Time Series Student Project
Fall 2013

I. Introduction

This student project examines the popularity of the name “Frederic,” its variants, and diminutives, in the state of New York from the years 1910-2012. The data was obtained from http://www.ssa.gov/OACT/babynames/ , the official website of the US Social Security Administration (SSA). The data presented is limited to those officially recorded by the SSA, and is originally expressed in terms of number of occurrences per name per year. This study defines popularity in terms of percentage of recorded names with respect to the total number of recorded names in a particular given year. For each given year, the study considers the total occurrences of names from both genders, and includes, but is not limited to, the following representative names: Frederic, Fredrick, Fred, Freda, Frederica, Federick, Freddy, Fredricka, etc. The complete data including all names can be located in the “Raw” tab of the accompanying Excel workbook. The filtered list of total occurrences of names of interest per year is filtered in the “Frederic” tab. All relevant steps in cleaning up the data are outlined in the remaining tabs: The “Summary Totals Per Year (Frdrk)” tab contains the total occurrences for each given year of all names of interest for both genders combined. The “Summary Totals Per Year (all)” tab contains the total number of registered names across both genders for each given year, and lastly, the “Popularity per year” tab takes the ratio of the results from the previous two arrays. The resulting form of the data is in percentage (%), and serves as the starting point of the time series analysis.

II. Objective

The objective of this project is to study and apply the fundamental techniques of time series analysis and arrive at a suitable time series model for the historical popularity of the aforementioned names of interest.

III. Calculations and Analyses

In order to determine the approach and time series techniques to be used, a preliminary plot of the popularity per year was determined. The resulting graph is seen below:

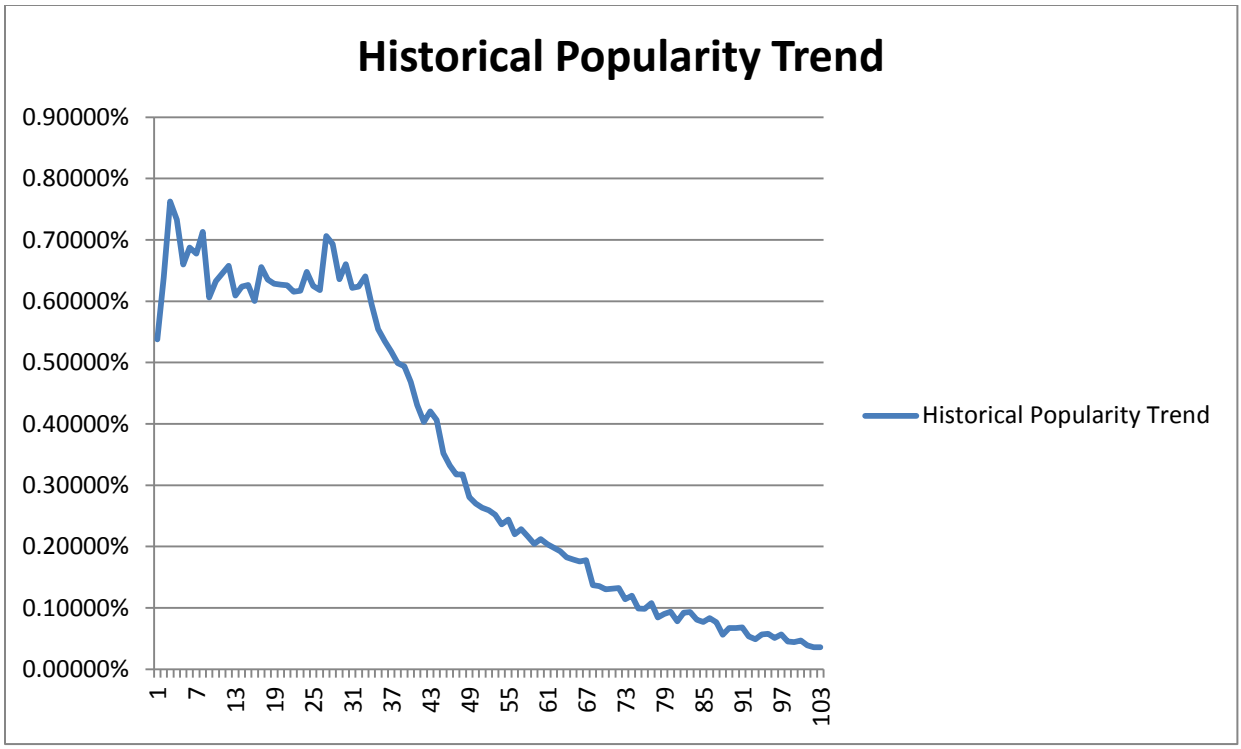

Figure 1: An initial plot of the historical popularity of the names of interest for the years 1910-2012

The above graph shows annual figures and no apparent seasonality. There is an observable downward trend in the plot starting from lag 33 onwards. This implies that the time series is not stationary. To examine this idea, the autocorrelation function was calculated for each lag and the corresponding correlogram was derived. All supporting calculations can be found in the "Calc_Initial + Correlogram" tab of the accompanying Excel workbook.

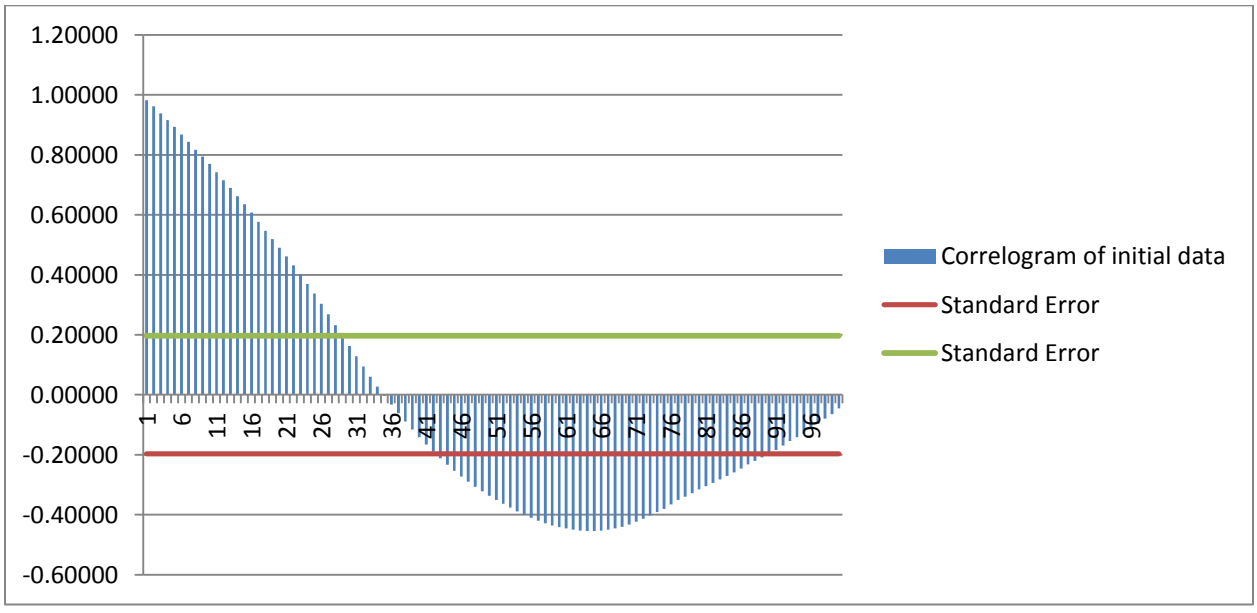

Figure 2: The correlogram of the initial data. The horizontal lines denote the standard error of $\pm 2/\sqrt{n}$.

It can be seen from Figure 2 that the autocorrelation function for the initial data does not approach zero until lags 34-35. In addition, the values are way beyond the standard error. This supports the claim that the original time series is not stationary. Taking first differences in order to achieve stationarity would be reasonable approach.

First differences were taken in order to produce a stationary time series. All supporting calculations can be found in the “Calc_1st Diff + Correlogram” tab of the accompanying Excel workbook. The results of the calculations are summarized below:

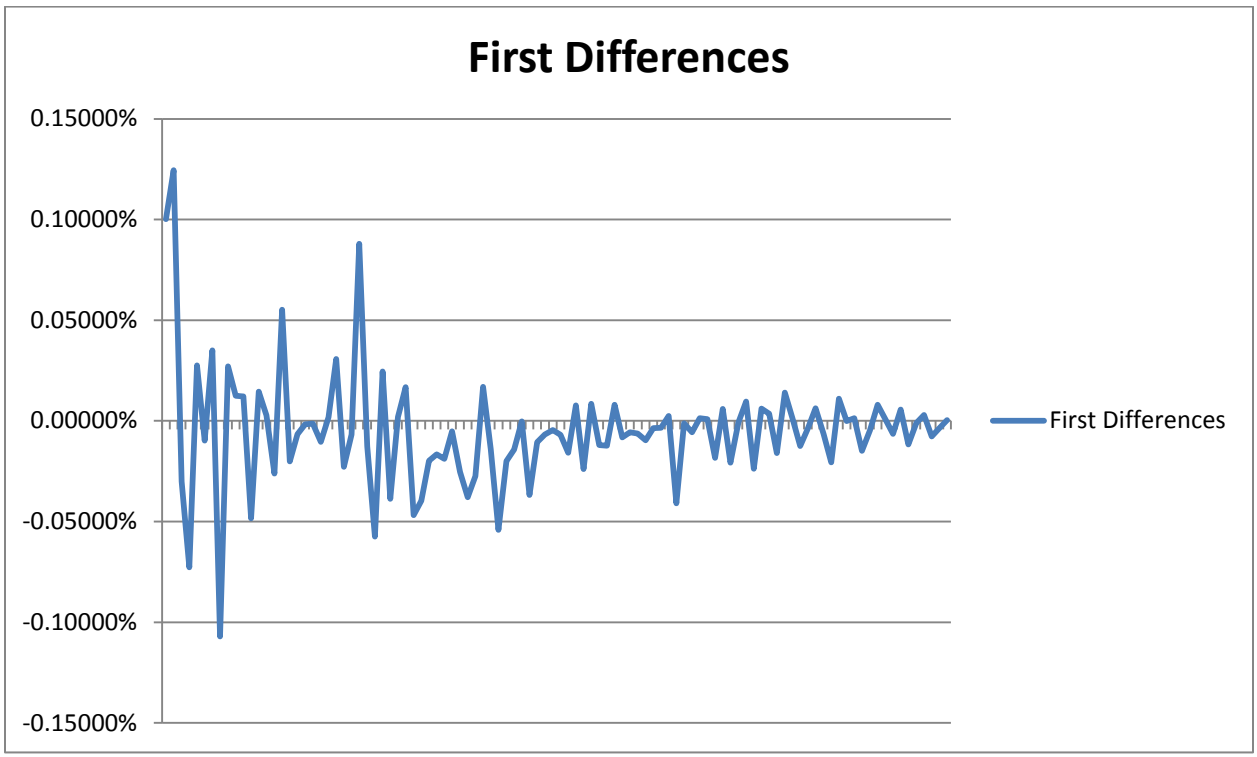

Figure 3: The time series plot obtained by taking first differences.

It can be seen from Figure 3 that there are no more observable trends in the graph, and the first differences serve as a viable candidate for a stationary time series. In order to examine stationarity, the autocorrelation function was calculated similar to the initial data, and the corresponding correlogram was plotted.

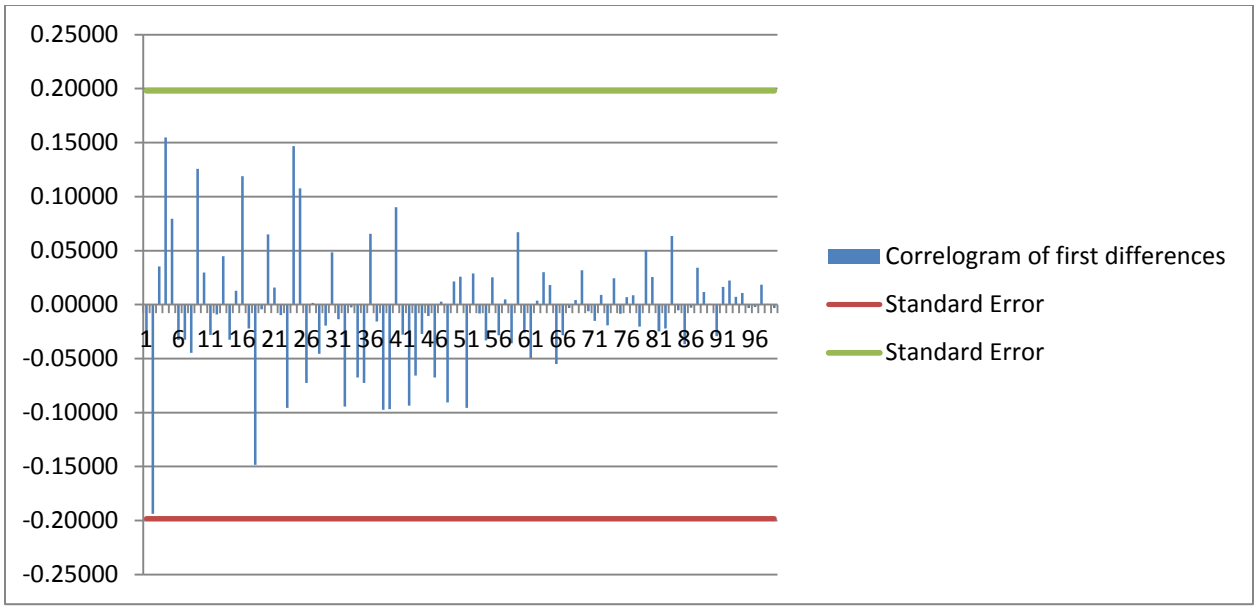

Figure 4: The correlogram of first differences. The horizontal lines denote the standard error of $\pm 2/\sqrt{n}$.

Based on the graph of the first differences and its corresponding correlogram, it was concluded that its time series is stationary. From here, the autoregressive (AR) model was selected as a possible fit. To begin doing this, the autoregressive time series of order 1, AR(1), was selected as a starting point:

AR(1): $Y_t = ø_1 Y_{t-1} + e_t$

The *Regression* add-in feature of Excel was utilized in order to determine an autoregressive fit and perform residual analysis on the data. The summary of the AR(1) model results are seen below. All supporting calculations are located in the "Calc_AR(1) + D.W. stat" tab of the accompanying Excel workbook.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.040434407							
R Square	0.001634941							
Adjusted R Square	-0.008449554							
Standard Error	0.000274128							
Observations	101							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	1.2183E-08	1.2183E-08	0.16212425	0.688076382			
Residual	99	7.43947E-06	7.51461E-08					
Total	100	7.45165E-06						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-6.15004E-05	2.76716E-05	-2.222511906	0.028524273	-0.000116407	-6.59E-06	-0.00011641	-6.594E-06
Yt-1	-0.037711899	0.093660055	-0.402646557	0.688076382	-0.223553767	0.14813	-0.22355377	0.14812997

Below is an excerpt of the residual data used to determine the Durbin-Watson (D-W) Statistic.

RESIDUAL OUTPUT					
			Durbin-Watson Statistic:	2.15742	
Observation	*Predicted Yt*	*Residuals*	*Residual Squared*	*Residual Difference*	*Residual Difference Squared*
1	-0.00993%	0.13446%	0.00018%		
2	-0.01085%	-0.01922%	0.00000%	-0.15369%	0.00024%
3	-0.00502%	-0.06771%	0.00005%	-0.04849%	0.00002%
4	-0.00341%	0.03097%	0.00001%	0.09869%	0.00010%
5	-0.00719%	-0.00255%	0.00000%	-0.03353%	0.00001%
6	-0.00578%	0.04089%	0.00002%	0.04344%	0.00002%
7	-0.00747%	-0.09953%	0.00010%	-0.14042%	0.00020%
8	-0.00211%	0.02915%	0.00001%	0.12868%	0.00017%
9	-0.00717%	0.01967%	0.00000%	-0.00948%	0.00000%
10	-0.00662%	0.01880%	0.00000%	-0.00087%	0.00000%
11	-0.00661%	-0.04187%	0.00002%	-0.06067%	0.00004%

The resulting AR(1) fit is a time series with equation

$$Y_t = -0.03771Y_{t-1} - 0.00006.$$

The graph of the actual versus predicted values is shown below.

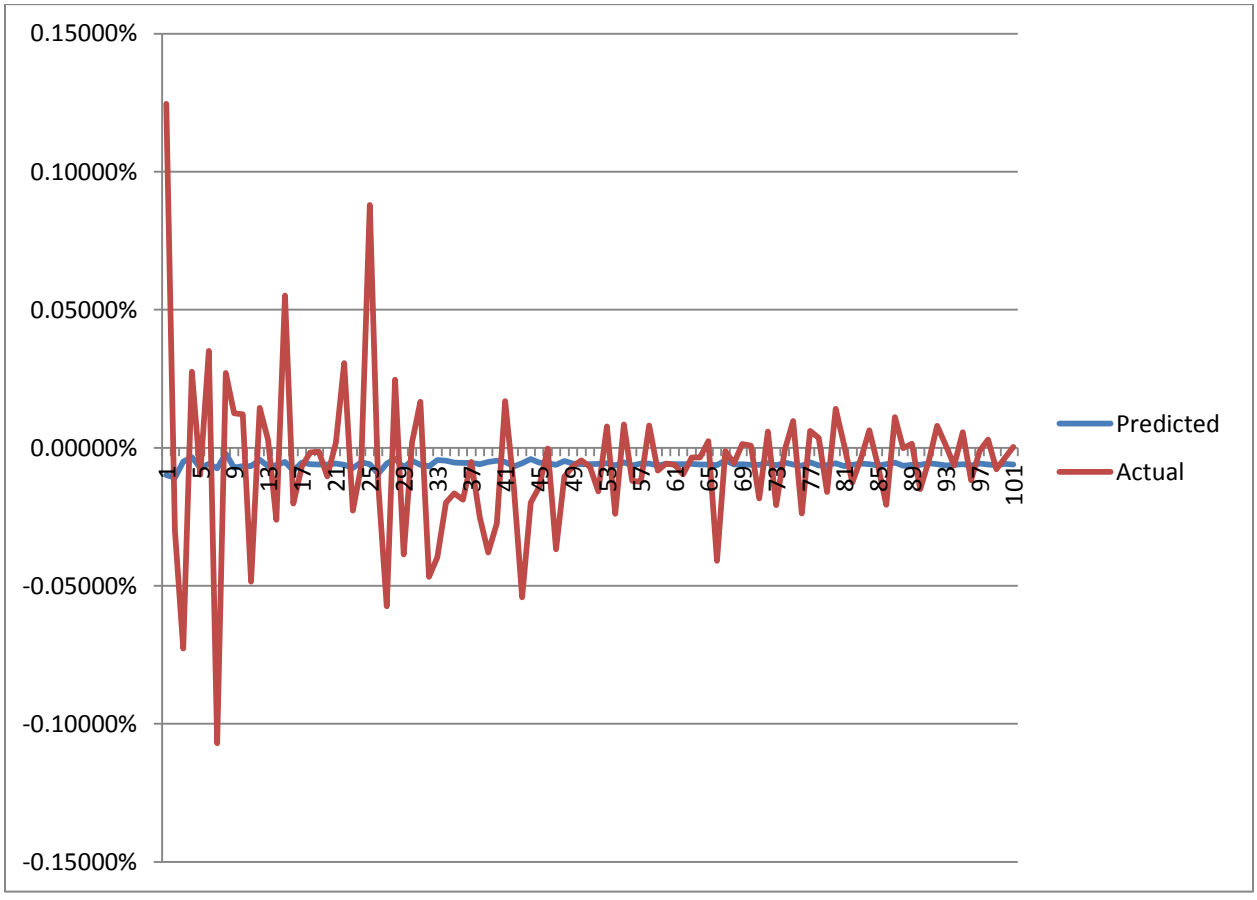

Figure 5: The actual versus predicted plot of the time series $Y_t = -0.03771Y_{t-1} - 0.00006$.

The D-W Statistic based on the above calculations is 2.15742. This is between 1.80 and 2.20, which suggests that the time series may be white noise. More important, however, is the notable lack of good fit provided by the AR(1) model, as shown in Figure 5. Thus, the AR(2) model is examined with the aim of providing a better fit. The following AR(2) equation is examined:

$$\text{AR(2): } Y_t = \text{ø}_1 Y_{t-1} + \text{ø}_2 Y_{t-2} + e_t$$

The summary of the AR(2) results are shown below. All supporting calculations are located in the "Calc_AR(2) + D.W. stat" tab of the accompanying Excel workbook.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.360601982							
R Square	0.13003379							
Adjusted R Square	0.112096342							
Standard Error	0.00022673							
Observations	100							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	7.45318E-07	3.72659E-07	7.249291671	0.001163766			
Residual	97	4.98641E-06	5.14063E-08					
Total	99	5.73173E-06						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-9.71693E-05	2.3573E-05	-4.122061493	7.91894E-05	-0.000143955	-5.03834E-05	-0.000143955	-5.03834E-05
Yt-2	-0.203050784	0.077530294	-2.61898637	0.010234841	-0.356926957	-0.049174611	-0.356926957	-0.049174611
Yt-1	-0.238463079	0.083148962	-2.867902042	0.005069763	-0.403490752	-0.073435405	-0.403490752	-0.073435405

Below is an excerpt of the residual data used to determine the D-W Statistic.

RESIDUAL OUTPUT					
			Durbin-Watson Statistic:	2.18128	
Observation	*Predicted Yt*	*Residuals*	*Residual Squared*	*Residual Difference*	*Residual Difference Squared*
1	-0.05977%	0.02970%	0.00001%		
2	-0.02783%	-0.04490%	0.00002%	-0.07460%	0.00006%
3	0.01373%	0.01383%	0.00000%	0.05873%	0.00003%
4	-0.00152%	-0.00822%	0.00000%	-0.02205%	0.00000%
5	-0.01299%	0.04810%	0.00002%	0.05632%	0.00003%
6	-0.01611%	-0.09089%	0.00008%	-0.13899%	0.00019%
7	0.00867%	0.01836%	0.00000%	0.10926%	0.00012%
8	0.00556%	0.00694%	0.00000%	-0.01143%	0.00000%
9	-0.01819%	0.03037%	0.00001%	0.02343%	0.00001%
10	-0.01516%	-0.03332%	0.00001%	-0.06368%	0.00004%
11	-0.00063%	0.01513%	0.00000%	0.04844%	0.00002%

The resulting AR(2) fit is a time series with equation

$$Y_t = -0.23846Y_{t-1} - 0.20305Y_{t-2} + 0.00010.$$

The graph of the actual versus predicted values is shown below.

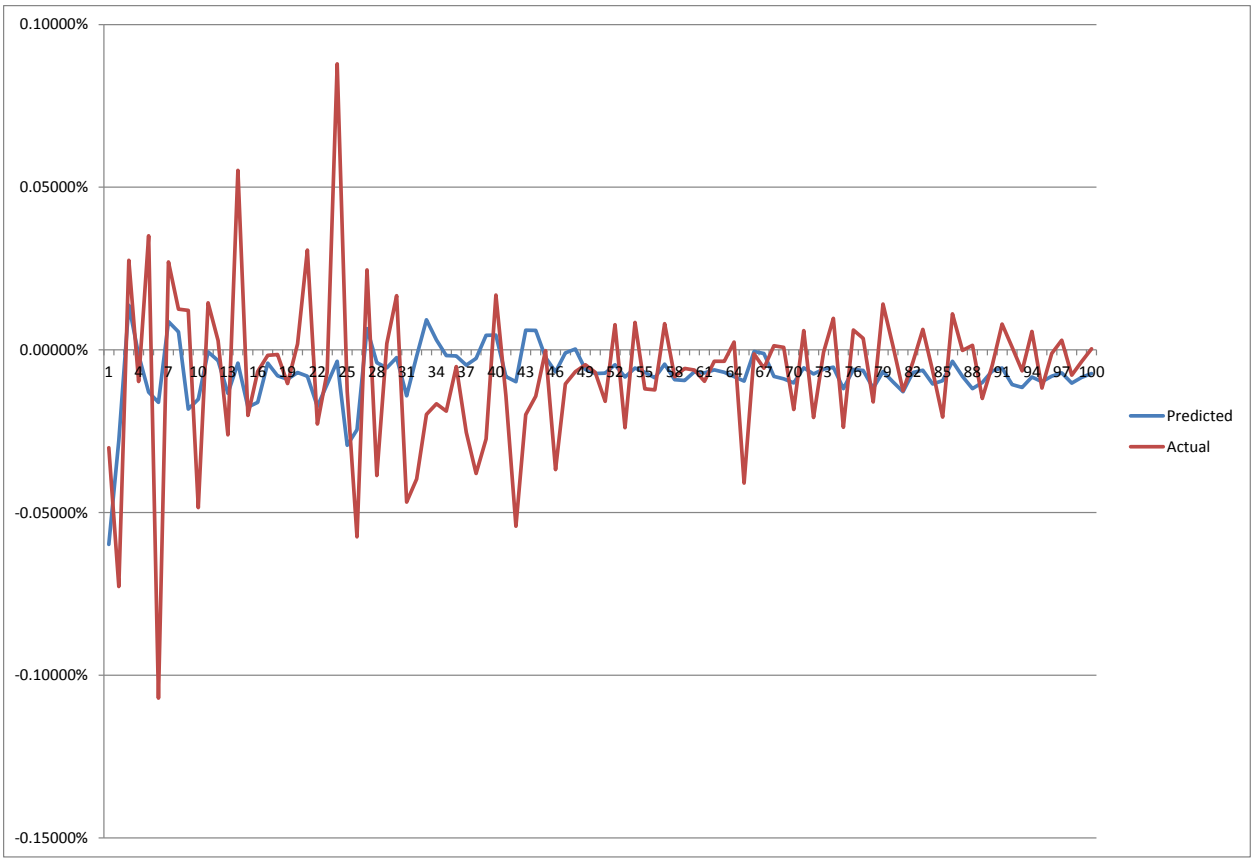

Figure 6: The actual versus predicted plot of the time series $Y_t = -0.23846Y_{t-1} - 0.20305Y_{t-2} + 0.00010$.

It can be seen from Figure 6 that the AR(2) model provides a significantly better fit than the AR(1). The D-W Statistic is equal to 2.18128, which is again between 1.80 and 2.20 just like in the AR(1) assumption. However, based on the above results, the AR(1) model will be discarded.

At this point, it is reasonable to adopt the AR(2) model as a fit for the given time series. As a final step, an AR(3) model is examined with the aim of invoking the Principle of Parsimony and retain the AR(2) assumption. In order to do this, the following AR(3) model is considered:

$$\text{AR(3): } Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + e_t.$$

The summary of the AR(3) results are shown below. All supporting calculations are located in the “Calc_AR(3)” tab of the accompanying Excel workbook.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.38660239							
R Square	0.149461408							
Adjusted R Square	0.122602294							
Standard Error	0.000225491							
Observations	99							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	8.48822E-07	2.83E-07	5.564644	0.001463854			
Residual	95	4.83038E-06	5.08E-08					
Total	98	5.6792E-06						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.000113605	2.55486E-05	-4.44663	2.37E-05	-0.000164326	-6.29E-05	-0.00016433	-6.28849E-05
Yt-3	-0.040477199	0.07978635	-0.50732	0.613106	-0.19887313	0.1179187	-0.19887313	0.117918731
Yt-2	-0.270261403	0.086142938	-3.13736	0.00227	-0.441276756	-0.099246	-0.44127676	-0.09924605
Yt-1	-0.338705477	0.101039524	-3.35221	0.001152	-0.539294294	-0.138117	-0.53929429	-0.13811666

The D-W Statistic was no longer calculated for the AR(3) model, since it is simply and extension study. It was determined that the D-W Statistic will only be calculated and further examined should the AR(3) model show any potential advantage over the AR(2).

The resulting AR(3) fit is a time series with equation

$$Y_t = -0.33871Y_{t-1} - 0.27026Y_{t-2} - 0.04048Y_{t-3} - 0.00011.$$

The graph of the actual versus predicted values is shown below.

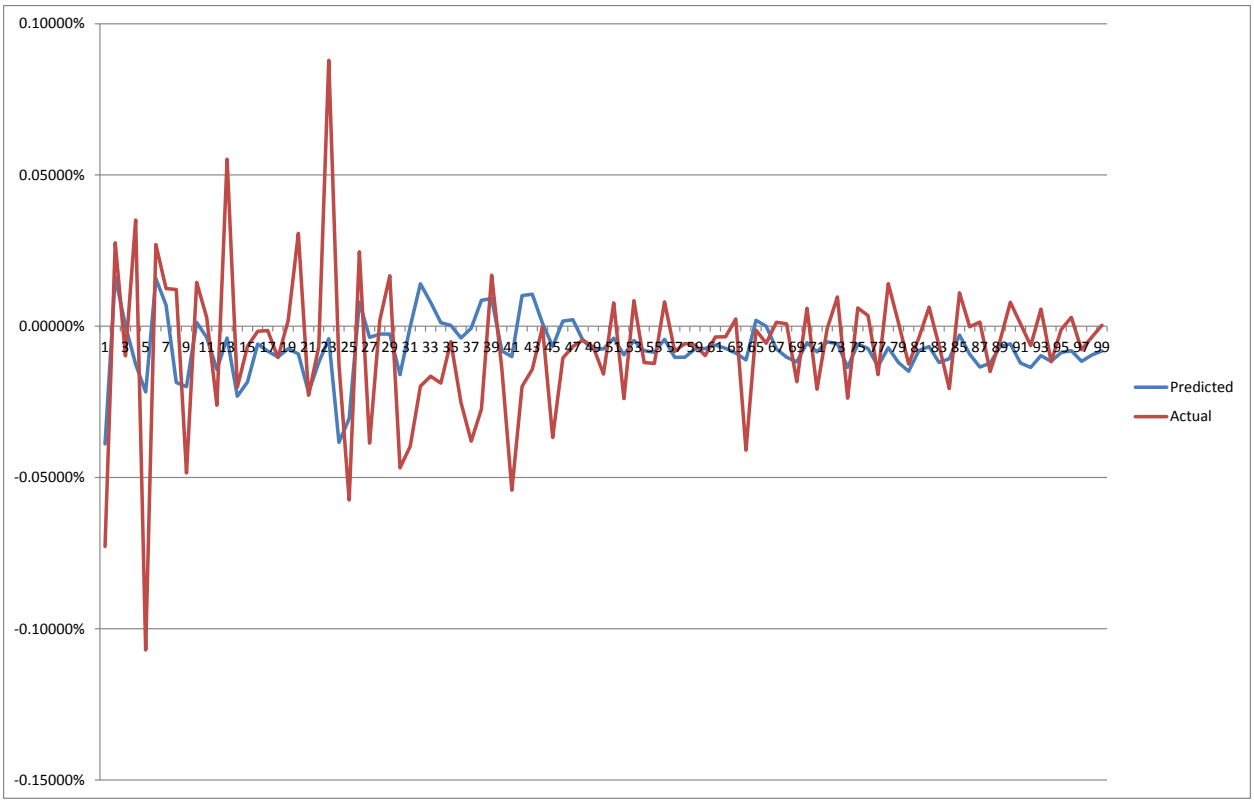

Figure 7: The actual versus predicted plot of the time series $Y_t = -0.33871Y_{t-1} - 0.27026Y_{t-2} - 0.04048Y_{t-3} - 0.00011$.

It can be seen from Figure 7 that the AR(3) model has no significant advantage over the AR(2) in terms of goodness-of-fit. In fact, the coefficient of determination, R^2, of the AR(3) model is 14.95% -- a mere 1.95% difference from the R^2 statistic of 13.00% from the AR(2) model.

IV. Conclusion

Since the AR(3) model proved no significant advantage over the AR(2) model, the Principle of Parsimony holds, and for pragmatic purposes, the AR(2) model is selected as the best-fit model among the time series that were studied.

V. Limitations

This study applies fundamental time series concepts and focuses on the autoregressive model in its statistical tests. It makes no use of MA, ARMA, and/or ARIMA models. Also, this study limits itself to AR models of relatively small degree, namely, orders 1-3. It remains open to the possibility that there are superior models in terms of goodness-of-fit, whether it be AR models of higher order, or of different nature (MA, ARMA, etc.) altogether.