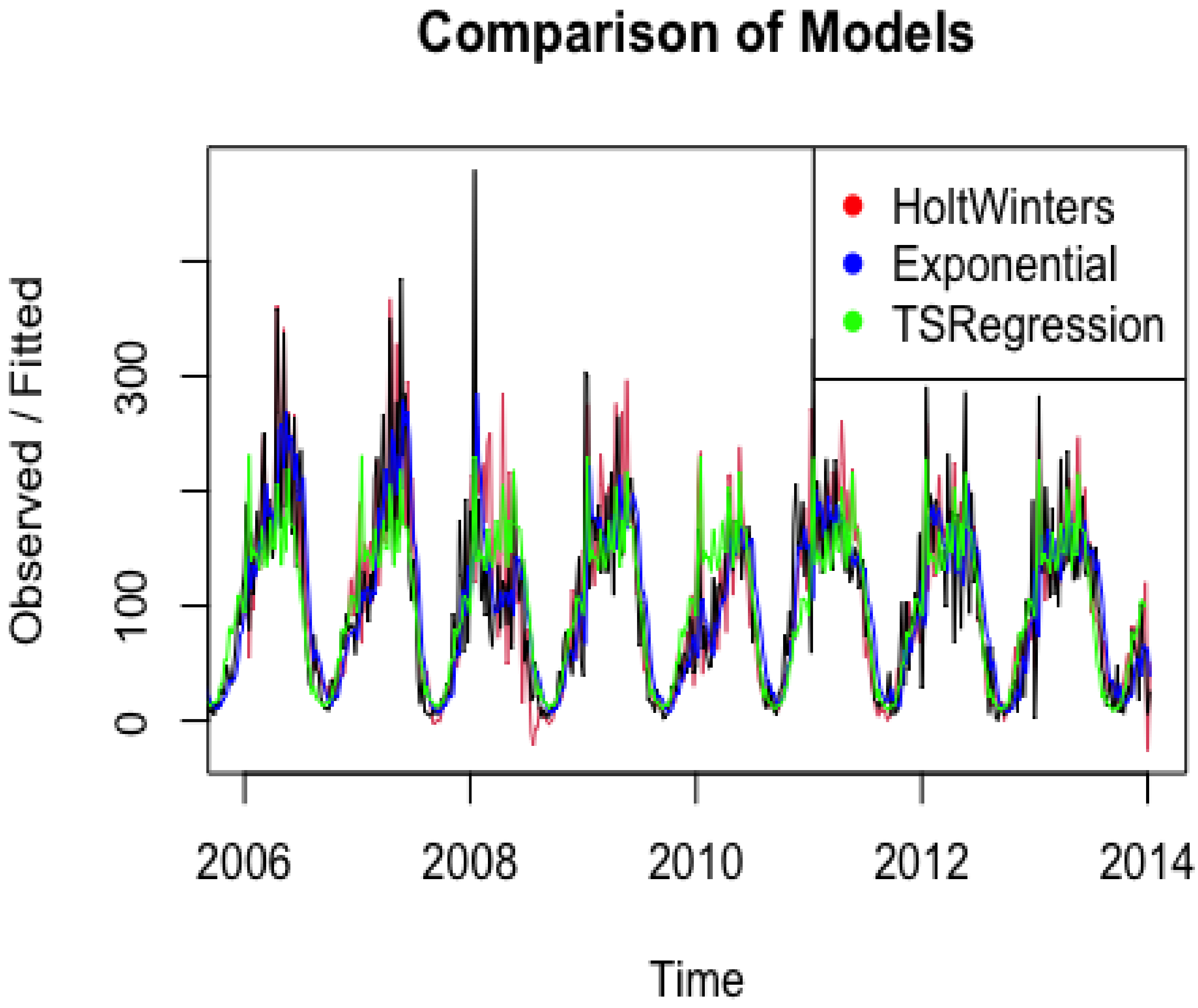
```
##
## Call:
## lm(formula = X ~ t + as.factor(nthWeek), data = NEWhungary_chickenpox)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.329  -22.141   -3.499   18.829  248.779
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             94.56066   16.32757   5.791 1.38e-08 ***
## t                       -0.01066    0.01648  -0.646 0.518316
## as.factor(nthWeek)2     57.56621   22.57078   2.550 0.011116 *
## as.factor(nthWeek)3    137.35465   22.57080   6.086 2.64e-09 ***
## as.factor(nthWeek)4     45.47642   22.57083   2.015 0.044567 *
## as.factor(nthWeek)5     55.59818   22.57087   2.463 0.014172 *
## as.factor(nthWeek)6     49.83106   22.57092   2.208 0.027808 *
## as.factor(nthWeek)7     49.17505   22.57099   2.179 0.029917 *
## as.factor(nthWeek)8     37.85238   22.57107   1.677 0.094289 .
## as.factor(nthWeek)9     74.30748   22.57116   3.292 0.001079 **
## as.factor(nthWeek)10    52.76258   22.57126   2.338 0.019882 *
## .                 .   .          .          .     .

## .                 .   .          .          .     .

## .                 .   .          .          .     .
## as.factor(nthWeek)51    14.75507   22.58582   0.653 0.513931
## as.factor(nthWeek)52     8.76573   22.58642   0.388 0.698143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.88 on 415 degrees of freedom
## Multiple R-squared:  0.6481, Adjusted R-squared:  0.604
## F-statistic:  14.7 on 52 and 415 DF,  p-value: < 2.2e-16
```

From the R output, we can write the estimated time series regression model for the number of cases of chickenpox in Budapest. Because there are 52 weeks, in our model, we have 52 categorical here defined as the nth week of the year where the first week of the year is considered the base case. Here in our model interpretation, we will only consider some of those.

$$\#\widehat{Cases} = 94.561 \cdot t - 0.010 \cdot t + 57.566 \cdot 2^{nd} week + 137.355 \cdot 3^{rd} week + 45.467 \cdot 4^{th} week + \cdots\cdots\cdots$$

The regression model is consistent with our prior findings. The coefficient of time t is very close to zero, which implies that the effect of the secular trend is weak. Another interesting result is that as the week changes from the first week of the year to the second week of the year, the average expected number of chickenpox cases increases by 57.566.

Follows a plot for fitted values by the three model vs the train time series for period 2005-1014.


Comparison of Models
HoltWinters
Exponential
TSRegression
Observed / Fitted
0
100
300
2006
2008
2010
2012
2014
Time

## Forecasting chickenpox cases for the first 4 weeks of 2014

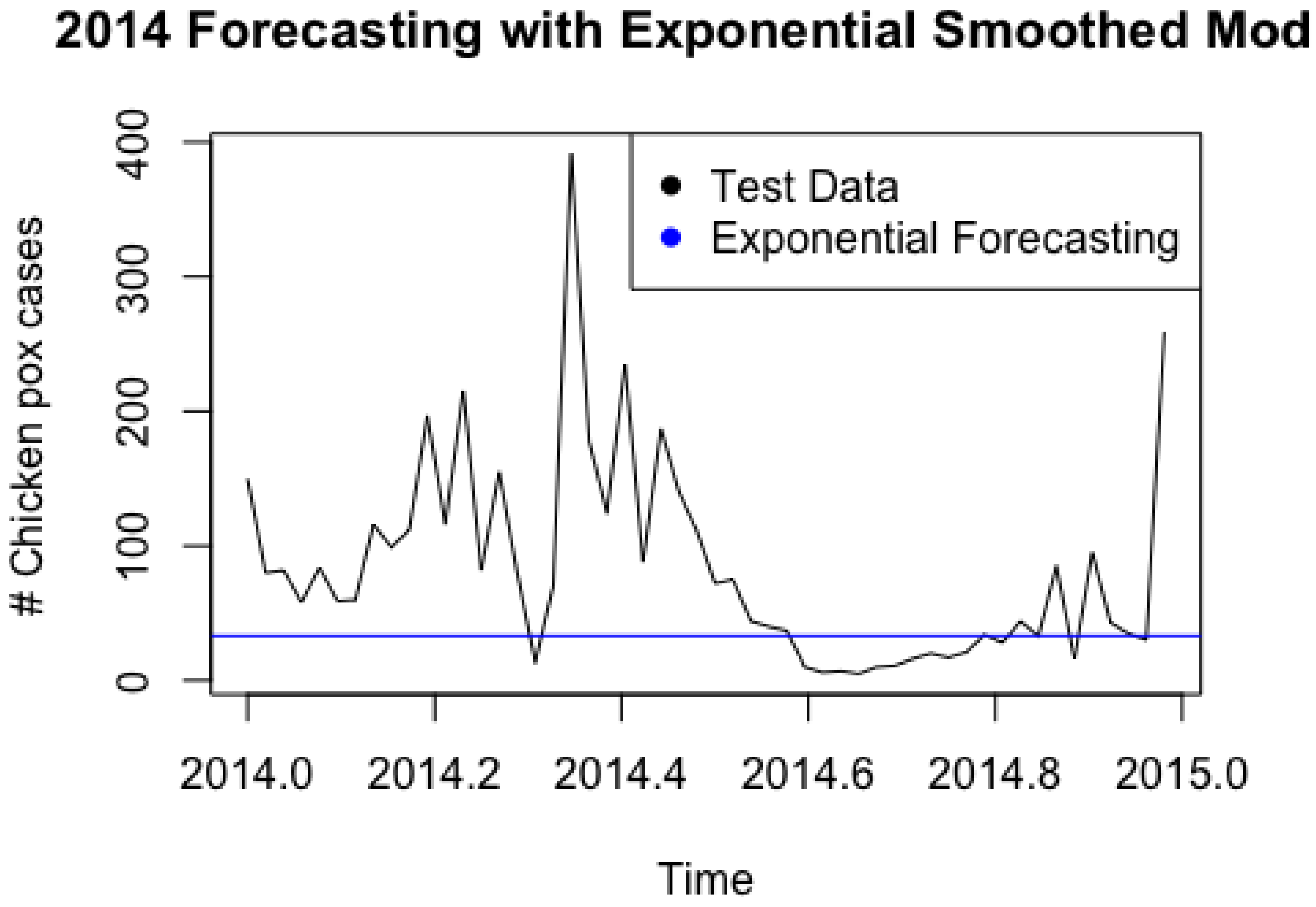
Now we want to forecast the first six weeks of chickenpox cases in Budapest by applying the three models we built before, the ones with exponential smoothing, Holt-Winters, and Regression.

### Forecasting using Model 1: Exponential Smoothing with alpha = 0.419

Forecasting using exponential smoothing gives constant values for all future n weeks starting from week 1 of the year 2014. The forecasting using our exponential smoothed model predicts 128.9 chickenpox cases for the first four weeks of the year 2014.

```
Forecast values first 4 weeks 2014 by Exponential Smoothing Model
## [1] 33.26326 33.26326 33.26326 33.26326
```

It follows forecasting for the whole 2014 year versus the test data set (last recorded year), along with MAPE, MAD, and RMSE tests as a measure of overall forecast accuracy for the first four weeks of 2014.



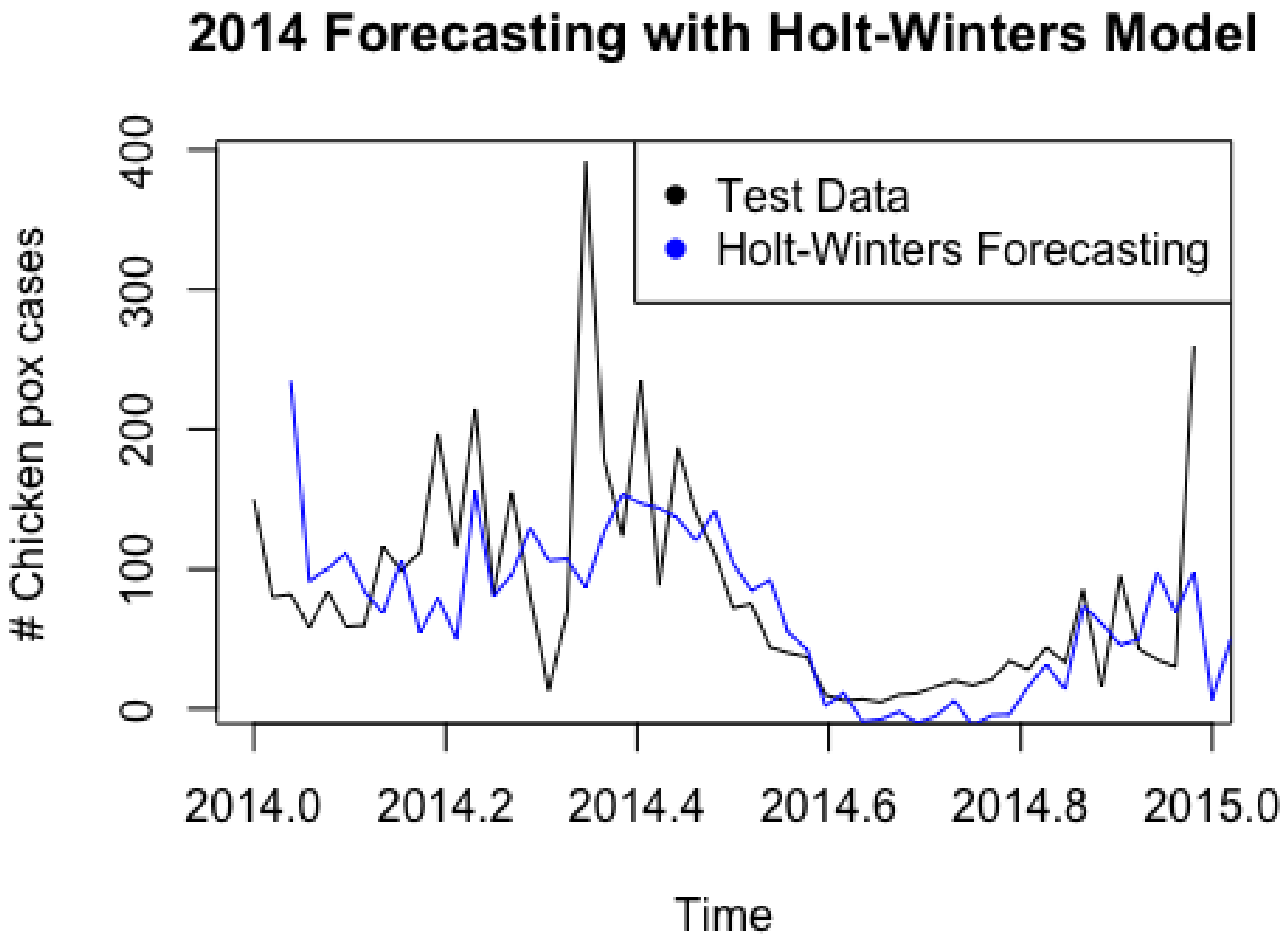
```
## [1] "MAPE:59.237,  MAD:59.583,  RMSE:68.555"
```

### Forecasting using Model 2: Holt-Winter with alpha = 0.214, beta = 0.002, gamma = 0.561

Forecasting using this method should forecast chickenpox cases for the first four weeks of 2014 closer to those from the test time series with respect to the exponentially smoothed model. Here is the forecasting using our Holt Winter model for the first four weeks of the year 2014.

```
Forecast values first 4 weeks 2014 by Holt-Winters Model
## [1] 234.39041  91.30484 100.36140 111.48155
```

It follows forecasting for the whole 2014 year versus the test data set (last recorded year), along with MAPE, MAD, and RMSE tests as a measure of overall forecast accuracy for the first four weeks of 2014.



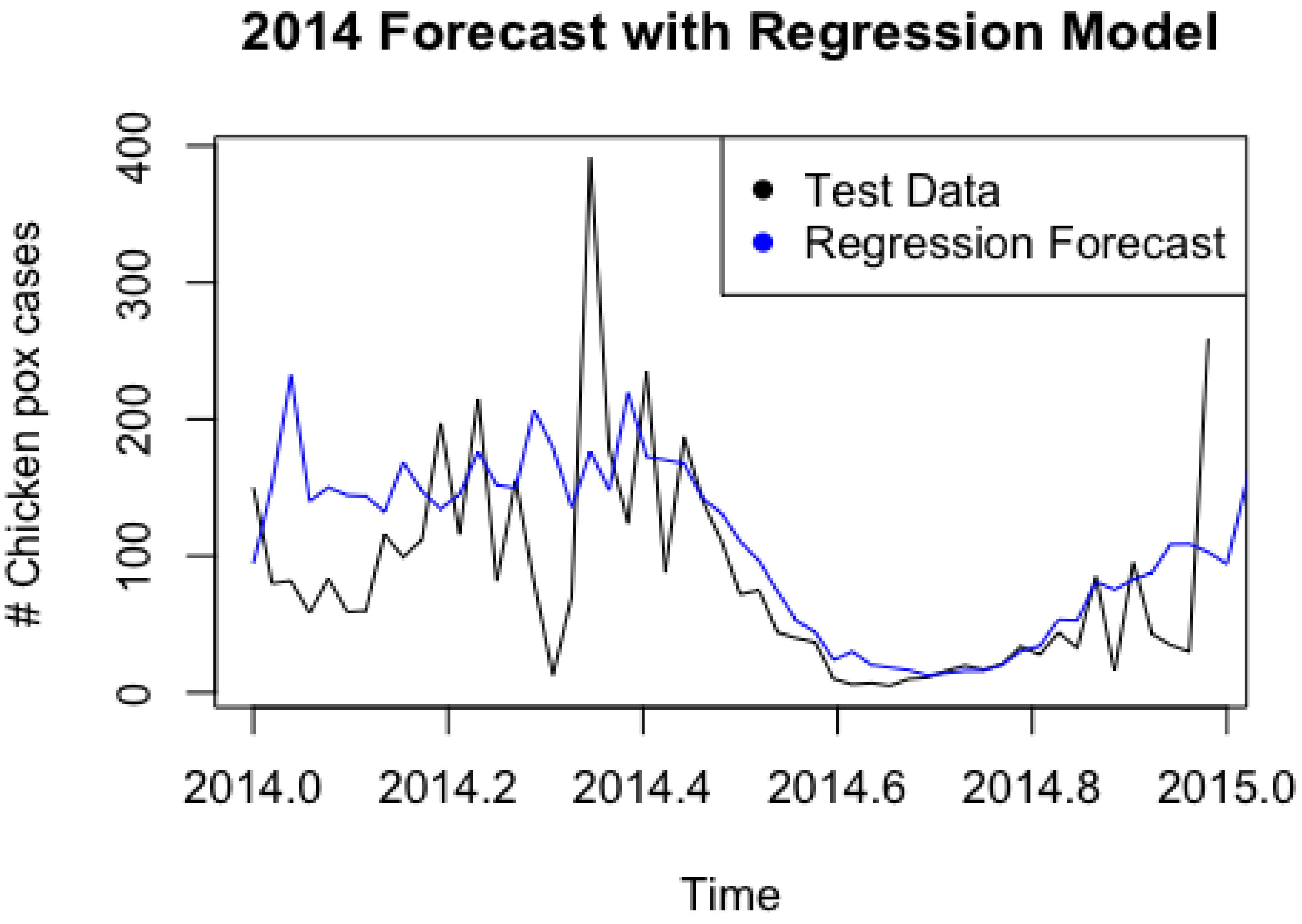
```
## [1] "MAPE:41.885,  MAD:46.248,  RMSE:51.105"
```

### Forecasting using Model 3: Regression Model

Here we want to use our regression model to forecast the number of cases of chickenpox in Budapest after the end of the year 2013. Here is the forecast for the first four weeks of 2014.

```
Forecast values first 4 weeks 2014 by Holt-Winters Model

## Week1 2014 Week2 2014 Week3 2014 Week4 2014
##   89.57316  147.12871  226.90649  135.01760
```

It follows forecasting for the whole 2014 year versus the test data set (last recorded year), along with MAPE, MAD, and RMSE tests as a measure of overall forecast accuracy for the first four weeks of 2014.



```
## [1] "MAPE:87.37,  MAD:108.425,  RMSE:93.658"
```

|  | MAPE | MAD | RMSE |
|---|---|---|---|
| **Exponential Smoothed Model** | 59.237 | 59.583 | 68.555 |
| **Holt-Winters Model** | 41.885 | 46.248 | 51.105 |
| **Time Series Regression Model** | 87.37 | 108.425 | 93.658 |

## Conclusions

In selecting the best model, we want the model that gives the lowest score for the measure of overall forecast accuracy by evaluating their respective MAPE, MAD, and RMSE. In this case, the Holt-Winters model with parameters alpha = 0.214, beta = 0.002, and gamma = 0.561 is the one we want to use to forecast the number of chickenpox cases in Budapest. For example, the test dataset for the second week of 2014 has an observed value of chickenpox cases of 81, while our model forecasts  91.3 chickenpox cases for the same period, only a 14% discrepancy between the predicted and the observed value.

The R-Studio project with the complete R-markdown with R chuncks, and original tables are in this GitHub repository that has been set public and can be visioned at https://github.com/Doriasamp/TimeSeriesProject