Searching for Cal:
An Analysis of Shortstop Range

SMT Data Challenge 2024
Team 005
Undergraduate Division

Abstract

A shortstop's range is an essential component of his value, but current attempts to understand the ins and outs of range tend to lack specificity, context, or both. To address this problem, an elastic-net logistic regression was used to model the probability of a shortstop successfully fielding a groundball, with predictors constructed from ball- and player-tracking data. The model outputs led to three conclusions: (1) There are limited number of plays where a shortstop's range influences success; (2) The angle between the shortstop and the ball is not only most important variable in determining success, but also something a shortstop may be able to control; (3) The speed of a shortstop, as measured by home-to-first times, did not seem to have an impact. These conclusions paint a clearer picture of a shortstop's range and can help team personnel gain a better understanding of their own players.

1. Introduction

Cal Ripken Jr. is one of the best defensive shortstops in baseball history. According to SABR's biography of the infallible Iron Man [1]:

"Cal's strength as a defensive player was his ability to play the batter. Knowing each hitter's tendencies at the plate and playing them accordingly allowed him to station himself at the right place on the field."

When we talk about a shortstop's range, it's often in reference to how much distance he can cover. By that definition, Ripken had poor range – he was slow, and didn't move much in pursuit of a ball. But it's precisely **because** his positioning was so good that he seldom moved. If we define range as a shortstop's ability to get to as many balls as possible, then Ripken had excellent range.

The question, then, is how to best incorporate and balance these two definitions, which current defensive metrics such as Defensive Runs Saved [2] and Outs Above Average [3] do not distinguish. As a result, they lack the specificity and context to be truly informative.

Using the information provided by this year's SMT Data Challenge, my goal was to identify and evaluate the components of a shortstop's range. In this paper, I detail my methods and results before suggesting how they should influence our evaluation of shortstops.

2. Data Acquisition

This year's data consisted of four farm system levels (1A to 4A), spanning two years. Home team players were part of that system, and away team players were re-anonymized each series. The data contained information about how each play in a game unfolded; corresponding timestamps; and the spatial locations of balls and players. See Appendix 9.1.

I was interested in grounders hit towards the shortstop, regardless of who ended up fielding them. Therefore, I considered plays meeting the following criteria:

- A ball-in-play (BIP) that **bounced at least once** and was **immediately fielded** by the shortstop, left fielder, or center fielder.

These criteria alone, however, do not guarantee that BIPs are groundballs rather than line drives. Therefore, I excluded balls with launch angles above five degrees. See Appendix 9.2.

Once I had the plays I wanted, I indicated the farm level and whether the shortstop fielded the ball. If a shortstop fielded the ball, I included ball and player position data until the moment of acquisition, but not after.

Even if someone else fielded the ball, we still need a stopping point. I looked at all the plays in which a shortstop got to the ball, then chose the maximum perpendicular distance relative to home plate as a limit.

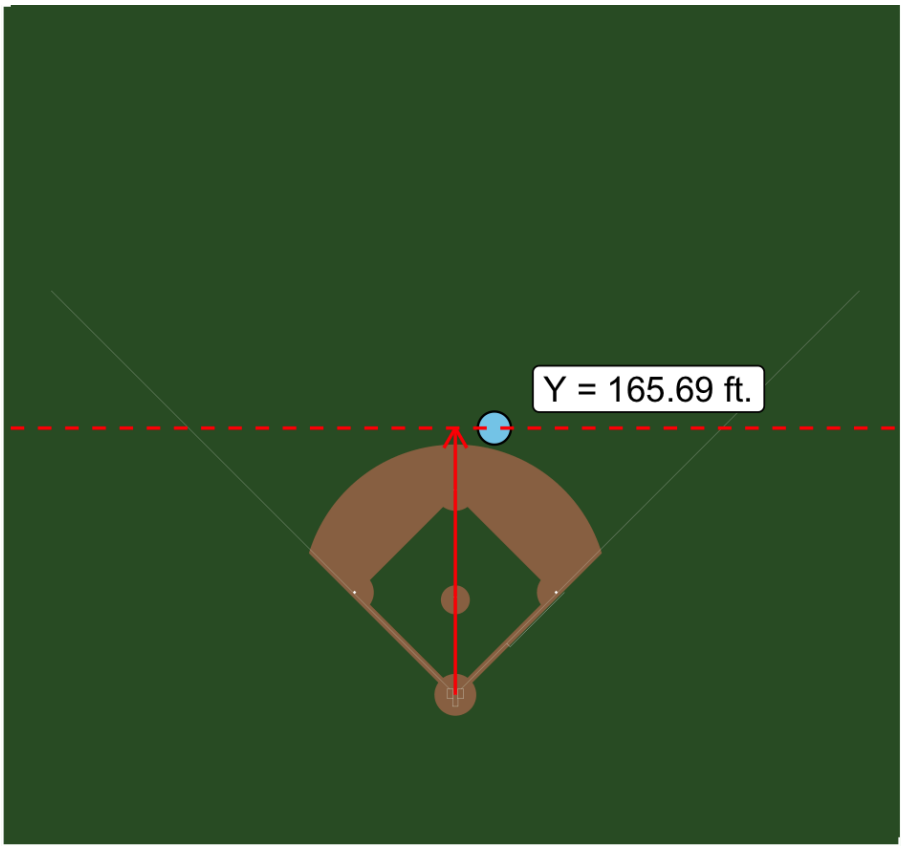

Figure 1: The highest recorded perpendicular distance of a shortstop, in feet, relative to home plate. The blue dot represents the shortstop.

There were a few plays with misaligned timestamps, so I estimated corrections – see Appendix 9.3. In total, I ended up with 902 grounders, of which 725 (80.4%) were fielded by the shortstop.

3. Model Building

My approach was to think from the perspective of a shortstop. Realistically, a shortstop doesn't move the instant the ball is hit, so it wouldn't make sense to start measuring at the moment of contact. Therefore, measurements of variables began 500 milliseconds **after** contact (Appendix 9.4). I considered these four variables:

Variable Name	Description
Initial Distance	The distance, in feet, between the BIP and the shortstop. (Fig. 2)
Angle Between	The angle, in degrees, between the BIP and the shortstop. (Fig. 2)
Distance Covered	How much distance, in feet, the shortstop covered in pursuit of the BIP. See Section 2. (Fig. 3)
Similarity	A measure of how the shortstop moved relative to how the BIP moved. See Section 2. (Fig. 4)

Table 1: Variables used for the shortstop range model, with descriptions.

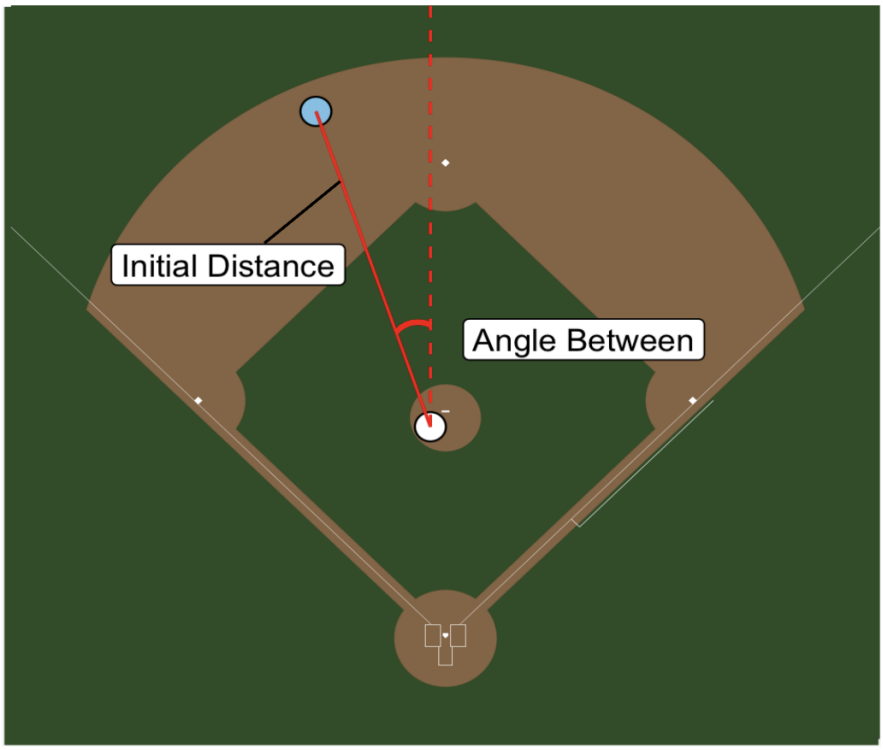

Figure 2: A visual representation of *initial distance* and *angle between*. (See Table 1.) The blue and white dot represent the shortstop and the BIP, respectively.

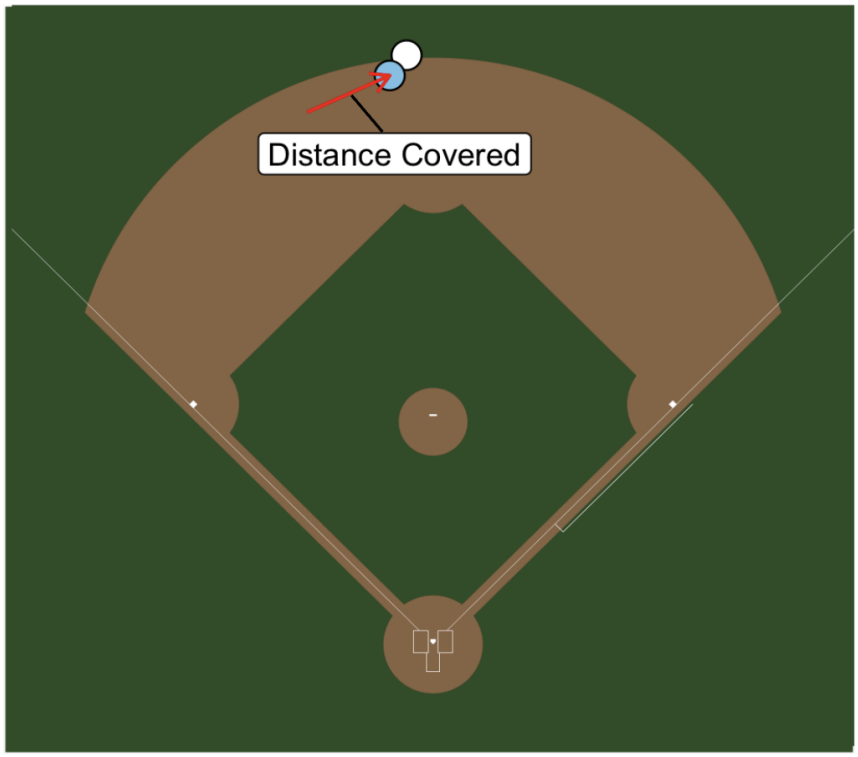

Figure 3: A visual representation of *distance covered*, at the moment of acquisition. The arrow indicates how much the shortstop traveled, and in what direction. (See Table 1.)

The last variable, *similarity*, is the cosine of the angle between the ball and player movement vectors. It is always between -1 (identical directions) and 1 (opposite directions). For a real-life example, imagine a shortstop is moving towards a ball that is moving directly towards him. Further, imagine representing these movements as arrows on a graph. Figure 4 shows what that would look like.

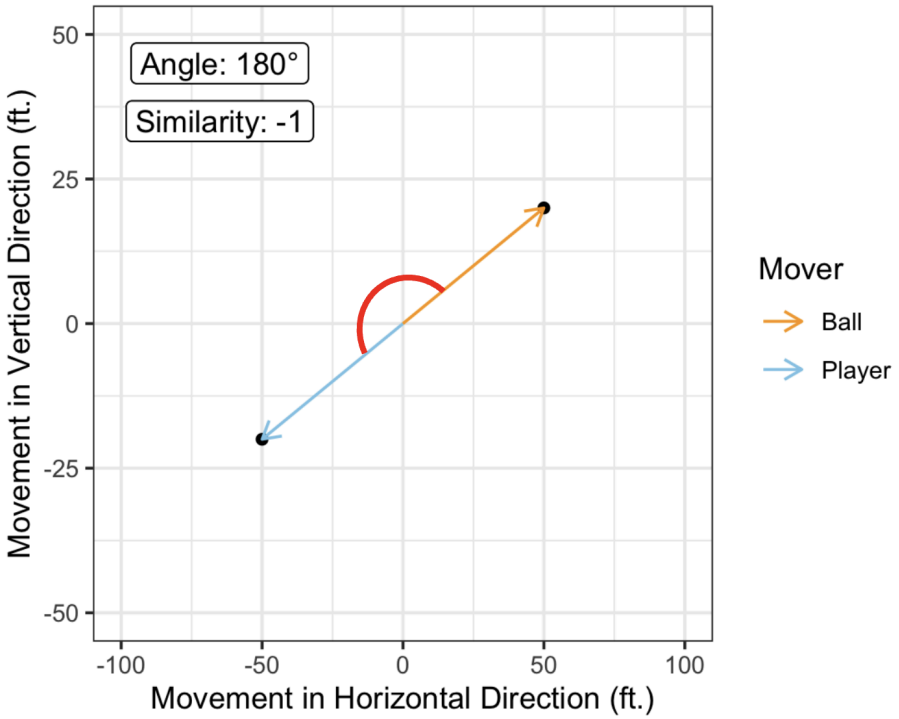

Figure 4: A visual representation of similarity. (See Table 1.)

The four variables were used to model the *probability* that the shortstop gets to the ball. The model of choice was an elastic-net logistic regression. A logistic regression is commonly used to model binary outcomes. In our case, 1 represents a successful fielding attempt, and 0 represents an unsuccessful one. Elastic-net regularization helps increase prediction accuracy by shrinking model coefficients and removing unnecessary variables.

Considering the disproportionate rate of eligible grounders fielded by the shortstop (80.4%), the model performed extremely well relative to a baseline naive estimate. See Appendix 9.4 for technical details.

4. Data Analysis

4.1. When Range Does (And Doesn't) Matter

Before anything else, let's explore how the different variables impact *expected success rate*, which is the modeled probability that a shortstop gets to the ball. For most variables, the relationship is straightforward:

- **The shorter the initial distance, the lower the success rate.** A shorter *initial distance* is often a signal for a high exit velocity, and shortstops have less time to react.
- **The greater the angle between, the lower the success rate**. A ball hit directly towards the shortstop is routine. A ball hit up the middle is considerably more difficult.

- **The greater the similarity, the lower the success rate**. Unfortunately, this isn't too informative. A high *similarity* generally means that the *angle between* is large.

However, the relationship between *distance covered* and success rate is more complex. To understand it, we also need to look at *angle between*. (Figure 5)

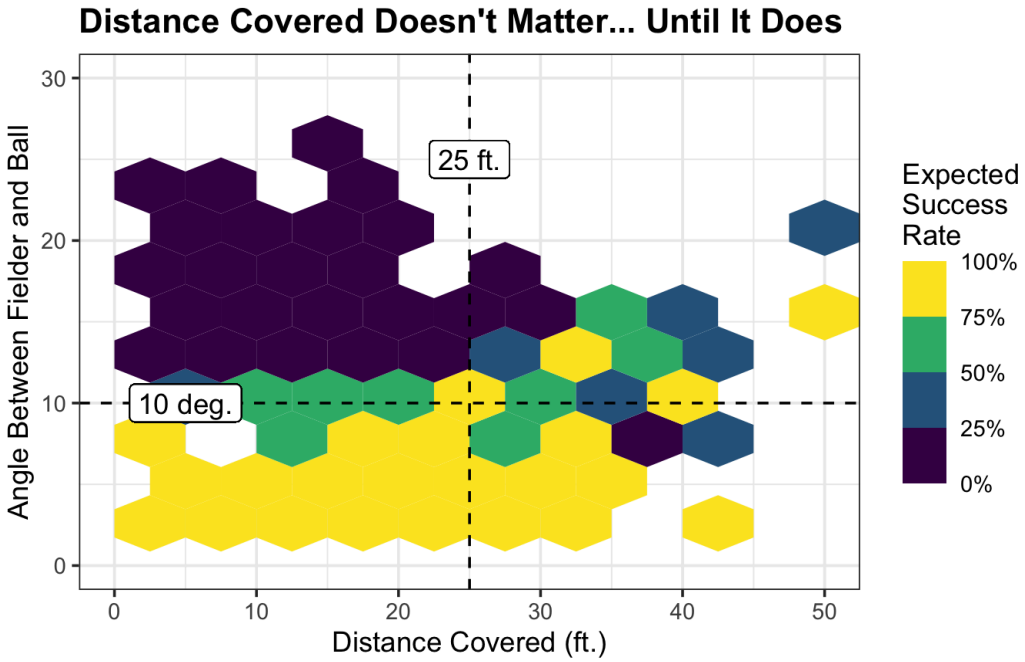

Figure 5: *Expected success rates* from the model, as determined by *distance covered* and *angle between*. Each hexagon is the average success rate of ten observations.

For angles < 10, it doesn't matter how much distance a shortstop covers – the play is usually routine. It's only on angles above ten degrees that *distance covered* has a significant impact on success rate. Even then, a shortstop must travel > 25 feet to have a chance at fielding the ball. The upper-right quadrant of the graph, then, is the opportunity space of a shortstop, where an individual's defensive capability matters most.

This is an eye-opening discovery: For most plays, a shortstop's "range" in the traditional sense does not determine whether he is able to field the ball. It suggests that any evaluations of "range" should only occur on plays within the aforementioned opportunity space.

4.2. Dr. Rangelove or: How I Stopped Worrying and Learned to Love the Angle

In Section 4.1, we saw that *distance covered* can only be understood in the context of *angle between.* As it turns out, *angle between* is the most important factor in determining a shortstop's *expected success rate* (Figure 6).

One way to measure the *importance* of a variable in a classification model is to see how much accuracy is gained by including it. Here, I used a random forest model, and determined *importance* by the Gini impurity index (Appendix 9.4).

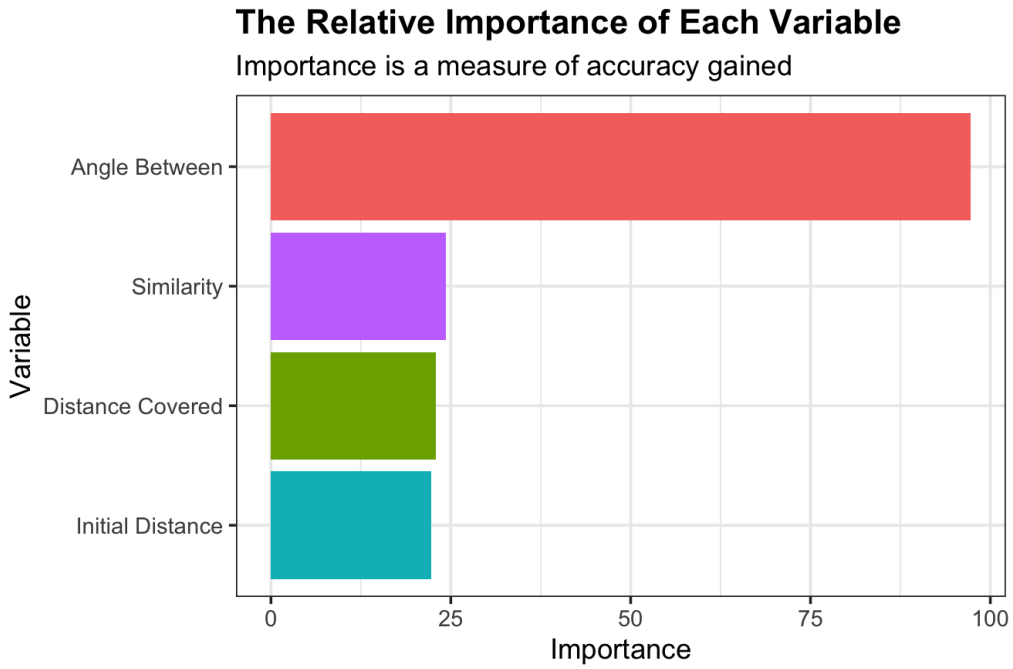

Figure 6: The importance of each variable, obtained from a random forest model. The scale is irrelevant; only the relative values matter.

Despite its importance, defensive metrics today do not include an equivalent of *angle between*. This is because *angle between*, like *initial distance*, is seen as something that is determined by the batter, not the fielder. *Distance covered* and *similarity*, on the other hand, are controlled by the shortstop.

However, there's evidence that shortstops may have, at the very least, some control over their angle to the BIP. Recall that our data comes from multiple levels of the same farm system. Assuming that farm level corresponds to defensive ability, we'd expect shortstops at higher levels to have a lower *angle between.*

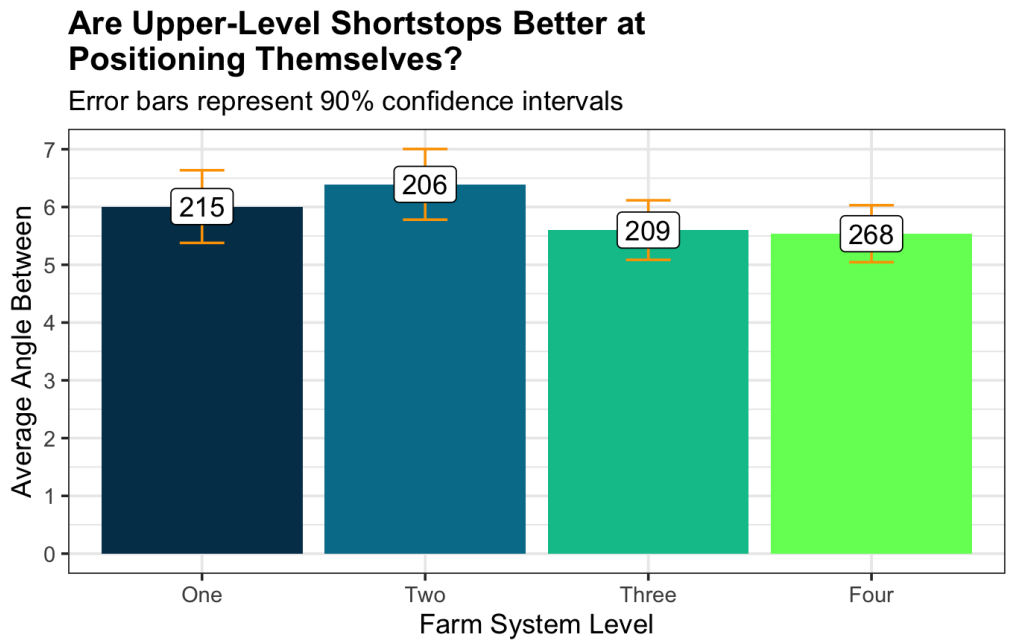

Figure 7: The average *angle between* by level. The numbers indicate the grounders from each level, totaling 902. See Section 2.

Figure 7 shows that the average *angle between* does seem to be lower at higher levels, although the difference is within the margin of error. However, while setting up a formal comparison of means, I noticed that the *variance* in each level was different; this difference was statistically significant (Appendix 9.5). A low *variance* indicates that the values are more concentrated around the average. The difference between levels is made more clear after sorting them into two groups.

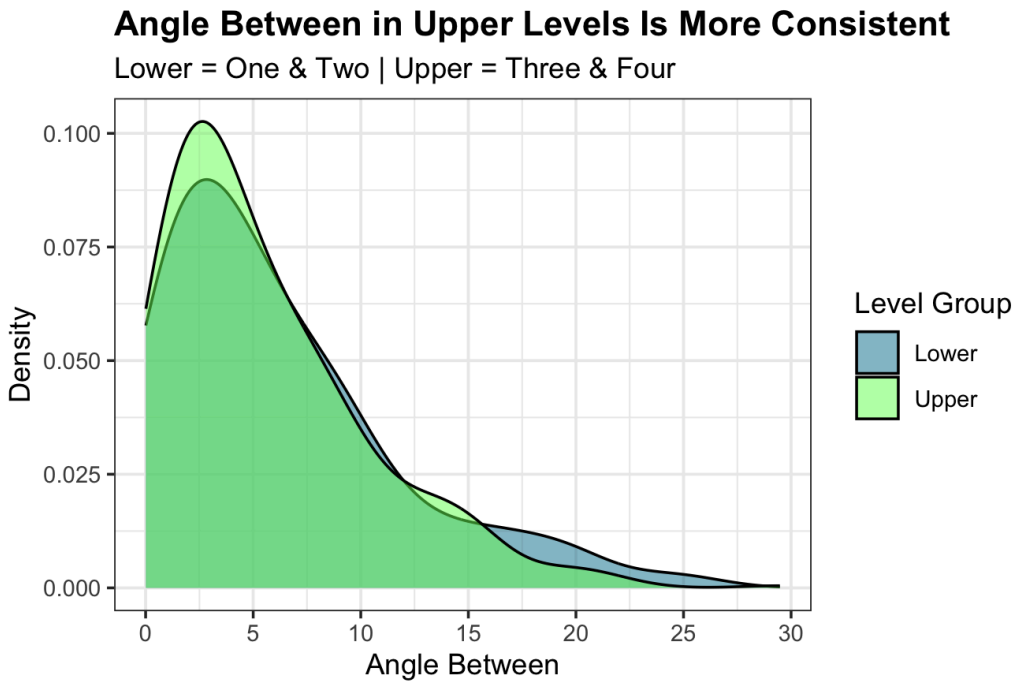

Figure 8: The distribution of *angle between* by level group.

As Figure 8 shows, there is less *variance* in the upper levels. This suggests that upper-level shortstops are better at positioning themselves (i.e. lowering their *angle between*) in order to maximize their fielding chances. Not surprisingly, the average *expected success rate* is higher in upper levels (83.0%) than in lower levels (78.8%).

4.3. (No) Need for Speed

In my initial modeling, two attempts were made to incorporate speed. First, I measured how much distance the shortstop covered over one second. The logic was that this would capture the "reaction speed" of the shortstop. Next, I considered *distance covered* divided by time elapsed. Both attempts significantly degraded the performance of the model.

I began to wonder if speed had **any** relationship with *expected success rate*. To quantify speed, I averaged each shortstop's 90th percentile-or-better home-to-first times, eliminating low-effort plays. This isn't as precise as the sprint speed metric on Baseball Savant, but it is a sufficient proxy. The relationship between speed and success rate is shown in Figure 9.

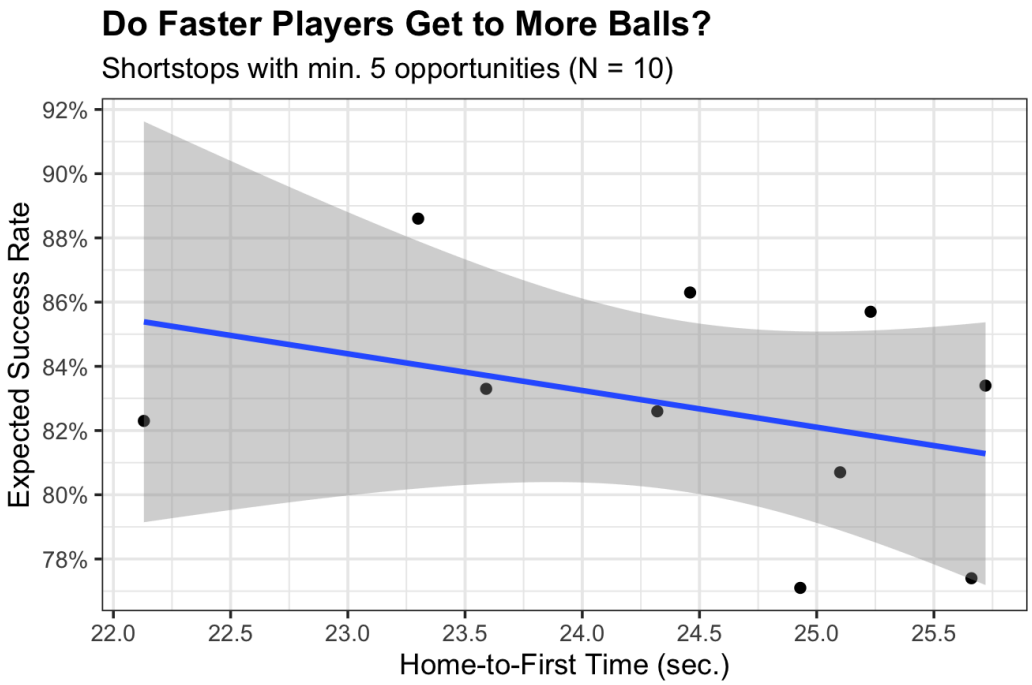

Figure 9: The relationship between speed and *expected success rate*. The blue line is the line of best fit.

Each dot represents a shortstop with five or more fielding attempts. If anything, home-to-first time has a slightly negative relationship with *expected success rate*. But given the small sample size, it's best to conclude that the existence of any relationship is dubious.

Perhaps the traditional ways to evaluate a player's speed have little-to-no applications to infield range. It would seem that shortstops can make up for their sluggishness with good positioning, but have significant trouble overcoming a disadvantageous angle with speed alone.

5. Applications

Let's compare two shortstops, identified in the data with the IDs 337 and 345.

Shortstop ID	Speed (ft./sec.)	Distance Covered (ft.)	Angle Between (deg.)	Expected Success Rate (%)
337	25.66 (1st)	14.27 (5th)	6.6 (3rd)	77.4 (9th)
345	23.3 (9th)	9.73 (10th)	5.2 (9th)	88.6 (1st)

Table 2: How the two shortstops fare by select metrics. Inside the parentheses are each player's ranks among ten shortstops with min. 5 fielding attempts.

Table 2 shows that 345 could be a shortstop in the mold of Cal Ripken Jr., making up for a lack of speed and range with solid positioning. By the heuristics commonly used to judge shortstops, 337, who is faster and covers more ground, would appear to be the better fielder. Yet, because his *angle between* is so low, 345 has a far better *expected success rate* than 337. While it remains to be seen if *angle between* is a metric that truly "stabilizes," it should not be dismissed.

Most of 345's plays were from level 3A. Meanwhile, despite his inferior success rate, 337 spent the bulk of his time at 4A. Although there may be other reasons for 337 playing at the higher level, there's a strong case for promoting 345 if his good results continue.

6. Conclusion

There are three takeaways from this paper:

- How much distance a shortstop covers is only relevant on certain plays.
- Positioning is extremely important, and there is evidence that it can be quantified.
- A shortstop's speed does not impact whether he can get to a BIP.

It took years for the league to recognize the defensive greatness of Ripken. Now, we can identify shortstops with profiles similar to that of the Hall of Famer, all the way back at the minor-league level. While the findings in this paper could be made more rigorous with non-anonymized data, they nonetheless present new avenues for research in infield defense, with numerous applications to teams, coaches, and players.

7. References

[1] Keenan, Jimmy. “Cal Ripken.” *sabr.org*, https://sabr.org/bioproj/person/cal-ripken/.

[2] Slowinski, Steve. “DRS.” *fangraphs.com*, https://library.fangraphs.com/defense/drs/.

[3] Petriello, Mike. “A new way to measure MLB's best infield defenders.” *mlb.com*, https://www.mlb.com/news/statcast-introduces-outs-above-average-for-infield-defense.

8. Acknowledgements

GIF animations of baseball plays were used extensively in crafting this paper. The code for those animations are courtesy of David Awosoga, who shared it in the SMT Data Challenge Slack channel.

Thank you to Dr. Meredith Wills and Cameron Adams for their invaluable feedback.

Thank you to Mom and Dad, who are the most supportive parents a child could ask for.

9. Appendices

9.1. Glossary of Data

Below is a guide to the tables and the variables within them that I used for this project. Note that this is not a comprehensive list of all the tables and variables that were made available.

GAME_EVENTS

- game_str: Each game string identifies the year and sequential day within a season.
- play_id: A numerical id for each play within a game, beginning at a 1 and listed consecutively.
- timestamp: Times, measured in milliseconds, that start at the beginning of each game and are in 50-millisecond increments, with a few exceptions.
- player_position: A number that corresponds to a player’s position.
- event_code: A number that corresponds to a unique ball event. (See Appendix 9.2.)

BALL_POS

- ball_position_x: The x-coordinate of the ball at a given timestamp, measured in feet.
- ball_position_y: The y-coordinate of the ball at a given timestamp, measured in feet.
- ball_position_z: The z-coordinate of the ball at a given timestamp, measured in feet.
- Also includes: game_str, play_id, timestamp.

PLAYER_POS

- field_x: The x-coordinate of a player on the field at a given timestamp, measured in feet.

- field_y: The y-coordinate of a player on the field at a given timestamp, measured in feet.
- Also includes: game_str, play_id, timestamp, player_position.

GAME_INFO

- home_team: All home teams are part of the same farm system. Team designations correspond to four consecutive minor league levels, ranging from Home1A to Home4A.
- away_team: Away teams are anonymized to be series-specific but are also from various levels of different farm systems, ranging from Vis1## to Vis4##, where ## indicates two consecutive random letters.
- play_per_game: A number indicating each play within a game, beginning at 1 and listed consecutively. This is equivalent to the play_id variable.
- top_bottom: Whether a play occurred in the top or bottom half of an inning.
- shortstop: A unique player ID that identifies a shortstop.
- Also includes: game_str.

9.2. Data Acquisition Details

A ball put into play corresponds to *player_position* = 10 and *event_code* = 4. A bounce corresponds to *player_position* = 255 and *event_code* = 16. A catch is denoted by *event_code* = 2, so in our case, the player position can be 6 (shortstop), 7 (left fielder), or 8 (center fielder). Querying for these player positions and event codes gave me the data I wanted.

It is surprisingly hard to verify that a bounce is the only event between the ball being put into play and being caught by a fielder. What I did was check for all one-bounce grounders, then two-bounce, three-bounce, and so on. For example, a one-bounce grounder is a play that begins with the *player_position*, *event_code* combination of 10-4, followed by 255-16, and ends with 6/7/8-2. A two-bounce grounder follows the pattern of 10-4, 255-16, 255-16, 6/7/8-2. No rows were returned in the six-bounce grounder query, so that was my stopping point.

Launch angle is defined as $arctan(\frac{\Delta\ ball\ position\ z}{\Delta\ ball\ position\ y})$. Initially, both $\Delta\ ball\ position\ z$ and $\Delta\ ball\ position\ y$ were measured 50 milliseconds after the time of contact, but this led to unrealistically high launch angles. Instead, they were measured 100 milliseconds after the time of contact, which “smoothed out” the angles and better accounted for interpolation errors.

The cutoff for groundballs is arbitrary, but it has to be. A popular threshold is ten degrees, which can be seen on Baseball Savant. However, because my custom launch angles had a narrow distribution centered around zero, a ten-agree threshold ended up including unwanted line drives. After some experimentation, a five-degree threshold was found to work best.

9.3. Coordinate Data Estimation

For three plays that occurred in the same game, the player position timestamps were 7 milliseconds higher than the ball position timestamps. Therefore, I needed to make an educated guess on where the ball was 7 milliseconds after the original listed time. Assuming constant

acceleration and a straight-line path, I added $.14 * \Delta\, ball\ position\ x$ and $.14 * \Delta\, ball\ position\ y$ to the existing $ball\ position\ x$ and $ball\ position\ y$, respectively. The constant is derived from the 50-millisecond intervals between timestamps. (i.e. $\frac{7}{50} = .14$)

9.4. Modeling Details

Let's (generously) assume that the hardest a minor leaguer can hit a baseball is 120 mph. After 500 milliseconds, the ball would have traveled roughly 88 feet – behind the mound, but still in front of the shortstop. This is why 500 milliseconds after contact makes sense as a starting point. If our starting point is too low, we fail to capture the information-gathering process. If our starting point is too high, there's a chance that the BIP has already passed the shortstop.

Similarity is formally known as "cosine similarity." One advantage of using cosine similarity over the dot product is that the magnitudes of the vectors do not influence cosine similarity. Thus, we would not have to account for the BIP and the player having different amounts of movement.

Both a random forest and an elastic-net logistic regression with interaction terms for all predictors were tested. A random forest was chosen because of its ability to capture complex relationships between variables. A logistic regression also made sense, as the response variable is binary (1 if the shortstop got to the BIP, and 0 otherwise). For the logistic regression model, interaction terms were essential due to the linear dependence between the predictor variables. Elastic-net regularization combines the benefits of LASSO and ridge regression by shrinking coefficients and removing unnecessary predictors to reduce bias, and therefore increase prediction accuracy. The predictors were also normalized so that the coefficients would be on the same scale. No preprocessing was required for the random forest model.

Training and testing data was established using a 75-25 split and stratified sampling to maintain the proportion of successful shortstop fielding attempts. 10-fold cross validation with 5 repeats was used to validate the models and find the optimal tuning parameters. Log-loss was the target metric, as other classification metrics such as accuracy do not account for the errors in predictions. While both model types seemed to return similar results, the logistic regression was chosen due to its interpretability and lower risk of overfitting the testing data. Indeed, running the random forest on the testing data produced a log-loss score of .190, whereas the logistic regression produced a log-loss score of .162.

The final logistic regression fit is shown in Table 3. Note that the euclid_dist variable refers to the "initial distance" used in the main text. An "x" between two variables denotes an interaction:

term	estimate
(Intercept)	3.12826437
euclid_dist	1.86626486
angle_between	-3.19887364
distance_covered	-0.56052285
similarity	0.23642969
euclid_dist_x_angle_between	0.41402912
euclid_dist_x_distance_covered	0.09172355
euclid_dist_x_similarity	0.15800323
angle_between_x_distance_covered	0.51933018
angle_between_x_similarity	-0.06750556
distance_covered_x_similarity	-0.78042767

Table 3: The coefficients of the logistic regression model. Each coefficient represents how a one-unit change in the variable affects the probability that a shortstop gets to the ball.

How do we know that the model is any good? A naive estimate is to assume that the probability of a shortstop getting to the ball is simply the proportion of balls fielded by the shortstop: .804. This would result in a log-loss score of roughly .500. Considering that our actual log-loss score (.162) is much lower, we have overwhelming evidence that the model is more informative than the data itself.

Gini impurity is a measure of how often a data point would be labeled incorrectly if it were labeled randomly. A random forest model aims to minimize the Gini impurity. A variable's importance is determined by how much Gini impurity is *subtracted* from the original impurity. Thus, if a variable is important, the amount subtracted will be high.

While I could have used the coefficients of the logistic regression as indicators for variable importance, they would have been somewhat misleading due to the interaction terms. For example, similarity alone has a positive relationship with expected success rate, but it's the interaction between distance covered and similarity that contributes to an overall negative relationship, as stated in the paper.

9.5. Equal Variance Testing

Analysis of Variance, or ANOVA, requires the subpopulations to have equal (or at least roughly equal) variances. There are many ways to verify the equal variance assumption; I used Levene's test due to its robustness against non-normal data. (The distribution of the angles between the ball and player has a heavy right-skew.) The test returned a p-value of .0255. Using the widely accepted alpha threshold of .05, we have sufficient evidence to reject the null hypothesis that the variance of the subpopulations are equal.