Playing the Odds: Defensive Positioning Strategies to Minimize Batting Average in Major League Baseball

Abstract

Over the last two decades tracking systems and advanced analytics have changed the way Major League Baseball (MLB) is played. One example of this is the increased implementation of defensive shifts and development of creative defensive alignments by MLB teams to better defend hitters. Due to the effects of these alignments on in-game outcomes, MLB will be placing restrictions on defensive alignments in 2023. Several studies and analyses have examined the effects of defensive alignment on hitter performance. We utilize, for the first time, a data set of starting coordinates for fielders from MLB that covers 3.5 seasons to quantify the effects of defensive alignment on batting average. Combined with batted ball data, we use the fielder coordinate data to create a position-agnostic description of defensive alignment. We then fit a gradient boosting model that estimates the probability a batted ball will be a hit given the alignment and features of the batted ball. The results of our method and model demonstrate the importance of defensive alignment in predicting whether or not a batted ball will be a hit and how defensive alignments can be improved to minimize a hitter's batting average. We use our model to explore the effectiveness of four-person outfields against three exemplary hitters and develop an optimization scheme to find optimal defensive alignment against those hitters. Our results indicate that batting average on balls in play would decrease by 14.0% for left handed hitters and 8.9% for right handed hitters, on average, if teams employed the best defensive alignment against each hitter. Despite the future restriction on defensive alignments, our method can be used to optimize a team's defensive strategy.

1. Introduction

Over the last two decades, the use of data science, statistics, and mathematics in sports has greatly increased. Teams from a wide variety of sports have been using data and analytics to optimize their on-field performance and in-game strategy to produce more wins. Bill James was a pioneer of this idea when he introduced Sabermetrics "the search for objective knowledge about baseball" {cite: Guide to SABR research}, in the early 1980s. James' way of thinking wasn't well known or popularized until the early 2000s when the book *Moneyball {cite}* was published.

Major League Baseball (MLB) was notably at the forefront of this analytics revolution due to its discrete nature and the data produced by its long regular season of 162 games. The creation and deployment of tracking systems have provided even more data about pitches, batted balls, and player positioning beyond traditional player hitting and pitching statistics. In 2006 Major League Baseball introduced PITCHf/x, a system that can track pitch characteristics {cite: FastPFxGuide article}. Shortly after PITCHf/x, HITf/x was introduced which tracked the velocity and angle of the ball off the bat {cite: FastPFXGuide article}. Since the first introduction of these tracking systems, new and improved tracking systems collect additional data that enables quantification of player attributes such as arm strength and sprint speed, among many others {cite: Statcast article}. The Statcast {cite} tracking system was introduced in 2015 and is now used by MLB to collect data from various tracking technologies. These new data sources have provided the opportunity for a different and deeper understanding of the game of baseball.

While MLB teams have traditionally used advanced metrics to identify undervalued players, the new sources of tracking data have provided the opportunity to optimize player performance and in-game strategy. One of these in-game strategies is defensive alignment. Outside of the catcher and pitcher, the other seven defensive players are free to move anywhere on the field. In recent years, teams have more frequently utilized non-standard defensive alignments to better defend against the tendencies of certain hitters. In 2016, an infield shift (defined as three infielders on one side of second base) was implemented for 13.7% of plate appearances, but that number more than doubled to 30.9% in 2021 {cite: Savant}. Teams have recently employed creative defensive alignments in an attempt to better position their defense. For example, the Tampa Bay Rays implemented a four person outfield with all three infielders positioned on the right side of second base in the 2019 American League Wild Card game against batter Matt Olson, leaving the entire left side of the infield vacant.

These defensive alignments can have a substantial impact on the results of an at-bat, and therefore the results of a game or season. While defensive alignment does not have an effect on several potential outcomes of an at-bat (strikeouts, walks, home runs, and hit by pitch), about 63% of plate appearances end in a batted ball hit into play that the defense has an opportunity to field {cite: Fangraphs}. Ben Lindbergh discussed the possibility that shifting may have cost the Atlanta Braves the National League Pennant in 2020 {cite: The Shift May Have…, 2020}. Sports Info Solutions {cite} estimated that 517 runs were saved throughout the league from infield shifts alone in 2021 {cite: Fangraphs}, indicating that defensive alignment is an important part of in-game strategy. The frequency and effectiveness of defensive shifts has prompted an agreement between MLB and the MLB Players Association to restrict defensive shifts that will take effect in 2023 with the goal of increasing the number of batted balls that are turned into hits {cite}.

Several previous research studies have examined the effects of defensive positioning on the outcomes of batted balls and optimal defensive alignments. Hawke Jr. {cite}, Model {cite}, and Gerlica {cite} all examined the effects of defensive alignment on the out/hit outcomes of batted balls as part of their senior capstone projects. Lewis and Bailey {cite} divided the infield into 9 zones and used information about the pitcher, batter, and count to find an optimal infield alignment. Montes et al. {cite} used fielder characteristics and batter's spray charts to optimize outfield alignment. Relatively little public, open-source research has been done on the effects of moving all 7 seven fielders together. Notable exceptions are Easton & Becker {cite} and

Bouzarth {cite} who both discretized the field into locations where fielders could be positioned and used hitter spray charts and integer programming to find an optimal defensive alignment. None of these studies have utilized the starting positions of the seven moveable fielders (excluding pitcher and catcher) and characteristics of batted balls to examine the effects of defensive alignment on hitter success. Additionally, several of the aforementioned studies make assumptions about fielder coverage.

We develop a novel analysis examining the effects of defensive alignment on hitter success in Major League Baseball. Our analysis uses data of the defensive player's starting locations which are not public and, to our knowledge, have never been used in an academic publication. We combine fielder position data with batted ball characteristics to develop a position-agnostic definition of defensive alignment that is used in a gradient boosting model to estimate the probability of a batted ball being a hit. We use the results of our model to identify important predictors of the outcome of batted balls,, and identify the best defensive alignment against a hitter and quantify its effects on the hitter's batting average on balls in play (BABIP). We also estimate the effectiveness of a four-person outfield, propose a method for exploring new, optimal defensive alignments, and create an interactive Shiny application that demonstrates our method. The results of our analysis provide novel insights into the effects of defensive positioning on hitter success in MLB. The remainder of the paper is outlined as follows: in Section 2 we describe the data and methodology, Section 3 outlines the results, and Section 4 provides discussion.

2. Materials and Methods

2.1. Data and data cleaning

We used data from Major League Baseball games recorded by Statcast from 2018 to 2021. Data about batted balls was obtained through Baseball Savant using the R package baseballr {cite: baseballr package}. We also obtained data of the fielders' coordinates in the field from Major League Baseball Advanced Media {cite: MLBAM} under a limited license agreement. Briefly, these data included information about fielder positioning and the batted ball information for approximately 284,000 batted balls in play after data cleaning. Additional details are provided about the data sources and data cleaning steps in the remainder of this subsection.

MLBAM shared data that contained each fielder's location at the moment of contact for every batted ball event (BBE) between March 29, 2018 and June 27, 2021. BBEs include any batted ball where the ball was hit into fair territory, foul territory if it leads to an out, or homeruns. These locations are given in X,Y coordinates, denoting the horizontal and vertical position of each fielder in feet where home plate is located at (0,0). Each fielder's location coordinates were identified by their defensive position, denoted 1 (pitcher) through 9 (right fielder). The fielder location data contained information from approximately 350,000 BBEs. Using the fielder coordinates, we calculated the angle between each fielder's coordinates and the vertical line passing through home plate, which is located at the origin (0, 0). Negative angles denote fielders positioned left of this vertical line. We also computed the distance each player was located from home plate and from first base, located at approximately (63.64, 63.64).The player coordinate data also contained information about the game identification number, season, date, home and away teams, at bat number within the game, batter and pitcher identification numbers, the result of the BBE, and the batted ball hit type (ground ball, flyball, popup, linedrive).

We obtained information about batted balls from Baseball Savant through the baseballr {cite} package. These data included characteristics of all pitches and batted balls, and the results of the batted balls, during the same time period as the fielder position data. Batted ball characteristics included launch angle, exit velocity, and the hit distance produced by Statcast. The Statcast hit distance is a computer-generated estimate of the distance that the batted ball would travel from home plate before landing on the ground. The batted ball data also included batted ball hit coordinates. For batted balls that remain in the field of play, these coordinates denote where the batted ball was first touched by a fielder. Although many of the batted ball features are derived from stadium tracking system data, the batted ball coordinates are estimated and recorded by a person {cite: Tom Tango, personal correspondence}. The data set also contained a column categorizing each batted ball put into play as either a ground ball, line drive, fly ball, and popup. Like the hit coordinates, this categorization is done by a human.

The batted ball data obtained from baseball Savant also included information about batters, situational descriptions (e.g., outs, runners on base), pitcher, the outcome of each pitch, and a categorical description of the defensive alignment. Batter information included the batter’s sprint speed and batting stand (right vs left). Batter sprint speed is estimated by each stadium’s tracking system taking the fastest two-thirds of a batter’s runs to first base, including them with a batter’s sprint speed on two base runs, and then taking the maximum amount of feet covered in a one second window {cite: BbSavant website}. Sprint speed was also obtained from Baseball Savant for all games from the beginning of 2018 to June 27, 2021 for each batter with at least 10 running opportunities. The batted ball data also included information on the outcome of the batted ball. This information included which fielder initially touched the batted ball, a short description of the play, batter stance, pitcher throwing arm, and other fields describing the pitch batter, and pitcher). The hit coordinates were transformed so that home plate was located at the origin and Y increases from home plate to the outfield wall, using the GeomMLBStadiums package {cite: GeomMLBStadiums}. Batted balls that did not have a Y coordinate greater than 0 were removed (approximately ~2,000 (2,152) batted balls met this criteria).

We joined the fielder location data with the batted ball data and created additional variables of interest for our analysis. We merged the data sets using game ID, batter ID, pitcher ID, at bat number in the game, batted ball hit type (ground ball, flyball, popup,or line drive), and game date. Our merged data set contained approximately 284,000 observations. We used the hit coordinates (where the ball was initially picked up) to compute the horizontal spray angle relative to home plate. Batted balls with a spray angle of 0 are hit directly up the middle, with -45 degrees along the third base foul line and 45 degrees along the first base foul line. We similarly computed the angle of each fielder relative to home plate. Finally, we computed the distance between each fielder’s starting location and home plate and the fielder’s location and first base.

We created a method to estimate the location where batted balls will first touch the ground. We refer to this location as the landing location. The aforementioned hit coordinates denote where the batted ball was first touched by a fielder. We created the landing location coordinates by utilizing the hit coordinates and the hit distance, a Statcast derived estimate of the distance the batted ball traveled, or would have traveled if first touched by a fielder, from home plate before hitting the ground. We define the landing location as the point located the same distance from home plate as the Statcast hit distance having the same spray angle defined by the hit coordinates. Rows without a Statcast hit distance observation were removed from the data set (approximately 17,600 observations were without a Statcast hit distance).

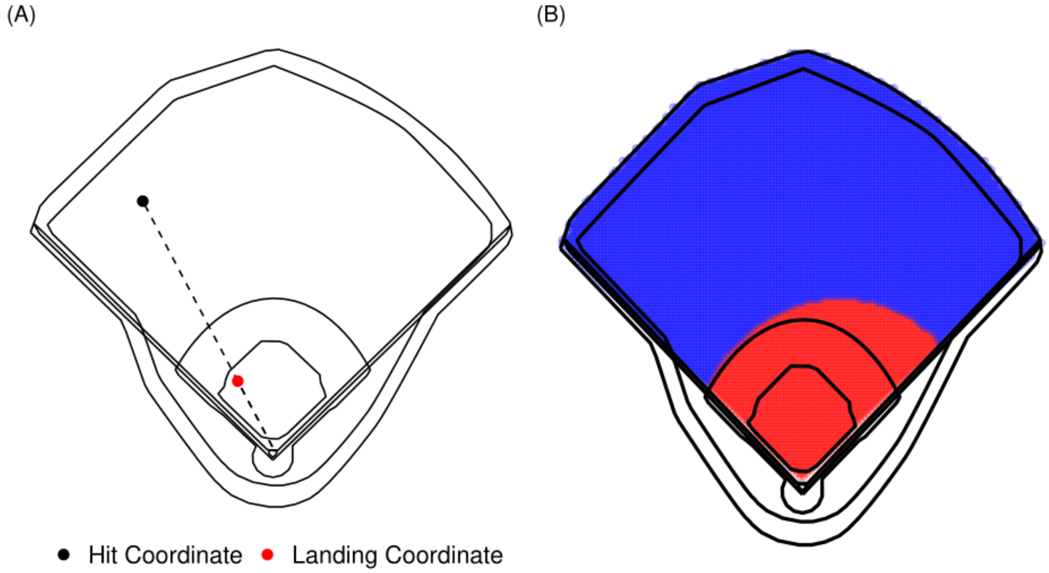


Figure 1: Two coordinates for batted balls and definition of the infield. Panel (A) shows the relationship between the hit coordinate and the landing location for a ground ball base hit. The hit coordinate is recorded by a human and shows where a fielder first touched the ball. We computed the landing location, which shows where the batted ball lands on the infield. Panel (B) shows the regions of the field we defined as infield (red) and outfield (blue). The determination of a fielder's intercept point depends on whether their starting location is in the infield or the outfield.

We used features in the aforementioned data sets to create new variables that we used to model the probability of a batted ball being a hit. We created a new variable to determine if a batted ball was a hit or not. If a batted ball had an outcome of a single, double, triple, or inside the park home run, it was considered a hit. If a batted ball had an outcome as an out, error, or fielder's choice, it was designated as not a hit. Out of the park home runs were removed from the data set because fielders have little opportunity to defend those batted balls. Occasionally outfielders catch fly balls that would otherwise be home runs, but that is a rare occurrence that we ignore for our analysis.

We performed several data cleaning steps to remove observations that were not of interest or that had unusual or erroneous features. Batted balls claimed to have a spray angle less than

-60 degrees or greater than 60 degrees were removed from the data set, as well as foul outs. We did not remove all batted balls beyond the foul lines (-45 and 45 degrees) because some base hits with a horizontal angle close to the foul lines were gathered by a fielder outside of fair territory. We wanted to include these batted balls because they landed in fair territory. We removed foul outs because defenses will position themselves to defend areas of the field where batted balls can land as hits, not where batters tend to hit foul balls. Through our exploratory data analysis, we discovered that there were accuracy issues with the hit coordinate information. This discovery was enabled via the use of MLB videos {cite: MLB film room} and confirmed by Tom Tango of MLB {cite: personal correspondence with Tom Tango(?)}. Mr. Tango confirmed that the hit coordinates given in the data set are manually tagged by a human to the best of their ability. This can lead to noisy spray angle, an important variable for this analysis.

As mentioned in more detail in the Discussion section below, our model for hit probability ignores several situational features that can affect defensive alignment. We did not use the ball-strike count, score, inning number, top or bottom of the inning, or runners on base. Many of these situational considerations often factor into a team's chosen defensive alignment. Among the batted balls in our final data set, 57% were hit with the bases empty. We anticipate that the results shown here could (a) be adjusted to account for these situational considerations and (b) be improved with future work that includes these factors.

2.2. Nearest Fielder Methodology

We use Tom Tango's intercept point methodology from the Outs Above Average statistic {cite} to identify the location that each defensive player is projected to intercept the batted ball to make an out. Our methodology considers characteristics of the batted ball (e.g., hit type classification, hit distance, spray angle) and the initial position of the seven movable defensive players to rank these players by their proximity to the line of the ball's travel. As we explain in further detail below, this methodology enables us to consider any defensive alignment in our models.

The fielders in the position data are identified by their fielding position (e.g., 5 for third baseman). Our initial statistical models used the fielder position specific coordinates as predictor variables. While these predictors provide good accuracy (0.845) for classifying hit vs no hit, they do not enable a more general approach to considering defensive alignments. For example, a team could implement a four-person outfield alignment by moving one of the 2B, SS, or 3B to the outfield. If coordinates of each defensive player are tied to a specific fielding position, then each alignment produced by moving one of these infielders to the outfield defines a unique four-person outfield alignment. These three unique alignments could produce different estimated hit probabilities for the same batted ball even if the starting position of the seven fielders are the same across the three alignments.

We developed a fielder position agnostic approach to identify the proximity of players to the batted ball's line of travel and rank them based on this proximity, which is termed their intercept point. Given a batted ball and its characteristics, we identify a theoretical intercept point for each defensive player. The intercept point is based on where we assume each fielder initially intercepts the ball, or where we assume they should have intercepted the ball if the ball got by them. A fielder's intercept point is dependent on where the fielder is initially located (infield vs outfield) and the characteristics of the batted ball (ground balls vs fly balls/popups, see below). Figure 1(B) shows the regions of the field that we designed as infield (red) vs outfield (blue). A player is considered to be in the outfield, and we refer to them as an outfielder, if the sum of the distance between a player's starting location and home plate and the distance between the starting location and first base is greater than 290 feet. If this combined distance is less than or equal to 290 combined feet, they are considered to be on the infield and we refer to them as an infielder.

We first describe the calculation of each fielder's intercept point for ground balls and line drives. For infielders the intercept point of these batted balls depends on the hit distance (landing location) and the fielder's starting distance from home plate. If the hit distance is closer to home plate than the starting point of an infielder, the infielder's intercept point is the point along the horizontal angle (spray angle) of the batted ball that is the same distance from home plate as the fielder's starting location. If the hit distance is greater than or equal to the infielder's starting distance, the intercept point is the point along the horizontal angle (spray angle) of the batted ball whose distance is given by the batted ball hit distance. For outfielders the intercept point is the landing location. By defining an outfielder's intercept point as the landing location, they are typically designated as being farther away from the ground ball or line drive than any infielder. We use this definition of outfielder intercept points because very few ground balls or line drives fielded in the outfield are converted to force outs at first base. We illustrate this concept in the following example.

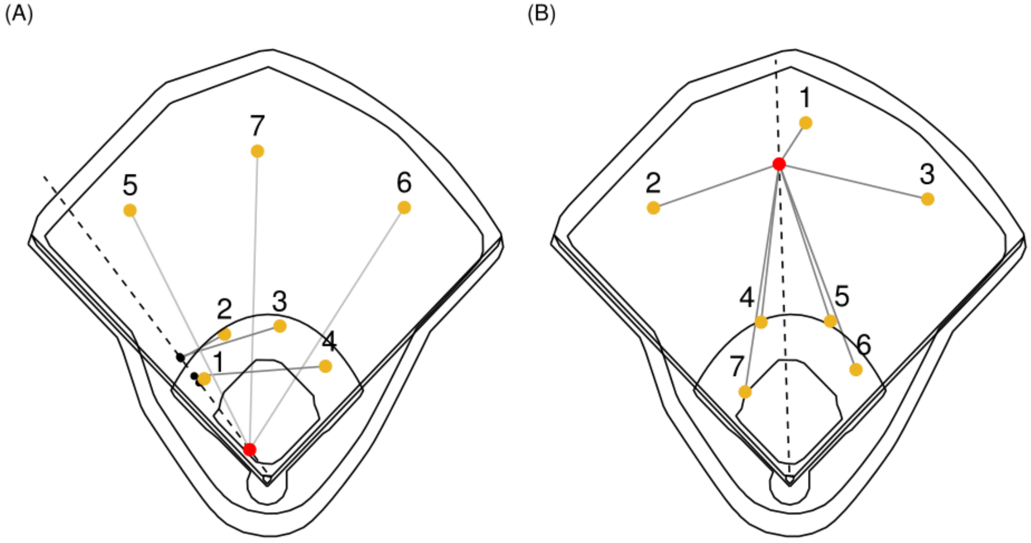


Figure 2: Intercept points for each fielder for a ground ball (A) and fly ball (B). The number next to each fielder indicates their rank for distance to their intercept point. Panel (A) shows each fielder's intercept point for a ground ball with a landing location near home plate. Each infielder's intercept point lies along the path of the ball's travel at the same distance from home as the

fielder’s starting location. Each outfielder’s intercept point is the landing location. Panel (B) shows each fielder’s intercept point for a fly ball to left-center field.

Figure 2(A) demonstrates the intercept points for each fielder for a ground ball. This batted ball’s landing location, shown in red, is not far from home plate. Each fielder is represented by a yellow point and connected to their intercept point, in black, by a line segment. The distance between home plate and the landing location is less than each infielder’s starting distance from home plate. As a result, the infielders’ intercept points are along the path of the ball, represented by the dotted line, at the same distance from home plate as their starting locations. This creates four unique intercept point locations, one for each infielder. Each outfielder is positioned at a location that makes it unlikely for them to record an out on this ground ball. In theory the only way they can make an out is by catching the ball. This makes each of their intercept points the same: the ball’s estimated landing location.

For fly balls or popups, every fielder’s intercept point is the landing location. We assume these batted balls must be caught to result in an out. Figure 2(B) shows the intercept point for a fly ball hit to left-center field. This fly ball’s landing location is deep in the outfield, so the only way for any fielder to make an out is to catch the ball. Thus, each player has the same intercept point at the landing location.

After determining each fielder’s intercept point we compute the distance between each fielder’s starting location and their intercept point. This distance is used to rank each player in their proximity to the batted ball. In Figure 2(A), the third baseman has the shortest distance between his starting location and intercept point giving him rank one. The shortstop has the second shortest distance giving him rank two. In Figure 2(B), the center fielder’s starting position is closest to the landing location, giving him rank one.

In addition to the fielder distance ranking, we compute other features describing each fielder and their positioning relative to the batted ball. These features include the angle and distance of each fielder’s starting position relative to home plate. To quantify how a fielder may need to move to field a ball, we compute the difference between each fielder’s angle and the spray angle. We also decompose the distance between a fielder’s starting location and the intercept point into horizontal and vertical components, where the vertical component is intended to reflect a player moving forward versus backward. Finally, we compute the distance between each fielder’s starting location and first base, and each fielder’s intercept point and first base.

We apply the intercept point methodology to the observed defensive alignments for all ~285,000 batted balls hit into play between March 28, 2018 to June 27, 2021. We also applied this process to theoretical defensive alignments created by changing fielders’ starting locations.

2.3. Statistical Modeling and Monte Carlo Simulation

We used gradient boosting (GB) to model the probability of a batted ball being a hit using batted ball characteristics and fielder positioning. We tried several classification methods and ultimately found the highest classification accuracy using GB. The GB model used 3 batted ball features and 70 fielder/fielder positioning features, and the batter sprint speed to estimate the probability a batted ball would result in a hit. We used 5-fold cross validation to train the model using 1,300 trees, shrinkage of 0.01, and interaction depth of 8. We examined the variable importance from the GB model and the partial dependence plots {cite: Friedman 2001, “greedy function approximation”} from the three most important variables. The partial dependence plots show the marginal effect of each feature on the probability of a batted ball being a hit {cite: Molnar, “Interpretable machine learning”}

Features of Batted Ball and Hitter	**Features of Fielder’s Starting Position Relative to Batted Ball**	**Features of Fielder Starting Position**
● Launch Angle ● Spray Angle ● Exit Velocity ● Hitter Sprint Speed	● Every fielder distance to their respective intercept point (7 features) ● Every fielder difference in horizontal angle in comparison to the ball’s travel (7 features) ● Every fielder’s intercept point distance to first base (7 features) ● Every fielder’s intercept point distance to home plate (7 features) ● Every fielder’s distance to their intercept point in the x-axis (7 features) ● Every fielder’s distance to intercept point in the y-axis (7 features)	● Every fielder distance to home plate (7 features) ● Every fielder distance to first base (7 features) ● Every fielder starting location in the x-axis (7 features) ● Every fielder’s starting location in the y-axis (7 features)

Table 1: Features used in the gradient boosting model. The GB model used features of the batted ball, fielder positioning relative to the batted ball’s landing location, and fielder starting positions to estimate the probability a batted ball will be a hit. Variable importance is discussed in Section 3.1.

We used the GB model to estimate the probability of a batted ball being a hit and model the efficacy of a defensive alignment against individual players. Our model can be used to estimate the probability of a hit for any defensive alignment, including manually specified alignments. We quantify the efficacy of a defensive alignment against an individual player by estimating the player’s batting average against that alignment. This estimate is derived using a Monte Carlo simulation where each batted ball is randomly designated a hit (1) or not a hit (0) using its estimated hit probability. The mean of these simulated 1’s and 0’s provides a player’s batting average on balls in play (BABIP). We repeat this simulation 10,000 times to account for unmodeled variability and uncertainty in our estimates. The mean batting average over these 10,000 simulations is our estimate of the player’s average against that defense.

2.4. Best Alignment, Alignment Optimization, and 4-Person Outfields

We used our GB model and a Monte Carlo optimization to search for the optimal defensive alignment against a hitter. We define optimal as the defense that minimizes a hitter's average BABIP. There are an infinite number of possible defensive alignments to employ. To make our optimization computationally tractable, we assume that hitters are well-defended with defensive alignments previously used by MLB teams against that hitter, and we constrain our optimization search. The first step in our optimization finds the best defensive alignment among the observed defensive alignments used against a given hitter. We note that switch hitters will have two best alignments, one for batting right handed and one for batting left handed.

After finding the best observed defensive alignment, we use a constrained Monte Carlo random walk with Metropolis acceptance criteria {cite: ??} to search for a better alignment that deviates slightly from the best observed alignment. We implemented several constraints on the random walk to limit proposed defensive alignments. We require that each fielder's starting position be no more than 10 feet away from their starting position of the best observed defense. We also require each fielder's starting position to be in fair territory. Finally, we require that the first baseman cannot play more than 45 feet from first base. We chose this value because it represents the 99th percentile of the distance first basemen have played from first base among all observed balls in play in the data. We ran the optimization for 5,000 iterations and reported the optimal defensive alignment. Due to the computational time required to find optimal alignments, we examined the effects of using the best alignment on BABIP for all hitters with at least 250 batted balls in play and demonstrated the same effect for optimal defensive alignment for just three hitters.

We performed a separate investigation of the effects of four-person outfields on BABIP. Four person outfields are categorized by a human and are designated in the dataset. Because four-person outfields were rarely employed by MLB teams, they were unlikely to be chosen as a best defensive alignment in our initial optimization, so we found the best observed four-person outfield against a given hitter among all four-person outfields used against any hitter in the data set. We hypothesized that some players will be better defended with a 4-person outfield than others.

2.5. Visualization with R Shiny

The method developed here enables the modeling of any defensive alignment's effect on a hitter's BABIP. We developed an R Shiny application that allows users to manually specify a defensive alignment and model a hitter's BABIP. The app can be found at this link: https://matt-boyd.shinyapps.io/defensive-positioning/ and is executed by choosing a hitter that has at least 250 batted balls in play between March 29, 2018 to June 27, 2021 in the 'Simulation' tab. Once a hitter is chosen, the user can manually change the coordinates of each fielder. When 'Simulate' is clicked, the application estimates the hit probability for every batted ball the chosen player has hit based on the user defined defensive alignment. Each batted ball is then designated as a hit or not a hit by simulating a Bernoulli random variable using the estimated hit and these realizations are used to compute a batter's BABIP. This procedure is repeated 10,000 times to create a distribution of possible BABIP's of the chosen player based on the user-defined defensive alignment. In the 'Optimization' tab, once a hitter is chosen and 'Optimize' is clicked, the best observed defensive alignment against the chosen hitter is found. Monte Carlo random walk with Metropolis acceptance criteria is then performed with the constraints listed in the previous section to find an optimal defensive alignment. Only 500 iterations are performed in the app to save time.

3. Results

3.1. Gradient Boosting

		Predicted Outcome		
		Hit	Not Hit	Percent Correct
Actual Outcome	Hit	74,047	17,527	80.9%
	Not Hit	9,157	183,618	95.3%
	Percent Correct	88.9%	91.3%	–

Table 2: The confusion matrix for the final gradient boosting model on the entire data set.

The CV classification accuracy of our GB model was 89.09%. The majority (64.2%) of the model's estimated probabilities for each batted ball were less than 0.10 or greater than 0.90, indicating that the model identified many batted balls as being very unlikely or very likely to result in a hit. The model tended to misclassify batted balls with probabilities between 0.10 and 0.90. A confusion matrix for the classification of all batted balls using the final model is given in Table 2. The model correctly classifies batted balls that are not hits about 95% of the time. Batted balls that are hits are classified correctly about 81% of the time.

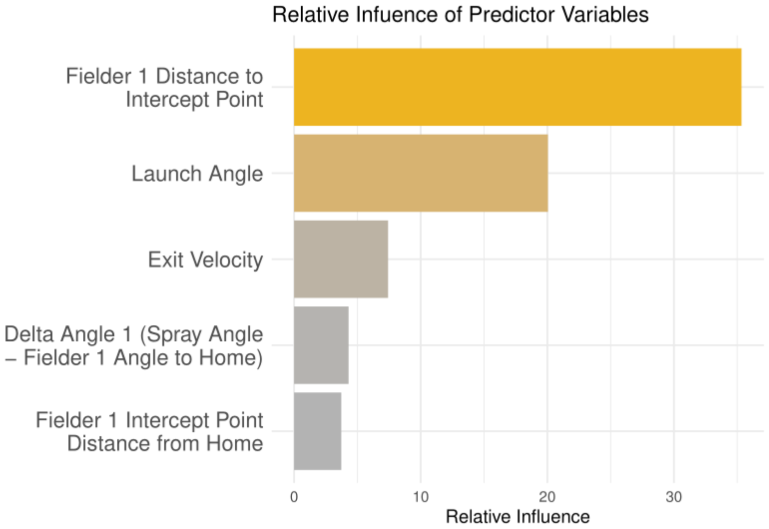


Figure 3: Variable importance from the gradient boosting machine model. Highlighting the importance of fielder positioning, the nearest fielder's distance to the intercept point is the most important feature. Launch angle is the second most important predictor, followed by exit velocity.

Figure 3 shows the five most important variables in the model by relative importance. Features describing the positioning of the nearest fielder and the characteristics of the batted ball data are shown to be the most important predictors. Fielder 1 is the fielder that the intercept point methodology identifies as the closest fielder to the batted ball's line of travel or landing location. Fielder 1's distance is the most important predictor because the shorter the distance a player has to travel to intercept the ball, the more likely an out will be made on both ground balls and batted balls that must be caught. The difference in horizontal angle relative to home plate between Fielder 1 and the batted ball spray angle was the fourth most important predictor in our model. The difference in angle indicates the direction (left or right) that the fielder has to go to reach their intercept point. It can also be an indicator of how far away the fielder is from their intercept point when considering the distance from home the fielder is positioned and their intercept point.

Exit velocity and launch angle of the batted ball are also very important in predicting if a batted ball will be a hit. The launch angle often determines which fielders can make a play. For example, a batted ball might be hit just high enough to be out of the reach of an infielder or just low enough for an infielder to intercept the ball. The exit velocity can determine if a batted ball will be tough to handle or be hit too slow to not have an opportunity at an out. If a ground ball is hit hard, fielders have less time to react and the speed of the ball will make it harder to field, or if a ground ball is hit too slow, a fielder will have less time to make a play before the runner reaches first base.

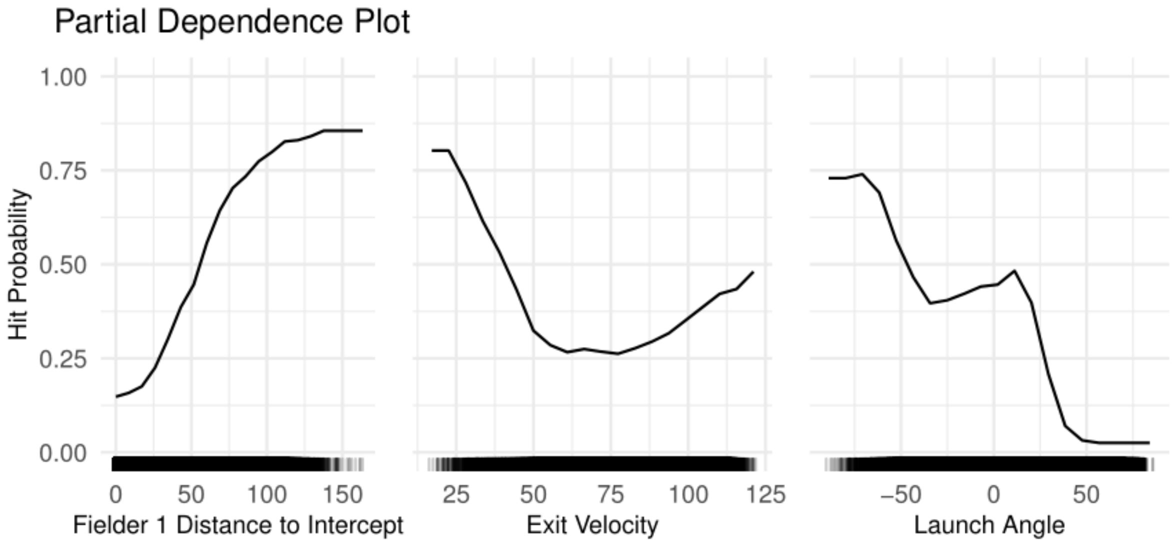


Figure 4: The partial dependence plots (PDP) for the three most important variables. The probability of a batted ball being a hit increases sharply as the distance to the nearest fielder increases.

Figure 4 shows the partial dependence plots for the three most important variables identified by the GB model. The partial dependence plot for Fielder 1's distance to their intercept point shows a sharp increase in the probability a batted ball will be a hit as the fielder's distance gets larger. This plot shows how the greater the distance the nearest fielder needs to travel to intercept the ball, the lower the chance of the fielder getting to the ball in time to make an out.

The partial dependence plot for exit velocity shows a U-shape relationship between exit velocity and the probability of a batted ball being a hit. Batted balls hit slowly off bat (< 38 mph) have the highest probability of being a hit, but this probability rapidly decreases with exit velocity, reaching a minimum in the 62-75 mph range. Beyond 75 mph, the probability steadily increases, exceeding 0.37 once exit velocity is greater than 100 mph. Balls hit slowly are difficult to field because infielders will have to rush in to make a play in time. Hard hit balls are difficult to field due to fielders' limited time to react and reach the intercept point.

The launch angle partial dependence plot shows the highest probability of a hit for batted balls that have a large downward launch angle (<-50 degrees). These batted balls that are hit with a very low launch angle are typically weak ground balls that are difficult to field in time to make an out. Hit probability decreases until launch angle reaches about -30, increases slightly from -30 to +10, then decreases again as launch angle increases. The increase in hit probability from -30 to 10 is likely due to solid contact being made by the hitter on these batted balls, which are hit more as a line drive rather than a ground ball. As a result, the ball is likely to be traveling quickly because it isn't being slowed by contact with the ground, giving infielders less time to react. The decrease seen after 10 degrees occurs because hang time increases with launch angle, giving fielders time to track and catch the ball.

PDP only shows marginal effects of each predictor, but Fielder 1 distance, launch angle, exit velocity and other factors work together to determine whether a batted ball will be a hit. Joint partial dependence plots can illustrate interactions between features. We do not include the joint PDPs because the nature of the large data set made the joint PDPs computational expensive.

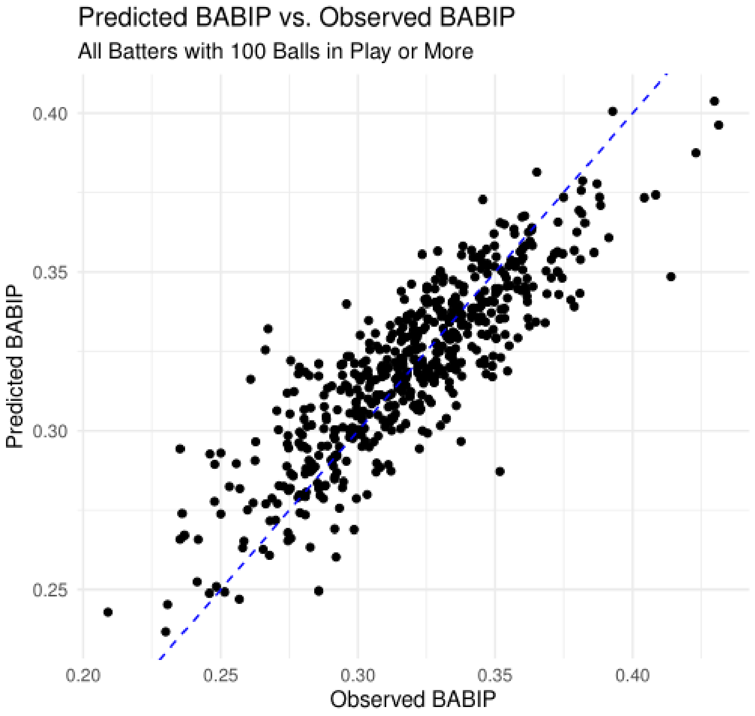


Figure 5: Predicted BABIP vs. observed BABIP for all hitters with at least 100 balls in play in the final data set. The model's predictions have a strong correlation with the observed results for each hitter.

To further explore our model's performance and ability to estimate a player's batting average, we compared the observed BABIP to the model predicted BABIP for 566 hitters with at least 100 balls in play in the final data set. The scatterplot showing the relationship between these two metrics is shown in Figure 5. The points tend to fall along the 1:1 line, and the correlation is 0.86. The agreement between the observed BABIPs and the model predicted BABIPs indicates that our model can provide reasonable predictions of a player's hitting performance.

We performed further testing of our method on three MLB hitters: Joey Gallo, DJ LeMahieu, and Hunter Renfroe. We chose these hitters because they differ in their batting stance (left vs right) and their hitting tendencies. Gallo and Renfroe are known to be pull hitters that hit for power. Gallo hits from the left side, so he tends to pull the ball to right field. Renfroe from the right side and tends to pull the ball to left field. LeMahieu also hits from the right side but is known as a contact hitter that has a tendency to hit the ball the other way (i.e., to right field, away from his pull side). We used these hitters to demonstrate the results of our defensive alignment optimization method and the effectiveness of four-person outfields.

We searched for optimal defensive alignments and tested the effectiveness of 4-person outfields against each of the three exemplary hitters. We first found the alignment that minimized each hitter's BABIP from among all observed alignments used against each hitter (i.e., best observed alignment). Then we used 5,000 iterations of the Monte Carlo optimizer to search for a similar, but improved, alignment resulting in a lower BABIP. We then performed a similar process to evaluate the effectiveness of a 4-person outfield. Because 4-person outfields were rarely used, we searched for the best 4-person outfield alignment against each hitter from among all 4-person outfields used during the time period (not just those unique to each hitter). For each of these demonstrations, we use Great American Ballpark, the ballpark of the Cincinnati Reds, to demonstrate our method. We choose this ballpark because it has a fairly standard outfield depth and wall configuration.

3.2. Defensive Alignment Optimization

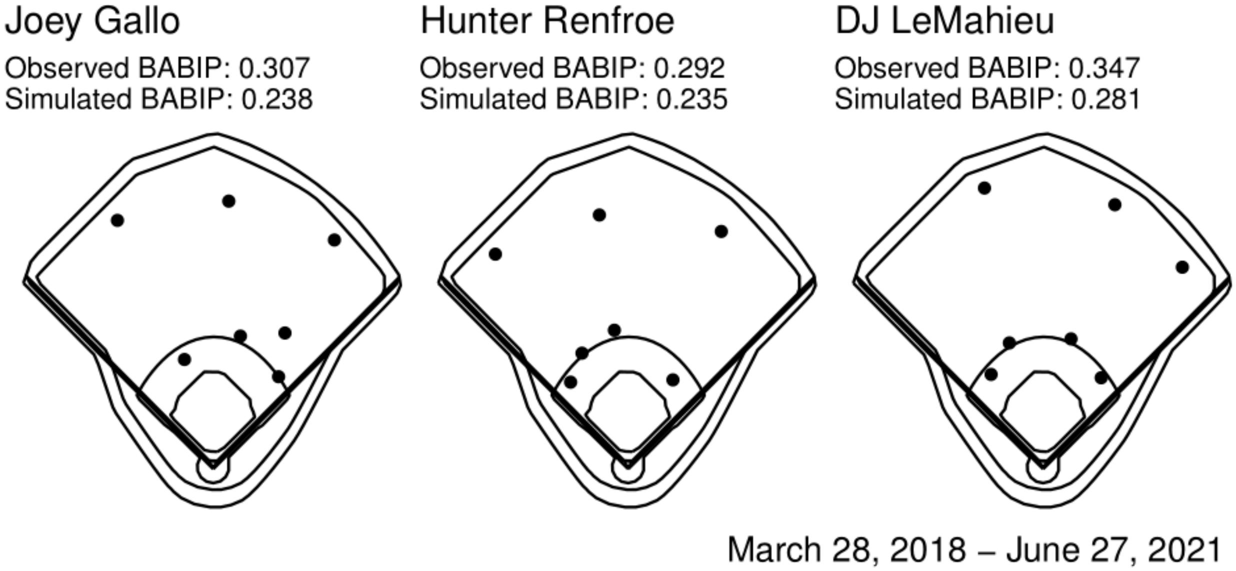


Figure 6: The optimal defense against three hitters based on the Monte Carlo optimization. The hitting tendencies of each batter are reflected in the optimal defense. If these defenses were used against the hitter over the time period used for this study, each hitter's BABIP would decrease by an estimated 0.055 points or more.

The optimal defensive alignments for Gallo, Renfroe, and LeMahieu are shown in Figure 6. For Gallo the optimal defense has the outfielders slightly shifted toward right field (Gallo's pull side). The infielders are shifted towards the right side with the exception of one player (e.g., the third baseman) on the left side of second base. One of the infielders (e.g., the second baseman) is positioned on the outfield grass. Gallo's actual BABIP for this time period in the cleaned dataset was 0.307. Our model estimates that his BABIP would decrease by 0.069 if this optimal defense had been used against him on every batted ball over that same period.

For Renfroe, the optimal alignment similarly has the outfield slightly shifted towards Renfroe's pull side. The infield is shifted to the left-field side of second base, which is Renfroe's pull side.. For Renfroe, all shifted infielders are positioned around the same distance from home. This is because the fielders are much further from first base and can't realistically play as deep because of the distance of the throw. The first baseman has to stay close to first base which leaves a big gap on the right side of the infield. This is different from the shift on Gallo because the third baseman has more room to close the gap on the left side of the infield. Renfroe's actual BABIP for this time period in the dataset was 0.292. Our model estimates that his BABIP would decrease by 0.057 if this optimal defense had been used against him on every ball over that same period.

For LeMahieu, the infield is playing in a relatively standard alignment with no shift in either direction. The outfield is heavily shifted towards right field and playing very deep. This is unusual but it shows the tendency of LeMahieu to hit the ball the other way in the air as a right handed hitter. LeMahieu's actual BABIP for this time period in the dataset was 0.347. Our model estimates that his BABIP would decrease by 0.066 if this optimal defense had been used against him on every ball over that same period.

3.3. Four Person Outfield

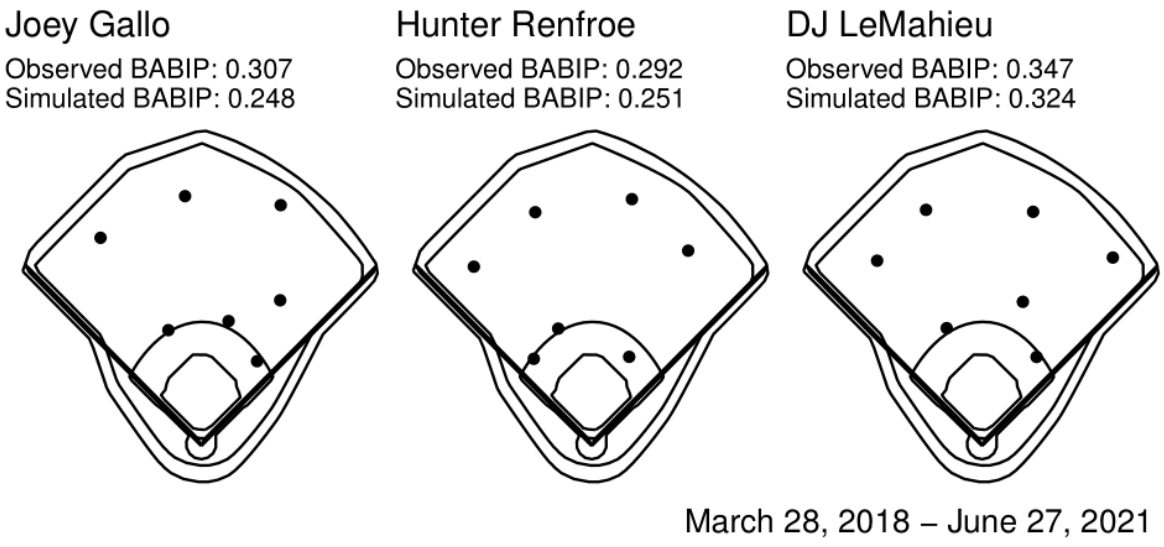


Figure 7: The optimal four-person outfield against three hitters.

The optimal four-person outfield defensive alignments for Gallo, Renfroe, and LeMahieu are shown in Figure 7. For Gallo, the optimal four person outfield looks quite similar to his overall optimal defense. The biggest difference is the fielder positioned in short right field is much deeper. This change allows the right fielder to move deeper and shade towards center field. Since the fielder in short right field is positioned far from home, this alignment is considered a

4-person outfield alignment. Our model estimates that Gallo's BABIP would decrease by 0.059 if this optimal defense had been used against him on every ball over that same period.

For Renfroe, four fielders are evenly positioned across the outfield, with the outfielders on the right side shading slightly deeper. Two infielders are positioned on Renfroe's pull side (left of second base) with the first baseman as the only fielder on the right side of second base. The first baseman is playing very far off first base to close the gap between him and the other infielders. Our model estimates that Renfroe's BABIP would decrease by 0.041 if this optimal defense had been used against him on every ball over that same period.

For LeMahieu, four fielders are evenly positioned across the outfield, with the farthest right fielder playing deeper than the others. The infield is positioned with only one fielder on the left side, in the region where the shortstop is traditionally positioned . The right side of the infield consists of the first baseman, playing very deep, and another fielder playing in short right field. LeMahieu frequently hits away from his pull side, so the infield of his 4-person outfiel is much different than the right-handed and pull-heavy Renfroe. Our model estimates that LeMahieu's BABIP would decrease by just 0.023 if this optimal defense had been used against him on every batted ball over the same period.

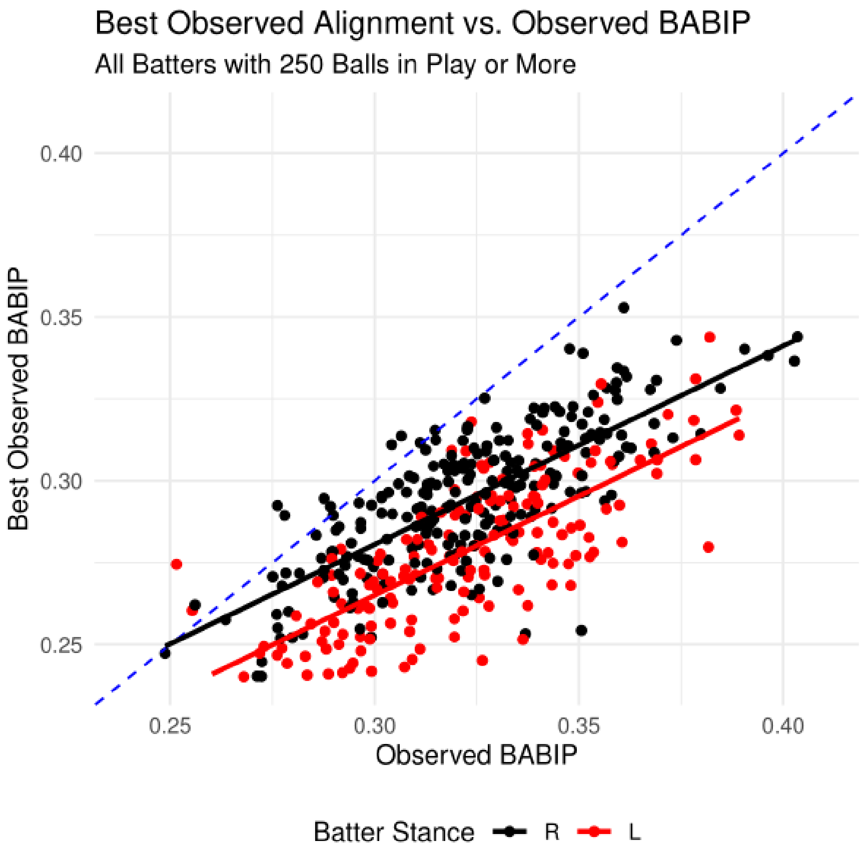


Figure 8: Observed BABIP vs. best observed alignment BABIP. We found each hitter's best observed alignment used against them. For all hitter's and their batting side (L/R) with over 250 balls in play in the final data set.

The results of our search for the best defense against each hitter indicated that almost all hitters would see a decrease in BABIP if teams used the best defensive alignment. Figure 8 shows the observed BABIP for each hitter against the predicted BABIP resulting from the best observed alignment. This figure shows that as a hitter's observed BABIP increases, the larger the difference between observed BABIP and best observed BABIP. It also shows that left-handed hitters (LHH) tend to have a larger difference between each BABIP metric than right-handed hitters (RHH). This may be in part due to shifts used against righties still having the first baseman positioned close to first base, effectively limiting the number of infielders that can be shifted toward a RHH's pull side. Because LHH are more often shifted against, it also suggests that MLB teams have room to explore new alignments against RHH that could further decrease BABIP. In 2021, defensive shifts were utilized against LHHs for 52.5% of plate appearances while just 16.2% of the time against RHHs {cite: Baseball Savant}.
Overall, we compute an average decrease in BABIP for LHH of 0.044 and 0.029 for RHH. These decreases are 14.0% and 8.9% of the observed average BABIP for LHH and RHHs, respectively. These decreases are larger than those reported for defenses with a shifted infield or four-person outfield found by Bouzarth et. al (2021) {cite}. The LHH with the largest decrease between observed BABIP and best observed alignment BABIP is Scooter Gennett with a decrease of 0.102 in BABIP. The RHH with the largest difference is Tom Murphy with a decrease of 0.096.

4. Discussion & Conclusions

We developed a method for predicting the probability a batted ball will be a hit based on the starting coordinates of fielders and batted ball and hitter characteristics. Our analysis is the first publicly available analysis utilizing the starting coordinates of fielders in MLB. We used our method to estimate a hitter's BABIP for any defensive alignment used against them. Our approach revealed important features, such as the distance the nearest fielder has to travel or launch angle, that predict whether or not a batted ball will be a hit. We demonstrated how our method can be used to find optimal defensive alignment strategies and evaluate the potential effectiveness of four-person outfields against individual hitters. Finally, we showed how improved defensive alignments could decrease BABIP across many hitters and how defensive alignments affect LHH more than RHH.

One of the biggest challenges of our analysis was relying on features that were recorded by a human. For example, the human-specified hit coordinates were noisy (verified using video). Similarly, batted balls are categorized as a line drive, fly-ball, pop-up, or ground ball by a human. A plot of launch angle versus exit velocity colored by batted ball categorization showed substantial overlap of categories. This indicates there is an opportunity to improve these categorizations using data from tracking systems.

There are several ways to improve or extend the methodology we developed. Our current method ignores in-game situations that affect hitter and defensive strategies. For example, we assume that there are no runners on base. Runners on base, especially runners on first and/or second base, can limit possible defensive alignments by requiring fielders to “hold” runners to prevent stolen bases or position themselves to enable double plays. Similarly, runners on base, the score, the inning, the number of outs, and the count can all influence a hitter’s approach.

Another extension of our method would be to include features of individual pitchers or fielders. For example, some pitchers are described as “ground ball” or “fly ball” pitchers. The most effective defense against a hitter may depend on the particular pitcher they are facing. This could be implemented by developing an “expected” spray chart for a given pitcher/batter matchup. Likewise, the most effective defense for a given team may depend on the skills of their fielders. For example, a team with a rangy center fielder may be able to place their corner outfielder closer to their respective foul lines. Not every fielder has the same skills. However, we are assuming every player has the same fielding skills in the outfield and infield. To make this more unique, individual player skills could be taken into account.

A third way to improve our method is to account for the dimensions and geometry of stadiums. Each stadium’s outfield has different wall configurations (e.g. height, corners) and dimensions (e.g., distance from home plate) which can lead to different outcomes of batted balls to the outfield. For example, the outfield alignment will be different if a team is playing in Fenway Park, with a very short left field compared to Coors Field, which has a much deeper left field. Ballpark specific effects could be important given that some outfielders will have to position themselves differently in different ballparks.

An final way to extend our method would be to substitute run value for hit/not hit as the response. Doubles and triples typically lead to more runs scored than singles. Positioning fielders to decrease the number of extra base hits could be beneficial to a team over the course of a season. This thinking is a reason for the increased use of four person outfields by MLB teams {cite: BaseballCloud article}. Among the three hitters examined closely here, our model indicates that a four person outfield would be most effective against Gallo and least effective against LeMahieu. None of the four person outfield alignments are estimated to be as effective as the optimal defensive alignments discussed in Section 3.2. As a result, we would recommend that four person outfields be only used strategically and not regularly.

An on-going discussion about defensive shifts is the ability of hitters to “beat” them. That is, to adjust their approach to hit the ball away from where fielders are positioned. Model {cite} examined the ability of hitters to adjust to defensive shifts and did not find evidence of successful hitter adjustments.

- Something about how alignments and strategy will continue to improve with more data and better sensors (e.g., information about spin rates, ball trajectories).

Despite the ban on infield shifts, there is still an opportunity to optimize defensive alignment, and our approach illustrates one possible way to find the best defense against a hitter.

5. Acknowledgments

The authors thank MLB Advanced Media for sharing the fielder positioning data and the Weller Stats Lab for providing feedback.

6. Citations

- http://tangotiger.com/images/uploads/History_of_the_Fielding.pdf
- https://03240aac-a3d1-4c7f-b69a-06f1eb1e80cb.usrfiles.com/ugd/03240a_608fba8fc66b4844a5752dcbd3463cb5.pdf
- Link
    - Shows Padres are shifting again LHH
    - LHH batting avg against Padres less than league average

Lewis, M., & Bailey, R. (2015). Batted ball spray charts: A system to determine infield shifting. 2015 Systems and Information Engineering Design Symposium, SIEDS 2015, 00(c), 206–211. https://doi.org/10.1109/SIEDS.2015.7116975

Becker, K. (2015). Optimizing baseball defensive alignments through integer programming and simulation. Angewandte Chemie International Edition, 6(11), 951–952., 10–27.

Model, M. W. (2020). Hitting around the shift : Evaluating batted-ball trends across Major League Baseball. May, 1–26. https://www.mhlmdl.com/_files/ugd/4e1ecc_dd865181ef5e41b8988abee0124bafd8.pdf

Gerlica, J., LaDuke, I., O'Shea, G., Pluemer, P., & Dulin, J. (2021). Quantifying the Outfield Shift Using K-Means Clustering. Industrial and Systems Engineering Review, 8(1), 18–23. https://doi.org/10.37266/iser.2020v8i1.pp18-23

Hawke, C. J. J. (2017). Quantifying the Effect of The Shift in Major League Baseball. Senior Projects Spring 2017. https://digitalcommons.bard.edu/senproj_s2017/191

Bouzarth, E., Grannan, B., Harris, J., Hartley, A., Hutson, K., & Morton, E. (2021). Swing shift: A mathematical approach to defensive positioning in baseball. Journal of Quantitative Analysis in Sports, 17(1), 47–55. https://doi.org/10.1515/jqas-2020-0027

Fast, M. (2010). What the Heck is PITCHf/x? The Hardball Times Baseball Annual 2010, 1–6. http://baseball.physics.illinois.edu/FastPFXGuide.pdf

Jensen, S. T., Shirley, K. E., & Wyner, A. J. (2009). Bayesball: A Bayesian hierarchical model for evaluating fielding in major league baseball. Annals of Applied Statistics, 3(2), 491–520. https://doi.org/10.1214/08-AOAS228

Montes, A., Argenziano, A., O'Sullivan, B., Orlinsky, C., Posner, D., Chagares, M., Flieder, A., Bookstein, B., Chernow, J., Brodsky, T., & He. (2021). Optimizing Outfield Positioning: Creating an Area-Based Alignment Using Outfielder Ability and Hitter Tendencies.

Bill Petti (2021). baseballr: Functions for Acquiring and Analyzing Baseball Data. https://billpetti.github.io/baseballr/, https://github.com/BillPetti/baseballr/.

Baseball Savant

History of Fielding

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Brandon Greenwell, Bradley Boehmke, Jay Cunningham and GBM Developers (2020). gbm: Generalized Boosted Regression Models. R package version 2.1.8., https://CRAN.R-project.org/package=gbm

Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret

Possible Journals:

- The Baseball Research Journal
- Journal of Quantitative Analysis in Sports
- Journal of Sports Analytics
- https://community.amstat.org/sis/journals
- https://www.sfu.ca/~tswartz/papers/publish.pdf

Notes Dump

Old Introduction

- P1: Data and Analytics in baseball and how they have rapidly expanded (player tracking, ball tracking, spin rates, etc..)
    - Started with the Moneyball book in the early 2000s.
    - Billy Beane used a more quantitative way to evaluate players outside of subjective thoughts of scouts and general managers
    - PITCHf/x introduced in 2006 playoffs, tracked pitch speed, location, movement, release point, spin - 2 cameras mounted in each stadium
    - 2009 HITf/x introduced to give trajectory, angle, and velocity of batted balls.
    - 2015 Statcast was introduced
    - 2020 Hawk-Eye (used for instant replay in tennis) - 12 cameras, 5 for pitch tracking (100 fps) and the other 7 for tracking players and batted balls (50 fps)
    - Allows for other ways to analyze player performance and on-field strategy in order to optimize results
    - Citations
- P2: analytics are important; large data science teams in MLB organizations; used to get an edge
    - Analytics teams continue to increase in size due to the massive amount of data to analyze
    - Teams that don’t utilize this information are further behind than those that do

- P3: Defensive alignment;
    - what is it? why it matters?
        - 7 players are open to freely move anywhere on the field
        - If a team can position themselves better, they will convert more hits into outs, which will mean less runs allowed, meaning more wins.
        - Different hitters have different tendencies
        - From 2018-2021, 63% (AB - HR - K + SH / PA) of plate appearances end in a batted ball that fielders can make a play on.
        - That means the fielding alignment matters in 63% of all plate appearances
    - Strategic alignments
    - Example of the 4 man outfield; 2B playing in shallow right field
        - Teams have opened up to more radical shifts ^
        - Gives me the idea that their can be even more shifts that are way against the norm that can optimize outs more than more common shifts

    - Teams are paying attention to alignments; more often we see the defense adjusting to the hitter
    - Shifting percent has gone from 13.7% In 2016 to 30.9% In 2021.
    - Defensive shifts have garnered more attention in media and analysis coverage over the past few years. For example, ?? discussed the possibility that shifting cost the Atlanta Braves some playoff wins in 2020.
    - Hawke Jr. (2017) examined the effects of the outfield shift and found that….
    -
    - Relatively little public, open-source research has been done on the effects of shifting all 7 seven fielders together
- P4: outline our research questions on defensive alignment; can we use batted ball characteristics and hitter characteristics to best position fielders?
    - Previous research used zones and the spray chart to create a better defensive alignment
    - Statcast allows for much more information that zones don't need to be used
    - Batted ball characteristics, and each player's position on the field allow for a different approach to the problem
- P5: We present an analysis of defensive positions; create models that show how defense positions matters

- We have data with the defensive players' starting position (never been published!)

Optimal defensive alignment??

(I think much of the rest of this section can be removed.)

- We didn't account for possible extrapolation → implementing a defensive alignment that has never been seen before.

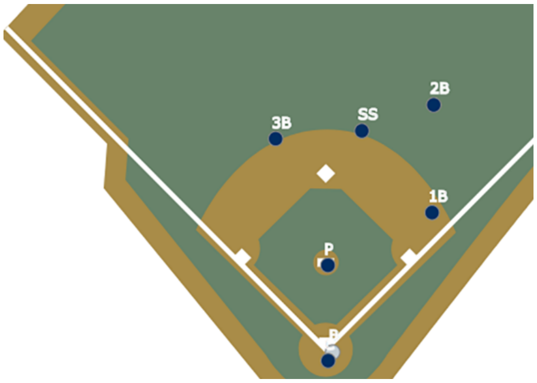
, but the recent development and implementation of player tracking systems enables new analysis of these features.

the recent development of new tracking systems and the implementation of defensive shifts, there are many open questions about the effects of changing defensive alignments on the success of batters, or conversely, the success of the defense at making an out. Additionally

- What we did
- The contributions of the work

- Preview of what is in the article

To give an example of how Major League teams are currently repositioning their fielders, a very common use of the shift is to pull heavy left handed hitters, where 3 fielders will position themselves on the right side of the infield as pictured below.



The reasoning behind this shift is that pull-heavy left handed hitters will hit a large portion of their ground balls into the wall of defenders on the right side. The probability of each batted ball becoming an out decreases compared to a straight up infield. If a team can play their fielders where the ball will most likely be hit, they will turn more hits into outs.

This has led to some creative shifts in recent years including the Padres playing Manny Machado in short right field and Nick Ahmed playing in short left field when Albert Pujols is at the plate. However, these creative alignments are still very rare and we want to know if they should be used more often. It's not easy to be at the forefront of something different than the norm, but one example of a team having success doing this is the Tampa Bay Rays. The Rays are known for their creative shifts on a nightly basis and their willingness to be creative has led to 110 runs saved from the shift since 2017 according to Sports Info Solutions, most in the league.

Old Data Cleaning Notes

- Describes the player position data (what I have)
    - Includes 9 position players X and Y coordinates at the moment of contact.
    - Each player's "book position" (e.g, 5 denotes third baseman)
- Data sources; what variables are included; what years?
    - MLB data and data descriptions
    - Batted ball characteristics
    - Fielder starting positions
- Batter characteristics (left vs right "stand")...
- Exact dates
- Total number of observations (all pitches)
- Batted ball information - "these data include hit coordinates that indicate where the batted ball was picked up."

^^Paragraph about data joining and cleaning

- Joining the data based on play ID
- Pitches that were claimed to be thrown in a count with 4 balls or 3 strikes and batted balls initially touched by the catcher or pitcher, according to the data set, were removed.
- Removing home runs
- Removing strange hit coordinates
- Removing rows that did not have a Statcast hit distance
- Final number of observations after data cleaning
- Cleaning the statcast data and issues with statcast data - creating new variables
    - hand record hit locations
    - “Through our exploratory data analysis, we discovered that there were accuracy issues with the hit coordinate information. This discovery was enabled via the use of MLB videos {cite: MLB film room}. We cleaned the data to remove these inaccurate hit coordinates”
    - Talk about the observations we go rid of because of coordinates
    -
- How did we handle runners on base?
    - Runners on base and those effects
    - Runners being on base will/can affect fielder position (1b holding the runner on, positioning for bunts or double plays)
    - Wanted larger sample size

Old Methodology Notes

Data from Baseball Savant includes a column describing three different alignments for the infield and outfield at the time of a pitch. The alignments include “Strategic”, “Standard”, or “Shift” for the infield and “Strategic”, “Standard”, and “4th Outfielder” for the outfield. For the sake of this analysis, all alignments categorized as “Strategic” were changed to “Standard”. Basic analysis can be done to find the batting average within these different alignments by choosing which combination of outfield and infield alignment is best against a hitter. However, knowing the exact locations of fielders can increase the accuracy of the analysis and the opportunity. We can move each fielder independently, helping us understand exactly where the best spot on the field is.

- Different methods of describing defensive positioning
    - Shift vs standard vs strategic: “coarse” description of defensive alignment, (can we tie in the idea that we need additional detail to do a good job? E.g., improvement in predictive ability?)

Literature Review

Hawke Jr. used probit regression to examine how the repositioning of defensive players affects their success rate of fielding a batted ball compared to if that player was positioned normally. {cite: Quantifying the effect…, 2017}.

- Capstone/thesis project (not peer reviewed)

- Hit velocity is 3 categories (arbitrary)
- Distance player traveled: estimate the starting coordinates of paper
- Direction player traveled to field the ball (binary, forward or backward)

Model {cite: Hitting around the shift: … 2020} used hierarchical Bayesian regression models to assess the effect of the defensive shifts on hitters. They did not find any batters that had a significant difference between their spray chart under the two defensive scenarios. As a result, they conclude that defensive shifts will continue due to a lack of adaptation by hitters.

- Capstone/thesis project (not peer reviewed)
- Used logistic regression to get probability shifts were implemented
-

Gerlica et al. {cite: Quantifying the outfield shift...2020} use the coordinate where the ball landed and K-Means Clustering to increase the outfield catch probability for the Air Force Academy Baseball Team. They set up K-Means Clustering to create three centroids to determine where the three outfielders would be located at the time of contact. From there they determine the catch probability of every batted ball from the hang time and the distance each fielder is to the landing spot. This led to a catch probability increase of 7.4% for a one-game sample of Air Force Academy Baseball.

- Limited to one game
- Limited to outfield alignment for 3-person outfield
- K = 3 k-means clustering
- Catch probability
- Capstone/thesis project (not peer reviewed)
- Tracking technology from University of Nebraska

Lewis and Bailey, 2015 {cite: Batted Ball Spray Charts: A System to Determine Infield Shifting} uses historical batted ball data for a given hitter and pitcher, the count, and the handedness of both the pitcher and batter to determine a distribution of horizontal angles in which a baseball will likely be hit. The horizontal angles are then placed in 9 different zones, ranging from the third base foul line, to the first base foul line. From there the model can be presented to coaches so fielders can be positioned accordingly.

- Infield is discretized into 9 zones
- Only infielder placement
- Accounts for base runners, outs, count, specific batter & pitcher
- Tampa Bay Rays & proprietary information

Montes et al. {cite: Optimizing outfield positioning:....2021} use outfielder's sprint speed and Outfield Jump from baseballsavant.com to create a circle of an outfielder's range for the 3 seconds after contact is made. A set of three outfielder's range circles are then placed in the field to optimize the total number of hit coordinate locations covered against a specific hitter's observed spray chart. They found an increase of batted balls reached by outfielders upwards of 12.7% when testing on several different fielding and hitting combinations..

- Home runs excluded

- field of play ignored
- hitter's are assumed to not adjust to the change in outfield alignment
- arm accuracy and arm strength ignored

Easton&Becker 2017, Becker {cite: OPTIMIZING DEFENSIVE ALIGNMENTS IN BASEBALL...2009}: discretized the playing field into defender locations and used integer programming with constraints to find the defensive alignment that minimizes a player's batting average based on their spray chart. They found a decrease in Derek Jeter's batting average from .317 to .309.

- They make assumptions about the radius in which fielders will make outs
- the expected costs of each hit from a batter to determine the optimal location of defensive players. A grid is placed over the field and each player is given a coordinate. Then, using a batter's hit coordinates, the batting average and slugging percentage are calculated.
- Defensive player can make out as long as they reach ball

Bouzarth et al 2021. {Cite: Swing shift: a mathematical approach to defensive positioning in baseball… 2020} use integer programming to create coverage zones for each fielder based on the historical batted ball locations (spray chart dots) of a given hitter. (Each fielder is given an area they "cover"). This method allowed for "non-traditional" alignments to be tested and improved upon. Their method lowered the predicted BABIP by 5.3% on average for right-handed hitters and 10.3% on average for left-handed hitters.

- Assume that 75 feet or less from home plate is covered by pitcher and catcher.
- No position labels
- Define 5-foot "patches" where defenders can be positioned; discretize the field into 4,229 patches
- Assume bases are empty
- Outfield is defined as 210+ feet from home
- Considers fielder placement considering their batted ball history and "risk areas" (areas with extra base hits can occur)

Article about beating the shift? Other articles about effective defenses?

- https://fivethirtyeight.com/features/dont-worry-mlb-hitters-are-killing-the-shift-on-their-own/
    - Hitters have a much lower ground ball percent when the shift is on
    - There is no effect to the ground ball percent when the shift is off
    - This means players are hitting more balls in the air, over the shift
-