Proceedings of the 2020 Annual General Donald R. Keith Memorial Capstone Conference
West Point, New York, USA
April 30, 2020
A Regional Conference of the Society for Industrial and Systems Engineering

Quantifying the Outfield Shift Using K-Means Clustering

Jeffrey Gerlica, Izaiah LaDuke, Garrett O’Shea, Pierce Pluemer, and John Dulin

Department of Operations Research
United States Air Force Academy, Colorado Springs, CO

Corresponding Author: GerlicaJ@gmail.com

Author Note: Jeffrey Gerlica, Izaiah LaDuke, Garrett O’Shea, and Pierce Pluemer are senior cadets at the United States Air Force Academy, collectively working on this project as part of a year-long operations research capstone course. The student authors would like to extend thanks to the advisors involved with this project as well as the client organization, the U.S. Air Force Academy Baseball Team.

Abstract: Sports teams constantly search for a competitive advantage (e.g. bidding for free agents or scouting nontraditional markets). As popularized by *Moneyball*, we focus on advanced analytics in baseball. These sabermetrics are employed to provide objective information to management and coaches to support player management and in-game strategy decisions. Though widely used at the professional level, analytics use in college baseball is limited. Air Force Academy Baseball has been one win short of qualifying for the Mountain West tournament three straight years, resulting in the loss of potential income from media payouts and exposure for future recruiting efforts. Using a K-means clustering method for defensive shifting, we calculate an overall catch probability increase of 7.4% with a shifted outfield in a one-game case study. Based on our analysis, we provide evidence that Air Force Baseball can benefit from an outfield defensive shifting scheme that drives a competitive advantage and additional wins.

Keywords: Baseball, Sabermetrics, Defensive Shifting, K-Means Clustering

1. Introduction

In collegiate sports, qualifying for a postseason conference tournament is a substantial accomplishment and eligible teams enjoy the many benefits that accompany it. Tournament qualification provides the opportunity to experience a potential championship run, the accolades to increase recruiting talent, and a substantial financial gain due to increased media exposure and attendance to participating schools. The United States Air Force Academy Baseball Team (AFB) has finished one win short of making the Mountain West Conference tournament three years in a row (2017-2019). The goal of this project is to account for that one win and help the team qualify for postseason play.

There are many ways baseball teams can look to gain a competitive advantage. Teams can focus on developing more strategic personnel decisions, such as recruiting players with broader skillsets, recruiting from non-traditional markets, and increasing the total number of recruits. Teams can also incorporate different training techniques to emphasize better strength, conditioning, mental preparation, or subtle positional skills such as pitch framing for catchers. Additionally, through advanced data analysis techniques, teams can develop in-game strategies such as pitch prediction, batter lineup optimization, and defensive shifting.

These in-game strategies are becoming more prevalent each year in Major League Baseball (MLB) due to advances in data collection and analytic techniques. Specifically, defensive shifts are increasing every season in MLB and “[d]ata indicates the shifts are working throughout baseball. The strategy saved 190 runs in the first half [of the 2015] season” for the Dodgers (Helfand, 2015).

Defensive shifting in baseball is positioning fielders in different locations than traditionally aligned. Figure 1 and Figure 2 demonstrate a standard defensive alignment versus a shifted defensive alignment, respectively, for the left fielder (LF), center fielder (CF), and right fielder (RF). These three fielders make up the traditional outfield (OF) positions.

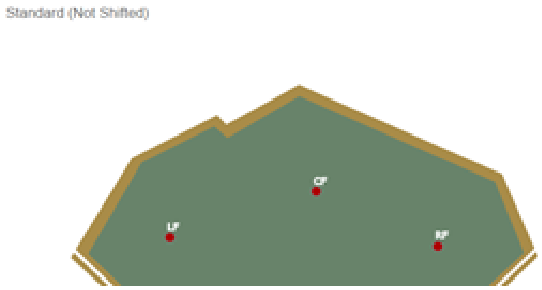

Figure 1. Standard Defensive Alignment

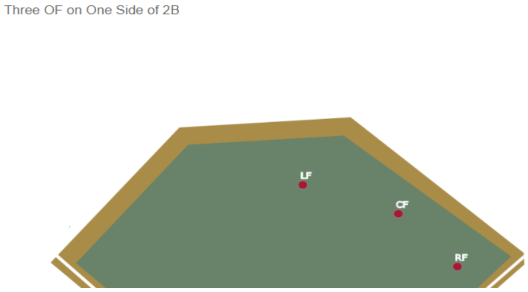

Figure 2. Shifted Defensive Alignment

1.1 Problem Statement

Air Force Baseball’s current approach to defensive shifting is qualitatively looking at trends and scouting reports for the team’s upcoming opponents in their previous games the same season to align defenders. Shifts are presently based on graphical displays, which do not provide precise locations and are open to interpretation. This leads to somewhat ambiguous decisions regarding positioning of outfielders and limits the ability to evaluate if the defensive shifts are better than having outfielders set in a traditional alignment. We quantify an approach to defensive shifting that eliminates human bias and more accurately determines where outfielders should be positioned.

Through this project, we provide proof of concept that AFB can incorporate a defensive shifting strategy into their game planning which will save runs from being scored against them by catching balls that previously fell for hits or preventing batters and runners from advancing extra bases on hits, leading to that additional win needed to qualify for the Mountain West Conference tournament. The focus of our analysis and this report is the use of defensive outfield shifting and the competitive advantage its use can gain for AFB.

Moreover, we evaluate the catch probability (the probability an outfielder will catch a ball for an out) for different alignments to determine the effectiveness of outfielder placement. With catch probability measured, we are able to justify that an overall increase in catch probability will lead to more outs, leading to fewer runs scored against and, as a result, more wins.

In this paper we ask, "How can advanced data analytics with outfield defensive shifting increase the effectiveness of the United States Air Force Academy Baseball team game planning in an effort to increase wins?" Using the data described in section 2, we measure catch probabilities, based on different locations of outfielders, and evaluate if the defensive shift is worthwhile. To determine if the shift is worthwhile, we calculate an average catch probability for a standard outfield alignment and compare it to a shifted outfield alignment. If the shift is deemed worthwhile through a higher average catch probability, we recommend where outfielders should be positioned with respect to the context of the game.

1.2 Related Work

Baseball is a data rich environment with many enthusiasts and professionals who have explored some of the nuances of defensive shifts. Unfortunately, as described by Lewis and Bailey (2015), many of the algorithms and models are proprietary information and are not available to examine. This is understandable as teams are not going to give away a competitive advantage they may have gained through advanced modeling or analysis. However, Lewis and Bailey (2015) do demonstrate that shifts are worthwhile based on runs saved as indicated by the Tampa Bay Rays incorporating an aggressive defensive shifting scheme. Because of the feasibility that runs can be saved, we show that AFB should pursue defensive shifting exploration and model development.

1.2.1 Defensive Shifting

Despite limitations in available models, there are still publications which describe methods to approach defensive shifting. For instance, Becker (2009) uses an integer programming model to minimize the expected costs associated with each hit based on the hit chart of Derek Jeter. He does not assign shifted locations to players on the field, but instead calculates the effect of a defensive shift on Derek Jeter’s batting average. The results show a substantial decrease in Jeter’s batting average when a defensive shift is employed. We use similar hit charts but assign positions to the outfielders.

1.2.2 Machine Learning

We use a machine learning method (K-means clustering) described in section 2 to position fielders, similar to James et al. (2017). The K-means clustering method does not require binning the field to assign shifted locations on the field like in Baumer (2014), other than separating the infield from the outfield. This allows for more flexibility in assigning shifted locations on the field which is a more efficient way to position players than a binning method. The centroid of each assigned cluster is where we determine the placement of each outfielder. See section 2.2 for more detailed information on the K-means methodology.

1.2.3 Catch Probability

We also incorporate different probabilities for catching a batted ball based on the distance a player will have to move to catch that ball and the amount of time it is in the air (hang time), similar to Hawke (2015). He uses a Probit model to determine the probability that a player will successfully catch a ball based on a number of variables. We simplify his approach, using the distance the fielder must travel to where the ball would land and the ball's hang time to determine the probability of catching the ball.

The probability of fielding a ball also factors into the performance evaluation of the shift. Pankin (1978), although dated, evaluates the offensive performance of a team with respect to the number of runs per game. We do not use a number of runs per game metric; however, an increase in catch probability leads to runs saved which is used to evaluate the effect of the shift against each player, each team, and over a season to assess if the shift is better than having players in their standard field locations.

2. Methodology

2.1 Data

In developing our model, we obtained a sample data extract that contains the following information in a discrete format with a separate row for every unique event (each pitch or pickoff attempt) in a 2019 game between the Air Force Academy and the University of Nebraska: an (x, y) coordinate for where every batted ball landed, who was batting and his handedness, who was pitching and his handedness, the type of pitch thrown (e.g. fastball, slider, etc.), an (x, y) coordinate of where the pitch was located over the plate, the stadium in which the contest took place, and many more descriptive data points. One of the key variables needed to determine catch probability is the batted ball's hang time. While this was not available in the data set, we were able to determine the values for each batted ball via a video clip associated with each particular ball in play. Because the video plays in real time we were able to hand-record hang time. We hope to work with our data provider to automate this process in the future for better accuracy and efficacy in data processing. These data provide key variables needed to identify trends on where players' batted balls land (Valero, 2016), and also coincide with variables identified by Beneventano, Berger, & Weinberg (2012) as the most important in predicting run production and run prevention in baseball.

Using data frame manipulation in R, we organized the fields required for our analysis and summarized them into a single table. Despite the limitation of having an extract of only a single game, there is such extensive data that we were able to provide proof of concept for the analysis requested by the AFB coaching staff. This proof of concept should provide justification for the procurement of additional data in order to perform more extensive analysis throughout the season. Additionally, we have developed a method to combine data across games in order to use additional observations from multiple games played by an opponent to refine our results in advance of a series against that opponent.

By using the data sample, we identify mathematical trends to determine outfielders' positioning that minimizes the distance to batted balls. This optimization of outfielders' locations can increase the catch probability for AFB, leading to fewer runs scored against and thus generating more wins. With this data, we also develop a user-friendly model which produces individualized shifts against all opponents based on game context and subject to certain constraints the coaching staff desires.

2.2 K-Means Method

The means to identify trends and conduct the analysis of outfield shifting was done through K-means clustering. K-means clustering is a method of grouping data into subsets in order to satisfy two properties: each observation belongs to at least one of the K clusters and each cluster is non-overlapping (James et al., 2017). This is done by minimizing the sum of the total within-cluster variation across all K clusters, as depicted in Equation 1.

$$\min_{C_1,\dots,C_k}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2\right\} \tag{1}$$

Where K is the number of clusters; C_k is the number of observations in the k^{th} cluster; i is a specific observation in the k^{th} cluster; j is a feature; and p is the number of features.

With the theoretical coordinates on the playing field of where each batted ball landed in the game, we plot each of those locations (Figure 3) to inform our K-means clustering model. These points are then used to determine the centroids of each cluster (Figure 4), which are in turn used to determine the optimal location for outfielders based on the pitcher, batter, or opposing team. Figure 4 shows how K-means clustering takes a theoretical batted ball plot and places the balls into clusters. The method then determines each cluster's centroid, which is where the outfielder should be positioned to minimize his distance to all of the balls he would be expected to field. We use K equal to three to account for the three outfielders. Unlike Becker (2009), our model goes a step further and assigns shifted locations on the field for each outfielder and uses the hit chart of each individual opponent of AFB to maximize the number of balls caught, thus increasing catch probability.

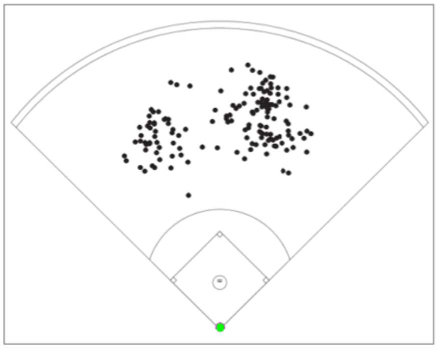

Figure 3. Ball Landing Locations

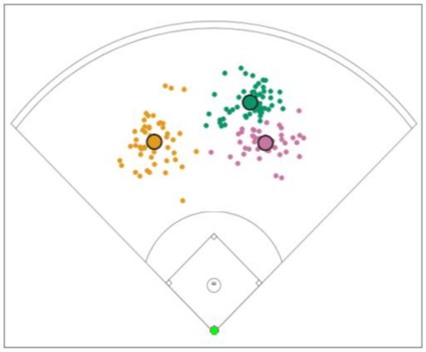

Figure 4. Ball Landing Locations with Clusters

2.3 Probability of Catching a Ball

We use a probability method to determine if the balls will be fielded (caught for an out) around each of the centroids. We determine the probability that the fielder will catch a particular ball based on the distance from the centroid and the hang time of the ball (time the ball is contacted at the plate to when it hits the ground). The further away from the centroid and less time a ball spends in the air, the less likely a fielder will be able to catch the ball. We determine the catch probability for each ball in play based on Statcast's graphical representation of catch probability, with respect to hang time and distance to the ball (Figure 5). While these numbers represent professional baseball fielders, we assume that college players have the same catch probability for simplification purposes. For example, a ball that would land 60 feet from an outfielder and has a hang time of 5 seconds is caught 90% of the time. However, if that ball would still land 60 feet from the outfielder but has a hang time of only 3.5 seconds, the catch probability drops to 20%.

2.4 Modeling

The model used in this analysis was built in RStudio, using the K-means function as well as the ggplot package. The data was pulled as a csv into RStudio to build the model. The data was condensed within RStudio to only include the coordinates of each outfield hit, then fed into the K-means function that uses Equation 1 to cluster the data. For this model, three clusters were used to represent the three traditional zones covered by outfielders. This data was then plotted along with the three centroids calculated by the K-means function. The plot was overlaid on an image of the baseball field for a better visualization that can be used as a final product for the baseball team.

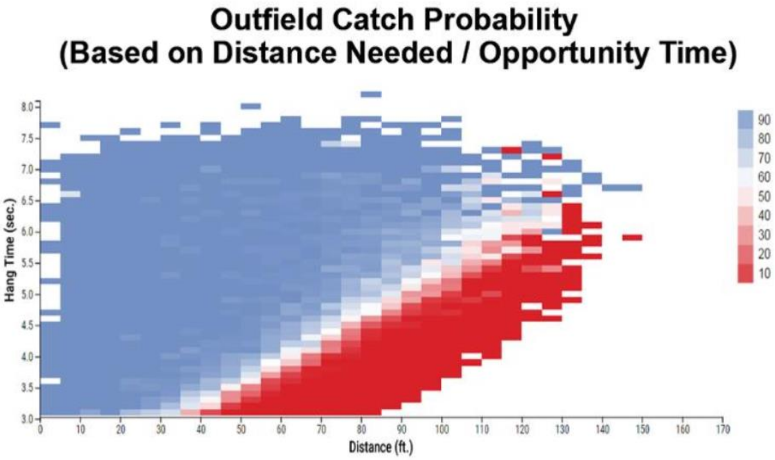

Figure 5. Statcast Catch Probability

3. Results and Analysis

3.1 Results and Analysis

Our model yields a visual tool to aid the decision makers (coaching staff) on the Air Force Baseball team. Using K-means clustering, the (x, y) coordinates on the field that represent the balls-in-play (BIP) have been classified into three clusters with a centroid in the middle of each. With the available data sample, the centroids have been placed in three unique positions, as seen in Figure 6 below. Note that we only included fly balls and pop-ups in our analysis (no ground balls or line drives), because these balls have a chance of being converted from a hit to an out. Line drives and ground balls that reach the outfield are assumed to produce a hit.

The three black dots in Figure 6 represent the minimal, centralized distance from the points of the same color. The practical application of the model is to place the outfielders where the black dots are. Each color represents the BIP for which a fielder is responsible. Through this positioning, the outfielders now have the best chance to convert balls in the outfield to outs. We recognize that with only a single game’s data we are unable to apply the model to specific players (thus generating a unique shift based on the batter). Further, because the data points are limited we understand that the proposed shifted locations vary fairly significantly from the traditional outfield alignment (Figure 7). Strictly based on the mathematics, these would be the optimal points to position the outfielders; practically speaking this demonstrates the value of the approach and suggests that additional data could yield valuable insight for the coaching staff.

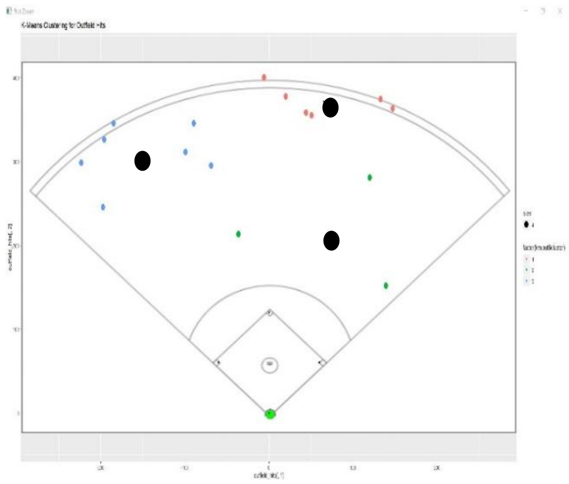
Plot Zoom
K-Means Clustering for Outfield Hits
outfield_hits[, 2]
outfield_hits[, 1]
size
4
factor(km.outfield.cluster)
1
2
3

Figure 5. K-Means Clustered Outfield

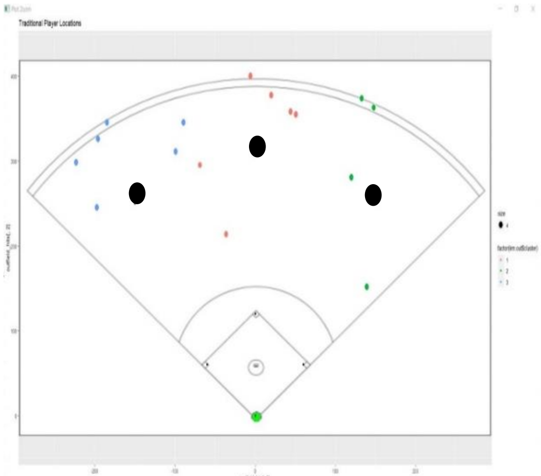

Figure 6. Traditional Outfield Alignment

Because each centroid has an (x, y) coordinate associated with it, these results could provide coaches with precise locations at which to align the outfielders using reference points. When using a standard outfield configuration, the estimated average catch probability for this game was 47.0%. When using the K-means method to shift the outfielders, the catch probability of balls to the outfield is increased by 7.4% to an average catch probability of 54.4%. This increase demonstrates the feasibility of the model and shows the value of placing outfielders at the centroids calculated using the K-means method.

4. Conclusions and Future Research

The precision of the K-means methodology allows for optimal alignment of the outfielders by maximizing catch probability, thereby reducing runs allowed and potentially generating more wins. Applying our recommended defensive alignment to the Air Force-Nebraska game in 2019 in a hypothetical case study, we identify two primary examples where the shifted outfield could have saved runs. The first example is in the third inning where an opponent hit a ball to deep center field, over the center fielder's head. The hit resulted in a double, scored a run, and this player eventually scored later in the inning. Under a traditional alignment, the ball had a catch probability of 30% but had the outfield been shifted, that probability increased to 70% and, assuming it was caught, would have prevented two runs from scoring. A second example occurs in the eighth inning with two outs. An opponent again hit a ball into deep right-center field with an estimated catch probability of 0.1%. This hit was also recorded as a double with one run scoring and the inning continuing. If the outfield was shifted, the catch probability would have increased to 60%, likely ending the inning and saving the run from being scored.

These two examples could have saved Air Force Baseball three runs from being scored against them in the game. While the team lost this game by more than three runs, there were several games throughout the season where saving as few as three runs could have been the difference between winning and losing. We are confident that defensive shifting could make a difference in AFB qualifying for the Mountain West Tournament.

Our findings provide the motivation to implement an outfield defensive shifting scheme and investigate more baseball data analytic strategies for Air Force Baseball. We recommend further investigation and sensitivity analysis on the effects of an outfield shift beyond catch probability, such as the vulnerability of new gaps for line drives to penetrate. Additionally, we recommend the Air Force Academy create a data analytics team to support baseball (and other sports) to evaluate the incorporation of an outfield shift and to research other data analytic techniques, such as infield defensive shifting, pitch heat mapping, lineup optimization, and pitch prediction.

5. References

Baumer, B. (2014) An Overview of Current Sabermetric Thought II Defense, WAR, and Strategy. In *The Sabermetric Revolution* (pp. 57–84).

Becker, K. W. (2009). Optimizing Defensive Alignments in Baseball through Integer Programming and Simulation. Kanas State University. Retrieved from https://krex.k-state.edu/dspace/handle/2097/2345?show=full

Beneventano, P., Berger, P.D., & Weinberg, B.D. (2012). Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics. *International Journal of Business, Humanities and Technology*. Vol. 2, No. 4, 2012, 67-75.

Hawke, C. J. (2017). Quantifying the Effect of the Shift in Major League Baseball. Kansas State Senior Projects Spring 2017. 191. Retrieved from http://digitalcommons.bard.edu/senproj_s2017/191

Helfand, Z. (2015, July 19). Use of Defensive Shifts in Baseball is Spreading - because it works. Retrieved October 21, 2019, from https://www.latimes.com/sports/la-sp-baseball-defensive-shifts-20150719-story.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: with applications in R.* New York: Springer.

Lewis, M., & Bailey, R. (2015). Batted Ball Spray Charts: a system to determine infield shifting. *2015 Systems and Information Engineering Design Symposium*, 2015, 206-211.

MLB Advanced Media. (n.d.). Catch Probability. Retrieved from http://m.mlb.com/glossary/statcast/catch-probability

Pankin, M.D. (1978). Evaluating Offensive Performance in Baseball. *Operations Research.* Vol. 26, No. 4, 1978, 610-619.

Valero S. (2016). Predicting Win-Loss outcomes in MLB regular season games – A comparative study using data mining methods. *International Journal of Computer Science in Sport, 15*(2), 91-112.