The Million Dollar Reversal: Using Baseball Data to Scout Cricket Talent

Graduate Team

August 2024

1 Abstract

This essay examines the feasibility of transitioning baseball players who are underperforming in their sport to cricket. Drawing inspiration from Million Dollar Arm, which successfully adapted cricket bowlers to baseball, the study investigates how metrics from baseball, such as RBIs and OPS, could align with cricket's scoring methods. The analysis utilizes advanced data techniques to correlate baseball hits with cricket runs and explores how elements like foul balls and pop-ups could influence cricket scoring. The findings indicate that converting baseball performance metrics could reveal potential cricket talent, offering new opportunities for players and contributing to cricket's global expansion.

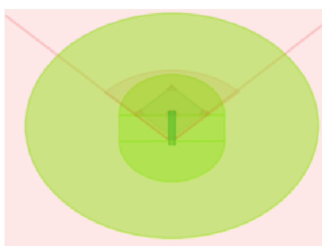
2 Introduction

Baseball and cricket, both originating from Western Europe, share several similarities in their pitching and batting approaches. For instance, while cricket allows two batsmen to be on the field simultaneously, baseball features only one batter per turn. Key performance metrics in baseball include Runs Batted In (RBIs), On-Base Plus Slugging (OPS), strikeout rate, chase rate, and fly-ball rate. In contrast, cricket evaluates a batsman's effectiveness through runs scored, batting averages, and the number of boundaries (4s and 6s) achieved. For more details on these sports and their similarities, check out this neat page on Explaining Cricket To Americans.

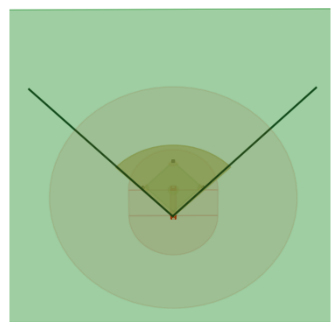
This comparison raises the intriguing question of whether skills from one sport could be effectively transferred to the other. Recent examples highlight this potential. Arjun Nimmala, the No. 20 overall pick in the 2023 MLB Draft for the Toronto Blue Jays, exemplifies such a transition. Initially a talented cricket player with a strong connection to the sport, Nimmala struggled with bowling but excelled as a hitter. His early experience in cricket contributed to his success in baseball, suggesting that skills developed in one sport could benefit performance in another. This raises the possibility that baseball players who fall short at certain aspects of their game might find success in cricket

This essay explores the potential for transitioning struggling baseball players to cricket. Inspired by the Disney film Million Dollar Arm, which successfully integrated cricket bowlers into baseball by evaluating their arm strength, we propose a reverse approach: transitioning baseball players to cricket. The original project identified cricket bowlers capable of pitching at least 80-85 miles per hour and adapted their mechanics for baseball. Despite initial skepticism, the project proved successful, leading to offers from Major League Baseball teams. This essay argues that a similar reverse approach could offer new opportunities for baseball players and enrich the sport of cricket, particularly in regions with a strong cricket culture, such as Southeast Asia and the former British Empire.

3 Data Analysis and Methodology



(a) Cricket Field with Baseball Outline



(b) Baseball Field with Cricket Field Outline

Figure 1: OverLaying The outlines of Baseball and Cricket Fields

Our study employs data analysis techniques using spatial coordinates and timestamps to assess player performance and ball trajectories. We examine how baseball singles, doubles, and triples might correspond to cricket runs by comparing player speeds and distances between bases and wickets. We also analyze the impact of foul balls and pop-ups, exploring how these baseball events could create scoring opportunities in cricket. This dataset encompasses sequences of events, labeled with specific codes denoting occurrences like a ball hit to second base or an out being recorded. It also includes detailed information about players across various farm system levels, providing insights into their performance.

To analyze this data, we interpret event codes to understand game dynamics to study player movements and ball trajectories. This approach addresses critical questions: how baseball performance metrics translate into cricket scoring, the impact of foul balls, and the significance of pop-ups. For a detailed explanation of answering the following questions and the data filtering process, refer to Appendix A.

3.1 Translating Baseball Hits to Cricket Runs:

To see how an extra-base hit in baseball might be translated into cricket runs, view the animation here. This animation illustrates an extra-base hit by showing a runner moving from first to third base. To simulate scoring two runs in cricket, the runner turns back from second base to first base. This visualization emphasizes the significance of mapping extra-base hits to calculate run values. By illustrating this comparison, the animation helps explain how baseball hits can be interpreted in the context of cricket scoring and highlights the need for further research on this topic.

To create a cricket scoring sheet based on baseball hits, we compare the speed of batters rounding bases in baseball with the time required to run between wickets in cricket. We filter the data to include only instances where batters safely reached a base, estimating how different types of baseball hits—singles, doubles, and triples—correspond to cricket runs. However, distinguishing between these types of hits proved challenging due to ambiguities in the dataset, such as the ball's landing position and fielding team actions. These issues, along with instances where runners were thrown out at first base potentially affecting speed estimates, are explored in more detail in the Appendix

3.2 Analyzing Foul Balls:

Foul balls in baseball, while less valuable as they do not contribute to scoring, can create additional scoring opportunities in cricket. Unlike baseball, where foul balls remain in play only within the diamond, a foul ball in cricket can potentially lead to extra runs if it reaches the boundary or if the batsman can score additional runs before the ball is retrieved. Referring to Figure 1, The larger size and different layout of cricket fields allow for more scoring possibilities from hits that would be considered foul balls in baseball. This analysis includes will aim to evaluate players who hit the most foul balls and how this might translate into cricket scoring.

A limitation of this analysis was the lack of detailed information on where the ball lands. The ball position dataset did not specify locations for foul balls, leading to assumptions that these hits would be considered in play for cricket, which may not always be accurate. This may affect the analysis of the ratio of foul balls to plate appearances.

3.3 Identifying Pop-Ups and Their Cricket Equivalent:

The third question explored how baseball pop-ups could be translated into sixes in cricket. Figure 1A illustrates this comparison by overlaying the dimensions of a baseball field with those on a cricket field. In Figures 1A and 1B, the home plate is positioned over the cricket wicket, and the boundary of the cricket field is shown extending beyond the baseball field's infield. This visual aid helps us understand how balls landing beyond the boundary line on the cricket field might correspond to sixes.

To analyze pop-ups, we need to compare the final position of the baseball to the dimensions of a cricket field, which has a diameter of approximately 450-500 feet. We set conditions to determine whether a pop-up exceeded the distance threshold for a six, using timestamps to track the ball's final vertical position, which needed to be above 200 feet. However, our analysis faced limitations, including missing player IDs in the dataset, which affected accuracy. Additionally, our method inadvertently included home runs in the results, which could skew the findings by showing plays where the ball was in play but also attributing home runs to players.

4 Results

play	double	single	triple
season_level			
1883_1A	3	1	4
1883_2A	3	1	4
1883_3A	3	1	4
1883_4A	3	1	4
1884_1A	3	1	4
1884_2A	3	1	4
1884_3A	3	1	4
1884_4A	3	1	4

Figure 2: Baseball Hits and Its Approximate Runs Scored in Cricket

4.1 Speed Analysis:

Using Matplotlib and the Pandas library, we analyzed the speeds of batters for different types of hits. Results revealed that players ran the slowest on singles and the fastest on triples, with the speed for triples aligning closely with the league average of 25 feet per second. We compared these speeds to the distance between cricket wickets, similar to the distance between home plate and the pitching mound in baseball. By converting these distances, detailed further in Appendix A, we estimated that baseball singles, doubles, and triples could be roughly translated into one, three, and four runs in cricket, respectively. Figure 2 illustrates that all teams in the baseball farm system had similar scorecards for each hit due to the relatively uniform speed of the runners.

Levels	total_plays	total_fouls	Other_plays	foul_pct
Home1A	8752	1469	7283	16.784735
Home2A	3090	541	2549	17.508091
Home3A	5345	940	4405	17.586529
Home4A	8972	1395	7577	15.548373

(a) Data Set 1883 Season

Levels	total_plays	total_fouls	Other_plays	foul_pct
Home1A	16042	2884	13158	17.977808
Home2A	16502	3060	13442	18.543207
Home3A	16987	2879	14108	16.948255
Home4A	19272	3086	16186	16.012868

(b) Data Set 1884 Season

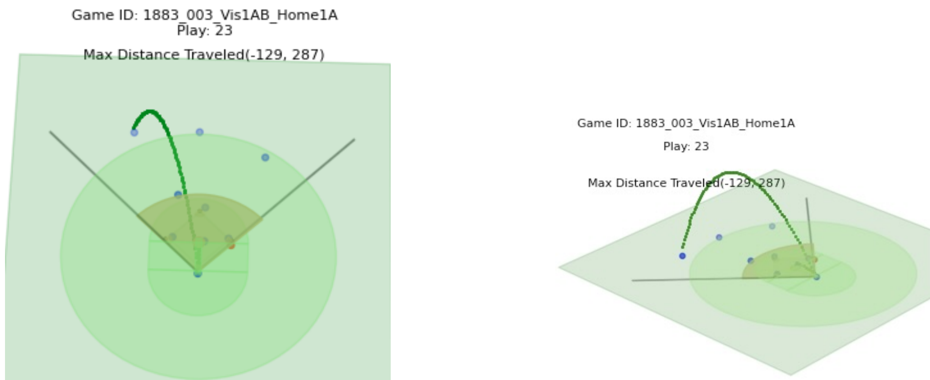
Figure 3: Foul Percentages for the entire Farm System

player_id	Total_Hits	Total_in_play_hits	total_fouls_hit	foul_rate_pct
586	768	584	184	23.958333
547	734	568	166	22.615804
702	665	505	160	24.060150
993	634	490	144	22.712934
495	624	459	165	26.442308
492	609	453	156	25.615764
798	601	455	146	24.292845
892	596	427	169	28.355705
963	567	412	155	27.336861
475	540	375	165	30.555556

Figure 4: Top 10 Batters Who Hit the Most Foul Balls

4.2 Foul Balls:

Figure 3 shows a correlation between higher foul ball percentages and increased run-scoring in cricket, where foul balls can contribute to the score. Conversely, Figure 4 lists the top ten players with the most foul balls over two seasons. Player ID 586 leads with 184 foul balls, while Player ID 547 follows with 166. Despite having the highest number of total hits, Player ID 586 has a relatively low foul ball percentage, suggesting that this player is effective at making contact with the ball. In contrast, Player ID 475, with 540 total hits and 165 foul balls, might need to assess their playing strategy, as their high foul ball count could indicate challenges in their hitting technique. Although foul balls have less significance in baseball scoring, in cricket, where they can contribute to the total score, the data implies that players like Player ID 586 could potentially perform well and players like Player ID 475 might look into changing to cricket.



(a) Flyout Front View (b) Flyout Side View

Figure 5: Flyout View from Two Angles

player_id	Total_Plays_Above_200ft	Total_HRs	Total_Flyouts	home_run_pct
973	39	2	37	5.128205
702	36	12	24	33.333333
626	33	3	30	9.090909
798	31	3	28	9.677419
630	30	4	26	13.333333
442	30	7	23	23.333333
586	30	3	27	10.000000
835	29	8	21	27.586207
565	29	1	28	3.448276
492	29	3	26	10.344828

Figure 6: Home Run Percentage Rate

4.3 Pop-Ups and Sixes:

Figure 5 highlights a comparison between baseball and cricket scoring by showing a play from the 1883 baseball season. In baseball, a flyout does not advance the runner and is generally considered less impactful. However, in cricket, a flyout that crosses the boundary is scored as a six, reflecting its high value. This comparison suggests that baseball players who excel at hitting flyouts, even if they don't hit many home runs, might have an advantage in cricket. Data from Figure 6 supports this idea: Player ID 973, who had 37 flyouts exceeding 200 feet and only two home runs, accumulated the most potential sixes in cricket. Conversely, Player ID 702, who also hit many flyouts over 200 feet, recorded the highest number of home runs. This suggests that while Player ID 702 may continue to excel in baseball, Player ID 973 could potentially achieve greater success in cricket. Overall, these findings indicate that baseball players with a high frequency of flyouts could leverage their skills to gain a competitive edge in cricket, where such hits are particularly valuable.

5 Conclusion

Transitioning struggling baseball players to cricket presents a promising opportunity for both the players and the sport. By analyzing baseball performance metrics and translating them into cricket scoring systems, we can identify potential new talent for cricket. A relevant example is Travis Bazzana, who was the top draft pick in the 2024 MLB draft. As an Australian who played cricket alongside baseball, Bazzana credits his batting success to the consistent contact skills developed in cricket. His experience suggests that transitioning from baseball to cricket could offer new opportunities for other players facing challenges in baseball.

Additionally, Jomboy Media, a prominent figure in the baseball community, has shown a growing appreciation for cricket. Jomboy noted, "Baseball is pretty brutal to watch when it's amateurs, but cricket, I could put on any league and I could enjoy the action. But that might be just because I'm not a snob yet." His engagement with cricket, particularly during the COVID-19 pandemic, and his efforts to educate sports fans about the game, reflect a rising interest in cricket within the U.S. Moreover, baseball teams like the Toronto Blue Jays, which host annual cricket nights, and events such as New York Yankees shortstop Didi Gregorius participating in cricket activities, highlight the increasing crossover between these two sports

As Major League Cricket gains traction in the US, exploring cricket metrics and player transitions could provide baseball players with new opportunities while enhancing cricket's global appeal. Such cross-sport talent transfers could benefit both sports and foster stronger international connections in the athletic world.

Appendix A

5.1 Data Filtering Prerequisites

To perform the Data Filtering Process efficiently, we utilize Pyarrow and Pandas, two powerful Python libraries. Pyarrow excels at quickly reading and writing large data files, while Pandas provides robust tools for data manipulation and analysis. By combining these libraries, we can effectively navigate extensive datasets and extract the specific information we need.

Our datasets include Game Information, Game Events, Player Positions, and Ball Positions. The first step in our process is to organize the data by Season and Level within the farm system. This organization facilitates better visualization and interpretation of the data. After organizing the data, we anticipate working with sizable datasets, particularly when examining each dataset individually.

The next stage involves filtering these datasets to isolate crucial variables. We will focus on Play IDs, which are integer values representing specific game events. For instance, Play ID 1 might indicate a pitch resulting in a strike, while Play ID 2 could denote a pitch leading to a foul ball. Additionally, we will extract Final Timestamps associated with each Play ID, as these timestamps are essential for answering our key questions. For the Player Position dataset, we will also select the Position Player ID of interest.

In the Ball Position and Player Position datasets, we will extract the final spatial coordinates at the Final Timestamp, along with the previously mentioned variables. This step is crucial before moving on to the next stages of our analysis. By focusing on these key variables, we can lay a solid foundation for addressing our analytical questions and performing more effective analysis to derive meaningful insights.

5.1.1 Speed Calculation

When it comes to calculating the speed of runners, we start by isolating the player of interest from the filtered Player Position dataset, focusing solely on the batter. We then create a new variable, "B," which includes all timestamps from the Game Events dataset. Our goal is to identify the Start Timestamp when a ball is hit into play by the batter, marking the moment the batter begins running. We will use event code "4" to filter games that include this event, as detailed in the metadata section of the Appendix.

With this information, we will merge the Player Position dataset with the timestamps in "B" to obtain the start and final times for plays matching our Play IDs and games where the batter stops at the final recorded location. The final step involves calculating speed by iterating through the merged dataset, computing the total distance traveled using spatial coordinates from the Player Position dataset, and dividing this distance by the total run time, determined by the start and final timestamps. This process will be repeated for all three types of infield hits, and users can select which play to analyze for speed calculations.

5.1.2 Foul Ball Count Calculaiton

To analyze foul balls, we use the filtered Game Events dataset. A foul ball is indicated by a sequence of event codes: 1 (pitch), 4 (ball hit into play), and 5 (end of play). By filtering for games that include this sequence, we compile a dataset of all games in the season where a foul ball was hit. Next, we filter the Game Information dataset to extract Batter IDs, Play IDs, and the Games being played. This step allows us to identify the players responsible for hitting foul balls.

After filtering, we merge the Game Events dataset with the Game Information dataset to match Game IDs, Play IDs, and Batter IDs. This merge provides a clear representation of how many foul balls were hit per game. Finally, we sum the total number of unique Batter IDs to determine how often each player hit a foul ball throughout the season, saving the aggregated data for further analysis.

5.1.3 Flyouts to Sixes Calculation

For analyzing fly outs, we use the Ball Position dataset along with the given Play IDs and games to extract final landing positions beyond the boundary. This filtered data is then merged with the Game Information dataset to create a new dataset that includes Batter IDs for balls hit beyond the 200 feet boundary. We finalize the analysis by summing the number of times each Batter hits the ball beyond this boundary and saving the data for further examination.

References

1. "CRICKET for BASEBALL PLAYERS." *Seattlecricket.com*, 2024, www.seattlecricket.com/history/crick.htm. Accessed 9 Aug. 2024.

2. "From a Cricket Field to the Top of the MLB Draft? Inside Arjun Nimmala's Unique Draft Journey." *ESPN.com*, 29 June 2023, www.espn.com/mlb/story/_/id/37929232/mlb-draft-2023-arjun-nimmala-cricket-baseball.

3. Farrell, Melinda. "Meet a Baseball Guy Making Cricket Fun in the US ahead of the T20 World Cup." *Al Jazeera*, Al Jazeera, 24 May 2024, www.aljazeera.com/sports/2024/5/24/who-is-jomboy-cricket-commentator-baseball-analyst-icc-t20-world-cup-2024. Accessed 9 Aug. 2024.

4. NZ, Sky Sport. "BASEBALL V CRICKET: New York Yankee Tries Out Cricket — SKY TV." *YouTube*, YouTube, 27 Jan. 2017, www.youtube.com/watch?v=gOIJd_oQC2s.

5. Savage, Nic. ""It Was Unbelievable": Aussie Cricket Prodigy's Sliding Doors Moment before MLB History." *Fox Sports*, FOX SPORTS Australia, 16 July 2024, www.foxsports.com.au/cricket/it-was-unbelievable-aussie-cricket-prodigys-sliding-doors-moment-before-mlb-draft-history/news-story/43b42f7469c61feed585c3b8bf620fb8. Accessed 9 Aug. 2024.

Acknowledgments

First and foremost, I would like to thank the organizers at SMT for providing me with the opportunity to gain experience on this Data Science project related to a sport I deeply love. I am especially grateful to Dr. Meredith Wills for her invaluable help, mentorship, and encouragement throughout this challenge. I would also like to express my gratitude to Kenny Bores for welcoming me and serving as an additional resource.

I appreciate Sahana for contributing to the foul ball calculation and for offering insightful solutions to coding issues with the pandas library. My thanks also go to Eddie Dew for supplying excellent source code that enabled me to view animations in two dimensions. Lastly, I would like to acknowledge the Data Challenge Participants in the Slack channel for their insights and assistance with various coding questions.