Name analysis of James and Abigail
Scott Basco
Spring 2015 VEE Time Series Project

Time series can be useful in predicting many things – one of which being the popularity of names. The Social Security Administration keeps track of how popular given names are every year and ranks the top one thousand. For the past hundred years, James has been amongst the most popular names for boys, though it has become less popular in the past two decades. Time series can be used to model this trend and forecast possible future popularity rankings of the name James.

On the other hand, the name Abigail has quickly gained popularity in recent years. Despite not appearing on the top one thousand list regularly until 1949, Abigail is now one of the ten most popular names for newborn girls. Again, time series may be useful in forecasting the possible continued popularity of this name. Data was analyzed in R and output is used to help create results.

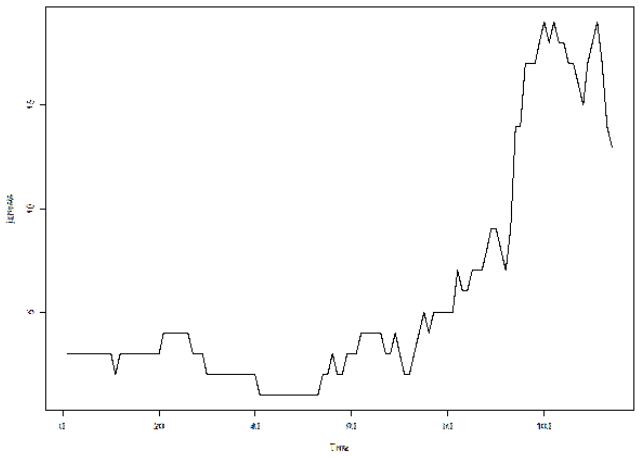

Figure 1: Popularity of name James, by year

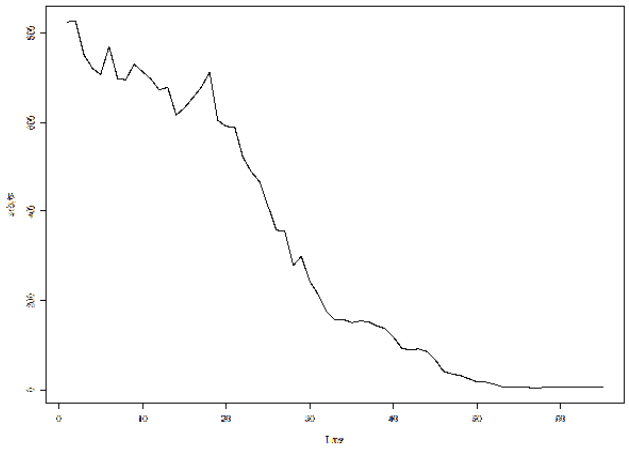

Figure 2: Popularity of name Abigail, by year

Figures 1 and 2 show this data in plots of rank by year (time 0 being the year 1900 for James and the year 1949 for Abigail) through 2013. Although James may have been stationary for many years, the activity in the most recent years changes this. Abigail's swiftly increasing popularity (since lower number means higher rank) is undoubtedly a non-stationary process. Differences are taken to turn the non-stationary processes into stationary processes.

Model for the name James

The first difference of the popularity of the name James is presented in Figure 3. Although it appears stationary, a Box-Pierce Q statistic is run to confirm this.

For the original series for James:

> Box.test(jamests,lag=40)

Box-Pierce test

data: jamests
X-squared = 1007.929, df = 40, p-value < 2.2e-16

The p-value is small enough that we must reject the null hypothesis that the series is stationary. The original series must be non-stationary.

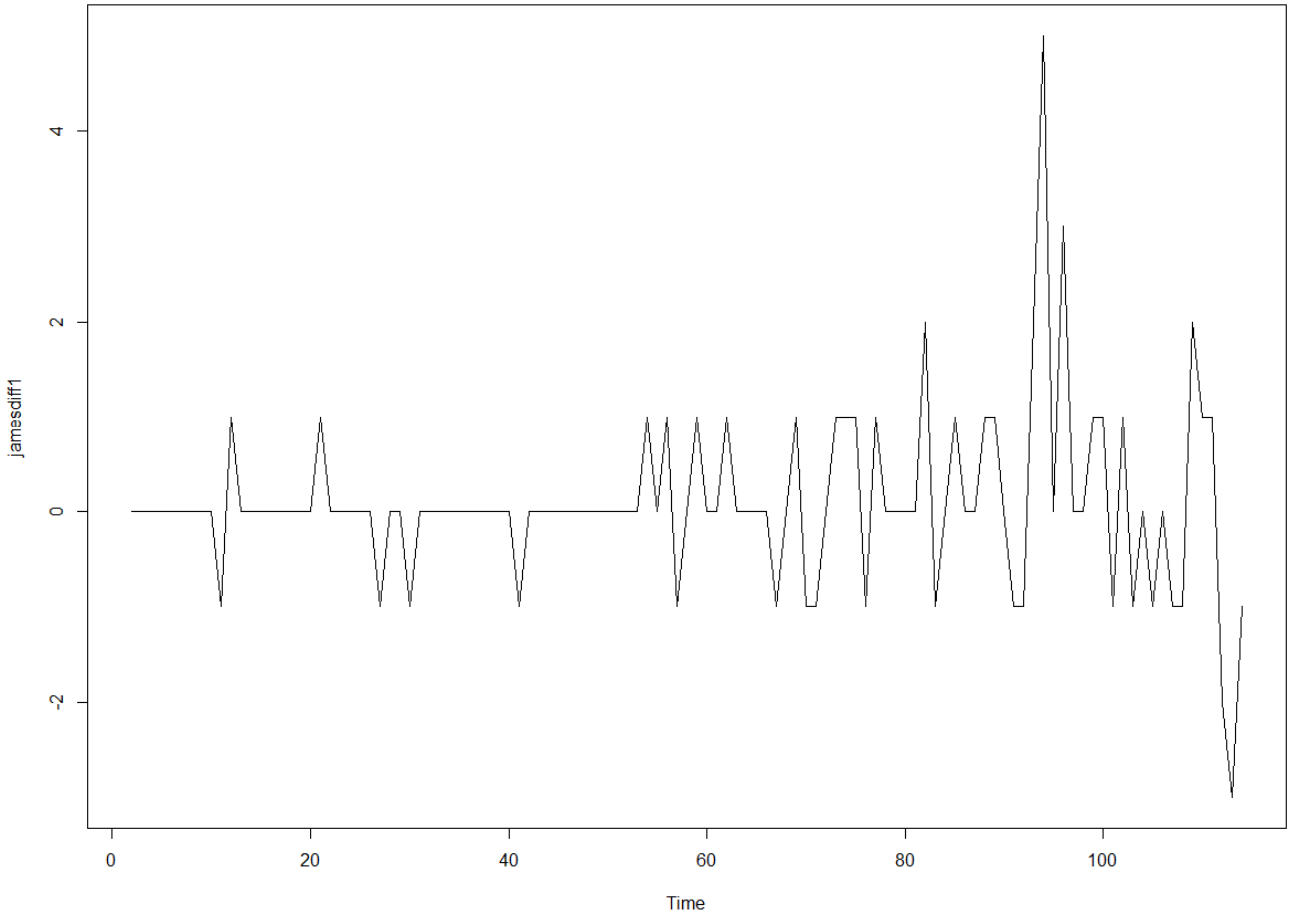

Figure 3: First difference of popularity of name James, by year

For the first difference:

> Box.test(jamesdiff1,lag=40)

Box-Pierce test

data: jamesdiff1
X-squared = 25.8541, df = 40, p-value = 0.9593

The p-value is well above .05, so here the null hypothesis cannot be rejected; the first differences are considered stationary.

In order to determine a proper ARIMA model for this data, correlograms and partial correlograms were used (Figures 4 and 5). Here, the high autocorrelation at lag 1 and low elsewhere implies a MA(1) process of the first-differences. The autocorrelation and partial autocorrelation at lag 18 are likely caused by noise, as the autocorrelation otherwise decays toward zero quickly. Thus, and ARIMA(0,1,1) model was used for the name James.

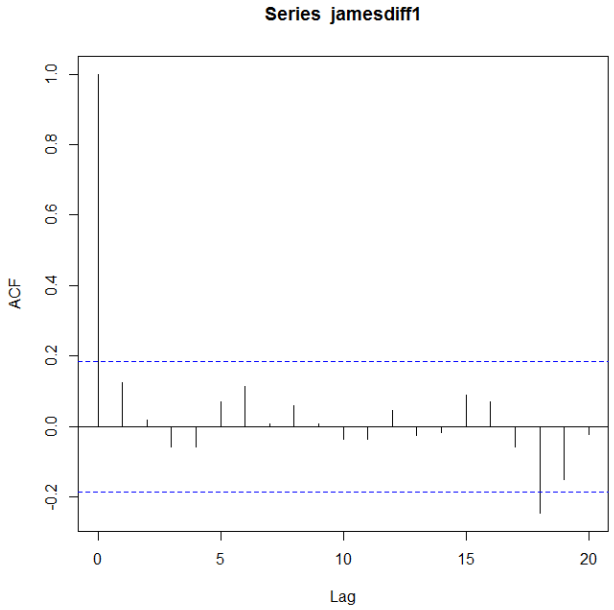

Figure 4: Autocorrelation for the name James

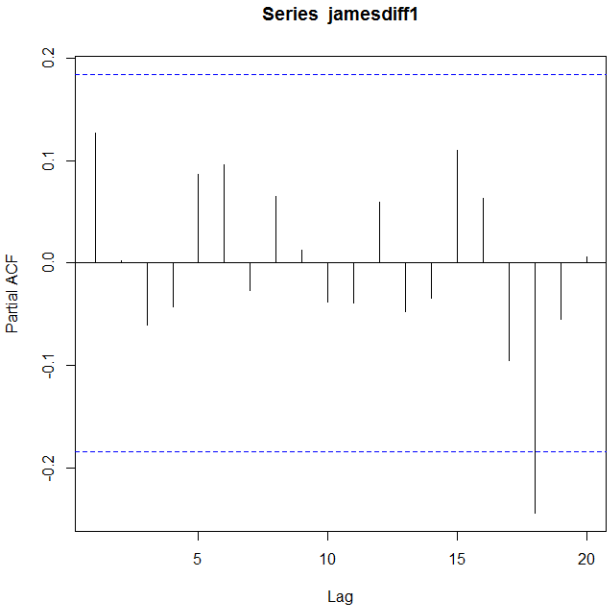

Figure 5: Partial Autocorrelation for the name James

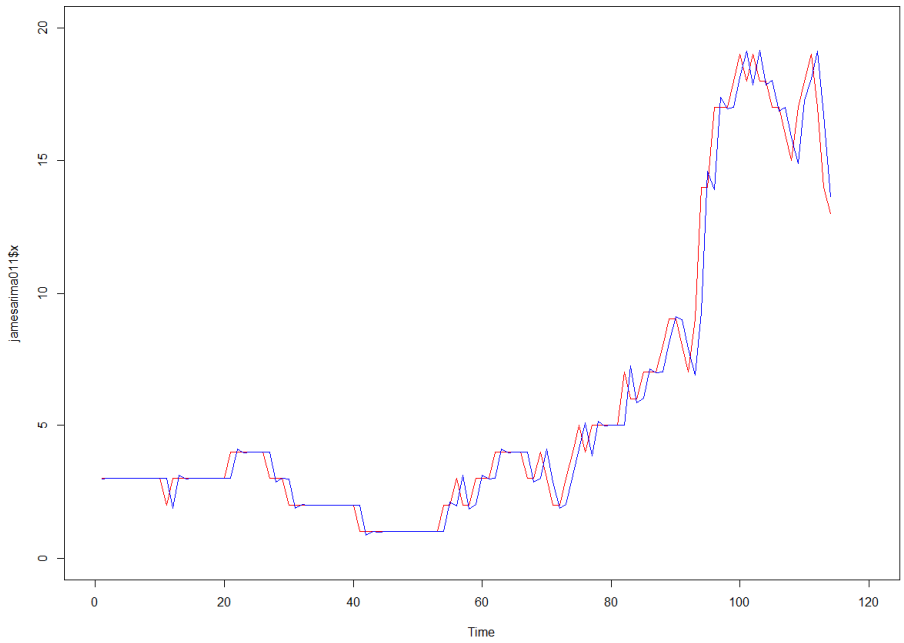

Figure 6: ARIMA(0,1,1) model (blue) for James compared to actual data (red)

Figure 6 shows the ARIMA(0,1,1) model compared to the actual ranks from 1900 to 2013 for the name James. The time series is indeed quite close to the actual data. The model is also parsimonious, as it uses as few parameters (1) as possible to closely model the data (along with using a first difference to create a stationary series). The actual model used is given by:

arima(x = jamests, order = c(0, 1, 1))

Coefficients:
ma1
0.1273
s.e. 0.0891

sigma^2 estimated as 0.8347: log likelihood = -150.14, aic = 304.28

which translates to the equation

$$Y_t - Y_{t-1} = e_t - 0.1273e_t - 1$$

where the $Y_t s$ are the rankings of the names at times t and t-1, respectively, and the $e_t s$ are the residuals at times t and t-1, respectively.

This model can be used to forecast future popularity of the name James. Using the ARIMA(0,1,1) process to forecast 5 future values results in:

Point	*Forecast*
115	*12.91696*
116	*12.91696*
117	*12.91696*
118	*12.91696*
119	*12.91696*

where the points 115-119 represent the years 2014-2018. Unsurprisingly, since the model is an MA(1) process, it reverts back to the mean quite quickly (the value for 2013 is actually a forecasted value and so it is the only one affected by the MA(1) process). Although of limited use for forecasting, the model does not provide a completely unreasonable forecast of near-future values.

In order to ensure that the model is appropriate, the residuals should be analyzed to ensure that they are a white noise process. A Box-Ljung test is run to determine this.

> Box.test(jamesforecast$residuals, lag=40, type="Ljung-Box")

Box-Ljung test

data: jamesforecast$residuals
X-squared = 28.7087, df = 40, p-value = 0.9081

With such a high p-value, the null hypothesis cannot be rejected and the residuals must be stationary. The ARIMA model is in fact useable.

Model for the name Abigail

The first difference of the popularity of the name Abigail was created, but the autocorrelation function showed no noticeable trends. While this may be indicative that no time series exists, it seems quite likely that this is not true. The second difference, as shown in Figure 7, is taken to see if a time series can be constructed. Although this may lead to over-differencing, it is possible that the second difference will show a trend.

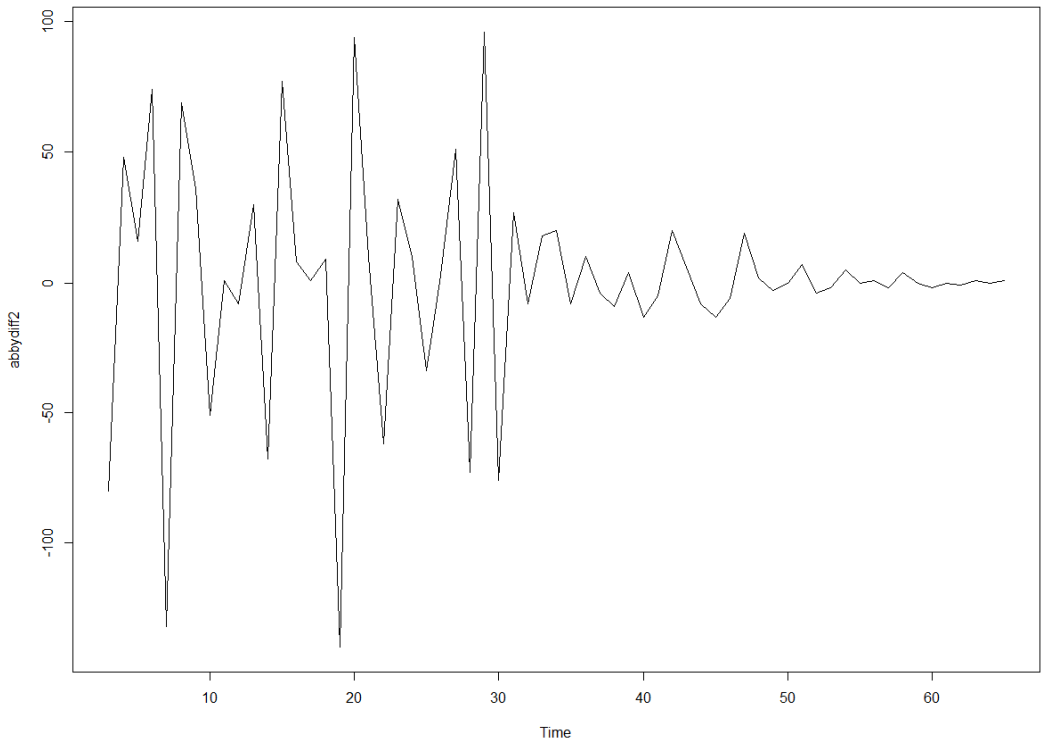

Figure 7: Second difference of popularity of name Abigail, by year

For the original data series:

> Box.test(abbyts,lag=40)

Box-Pierce test

data: abbyts
X-squared = 612.245, df = 40, p-value < 2.2e-16

Just like with the data for the name James, a Box-Pierce test was run to determine stationarity of the data. Similarly, the Abigail data is clearly not stationary as the p-value is so low that we must reject the null hypothesis.

For the second differences:

> Box.test(abbydiff2,lag=40)

Box-Pierce test

data: abbydiff2
X-squared = 54.2768, df = 40, p-value = 0.06543

For the second difference, although the p-value is close to .05, it is still above it and we cannot reject the null in favor of the alternative; the series is stationary.

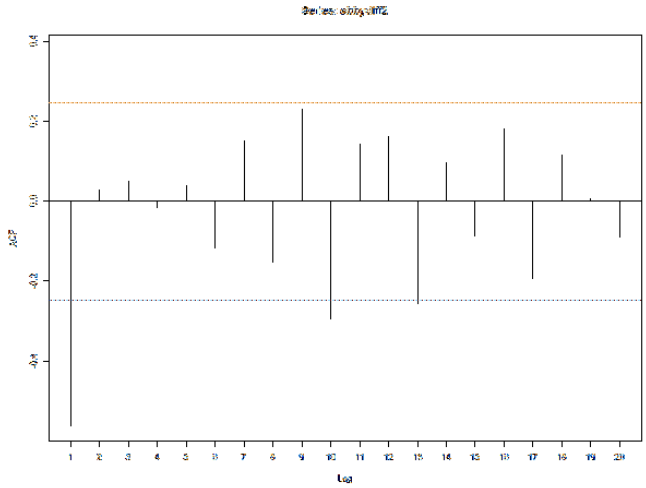

Figure 8: Autocorrelation for the name Abigail

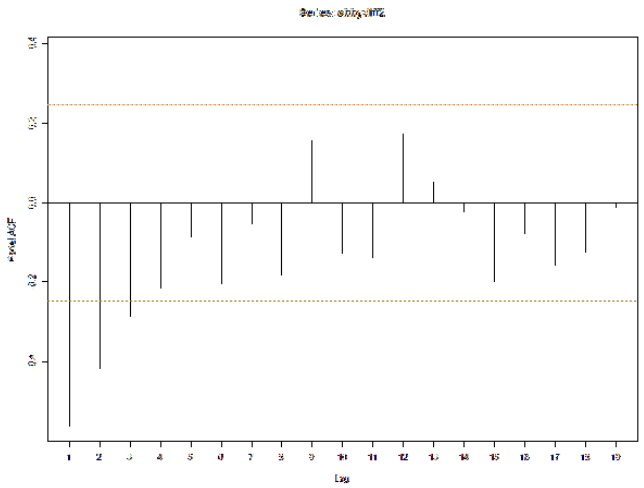

Figure 9: Partial Autocorrelation for the name Abigail

Correlograms and partial correlograms were again used to determine what the appropriate ARIMA model would be for the Abigail data. Here, unlike for James, the choice was not quite so obvious (Figures 8 and 9). The correlogram is likely indicative of an MA(1) relationship, though the partial correlogram shows that the process may be more complicated. For the sake of comparison, three different models were constructed: an ARIMA(0,2,1), and ARIMA(1,2,0), and an ARIMA(1,2,1). They are, in order:

> Arima(abbyts, order=c(0,2,1))
Series: abbyts
ARIMA(0,2,1)

Coefficients:
ma1
-0.9210
s.e. 0.0634

sigma^2 estimated as 835.5: log likelihood=-302.27
AIC=608.54 AICc=608.74 BIC=612.82

> Arima(abbyts, order=c(1,2,0))
Series: abbyts
ARIMA(1,2,0)

Coefficients:
ar1
-0.5852
s.e. 0.1041

sigma^2 estimated as 1194: log likelihood=-312.78
AIC=629.56 AICc=629.76 BIC=633.85

> Arima(abbyts, order=c(1,2,1))
Series: abbyts
ARIMA(1,2,1)

Coefficients:
ar1 ma1
-0.2006 -0.8635
s.e. 0.1425 0.0949

sigma^2 estimated as 812.1: log likelihood=-301.3
AIC=608.59 AICc=609 BIC=615.02

All of these ARIMA models produced a graph that was very close to that of the actual data. To decide which one was the most appropriate, the next five values in the series were forecast with each ARIMA model. The results made more sense for some models than for others. In the same order, the forecast values are:

ARIMA(0,2,1)
66 3.97664414
67 -0.04671173
68 -4.07006759
69 -8.09342346
70 -12.11677932

ARIMA(1,2,0)
66 8.414769
67 9.172033
68 9.728858
69 10.402986
70 11.008465

ARIMA(1,2,1)
66 6.1921120
67 4.9474764
68 3.5898543
69 2.2548969
70 0.9153931

Here, the points 66-70 represent the years 2014-2018. For the forecasts, the ARIMA(0,2,1) series makes no sense as it's predicting negative future values (which, of course, we cannot have). The other two models are providing sensible forecasts. The model settled on was the ARIMA(1,2,1) model, as the original thought was that an MA(1) process was part of the overall ARIMA model (though the other model seems perfectly sensible as well). This ARIMA model is written as

$$Y_t - 2Y_{t-1} + Y_{t-2} = -0.2006(Y_{t-1} - 2Y_{t-2} + Y_{t-3}) + e_t - 0.8635e_{t-1}$$

Finally, a Box-Ljung test is run to ensure the residuals are stationary.

> Box.test(abbyforecast$residuals, lag=40, type="Ljung-Box")

Box-Ljung test

data: abbyforecast$residuals
X-squared = 24.1991, df = 40, p-value = 0.9771

The p-value above .05 means the null hypothesis should be rejected and the residuals are stationary. The ARIMA model is appropriate. It is shown together with the actual data in Figure 10.

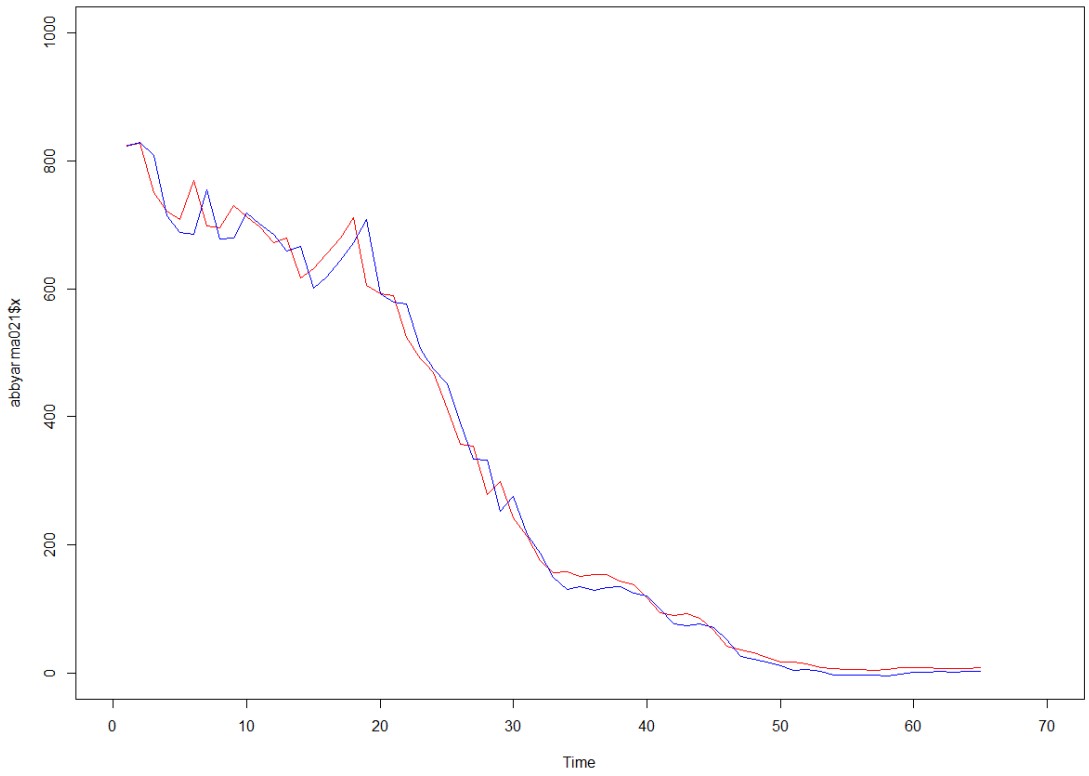

Figure 10: ARIMA(1,2,1) model (blue) for Abigail compared to actual data (red)

Conclusion

The ARIMA models create plausible models for approximating actual values and forecasting future values for both the names Abigail and James. Despite the fact that both names follow different patterns, they can both be modeled with ARIMA time series. As expected, the models are quite different, but this is because the names are quite different in their historical popularity.