Analytics for Local Collegiate Baseball League:
Improved Statistics and Favorable Factors

Researchers: Hunter Hartke and Brett Schulte, with Faculty advisor: Dr. Jessica Kraker

University of Wisconsin – Eau Claire, Mathematics Department

The Power of AND

University of Wisconsin Eau Claire

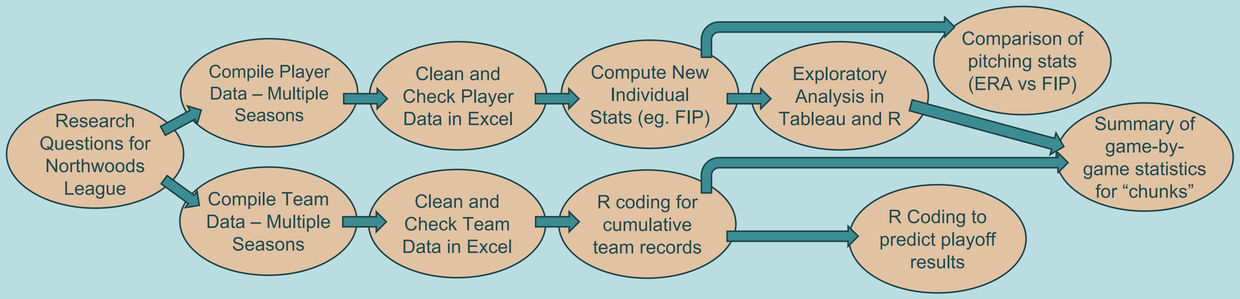
Abstract

This project focuses on analytics methods based on traditional, historic statistics gathered for baseball players, as well as team win-loss records within a defined competitor framework. Methodologies for both team-level and player-level analyses were adjusted for the Northwoods League, including the local team Eau Claire Express, using historical data. We hope to be able to provide value to the local community by sharing some of the insights gained.

Assessments of individual player batting and pitching strengths were computed, based on statistics developed recently within Major League Baseball; explanation of these metrics are available on sites such as at FanGraphs. Comparisons of these newer metrics are made to historical assessment measures.

Summaries of team records were gathered across the most recent four seasons, for 18-20 teams in the league. Various recursive record-updating methods were considered for predictive purposes. The current analysis examines summary statistic values that appear to be most associated with streaks of wins or losses. Methods for modeling streaks by incorporating team statistics and other metrics are examined.

Goals and Process This outline to the right tracks the research, data management, and programming used to gather, clean, compute, organize, and evaluate the team and individual player data for the Northwoods League.



Data Compilation

All individual and team statistics were compiled directly from the Northwoods League website for each season. For individual players that played on multiple teams during a given season, his full season's data was compiled for use. Data was parsed to include only useful statistics. New statistics were created from this data modeled off of MLB statistics. Two metrics used frequently, FIP and wOBA, are relatively new statistics in the MLB to better represent the value of pitchers and batters, respectively. Following the formula outlined on FanGraphs, we created these statistics for the Northwoods League players and teams.

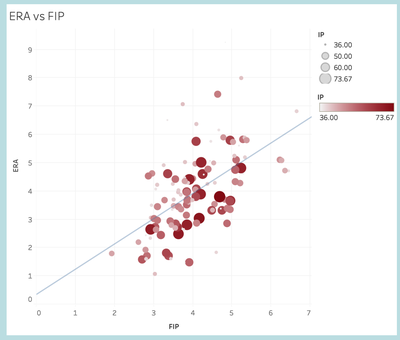
$$FIP = \frac{(13*HR)+(3*(BB+HBP))-(2*K)}{IP} + FIP\ Constant$$

$$FIP\ Constant = lgERA - \frac{(13*lgHR)+(3*(lgBB+lgHBP))-(2*lgK)}{lgIP},$$

where each weight is specific to the 2018 MLB season

$$wOBA_{2018} = \frac{(.69*BB)+(.72*HBP)+(.88*1B)+(1.247*2B)+(1.578*3B)+(2.031*HR)}{AB+BB+HBP-IBB+SF+SH}$$

Plot of historic measure (ERA) vs. new measure (FIP) summarizing pitcher effectiveness, for Northwoods League pitchers in 2018.



For team data, each team's game schedule was copied and evaluated in order to look at the team's overall progression of their record throughout the season. A table was put together for each team of a cumulative record for any given team - the organization and computation of the cumulative records was done with for loops in *R*. The team's record could be obtained for any given point of the season. The win-loss records across all teams were further used to calculate Bradley-Terry ratings. These ratings can be used as a way to calculate the probability of each team winning in a head-to-head game, based on all the past games they have played ("strength of record").

Methods

We were interested to identify reasons why teams performed well or poorly during specific parts of the season. Initially, we wanted to look at winning and losing streaks of teams during the 2018 season, but we soon realized that this might not capture the whole picture. Teams can win five games in a row, lose one, and then win the next four games, but looking at streaks only would not capture this 10-game span. Thus we developed the idea of a chunk. A chunk is a number of games (we looked at 15 and 20 game chunks), in which a team had a winning or losing percentage above a minimum value (in our case, we used .8). From this we performed exploratory analyses on one 20 game winning chunk, one 20 game losing chunk, one 15 game winning chunk, and one 15 game losing chunk. Unsurprisingly, the 20 game winning chunk came from the team with the best overall record, Madison, and the 20 game losing chunk came from the team with the second worst overall record, Thunder Bay.

Our exploratory analysis consisted of looking at various individual and team metrics during the chunk and comparing that to their season averages. For the team statistics, we made a game-by-game comparison. For the individual statistics, we compared values from the chunk with the season, weighted by playing time. To evaluate a team's offense, we used wOBA, OPS, BB/K, extra base hit percentage, and team efficiency. To evaluate a team's pitching, we used FIP, ERA, K, BB, K/BB, and percentage of unearned runs give up. Due to the limited defensive data available, no conclusions can be made about the effect of defense during a chunk.

Another initial hypotheses was about the variability of certain metrics. FIP (Fielding Independent Pitching), a relatively new statistic developed for Major League Baseball, is thought to provide a better (compared to the classic statistic ERA) idea of how good a pitcher truly is because it only factors in results the pitcher has direct influence over during the season (or other specified timeframe). Because FIP takes away the variability of the defense for a pitcher, it is thought to have less overall variability between seasons for the same pitcher. So, we decided to test the hypothesis that FIP is a less variable statistic than ERA when evaluated across seasons. Pitching data was gathered for the 2015-2018 seasons, considering only those pitchers who pitched in multiple seasons. Each statistic, FIP and ERA, was computed for these pitchers; the standard deviation, $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$, as well as the simple range, of the statistic was computed across the repeated seasons. In total, s_{FIP} and s_{ERA} were compared for 258 players (218 were evaluated across two seasons, 37 across three seasons, and 3 across four seasons).

The final methodology involves a rating computed from team records, to allow us to predict future game results. The model for Bradley-Terry Ratings is assumed to have a set number of items, K. We compare items *i* to *j* assuming one wins and the other loses. A rating π for each item can be calculated with each iteration with the equation $\pi_i^{(n+1)} = \frac{W_i}{\sum_{j=1}^{K} \frac{n_{ij}}{\pi_i^{(n)}+\pi_j^{(n)}}}$, where π can be calculated for period n+1 by using the π from n to calculate it in the iterative update. Note that W_i, w_{ij}, and w_{ji} are the total number of comparison wins for team *i*, total number of wins for team *i* versus team *j*, and the total number of wins for team *j* versus team *i*. Thus $W_i = \sum_{j=1}^{K} w_{ij}$ equals the total number of comparisons won by item *i*, and $n_{ij} = w_{ij} + w_{ji}$ (with $n_{ii} = 0$) are the number of comparisons between *i* and *j*.

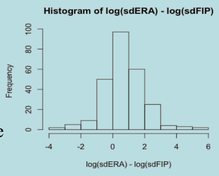
In each period n+1, the π_i for each team is calculated with each game outcome impacting all π_i calculations for all teams. Each team, at n=0, will start with a $W_i = 1$ and $n_{ij} = 1$ as well, which corresponds to an equal start for all teams, and we used the constraint $\pi_1 + \pi_2 + \pi_3 + \cdots + \pi_{20} = 1$. This constraint implies that π_i for each team is equal to .05 before the season starts because each team has the same chance of winning a game as another team, with K=20 representing the total number of teams playing. Finally, these are used to calculate: $p_{ij} = \mathbb{P}(i\ beats\ j) = \frac{\pi_i}{\pi_i+\pi_j}$, which is the probability *i* beats *j*, with π (between 0 and 1) reflecting the iteratively-updated rating incorporating the team's record to date. These constants π_i were computed for the 2018 regular season to integrate strength of teams' season records into predicting playoff results.

Some Conclusions

Comparison of pitching stats (ERA vs FIP):

If there was no difference in variability between the two metrics, then the histogram should be centered around 0. The histogram, however, is not centered around 0, instead around some positive value between 0 and 1.

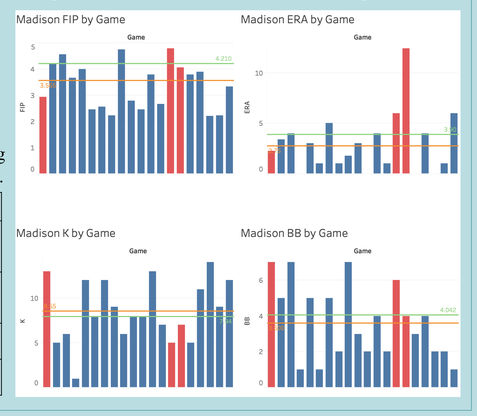
From this graph, it is clear that ERA *tends to* be a much more variable statistic than FIP for pitchers in the 2015-2018 seasons.



If $\frac{s_{ERA}}{s_{FIP}} > 1$ then $\log(s_{ERA}) > \log(s_{FIP})$; hence, we run a simple sign test to determine if the $\log(s_{ERA})$ *tends to* be larger than $\log(s_{FIP})$. This test results in a P-value of $3 \cdot 10^{-15}$, meaning that there is extremely strong evidence that pitchers' ERA tends be more variables than their FIP. The values of s_{ERA} also tend to be more extreme than those for s_{FIP}, hence the need to examine the logarithms. We discuss possible extensions of this analysis in Future Work.

Summary of game-by-game statistics for "chunks": The results from the exploratory analysis on chunks (whether winning or losing) showed teams benefitted from (or were disadvantaged by) different aspects of the game. However, the direction of the pitching or batting performances during the chunk, as compared to the season average, did align with the team's performance over the chunk.

The graphs on the right provide an example of the statistics used to evaluate team performance. Shown are four pitching metrics for Madison during its 20 game chunk, with the chunk and season averages for each statistic shown in orange and green, respectively.

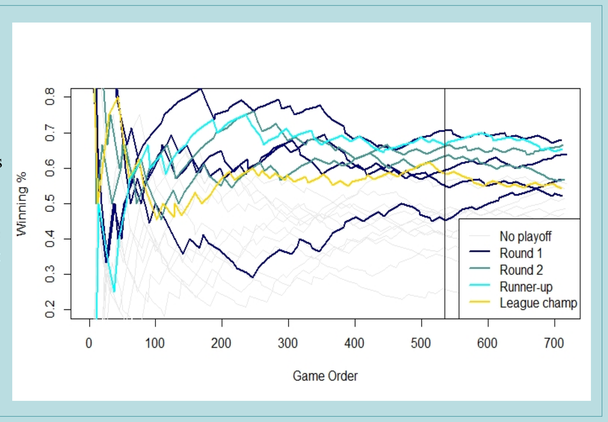


The table below summarizes the performance results for each team during the chunk, in comparison to their season.

	MAD	DUL	TB	WAT
Chunk Type	Win	Win	Lose	Lose
# of Games	20	15	20	15
Pitch	↑	---	↓	↓
Bat	---	↑	---	↓

Predicting playoff results: The Bradley-Terry ratings (calculated with the *BradleyTerry2* package in *R*) across the full season were not accurate for predicting team performance in the playoffs. Teams that were expected to win in the first round of the playoffs were beaten by teams that weren't predicted to win the game. However, if the Bradley-Terry ratings from about 180 games before the playoffs (after the season break) are calculated, it's more accurate than the overall season ratings. Going back 150 games before the playoffs gives an even more accurate prediction of how the playoffs will turn out.

Though we concluded that the iterative ratings work better for the latter part of the season, the winning-percentages in the "chunks" directly prior to the playoffs do not provide meaningful insight. This may be due to several factors: teams may rest their better players in the last few regular-season games; or, better players may be absent during the playoff games (due to the players returning to college); or, the one-elimination format of the playoffs may not give teams the same opportunity that a series would. More playoff games would give a better overall picture of which team is better, and further examination of individual-player effects is discussed below.



Future Work

To continue to test the variability of FIP and ERA, we would like to consider number of innings pitched as a weight in our test in order to factor in players that pitched more, and include less those who pitched such few innings. Future work may also include tests about advantages of new batting metrics compared to historic statistics.

Our KRACH ratings did a poor job at predicting playoff results, which leads us to believe individual players have a large effect on the playoffs, especially because most rounds are single elimination. In order to incorporate individual players into a predictive model, we need a value to assign to each player. The MLB uses WAR, defined below, as a single number to give explain the amount of value each player provided to his team that season. We would like to pair WAR with our KRACH ratings to predict the playoffs, but many pieces to WAR are difficult to identify at a collegiate level.

$$WAR = \frac{Batting\ Runs + Base\ Running\ Runs + Fielding\ Runs + Positional\ Adjustment + League\ Adjustment + Replacement\ Runs}{Runs\ Per\ Win}$$

Review of how WAR- "like" statistics are computed for players with only traditional statistics could allow us to devise a similar measure.

Finally, we would like to continue our exploratory analysis of winning and losing chunks by analyzing more chunks to look for consistent factors that led to both success and failure over periods of time, and devise tests to assess the strength of those factors.

Acknowledgements

- This research project was funded by the Office of Research and Sponsored Programs (ORSP) during the 2018-19 academic year.
- Computing support and research space from the Mathematics Department.

Data All data gathered during October 2018-April 2019 from https://northwoodsleague.com/eau-claire-express/, open to public access.

References

- References for *BradleyTerry2* package in R: https://cran.r-project.org/web/packages/BradleyTerry2/vignettes/BradleyTerry.pdf.
- Definitions for various statistics available at: https://www.fangraphs.com/, or notes available upon request.
- Computation: Tableau v. 10, R v.. 3-5-0 to 3-5-3 (2018-2019 versions), and Excel.