Center for Jewish History, NYC, Public domain, via Wikimedia Commons

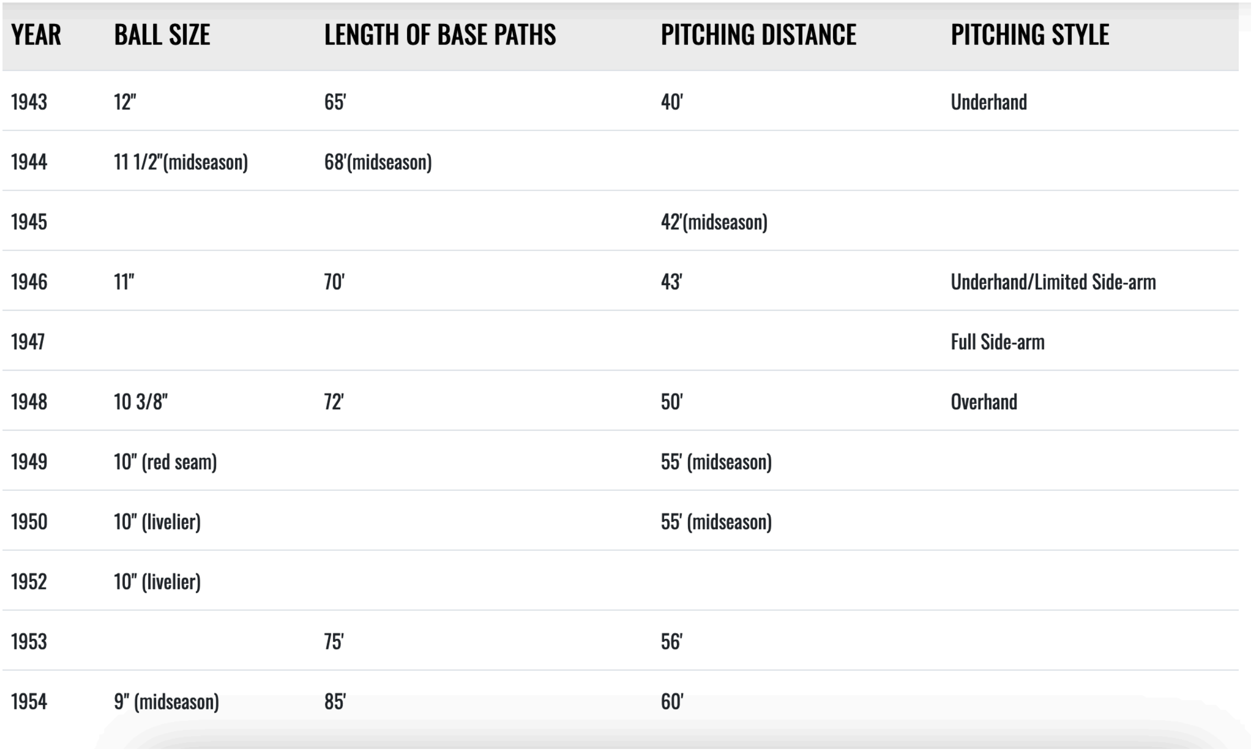
Modified Advanced Statistics for the All American Girl Professional Baseball League

Kacey F · MSDS 456 - Sports Performance Analytic

Introduction

Only two women’s baseball leagues have played in an organized, high level fashion; The All American Girls Professional Baseball League (AAGPBL) and the Japanese Women’s Baseball League. There are few deep dives into women’s baseball from a statistical lens as well as attempts to understand the AAGPBL from a baseball perspective. The few examinations we have of the AAGPBL are based largely on men’s baseball as the standard, rather than adapting statistical methods for the unique type of baseball that the AAGPBL played due to changing rules, gender barriers, lack of typical player development available, and time period.

The AAGPBL experienced a lot of rule changes through the years that drastically affected performance. Over time, the rules got more similar to baseball, affecting the run environment.



YEAR	BALL SIZE	LENGTH OF BASE PATHS	PITCHING DISTANCE	PITCHING STYLE
1943	12"	65'	40'	Underhand
1944	11 1/2"(midseason)	68'(midseason)		
1945			42'(midseason)	
1946	11"	70'	43'	Underhand/Limited Side-arm
1947				Full Side-arm
1948	10 3/8"	72'	50'	Overhand
1949	10" (red seam)		55' (midseason)	
1950	10" (livelier)		55' (midseason)	
1952	10" (livelier)			
1953		75'	56'	
1954	9" (midseason)	85'	60'	

AAGPBL Players Association. (2014). AAGPBL Rules of Play. AAGPBL.org. https://aagpbl.org/history/rules-of-play

The main objectives are to:

- Calculate a modified version of Weighted On Base Average to understand the overall offensive performance for players, especially in times where slugging was low, and to account for the heightened importance of the running game in the league
- Calculated Weighted Runs Above Average to compare player performances from different season as the run environment changed

Literature

Concepts, data, and formulas from in this paper were either taken from or inspired by the following research. The sources are a mix of research into calculating linear weights, wOBA, and wRAA in the MLB and historical data for the AAGPBL.

Weinberg, N. (2016)

The article *The Beginner's Guide To Deriving wOBA* on Fangraphs was the main guide for seeing how MLB wOBA was calculated in modern day and what parts of it could be derived for historical data missing play by play. Specifically, the methods for scaling the data were taken from this article.

Klaassen, M. (2011)

Klaassen's article from *Beyond the Boxscore* provides SQL script for finding the linear weights for MLB seasons. While the exact script wasn't used for this project, it served as inspiration for the coding and helped provide further insight on linear weights conceptually. Klaassen also provides all wOBA Coefficients from 1871 to 2010, which provides historical comparison to the All American Girl Professional Baseball League.

Winston, W. (2009)

Chapter 3 of *Mathletics* provides guidance to calculating linear weights for predicted runs scored. His method of a Multiple Linear Regression model to derive linear weights is the method used here, although with R as the vehicle for calculating it rather than Excel.

Additionally, the formula for predicted runs scored was part of the inspiration for the modified wOBA formula used in this situation.

AAGPBL Players Association Staff, 2014

The All American Girls Professional Baseball League Player Association uploaded the rules in terms of ball size, base path length, mound distance, and pitching style (underhand vs. sidearm vs. overhand.) The rulebook provides important context for the run environment and why certain stats may fluctuate.

Tango, T., Lichtman, M., Dolphin, A., 2006

The Book is one of the oldest guidebooks for baseball analytics, which the website *Inside the Book* takes excerpts from and elaborates on. This gives the basics of the wOBA stat and explains it conceptually while giving the basic formula. Similar, but more base level, to the article from Weinberg.

Slowinski, P. (2010)

Piper Slowinski's article on Fangraphs gives the basis for Weighted Runs Above Average. The formula provided here is used in this paper as well as conceptual concepts such as using wRAA to compare between years and eras.

statscrew.com

The majority of data featured in this paper is derived from Statscrew, which has all counting stats as well as on base percentage, slugging, and on base plus slugging for all years of the All American Girl Progression Baseball League.

Madden, W.C., (2000)

The data that could not be found online was sourced from Madden's book, *All-American Girls Professional Baseball League Record Book.* All but a small handful of instances matched the statistics present on statscrew.

Methods

Data Acquisition and Cleaning

Data was sourced primarily from StatsCrew, supplemented with the All American Girl Professional Baseball League website and *All-American Girls Professional Baseball League Record Book*. Statscrew doesn't provide CSV downloads, and while web scraping with R and Python is certainly possible, with the there were so few seasons and players that doing it manually was faster.

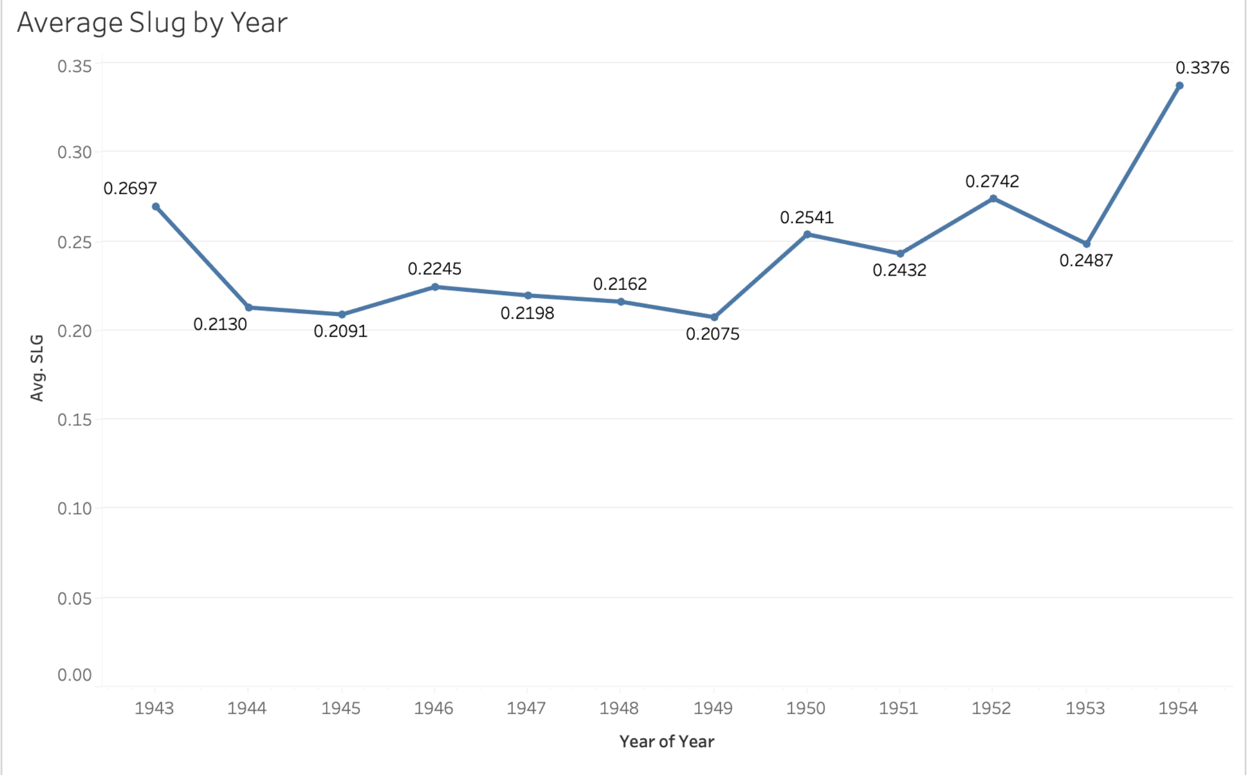
To view the original data, the file labeled as "AAGPBLcountingstats.xlsx" compiles all the stats from Statscrew with a few modifications. Cleaning the data first required inspecting players with incomplete data and either removing them or searching for the missing data. Most of these cases were players who had less than ten at bats and blank data could either reasonably be assumed to be a 0. Next, for the formula we plan to use, a column for singles needed to be calculated, which required adding a column for subtracting doubles, triples, and home runs from total hits. Another column was added indicating the year and team in which each stat line occurred, to be able to combine data in the future. Data cleaning was done on Excel as well, but can be done on R. The excel file with base level data is recalled in the R code provided for the project execution.

Execution

The traditional wOBA formula can not be used for this data. Counting stats required to run a traditional wOBA formula like intentional walks, sacrifices flies, or caught stealing are not in any surviving AAGPBL documents. Situational play by play data such as runners on base or outs when runs occurred is also missing, thus leaving us with the inability to do a run expectancy matrix. As a result, we cannot calculate wOBA. Instead, we will calculate an altered version of wOBA and predict runs that we will call Run + wOBA Lite, or RunwOBALite

$$\underline{\text{Run + wOBA Lite} = ((B1)*BB + (B2)*1B + (B3)*2B + (B4)*3B +(B5)*HR + (B6)*SB) / (AB + BB\ )}$$

The decision to include stolen bases, not traditionally included in wOBA, is in part due to how big of a role it played in AAGPBL offense and served as a distinct offensive event that had as much, if not more, weight than some other ways to reach base. For example, the highest stolen bases per game in any MLB season was in 1887 with three steals per game. Steals per game in any given AAGPBL season were often higher than four, with the highest being 6.02 steals a game in 1946. Additionally, slugging was significantly lower than it was in the MLB until the final few years. Most years, extra base hits were slim, with few players reaching double digit doubles at times. The chart below shows the average slugging percentage for the league over time. In comparison, the MLB slugging never dipped below .303 league wide.



Thus, inventing a formula and modifying wOBA to properly showcase the offense of the league is required.

The following process will be done for each season of the AAGPBL individually. To follow along with the section, there is a RMarkdown file titled “finalproject.Rmd” will walk a user through the code and why each step was done, and a file titled “finalproject.html” will do the same, but without opening R. The file titled “finalprojectcodeonly.R” will open to just the R code.

A multi regression model was used to calculate the impact of our offensive events (walks, singles, doubles, triples, home runs, stolen bases) to linear weights. To check the validity of our model, we can create a histogram of the residuals to check skew as well as check P Value and F-statistic to verify the statistical significance of the model. For all years, the histogram will skew right, but not overwhelmingly enough to invalidate the model.

We can use those weights to run our initial wOBA light formula with the total numbers of walks, singles, doubles, triples, home runs, stolen bases that occurred that year for a league wOBA. Then, we can divide the league OPS by the league wOBA to get the “wOBA scale,” the number we need to multiply our linear weights by in order to get linear weights that will produce wOBA numbers that resemble OPS. Our wOBA lite won’t end up resembling OPS a ton in the end, due to the inclusion of stolen bases, but provides still something closer to the OPB numbers. Once we have the wOBA scale and calculate our new, scaled linear weights, we can run our wOBA Lite expression with individual player performances.

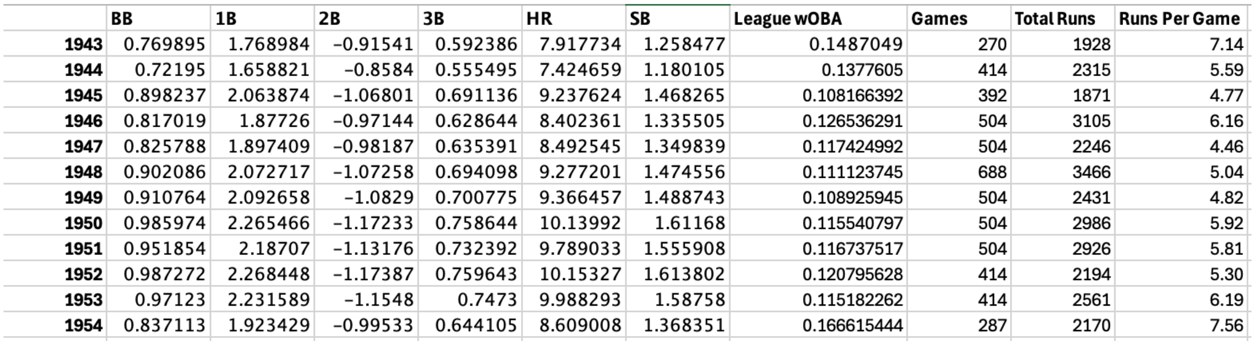
Getting individual wOBA will allow us to calculate wRAA using the following formula. Since it’s using a different wOBA, we will call it modified weighted runs above average or MwRAA

MwRAA = ((RunwOBALite – league wOBA) / wOBA scale) × PA

With linear weights, wOBA, and wRAA calculated from each player from every year, we can now move on to analysis.

Results

The coefficients used for walks, singles, doubles, triples, home runs, and stolen bases, as well as the league wOBA (before scaled), games, total runs, and runs per game are included.



	BB	1B	2B	3B	HR	SB	League wOBA	Games	Total Runs	Runs Per Game
1943	0.769895	1.768984	−0.91541	0.592386	7.917734	1.258477	0.1487049	270	1928	7.14
1944	0.72195	1.658821	−0.8584	0.555495	7.424659	1.180105	0.1377605	414	2315	5.59
1945	0.898237	2.063874	−1.06801	0.691136	9.237624	1.468265	0.108166392	392	1871	4.77
1946	0.817019	1.87726	−0.97144	0.628644	8.402361	1.335505	0.126536291	504	3105	6.16
1947	0.825788	1.897409	−0.98187	0.635391	8.492545	1.349839	0.117424992	504	2246	4.46
1948	0.902086	2.072717	−1.07258	0.694098	9.277201	1.474556	0.111123745	688	3466	5.04
1949	0.910764	2.092658	−1.0829	0.700775	9.366457	1.488743	0.108925945	504	2431	4.82
1950	0.985974	2.265466	−1.17233	0.758644	10.13992	1.61168	0.115540797	504	2986	5.92
1951	0.951854	2.18707	−1.13176	0.732392	9.789033	1.555908	0.116737517	504	2926	5.81
1952	0.987272	2.268448	−1.17387	0.759643	10.15327	1.613802	0.120795628	414	2194	5.30
1953	0.97123	2.231589	−1.1548	0.7473	9.988293	1.58758	0.115182262	414	2561	6.19
1954	0.837113	1.923429	−0.99533	0.644105	8.609008	1.368351	0.166615444	287	2170	7.56

How does the Run + wOBA lite and MwRAA scale come out, since we include base running in it and use a modified version? Here is the scale we will use.

Rating	**Run + wOBA Lite**

Excellent	>.700
Great	0.600-0.700
Above Average	0.500-0.600
Average	0.450-.500
Below Average	0.400-0.450
Poor	0.300-0.400
Awful	<.300

Rating	**MwRAA**
Excellent	60
Great	50
Above Average	40
Average	30
Below Average	20
Poor	10

Awful	0

An interactive chart to see Run + wOBA Lite per year, MwRAA by year, all time MwRAA, OPS by year, and wOBA Lite variance can be seen at the following public link: https://public.tableau.com/views/AAGPBLStatSheet/Dashboard6?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

A few main concepts can be drawn from these charts and linear weights:

- Doubles were so rare that they had very little correlation on runs. In 1150 instances of a player season, only 150 players ever had 10 or more doubles. This further puts emphasis on the stolen base.
- Joanne Weaver's 1954 season is the single best season in AAGPBL history, but Sophie Kurys has the highest wRAA of all time due to a mix of being able to play for several seasons and consistently being one of the best players when doing so. Kurys impact was previously hard to visualize due to not having base running in OPB stats, but with her multiple 100-200 stolen base seasons, she was a major driver of offense.
- A common tactic for AAGPBL players was to get on base any way possible (walks, singles, bunts) and steal bases to create offense, and both statistics pictured elevate players who did so successfully. It was a lot more likely a player would get to second by hitting a single rather than stealing rather than hitting a double, and this stat does a better job accounting for that impact.

- The impact of players like Thelma Eisen, Inez Voyce, Jaynie Krick, and Dorothy Key who did not win accolades or make all star teams during their career are better visualized.
- Home runs were extremely rare, which made players who could slug even more impactful. Betty Foss was able to accumulate 353.7 MwRAA in just four seasons due to being one of the extremely rare sluggers in the league.

Conclusion

Limitations

With more time, messing with the scaling to make it similar to OPS would be more effective/easier to visualize. The current version is not suitable for comparison to the MLB, if one is looking to do that.

Not having caught stealing is also a major limitation, as how many outs a player made on the base paths is an important component to consider if we are considering stolen bases.

Recommendations

For future recommendations, entities like the Women's Baseball World Cup and other women's baseball entities should keep better data. We still do not have play by play data for modern women's competition in baseball. Anything more complex than counting data is not available. If one wanted to calculate actual wOBA for the most recent Women's Baseball World Cup this past

July, they would not be able to. Additionally, how women understand their own talents - from slugging, to things not featured in this article like fastball velocity, is still done using the averages of MLB players. There is little data available on women's baseball for players to be able to evaluate themselves. Research into the AAGPBL shows just how unique the environment of the league was compared to their MLB counterparts and how comparing to or using statistical evaluation methods from the MLB would not have painted the full picture of a players talents and contributions. More data needs to be compiled to keep the same from happening today.

References

Weinberg, N. (2016, April 11). *The Beginner's Guide To Deriving wOBA | Sabermetrics Library*. Sabermetrics Library. https://library.fangraphs.com/the-beginners-guide-to-deriving-woba/

Klaassen, M. (2011b, January 4). Custom WOBA and linear weights through 2010: Baseball Databank Data Dump 2.1. *Beyond the Box Score*. https://www.beyondtheboxscore.com/2011/1/4/1912914/custom-woba-and-linear-weights-through-2010-baseball-databank-data

AAGPBL Players Association. (2014). *AAGPBL Rules of Play*. AAGPBL.org. https://aagpbl.org/history/rules-of-play

Tango, T., Lichtman, M., & Dolphin, A. (2006). *THE BOOK*. https://insidethebook.com/

Slowinski, P. (n.d.). *WRAA | SaberMetrics Library*. Sabermetrics Library. https://library.fangraphs.com/offense/wraa/

All-American Girls Professional Baseball League Minor League baseball statistics. (n.d.). https://www.statscrew.com/minorbaseball/l-AAGL

Major League Batting Year-by-Year Averages | Baseball-Reference.com. (n.d.). Baseball-Reference.com. https://www.baseball-reference.com/leagues/majors/bat.shtml

All-American Girls Professional Baseball League Record Book – W. C. Madden. Publisher: McFarland & Company, 2000. Format: Paperback, 294pp. Language: English. ISBN 0-7864-3747-2