

Runs per Game in Baseball 1871 - 2010

The data and questions to consider

The data used in this project consists of the average runs scored per game (R/G) by a team over the course of each Major League Baseball season (1871 - 2010). Over the 100+ seasons of professional baseball, there have been historical periods of play where rules or style of play fluctuate in favor of either pitchers or hitters. Thus it may be possible high or low R/G seasons are seasonal due to the change in rules or environment that favors hitters or pitchers. It is also important to see whether or not these fluctuating periods of time can be predicted through a time series model.

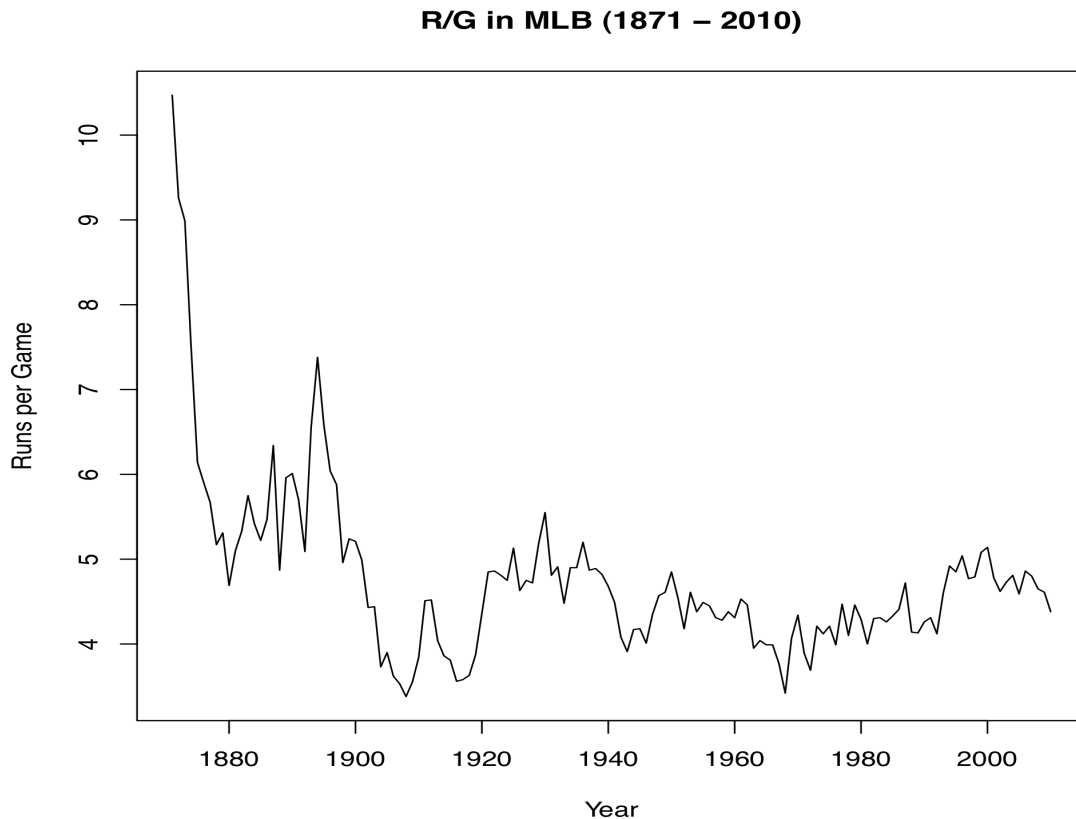


Figure 1: R/G in MLB (1871 - 2010)

Are we able to predict future season dominance by hitters or pitchers? Figure 1 plots the time series data. In the beginning we see a lot of fluctuation as the league was just forming. New rules were experimental and the game was beginning to take shape. After 1930, R/G continues to fluctuate over decades, but doesn't seem as volatile as prior years.

ACF/PACF

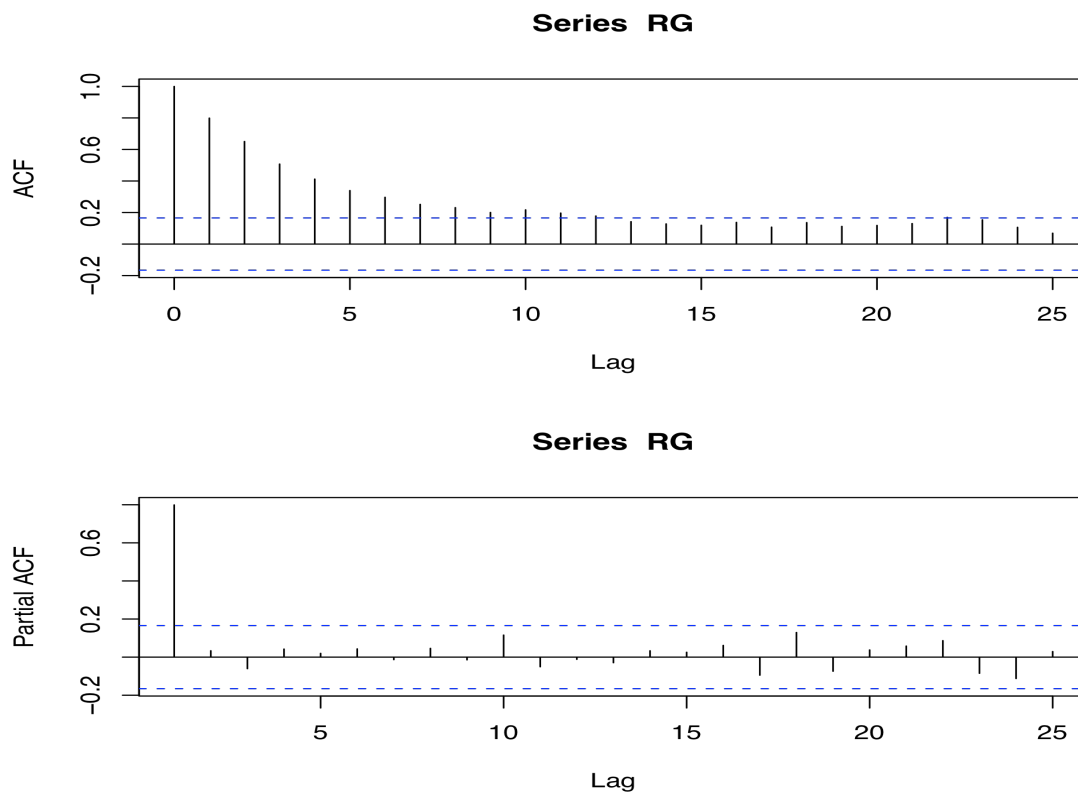


Figure 2: ACF and PACF of R/G

From Figure 2 we see how the ACF of the data is tailing off as the lags increase. The PACF also seems to drop to near zero after the first lag (which is pretty high at around 0.8). This leads to a belief that a good model to fit the data would be an AR(1) model. Going against my thought of possible seasonality, the ACF/PACF graphs show now evidence for such instances.

ARMA Model

As stated, I used an AR(1) model to fit on the time series data due to the characteristics of the estimated ACF and PACF in Figure 2. I also used an ARMA(1,1) to compare results. When comparing their AIC, the AR(1) model is slightly better (167 to 169). Thus the model fitted is $X_t = 0.981_{(0.02)} X_{t-1} + W_t + 5.884_{(1.59)}$ where $W_t \sim WN(0, 0.1818)$. This model could have predictive powers, although it is telling us that the next year's R/G average will be close to the previous year's mark. Looking at the diagnostics however we see the model fits the time series data pretty well.

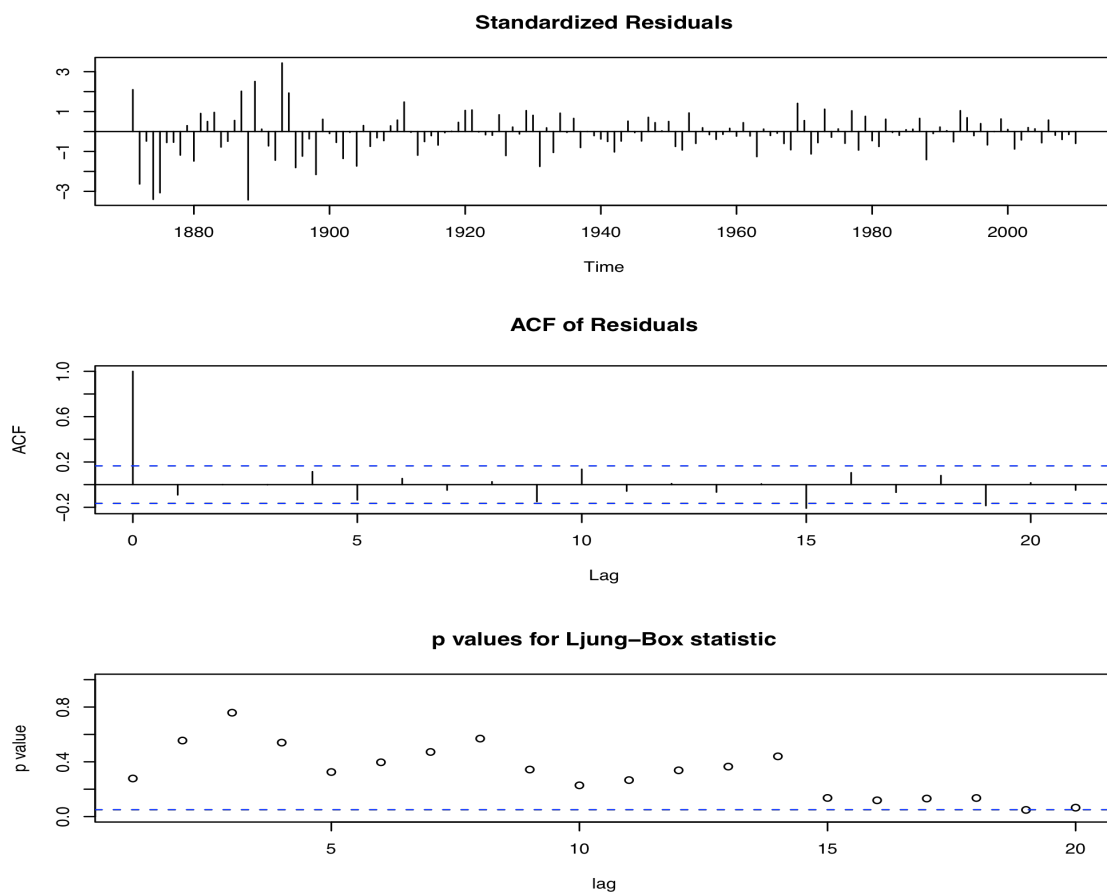


Figure 3: AR(1) Diagnostics

As shown in Figure 3, the residuals don't show any sort of pattern besides the high volatility of runs scored per game during the early years of the league mentioned earlier. The ACF of the standardized residuals also show no patterns, as the lags are close to zero. The Ljung-Box statistic test also show no significant p-values until higher lags greater than 15.

Spectral Density

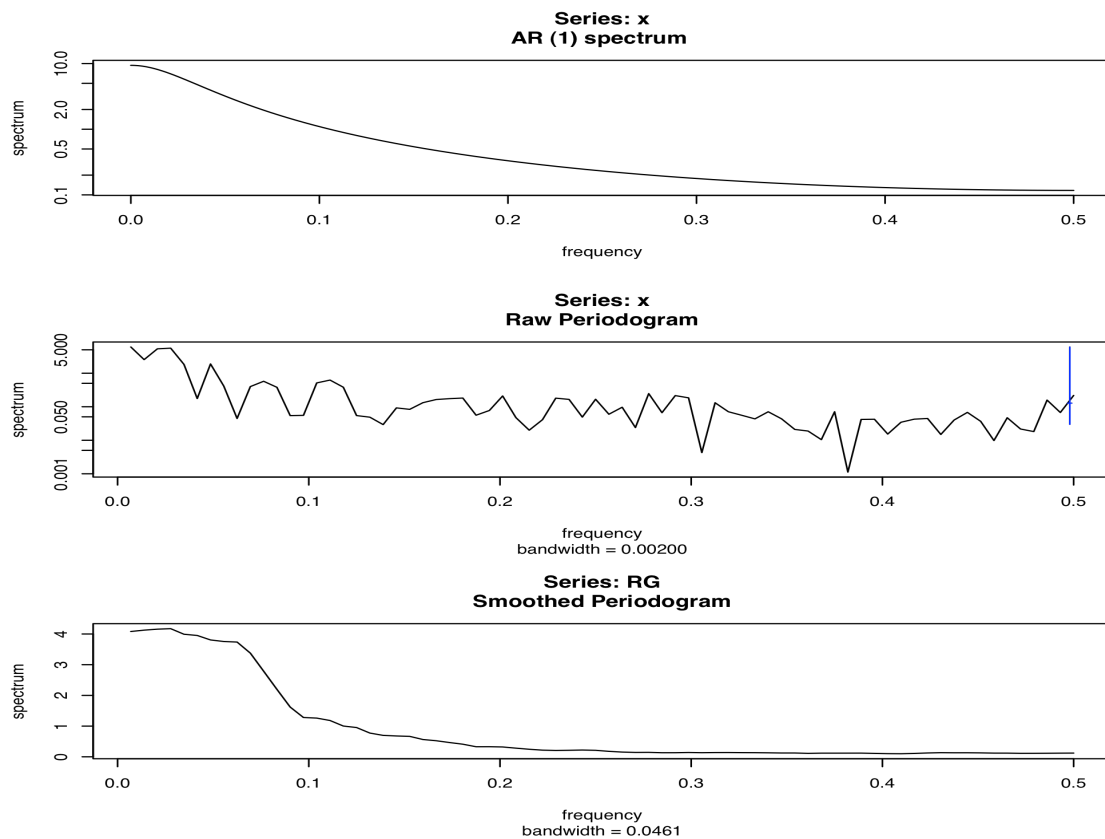
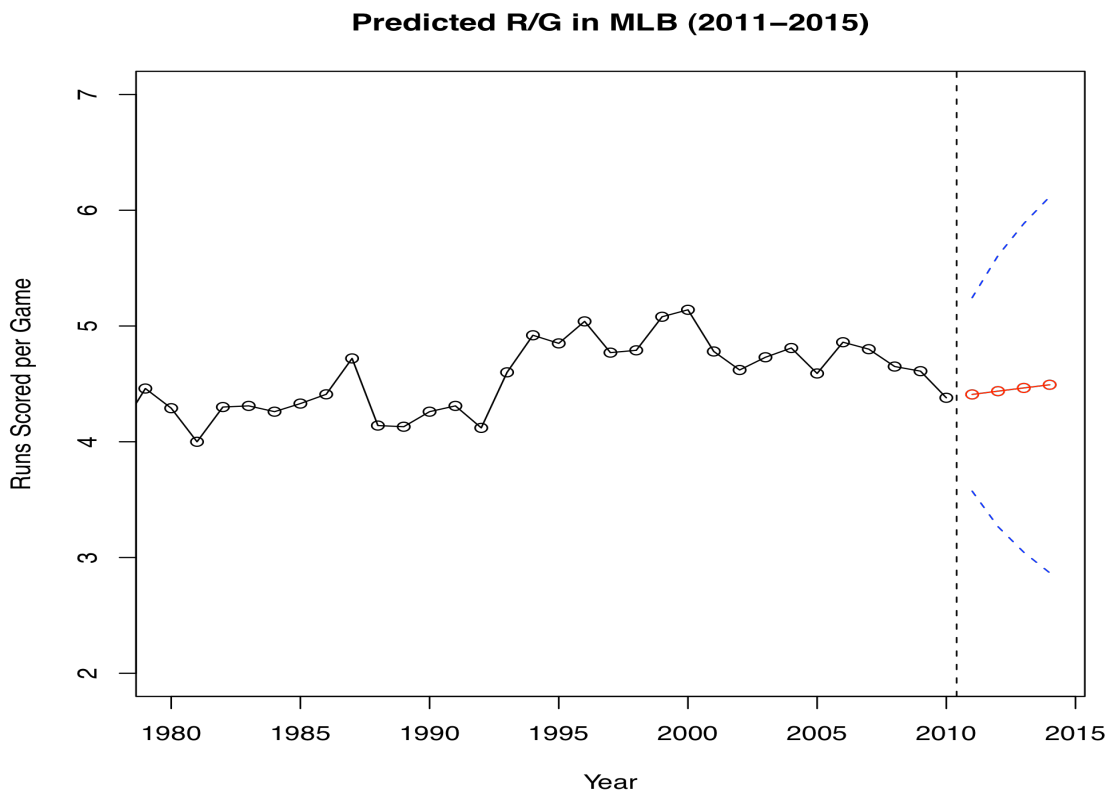


Figure 4: Spectral Density (AR, Raw Periodogram, Smoothed Periodogram)

As you can see in Figure 4, the estimated spectral density (using a raw periodogram approach) doesn't fully show the general shape of the theoretical spectral density (the top graph). Using a smoothed periodogram the shape is more clearly seen.

Conclusion

As I mentioned in the beginning, the most important aspect of using time series modeling for R/G in MLB was its predictive powers. While the model found to best fit the data was significant, its predictive powers are not so telling. A standard 95% confidence interval for the 2011 season tells us that average R/G will be between 3.56 and 5.23. This interval is still very big and inconclusive on what kind of years would pan out. Below is a graph of the next 5 years graphed out with confidence intervals.



Reference: Baseball-Reference.com, 2010,

<http://www.baseball-reference.com/leagues/MLB/pitch.shtml>