A BOTTOM-UP APPROACH TO EVALUATING INFIELDER RANGE ON GROUND BALLS

A project report for the

2023 SMT Data Challenge

By

ANONYMIZED

August 2023

31 August, 2023

Anonymized

ABSTRACT

The most important aspect of an infielder's defensive play is their ability to make outs on ground balls. I have created a method to capture this ability (which I simply call their "range") through the meticulous deconstruction of ground ball outs into three discrete phases: (1) contact-to-glove, (2) transfer and (3) throw, along with a method to estimate whether an infielder can reach a ground ball. My approach is interpretable and effective with clear applications in the evaluation of individual defensive ability. However, the most useful and unique application is for optimally positioning infielders based on the interaction of infielders' ranges and batter tendencies.

Github Link

ACKNOWLEDGEMENTS

Thank you to Dr. Meredith Wills, Dr. Mohammad Reza Rajati, and Dr. Ron Yurko for their help with this project.

August 31, 2023

TABLE OF CONTENTS

CHAPTER

CHAPTER 1

Introduction

1.1 Background

An infielder's primary defensive responsibility is to record outs on ground balls. While sounding simple, the defender's task can be surprisingly tricky. The most difficult of these plays are some of the greatest displays of athleticism on a baseball diamond, like in the play by Dansby Swanson shown in Figures 1.1a, 1.1b, 1.1c below. Swanson displays great raw speed to chase down the ball, cleanly fields and swiftly and fluidly transfers the ball to his throwing hand, and generates immense torque to make a strong throw from an awkward position. It's these plays that remind naysayers that, indeed, baseball players are great athletes too.

(a) Dansby Swanson fields a Wil Myers ground ball

(b) Dansby Swanson makes an on off-platform throw on the same play

(c) Wil Myers is out at first base, saving a Max Fried CG shutout

Figure 1.1: MLB Film Room Link to Dansby Swanson play, 9/24/2021

1.2 Motivation

For a majority of baseball history, teams positioned their infielders uniformly across the league. More recently, using large amounts of data on batter tendencies like in Figure 1.2, teams "shifted" their infielders in nontraditional alignments to great effect. This took away many brilliant plays, like the one above, as ground balls were increasingly hit right to where defenders were standing at the start of the play. However, teams became so adept at this that MLB placed restrictions on how team's could shift last year - it was simply too free of a square.

Motivated by the clear advantage in having optimally positioned defenders and MLB's new restrictions on shifting, I wanted to explore how this strategy could be improved even more. For example, without the ability to perfectly place three defenders on one side the infield against a pull hitter, how can two defenders on each side of the infield be used to their maximum effectiveness? Figure 1.2 conveniently shows a legal alignment in 2023 and begs the question: is it *really* optimal for Trevor Story and Nolan Arenado to be standing so close to the exact halfway point of their "zone"? Might each player's

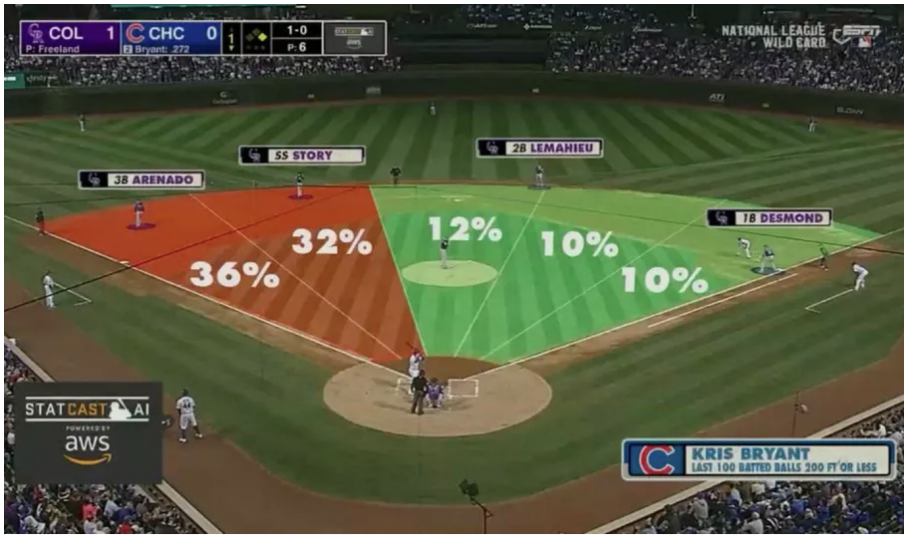


Figure 1.2: Rockies shifted against Kris Bryant overlaid by his batted ball spray data

individual unique ability inform an even better way to position them? In other words, can this free square become even more free?

1.3 Problem Statement

As fans, we have little understanding of exactly how front offices are making defensive positioning decisions. While it's clear that teams care about where hitters tend to hit the ball, this is a purely reactionary strategy. This project injects proactivity into this decision-making process by considering each defender's ability in three categories: (1) movement, (2) transfer time, and (3) arm strength.

For each of these categories, I start from two simple assumptions.

1. **If a player *has* done something then they *can* do it again.**
2. **A player's defensive ability is defined by the limit of what they *can* do.**

Therefore, I only look at the best of what players have ever done in each category.

Returning to the idea of ground ball plays as a race between defenders and batters, I needed to measure the time it takes a defender to get the ball to first base compared to the time it takes a runner to reach first base. To do this, I deconstruct ground balls into discrete phases: (1) contact-to-glove, (2) transfer, and (3) throw. To determine whether a play results in an out or safe call, I use a piecewise function that compares the sum of times for each phase on defense and a runner's time to first base.

$$f(x) = \begin{cases} 1 & \text{if } T_{cg} + T_{tt} + T_t < T_R \\ 0 & \text{otherwise} \end{cases}$$

$$T_{cg} = \begin{cases} T_{bp} & \text{if } T_{f90} <= T_{bp} \\ \infty & \text{otherwise} \end{cases}$$

Variable	Meaning
T_{cg}	Time from contact to defender's glove (infinity if cannot be fielded)
T_{tt}	Min transfer time for each defender
T_t	Time for throw to reach first baseman given defender's hardest throw speed
T_R	Time for runner to reach first base
T_{bp}	Time for ball to reach x feet away from home plate
T_{f90}	Time for defender to reach points 1-90 feet away from them

Table 1.1: Variable notation

Throughout the project, I use player's minimum transfer times and models I built for throw time and contact-to-glove time. I determine whether a defender can reach a ground ball by considering a defender's maximum speed and fastest times to travel certain distances. If the player would not be able to field the ground ball, $T_{cg} = \infty$ and f(x) = 0.

1.4 Research Aims and Applications

The envisioned application of this project is creating even better defensive alignments. In the Rockies example above, could Nolan Arenado's great hands and arm strength justify playing him and Trevor Story a step or two more towards second base? This would allow Story, who was in the 6th percentile in Outs Above Average in 2021, to field more balls up the middle. Or on the brilliant play that Dansby Swanson made, could Austin Riley also have been playing slightly farther from third base? This would turn a extremely difficult play into a routine play for Riley and, over the course of a season, turn several base hits into outs.

CHAPTER 2

METHODS

All data in this section comes from a 97 game sample of minor league games provided by SMT. My analysis is limited to plays within this dataset where a ground ball was hit to an infielder (2B, SS, 3B) who made a throw to first base. However, for the T_{bp} model, I use all ground ball data because this quantity is defender-agnostic.

2.1 Transfer Time, T_{tt}

Transfer time refers to the amount of time between a defender acquiring and throwing the ball which are labeled events in the data. Using the labeled events, I calculated each player's minimum transfer time. Table 2.1 shows the ten players with the lowest transfer times.

Player ID	Position	Transfer Time (ms)
9880	2B	350
9087	2B	350
3726	SS	400
6761	SS	450
1650	SS	450
1748	SS	450
6327	SS	500
1650	2B	500
6528	2B	500
1628	2B	500

Table 2.1: Ten fastest transfer times by player

Encouragingly, the ten fastest transfer times all came from middle infielders and not third basemen which reflects a real selection bias in professional baseball where only the most skilled infielders play the middle infield.

2.2 Throw Time, T_t

To quantify the time it would take for a defender's throw to reach the first baseman, I constructed a model using two features—distance (D_{throw}) and initial instantaneous speed (S_i) of the throw. Figures 2.1 shows the relationships and distributions of the data.

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

c

$$S_i = \frac{\Delta P}{\Delta T} = \frac{\sqrt{(x_{release+2} - x_{release+1})^2 + (y_{release+2} - y_{release+1})^2}}{t_{release+2} - t_{release+1}}$$

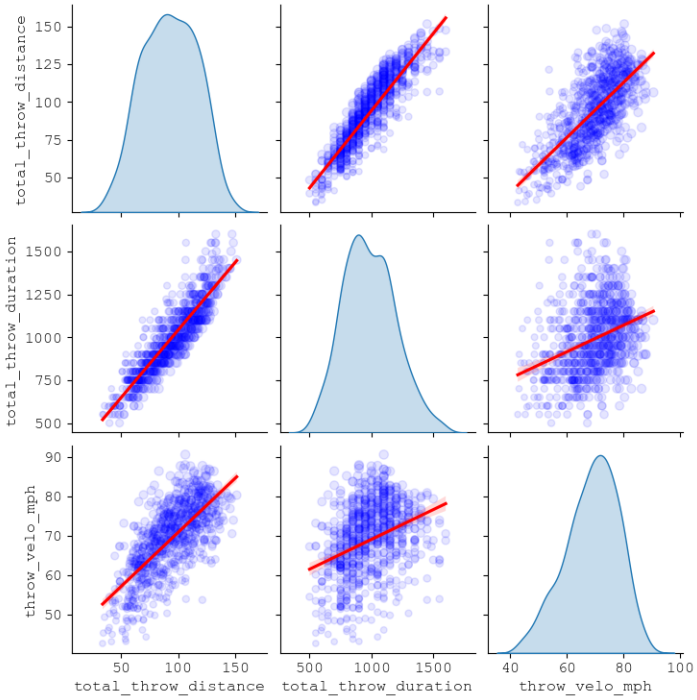
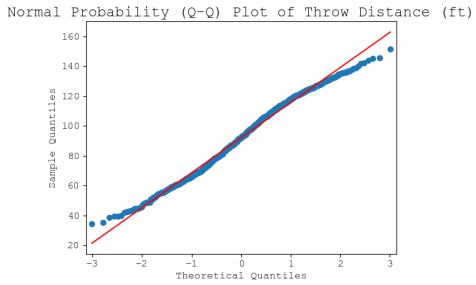
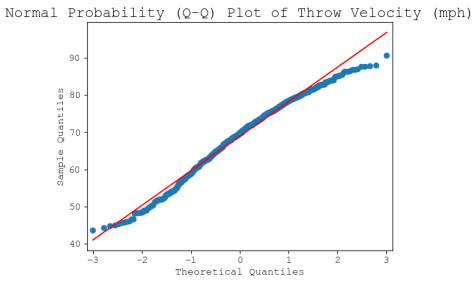


Figure 2.1: Pairplot shows bell-shaped curves and strong linear relationships for all data. The relationship between initial velocity and throw time is weaker and increases slower than distance vs. throw time. There is also collinearity between the predictors.



(a) Throw distance data closely follow the standardized diagonal line



(b) Throw velocity data closely follow the standardized diagonal line

Figure 2.2: Q-Q plots show an approximately normal distribution of the features

Based on approximately normal distributions evidenced in Figures 2.2a and 2.2b and linear relationships of the predictors and response, linear regression was a plausible modeling choice. However, a defender may have good reason to throw a ball harder or softer from the same spot depending on the runner's speed or how hard the ball was hit to them. At the same time, farther throws generally require harder throws. These complexities motivated the use of k-nearest neighbors (KNN), and bagged trees which might capture these effects. The prediction accuracies for each model are reported in Table 2.2.

Model Type	Mean absolute error (ms)
Linear Regression	23.267
K-Nearest Neighbors	23.891
Bagged Trees	24.817

Table 2.2: Mean absolute errors for models. MAE is the average difference between the model's predictions and true values. For KNN, k = 2 with distance weighting using 5-fold cross-validation. For bagged trees, # trees = 100 using 10-fold cross-validation. n = 778.

Ultimately, I chose the linear regression model which had the best out-of-sample accuracy (predictions show in Figure 2.3), robust predictions for outliers (a weakness of the KNN model), and resulted in an interpretable (a weakness of bagged tress) equation: $T_t = 11.467 \cdot D_{throw} - 13.267 \cdot S_i + 833.656$

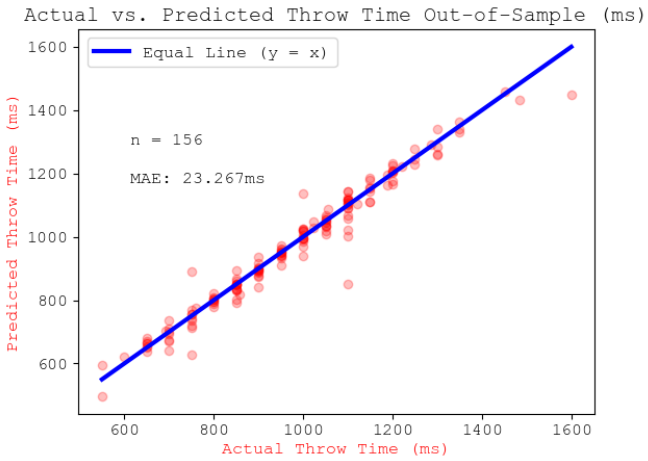


Figure 2.3: Predicted vs. Actual Throw Times (Linear Regression)

2.3 Contact to Glove Time, T_{cg}

Determining whether a player can reach a ground primarily depends on fielder speed, ball speed, and how far the ball is hit from their initial position.

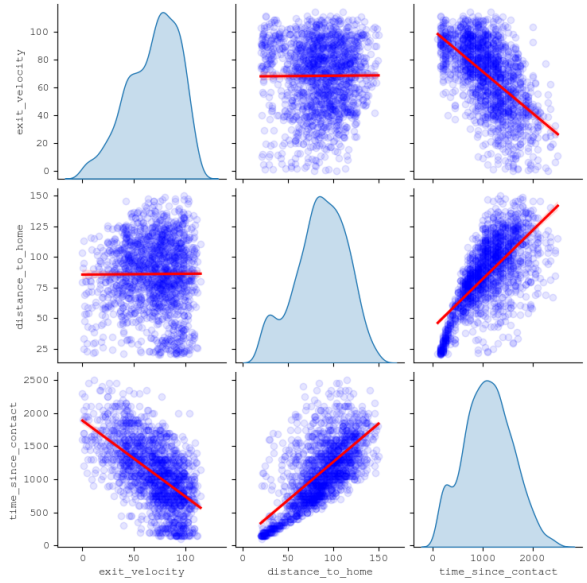
2.3.1 Ball to Point Time, T_{bp}

I used two features-distance from home plate (D_{home}) and initial instantaneous exit velocity (EV_{mph})-defined similarly to above. To supplement the training data, I also added data from the first three bounces of any ground ball.

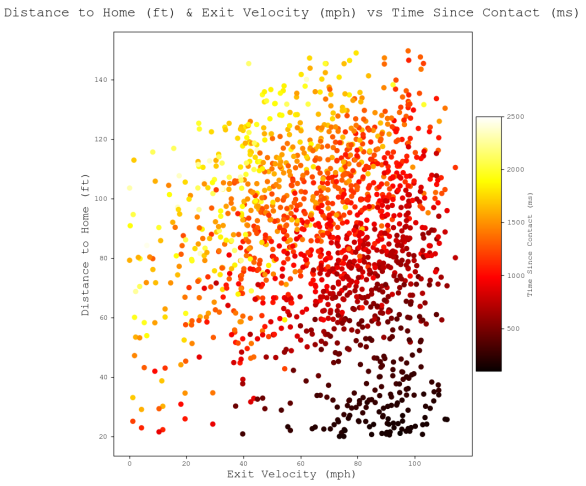
$$D_{home} = \sqrt{(x_2)^2 + (y_2)^2}$$

$$EV_{mph} = \frac{\Delta P}{\Delta T} = \frac{\sqrt{(x_{contact+2} - x_{contact+1})^2 + (y_{contact+2} - y_{contact+1})^2}}{t_{contact+2} - t_{contact+1}}$$

From the plots in Figures 2.4b and 2.4a we see intuitive relationships where balls farther from home took longer to reach that point, and harder hit balls take less time to reach any distance.



(a) Pairplot of (top to bottom, left to right): exit velocity, distance to home, time since contact. No collinearity in the features with strong linear relationships by themselves.



(b) All three variables on one plot

Based on the above plots, I used linear regression to model the relationship between the predictors and response. This model achieved an MAE of 154 milliseconds and the equation of the regression line was: $T_{bp} = 11.63 \cdot D_{home} + -11.61 \cdot EV_{mph} + 892.94$.

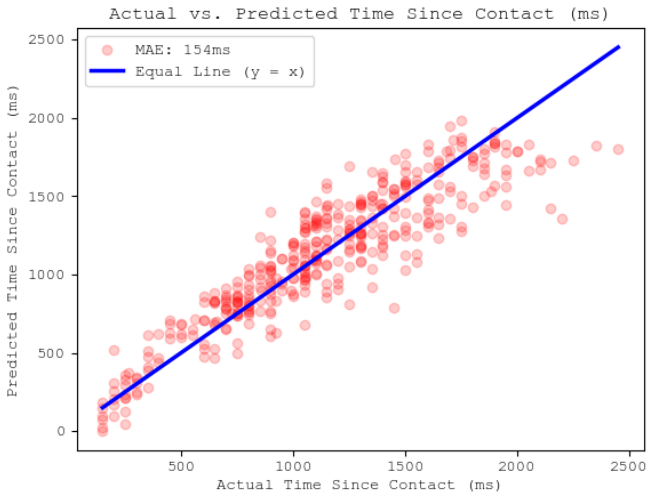


Figure 2.5: Predicted vs. Actual Ball Times

With a model with good predictions on T_{bp}, I needed a method to determine whether a player could beat a ball to a spot on the field.

2.3.2 Fielder Run Time, T_{f90}

My first approach only considered movement on defensive plays which are most directly related to my analysis. On these plays, I measured a player's max speed, the distance they traveled, and the time it took them to reach the ball. Then I determined whether they could reach a ground ball based on the fastest time anyone with a lower maximum speed reached a similarly hit ball. This approach failed due to a lack of data as shown in Figure 2.6.

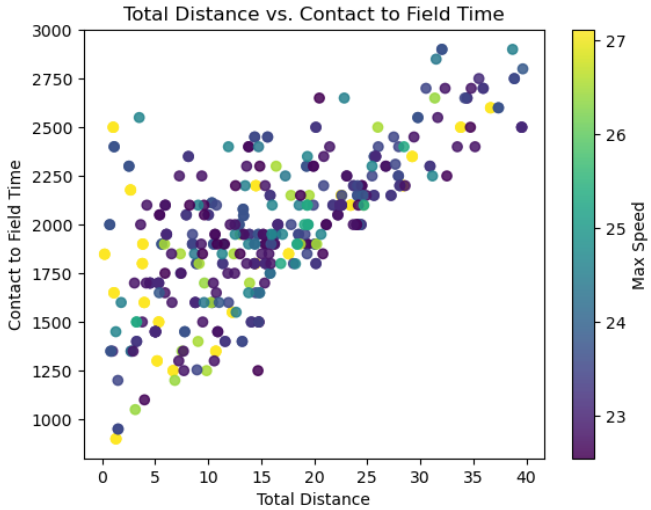


Figure 2.6: Many (slow) purple dots are the fastest to reach a ball hit any distance away from a player's initial position

My second approach looked at plays where a player was the batter. I measured a player's maximum speed and the fastest times they could travel between 1-90 feet. Using maximum speeds, I could differentiate between players who could reach far away ground balls better or worse as seen in Figures 2.7 and 2.8.

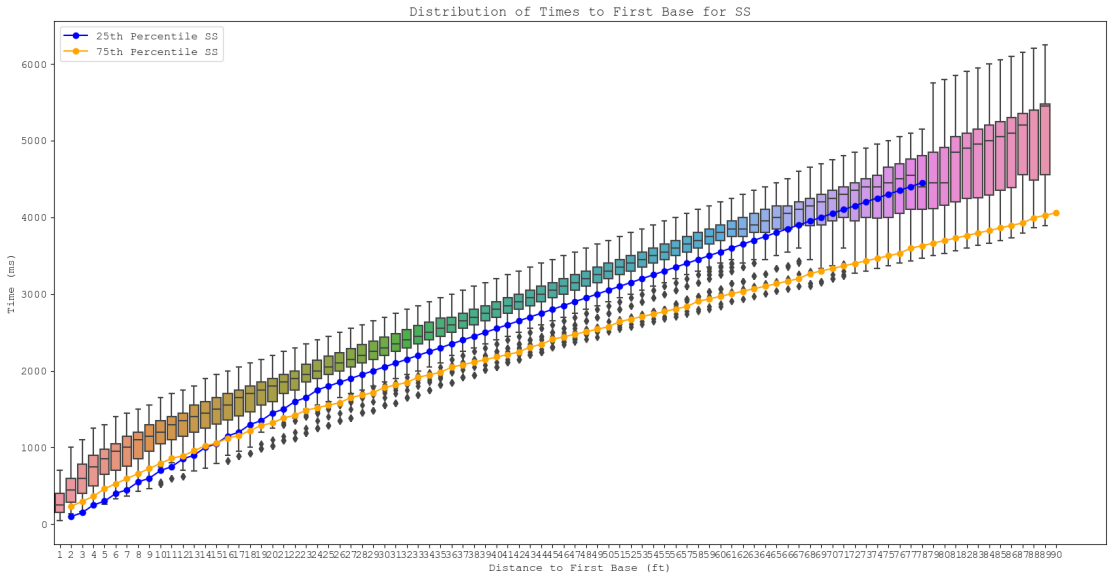


Figure 2.7: SS 25th Percentile (blue) vs. 75th Percentile (orange) Max Speed

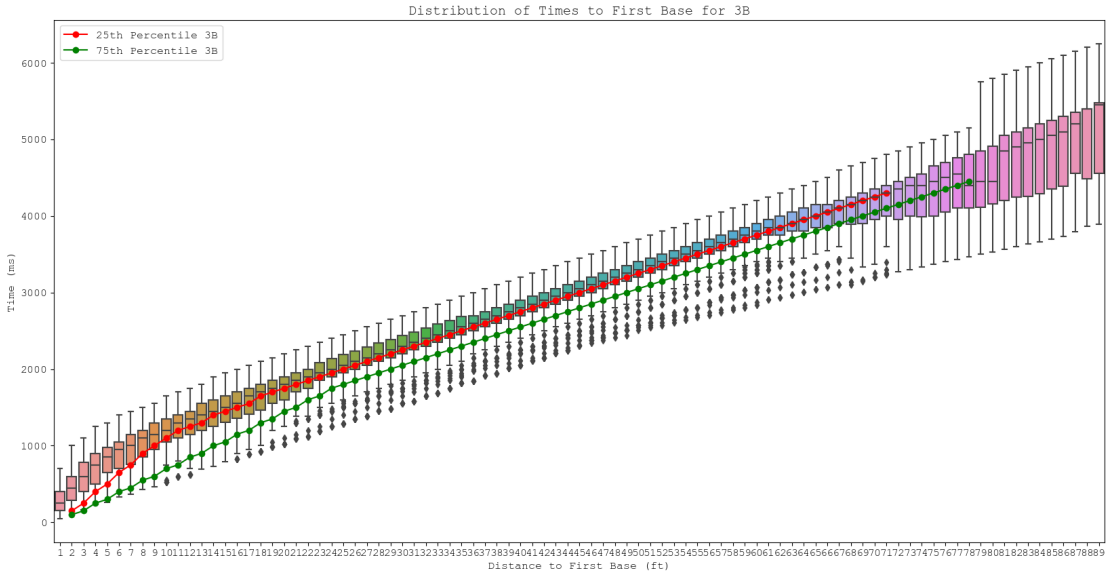


Figure 2.8: 3B 25th Percentile (red) vs. 75th Percentile (green) Max Speed

2.4 Infielder Range, $T_{cg} + T_{tt} + T_t$

With data on players' minimum transfer and run times and models capable of predicting ground ball and throw times, I could finally combine each factor to make a determination of infielder range.

2.4.1 Play Context

To highlight the differences in range and optimal alignments, I used Patrick Wisdom's 2022 batted ball data (Table 2.3). As a pull hitter, defenses are unlikely to play a standard alignment. In addition, I used the average MLB home to first time, 4474 milliseconds, for T_R.

Exit Velocity (mph)	Pull%	Straight%	Oppo%
91.2	48.1	32.2	19.7

Table 2.3: Patrick Wisdom Statcast Data, 2022. Note: 81st percentile AvgEV = 91.2 mph in the ground ball data.

2.4.2 Constructing Good and Bad Defenders

I constructed good and bad defenders, shown in Table 2.5, based on approximately the 25th and 75th percentiles for each position, which are shown in Table 2.4.

Metric	2B	3B	SS
Max Throw Velo (mph)			
	Mean: 69.263	Mean: 77.808	Mean: 79.822
	25%: 65.975	25%: 73.805	25%: 75.667
	75%: 73.936	75%: 81.773	75%: 83.007
Min Transfer Time (ms)			
	Mean: 791.353	Mean: 830.759	Mean: 752.861
	25%: 625.000	25%: 650.000	25%: 600.000
	75%: 900.000	75%: 1000.000	75%: 850.000
Max Speed (ft/s)			
	Mean: 22.051	Mean: 23.764	Mean: 23.042
	25%: 19.100	25%: 21.585	25%: 20.440
	75%: 24.383	75%: 26.289	75%: 25.515

Table 2.4: Summary Table by Position. Because the data didn't span enough games for every player to showcase their best ability, the above values are not realistic. For example, the mean *maximum* sprint speeds values would be considered "poor" according to 2023 Statcast sprint speed leaderboards

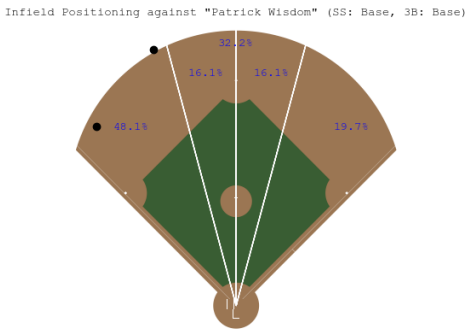
Because I had access to raw sprint speeds from Statcast, I used this to find players with close to MLB 25th and 75th percentiles and thus I also used those player's T_{f90}'s.

Label	Transfer Time (ms)	Throw Velocity (mph)	Sprint Speed (ft/s)
"Bad" SS	850	75.667	27.3
"Bad" 3B	1000	73.805	26.2
"Good" SS	600	83.007	28.6
"Good" 3B	650	81.773	28.0

Table 2.5: "Good" and "Bad" defenders based on data

2.4.3 Range Calculation

To determine a defender's range, I set an initial position based on where teams played their shortstop and third baseman against Patrick Wisdom (Figures 2.9a and 2.9b).



(a) My intial positions based on how defenses played Patrick Wisdom in 2022. This was determined using the angles relative to the bases and counting pixels for distances.

(b) Actual alignments against Patrick Wisdom from Statcast

Figure 2.9: Observed vs Modeled alignments against Patrick Wisdom

Then, I moved away from first base, on an axis perpendicular to home plate, one foot at a time from the initial position and input the distance from home and Wisdom's exit velocity (91.2 mph) to my T_{bp} model to make a ball time prediction. If the defender's T_{f90} at that distance from their initial position was lower than T_{bp}, i.e. they could field the ball, I used their distance to first and throw velocity in the T_t model to predict the throw time. Then I used their transfer time and the piecewise function from section 1.3 to determine if they could make the out.

I move perpendicularly away from first base for two reasons. Firstly, ground balls to a player's right are harder to make an out on than to the left. Secondly, a player only charges a ball if they *know* they can field it. Therefore, to examine the limit of a player's range, perpendicular makes sense as it's approximately the minimum angle a player would take to field a ground ball they're unsure they can field. For context, this is a 92 mph ground ball where Trea Turner takes an approximately perpendicular route and requires a slide to field.

CHAPTER 3

RESULTS AND DISCUSSIONS

3.1 Applications

The results of the above process are shown in Figures 3.1 and 3.2.

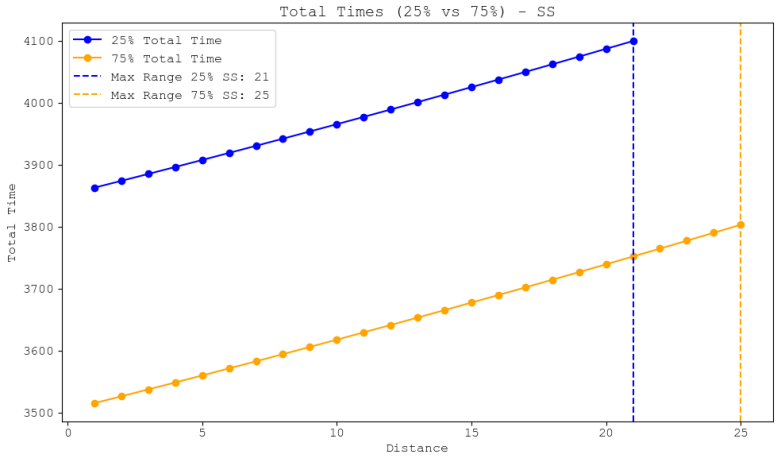


Figure 3.1: Based on transfer time, arm strength, and speed, a 75th percentile SS can make outs on 91.2 mph ground balls 4 feet farther to his right than a 25th percentile SS. The lines are perfectly linear because throw and ball time are being predicted using linear regression models and transfer time is constant. A good shortstop makes a play on a 91.2 mph ground ball around 350 milliseconds faster than a bad one.

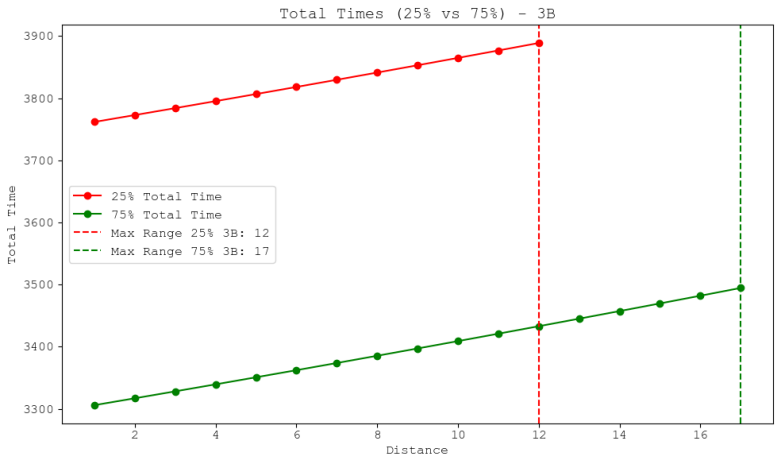


Figure 3.2: A 75th percentile 3B can make outs on 91.2 mph ground balls 5 feet farther to his right than a 25th percentile 3B. A good third baseman makes a play on a 91.2 mph ground ball around 450 milliseconds faster than a bad one.

Based on this result, teams with a 75th percentile shortstop can position him 4 feet farther up the middle against a hitter like Patrick Wisdom. Additionally, a good third baseman can be 5 feet farther off the line (a big difference on the original Austin Riley, Dansby Swanson play!).

However the most impactful finding is the interaction between defenders that occurs when you have different combinations of good and bad shortstops and third baseman. In section 2.4.3, I showed the base alignment used against Patrick Wisdom in 2022. Using those initial positions and the max ranges from above, Figures 3.3, 3.4, 3.5, 3.7, and **??** are the ranges of different SS, 3B combinations.

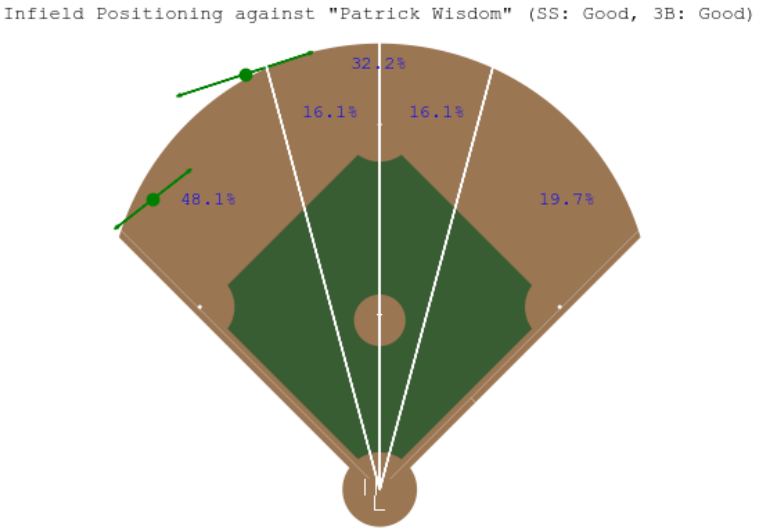


Figure 3.3: SS Range: 25ft, 3B Range: 17ft

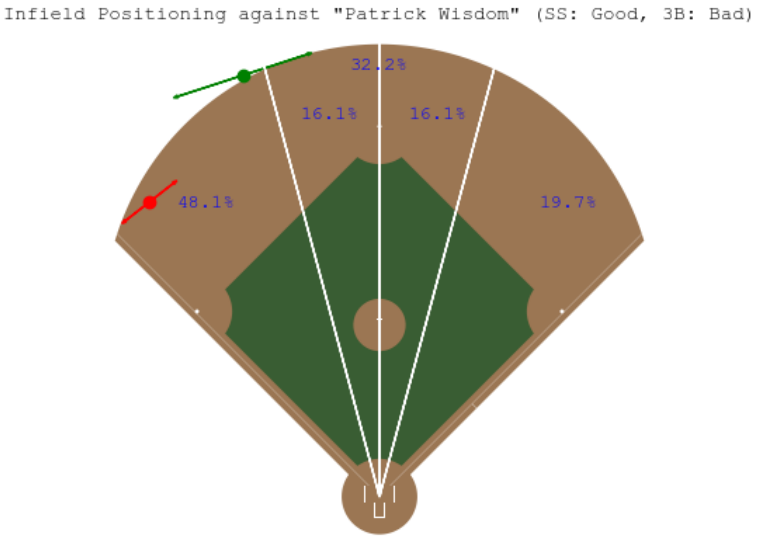


Figure 3.4: SS Range: 25ft, 3B Range: 12ft

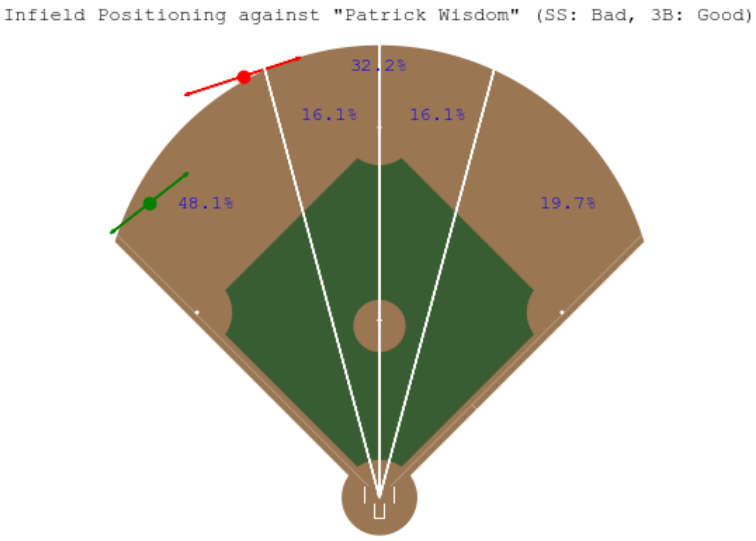


Figure 3.5: SS Range: 21ft, 3B Range: 17ft

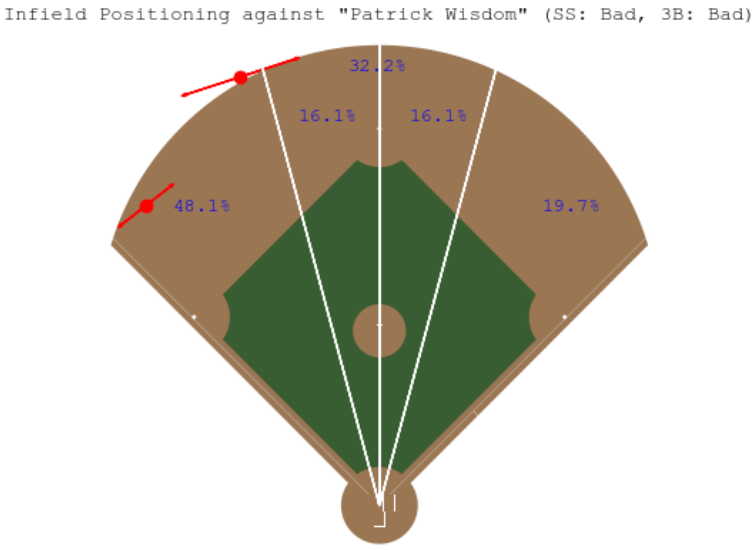


Figure 3.6: SS Range: 21ft, 3B Range: 12ft

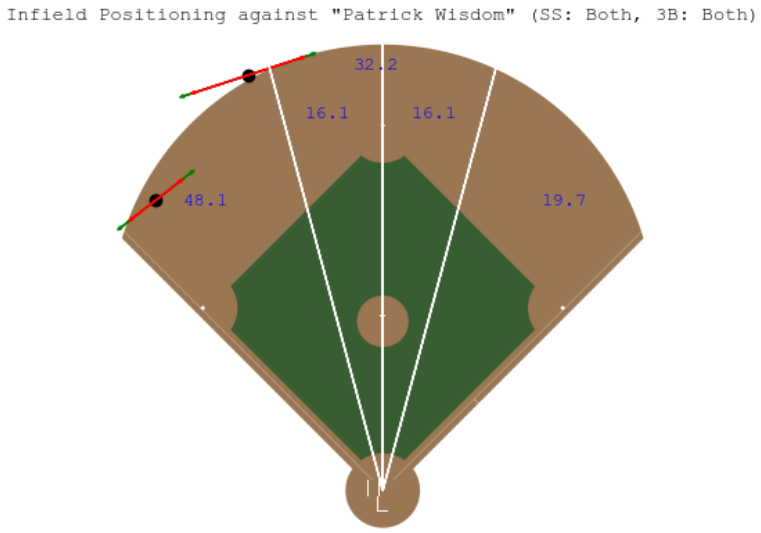


Figure 3.7: Combined Good and Bad

Based on the eye test, only teams with at least 75th percentile defenders at *both* SS and 3B would prevent the average ground ball hit by Patrick Wisdom from going between their shortstop and third baseman. Since the limiting factor was fielder's speed, we can use 2022 Outs Above Average (OAA) range percentiles as an analogy for my percentiles. Within the NL Central (Wisdom's division), the Cardinals (SS Edman 100% OAA, 3B Arenado 99% OAA) and Brewers (SS Adames 96% OAA, 3B Peterson 91% OAA) could afford to play the base alignment while the Reds (SS Farmer 28% OAA, 3B Drury 71% OAA) and Pirates (SS Cruz 3% OAA, 3B Hayes 100% OAA) could not.

3.2 Improvements

This project could be improved with a larger quantity and more precise data. However, methodological improvements include:

1. Instead of including the first 3 bounces, using all bounces and adding a feature for number of bounces that have occurred up to any point.

2. Using lateral speed instead of straight line speed to first. This could be done by limiting speed/time measurements to plays where a defender moved within a range of angles from their start position.

3. Differentiating between a fielder moving left vs. right. A fielder should have more range moving to their left because the throw is shorter and likely, on average, harder.

4. Not only considering routes perpendicular to the initial position. One approach could be that once a fielder can no longer reach a ball by going perpendicular, doing a similar iterative foot-by-foot process going away from home plate which would buy them more time to field the ball. Just reaching the ground ball was by far the "hardest" part since at the 75th percentile SS's max range there was still 750 milliseconds until the average runner would touch first base.

5. Predicting probabilities of whether a fielder can make an out rather than a binary outcome.

6. Predicting BABIPs based on an alignment

7. Ability to provide recommendations on alignments by testing different initial positions and minimizing BABIP