Case Study: Data Analytics in Major League Baseball

Mike Ogrysko

Loyola University Maryland

DS 795: Data Science Project Design

Dr. Bu Hyoung Lee

November 8, 2022

Major League Baseball (MLB) is a multi-billion dollar-business, nearing $11 billion in gross revenues in 2022 (Brown, 2022). Its 30 teams have player payrolls that range from $61 million to $282 million with an average of $163 million (Spotrac, 2022). The average value of an MLB team is $2 billion (Ozanian & Teitelbaum, 2022). Given the enormity of these revenues and values, MLB and its teams are relying on data now more than ever to build their on-field product. MLB continues to evolve its technology for gathering game and player data and teams are increasing their expertise in data analytics, with some shifting from no formal analytics team as recently as 2014 to having a full staff in 2022 (Baumer & Zimbalist, 2014, p. 26; Groke, 2022). This case study will explore how MLB is collecting data. In addition, this analysis will discuss how teams are using the data that is collected. Finally, this case study will examine Wins Above Replacement (WAR), a metric aggregated from collected data to evaluate player performance and influence player personnel decisions.

MLB has been tracking game data since the first box score was documented in 1845 (Schwarz, 2004). Since then, the sport has evolved in the data types collected and methods used to collect data. Today's modern data tracking started in 2006 with the installation of Pitch f/x, three cameras mounted in each stadium that collected data on pitch speed, trajectory, type, and location (Dimeo, 2007). In 2015, MLB implemented its Statcast camera and radar system in 30 stadiums to track and collect data. In 2020, MLB implemented Hawk-Eye, a new 12-camera system installed in every stadium. Five of the Statcast Hawk-Eye cameras focus on pitch tracking, while the remainder track players and batted balls – tracking 99% of all batted balls (MLB, 2022). According to MLB, Statcast captures data on the following areas:

- Pitching (velocity, spin rate and direction, movement),
- Hitting (exit velocity, launch angle, batted ball distance),

- Running (sprint speed, base-to-base times), and
- Fielding (arm strength, catch probability, catcher pop time).

This equates to 25 million data points captured during a single game (Google Cloud Tech, 2022). MLB has partnered with Google Cloud to create a pipeline for collection, storage, and deployment of this data. The pipeline begins in each stadium with Google Cloud Anthos. The data is pushed to a Kubernetes cluster running in Google Cloud. From there, data is scaled and stored Cloud SQL for a PostgreSQL Database. Once in the database, the data is read into MLB's Stats API for consumption by the teams (Google Cloud Tech, 2022).

As more data is captured, teams are increasing their analytics departments within the front office and placing greater prominence on its use. For example, in 2014, the Baltimore Orioles had three employees on their analytics staff (Baumer & Zimbalist, 2014, p. 26). By the end of the 2021 season, the team had twelve full-time staff members dedicated to analytics (Meoli, 2021). Even the Colorado Rockies, a team that had no analytics department in 2014 (Baumer & Zimbalist, 2014, p. 26), has a dedicated staff of eight analyzing data today (Groke, 2022).

How are these teams using the data they collect? Some teams are using this data as a development tool to measure performance and adjust technique as necessary. For example, pitching coaches for the Boston Red Sox used video and data collected from high-speed cameras and ball trackers to diagnose biomechanical delivery flaws of two of their relief pitchers (Schrage, 2019). Players are using analytics to market themselves during free agency. In the winter leading up to the 2020 season, outfielder Marcell Ozuna sent a 43-page booklet to teams highlighting metrics where he excelled like the exit velocity of his hits and his Expected Weighted On-base Average (xwOBA) (Davidoff, 2019), which takes exit velocity, launch angle,

and sprint speed into account to create a measure of a player's offensive contributions (MLB, 2022). Teams use data about where a player hits the ball to position their defensive players on the field. In 2013, the Tampa Bay Rays and Milwaukee Brewers were estimated to have saved their teams between 10 and 15 runs over the course of the season just by shifting their players to specific areas on the field based on data they evaluated for each hitter (Common, 2014). Teams are using data to project the results of individual pitcher and hitter matchups. The Yankees estimate hitters' performances against their pitchers' specific pitches – using measured pitcher velocity and spin as well as a hitter's typical bat path – to simulate an expected result. This allows the team to map out specific scenarios and plans for each game (Carig, 2018). Teams also use this data to generate metrics that facilitate the creation of their season rosters and win projections. As an example, every team calculates a version of Wins Above Replacement (WAR) to estimate the contributions of their players (Castrovince, 2019).

The general idea behind Wins Above Replacement (WAR) as a metric is placing a value on a player that accounts for the number of team wins that he alone represents when compared to a replacement player at the same position (Castrovince, 2020, p. 189). As an example, a team comprised of only Minor League players is estimated to win 48 games. If a player from that team was replaced with a free agent, and the team's win total jumped to 54 games, the new player would have a 6 WAR (Castrovince, 2020, p. 191). Interestingly, there is no single calculation for a player's WAR. Public facing statistical reference sites, Baseball-Reference, FanGraphs, and Baseball Prospectus have their own calculations for WAR that use different metrics in their computations (Baseball-Reference, 2022). As an example, to measure defense, FanGraphs uses a metric called Ultimate Zone Rating, while Baseball-Reference uses another metric called Defensive Runs Saved (Castrovince, 2020, p. 190). The individual teams have their own WAR

calculations as well, factoring in different public inputs and other considerations such as scouting reports, budget, and other market-specific information (Castrovince, 2019). In other words, everyone (individual teams, media, and fans) uses different methods for player valuation despite the common moniker.

There are commonalities among all WAR calculations. First, all calculations employ different methods for positional players and starting pitchers. In addition, despite the different calculations, the scale is widely agreed upon (see Table 1).

Table 1

WAR Categorization for Positional Players and Starting Pitchers

Quality of Player	WAR
MVP Material	8 or higher
Superstar	6-8
All-Star	4-6
Solid Regular	2-4
Role Player	1-2
Bench Material	0-1
Triple A Material	0 or below

Note. Reprinted from *A Fan's Guide to Baseball Analytics* (p. 192) by A. Castrovince, 2020, Sports Publishing. Copyright 2020 by Anthony Castrovince.

WAR is not used as a viable metric for relief pitchers because they generally do not accrue enough innings pitched to generate a WAR that aligns with this scale (Castrovince, 2020, p. 192).

In general, a WAR model for positional players sums the following elements (Weinberg, 2014):

- Runs contributed through hitting relative to average,
- Runs contributed through baserunning relative to average,

- Runs contributed through fielding relative to average,
- Runs contributed based on position played (e.g., Shortstop, Centerfield, Catcher, etc.) relative to average,
- Runs contributed based on league adjustment relative to average, and
- Runs relative to a replacement-level player.

This sum represents the total number of runs contributed by the player beyond what a replacement player would have contributed (Baumer & Zimbalist, 2014, p. 73). It is divided by an estimate of the number of runs a team needs to score to add a single win to their total (Weinberg, 2014).

Baumer and Zimbalist (2014, pp. 72-73) provide an example using former Mets third baseman, David Wright's 2008 season and data from FanGraphs as an example:

- Contributed 44.0 runs above average to the Mets in the 2008 season,
- Cost the team 4.5 runs as a baserunner,
- Contributed 5.1 defensively,
- Contributed 2.3 runs as a third baseman, and
- Is credited with being 24.5 runs better than the average replacement player.

Wright's total number of runs beyond what a replacement player would contribute was 71.4. Baumer and Zimbalist found FanGraphs' league adjustment to be negligible, so this was left out of the total. For this estimate of WAR, they estimated 10 runs to produce a single win. Dividing Wright's 2008 total by 10, provides a WAR of 7.1 – meaning that he was worth 7.1 wins above a replacement player for the Mets in 2008.

For starting pitchers, FanGraphs sums the following metrics to calculate WAR (Weinberg, 2017):

- League Fielding Independent Pitching (FIP) less the pitcher's FIP divided by the pitcher's specific runs per win,
    - FIP focuses on the "events a pitcher has the most control over – strikeouts, unintentional walks, hit by pitches, and home runs (Castrovince, 2020, p. 109)."
    - This difference is the number of runs above average per nine innings relative to a specific league.
- Difference between an average pitcher and a replacement-level pitcher multiplied by innings pitched divided by 9, and
- League Correction.

Weinberg (2017) provides an example of starting pitcher Marcus Strohman's 2016 season using data from FanGraphs:

- American League FIP was estimated at 4.56 less Strohman's FIP at 4.12 is equal to 0.44 divided by Strohman's runs-per-win of 9.61 provides a wins-per-game-average at 0.046,
- The difference between an average pitcher and a replacement-level pitcher is estimated at 0.12 added to Strohman's wins-per-game-above-average of 0.046 is 0.166 multiplied by his 2016 innings pitched of 204 divided by nine innings gives an initial WAR of 3.76, and
- League correction is estimated at -0.14.

Summing 3.76 and -0.14 together produces a 2016 WAR of 3.6 for Marcus Strohman.

Teams generate their own version of WAR for their current players and pitchers as well as free agents. They use these numbers to build a roster, make decisions on acquisitions, and generate projections for team success. WAR is one of several metrics that teams will create with the collected data for these purposes.

As MLB continues to improve its methods and technologies for data collection, teams are using this data now more than ever to create their on-the-field products. Improving player performance, developing in-game strategy, simulating game scenarios, and projecting player outcomes are just a few ways that teams are leveraging the data available in today's game. In addition, the use of this data to create metrics to measure player performance has changed the way teams build their rosters, evaluate players, and project team success. With the increase in data availability, teams are re-engineering their front offices to leverage the data and make data-driven decisions, making them more competitive on the field and increasing their franchise values.

References

Andres, A. (2016). *Sabermetrics 101: Baseball analytics.* [Video]. YouTube. https://www.youtube.com/watch?v=wu2twZdrM-E

Baseball-Reference. (2022). *WAR comparison chart*. https://www.baseball-reference.com/about/war_explained_comparison.shtml

Baumer, B. & Zimbalist, A. (2014). *The sabermetric revolution*. University of Pennsylvania Press.

Brown, M. (2022, Apr 7). *How Major League Baseball could crack $11 billion in revenues in 2022*. Forbes. https://www.forbes.com/sites/maurybrown/2022/04/07/how-major-league-baseball-could-crack-11-billion-in-revenues-in-2022/?sh=68098b77f63a

Carig, M. (2018, Sep 27). *The Yankees have paired the best bullpen ever with cutting-edge tools to optimize its usage. Will it be enough in the playoffs?* The Athletic. https://theathletic.com/552336/2018/09/27/the-yankees-have-paired-the-best-bullpen-ever-with-cutting-edge-tools-to-optimize-its-usage-will-it-be-enough-in-the-playoffs/?redirected=1

Castrovince, A. (2019). *A fan's guide to baseball analytics*. Sports Publishing.

Castrovince, A. (2019, Feb 3). *The influence of WAR on modern front offices*. MLB.com. https://www.mlb.com/news/war-embraced-by-mlb-front-offices-c303484670

Common, D. (2014, Aug 18). *How the defensive shift and big data are changing the game*. CBC. https://www.cbc.ca/news/world/how-the-defensive-shift-and-big-data-are-changing-baseball-1.2739619

Davidoff, K. (2019, Nov 19). *Marcell Ozuna among players smartly embracing analytics in MLB free agency.* New York Post. https://nypost.com/2019/11/19/marcell-ozuna-among-players-smartly-embracing-analytics-in-mlb-free-agency/

Dimeo, N. (2007, Aug 15). *Baseball's particle accelerator*. Slate. https://slate.com/culture/2007/08/pitch-f-x-the-new-technology-that-will-change-baseball-analysis-forever.html

Google Cloud Tech. (2022). *How MLB analyzes data with Google Cloud.* [Video]. YouTube. https://www.youtube.com/watch?v=O_W_VGUeHVI

Groke, N. (2022, Aug 5). *Rockies hire a new director of analytics, familiar face in Brian Jones.* The Athletic. https://theathletic.com/3481874/2022/08/05/rockies-hire-brian-jones/

MLB. (2022). *Statcast*. https://www.mlb.com/glossary/statcast

MLB. (2022). *Expected Weighted On-base Average (xwOBA)*. https://www.mlb.com/glossary/statcast/expected-woba

Meoli, J. (2021, Dec 15). *The Orioles' analytics department under assistant Sig Mejdal keeps growing. Their ideal candidates reflect the organization's direction.* The Baltimore Sun. https://www.baltimoresun.com/sports/orioles/bs-sp-orioles-sig-mejdal-analytics-20211215-2t7aendtijfdhmgg3n3zrrspra-story.html

Ozanian, M. & Teitelbaum, J. (2022, Mar 24). *Baseball's most valuable teams 2022: Yankees hit $6 billion as new CBA creates new revenue streams*. Forbes. https://www.forbes.com/sites/mikeozanian/2022/03/24/baseballs-most-valuable-teams-2022-yankees-hit-6-billion-as-new-cba-creates-new-revenue-streams/?sh=244684d600a2

Schrage, M. (2019, July 15). *What baseball can teach you about using data to improve yourself.* Harvard Business Review. https://hbr.org/2019/07/what-baseball-can-teach-you-about-using-data-to-improve-yourself

Schwarz, A. (2004, Jul 8). *A numbers revolution*. ESPN. https://www.espn.com/mlb/columns/story?columnist=schwarz_alan&id=1835745

Spotrac. (2022). *MLB team payroll tracker*. https://www.spotrac.com/mlb/payroll/2022/?ref=trending-pages

Weinberg, N. (2014, Sept 19). *Calculating position player WAR, A complete example*. FanGraphs. https://library.fangraphs.com/calculating-position-player-war-a-complete-example/

Weinberg, N. (2017, Apr 17). *Calculating pitcher WAR, A complete example*. FanGraphs. https://library.fangraphs.com/calculating-pitcher-war-a-complete-example/