Run Value of MLB Swing Decisions Derived from Player Strengths

Johnny Nienstedt[1]

[1] *University of California, Santa Barbara*

(Dated: March 19, 2024)

INSTALLATION AND USAGE

This project requires one library which is not installed on a PHYS 129L Raspberry Pi: the pandas library, which can be installed by running *pip install pandas* from the terminal.

This project requires no external hardware to run. Since the data acquisition scripts take several hours to run, I have included copies of the resulting csv files which will allow the main script to run posthaste.

INTRODUCTION

For those not familiar with the intricacies of the sport of baseball, I will now share a brief explanation of the impetus behind this project. Hitting a baseball is often described as the single most difficult thing to do in sports, and for good reason; however, there are certain things that batters can do to increase their odds of making solid contact. The first and most obvious of these is to swing at 'good' pitches, and take (don't swing at) 'bad' pitches. But how to define good and bad pitches?

For the most part, in the public sphere at least, this distinction is made via the strike zone, the region above home plate extending from the hollow of the batter's knees up to the point midway between the belt and the letters on his jersey (yes, that is the official rulebook definition).

So, publicly available swing-decision statistics are generally limited to evaluating how often the batter swung at pitches inside and outside the strike zone, where the former is far more preferable to the latter. But there is an obvious flaw in this assumption: not all batters find the same pitches easiest to hit. Some batters prefer the ball inside, some outside; some high, some low. Therefore, swing-decision statistics should be based on how often the batter swings at pitches that *he specifically* can and cannot hit, regardless of whether they are in the strike zone.

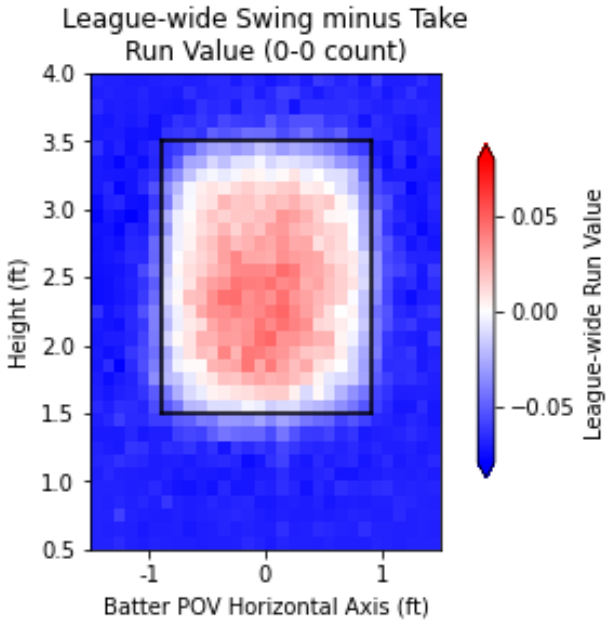

FIG. 1: *League-wide run value heat map for 0-0 pitches from 2021-2023. The strike zone is overlaid in black.*

The previous sentence is the stated goal of this project. To achieve that goal, I needed to compile a truly staggering amount of data; this is, by my estimation, the main reason why such a statistic is not publicly available. First, in the script *league_data.py*, I scraped data from baseballsavant.com on every single pitch over the last three years, and found the baseline run value (per pitch) based on location and count (see FIG. 1. 'Run value' may seem like a bewilderingly arbitrary statistic, but it is actually very elegantly defined:

based on the count, one can quantify how many more runs the offense is likely to score depending on whether the pitch is a ball, a strike, or put into play. If you're interested in reading about this in more detail, this article is a good place to start.

The player data was much more involved. Since even an everyday player only swings at a few thousand pitches per year, I couldn't use a 30x35 grid as I had for the league data. However, I still wanted a smooth distribution to get an adequate sensitivity to areas of strength and weakness. To work around this, I used data from 14 large zones, and then solved Laplace's equation to smooth out the distribution (see FIG. 2).

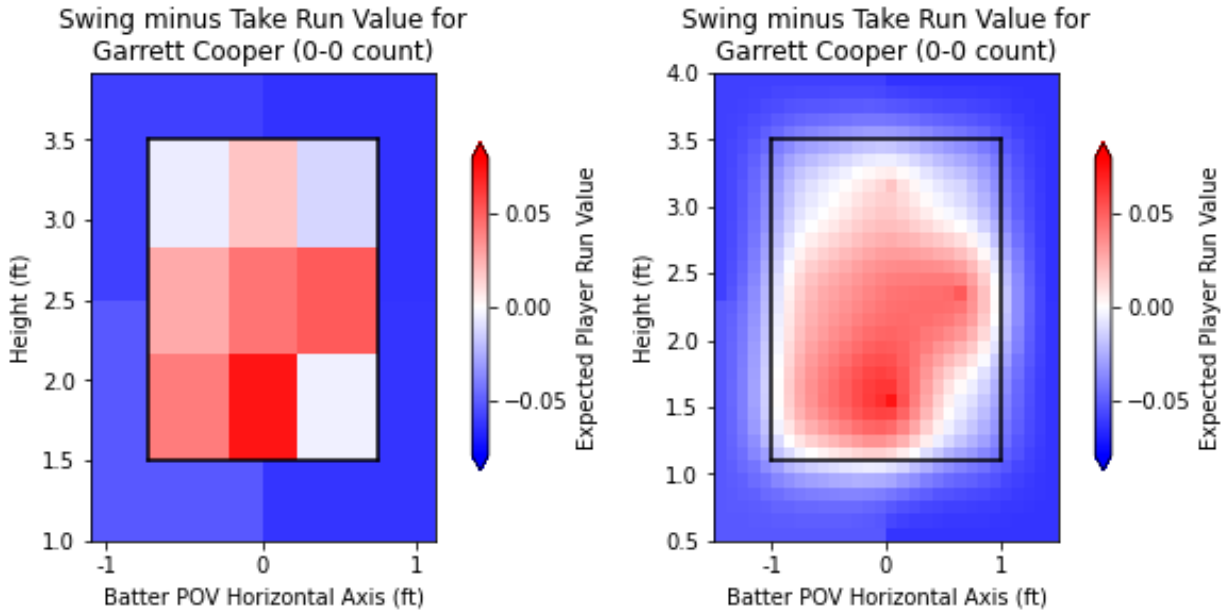

FIG. 2: *Raw and merged run value heat maps for Garrett Cooper of the Miami Marlins. We can see that Cooper likes the ball low and inside and has weak spots up and in, up and away, and low and away.*

The zone data was relatively easy to acquire (see *player_data.py*), but storing the solutions to so many differenital equations – 640 players times 12 counts times a 30x35 grid – took almost a gigabyte just in csv form (see *player_heatmap.py*).

Then, to actually make the swing-decision evaluations, I had to determine which pitches the batter swung at, and which he didn't (see *swing_take.py*). For each pitch he swung at, I assigned the corresponding swing run value, and for each pitch he took, I assigned the corresponding take run value (see FIG. 3). The values were cumulatively summed for each and every pitch thrown to that batter over the last three years, returning the total value, in runs, that the batter contributed to his team solely from their swing-decisions, irrespective of their other batting skills.

RESULTS

You can view the results of my analysis using the *project.py* program. The program will first display the league average baseline from FIG. 1, and then prompt you for a player. You should be able to enter either a first or last name (or both), but suffixes and multi-word last names do make things tricky. If you don't know any players off the top of your head, that's fine; you can simply enter 'random,' and the program will choose a player for you. It will display that player's zone and differential delta RV heat maps for 0-0 by default, after which it will display that batter's swing-decision statistics, both in terms of total run value, and in terms of a percentile ranking. Then the program will allow you to choose a different count and/or action (swing, take, or delta), or you can simply exit and start over with another player.

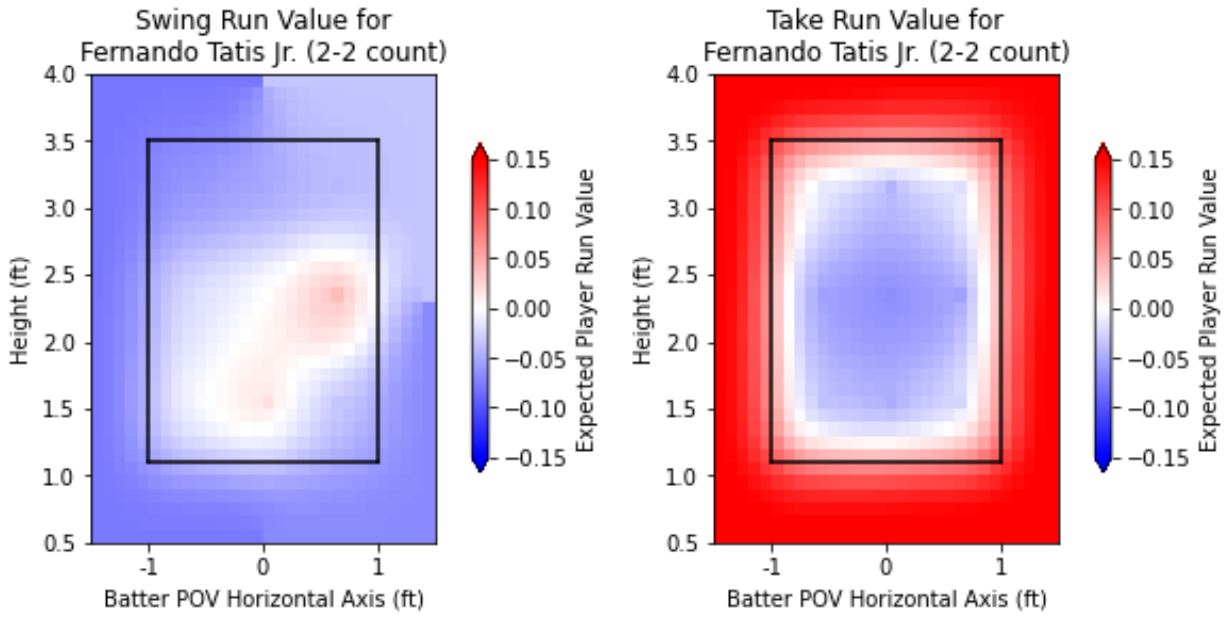

FIG. 3: *Swing and take run values for Fernando Tatis Jr. in a 2-2 count. For any given pitch, if he swings, he is assigned a value from the left plot, and if he takes, he is assigned a value from the right plot. While it may seem that Tatis shouldn't swing at all (since there is very little red in his swing RV heatmap), this is offset by the negative value of taking a pitch for strike three; this is why the delta heatmap is much more visually useful. We can also see from the scale that a 2-2 count is much more pivotal than a 0-0 count, comparing to FIG. 2.*

The statistics won't change with count, as they incorporate all pitches from all counts. The full rankings can be found and sorted in the *player_st.csv* file. The top three players by swing-take runs are disciplined mashers Matt Olson, Ronald Acuña, Jr., and Juan Soto, while the bottom thee are noted free-swingers Javier Baez, Oscar Gonzalez, and Cody Bellinger. This is a good sign that my statistic has been properly calculated.

These results are very valuable for player evaluation, because they highlight a skill that is not captured well by other publicly available statistics, as explained above. My original hope was to compare this statistic to the primary available swing-decision metric, z-o swing%, but I ran out of time. I will continue that work over the next few weeks. That will be beneficial for determining which players are truly able to descriminate between pitches that they can and cannot hit versus simply strikes and balls. Nevertheless, these results are still a good indicator of selective aggression, and may be useful for identifying players primed for breakout or bounceback seasons. Also, players who make good swing-decisions are more likely to age gracefully, since they are less reliant upon their ability to make powerful contact, and can still provide value from their plate discipline even as the former ability fades.

OTHER NOTES

Unfortunately, the of the two main portions of this project, one (the data collection) is impractical for you (the grader) to run due to time constraints, and the other (numerical solving) turned out to be a lot easier than I expected. Nevertheless, please understand that I put in a huge amount of work on this project and I am very proud of the final product. With a few modifications, I will be including this in my application portfolio for MLB teams.