Using Estimated Expected Runs to Evaluate Player Base Running and Fielding

1. Abstract

Estimating expected runs could provide us a different and effective angle in player evaluation. Based on different player and ball locations, we could calculate expected runs from various aspects of the baseball game, including batting, pitching, and fielding. With that result, we design a formula to improve player evaluation.

2. Introduction

As a team from a country that knows little about baseball, we would love to see baseball become popular in our country. Even though we know baseball well, when we introduce it to other people, they always say, "The rules of baseball are too complicated to study." It is hard for a rookie to evaluate a player just based on data. Therefore, our research goal is not only to complete this project, but also to provide a most intuitive way for novice fans to evaluate players based on data. Because for rookies, they just want to know who the most powerful player in the league is.

Our research question is what would you estimate expected runs throughout the course of a play based on player and ball locations, and how would you use that to evaluate player base running and fielding? Expected run is an interesting statistic. Conceptually, the expected run can be recognized as an average run. It means that in the long term, the average runs that a play can score. For example: the expected run when the ball reaches the right field means that if we encounter millions of the same situation, the ball in the right field, how many runs we could expect for a single play on average. In a similar manner, we could separately estimate the expected runs from pitching, batting, and fielding with the help of ball and player tracking data, and the higher expected run could be considered as better batter in offense but worse pitcher and fielder in defense and vice versa.

3. Method and Result

1) Estimate outs and runs in each play

In the first step, we choose to estimate the number of outs and runs. Since we don't have play-by-play data, we need to use player and ball tracking data to extrapolate the out counts and run counts of each play to continue the analysis and get our desired results.

2) Calculate expected runs based on base running and out counts

In this part of analysis, we make a very detailed distinction in the calculation. As shown in the resulted form, we list different situations in base running (different base loaded and different out counts) and calculate the expected runs separately. Because we think different situations will lead to different things happening, such as the changing mentality of the players or the tactics of the team.

	0 out	1 out	2 outs
Base 1	0	0.011583	0.0087912
Base 2	0	0.0043859	0.0143799
Base 3	0.066667	0.0574713	0.1171994
Base 1 and 2	0	0.00566	0.0132626
Base 1 and 3	0.185185	0.083969	0.146067
Base 2 and 3	0.333333	0.17	0.146771
Base 1, 2, and 3	0.1463415	0.0960452	0.172949

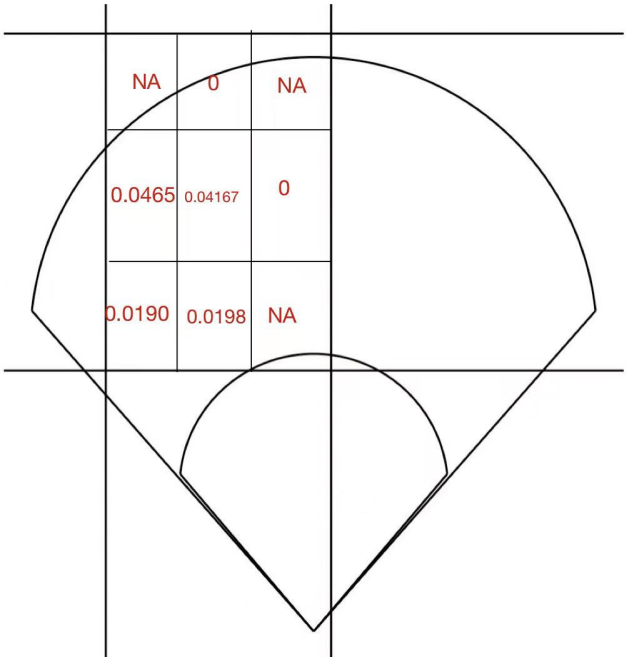
According to our results we calculated, we found that more bases loaded and more bases close to the home plate will potentially lead to a higher chance of a run. We see a trend that when teams have 2 outs, in some situations they will have higher expected runs than situations when teams have 1 outs or 0 outs. A possible explanation for this could be when 2 outs are reached, batters

would focus more and tend to swing and hit the ball instead of hesitating about whether a good or bad ball is pitched.

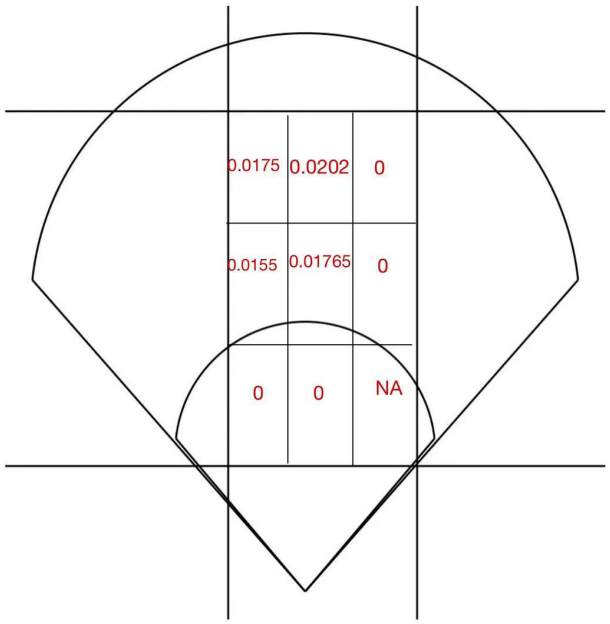
Another thing to notice is that there are some “0” s in the table. Our guess is that the data that we got contains too few plays in those situations, and unfortunately, none of those plays ends with a run. If we could get more data, the result will be closer to the truth.

3) Calculate expected runs based on fielder initial lineup when play starts

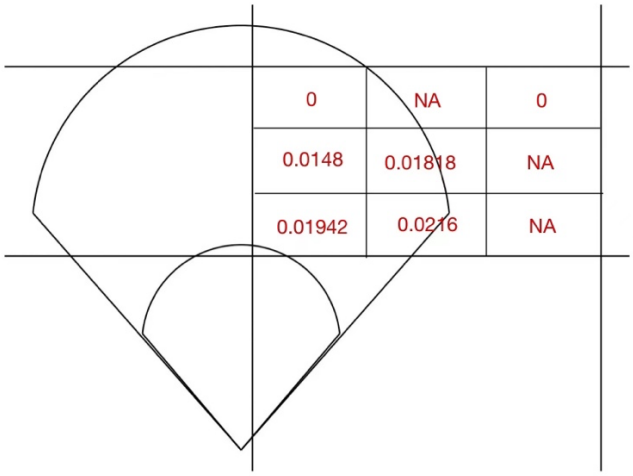
Our initial inspiration to study this subtopic is that we think a different fielder lineup when each play starts could potentially influence runs achieved by batters, since coaches or players might have some specific strategies in fielding defense. To that end, we separately analyze the initial lineup of left, center, and right fielders.



• Expected runs of left fielders based on initial lineup



• Expected runs of center fielders based on initial lineup



• Expected runs of right fielders based on initial lineup

According to the diagrams, we can conclude:

For the left fielders, getting closer to the Homeplate a bit, to the down left and down middle areas on the diagram, would be better, since the expected runs are lower in down left

(0.0190) and down middle (0.0198), which means that a more Effective Defense would happen if the fielders were initially set in that area, and the runners are not as easy to get a run.

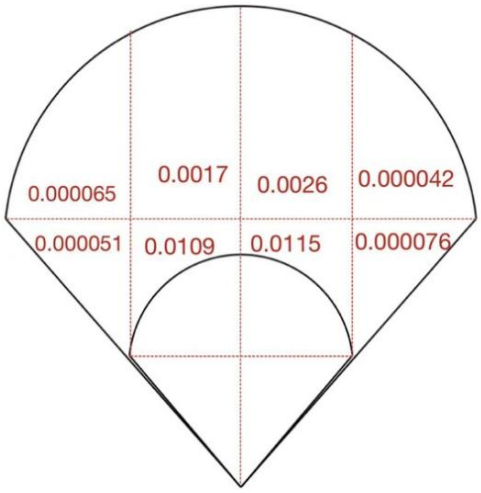
Same for the center player, they need to get closer to the Homeplate, to the middle left and middle areas, because of the lower expected runs (0.0155 and 0.01765).

For the right player, getting farther away from the Homeplate a bit, to the middle left and middle areas, would be better for the similar reason of low expected runs (0.0148 and 0.01818).

Once again, we notice some "0" s. 0 doesn't mean it's impossible to get a run; it's very likely because too few plays happened in that area, which unfortunately happens to have no runs achieved and leads to a large noise that influences the result. In this situation, we could simply ignore the appearance of 0 and consider only those meaningful values, and all the areas we've talked about are based on those values rather than 0 or NA.

4) Calculate expected runs based on ball position—fielding

Left off from the expected runs of player position, we continue to calculate the stats with ball position provided. The first part is fielding: we divide the entire field into 8 sections, and we calculate the expected runs in each section if the ball, in any play, goes through that section.



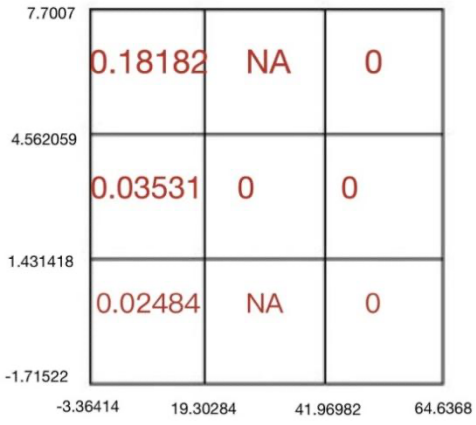
• Expected runs based on ball position in different sections of field

Based on the result, we can compare the expected run according to the field as the ball moves with the expected run according to the fielder initial lineup which we've discussed before. Our conclusion from the "fielder initial lineup (player position) part" is that the LEFT fielders and the CENTER fielders can get CLOSER to the Homeplate initially due to a lower expected run (a better defense) achieved as they get closer to the Homeplate; the RIGHT fielders can get FARTHER AWAY due to the same reason. In "fielding (ball position) part", we can also see on the left side, the area below has a lower expected run than the area up, which means when the ball reaches the area below, the fielder can defend on it better. Similarly on the right, the area up has a lower expected run than the area below, which means that the fielder can defend on the ball better if it reaches the area up. Both left and right correspond to the conclusion we've discussed. But for the center area, both areas up have lower expected runs than the areas below. There are two possible explanations for this difference in conclusion. First, the division of field is not exactly the same, which means that the expected runs could be different if we change the sections. Second, center fielder is believed as a position that needs to move the most. In other words, we can have our center fielder standing closer to the Homeplate initially in the given section, but during the game, they need to move fast up to catch high and far fly balls or down to catch the short but fast balls. So, the initial lineup we've mentioned before actually gives us a balanced position, where center fielders don't need to run too far away to catch both high and short balls. Moreover, unlike the high balls which might end up with a catch out, the short balls are harder to defend, which also explains why the center down areas have higher expected run than the center up areas.

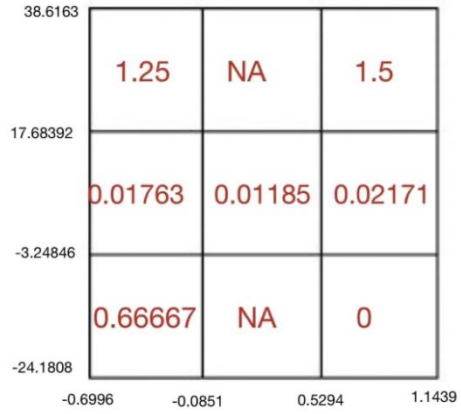
One more thing to notice is that the center areas tend to have higher expected runs than the left or right. We believe short but fast balls in the middle, which are difficult to defend, could be one reason for that. For left and right areas, the high balls easily end up catching out, and fast balls are easily out of bounds. But we can utilize such a difference. For the coaches' tactics, if some fielders have obviously bad ability in defense, they could be assigned to left or right areas, since those areas, based on our discussion, have lower expected runs naturally (very few batters can get a run easily if the ball are in those areas).

5) Calculate the expected runs based on ball position—pitching

The other topic that we analyzed using ball position data is pitching. We want to figure out whether different locations of pitching would lead to different results of expected runs. Here we sort all the plays into two categories: the batter hits the ball, and the batter doesn't hit the ball.



• Expected runs when the batter hits the ball



• Expected runs when the batter doesn't hit the ball

Based on the result, we found that there are some interesting points to notice. First, when the ball is not hit by the batter, the left upper and right upper areas have larger expected runs, which could potentially be because of walks or stealing bases. But we think it's also probably because

of the lack of data, since when we are dealing with those areas, we see very few plays, and some of them happen to end up with a run scored, which adds noise into our study. So, like we've mentioned before, more data would bring us closer to the truth. Second, if the pitcher MUST throw a good ball in the next play, he could aim to the left middle or left down area (the coordinate is shown to see the exact range) due to the lower expected runs.

Once again, we notice some "0" s. As we've mentioned, we believe the potential reason for all those "0" s is lack of data, since there are too few plays that satisfies the position where the ball is pitched, and none of them happen to end up with a run. But we cannot say that if we pitch toward that location, nobody will score a run. It's only because of the noise brought by the lack of data.

6) Formula & ways to evaluate players

Based on all the different types of expected runs we've calculated and estimated before, we can utilize them to build a novel formula or way to evaluate players

The first part is to evaluate player base running, and we've designed a formula:

$$Batter\ Ability\ Score = \frac{Average\ Expected\ Runs\ of\ Batter}{Average\ Expected\ Runs\ of\ Pitching * Average\ Expected\ Runs\ of\ Fielding}$$

Basically, the formula is built to address the potential influencers or confounders when we evaluate batter abilities—pitching and fielding, since a bad pitching or fielding will make batting look good. In this case, we can utilize the result in the table of Part 2) to calculate the numerator, Average Expected Runs of Batter, result in Part 5) to calculate Average Expected Runs of Pitching, and result in Part 3) and 4) to calculate the Average Expected Runs of Fielding.

$$Average\ Expected\ Runs\ of\ Batter = \frac{Total\ Expected\ Runs\ in\ Different\ Baserunning\ Situations}{Total\ Plays\ of\ the\ Batter}$$

$$Average\ Expected\ Runs\ of\ Pitching = \frac{Total\ Expected\ Runs\ in\ Pitching}{Total\ Plays\ of\ the\ Batter}$$

$$Average\ Expected\ Runs\ of\ Fielding$$

$$= \frac{Average\ Expected\ Runs\ in\ Fielder\ initial\ lineup + Average\ Expected\ Runs\ of\ Ball\ Location\ in\ Fields}{2}$$

For instance, we calculate the Batter Ability Score of Player No. 7225:
Player No.7225 = Average Expected Runs of Batter/ (Average Expected Runs of Pitching* Average Expected Runs of Fielding) = 4.50823

This formula can balance the ability of pitching and fielding when we evaluate players and help us get a "score" for each player. When the batter is against a good pitch with low expected runs, his ability score will be reasonably balanced to be larger. The higher score that a batter can get, the better baserunning ability that the batter has.

The second part is to evaluate fielders. The method we have is to compare fielder average expected runs (the average expected runs of initial lineup and ball position in the field) with the average of their "zone" in Part 4).

For instance, we've calculated the average expected runs of two center fielders:
Player No.5535 (center fielder) = 0.015105
Player No.1177 (center fielder) = 0.006197
Average Expected Runs of Center Field = 0.006675

As we compare those values, Player No.5535 has a higher average expected runs than the average value of the center field, whereas Player No.1177 has a lower average expected runs. In this case, lower expected run means higher fielding ability and better defense ability to limit run scored.

Appendix

Methodology, Acknowledgement, and Discussion

• Our research is conducted using RStudio and MATLAB to achieve data manipulation, calculation, and visualization.

• Specifically in the result of the situation when the batter doesn't hit the ball in Part 5) of pitching, the coordinate range is being calculated using 1st and 3rd quartile value, since the max and min value is too extreme, and we think that it is probably because of prevention of stealing bases.

• There are several things, we think, that we can improve on. First and foremost, we could potentially work on more data, since in several parts of analysis, for example, the calculation of expected runs based on base running and pitching, we get some results of "0" s. As we've discussed in the main part of the report, the "0" s don't necessarily mean that it's impossible to score a run in those situations; it's because the data that we have is not enough, and they happen to end up with no run scored. Therefore, with more data in our hand, we could reduce the noise in the study and provide a better result. Secondly, algorithms cannot account for everything. For instance, since we don't have the play-by-play data, our counting of out and run might be slightly inaccurate. When we are conducting the part of counting using written algorithms, we see some out counts larger than 3. Maybe the batter is injured, and players are changed, but the algorithms count that as an "out". It would definitely be better to have the play-by-play data, but we think if not, there is a room where we can upgrade our algorithms.