*Time Series Analysis of Major League Baseball Annual Home Run and Attendance*

By Won Ha Gonzalez

**I. Introduction**

The United States has been blessed over the years with being privileged enough to be able to spend time consuming many types of entertainment products. Among the many options for entertainment that has engrained itself in the American culture is sports. One particular sport has the honor of being titled “America’s pastime”, baseball.

Baseball has a long history in the United States with professional leagues developing in the mid-1800s and continuing to the present. The two leagues(the National League and the American) that make up the premier professional baseball league in the US, Major League Baseball (MLB) were established in 1876 and 1901 before merging in 1903. Additionally, the nature of the game means that each play that occurs in baseball is a set play. The start-stop nature of the sport naturally allows followers of the game to create and maintain statistics related to the game. Because of this, baseball has a rich history of statistics gathering, making it an ideal venue for all sorts of statistical analysis.

Much of the focus of statistical analysis in MLB naturally is built around the question of how to build winning baseball clubs. While, this is clearly important in terms of building an entertaining product and understanding resource allocation, there is much less discussion and work on analyzing time series trends in MLB. By analyzing time series data, MLB would be able to gain understanding of the nature of key statistics over time. This would allow MLB to project into the future and give them a basis of understanding to plan for their future needs.

In this paper a time series analysis of two long kept variables will be analyzed, home runs and attendance records. While the author has general understanding of trends in both areas, a naïve approach will be taken in this paper. The reason for this is because there is little to no public information or models to project forward based on past values. Having a better understanding of how we can project forward based on past values could greatly help MLB to increase the bottom line.

**II. Data Characteristics**

There are now several sources for baseball statistics that span large swaths of baseball history. One of the preeminent collectors of baseball data is Sean Lahman. At his website, Lahman provides detailed data that covers offensive, defensive, post-season, and more specific statistics (ie. all-star, manager, etc.). These databases are updated annually, with the latest version (year 2019) detailing statistics from 1871 to 2018. The analysis does not take into account other leagues with well-known professional players (often African American) early in baseball history, such as the Negro leagues.

The data set used for this analysis is the Teams data sets. This data set includes the annual totals of various statistics from each team. It also includes descriptive information about each team for a given year. From these, I will use the Home Runs statistics and the Attendance records.

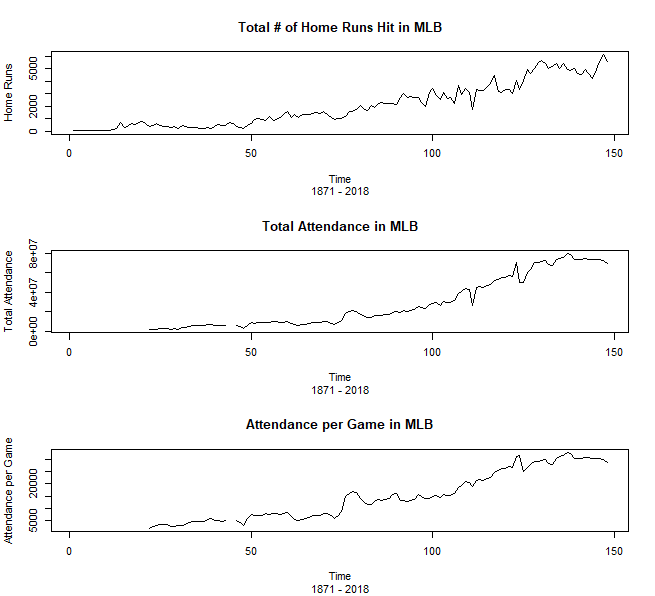
Because my analysis will focus on what is occurring throughout MLB baseball throughout the years, I combine to the annual data from each team into a single value to record for MLB for each individual year. The annual sum of home runs makes up a time series from 1871 to 2018, a span of 148 years.

Attendance records are not as immaculately recorded throughout baseball history. For approximately the first 20 years of data, attendance records are missing. After, this attendance records are much more reliable, though there are missing data points for the years 1914 and 1915. Upon conducting research, I was not able to determine the cause of the inconsistent attendance record keeping throughout the league. There were no obvious external reasons for why this would occur, and other records, like offensive and defensive statistics were kept well. The United States of America did not enter World War I until April 6, 1917 and attendance records were kept throughout the duration of the Spanish Flu pandemic in 1918.

Because of the inconsistent record keeping for attendance, the first 21 years in the data set will be dropped as there are not attendance records for these years. The values for 1914 and 1915 are imputed using the “imputeTS” package in R, specifically using imputation by ARIMA State Space Representation & Kalman Sm. (Moritz and Bartz – Beielstein, 2017). Therefore, the attendance time series spans from 1892 to 2018. Additionally, the attendance information was transformed into a per game rate making the time series analysis on MLB annual attendance per game.

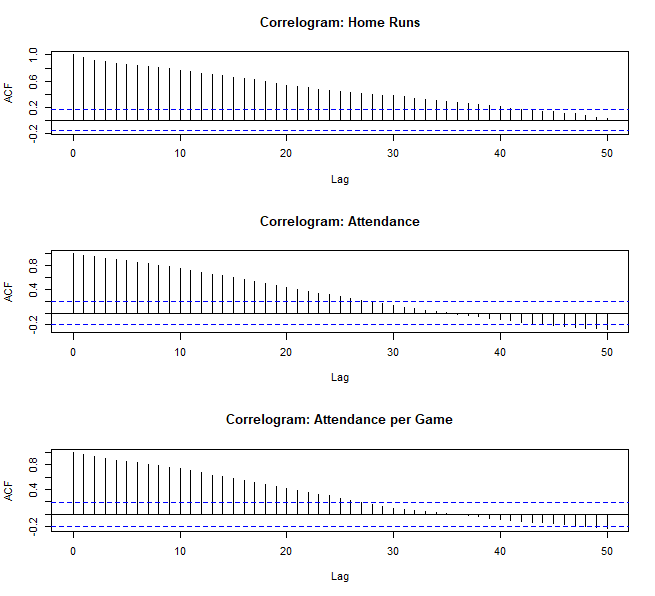
**III. Analysis and Model Selection**

In order to understand what is happening throughout the time span in question, I first create a line plot of each of the time series: home runs, attendance, and attendance per game. With each of these variables, there is an increasing trend. While growth in all three is slow towards the beginning of the time series, we do see that there is a steady, upward, additive trend. There is no obvious seasonality that appears in the line plots, though it cannot be ruled out.



*Fig. 1*: Line plots of annual figures for home runs, attendance, and attendance per game from 1871 to 2018.

I now use a correlogram plot in order to get a sense of how much autocorrelation there is in these time series. In each of the time series, there is slow dampening of the autocorrelation values. The correlogram for the home run time series shows a very slow dampening effect, where the autocorrelation is significant until about lag 40 (k = 40). This may be an indication that this time series is a candidate for fractionally differenced ARIMA model (ARFIMA).



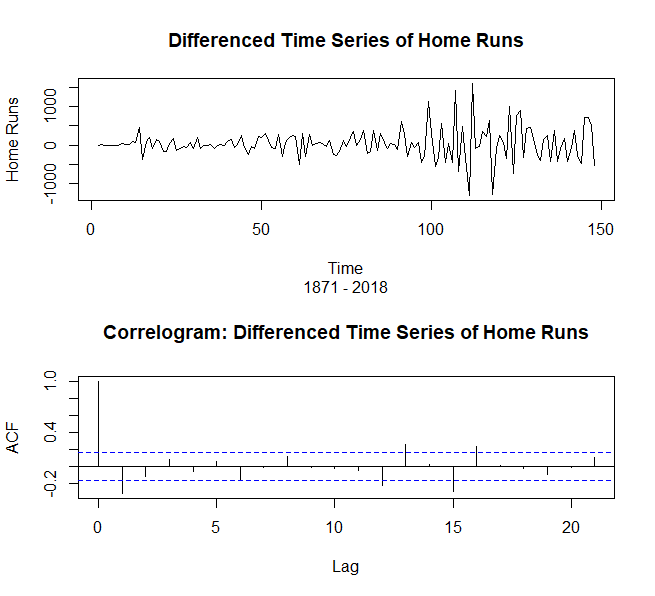
*Fig. 2*: Correlograms to display autocorrelation of time points in the home run, attendance, and attendance per game time series.

Because we know very little about the behavior of these time series, this analysis will take a naïve approach to the analysis. By this, I mean that we start from simple time series modeling approaches to determine if they can account for the behavior that appears in the time series. If it fails to do so, I will move on to more complex methods until all the variance can be explained by the model, and we are left with white noise.

**III. ii.) Differencing**

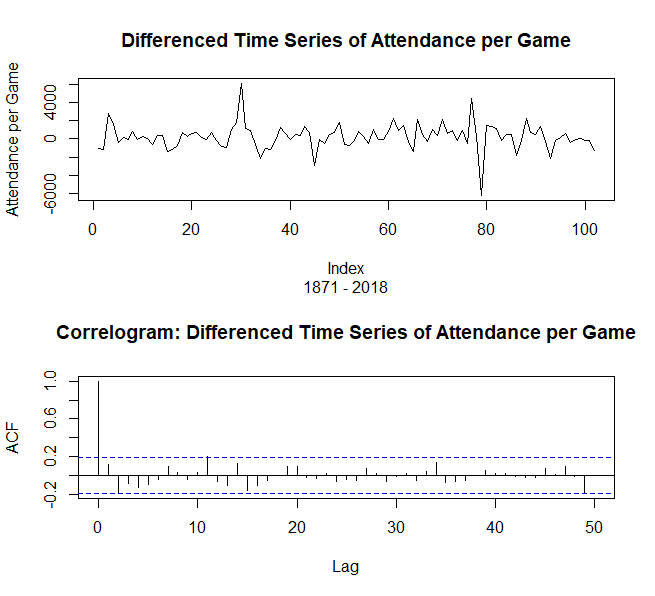
I start with differencing of the times series of the home runs and attendance per game time series. The reason for differencing is to eliminate the any trend or seasonality. Trends needs to be removed as the mean may vary over time and seasonality needs to be removed as seasonality allows for changing variance over time. The removal of both trend and seasonality allows for the time series to be stationary, and thus easier to model.

In applying the differencing to the time series of home runs and plotting the resulting time series and correlogram, the mean appears to be stable. However, the variance is not stable over time. The correlogram also gives indications that more complex methods for removing trend and seasonality are necessary, as we see that the first lag is significantly correlated and we see growing significance as we look further to the right of the plot.



*Fig. 3*: Top – Line plot of differenced time series of home runs; Bottom – Correlogram of differenced time series of home runs.

In applying the differencing method to the attendance per game time series, the plots indicate that attendance per game may be model using less complex methods. There may be some changes in the mean as we see a large spike early in the line plot and a large valley later in the line plot. The correlogram indicates that differencing removed trend and seasonality from the time series, and we are left with a series that does not have autocorrelation.



*Fig. 4*: Top – Line plot of differenced time series of attendance per game; Bottom – Correlogram of differenced time series of attendance per game.

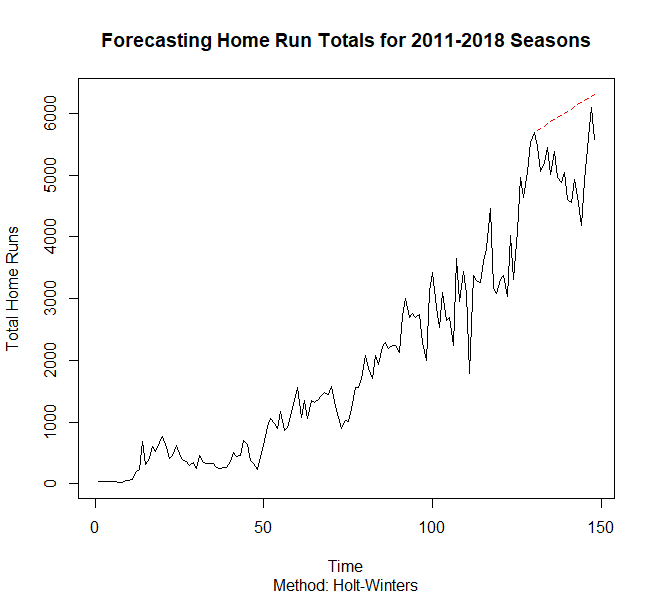
This analysis will now focus on modeling the time series behavior of the home run time series given that early indications are that this will require more complex modeling to explain the variation within the time series. I will return to the analysis of the attendance per game time series following the modeling of the home run time series.

**III. iii) Home Runs**

**III. iii a.) Holt – Winters Model for Annual Home Runs**

The first approach taken is the use of the Holt – Winters method. This method attempts to account for the level, trend, and seasonality by using an exponentially weighted smoothing parameter. In order to see how well this model estimates annual homerun totals, I used the first 130 years in the time series to measure the level, trend, and seasonality. Regardless of whether the seasonality setting is set as “additive” or “multiplicative”, the coefficient estimates for the level and the slope are the same.

I then use these estimates to predict the next 18 years and overlay this on the original home run time series to see how well the Holt – Winters model does in projecting annual home run totals. The Holt – Winters projection does a nice job of capturing the overall trend from the last time observation in the 130 years time series. However, it does not capture the seemingly randomness that could be attributed to seasonality. Because of this, I am moving on to more complex models as the Holt – Winters projection is not adequate for capturing the true nature of this time series.

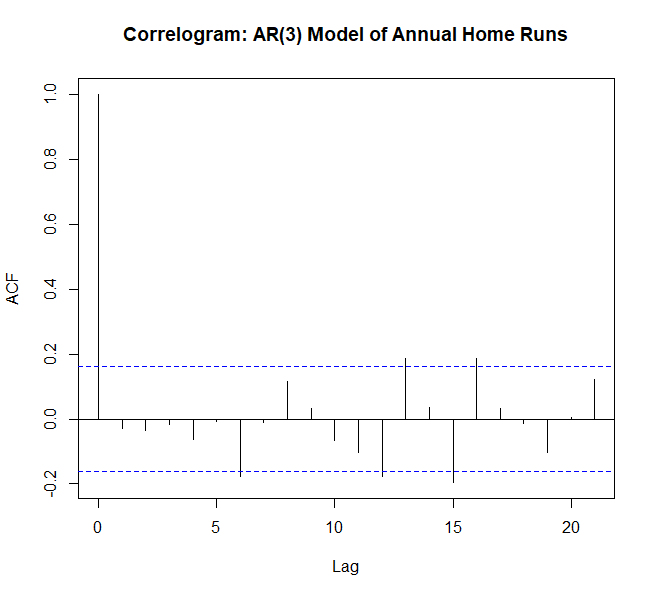


*Fig. 5*: Holt – Winters projections for last 18 years overlaid on original home runs time series. Prediction of 2001 – 2018 in red dotted line.

**III. iii b.) AR Modeling**

Assuming that the home run time series is stationary (stable in the mean and variance), an Auto Regression (AR) model should be able to capture the behavior that exists in the time series. This model is a regression of a time point on past terms of the same series.

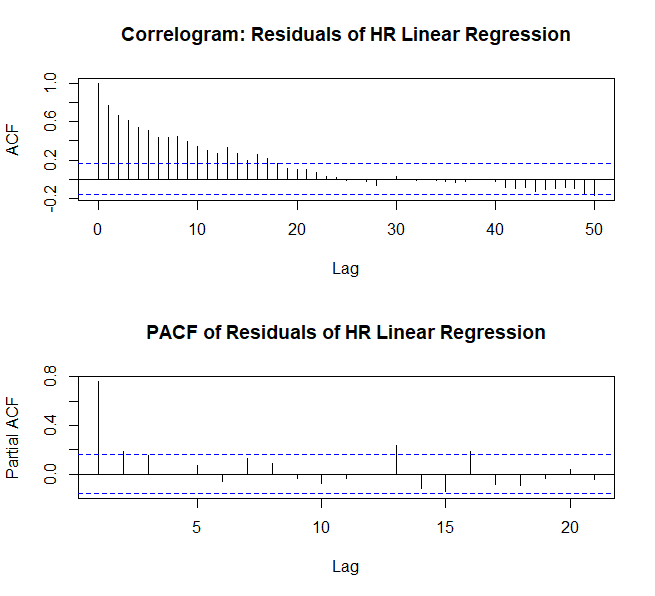
Here, the home run time series does not fit will with AR modeling. The technique states that an AR(3) model adequately explains the variation that appears in the home run time series. However, the correlogram is problematic. While the early lag points are insignificant, the autocorrelation seems to grow in a pattern is one looks at later lag values. This pattern indicates that this model is not appropriate for this home run time series.



*Fig. 6*: Correlogram of an AR(3) model on the home run time series. Early lag points show no significant autocorrelation, but later lag points increase slightly over time. This indicates that a more complex model is needed.

**III. iii c.) Regression Model**

The next method to try modeling the home run time series will be a linear regression model. The idea here is to treat time as the predictor variable against the annual home run totals. For the time being, I will use a simple linear regression, as the Holt – Winters projection indicated that the home run time series had a clear linear trend. Upon completion of the linear regression, the correlogram and the partial autocorrelation function shows that the linear regression model is not adequate for explaining the variations that exist in the home run time series. The correlogram of the residuals again shows a slow dampening of the autocorrelation values. Additionally, behavior seen in the partial autocorrelation is odd. The correlation of the residuals on the next lagged point is inconsistent throughout the plot. This is concerning, so the analysis shall move on to the next method.

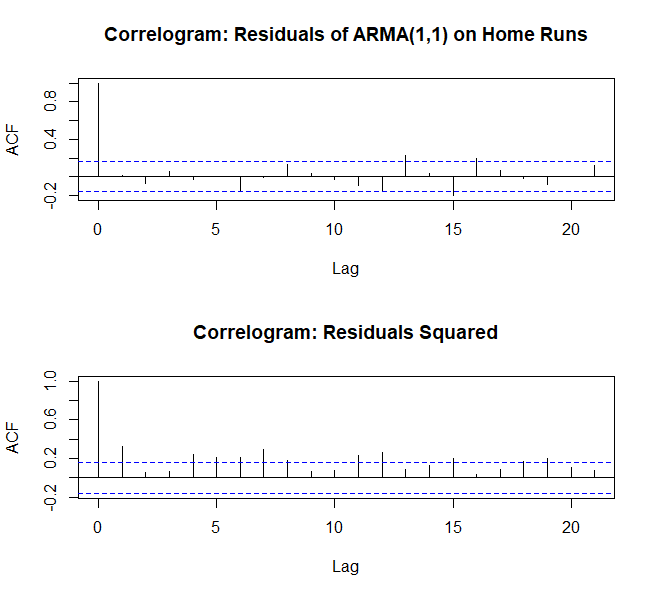


*Fig. 7*: Top – Correlogram of the residuals of a linear regression model on the home run time series shows slow dampening. Bottom – Partial autocorrelation function plot shows inconsistency throughout the plot.

**III. iii d.) ARMA Modeling**

Because the time series may have a moving average, the next method chosen is the ARMA modeling technique. When implementing this model, I commanded R to perform the model with several different parameters against the residuals of the linear regression model above. In choosing the AR(p, q) model, I set the function to restrict the selection of p(0 to 4) and q(0 to 2). I selected the best ARMA model by using the Akaike Selection Criteria (AIC). This is a selection criterion that penalizes adding more parameters to a model, thus lower values are better scores. As a side note, in selecting a final method as a best model, this paper will also use the AIC score to determine the final model.

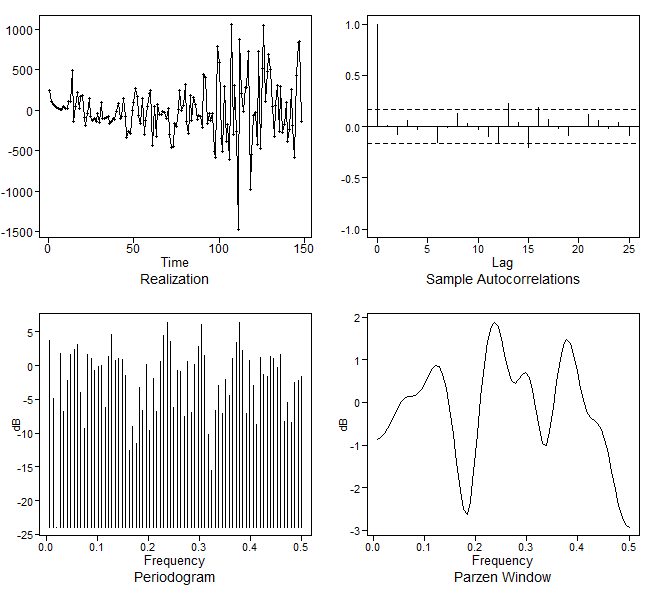
The ARMA model selected to best explain the variations that occur in the home run time series is an ARMA(1, 1). To determine with confidence if this is the final model, the correlogram is again plotted. The correlogram seems to indicate that the variations have been explained away by the model, though the continued slowly increasing autocorrelations are still of concern. A correlogram is plotted using the squared residuals of the ARMA(1, 1) model to check to see if there is conditional heteroskedasticity. It the wave pattern of increasing and decreasing autocorrelation on the residuals squared is clear indication of conditional heteroskedasticity.



*Fig. 7*: Top – Correlogram of the residuals of an ARMA(1, 1) model from regression residuals. Bottom – Correlogram of the squared residuals of an ARMA(1, 1) model.

Additionally, a frequency/spectral analysis is conducted for this ARMA model, using the residuals from the linear regression model. This analysis is used to learn more about the frequency that occurs in the home run time series, and to verify the findings in the ARMA(1, 1) modeling. This is particularly useful as this analysis started with a naïve approach that had no understanding of the frequency that exists in the time series. The spectral analysis also tests several parameters for (p, q) and again found that the best performing ARMA model is an ARMA(1, 1) model.

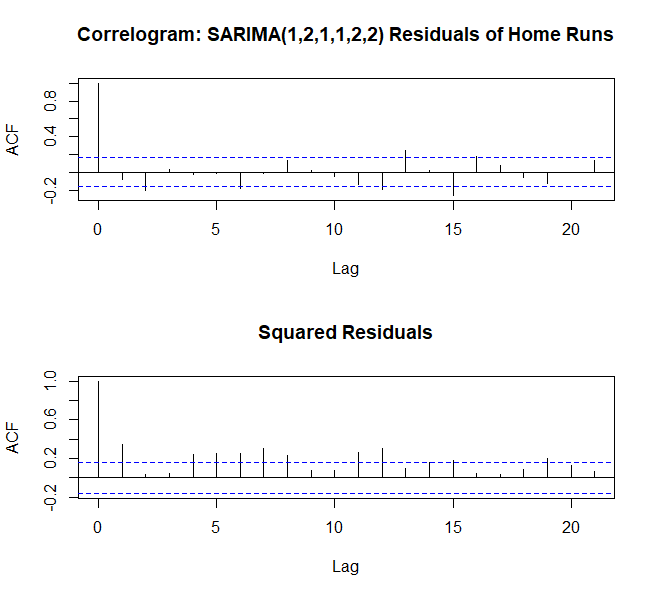
Plotting the residuals of the spectral ARMA(1, 1) model, the Parzen window in Figure 8 shows multiple peaks and a deep valley, which shows the combination of the effects of each parameter. The fact that there are multiple peaks is a worrying sign that the ARMA(1, 1) mode does not adequately explain the variance in the home run time series. To test whether the spectral analysis with an ARMA(1, 1) model explains the variation that exists, a Llung – Box test is run against the residuals estimations of the ARMA(1, 1) model. P-values under 0.05 indicate that there is still variation that needs to be modeled for. The Llung – Box test gives a p-value of 0.008, indicating that a different model needs to be used.



*Fig. 8*: Spectral Analysis of the residuals of the ARMA(1, 1) model. Patterned growth in the correlogram and the multiple peaks in the Parzen window give concerns that this model is not adequate.

**III iii e.) SARIMA Modeling**

It appears to be evident at this point that the home run time series is not stationary, so non-stationary modeling techniques should be used in order to explain the behavior in the time series. A seasonal ARIMA model is now implemented to attempt to address the non-stationarity and seasonal effects that seem to exist in the home run time series. As with the ARMA modeling, a function is created to try several parameters for (p, d, q, P, D, Q). A maximum value of 2 was set for each of these parameters and the best model was the model that had the lowest AIC score. The best SARIMA model was a SARIMA(1, 2, 1, 1, 2, 2). The correlograms of the residuals of the model and the residuals squared are again checked to see if there is any leftover unexplained variation in the residuals. Just as seen with the correlograms of the ARMA model, a slowly growing significance appears in the correlogram of the residuals, and a wave pattern appears in the correlogram of the residuals squared. Another approach needs to be taken to address the behavior in this time series.

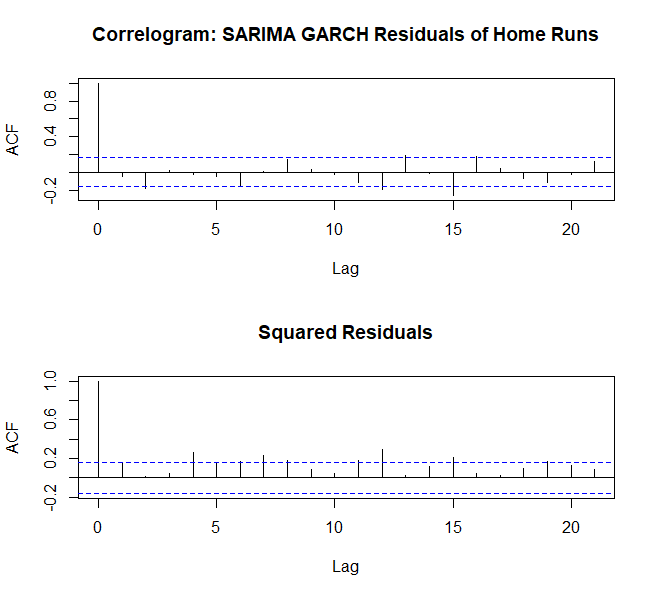


*Fig. 9*: Top – Correlogram of the residuals of a SARIMA(1, 2, 1, 1, 2, 2) model. Bottom – Correlogram of the squared residuals of a SARIMA(1, 2, 1, 1, 2, 2) model.

**III. iii f.) SARIMA GARCH Modeling**

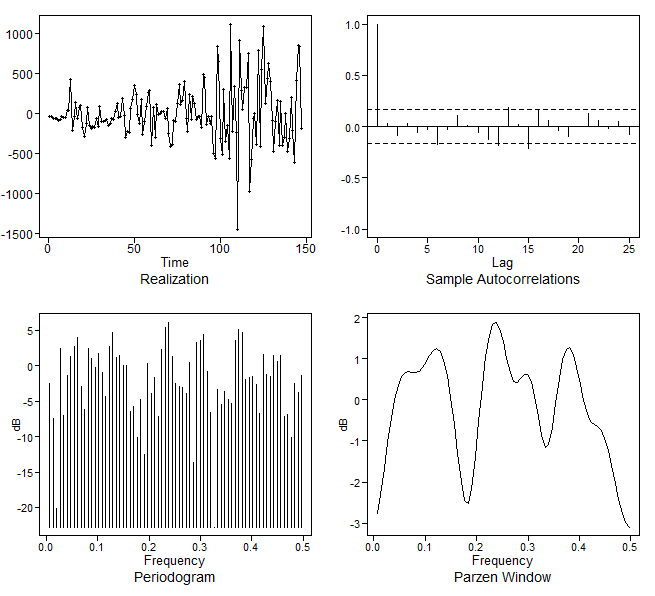
The wave pattern apparent in the correlogram of the squared residuals in both the ARMA and SARIMA model indicates that the home run time series has conditional heteroskedasticity. This can be addressed using GARCH modeling techniques. By eliminating the conditional heteroskedasticity, it is likely that there will only be white noise left over, thus explaining all the variation that exists in the series.

A GARCH model is implemented using the residuals from the SARIMA model above. The residuals of this GARCH model are then used to gauge adequacy of this model. While, there appears to be some improvement in the correlograms of the residuals in the SARIMA – GARCH model, the same general patterns emerge.



*Fig. 10*: Top – Correlogram of the residuals of a SARIMA – GARCH model. Bottom – Correlogram of the squared residuals of a SARIMA – GARCH model.

I now check this against spectral/frequency analysis that does not assume stationarity. I start by taking the difference of the residuals of the linear regression (making d = 1). The spectral analysis selects values from (p, q) to be (0, 1). This means the spectral analysis here has a model of ARIMA(0, 1, 1). This does differ from what we saw above. However, we see similar patterns in the plots of the correlogram. Also, has multiple peaks and valleys, indicating that there is more going on in this time series. The Llung – Box test verifies this, giving a p-value of 0.007. This means there is more variation to be explained.

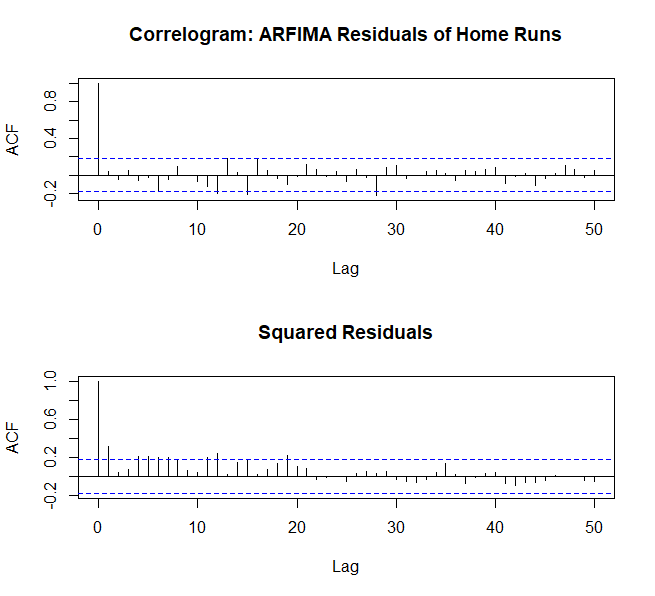


*Fig. 11*: Spectral Analysis of the residuals of the ARIMA(0, 1, 1) model. Patterned growth in the correlogram and the multiple peaks in the Parzen window give concerns that this model is not adequate.

**III. iii g.) Fractional Differencing Modeling/ ARFIMA – GARCH**

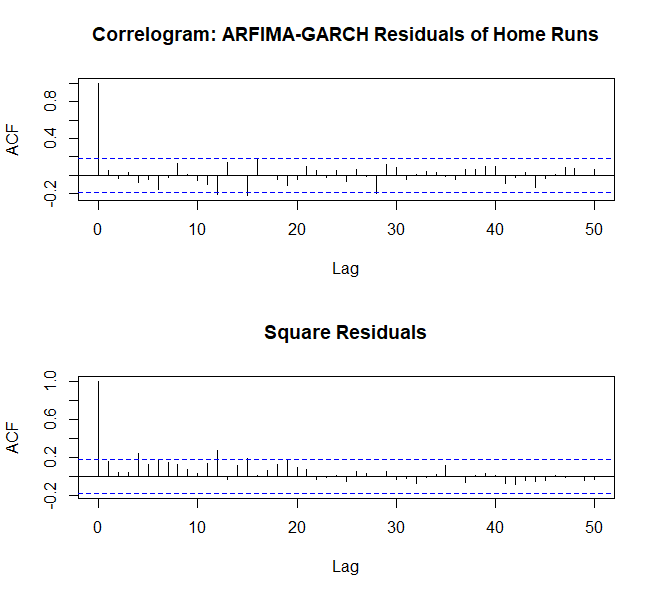
By this time, I suspect that the home run time series has trends, seasonality, and is non-stationary. However, taking the difference in multiple methods (d = 1) did not properly address the variability that appears in this time series. This indicates that the differencing occurring at a slower rate that cannot be picked up by modeling techniques that use differencing of 1. Fractional differencing may address the problem here.

Implementing a function to take the fractional difference of the original home run model, I then fit an ARMA model on the residuals of the fractional differencing. After checking for multiple values in the parameters (p, q), the best model is an ARFIMA(1, 0., 1). Checking the correlograms for the ARFIMA models shows a great improvement over the SARIMA – GARCH model. However, we do continue to see the wave pattern in the correlogram of the squared residuals. To address this, a GARCH model is implemented on the residuals of the ARFIMA model.



*Fig. 12*: Top – Correlogram of the residuals of an ARFIMA model. Bottom – Correlogram of the squared residuals of an ARFIMA model.

Once implementing the GARCH model on the ARFIMA, there is even greater improvement. Visual inspection on the correlograms of the residuals of the ARFIMA – GARCH model show that all autocorrelation has been accounted for. Additionally, the correlogram of the square residuals now show that conditional heteroskedasticity has also been modeled out of the home run time series. A Llung – Box test on the residuals of this model also verifies that we are left with white noise, with a p-value of 0.079.



*Fig. 13*: Top – Correlogram of the residuals of an ARFIMA – GARCH model. Bottom – Correlogram of the squared residuals of an ARFIMA – GARCH model.

**III. iii h.) Home Run Time Series Model Selection**

Several time series modeling techniques were used throughout the course of this analysis on the home run time series. Along the way, we visually inspected the residuals of each method and determined whether other methods were needed to find good fits for the time series. Additionally, AIC scores from each model have been recorded to compare the many models. AIC penalizes added parameters, so low scores are considered better models. Table 1 provides the AIC values for each model. It is clear that our model was improving along the way and that the final model tested, the ARFIMA – GARCH model, provided for the best fit with the home run time series.

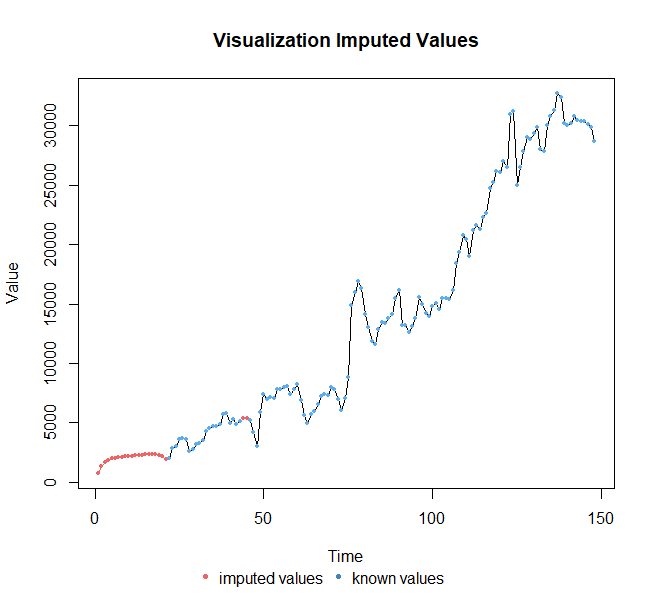
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Models: | Linear Model | AR(3) | ARMA | SARIMA | SARIMA-GARCH | ARFIMA | ARFIMA-GARCH |
| AIC | 2306.504 | 2163.371 | 2162.824 | 2137.78 | 2132.806 | 1756.905 | 1727.594 |

*Table 1*: Akaike Information Criteria scores for the various methods used to fit the home run time series.

The final model for the home run time series is ARFIMA – GARCH (1, 1, 1):

**III. iv) Attendance Per Game**

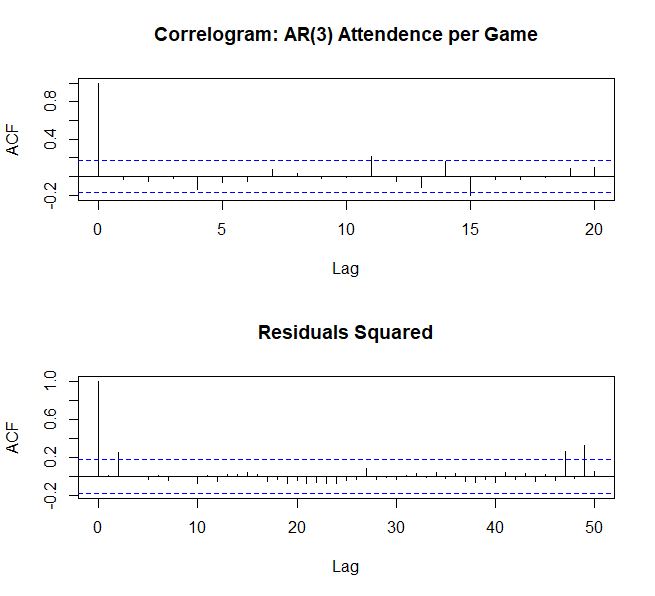
As stated above, the attendance records are not complete, as was the home run records. Because of this, I imputed missing values for the years 1914 and 1915. The first 21 years were missing. It was determined that having a long period be imputed may negatively affect the analysis, so it was not included in the analysis.



*Fig. 14*: Attendance per game time series with known values and imputed values. This analysis moved forward with the elimination of the first 21 time points, which were all missing.

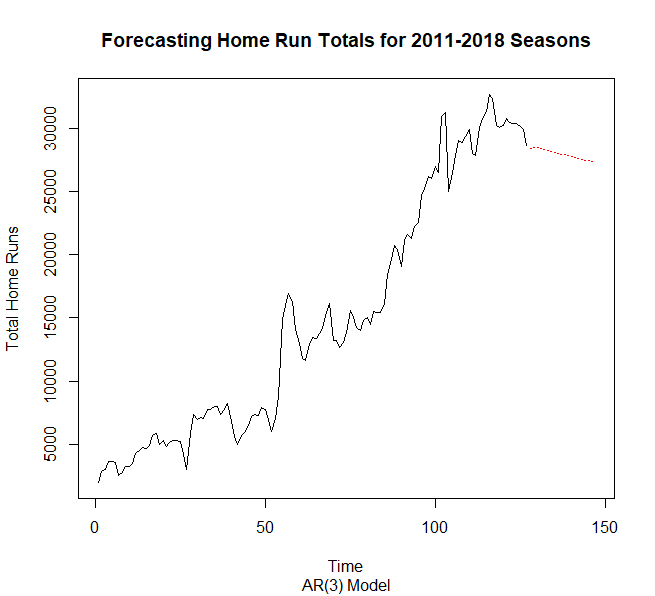
**III. iv a.) AR Modeling**

The analysis of the attendance per game will follow a similar path as was followed in the home run time series, starting with the AR model. In fitting an AR model to the attendance per game time series and analyzing the correlogram outputs, this seems to be a good fit. The correlogram shows that there is no autocorrelation left in the residuals of the model. The correlogram of the squared residuals also shows no autocorrelation, indicating that there is no conditional heteroskedasticity.



*Fig. 15*: Top – Correlogram of the residuals of an AR(3) model. Bottom – Correlogram of the squared residuals of an AR(3) model.

To test how well the model fits, I now forecast for the next 20 years using the fit from the AR(3) model. The forecast looks fairly good given where the time series left off. However, as can be expected from AR models, the forecast tends to the mean of the time series.



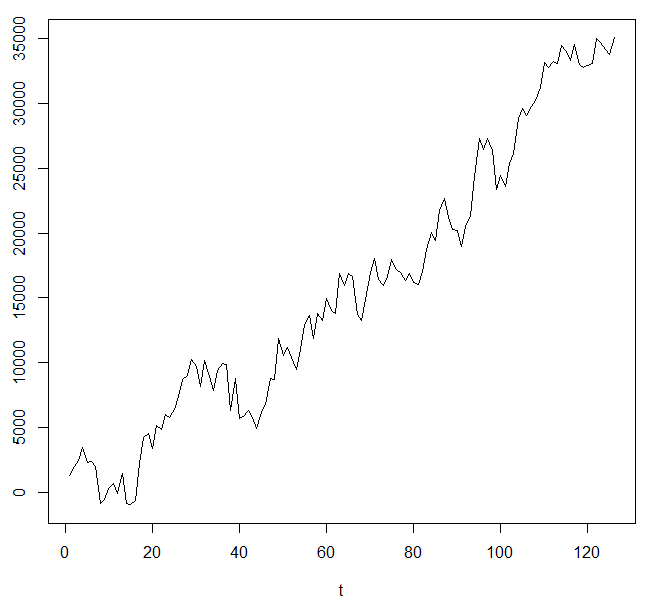
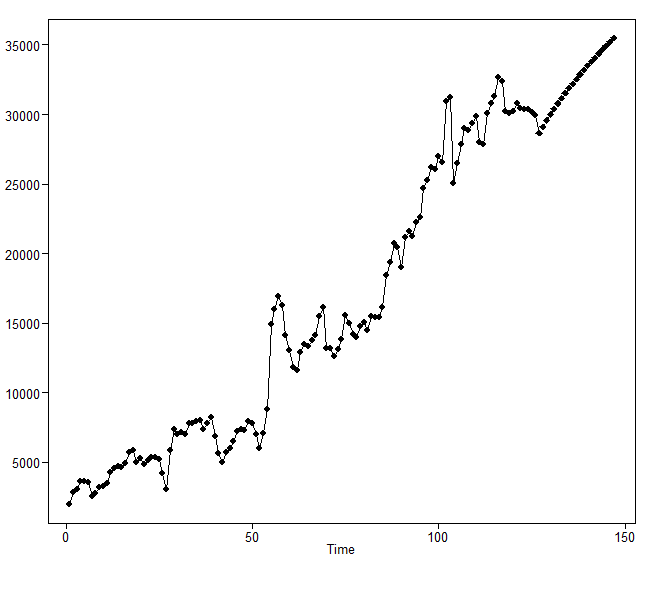
*Fig. 16*: Attendance per game time series with known values and imputed values. This analysis moved forward with the elimination of the first 21 time points, which were all missing.

**III. iv b.) Spectral/Frequency modeling**

I will now use spectral analysis to back what we have found in the AR(3) modeling above. While taking a naïve approach to analyzing both the home runs and attendance data set, it seems to be clear that modeling either should not be so straight forward. Home runs were particularly difficult to model, given the very random nature of baseball. However, sales figures should be a little less random. That being said, it is still likely that attendance records contain both signal and noise. From Figure 16, we see that it is difficult to see any clear seasonality. It is clear there is a trend, but it is also likely that there is some change in the mean and variance. Because of this, I will use a signal + noise approach to analyzing the attendance per game time series.

When implementing this model technique, R was tasked with selecting the best *p* parameter from 0 to 6. The signal plus noise analysis model selected *p* = 1 as the best model fit. Continuing with an AR(1) signal plus noise fit, the coefficient of *p* is pulled from the summary and a model can be specified. In order to check the accuracy of the model, we again check the correlogram of the residuals of the model and use the Llung – Box test. In both, we see that variation is fully accounted for.

Additionally, a simulation is generated based on the parameters of the model to see if it generates a similar looking time series. The resulting plot is very close to what appears in the imputed time series.

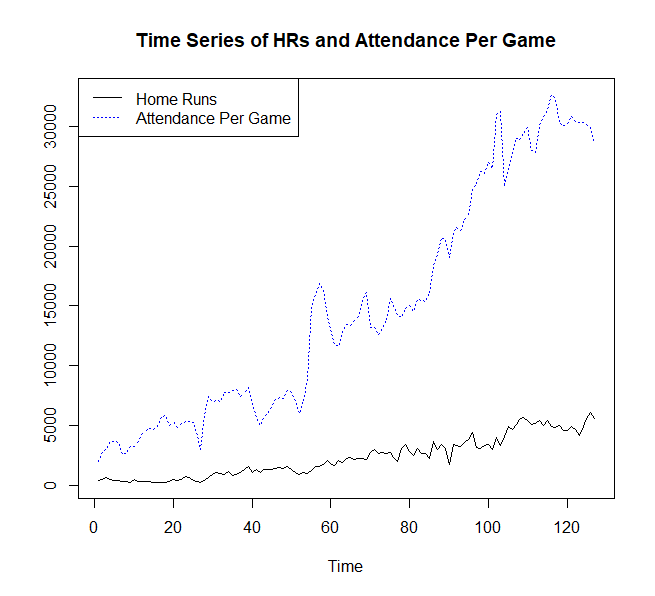
*Fig. 17*: Top – Simulated attendance per game time series using signal plus noise spectral analysis. Bottom – Projection of attendance per game for the next 20 years.

Following the simulation, a projection was made 20 years into the future based on the signal plus noise model. The resulting projection does show a trend of continued growth. The projection also has a slight curve to it, indicating that the growth in attendance per game does slow down slightly with each passing year.

The final spectral analysis AR(1) model is:

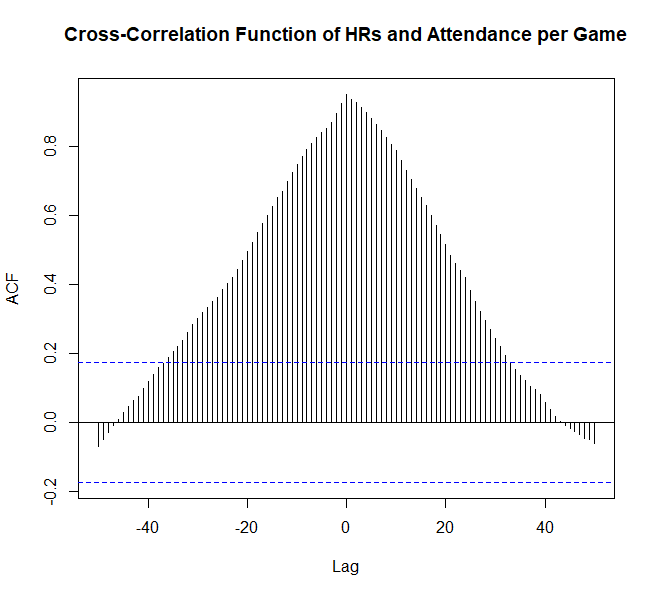
**III. v) Vector Autoregressive (VAR) analysis**

It is often said in most sports that offensive output drives entertainment. We can take attendance records as an indicator for how popular is game is. Because this analysis is already ready using the most entertaining offensive play in baseball as a variable of interest, we can use vector autoregressive analysis to see if increases in the number of home runs hit in MLB has any effect on the attendance per game records. In a plot with both variables, it is clear that both have increased steadily over time.



*Fig. 18*: Time series of both annual home runs and attendance per game.

The next step is to then find the correlation between the two variables across the time series. The correlogram of the both time series give clear indication that the number of home runs hit is highly correlated with number of people in attendance per game. The correlation is also slowly dampening, which indicates that the value of the home run totals has a long lasting effect on attendance numbers.

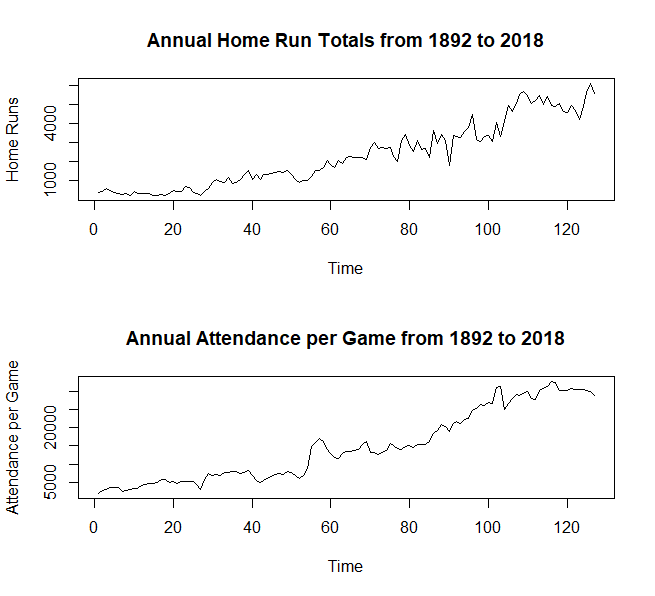
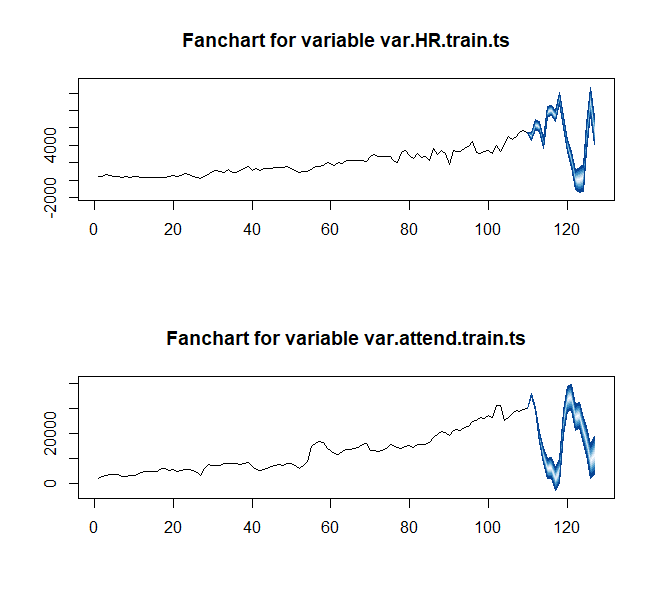


*Fig. 19*: Cross – correlation function of home runs as a leading indicator for attendance per game.

To help identify the relationship between the two variables, VAR analysis is implemented. The following analysis will use a portion of the home run and attendance time series, to predict the values of another portion of each time series. The training time series for both time series will be from 1892 to 2002, a span of 110 years. Using the training model, the next 17 years will be tested.

Allowing the “vars” package to select the number of lags, *p* = 35 is selected to tune the model fit. In Figure 19, the cross-correlation function loses significance between k = 30 and k = 40, so the model tuning estimation is good. The model fit with continue with a VAR(35).

Using a VAR(35) model was inadequate for the purposes of predicting each of the variables based on their relationship with each other. This indicates that the relationship is more complex than the VAR(35) model could account for. The home run predictions start off promisingly. However, the prediction for home run totals quickly becomes volatile, going from an extreme crash, to and extreme climb. The attendance per game predictions immediately become volatile.

*Fig. 20*: Left – Actual home run and attendance per game records for the years 1892 to 2018. Right – Using VAR(35) fit to predict the years 2003 to 2018. VAR(35) was inadequate in fitting complexity of correlation.

**IV. Conclusion and Potential Future work**

By starting with a naïve approach and working through various methods of analyzing time series data, this project was able to gain some insight into the nature of baseball, both in game play and in sales. The times series for the annual home run totals is quite complex, as would be expected for a particular type of play in a game very well known for its randomness. Ultimately, a fractionally differenced ARIMA – GARCH (or ARFIMA – GARCH) model was best able to account for the variation that existed in the time series; indicating that the it is non-stationary and is very autocorrelated, with conditional heteroskedasticity.

While there was great progress made in selecting the best model, this analysis was not able to implement prediction and projection analysis with the ARFIMA – GARCH model. Future work should include these types of analysis to check how well the model fits with the data and what can be learned projecting into the future.

The attendance per game was seemingly “less complex” to model. An AR(3) model was adequate to explain the variation of this time series. A spectral/frequency signal + noise analysis was also conducted and an AR(1) model was selected. Projection analysis of both models leave the verdict of which model performs better in the hands of each individual. Because there is a finite number of seats, there could be a natural tailing off of attendance, as projected by the AR(3) model. However, it is important to note that AR models often tend towards the means. The signal plus noise AR(1) model projected continued growth, with some slow flattening out. The conclusion here, is that the signal plus noise AR(1) model best approximate realistic attendance figures.

Future work on attendance data should include daily attendance figures both for individual teams and for MLB. The additional data should help to create more accurate models. Additionally, daily figures would allow for the analyst to start with an understanding of frequency. Current major league seasons have 162 games. Because this analysis had no information of any type of natural frequency that occurs in annual figures, no informed decision could be made on that aspect of the research and no adjustments could then be made on the analysis.

Additionally, a VAR analysis was conducted to determine the relationship between home runs and attendance per game. The cross – correlation function indicated there was cross correlation for a lag of up to 35 years. This immediately will cause some concern. The concern was then validated using prediction analysis. Using 110 years of annual figures as a training time series for both home runs and attendance per game, a VAR(35) model was fitted and predictions for the next 17 years were made based on the fit. The results of this analysis left much to be desired. Both home runs and attendance per game predictions were extremely volatile, though home run predictions looked promising for about 6 years.

Future consideration for multivariate analysis should look further into specific statistics and their relationship with attendance totals throughout the league. By knowing how the games have turned out, MLB and MLB teams would then be able to project forward on how many fans they can expect to attend games, making this economically vital analysis. Additionally, more complex models for may be necessary to model multivariate time series in baseball, particularly when looking at an often random statistic (like home runs) and a much less random figure (like attendance.

**References**:

Lahman, Sean. “Lahman’s Baseball Database.” *SeanLahman.com.* <<http://www.seanlahman.com/baseball-archive/statistics/>>

Moritz, Steffen and Thomas Bartz – Beielstein. “imputeTS: Time Series Missing Value Imputation in R.” *The R Journal*. Vol. 9, Iss. 1. June 2017.