Closing the Gap: Outfield Optimization

https://github.com/robo515/SMT_Data_Challenge_2023_Final

Abstract:

In baseball, the margins between wins and losses can be incredibly slim. In the era of the shift ban, analysts looking to improve baseball defense are left with only outfield positioning to optimize. This paper looks to optimize outfield positioning to secure outs in pursuit of wins, and in the bigger picture, a World Series Championship.

To determine issues with current outfield positioning, we applied the K-means clustering algorithm to categorize all outfield balls in play (OBIP) into three clusters: left, center, and right field. These clusters guide outfielders to optimal starting positions that minimize the distance they need to cover for each batted ball within their assigned cluster. We determined the issue with current positioning is that the right and left fielders often play too wide and too deep, allowing preventable hits in front of them.

Outfielders, especially those in right and left field, can improve their performance by adjusting closer to center field and home plate. By moving outfielders to their optimal positions, we project an increase in total outs generated. While some potential outs may be lost due to a change in position, the net gain is substantial, resulting in 76 additional outs over a season, with only 61 outs lost, yielding a net positive of 15 outs.

This paper underscores the value of data in improving outfield defense, offering teams a tangible advantage as they vie for the title of World Series Champion.

Introduction:

They say football is a game of inches. Unfortunately, games of baseball can be lost for less. Take, for example, the 2011 World Series. Game Six: Win or go home for the Cardinals. Bottom of the 9th. Two outs. Texas sits up two scores, just a strike away from a World Series Championship. The Rangers have allowed two men on base, but with a fresh at-bat, a win is well within reach. David Freese steps up to the plate and launches a ball to Rangers right fielder Nelson Cruz. Well, *almost* to Cruz.

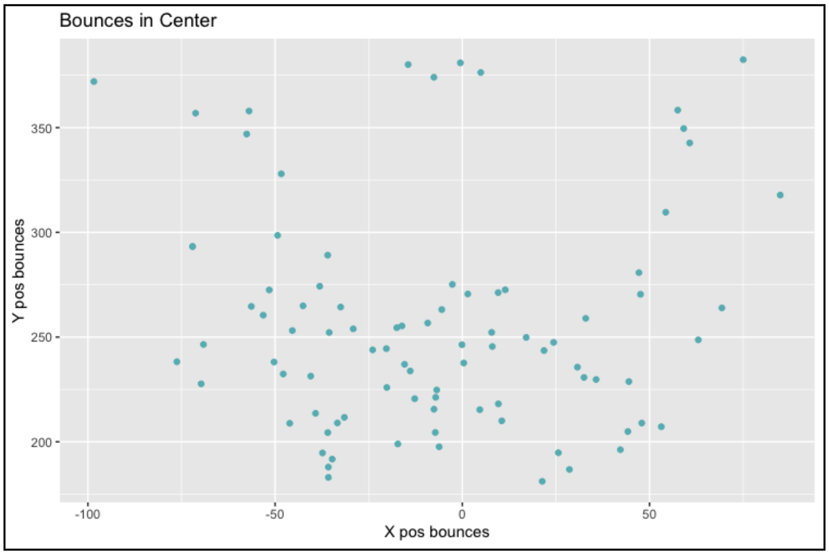
As Cruz jumps backward, the ball narrowly falls behind his glove, bouncing off the wall and back into the outfield. With a catch, he would've ended the series. Instead, the Cardinals tied the game on this play and won in extra innings. Instead of emerging victorious, Cruz is still searching for his first World Series ring.

With so many of baseball's greatest (or most heartbreaking) moments coming from small-margin plays, there must be a way to cover these margins. This is just one example of the type of event we look to prevent. By optimizing the positioning of the outfielders, we can predict generated outs from better positioning, improving defensive performance and increasing our odds of taking baseball's most coveted title: World Series Champion.

Issue:

To address how to get more outs, we first must understand where we *aren't* getting outs. At the moment, most base hits come from finding gaps in the defensive structure. Whether these are narrow infield gaps targeted by ground balls or well-placed fly balls in the outfield, the seams between defensive players are gold mines for hitters. These seams themselves are defensive interactions, hidden until well placed hits bring them to the surface. With these seams being so important, it begs the question: Are players currently positioned optimally to generate outs? Are the seams they form unknowingly allowing offensive production? This is the driving curiosity behind this paper.

Most players adopt starting positions as they begin their forays into baseball, and only tweak their placement with coaching/managing suggestions. They aren't deciding where to stand based on data-backed calculations. For example, according to this scatterplot of ball landing positions in center field, there is a large number of hits that would be catchable if the center fielder had only started a few feet closer to the diamond.



Centerfield bounces, Fig. 1

Each outfielder has an optimal position based on maximizing outfield balls in play (referred to as OBIP) coverage and minimizing the time they take to get to each OBIP. As outfielders change positioning, their ranges, or the area they can cover in all directions, may then overlap with other outfielders. With proper adjustment, the interaction and overlap of these defenders can be optimized to both place outfielders in high-volume catch areas and place their seams in lower-volume areas. By improving both, teams can generate more outs and win more games.

Data Usage:

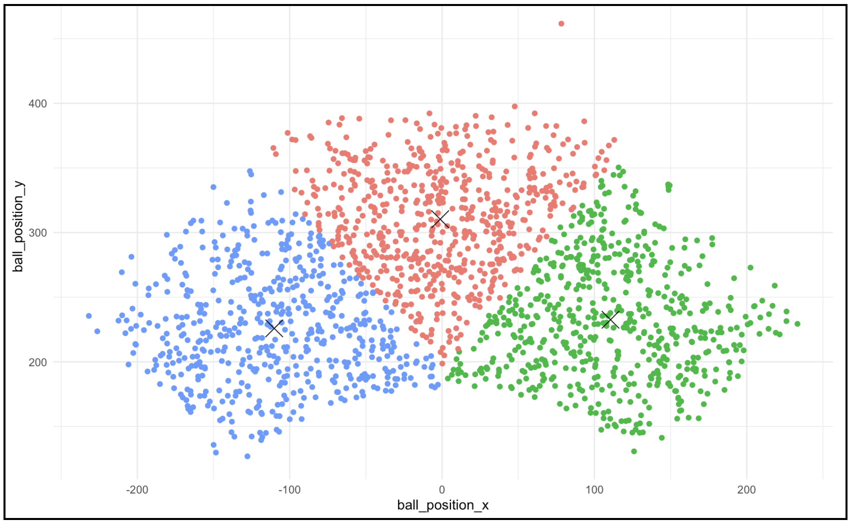
Along with utilizing the data we were provided, we included one dataset from Statcast to make key assumptions for our analysis. Statcast provides data on players' 90-second running splits, and we used this dataset to calculate the average ft/sec it takes for a player to run 0-60 ft. Although this varies by a wide range depending on age, player, and position, we found that on average players can run around 20.314 ft/sec. This is the speed used for our calculations on determining whether each OBIP can actually be caught.

Another key assumption we made was that these players have omnidirectional speed. For ease of calculation, we are assuming that the outfielders can run in any direction at the same speed. While this may seem conspicuous, we aired on the side of caution and kept our speed estimate rather conservative considering that outfielders are some of the fastest players in baseball.

Our Solution:

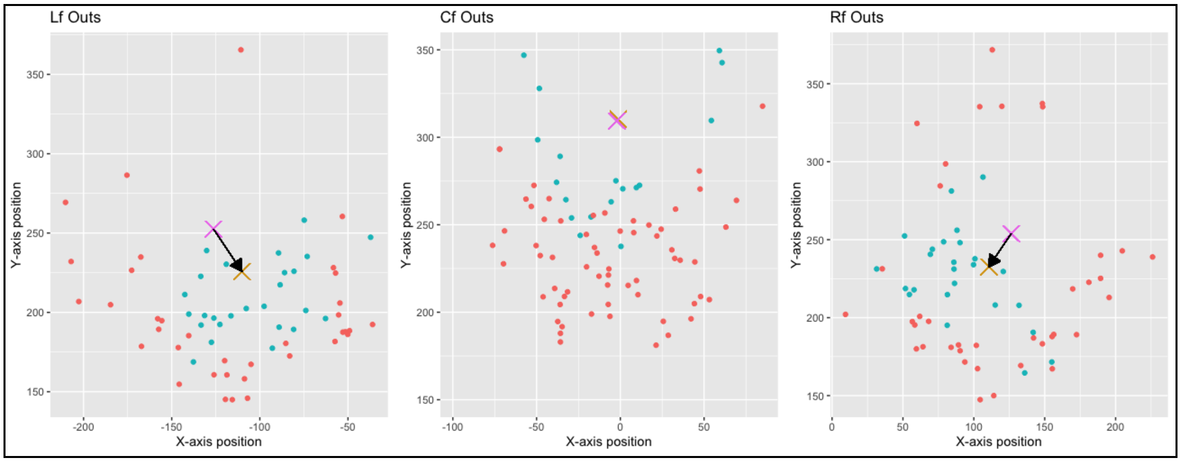
The first step to our solution was using the data to find all the balls that were hit into the outfield and acquired by an outfielder. This was accomplished by determining the plays in which balls in play were acquired by an outfielder as well as the ball position location being outside the radius of the infield. Then we utilized the K-means strategy to cluster all of these hits into three clusters that each outfielder should be able to obtain.

To prevent runs, outfielders should be positioned where they can cover the maximum number of batted balls. We know the locations of every batted ball caught in the outfield or bounced in the outfield should identify locations an outfielder would want to cover. To maximize OBIP coverage, each outfielder should be placed in a position that minimizes the total distance between the ideal starting position and each batted ball. This is precisely what the K-means clustering algorithm accomplishes, clustering every OBIP into three distinct groups (left, center, and right field) that minimize the total distance from the center of each cluster to every OBIP within that cluster.



K-means clusters for left, center, and right field, Fig. 2

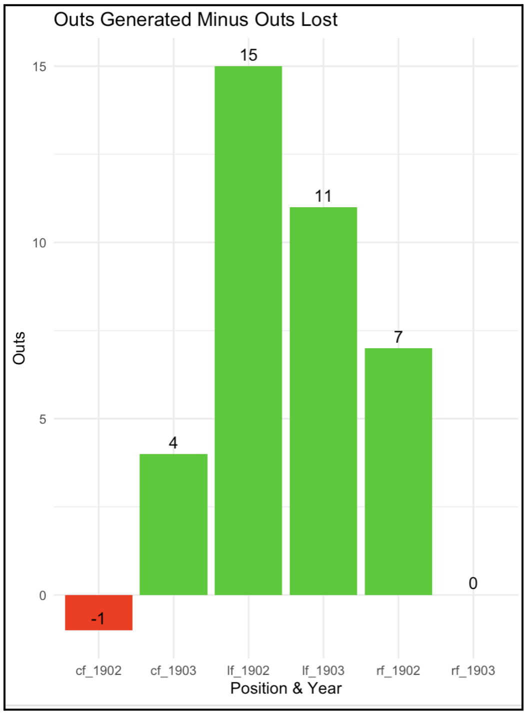
Entering the position of every OBIP from the 1903 games into the K-means algorithm, we obtain three distinct clusters (Fig. 2: RF, CF, LF) that can help inform managers where to put outfielders to maximize the number of fieldable balls. Each point in the plot in Figure 2 represents a different landing spot, including hits, catches, and bounces. The result of this clustering gives us a similar starting position to the average for center fielders, but tells us that right and left fielders should start closer to center field and home plate to maximize fieldable OBIP.

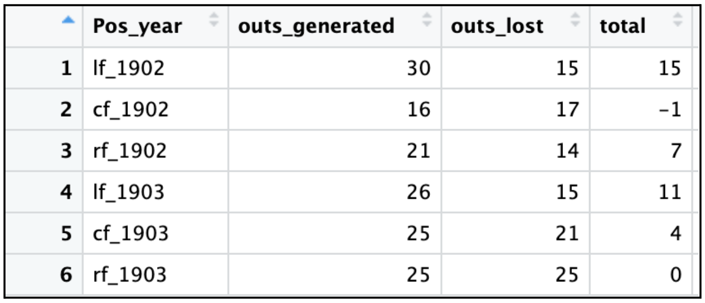


Blue: Newly-Generated Outs, Red: Hits, Gold X: K-means starting position, Purple X: Average starting position

Arrow: Suggested movement for each outfield position, Fig 3

After adjusting each outfielder to their optimal position, we can calculate our total outs generated to be 143 across the 1902 and 1903 seasons (See Appendix B for calculation). While this number seems quite high, the adjustments also move the outfielders out of range of some potential outs, ‘sacrificing’ them in pursuit of higher volume. With the adjustments, we still generate positive outs, sacrificing only 107 outs, creating a net gain of 36 outs across these 2 seasons due to our defensive adjustments.





	Pos_year	outs_generated	outs_lost	total
1	lf_1902	30	15	15
2	cf_1902	16	17	-1
3	rf_1902	21	14	7
4	lf_1903	26	15	11
5	cf_1903	25	21	4
6	rf_1903	25	25	0

Outs generated vs outs lost, chart and graph, Figs. 4 & 5

Limitations:

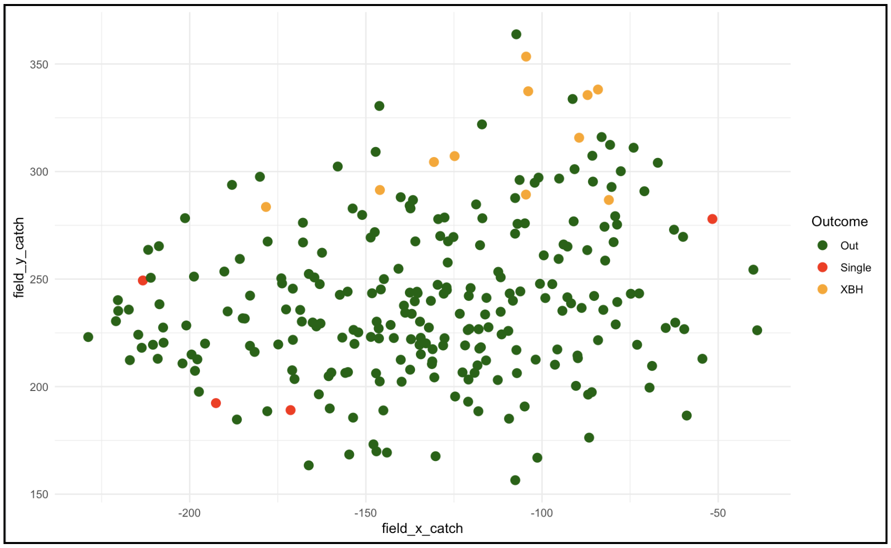
There are key limitations to this method. The most glaring and obvious limitation is the trade-off between outs generated throughout the season and the newly added extra base hits (XBH) that would occur based on our new position. This method did not accomplish placing some sort of hit value on each ball. Outfielders often play deeper so they can limit the number of balls that get behind them, thus eliminating some threat of extra-base hits. Our analysis shows that defenders should move forward if they want to generate more outs throughout the season, but this does mean sacrificing some extra bases. For instance, in the case of Nelson Cruz, allowing an extra-base hit was the difference between winning and losing the World Series. Our analysis would not have accounted for the difference in allowing a single versus a double.

The second key limitation to our analysis is that we fail to account for the exact game state in all scenarios. We recommend this outfield strategy during simple scenarios such as bases empty or one runner on. However, baseball is an ever changing game and this strategy may not be the optimal choice in different situations. For example, if the bases are loaded while your team is protecting a 3 run lead, outfielders may want to play deeper to prevent any game changing plays from occurring. For situational baseball, we would rather let the ball drop in front of us for a single, rather than risking the possibility of a game-tying double.

Finally, while we've discussed the strengths of K-means, we haven't been able to explore more complex methods of spatial optimization. The allocation of 3 outfielders in open space to maximize balls covered is a question that could be properly modeled via linear programming, utilizing an objective function and constraints to create a set of feasible solutions. However, these models tend to be NP hard by nature, requiring far more computation power and time than we had available.

Improvements:

In future work, we could explore more regarding the expected run value, use more detailed speed metrics, explore other machine learning techniques, and clean up our parameter definitions of the outfield.



Changes to former outs due to new positioning, Fig. 6

Using the same average player speed and the new ideal starting position (Left Field, 1903), we calculated that we would generate 26 extra outs, but give up 15 extra hits. While this is a positive change, not all hits are created equal. Using a rough estimate of 280 feet as a benchmark for XBH, we estimate that 11 of these 15 new hits would likely go over the left fielder's head and result in extra bases (Figure 6). In the future, we want to try and quantify the idea of outs generated versus run value. Is catching 26 extra balls in left field throughout the season worth giving up an extra 11 doubles?

Another improvement to be made would be individual player tracking to determine more exact speed metrics over different distances. This would allow a more effective application of acceleration and direction into our range calculations, improving the accuracy of generated outs.

Finally, we used a radius of balls landing 180 feet away from home plate as outfield balls in play, which is an important parameter that we used our best judgment to estimate. We could look at calculating a better estimate of outfielder speed in different directions to improve our outfield optimization.

Conclusion:

Despite the limitations of our data, we were able to develop a novel method for identifying optimal positional adjustments in the outfield. Due to the simplicity of the K-means clustering algorithm, we can claim with some degree of confidence that these optimized positionings would have prevented more hits. Over the course of a 30-game minor league season, we predict we could have prevented 15 more outs. Across 2 seasons, the model predicts we would have prevented 36 outs. As a result of the right and left fielders standing further in and towards the center, the seams between them and the centerfielder narrow, thereby preventing more OBIP from dropping. These incremental changes, nearly invisible to the naked eye, are what make baseball beautiful.

Nelson Cruz has spent the last 18 years competing for a World Series Championship. In this time, the league has changed completely. From pre-Moneyball management styles, to the rise of sabermetrics, to the eventual overwhelming use of data in the sport, Cruz has been there for every step. Now, with these changes to outfield placement, he may finally close the gap he found himself in back in that 2011 World Series Game.

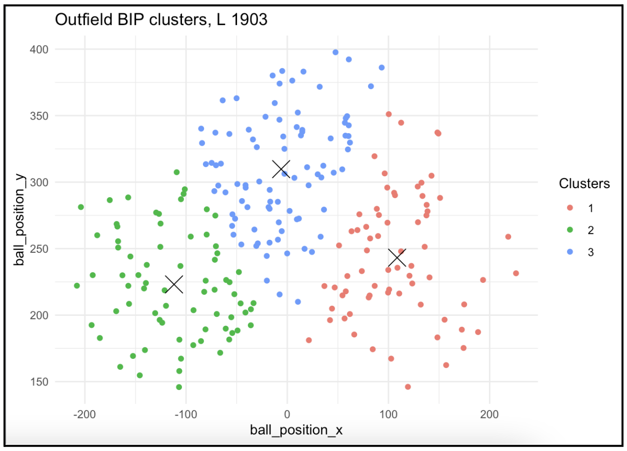
References:

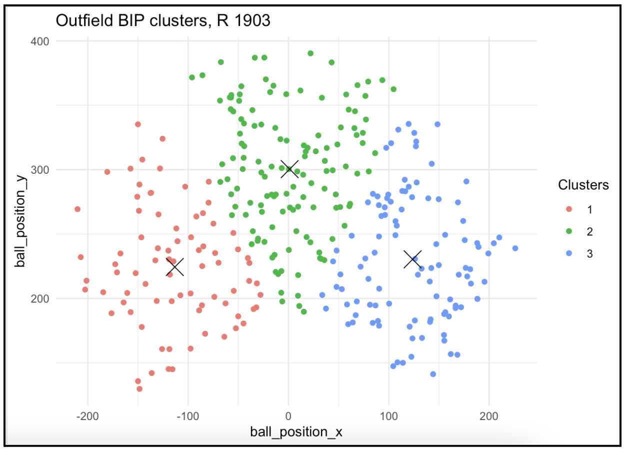
Kepner, T. (2019, October 4). *October dreams, and nightmares, have kept Nelson Cruz coming back*. The New York Times. https://www.nytimes.com/2019/10/04/sports/baseball/nelson-cruz-twins.html

Statcast Sprint Speed Leaderboard. baseballsavant.com. (n.d.). https://baseballsavant.mlb.com/sprint_speed_leaderboard

Statcast running splits leaderboard. baseballsavant.com. (n.d.-a). https://baseballsavant.mlb.com/leaderboard/running_splits

Appendix A

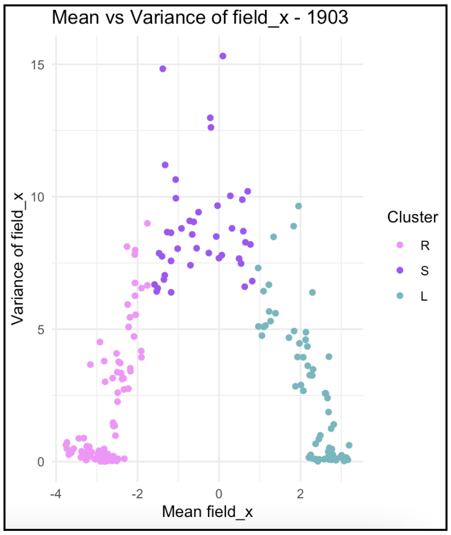




Outfield K-means clusters against LH batters, Fig. 7

Outfield K-means clusters against RH batters, Fig. 8

A crucial factor which affects positioning of outfielders is the handedness of the batter. The spray chart of batted balls across righties and lefties tends to be very different, meaning our analysis of optimal outfielder positioning is affected by batter handedness. However, batter handedness was not provided to us explicitly in the data. To work around this limitation, we implemented the same K-means clustering algorithm on the initial positions of each batter as they stand in the batter's box.



Batter handedness groupings, Fig. 9

We plotted each unique batter's mean x-position against the variance in their x-position at the beginning of each at-bat. Then we applied the K-means algorithm on these starting positions in hope we could find a reasonable classification for righties, switch hitters, and lefties. Figure 9 shows that this method seems to produce reasonable results.

Righties have a negative starting x-position with low variance. Lefties have a positive x-position with low variance, and switch hitters are in the middle with high variance. Intuitively, this makes sense because as you tally up where switch hitters stand in the box throughout the season, it'd be reasonable to expect their average starting position to be somewhere in the middle of the batters boxes with high variance. However, this analysis is at best an educated guess. Furthermore, it likely overestimates the amount of switch hitters in the dataset because the K-means algorithm equalizes the spatial variance of the clusters, not taking into account the real-world relative rarity of each grouping like switch hitters which are far rarer than lefties or righties. While an important factor in outfield arrangement, because of these issues we chose not to include batter handedness into our final analysis.

Appendix B:

The method we utilized to find the newly found outs was a very simple calculation. First, using Statcast's player speed data, we were able to find that on average players are able to go from 0 → 60 feet at around 20.314 ft/sec. Next, after finding our ideal new starting location for our outfielders, we used the straight line distance formula to find the distance between the new starting position and the location the ball bounced.

$$Distance = \sqrt{\,((x_2 - x_1)^2 + (y_2 - y_1)^2)}$$

Next, using the timestamp data provided, we were able to find every single outfield hit's hangtime by subtracting the timestamp from when the ball was hit (event_code 4) from when the ball bounced in the outfield (event_code 16) and converting to seconds. Finally, we multiplied hangtime for every ball by the average player speed to calculate a number that defines how far the average outfielder can go in that time span—we will call this expected distance. If this expected distance was greater than or equal to the straight line distance between the ball and new starting position we classified this ball as an out. Then we simply repeated this same process to be able to calculate what previously caught balls would become hits with the outfielder catch dataset.