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

Matthew Boyd*, Zachary Weller, Aaron Nielsen
Department of Statistics, Colorado State University
*mattboyd@rams.colostate.edu

Abstract

Over the last two decades tracking systems and analytics have changed the way Major League Baseball (MLB) is played. One example of this is the increased use of defensive shifts and development of creative defensive alignments to better defend hitters. The effects of these alignments has prompted MLB to place restrictions on defensive alignments starting in 2023. Several studies have examined the effects of defensive alignment on hitter performance. We utilize, for the first time, data that provides the starting coordinates of MLB fielders covering 3.5 seasons to quantify the effects of defensive alignment on batting average. Combined with batted ball data, we used the fielder coordinate information to create a position-agnostic description of defensive alignment. We then fit a gradient boosting model to estimate the probability a batted ball will be a hit given the alignment and batted ball features. Our results demonstrate the importance of defensive alignment in predicting whether a batted ball will be a hit and how defensive alignments can be improved to minimize a hitter's batting average. We used 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, and we developed an interactive R Shiny app that can be used to implement our method.

1. Introduction

Over the last two decades, the use of data science, statistics, and mathematics in sports has greatly increased. Teams from a 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" (Birnbaum 2021), in the early 1980s. James' way of thinking wasn't well known or popularized until the early 2000s when the book *Moneyball* (Michael Lewis 2004) 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 recent creation and deployment of tracking systems has provided even more data about pitches, batted balls, and player positioning beyond the traditional hitting and pitching statistics. In 2006 Major League Baseball introduced PITCHf/x, a system that can track pitch characteristics (Fast 2010). Shortly after PITCHf/x, HITf/x was introduced which tracked the velocity and angle of the ball off the bat (Fast 2010). 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 (Major League Baseball Advanced Media 2021a). The Statcast (Major League Baseball 2021b) 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 hitters. In 2016, teams implemented an infield shift (defined as three infielders on one side of second base) for 13.7% of plate appearances, but that number more than doubled to 30.9% in 2021 (Major League Baseball 2021b). Teams have employed creative defensive alignments 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 playoff 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 (FanGraphs 2021). Ben Lindbergh discussed the possibility that shifting may have cost the Atlanta Braves the National League Pennant in 2020 (Lindbergh 2020). Sports Info Solutions (Simon 2019) estimated that 517 runs were saved throughout the league from infield shifts alone in 2021 (FanGraphs 2021), 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 (Verducci 2022).

Several previous research studies have examined the effects of defensive positioning on the outcomes of batted balls and optimal defensive alignments. Hawke Jr. (Hawke Jr 2017), Model (Model 2020), and Gerlica (Gerlica et al. 2020) 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 (Myles Lewis and Bailey 2015) divided the infield into 9 zones and used information about the pitcher, batter, and count to find an optimal infield alignment. Montes et al. (Montes et al. 2021) 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 (Easton and Becker 2017) and Bouzarth et al. (Bouzarth et al. 2021) 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 these studies make assumptions about the size of the area that a defensive player can cover to field a batted ball.

We develop a novel analysis examining the effects of defensive alignment on hitter success in MLB. Our analysis used 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 combined 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 used 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 estimated the effectiveness of a four-person outfield, propose a method for exploring new, optimal defensive alignments, and created 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 MLB games recorded by Statcast from 2018 to 2021. Data about batted balls was obtained through Baseball Savant using the R package baseballr (Petti 2021). We also obtained data of the defensive players' coordinates in the field from MLB Advanced Media (MLBAM) under a limited license agreement (Major League Baseball Advanced Media 2021b). Briefly, these data included information about fielder positioning and the batted ball information for approximately 284,000 batted balls in play after data cleaning. We provide additional details 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 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, line drive).

We obtained information about batted balls from Baseball Savant through the baseballr 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 (Tango 2021). The data set also contained the human-entered feature hit type, which categorizes each batted ball as either a ground ball, line drive, fly ball, or popup.

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 (standard, strategic, or shifted) of the infield and outfield defensive alignment. Batter information included the batter's sprint speed and batting stand (right vs. left). 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 (Dilday

2021). Batted balls that did not have a Y coordinate greater than 0 were removed (approximately 2,200 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, hit type (ground ball, flyball, popup, or line drive), and game date. Our merged data set contained approximately 284,000 batted balls in play. 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 caught 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 derived from 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).

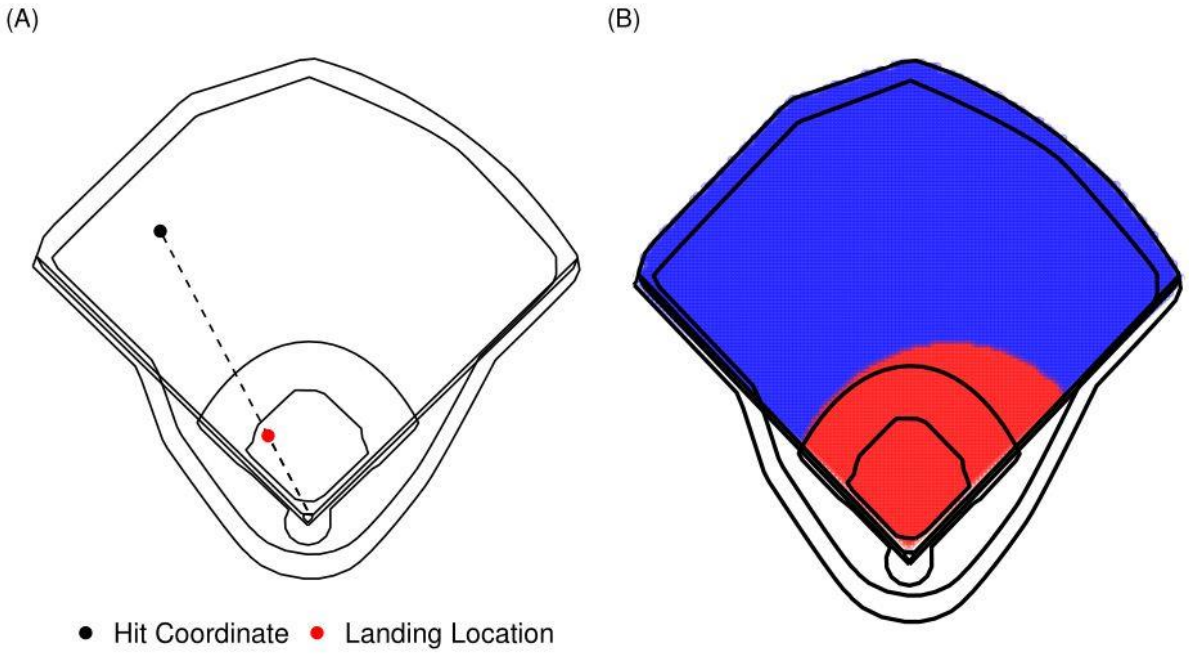

Figure 1: Two coordinates for batted balls and definition of the infield. Panel (A) shows the relationship between the hit coordinate and the landing location for a ground ball base hit. The hit coordinate is recorded by a human and shows where a fielder first touched the ball. We computed the landing 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 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 included these batted balls because they landed in fair territory. We removed foul outs because defenses 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 accuracy issues with the hit coordinate information. This discovery was enabled via the use of MLB videos (Major League Baseball 2021a) and confirmed by Tom Tango of MLB (Tango 2021). 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 leads to random error in the spray angle.

2.2. Nearest Fielder Methodology

We use the intercept point methodology from Tom Tango's Outs Above Average analysis (Tango 2020) to identify the location that each defensive player is projected to field a batted ball to make an out. Our approach 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 and is position agnostic. The defensive players in the fielder coordinate data are identified by their fielding position (e.g., 3 for first baseman, 5 for third baseman, etc.). Our initial statistical models used the fielder position specific coordinates as predictor variables. While these predictors provided good accuracy for classifying hits vs not hits, they do not enable a more general approach to characterizing 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. Given a batted ball and its characteristics, we identified a theoretical intercept point for each defensive player. The intercept point is based on where we assume each fielder initially intercepted 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). We defined the infield as any location in fair territory where the sum of the distance from that point to home plate and the distance from that point to first base is less than or equal to 290 feet. All other locations in fair territory are defined as the outfield. We refer to defensive players with starting coordinates in the infield as infielders, and players with starting coordinates in the outfield as outfielders.

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 location 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 used this definition for outfielder intercept points because very few ground balls or line drives fielded in the outfield are converted to force outs at first base.

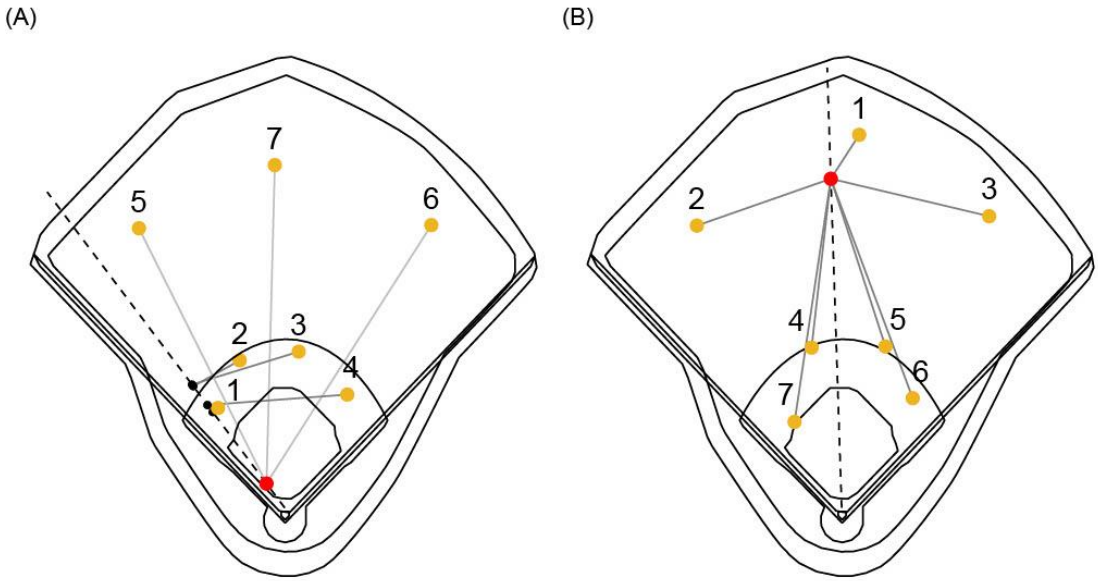

Figure 2: Intercept points for each fielder for a ground ball (A) and fly ball (B). The number next to each fielder indicates their rank for distance to their intercept point. Panel (A) shows each fielder's intercept point for a ground ball with a landing location near home plate. Each infielder's intercept 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 near 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 predicted path of the ball's travel, 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 computed 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's predicted line of travel. In Figure 2(A), the third baseman has the shortest distance between his starting location and intercept point giving them rank one. The shortstop has the second shortest distance giving them rank two. In Figure 2(B), the center fielder's starting position is closest to the landing location, giving them rank one.

In addition to the fielder distance ranking, we computed other features describing each fielder and their positioning relative to the batted ball. These features included 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 computed the difference between each fielder's angle and the spray angle. We also decomposed the distance between a fielder's starting location and their intercept point into horizontal and vertical components, where the vertical component is intended to reflect a player moving forward versus backward. Finally, we computed the distance between each fielder's starting location and first base, and each fielder's intercept point and first base. We applied the intercept point methodology to the observed defensive alignments for all

approximately 284,000 batted balls hit into play between March 28, 2018, to June 27, 2021. We also applied this process to new defensive alignments created by changing fielders' starting locations.

2.3. Statistical Modeling and Monte Carlo Simulation

We used batted ball characteristics and fielder positioning features to train a gradient boosting (GB) model to estimate the probability of a batted ball being a hit. We tried several classification methods and ultimately found the highest classification accuracy using GB. The GB model used three 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. The features used our GB model are listed in Table 1. We used 5-fold cross validation to train the model, which used 1,300 trees, shrinkage of 0.01, and an interaction depth of 8. We examined the variable importance from the GB model and the partial dependence plots (Friedman 2001) for 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 (Molnar, Bischl, and Casalicchio 2018).

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 user-specified alignments, and does not rely on assumptions about the area that a defensive player can cover to make an out. We quantify the efficacy of a defensive alignment against an individual player by estimating the player's batting average on balls in play (BABIP) 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 an estimate of a player's BABIP given the alignment. 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 BABIP against that defense.

2.4. Best Alignment, Alignment Optimization, and Four-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 search computationally tractable, we assumed that hitters are well-defended with defensive alignments previously used by MLB teams against that hitter, and thus constrained our optimization search to examine similar defensive alignments. 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. 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.

After finding the best observed defensive alignment, we use a constrained Monte Carlo random walk with Metropolis acceptance criteria (Spall 2005) 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 required that each fielder's starting position be no more than 10 feet away from their starting position of the best observed defense. We also required each fielder's starting position to be in fair territory. Finally, we required 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.

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 four-person outfield than others.

We searched for optimal defensive alignments and tested the effectiveness of four-person outfields against three exemplary 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 tends 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 used Great American Ballpark, the ballpark of the Cincinnati Reds, to demonstrate our method. We choose this ballpark because it has a standard outfield depth and wall configuration.

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 classified batted balls that are not hits about 95% of the time. Batted balls that are hits are classified correctly about 81% of the time.

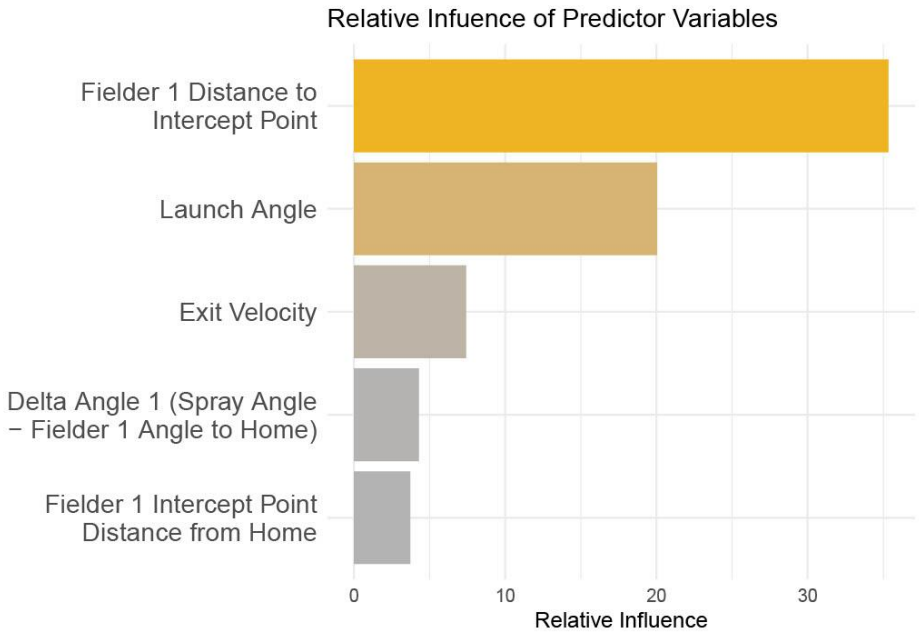

Figure 3: Variable importance from the gradient boosting machine model. Highlighting the importance of fielder positioning, the nearest fielder's distance to the intercept point is the most important feature. Launch angle is the second most important predictor, followed by exit velocity.

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 identified as 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 must 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 must go to reach their intercept point. It can also be an indicator of how far away the fielder is from their intercept point when also considering the fielder's distance from home 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.

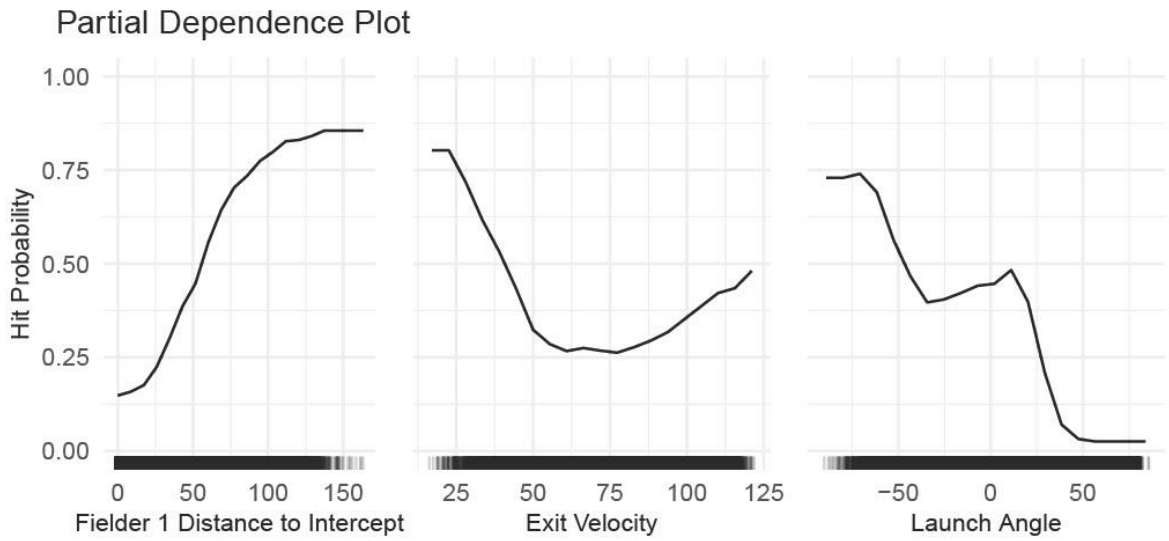

Figure 4: The partial dependence plots (PDP) for the three most important variables. The probability of a batted ball being a hit increases sharply as the distance to the nearest fielder increases.

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 very slowly off the 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.

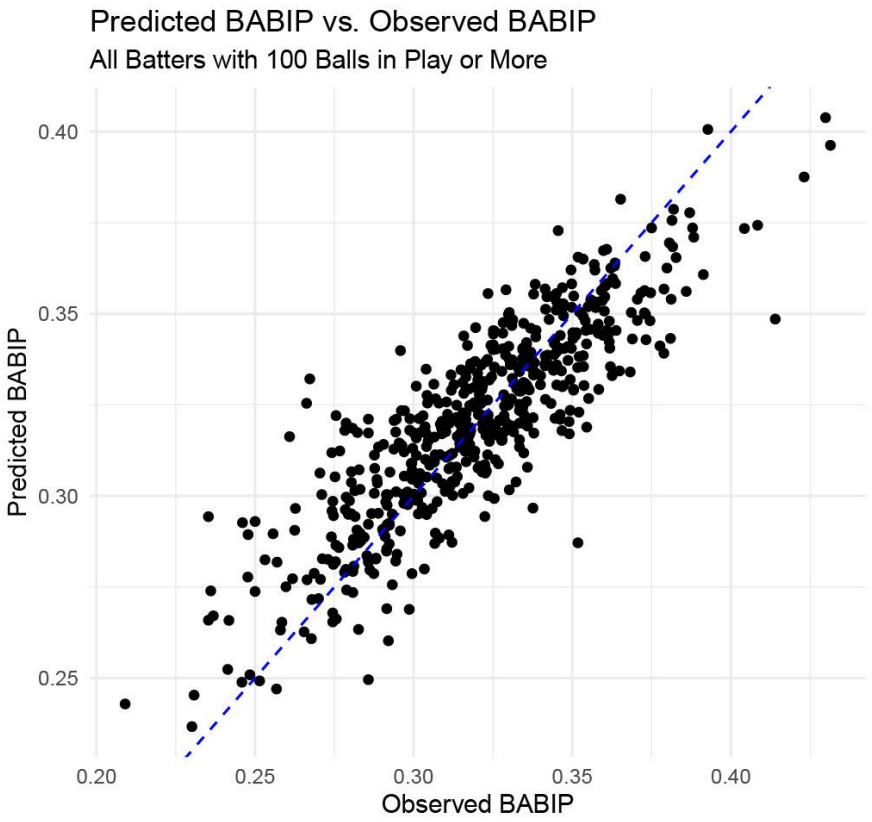

Figure 5: Predicted BABIP vs. observed BABIP for all hitters with at least 100 balls in play in the final data set. The model's predictions have a strong correlation with the observed results for each hitter.

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.

3.2. Best Defensive Alignments

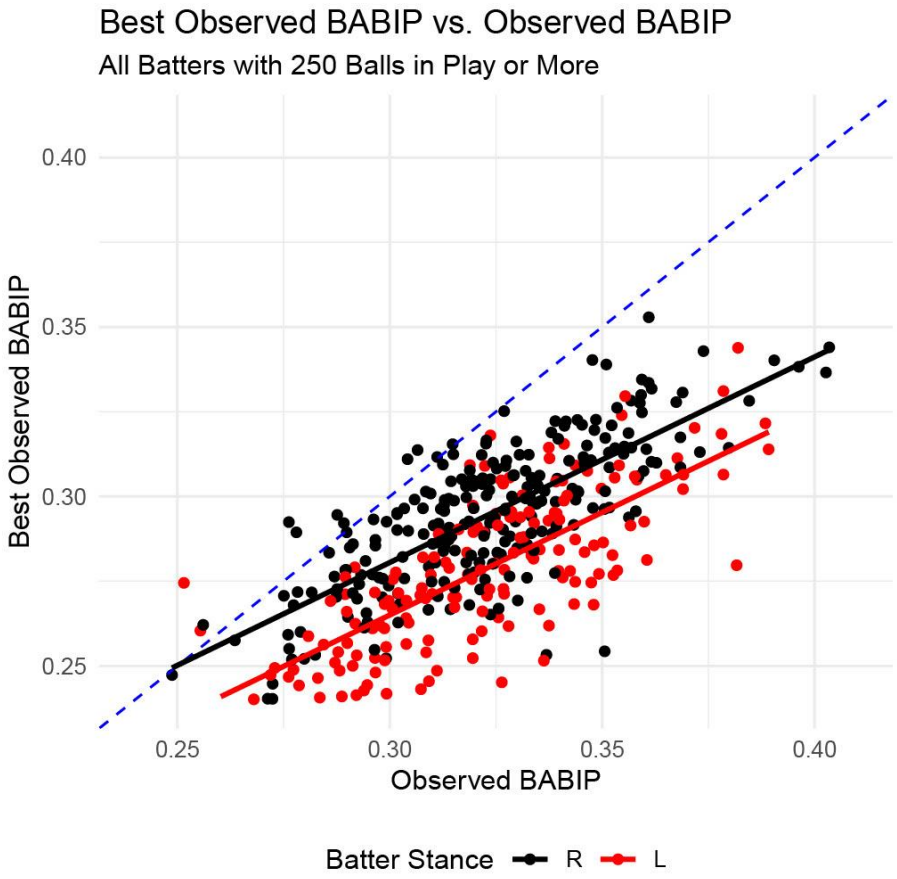

Figure 6: Observed BABIP vs. best observed alignment BABIP. We found each hitter's best observed alignment used against them. For all hitter's and their batting side (L/R) with over 250 balls in play in the final data set.

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 (Major League Baseball Advanced Media 2021a).

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 (Bouzarth et al. 2021). 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.

3.3. Defensive Alignment Optimization

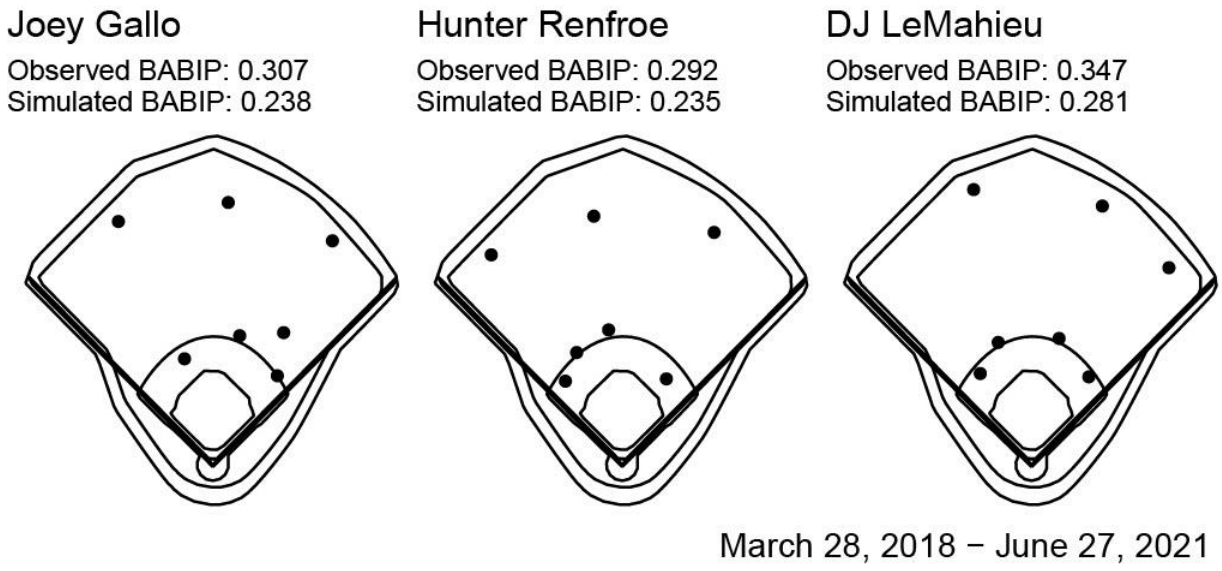

Figure 7: The optimal defense against three hitters based on the Monte Carlo optimization. The hitting tendencies of each batter are reflected in the optimal defense. If these defenses were used against the hitter over the time period used for this study, each hitter's BABIP would decrease by an estimated 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 except for 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 must 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 shifted towards right field and playing very deep. This is unusual but it reflects 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.

3.4. Four-Person Outfield

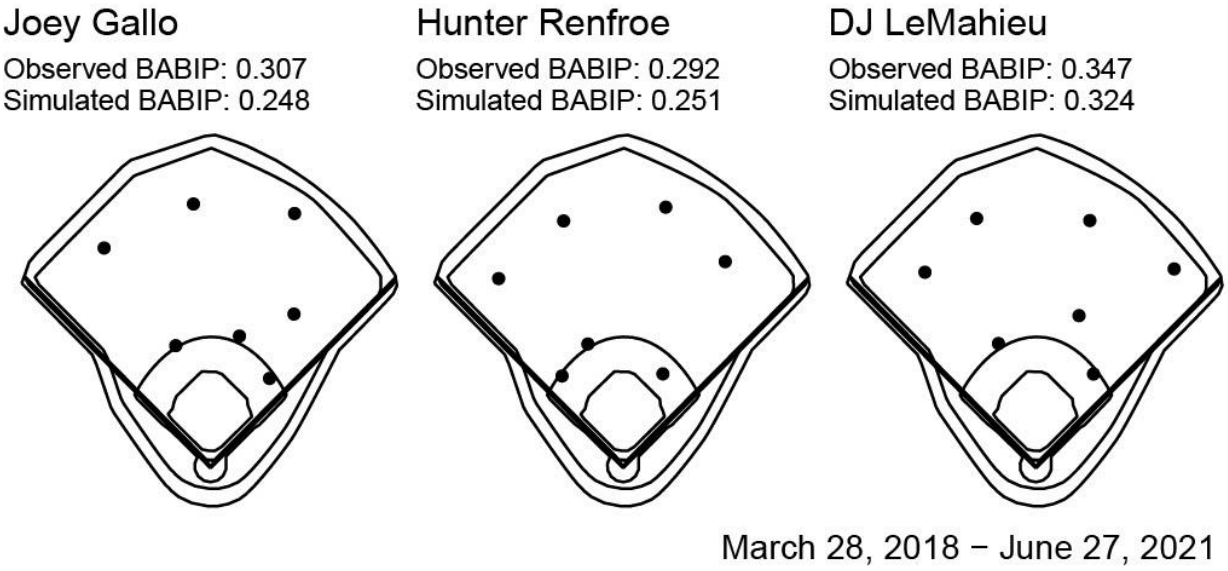

Figure 8: The optimal four-person outfield against three hitters. The optimal defense for each player reflects the tendencies in the location of their batted balls.

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

For LeMahieu, four outfielders are evenly positioned across the outfield, with the farthest right fielder playing deeper than the others. The infield is positioned with only one infielder 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 four-person outfield 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 four-person outfield had been used against him on every batted ball over the same period.

3.5. 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.

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 must travel or launch angle, that predict whether or not a batted ball will be a hit given the defensive alignment. We demonstrated how our method can be used to find optimal defensive alignment strategies and evaluated 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. Among the batted balls in our final data set, 57% were hit with the bases empty. 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. We anticipate that the results shown here could be adjusted to account for these situational considerations and be improved with future work that includes these factors.

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 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 considered.

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.

A 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 (Seibold 2020). 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.

As tracking technology continues to advance and more data becomes public, defensive positioning and strategy will continue to improve. Although the new 2023 rules on defensive shifts will impose limits on defensive alignment, there will still be an opportunity to optimize defensive positioning within the constraints of the new rules. Those constraints could be readily incorporated into our optimization method, and thus it could be used to find the best defensive strategy against a batter.

5. Acknowledgments

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

6. Citations

Birnbaum, Phil. 2021. “A Guide to Sabermetric Research.” Sabr.Org. 2021. https://sabr.org/sabermetrics.

Bouzarth, Elizabeth, Benjamin Grannan, John Harris, Andrew Hartley, Kevin Hutson, and Ella Morton. 2021. “Swing Shift: A Mathematical Approach to Defensive Positioning in Baseball.” *Journal of Quantitative Analysis in Sports* 17 (1): 47–55.

Dilday, Ben. 2021. “GeomMLBStadiums: GeomMLBStadiums: Draw Major League Baseball Stadiums with Ggplot2.”

Easton, Todd, and Kyle Becker. 2017. “Optimizing Baseball Defensive Alignments through Integer Programming and Simulation.” *International Journal of Modelling and Simulation* 37

(2): 82–87.
FanGraphs. 2021. “FanGraphs Leaderboards.” Fangraphs.Com. 2021. https://www.fangraphs.com/.
Fast, Mike. 2010. “What the Heck Is PITCHf/X.” *The Hardball Times Annual* 2010: 153–58.
Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics*, 1189–1232.
Gerlica, Jeffrey, Izaiah LaDuke, Garrett Oâ€™Shea, Pierce Pluemer, and John Dulin. 2020. “Quantifying the Outfield Shift Using K-Means Clustering.” In *Proceedings of the 2020 Annual General Donald R. Keith Memorial Capstone Conference. West Point, New York, USA. Accessed*, 7:16–21.
Hawke Jr, Christopher John. 2017. “Quantifying the Effect of The Shift in Major League Baseball.”
Lewis, Michael. 2004. *Moneyball: The Art of Winning an Unfair Game*. WW Norton \& Company.
Lewis, Myles, and Reid Bailey. 2015. “Batted Ball Spray Charts: A System to Determine Infield Shifting.” In *2015 Systems and Information Engineering Design Symposium*, 206–11.
Lindbergh, Ben. 2020. “The Shift May Have Cost the Braves a Pennant. Will It Cost Someone a World Series?” Theringer.Com. 2020. https://www.theringer.com/2020/10/21/21526283/defensive-shift-world-series-los-angeles-dodgers-will-smith-tampa-bay-rays.
Major League Baseball. 2021a. “MLB Film Room.” Mlb.Com. 2021. https://www.mlb.com/video/search.
———. 2021b. “Statcast.” Mlb.Com. 2021. https://www.mlb.com/glossary/statcast.
Major League Baseball Advanced Media. 2021a. “Baseball Savant.” Baseballsavant.Mlb.Com. 2021. https://baseballsavant.mlb.com/.
———. 2021b. “Major League Baseball Advanced Media.” Mlb.Com. 2021. mlb.com.
Model, Michael W. 2020. “Hitting around the Shift: Evaluating Batted-Ball Trends across Major League Baseball.”
Molnar, Christoph, Bernd Bischl, and Giuseppe Casalicchio. 2018. “Iml: An R Package for Interpretable Machine Learning.” *JOSS* 3 (26): 786. https://doi.org/10.21105/joss.00786.
Montes, Anthony, Anthony Argenziano, Brian OSullivan, Charles Orlinsky, Drew Posner, Matthew Chagares, Anna Flieder, Bennett Bookstein, Jack Chernow, and Teddy Brodsky. 2021. “Optimizing Outfield Positioning: Creating an Area-Based Alignment Using Outfielder Ability and Hitter Tendencies.” *The Baseball Research Journal* 50 (1): 92–104.
Petti, Bill. 2021. “Baseballr: Functions for Acquiring and Analyzing Baseball Data.”
Seibold, Chase. 2020. “Examining 4-Man Outfields.” Baseballcloud.Blog. 2020. https://baseballcloud.blog/2020/11/03/examining-4-man-outfields/.
Simon, Mark. 2019. “What Is Team Shift Runs Saved?” Sportsinfosolutions.Com. 2019. https://www.sportsinfosolutions.com/2019/04/01/what-is-team-shift-runs-saved/.
Spall, James C. 2005. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Vol. 65. John Wiley \& Sons.
Tango, Tom. 2020. “History of the Fielding.” 2020. http://tangotiger.com/images/uploads/History_of_the_Fielding.pdf.
———. 2021. “Personal Communication.” 2021.
Verducci, Tom. 2022. “MLB’s Modernization Is Underway, With More Radical Changes to Come.” Si.Com. 2022. https://www.si.com/mlb/2022/04/12/baseball-radical-changes-coming.