

Yankees and Red Sox:

A Time Series Analysis of Win Percentage Correlations between Professional Baseball Teams

Grant Johnson, University of Louisville, Louisville, KY

ABSTRACT

Objective. The purpose of this study is to demonstrate how time series can be used to analyze the correlations of win percentages between the Red Sox and Yankees rival professional baseball teams.

Method. The data to be analyzed in this paper were collected from Baseball-Reference.com or were calculated from data from said resource. The data of primary interest include wins and win percentages for each team. Wins were projected as if each season consisted of the current standard of 162 games. The relevance of additional inputs in creating the time series model was investigated, including a study of head-to-head match-ups between the two teams and the sale of Babe Ruth from the Red Sox to the Yankees.

Results. The ratio of the Yankees winning percentage to the Red Sox winning percentage begins to more closely approximate 1.0 starting in 1968, paralleling a general historical trend of increased competitiveness in Major League Baseball. The Yankees and Red Sox have a sharp increase in wins in 1994 and 1995 respectively. These factors have continued to ensure that this rivalry is arguably the greatest in all of sports.¹ Both head-to-head match-ups between the teams and the sale of Ruth seem to help us obtain a more accurate model, but these results are tempered by a conflicting outcome when a hold-out sample is used.

Conclusion. The methods shown here can be applied to win percentages of professional baseball teams to analyze competitiveness, likelihood of rivalry and historical winning trends. This work could be extended to measure the effect of an acquisition of a player, a geographic move, a change in ballparks, etc.

INTRODUCTION

It is the purpose of this paper to analyze the correlations of win percentages between the Red Sox and Yankees rival professional baseball teams to determine if there are trends in win percentages between the teams and how they relate to historical competitiveness. We use time series to study such correlations and their implications

The rivalry between the Yankees and Red Sox professional baseball teams is considered by some to be the greatest rivalry in all of sports.¹ Historical aspects of this rivalry include the sale of Babe Ruth, nicknamed "The Bambino," from the Boston Red Sox to the New York Yankees in 1920 for the sum of \$100,000.² Ruth, who is considered the greatest baseball player of all time by the Society for American Baseball Research,³ went on to win four championship rings during his career with the Yankees. The Red Sox, on the other hand, went 83 years without winning a championship after the sale of Ruth – a drought that has been referred to as "The Curse of the Bambino."² With the Red Sox championship win in 2004, the Yankees conceded their monopoly on World Series championships.² The data reflect a surge of competitiveness to this storied rivalry as well as the entirety of Major League Baseball. We will look at the statistical significance of the Ruth sale. The importance of head-to-head match-ups between these fierce competitors will also be examined.

Time series methods show a continuing convergence of win percentages between the two teams along with a recent upward trend in win percentages for each team. This reflects the sustainability of the famous rivalry in recent years. That is, both teams have won more games and have had similar win percentages in recent years, causing an increase in the quality of competition.

METHOD

The dataset used in this paper consists of data collected from the Baseball-Reference.com website. Wins and losses for each team were taken directly from this resource. Winning percentages were calculated from the original data and rounded to three decimal places, as is customary in baseball statistics.

A glance at the data of total wins or losses for either team can be misleading since the current standard of a 162-game season does not appear in the data until 1961. Strike-shortened seasons, cancelled games and other factors also created seasons in which fewer than 162 games were played. To account for this, the total number of wins each year was projected to a 162-game season based on the teams' win percentages for that season. These numbers were rounded to whole values to reflect possible, real-world values. Losses were also projected to a 162-

game season. It is important to note that win percentages for each season were calculated using the original data and not the projected data – this ensures greater accuracy in the model.

The wins and losses for each team in head-to-head match-ups were not directly studied because of the limited number of head-to-head games played each season. The ratio of Yankees to Red Sox head-to-head wins was studied as a regressor for the model.

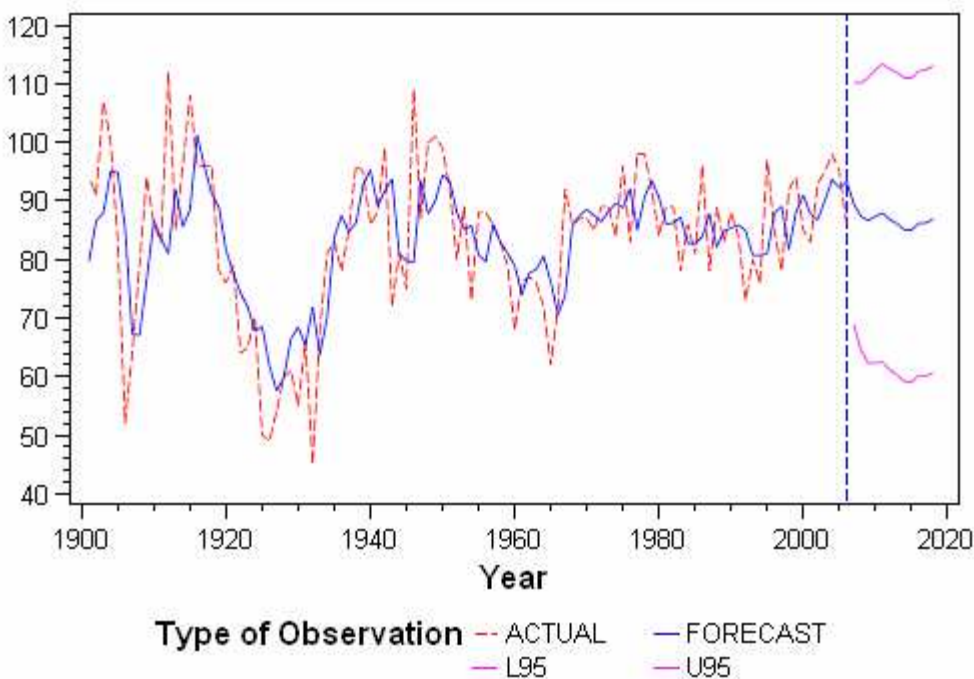
Neither the ratio of the Yankees win percentage to the Red Sox win percentage nor the ratio of Yankees to Red Sox head-to-head wins were rounded to three decimal places, as such ratios are not a traditional baseball statistic, and in this study, we need not adhere to formal formatting.

There are no missing seasons; every year from 1901 to 2006 is accounted for in the dataset. This makes the dataset conducive for analysis using time series methods. It is hypothesized that autoregressive forecasting methods would be more relevant than seasonal forecasting methods since there is no obvious reason for seasonal trends in winning percentages of professional baseball teams. The results obtained using SAS® and SAS Enterprise Guide® support this hypothesis.

RESULTS

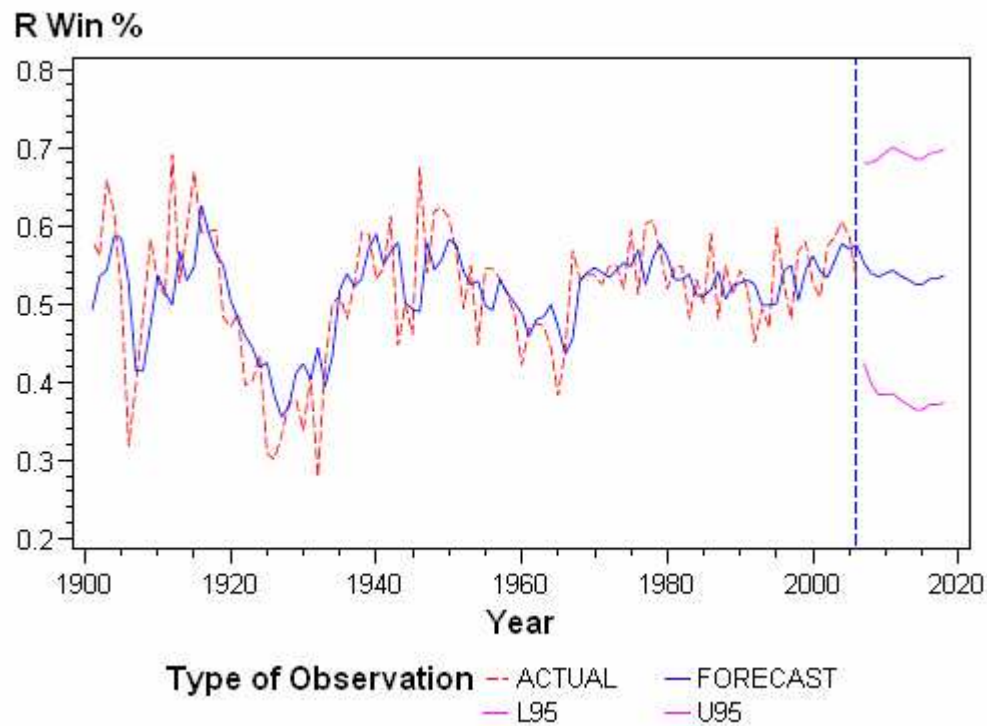
Figure 1. Stepwise Autoregressive Forecast of Normalized Red Sox Wins

R Wins (162)



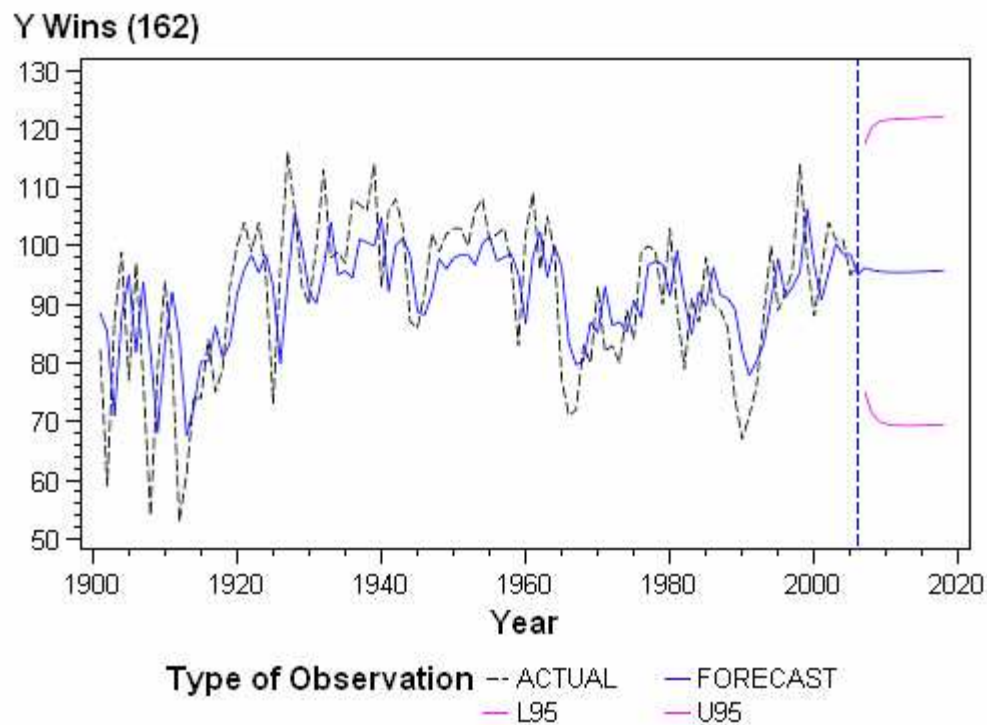
The data for the Red Sox wins, and all following data, are normalized for a 162-game season. There is a sharp decline in wins around 1920 when Babe Ruth was sold to the Yankees – the beginning of an unsuccessful period for the Red Sox. A sharp increase in wins occurs around 1995. There is no hint of seasonality in the data, and seasonality would not be expected for the sort of data we are looking at. The forecast predicts a general decline in wins for the Red Sox over the next 12 years.

Figure 2. Stepwise Autoregressive Forecast of Red Sox Winning Percentage



We see the same graph as Figure 1 when looking at Red Sox winning percentage, save the fact that the vertical axis represents winning percentage instead of number of wins.

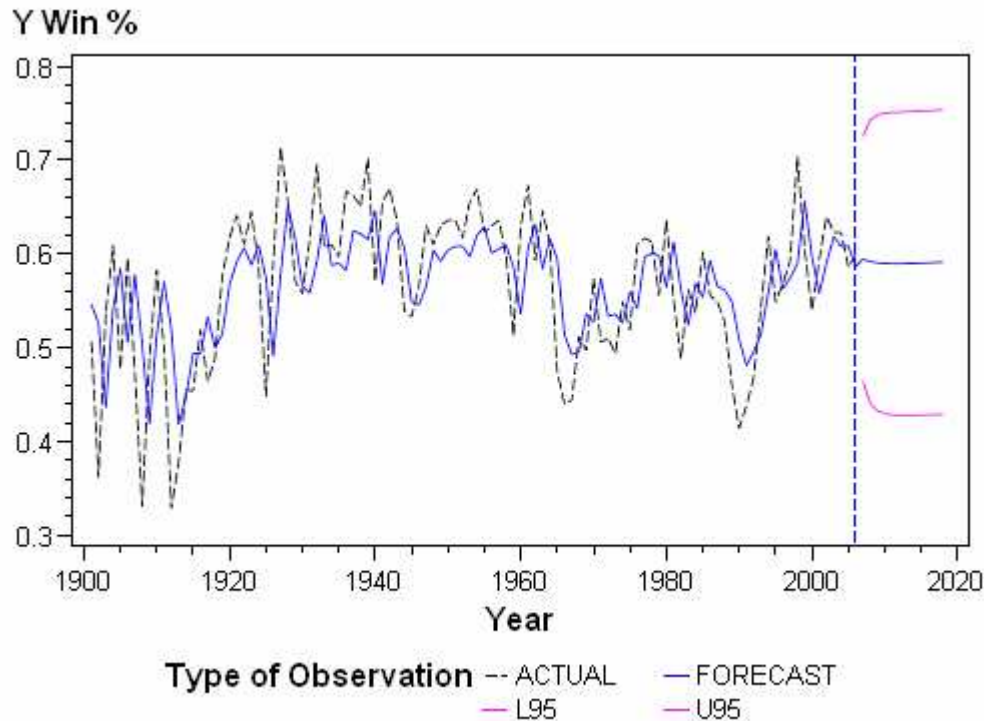
Figure 3. Stepwise Autoregressive Forecast of Normalized Yankees Wins



The data for normalized Yankees wins shows a sharp increase in wins around the time Babe Ruth was acquired from the Red Sox. The high number of wins continues for most of the Yankees' history. A sharp increase in wins occurs in 1994 for the Yankees. This, coupled with the fact that the Red Sox had a sharp increase in wins in 1995, reflects a paralleled increase in quality for both teams. We will see later on how these increases influence each other.

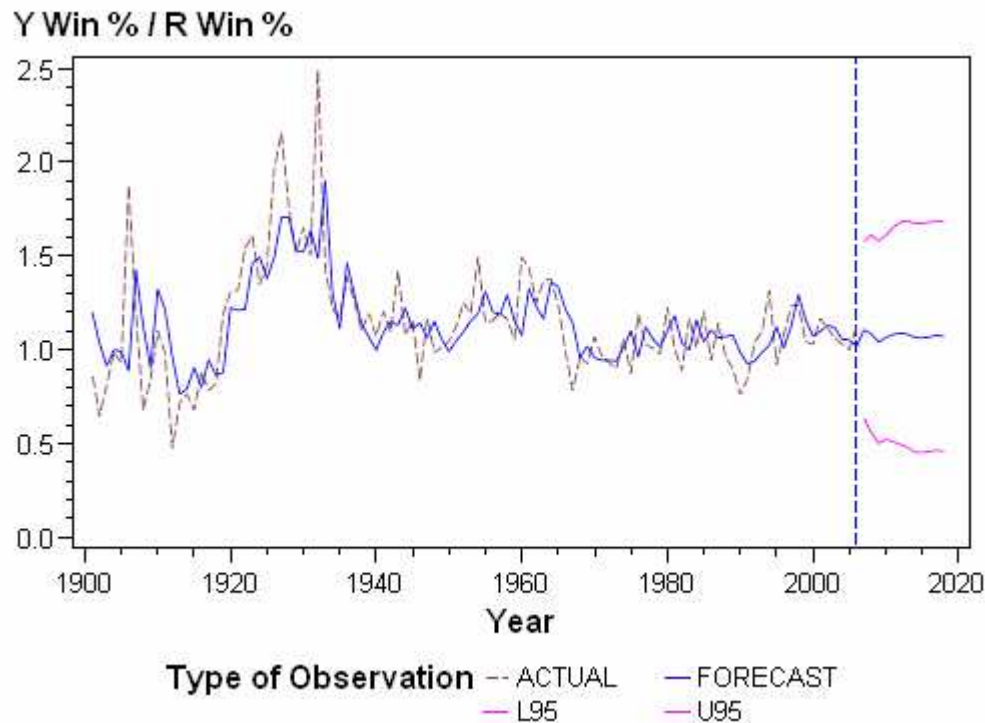
The forecasted data here are nearly horizontal, predicting values only within a limited range, especially when compared to the historical range of the data. This does not necessarily imply that the model is suboptimal or faulty.

Figure 4. Stepwise Autoregressive Forecast of Yankees Winning Percentage



We see the same graph as Figure 3 when looking at Yankees' winning percentage, save the fact that the vertical axis represents winning percentage instead of number of wins.

Figure 5. Stepwise Autoregressive Forecast of the Ratio of Yankees Winning Percentage to Red Sox Winning Percentage

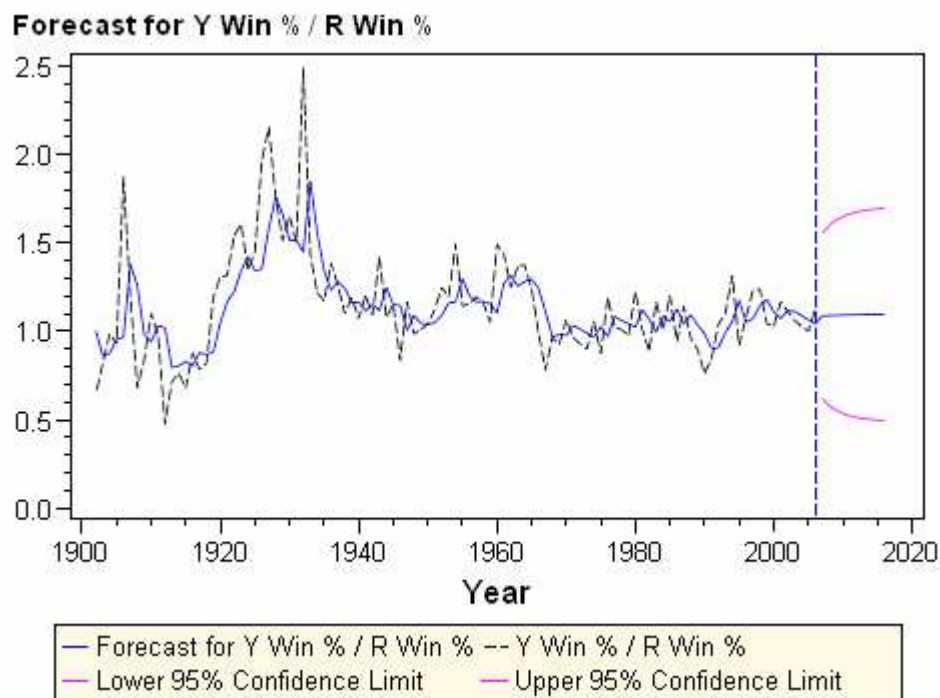


The ratio of Yankees winning percentage to Red Sox winning percentage shows the dominance of the Yankees over the Red Sox from approximately 1920 to 1940. Noting that any value above 1.0 in the graph implies a greater winning percentage for the Yankees than the Red Sox; the Yankees have been a better team historically than the Red Sox.

The ratio of the Yankees winning percentage to the Red Sox winning percentage begins to approximate more closely the value 1.0 starting in 1968. When we combine this information with the knowledge that the Yankees and Red Sox have a sharp increase in wins in 1994 and 1995 respectively, we can make the determination that the rivalry is well and alive. The competitiveness needed to keep fans interested and keep play balanced has thrived in recent years.

The model forecasts that this rivalry should continue in the near future. As long as the ratio is near 1.0 and both teams are winning a high percentage of their games, the rivalry will be strong.

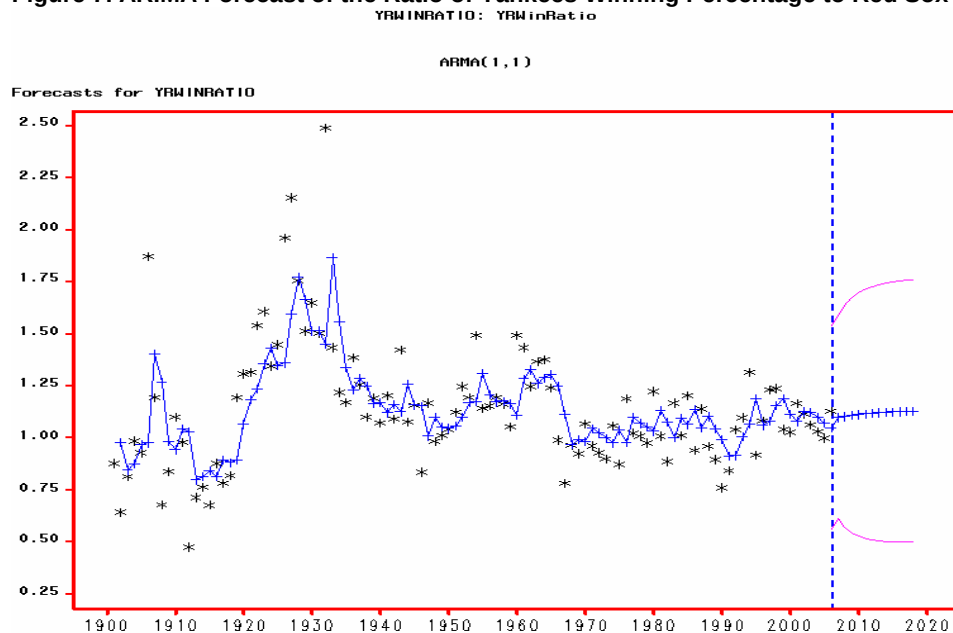
Figure 6. ARIMA Forecast of the Ratio of Yankees Winning Percentage to Red Sox Winning Percentage



The graph of the ARIMA model ($p=1$, $d=0$, $q=1$) for the ratio of Yankees winning percentage to Red Sox winning percentage appears to be similar to that of Figure 5. There appears to be a slight lag in the forecast, but this could not be compensated for with attempted changes in autoregressive and moving average parameters.

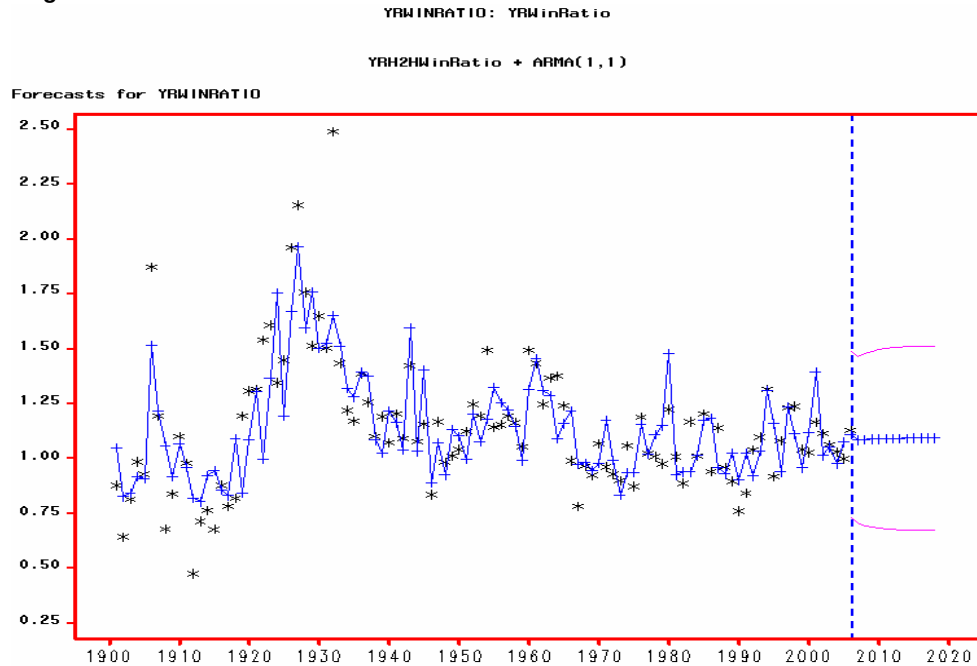
The forecast for the future ratio is nearly horizontal, predicting a slight and consistent edge for the Yankees over the Red Sox in winning percentage. This model, like the one in Figure 5, forecasts that the rivalry should continue to be strong in the near future.

Figure 7. ARIMA Forecast of the Ratio of Yankees Winning Percentage to Red Sox Winning Percentage



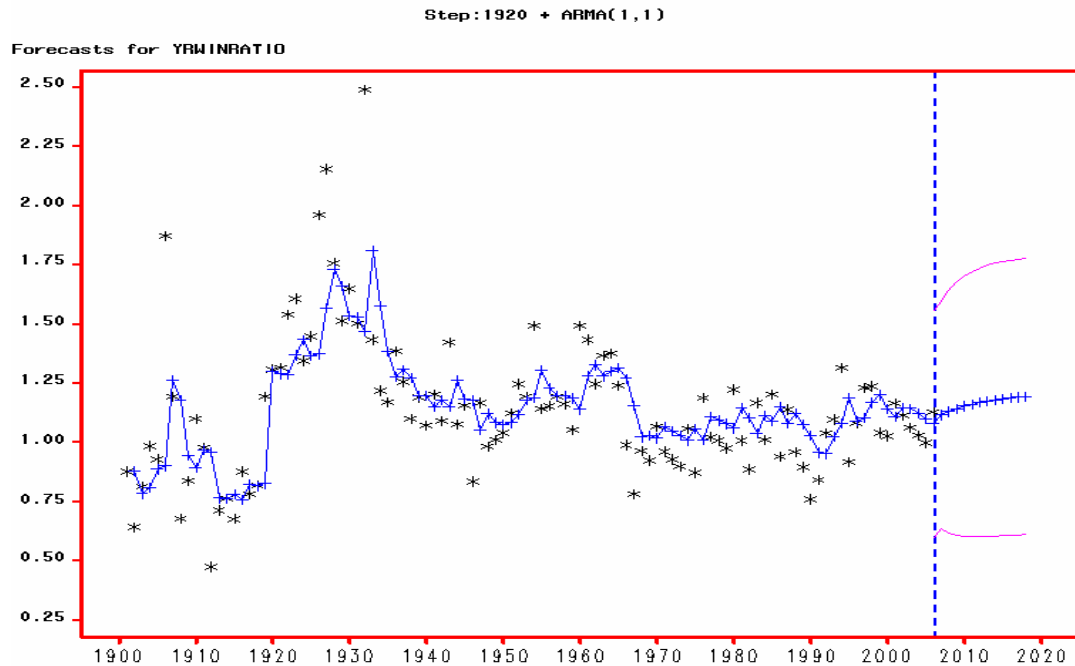
The graphs in Figure 7 and Figure 6 are the same – Figure 7 was created using SAS® while Figure 6 was created using SAS Enterprise Guide®. This model has a Root Mean Square Error (RMSE) of 0.23408. We will look to improve the model by using data from head-to-head match-ups and by taking the Ruth sale into account.

Figure 8. ARIMA Forecast of the Ratio of Yankees Winning Percentage to Red Sox Winning Percentage with Regressor of the Ratio of Yankees to Red Sox Head-to-Head Wins



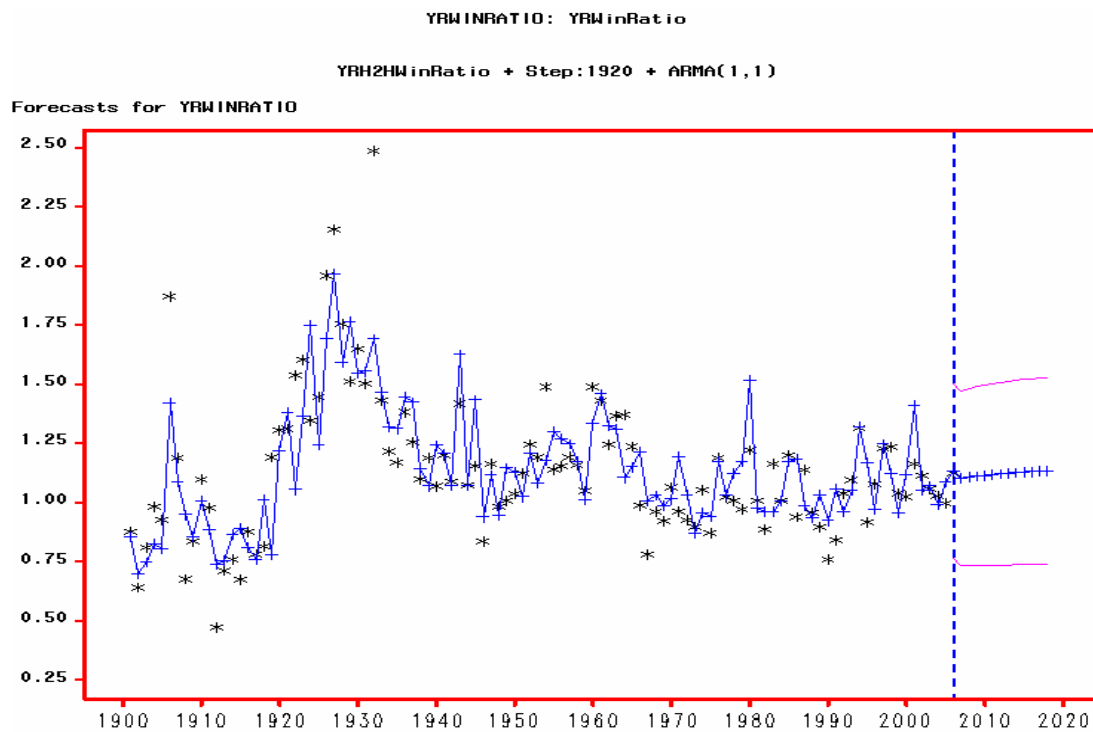
When the ratio of Yankees to Red Sox head-to-head wins is added as a regressor to the model corresponding to Figure 7, we get a more chaotic model with a larger range of predictor values. This modified model appears to better approximate more extreme values and is more accurate overall. Here, we have RMSE of 0.18309 as opposed to the original 0.23408. It is interesting to note that if all games were played head to head, the regressor would perfectly duplicate the original model.

Figure 9. ARIMA Forecast of the Ratio of Yankees Winning Percentage to Red Sox Winning Percentage with Intervention at 1920 Corresponding to Sale of Babe Ruth to Yankees from Red Sox
YRWINRATIO: YRWinRatio



We will look at adding an intervention to the original ARIMA model without using a regressor. If we add a step intervention at 1920 to account for the sale of Babe Ruth from the Red Sox to the Yankees, we get the results from Figure 9. This modified model does not appear to be significantly different, as a sharp increase already occurs in the original ARIMA model. The RMSE is slightly better when compared to the original ARIMA model. Here it is 0.22988 compared to the original of 0.23408. What is interesting about this modified model is that the forecasted values show a logarithmic-like increase of the ratio of Yankees winning percentage to Red Sox winning percentage; this is opposed to the nearly flat-line forecast of the original ARIMA model.

Figure 10. ARIMA Forecast of the Ratio of Yankees Winning Percentage to Red Sox Winning Percentage with Regressor of the Ratio of Yankees to Red Sox Head-to-Head Wins and Intervention at 1920 Corresponding to Sale of Babe Ruth to Yankees from Red Sox



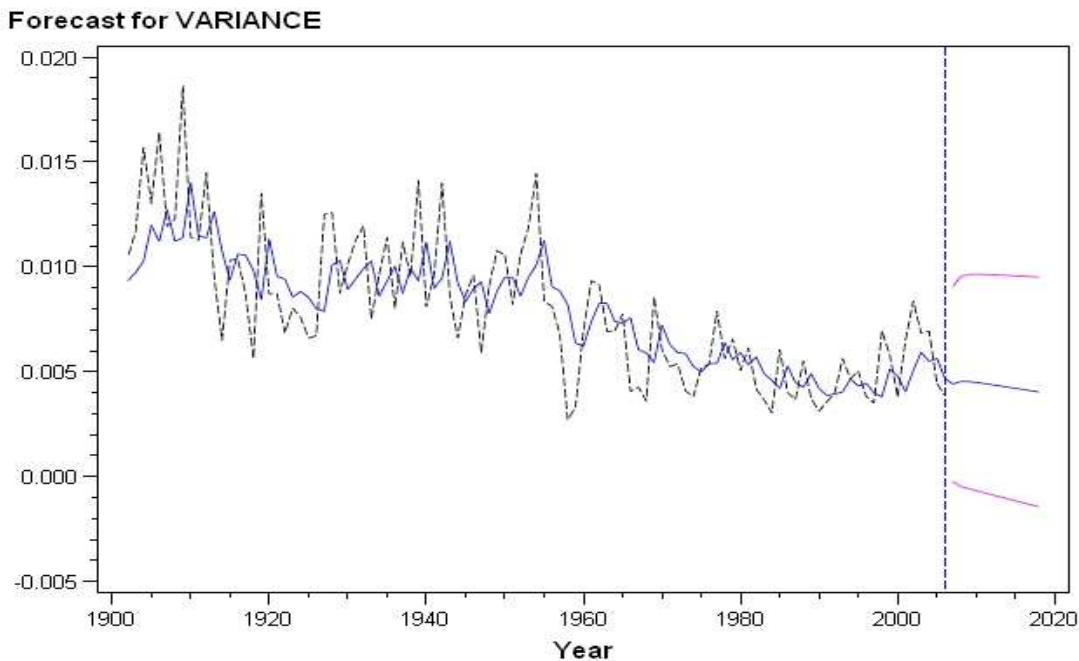
If we add both the ratio of Yankees to Red Sox wins as a regressor and the sale of Babe Ruth in 1920 as an intervention, we obtain the graph of the modified ARIMA model shown in Figure 10. These two modifications to the model each resulted in an increase in accuracy on their own accord. When we combine these two factors, we achieve our most accurate model. The RMSE of the model shown in Figure 10 is 0.17734, a significant improvement over the original ARIMA model with RMSE 0.23408.

Table 1. RMSE of Models with and without Hold-Out Sample of 12 Periods

Model	RMSE without hold-out sample	RMSE with hold-out sample
ARIMA	0.23408	0.11542
ARIMA + Regressor	0.18309	0.12260
ARIMA + Intervention	0.22988	0.11719
ARIMA + Regressor + Intervention	0.17734	0.12173

When a hold-out sample of 12 periods is applied to the models, the rank in accuracy of the models is significantly different. While adding our selected regressor and intervention improved the original model, we get nearly opposite results with a hold-out sample of 12 periods. When a hold-out sample is used, the original ARIMA model is the most accurate with RMSE of 0.11542, which is far lower than the same model without a hold-out sample. Adding a regressor or intervention when using a hold-out sample results in a decrease in accuracy of the model. This dramatic change in results should make us very cautious when using these models for forecasting purposes.

Figure 11. ARIMA Model of Variance in Winning Percentages Each Year between all Major League Baseball Teams



The increase in competitiveness between the Yankees and Red Sox parallels a historical trend of growing parity in Major League Baseball. The ARIMA model ($p=1$, $d=1$, $q=1$) of the variance in winning percentages each year between all teams forecasts a continuation of this trend. There is much room for further investigation.

CONCLUSION

The correlations of win percentages between the Red Sox and Yankees in recent years reflect a boost in the rivalry and increase in competitiveness between the two teams. Notable recent trends include the ratio of the Yankees winning percentage to the Red Sox winning percentage becoming close to 1.0 starting in 1968. Also, a sharp increase in wins occurred in 1994 and 1995 for the Yankees and Red Sox respectively. This general trend of a higher win percentage for both clubs in recent years has helped fuel the storied rivalry between the teams.

The results obtained using SAS Enterprise Guide® imply a forecast of continued competitiveness between the Yankees and Red Sox. A number of the models forecast a limited range of values for the future, especially when compared to the historical range of the data. This is true for the Yankees' normalized wins, normalized losses and winning percentage using the Stepwise Autoregressive Method. We have the same result for the ARIMA model of the ratio of the Yankees winning percentage to the Red Sox winning percentage.

The fact that the models have nearly flat-line forecasted values does not necessarily imply that they are poor or incorrect. They seem to fit the data quite well, especially in contrast to seasonal or exponentially-smoothed models. The results obtained when using a hold-out sample may be more significant when studying model accuracy.

Though adding data from head-to-head match-ups as a regressor and creating a step intervention corresponding to the sale of Ruth originally gave us a more accurate model, both showed contradictory results when a hold-out sample was used. A few more decades of statistical data may yet reduce this problem. At the least, we can say the selected regressor and intervention showed strong correlations to the data of primary interest. In the case of the latter, we obtained a notable difference in trend of forecasted values.

The methods shown here can be applied to win percentages of all professional baseball teams to analyze competitiveness, likelihood of rivalry and historical winning trends. This work could be extended to measure the effect of an acquisition of a player, a geographical move, a change in ballparks, etc. There are other sports for which time series methods would be particularly relevant, especially those with a high number of games in a season.

Time series methods do not seem to be heavily used in the area of sports statistics. These methods could be particularly useful for sports team owners who wish to identify what factors are influential in creating a team that can win consistently. Sports leagues could try to spot and hasten the development of rivalries by looking at winning percentages and then scheduling games accordingly.

REFERENCES

1. Sheinin, Dave. "Red Sox-Yankees Is Difficult to Top." *The Washington Post*. 30 Mar. 2005. 19 Feb. 2007. <<http://www.washingtonpost.com/wp-dyn/articles/A8479-2005Mar28.html>>
2. "Boston Red Sox (1903-Present)." 8 May 2006. 21 Feb. 2007. <<http://www.sportsecyclopedia.com/al/bosox/redsox.html>>
3. "100 Greatest Baseball Players." 20 Feb. 2007. <<http://www.baseball-almanac.com/legendary/lisab100.shtml>>

ACKNOWLEDGMENTS

Special thanks to Dr. Patricia Cerrito.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Grant Johnson
Student – University of Louisville
3116 Hill Park Ct APT 272
Louisville, KY 40220
Work Phone: (502) 475-2549
E-mail: gajohn03@louisville.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.