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 Krakor

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 toward 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.
  - 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 modeling.

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 Relenium. 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 were 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 inning in which they occurred.

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

<table>
<thead>
<tr>
<th>Bases/Outs</th>
<th>0 Outs</th>
<th>1 Out</th>
<th>2 Outs</th>
</tr>
</thead>
<tbody>
<tr>
<td>score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0B-0B</td>
<td>0.579</td>
<td>0.293</td>
<td>0.102</td>
</tr>
<tr>
<td>1B-0B</td>
<td>0.997</td>
<td>0.57</td>
<td>0.251</td>
</tr>
<tr>
<td>2B-0B</td>
<td>1.243</td>
<td>0.64</td>
<td>0.264</td>
</tr>
<tr>
<td>1B-B</td>
<td>1.548</td>
<td>1.015</td>
<td>0.402</td>
</tr>
<tr>
<td>2B-B</td>
<td>1.561</td>
<td>0.96</td>
<td>0.458</td>
</tr>
<tr>
<td>3B-B</td>
<td>1.908</td>
<td>1.244</td>
<td>0.581</td>
</tr>
<tr>
<td>1B-2B</td>
<td>2.043</td>
<td>1.36</td>
<td>0.56</td>
</tr>
<tr>
<td>2B-3B</td>
<td>1.829</td>
<td>1.268</td>
<td>0.675</td>
</tr>
</tbody>
</table>

Describing Changes to WAR
WAR is Wins Above Replacement. This attempt to give an estimate of a player's value in comparison to a replacement (bench player or free agent).

**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 toward a WAR-analogy estimation (using a model connecting team performance and player appearance) will be included.**

Using Field Location
Given game 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 these 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
**Any measure of player utility would effectively measure which player each data set to each game, averaged out over a season. Using indicator variables for each position and the league, the i denotes player within team. Then, we could model a response of “team winning-ness” as a linear combination of indicator variables:**

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

**Predicting binary win-loss**
- The best model must be very large and complex to capture all the interactions in the data.

**Using Results for further Modeling**
- The intercept can be interpreted as the “home team effect”.
- **Type 2 evidence**

Using Results for further Modeling:
- Number of games (n > p problem), a modified regression model is used. A penalized regression model is fit using glmnet package in R.

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 perform over-fitting.
- Implementing other predictors as part of the response-modeling process will be considered: further information for modeling these player utilities. Further connections will be made to these statistics through a nonlinear model.

References
- [Definitions for various statistics available at: https://www.fangraphs.com/](https://www.fangraphs.com/)
- [Using Results for further Modeling](https://www.fangraphs.com/)
- [Computing other predictors as part of the response-modeling process will be considered: further information for modeling these player utilities. Further connections will be made to these statistics through a nonlinear model.](https://www.fangraphs.com/)
- [Future Work](https://www.fangraphs.com/)