Analytics for Summer Collegiate Baseball:
Connecting Individual and Team Results

Researchers: Bryce Damman, William O'Brien, and Brett Schulte

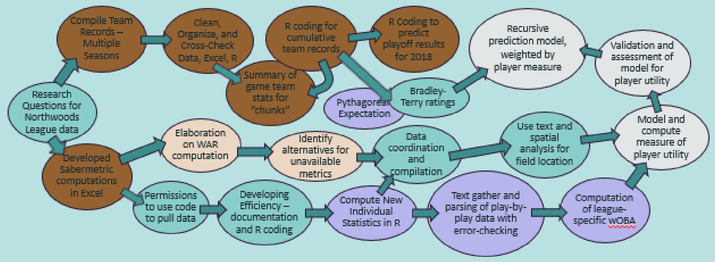
University of Wisconsin – Eau Claire, Mathematics Department with Faculty advisor: Dr. Jessica Kraker

Abstract

A prior study on baseball analytics for the Northwoods League summarized the available components needed to calculate a "Win Above Replacement" (WAR) metric for individual players, as well as identified missing and potential replacement measures in this league.

Some of the inputs were found to be unobtainable due to lack of technical equipment in the league, but research into the origins of the measure and historic records allowed us to identify some suitable substitute measures.

Obtaining these measures is possible due solely to numerous programming and data-gleaning achievements. These include code to pull and summarize play-by-play information from the original online text source; back-computing the physical locations of defensive plays from an image of the playing field; and creating a system both to extract and to compute player and team metrics beyond those automatically provided for the Northwoods League data. Summary of these methods will be included in the presentation, along with a discussion of the structure of the usable data, including play-by-play output, individual-player, and league-level summaries. Additionally, progress towards a WAR-analogy estimation (using a model connecting team performance and player appearance) will be included.



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 Northwoods League.

Data Principles and Permissions

- A large amount of the work of this project consists of data collection. In the past, we gathered our data from the Northwoods League site by hand. However, this severely limited the scope of the data we were able to use, as it would be unfeasible to gather play level data from multiple seasons, over all games in those seasons, and over all games in those innings. In order to get such amounts of low-level information, the process needed to be automated.
- We received permission early in the year to secure permission to automate the data aggregation process. We were able to automate this process using a web automation tool, Selenium, the process of trying to aggregate each different type of data for each season, and put it into a usable format, was a trial of its own. Our main system for pulling data pulls nine different data sets: specifically:
  - The individual games, including factors like winning team, final scores, weather information, etc.
  - The list of balls in play for each game, with an approximate location of where the ball was hit to.
  - The list of players in each game, or the set roster for each game, including their positions on the field.
  - Text data regarding what happened during each at bat during each game.
  - Player data, including relevant statistics for pitchers, on a per season basis.
  - Player data, including relevant statistics for batters, on a per season basis.
  - Player data, including relevant statistics for pitchers, on a per game basis.
  - Player data, including relevant statistics for batters, on a per game basis.
- Utilizing Northwood's system for uniquely identifying players and games by a set ID, we are able to effectively match these data sets to each other, as shown in the Data Structure and Organization section. This allows us to create many useful factors for modelling.

Play-by-Play

- One of our main goals was to see if it was possible to develop league-specific metrics as the previous work for this study only utilized constants and weights that were originally created for Major League Baseball (MLB). To actualize these league-specific values, we would need to analyze individual play-by-play results across multiple seasons. We define play-by-play to be information within either a single plate appearance or a substitution by either the offense or defense during an inning. We transferred play-by-play from just over 10 seasons worth of games using *R* packages *rvest* and *RSelenium*. In addition to what was already present, we also added identifiers of inning number, game ID, inning half (top or bottom), and the order of a plate appearance or substitution within an inning. Once adjustments to fix or remove errors were made, we finally had suitable play-by-play data that could assist in creating weights and matrices specific to Northwoods League (NWL). One example of this is a Run-Expectancy matrix.

- A Run-Expectancy matrix gives an idea of how many runs on average will be scored by the end of an inning given the current base-out state. So if all we knew was there are baserunners on 1st and 3rd with 1 out, we expect 1.244 runs to be scored by the end of an inning.
- To develop this matrix, we needed to find how often each base-out state occurred, and the total number of runs scored from the time that base-out state occurred until the end of the innings in which they occurred.

Bases/Outs	0 Outs	1 Out	2 Outs
_ _ _	0.579	0.293	0.102
1B _ _	0.997	0.57	0.231
_ 2B _	1.243	0.706	0.314
_ _ 3B	1.568	1.015	0.402
1B 2B _	1.561	0.96	0.438
1B _ 3B	1.908	1.244	0.551
_ 2B 3B	2.043	1.36	0.6
1B 2B 3B	1.829	1.268	0.675

$$RE_h = \frac{\sum Runs_h}{\sum Instances_h}$$ for $h = 1,2,\ldots,24$

League Specific wOBA

One metric that we can create a league specific version of using play-by-play is weighted On-Base Average (wOBA).

$$wOBA_j = \frac{(wBB_j * BB) + (wHBP_j * HBP) + (w1B_j * 1B) + (w2B_j * 2B) + (w3B_j * 3B) + (wHR_j * HR)}{AB + BB - IBB + SF + HBP}$$ where j = League Year.

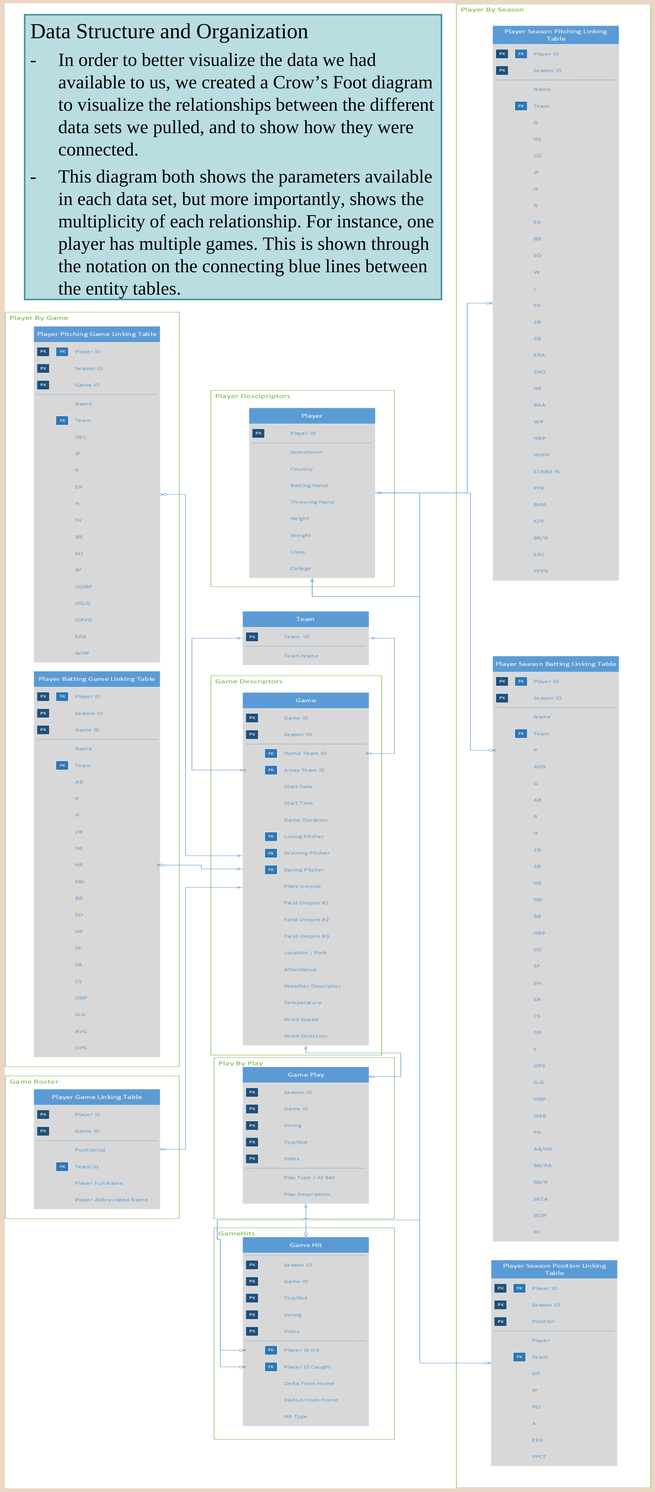
Using the play-by-play and Run Expectancy matrix, we found the following weights:

Year	Walk	Hit By Pitch	Single	Double	Triple	Home Run	wOBA Scale
2010	0.825	0.838	1.051	1.493	2.09	2.429	1.33
2011	0.818	0.844	1.049	1.496	1.923	2.32	1.328
2012	0.797	0.83	1.046	1.502	1.897	2.366	1.284
2013	0.83	0.829	1.082	1.561	1.976	2.484	1.326
2014	0.896	0.937	1.181	1.712	2.264	2.612	1.477
2015	0.806	0.824	1.057	1.525	1.897	2.41	1.319
2016	0.813	0.816	1.041	1.489	1.944	2.328	1.298
2017	0.813	0.825	1.057	1.503	1.968	2.329	1.295
2018	0.805	0.819	1.053	1.487	1.974	2.274	1.29
2019	0.825	0.831	1.072	1.561	1.956	2.341	1.308

In comparison to the MLB weights found on *FanGraphs*, the NWL weights for each category are higher than their professional counterparts for each season. This is due to NWL having a shorter season, and thus a higher variability of our data. This is especially obvious with the 2014 weights, as the 2014 season had a higher proportion of games missing play-by-play (17%) than other seasons (7-10%).

Data Structure and Organization

- In order to better visualize the data we had available to us, we created a Crow's Foot diagram to visualize the relationships between the different data sets we pulled, and to show how they were connected.
- This diagram both shows the parameters available in each data set, but more importantly, shows the multiplicity of each relationship. For instance, one player has multiple games. This is shown through the notation on the connecting blue lines between the entity tables.



Describing Changes to WAR

WAR is Wins Above Replacement. This attempts to give an estimate of a player's value in comparison to a replacement (bench player or free agent).
WAR = (Batting Runs + Base Running Runs + Fielding Runs + Positional Adjustment + League Adjustment +Replacement Runs) / (Runs Per Win)

Below are the calculations used for our own version of WAR.
Changes were made to Baserunning Runs and Fielding Runs as well as getting rid of League Adjustment and Replacement Runs

Position Specific Run Values	
Catcher	+12.5
First Base	- 12.5
Second Base	+2.5
Third Base	+2.5
Shortstop	+7.5
Left Field	-7.5
Center Field	+2.5
Right Field	-7.5
Designated Hitter	-17.5

F6 Based on Position	
Catcher	1
First Baseman	2
Second Baseman	1.25 * (PO+A)/GP
Third Baseman	(4/2.65) * (PO+A)/GP
Shortstop	(7/4.6) * (PO+A)/GP
Outfielder	3 * (PO+A)/GP

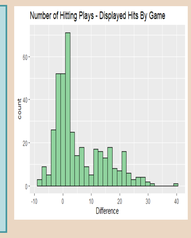
BIZ	Balls in Zone
PM	Plays Made
ZR	Zone Rating = PM/BIZ
OOZ	Out of Zone Plays Made
Chances	BIZ + OOZ
Total PlaysMade	PM + OOZ

F1	20 * (((SB+3)/(SB+CS+7))/4)
F2	(1/.07) * sqrt((SB + CS)/(1B + BB + HBP))
F3	625 * 3B/(AB-HR-K)
F4	25*(((R-HR)/(H+BB+HBP-HR))-.1)
F5	(1/.007) * (.063 – (GDP/(AB-HR-K)))

Batting Runs	wRAA + (lgR/PA – (PF*lgR/PA))*PA
wRAA	((wOBA – lgwOBA)/wOBA Scale) * PA
Baserunning Runs (Speed Score)	(F1 + F2 + F3 + F4 + F5 + F6)/6
Fielding Runs (PAA)	Total Plays Made – lgZR * Chances
Positional Adjustment	((IP / 9) / GS) * Position Specific Run Value
Runs Per Win	9 * (lgR / lgIP)*1.5 + 3

Using Field Location

Given the text based play-by-play data, and the hit location data we were able to pull, we attempted to determine where each ball-in-play landed, and whether that ball in play was converted into an out. However, as we did not have a linking key between those two data sets, we were unable to do so directly. Using an approximating match did not work either, as the number of instances with a ball in play, and the number of ball-in-play locations were not equivalent in most games. Instead, we focused on utilizing the text based play-by-play data to determine the position that would be responsible to convert each ball-in-play into an out, and whether it was converted. We were able to create a metric to describe the percentage of plays where a player could have made an out and did so.



Modeling

Idea: Any measure of player utility would effectively measure what each player adds to any game, averaged out over a season. Using indicator $x_{Team,i} = \begin{cases} 1 & \text{for } played\ some \\ 0 & \text{for } didn't\ play \end{cases}$ is a simple model of "presence"; subscript *Team* is *Home* or *Away*, and the *i* denotes player within team. Then, we could model a response *y* of "team winning-ness" as a linear combination of $x_{Team,i}$: $Response = \beta_0 + \beta_{H1}x_{\text{Home},1} + \beta_{H2}x_{\text{Home},2} + \cdots - \beta_{A1}x_{\text{Away},1} - \beta_{A2}x_{\text{Away},2} - \cdots$

Reasoning: allows for players who are playing multiple positions and/or for multiple teams within a season.

Predicting binary win-loss

- The best model cut a very large proportion of players from modeling.

Games per season	% players cut
Few (<10%)	44%*
Moderate (10-30%)	50%
Many (> 30%)	63%*

*opposite trend of what we want to retain

- Coefficients are less readily interpretable

Predicting score differential *y*

- The model over-shrank the prediction; twice the prediction aligned well with *y*:
- The inability to fit zeroes (no ties) would need to be re-worked: *y* as 9-inning score outcome or utilizing Poisson-regression.
- Coefficients are interpreted as the player contribution to average per-game *y*

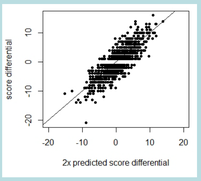
The intercept can be interpreted as the "home team effect".

Type of *Response*:

Option 1: numeric *y* = score differential (home-away).

Option 2: log-odds of winning as response.

Initial Modeling method: Since the number of players is greater than the number of games ($n < p$ problem), a modified regression model is used. A penalized regression model **[ISLR]** is fit using **glmnet** package in R. This constrains a function of the coefficients by an upper bound.



Using Results for further Modeling:

With the adjustments and additional back-computations of metrics for individual players, we now have sufficient available information for modeling these player utilities. Further connections will be made to these statistics through a nonlinear model.

Future Work

- The primary focus of future work is to identify the most effective model for obtaining a measure of player utility. This entails working through further modeling considerations (type and trimming of predictors and/or nonlinear modeling), as well as (and, more importantly) properly validating the model-selection to prevent over-fitting.
- Implementing other predictors as part of the response-modeling process will be considered: further information for parks will been integrated; minute play-by-play data is used to finalize league-specific wOBA and weights.

References

•Definitions for various statistics available at: https://www.fangraphs.com/, upon request.
•Computation: R (2019 software versions, open-source) with primary packages: **rselenium, glmnet**, **stringr**. Current version of Excel was also used for examination of data.
•**[ISLR]** Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An introduction to statistical learning : with applications in R*. New York :Springer, 2013.

Acknowledgements

- Funded by Office of Research and Sponsored Programs (ORSP) during 2019-20.
- Computing support and research space from the Mathematics Department.
- Data gathered during Nov. 2019 – April 2020 from publicly-accessible site: https://northwoodsleague.com/
- Dr. Michael Axelrod for consults on local historical baseball statistics.