To Shift or Not to Shift? Using Player & Ball Motion Data to Build a Highly Flexible Defensive Positioning Algorithm in Baseball

Cameron Grove

Email: cameron.grove@durham.ac.uk

UCSAS 2022 - SMT Data Challenge

July 2, 2022

Abstract

In this project, I have measured many aspects of baseball player defense. These include catch probability, infield range, arm strength, & transfer speed. Combining models for these actions with batted ball distributions, I have built a general model for the likelihood of a ball in play resulting in an out. This model can be used to assess the efficiency for hypothetical fielding alignments, and a fitting method can be used to produce an estimate for the ideal defensive shift. The variation of these fits with fielder ability, batter spray distribution, and batter speed is presented. A website where any custom configuration can be tested has been hosted online[1].

[1] https://pitching.shinyapps.io/ShiftTester/

1 Introduction

Finding the ideal positions for fielders is a problem which has persisted in baseball since its inception, back when the rover roamed no-man's land between the infield and the outfield[1]. Positioning adjustments have evolved from the"Boudreau Shift"[2] to a menagerie of modern MLB methods aimed at removing hits.

This change is being initiated by the wealth of information provided to MLB front offices, and every MLB team now shifts at least 20% of the time, with some shifting on more than half of plate appearances[3].

As sabermetric knowledge becomes more complete over time, the exact impact of defensive shifts is a topic which has resisted a definite answer in public analysis. Even between MLB organisations there is a large divergence in opinion over how to place fielders, especially with decisions on how & whether to shift on left vs right handed batters.

In this project I have used player and ball tracking data to produce an algorithm which predicts the BABIP of a particular defensive shift, while accounting for the individual qualities of the batter and fielders. This can be used to estimate the ideal defensive positioning in a given situation, which can inform how fielders should be placed.

There is a lot of groundwork underlying this algorithm, including models of how likely players are to catch flyballs, and multiple facets of infield defense which is very sensitive to timing.

Section 2 describes the underlying models, including how they are built, tests of their accuracy, and which assumptions are used. In Section 3 I describe how I create a sample distribution of batted ball profiles and link this to create the general BABIP algorithm. Section 4 shows the results of tests of this model and how the suggested defensive shifts align with conventional baseball wisdom. Limitations and improvements are discussed in Section 5 before conclusions in Section 6.

The code associated with this project can be found on Github[2].

Before starting the project, I spotted some mismatches/errors in the initial data release and these were brought to the attention of the challenge organisers.

2 Defensive Models & Measurements

The core of a defensive positioning algorithim is to work out the likelihood of an out on a play. This can then be maximised to find an ideal fielding alignment.

There are two main types of out which I shall model in this section, flyouts and groundouts.

2.1 Catch Probability

A catch probability model finds the likelihood of a ball being caught by a specific fielder. The most important inputs to such a model are hang time and distance to ball landing location, effectively turning the question of catch probability into a timing comparison: Can the fielder reach the landing spot before the ball drops? Fielder speed, jump quality, and sure-handedness are also important, but these should be treated as residuals to a model which aims to find catch probability for an average fielder.

To build this model I filtered the input data to balls in play received by outfielders. These were tagged as catch or no-catch based on whether there was a ball bounce or ball deflection event before the fielder reception. A logistic regression for catch probability was created with hang time and distance as the only inputs.

The results of this model can be seen in Figure 1. Residuals of direction of travel were considered but didn't provide a strong enough effect to justify including this variable in the model. Spin can cause hooking/slicing/rising/sinking motion on the ball which could affect catch probability but these factors were also not considered.

2.2 Infield Range

When modeling the probability of a groundball out there are several steps in the process. The first is that the infielder has to reach the ball and stop it rolling into the outfield. All balls in play from

[2] https://github.com/PitchingBot/SMTDataChallenge - This is a private repository, email for access

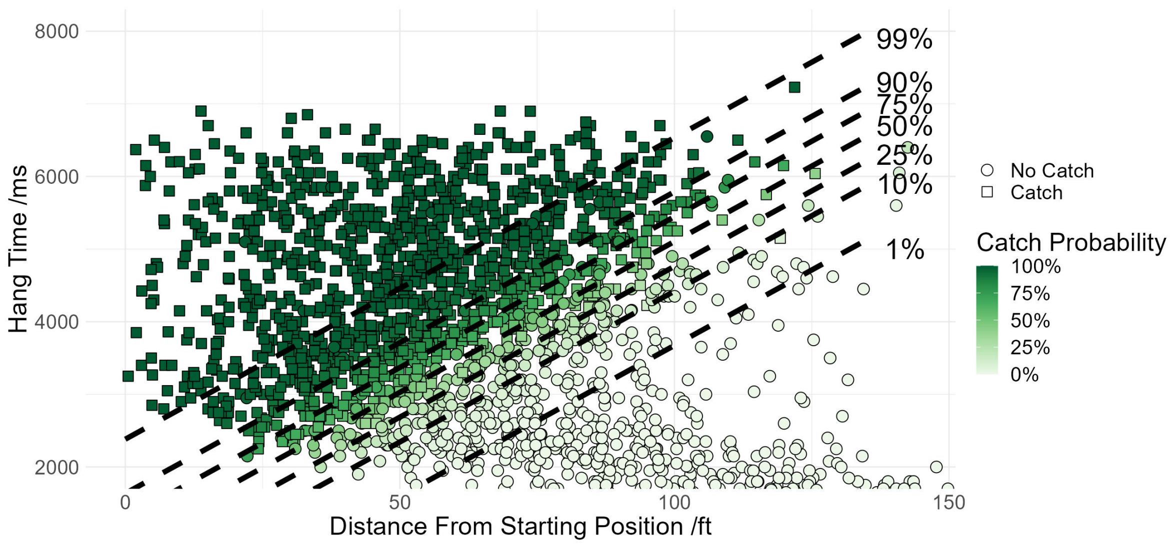


Figure 1: Results from the catch probability model. The clear relationship between hang-time and distance to landing spot can be seen. Dashed lines of constant catch probability are shown. This model captures the observed distribution of outfield catches well as seen by the mostly white circles and green squares.

the input data were considered, with the first fielder to reach the ball labeled in each case. A logistic regression was used with ball speed, launch angle, inverse ball speed, angle to fielder, fielder initial depth, and a couple of binary flags for locations in speed/angle space for locations which weren't being modeled well initially.

Launch angle and exit velocity were estimated using ball tracking values from several timesteps after the ball was hit into play as this was the most reliable method available to me. This results in launch angles substantially closer to 0 on groundballs than would be produced from Statcast-like data.

Results from this model and comparisons to the input data are shown in Figure 2. The fielding probability drops off sharply as the angle increases. The peak fielding probability is for a ball hit with intermediate speed directly at the fielder.

The next step is to model the time it takes for a fielder to reach the ball. This is achieved by using a linear regression model with exit velocity, inverse exit velocity, launch angle, angle to fielder, and initial depth as inputs. The results of this model as a function of angle and ball speed are shown in Figure 3.

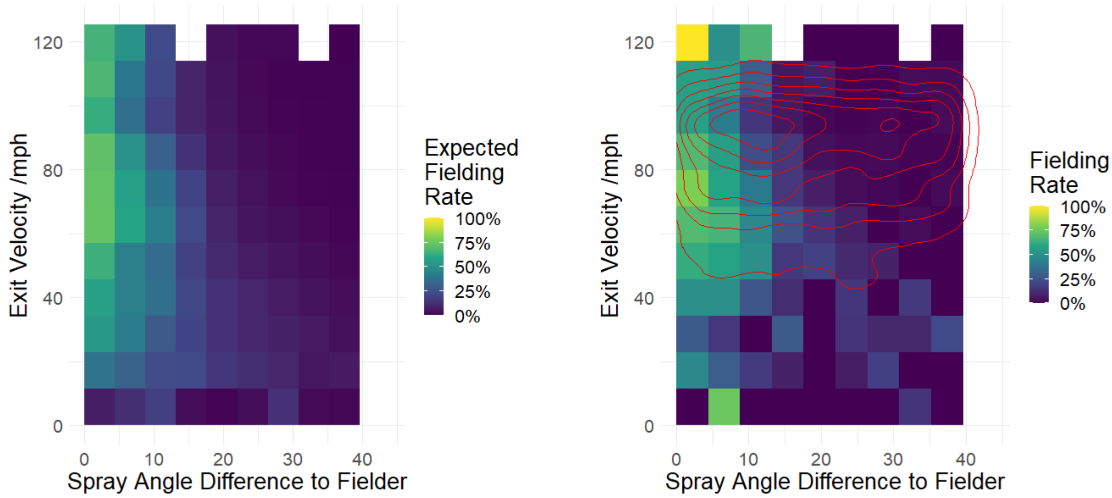
Finally, a model for the position at which the ball is fielded is needed. This tells us where the throw to first base comes from. Since the ball's spray angle is already known, only the depth needs to be modelled. Slowly hit balls tend to be charged by the fielder and fielded at a shallower depth than the player's initial position. A linear regression model was used for this aspect of infield defense, with results shown in Figure 4.

2.3 Infield Arm, Transfer Times, & Batter Speed

After a ball has been fielded, there is a race between the batter and fielder to first base. The batter has to sprint there meanwhile the fielder must transfer the ball to their throwing hand and fire it across to the waiting first baseman. Infield plays are very sensitive to this timing, which means that modelling it accurately is imperative.

I have already simulated how long it takes for a fielder to reach the ball, but the next step is finding transfer times and throw speeds and how they vary by position. Details of how these quantities were measured from the data can be found in the code on Github but are not included here for brevity. Only direct throws to first-base were considered.

Figure 5 shows the distribution of transfer times by infield position. Third basemen can take longer than middle-infielders for a couple of reasons. Firstly, their position may not select for the quickest hands generally as they are not involved in most double plays, secondly, they may need to



(a) Model Fielding Rates

(b) Data Fielding Rates

Figure 2: (a) Expected fielding rates from the infield range model. (b) Fielding rate from the input data in the same space as panel (a), the density of balls-in-play within this space is represented by the red contours.

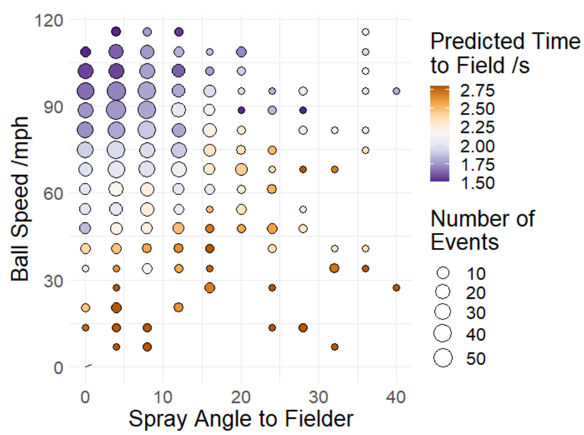


Figure 3: Results from the infield fielding time model. Harder hit balls are predicted to be quicker to field. Balls where the fielder has to travel at an angle to reach the ball take longer to field on average.

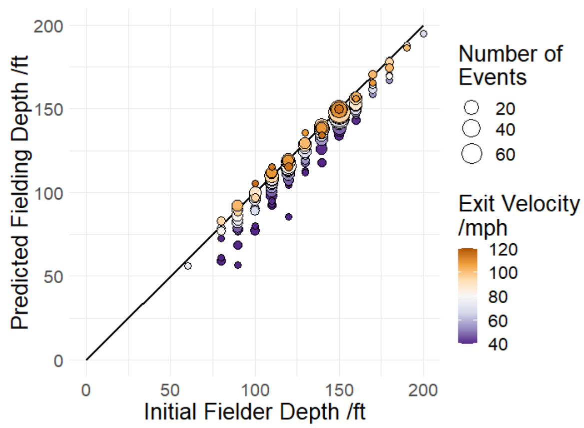


Figure 4: Results from the infield fielding depth model. Slowly hit balls are charged down and predicted to be fielded at a shallower depth than the initial fielder's position.

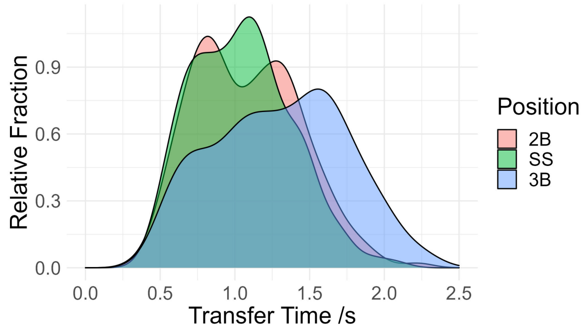


Figure 5: Density chart of the time between fielding and releasing the ball grouped by fielding position. Third basemen take slightly longer than the middle-infielders to release the ball.

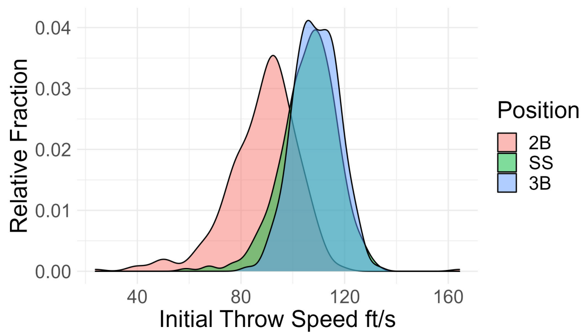


Figure 6: Density chart of initial throw speed from infielders. Third basemen throw the hardest, closely followed by shortstops, with second basemen having significantly weaker measured arm strength.

set themselves securely to make strong and accurate throws over a longer distance than the other positions.

Figure 6 shows the distribution of throw speeds from the three infield positions, the average second baseman has a weaker throwing arm than the average shortstop or third baseman.

A simple linear equation was used to convert throw speed and distance to throw time. Ideally a physical model including drag would be used but the accuracy and simplicity of this model is appropriate for this purpose.

$$t = -0.11 + 1.17\frac{d}{v} \tag{1}$$

Where t is throw time to first base, d is the distance from the throwing location to first base, and v is the initial throw speed.

The final step in modelling groundball outs is batter speed, and how home-to-first time along with the time of arrival of the throw at first base relates to the probability of an infield hit.

Home-to-first time was measured from the time when the ball is first put into play, to when the batter first reaches a distance of 90ft from home plate ((0,0) in the data coordinate system).

As there was no play-by-play data provided, hits were found by taking plays where a ball in play resulted in a change from first base unoccupied, to first base occupied. The time difference between the first baseman receiving the ball and the runner's home-to-first time was found and the best fit logistic curve was used to provide the probability of a hit based on the time difference between the two events as shown in Figure 7.

3 Batted Ball Profiles & BABIP Estimation

I have models for predicting the likelihood of an out on any ball in play, based on average fielder abilities in the minor-league data provided. However in order to simulate the efficiency of a fielding

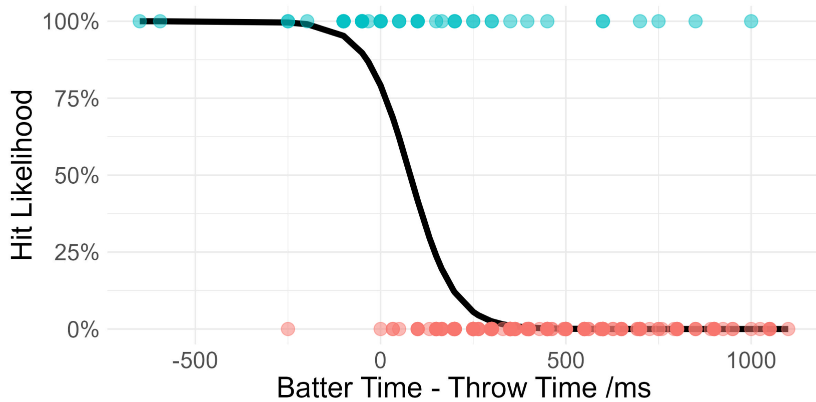


Figure 7: The best fit logistic curve for hit probability (black line) as a function of time difference between the batter arriving at first base and the throw being received by the first baseman. Individual hits/outs are shown as translucent dots in blue and red respectively.

alignment it is also necessary to simulate the distribution of ball-in-play trajectories that are likely to arise from the batter at the plate.

3.1 The Average Hitter Profile

The minor league ball tracking data provided by SMT proved too difficult to use for this purpose, the difficulty in estimating launch angle and exit velocity made it inappropriate. Instead, I used data from MLB's Statcast system which has a very large sample size and high accuracy. The batted ball data was tabulated in three dimensions: exit velocity, launch angle, and spray angle. A parameterised distributional approach was considered but the covariance between the variables meant that this would have become complex to construct.

Summary graphs of the batted ball data used are shown in Figures 8-10. These show the relationship between launch angle, spray angle, and exit velocity in the batted ball data used.

3.2 Adjusting To Different Hitter Tendencies

To make the simulation flexible to different styles of hitter, it is necessary to adjust this binned batted ball data. The quantities which I decided to adjust for are: average launch angle, average exit velocity, groundball & flyball pull percentages, handedness, and speed. Pull percentage in this work is defined as the fraction of batted balls hit anywhere to the pull side of second base. Groundballs and flyballs are defined as a launch angle of below and above 10 degrees respectively.

Modifying for handedness is easily done by inverting the spray angle bins for left handed hitters. When modifying GB pull%, FB pull%, and launch angle, this is done by shifting the occupation of each bin in each dimension and taking the shifted distribution which best fits the desired quantities. The average exit velocity can be matched exactly by adding a constant to the shifted distribution.

Batter home-to-first speed is input directly and assumed to be constant on all plays.

3.3 BABIP Estimation

The estimated **B**atting **A**verage on **B**alls **I**n **P**lay (**BABIP**) is the key output from the model. I am not considering the impact extra-base hits here, however this modification could be made by including a model for the likelihood of doubles and triples on a play. The models from section 2 are used to find the individual catch probability for every player on each possible ball in play type. These are then aggregated such that the probability of an out is defined as:

$$P(\text{Flyout}) = 1 - \prod_{i=3}^{9}(1 - P(\text{Catch}_i)) \tag{2}$$

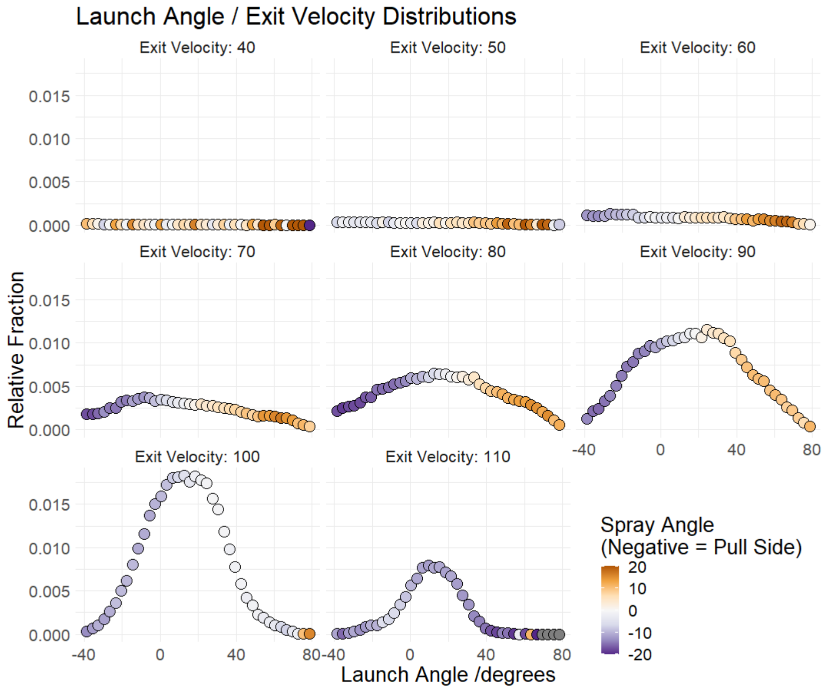


Figure 8: The distribution of launch angle (points), split by exit velocity in mph (individual panels), coloured by the average spray angle in each bin. Weakly-hit balls tend to be hit at launch and spray angles further from zero than hard-hit balls

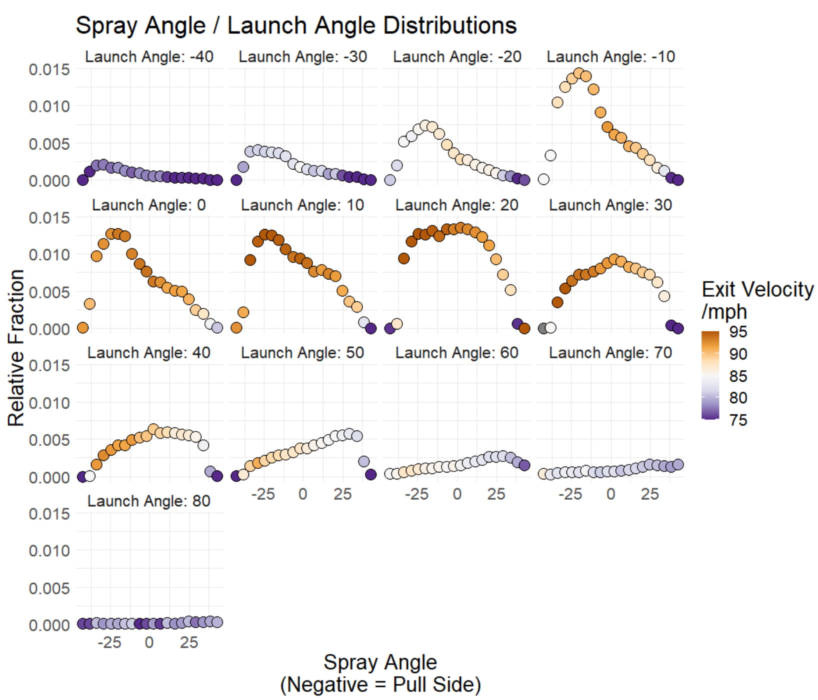


Figure 9: The distribution of spray angle (points), split by launch angle (individual panels), coloured by the average exit velocity in each bin. Groundballs tend to be pulled, while flyballs are sprayed to the opposite field. Hard hit balls tend to be pulled and hit at launch angles between -10 to 30 degrees.

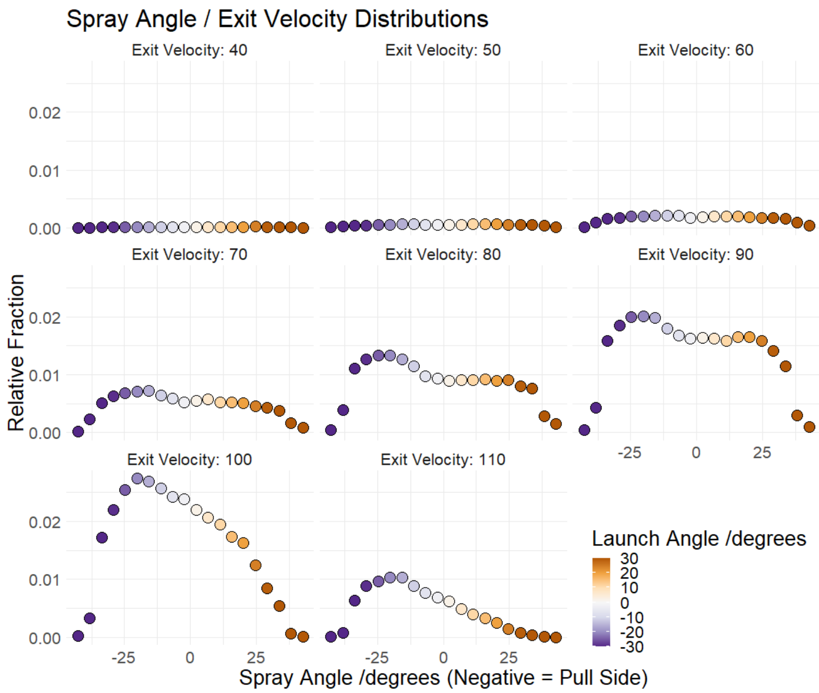


Figure 10: The distribution of spray angle (points), split by exit velocity in mph (individual panels), coloured by the average launch angle in each bin.

Where $P(\text{Catch}_i)$ is the probability of the fielder with position i catching the ball. i.e. the probability of a flyball hit is the probability of no-one catching the ball using independent probabilities.

The same approach can be used to calculate the likelihood of an out on a groundball. First I find the probability of each infield position fielding the ball. Then add up the timing components of the play (field, transfer, throw) and find the difference to the batter's home-to-first time. Apply this to the sigmoid function shown in Figure 7 to find the probability of a hit if each position fields it. These are then combined such that the probability of a groundball hit is the probability of every position failing to make the play independently.

$$P(\text{Field}_i) = \text{RangeModel}(\text{angle}_i, \text{speed}_i, ...) \tag{3}$$

$$T_i = \text{FieldTime}_i + \text{TransferTime}_i + \text{ThrowTime}_i \tag{4}$$

$$P(\text{Groundout}_i) = P(\text{Field}_i) \times (1 - \text{HitProbability}(T_i, \text{BatterTime})) \tag{5}$$

$$P(\text{Groundout}) = 1 - \prod_{i=3}^{6}(1 - P(\text{Groundout}_i)) \tag{6}$$

In the equations above, $P(\text{Field}_i)$ is the probability of each player reaching the ball, RangeModel is the model shown in Figure 2. T_i is the total time for each fielder to make the play if they field it, BatterTime is the batter's home-to-first time. HitProbability is a function for the logistic curve shown in Figure 7. $P(\text{Groundout}_i)$ is the probability of each fielder converting the out.

The total out probability is the sum of these two out probabilities across the entire batted ball sample, weighted by the expected frequency of each batted ball.

3.4 Simulating levels of fielder ability

A useful aspect of this model design is that it is possible to modify fielder ability in diverse ways and see how the best alignment changes.

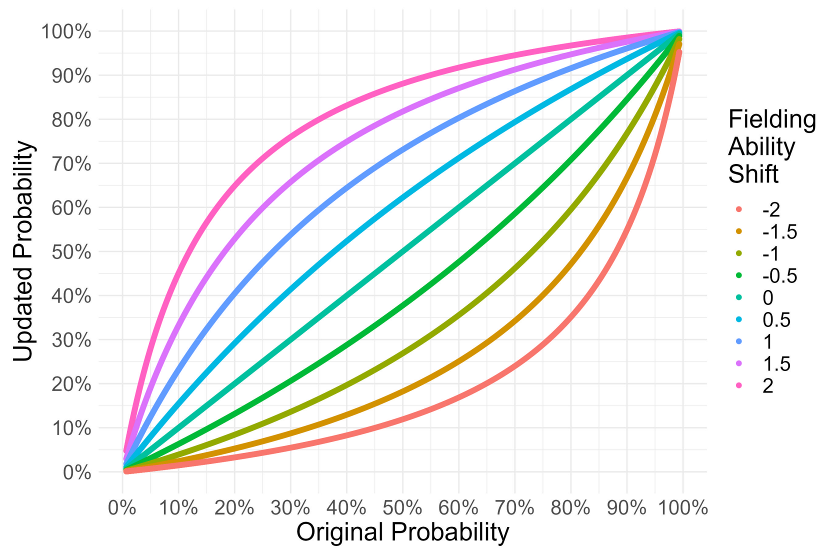


Figure 11: The effect of changing fielder ability on the probability of events.

Infield throw speed and transfer time are easily changeable by simply using a different number in the calculations.

To modify infield range and catch probability, a shift in the underlying logistic curve is applied to the probability of making the play. This ensures that impossible plays remain impossible, and certain plays remain certain, however the units are quite unintuitive. Examples of how fielding probability change with these ability ratings are shown in Figure 11. For example an outfielder with +2 range will make 50/50 catches at a 90% rate. Or an infielder with -1 range will reach balls at a 50% rate that the average infielder would reach 75% of the time.

3.5 Ad-hoc Modifications to Existing Systems

Several ad-hoc changes to the model were made to produce realistic results on infield plays as the original models allowed for too many infield hits or fielders playing closer than they could react to the ball.

A "cliff" in fielding ability was introduced as a function of initial depth, to reduce fielding ability when playing closer than standard alignments allow. This was needed because the effects of infielders playing very shallow weren't captured in model training data (because no one does it!). The effect of this is seen in Figure 12.

The output of the fielding time model was reduced by 20%, this allowed for more reasonable BABIP when testing the model and reduced the number of infield hits to a more realistic level.

Finally, infielders were given -1 catch ability to adjust for the fact that they have much less reaction time on line drives and so their range is more limited than outfielders.

4 Fitting Algorithm & Ideal Defensive Shifts

To find an ideal defensive alignment it is necessary to test many alignments then choose the one with the lowest BABIP. For my fitting method, I used an algorithm which shifted fielder positions in the direction which lowered BABIP, stepwise, until there was no further improvement and took this as the best alignment for the chosen batter/fielders. The first baseman was pinned to a location within 45ft of first base.

In the next sections the best defensive shift and expected BABIP, groundout% and flyout% are shown while varying a variety of input parameters. These are presented without significant comment for concision. Unless otherwise stated, the default parameters used are shown in Tables 1 & 2. When varying fielder ability the curve described in Figure 11 is used. For ability change a, infielder transfer

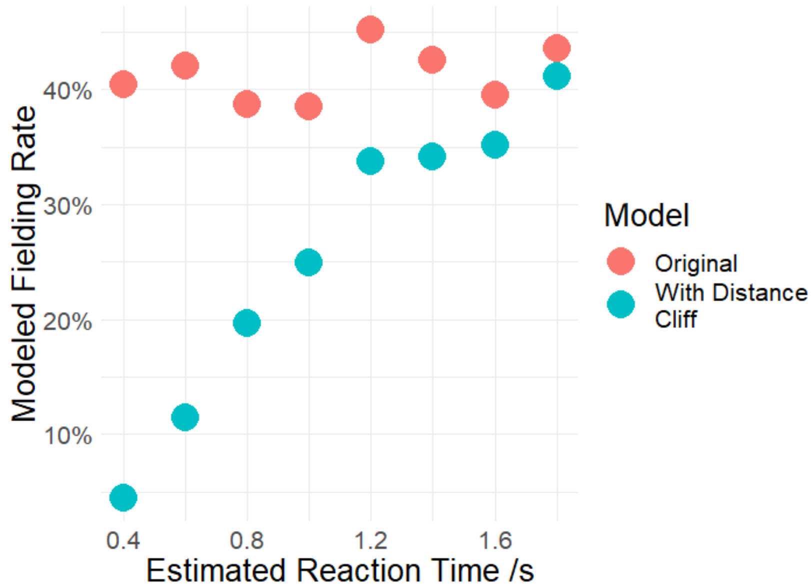


Figure 12: A reduction in fielding ability at shallow depths was introduced to limit effectiveness with very low reaction time. Effect on dummy data is shown here as no real infielders play extremely shallow.

times are also modified by $-0.2a$ seconds and throw speed is modified by $+10a$ ft/s from the default values.

The hope is that these tests can demonstrate the versatility of this method for placing fielders and estimating BABIP in diverse situations.

Average Exit Velocity	Average Launch Angle	GB Pull%	FB Pull%	Home-to-First Time
87 mph	12 degrees	70%	45%	4500ms

Table 1: Default Batter Parameters

Fielder	Catch Ability	Infield Range Ability	Transfer Time	Arm Strength
1B	-1	-0.5	1.1	80
2B	-1	0	0.9	90
3B	-1	0	1.1	110
SS	-1	0	0.9	105
LF	0	-	-	-
CF	0	-	-	-
RF	0	-	-	-

Table 2: Default Fielder Parameters

4.1 Batter Speed

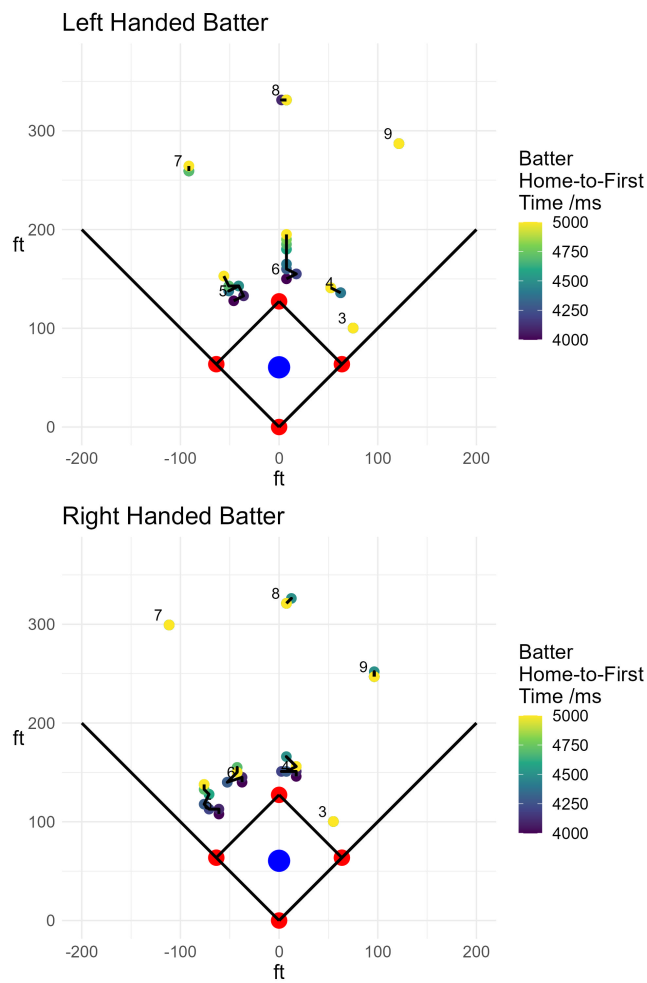


Figure 13: The best-fit fielding alignment as a function of batter speed and handedness. Infielders should play shallower against faster batters, especially third basemen.

Home-to-First Time /ms	BABIP(LHB)	BABIP(RHB)
4000	0.309	0.345
4100	0.296	0.330
4200	0.293	0.324
4300	0.298	0.315
4400	0.296	0.300
4500	0.289	0.293
4600	0.289	0.289
4700	0.288	0.287
4800	0.288	0.285
4900	0.287	0.283
5000	0.286	0.283

Table 3: Statistics using the best defensive alignment with different batter speeds. High speed is best for RHBs who hit most of their groundballs to the left side of the infield which results in longer throws and more infield hits.

4.2 Average Exit Velocity

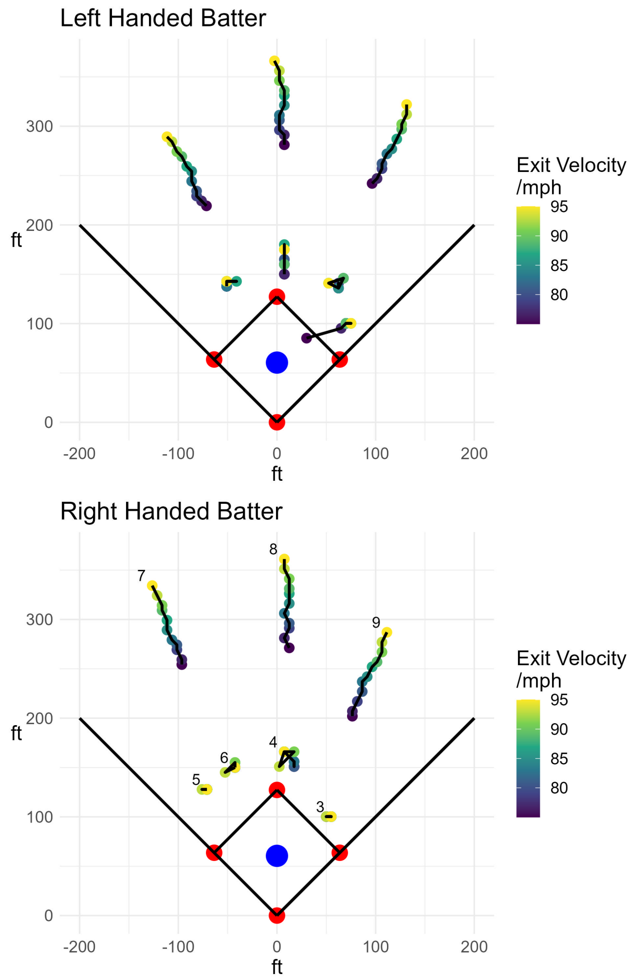


Figure 14: The best-fit fielding alignment as a function of batter exit velocity and handedness. Outfield depth is most strongly affected by how hard the batter hits the ball.

Average Exit Velocity /mph	BABIP(LHB)	BABIP(RHB)
75	0.216	0.225
77	0.222	0.229
79	0.234	0.240
81	0.249	0.253
83	0.264	0.268
85	0.280	0.283
87	0.289	0.293
89	0.306	0.308
91	0.320	0.322
93	0.334	0.336
95	0.347	0.344

Table 4: Statistics using the best defensive alignment with different batter exit velocities. High velocity is best to increase BABIP, as expected.

4.3 Batter Launch Angle

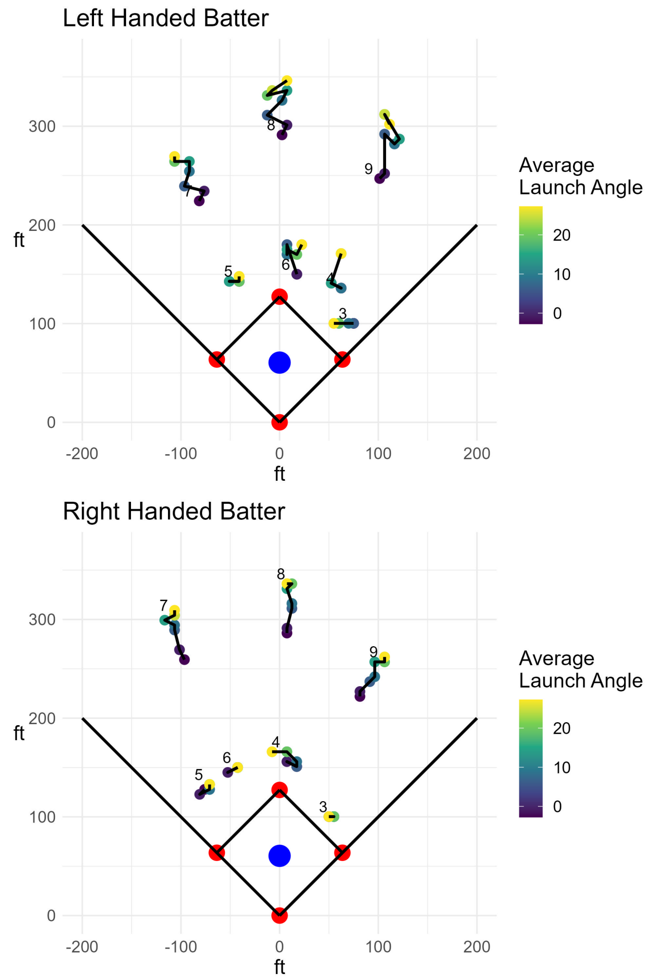


Figure 15: The best-fit fielding alignment as a function of batter launch angle and handedness. Outfielders play deeper against batters with higher average launch angles.

Average Launch Angle /degrees	BABIP(LHB)	BABIP(RHB)
-3	0.290	0.313
-1	0.293	0.319
6	0.302	0.307
9	0.299	0.300
14	0.276	0.281
20	0.245	0.255
25	0.212	0.223
27	0.195	0.206

Table 5: Statistics using the best defensive alignment with different batter launch angles. Extreme flyball hitters have a significant disadvantage in predicted BABIP.

4.4 Batter Pull Tendency

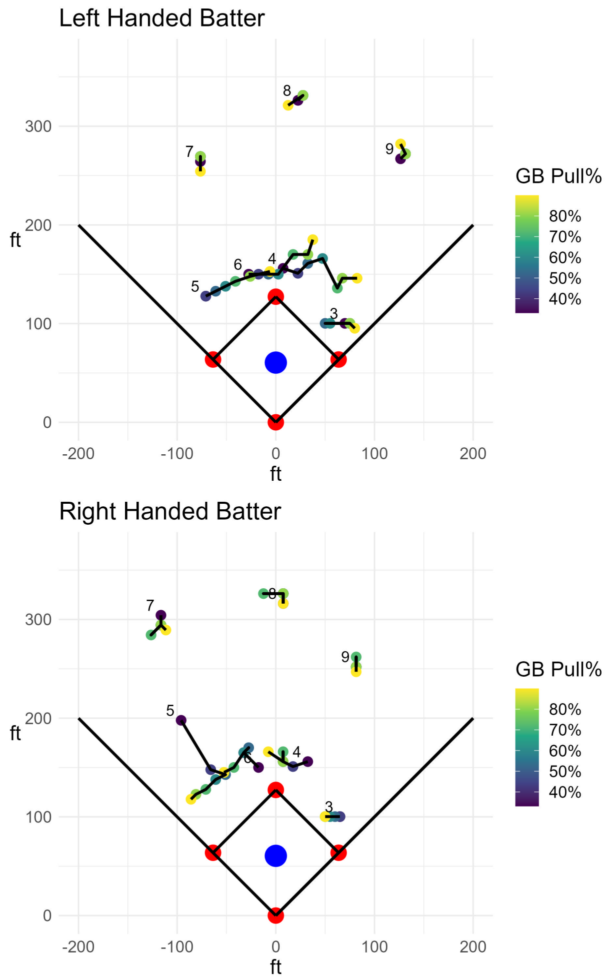


Figure 16: The best-fit fielding alignment as a function of batter groundball pull% and handedness. The benefits of strongly shifting the infield to the pull side can be clearly seen, with no-one occupying the left side of the infield for the strongest LHB groundball pull%

Groundball Pull% /degrees	BABIP(LHB)	BABIP(RHB)
0.33	0.285	0.295
0.44	0.289	0.301
0.53	0.283	0.290
0.60	0.282	0.287
0.72	0.286	0.292
0.80	0.270	0.292
0.90	0.251	0.281

Table 6: Statistics using the best defensive alignment with different batter groundball pull percentages. High pull% LHBs see the largest decrease in BABIP as infielders can flood the right side of the infield in the best alignment.

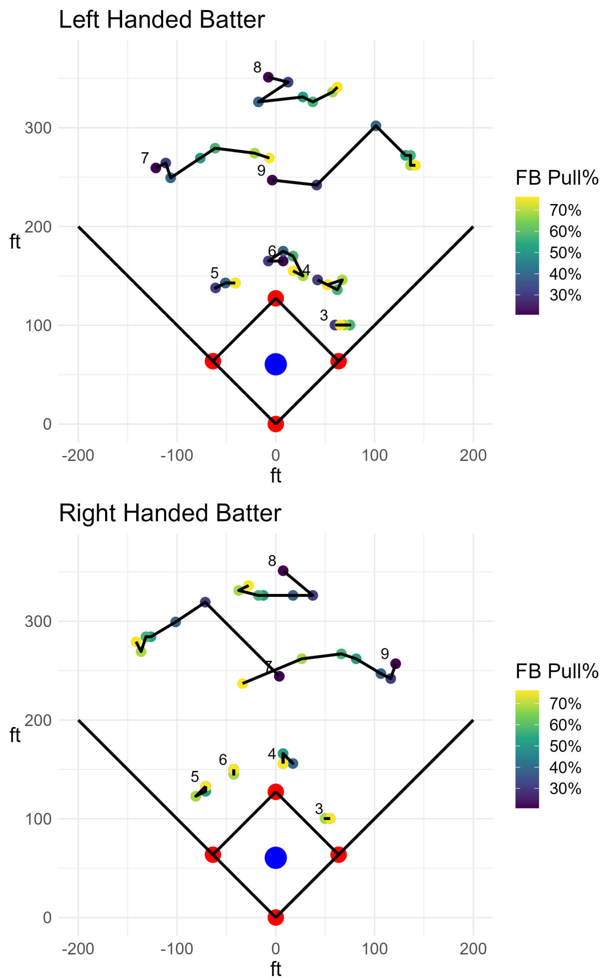


Figure 17: The best-fit fielding alignment as a function of batter flyball pull% and handedness.

Flyball Pull% /degrees	BABIP(LHB)	BABIP(RHB)
0.21	0.265	0.270
0.31	0.285	0.282
0.39	0.290	0.292
0.52	0.286	0.292
0.57	0.282	0.294
0.69	0.273	0.288
0.76	0.266	0.268

Table 7: Statistics using the best defensive alignment with different batter flyball pull percentages. BABIP is generally maximised with pull% around 50% as this forces the outfielders to spread out the most.

4.5 Fielder Ability

4.5.1 Center Fielder

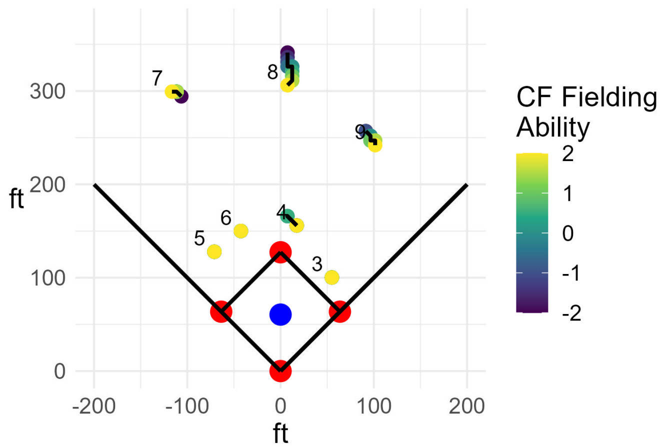


Figure 18: The best-fit fielding alignment as a function of center fielder ability. As CF range increases, they can play shallower and the corner outfielders are shifted towards the foul lines.

CF Ability	BABIP(RHB)
-2.0	0.304
-1.5	0.301
-1.0	0.298
-0.5	0.295
0.0	0.293
0.5	0.290
1.0	0.286
1.5	0.283
2.0	0.279

Table 8: Statistics using the best defensive alignment with different center fielder ability. .

4.5.2 Second Baseman

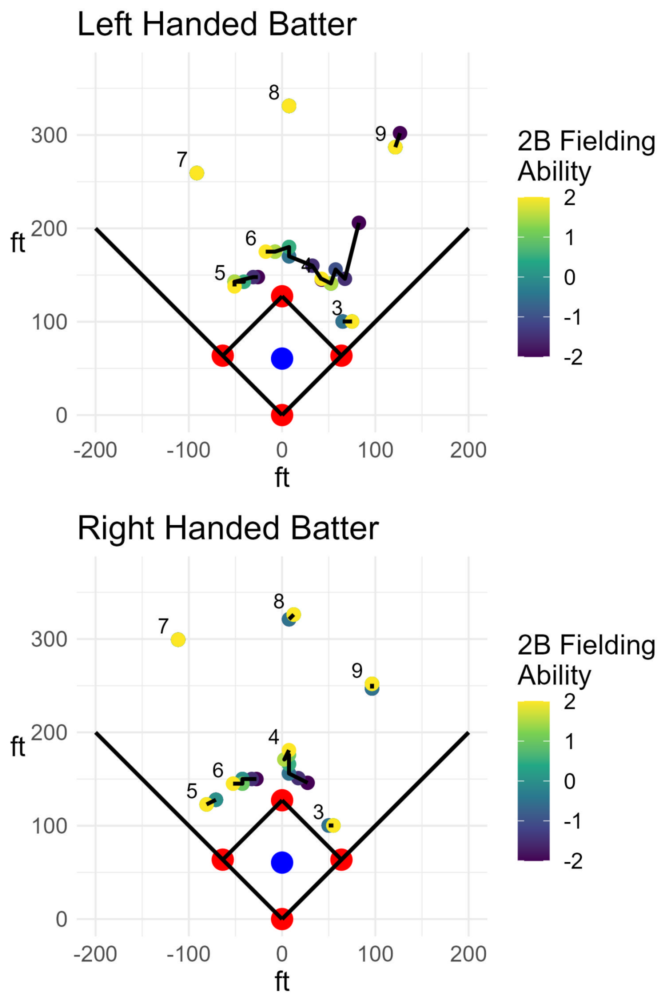


Figure 19: The best-fit fielding alignment as a function of second baseman ability and batter handedness.

Second Baseman Ability Change	BABIP(LHB)	BABIP(RHB)
-2.0	0.329	0.337
-1.5	0.319	0.327
-1.0	0.327	0.320
-0.5	0.307	0.305
0.0	0.289	0.293
0.5	0.269	0.284
1.0	0.256	0.271
1.5	0.237	0.255
2.0	0.231	0.243

Table 9: Statistics using the best defensive alignment with different second baseman ability.

4.5.3 Third Baseman

Figure 20: The best-fit fielding alignment as a function of third baseman ability and batter handedness.

Third Baseman Ability Change	BABIP(LHB)	BABIP(RHB)
-2.0	0.338	0.353
-1.5	0.332	0.344
-1.0	0.326	0.331
-0.5	0.317	0.317
0.0	0.289	0.293
0.5	0.273	0.274
1.0	0.262	0.255
1.5	0.252	0.236
2.0	0.240	0.220

Table 10: Statistics using the best defensive alignment with different third baseman ability.

4.5.4 Shortstop

Figure 21: The best-fit fielding alignment as a function of shortstop ability and batter handedness.

Shortstop Ability Change	BABIP(LHB)	BABIP(RHB)
-2.0	0.330	0.339
-1.5	0.329	0.339
-1.0	0.323	0.326
-0.5	0.313	0.309
0.0	0.289	0.293
0.5	0.275	0.276
1.0	0.257	0.264
1.5	0.238	0.251
2.0	0.231	0.236

Table 11: Statistics using the best defensive alignment with different shortstop ability.

4.6 Shiny App

Any custom prediction can be generated using a shiny app which is deployed at the following URL: https://pitching.shinyapps.io/ShiftTester/.

To comply with the NDA, this website should not be shared beyond the Data Challenge.

5 Limitations & Improvements

In this section I shall shortly discuss some limitations to this model, and how improvements could address these issues.

The first issue is that the batted ball data and the training data for the fielding models come from separate sources. For consistency, these would ideally come from the same source and be treated identically to reduce any systematic errors. Improved accuracy of my launch angle and exit velocity measurements would also increase the power of the infield fielding models.

Another issue is that the impact of foul balls, pitcher/catcher fielding, outfield fences, and bunting are not considered. Each of these factors could shift how the defense should be aligned. Also, the value of extra base hits has not been implemented and so the model currently sees preventing a double as the same value as preventing a single, but the former is more valuable for saving runs.

Finally, the sampling method to estimate probabilities of different balls in play does not contain samples outside the foul lines. This means that when shifting the pull tendency of the batter, some extreme spray angles can be unsampled and given zero weight which is unrealistic.

There is potential to expand the scope of the algorithm outputs beyond BABIP. The likelihood of any fielding event could be generated, and its dependence on the input parameters can be estimated. For example one could look at the predicted rate of infield hits to the third baseman being reduced with a stronger arm and quicker transfers, or flyout rate to right field as a function of batter launch angle.

6 Conclusions

I have developed a general defensive positioning algorithm for baseball. This algorithm is flexible to varying batter styles, with regards to handedness, exit velocity, launch angle, spray angle and speed. In addition, fielder abilities such as range, transfer speed and arm strength can be varied.

The spatio-temporal nature of this problem means that ball and player tracking time-series data was essential to build the relevant models. Modelling of catch probability, and infield plays has been described in detail. The relevant code is shared on Github.

A fitting method was implemented to derive the best-fit defensive alignment for any chosen input parameters. The variation in defensive positioning and expected BABIP has been shown while varying many of the inputs. Model predictions and best-fit positioning can be explored on an app online.

Public analysis of defensive shift effectiveness has so far been limited to simplistic analysis of outcomes under the binary change of shift-on or shift-off[4]. This work is a first step to model shift effectiveness and defensive positioning in a more granular way, with many potential applications for both teams and analysts.

References

[1] E. Nash, "From 4th Outfielder to Shortstop." https://perma.cc/TFW6-BLQ9,https://bosoxinjection.com/2013/11/15/history-of-ss/.

[2] J. Posnanski, "The Boudreau Shift." https://perma.cc/84WY-A3D6,https://joeposnanski.substack.com/p/the-boudreau-shift. Accessed: 2022-07-01.

[3] "Statcast Team Positioning Leaderboard." https://perma.cc/QV4P-9DVC,https://baseballsavant.mlb.com/visuals/team-positioning.

[4] "Examining shift effectiveness with batted ball data." https://perma.cc/4N6D-8NKP,https://baseballcloud.blog/2020/10/02/examining-shift-effectiveness-with-batted-ball-data-part-1/.