An Investigation Into Minor League Fielding Through Two Key Metrics

Ben Wieland and Ajay Patel

I. Introduction

Our understanding and ability to capture fielding performance has improved vastly as more and more information has become available. MLB teams have greatly detailed metrics to evaluate their players in each aspect of defense, whether it be through their routes, arm, or reads. We haven't seen the same translation, at least publicly, yet, for minor league players. Given that this challenge focused on minor league data, we saw an opportunity to apply some major league metrics to minor leaguers, giving us a phenomenal level of insight into every player with significant available data..

Two metrics jumped out to us initially: route efficiency and outs above average, both propelled by Statcast. The tracking data made available to us allowed for the possibility of these metrics to be recreated through our own unique methodology for the given minor league players. Analyzing the two stats' relationship and their individual standing was our main goal. By creating the two metrics, we would not only get a good grasp of which fielders perform best, but we would develop an intricate understanding of each metric and its strengths and flaws — and an understanding of how front-office decision makers can apply each statistic in their own player evaluation processes.

II. Collecting the Data

To evaluate outfielder effectiveness, we collected a subset of play-by-play data that met the following criteria:

1. The batter put the ball in play
2. The ball traveled 150 feet from home plate before either being caught, fielded, or bouncing off the wall or the ground
3. The ball's launch angle was greater than 10 degrees, and would have been classified as a line drive, fly ball, or pop-up by MLB's Statcast guidelines.

For each of these plays, a variety of useful variables were collected, including whether the batter reached base, the fielder's location at the moment the ball was hit and the moment the ball was fielded, which fielder was responsible for each recorded out, and the location of the ball at the moment it was caught, fielded, bounced off the wall, or bounced off the ground.

The final dataset contained 1,915 plays from 97 games involving 37 different teams: four home teams and a rotating set of 33 away teams.

III. Route Efficiency

Route efficiency is a fairly simple, yet truly indicative metric. It gives us insight into which fielders, in this specific case, outfielders, are taking the most successful routes to the ball and who may be wasting time taking inefficient paths.

And in certain situations it is extremely meaningful. If there is a runner tagging up from second, the outfielder needs to get the ball in as quick as possible. A long route may prevent this, but an efficient one would mean the fielder did everything possible to prevent the runner from advancing. While route efficiency is most important for turning hits into outs, it does provide value in other cases as such.

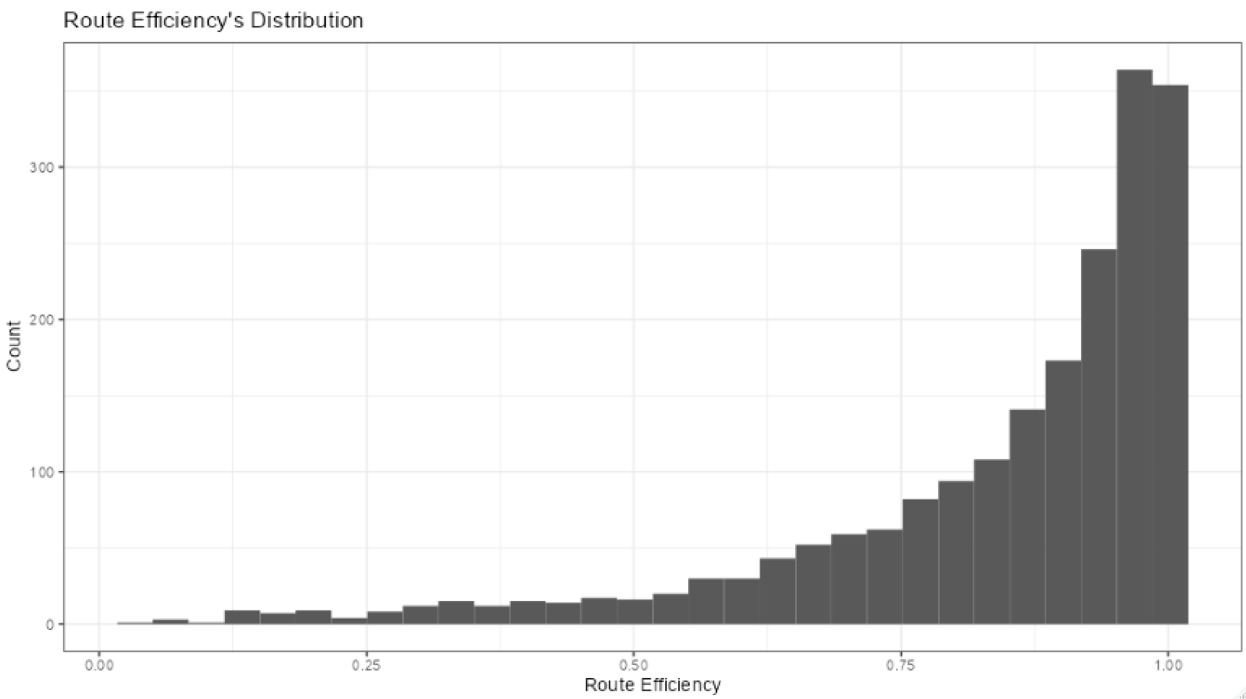
To create the metric, we had to narrow down our aforementioned dataset. It was filtered to only balls hit to outfielders or ones where the ball dropped in for a base hit. This ensures our dataset is not only focused on flyouts, which would lead to selection bias issues as of course fielders will have efficient routes when making outs.

Now that we had our ideal dataset, we had to add our tracking data to it. To do this, the player_pos csv file was utilized. A distance_x and a distance_y variable were created based on the distance covered in each direction after each tracking capture for each play and each fielder. A little more manipulation was required: grouping by each play allowed us to find a total distance traveled by summing up the two distance variables for each fielder, and it was then assigned to a new dataframe called total_distance_traveled. This process was repeated, but instead of total distance traveled, straight line distance was found by subtracting a player's starting point from his end point in the x and y direction. This is what was considered the ideal route, and it was assigned to a new dataframe called straight_distance.

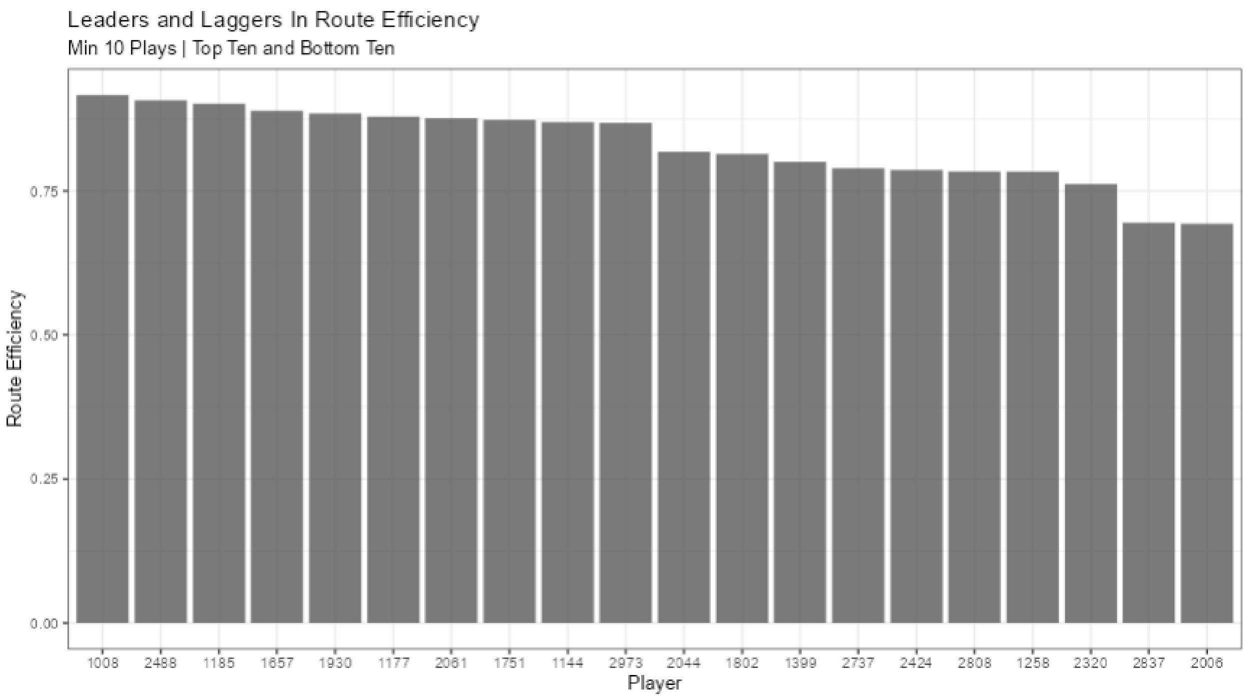
The two new dataframes, total_distance_traveled and straight_distance, were joined into a new dataframe. A few outlier plays were removed where total distance was somehow less than the straight line distance, which we chalked up to errors in the data. Our last steps were simple: using the distance formula on the x and y variables we found the total distance covered and the straight line distance for every play, and then to find route efficiency, straight line distance was divided by total distance traveled.

To present our final results, the route efficiency dataset was joined to the final dataset we created, mentioned in section II of the paper, to focus on action plays and the outfielders involved.

The first thing worth validating was the distribution of route efficiency. MLB players are usually around 0.8-0.95 for route efficiency, skewed to the left. This distribution tracked with our priors and intuition, as most plays involved fairly efficient routes with the highest frequencies occurring at great route efficiency values.

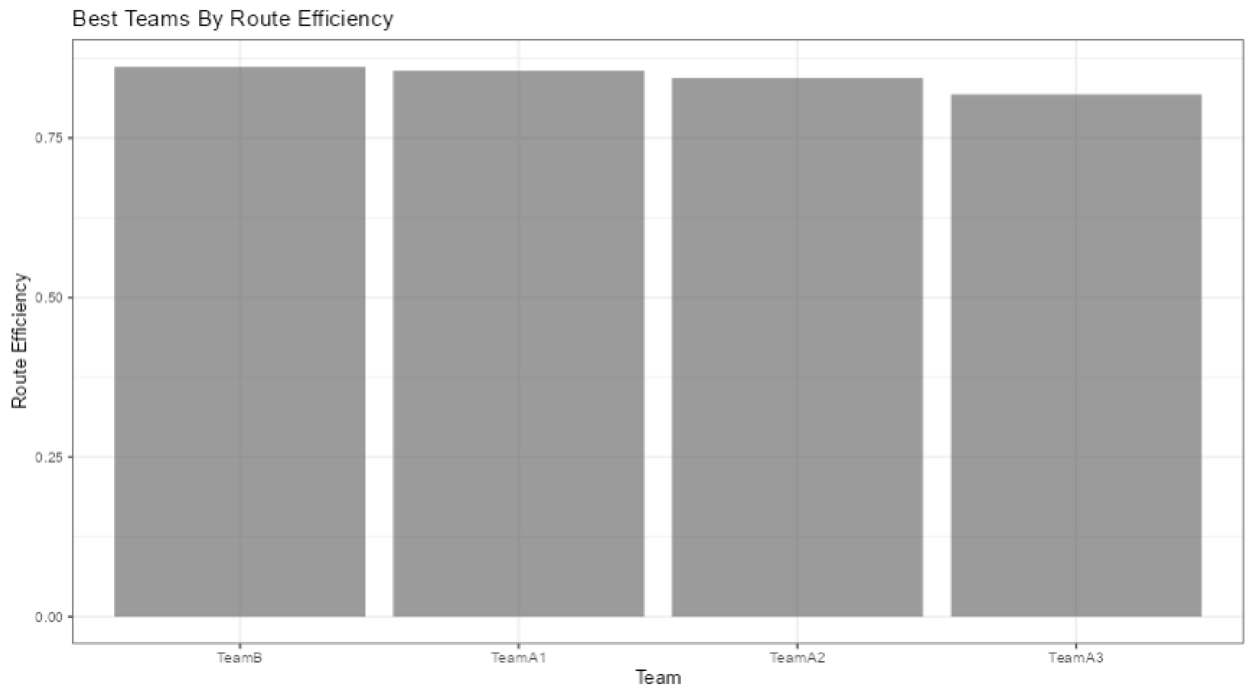


Route efficiency's usefulness largely comes on a player by player basis. Which players are getting to balls as efficiently as they can and who may be taking unideal routes? The graph below visualizes the top ten and bottom ten players by route efficiency, minimum ten plays to remove extreme outliers.



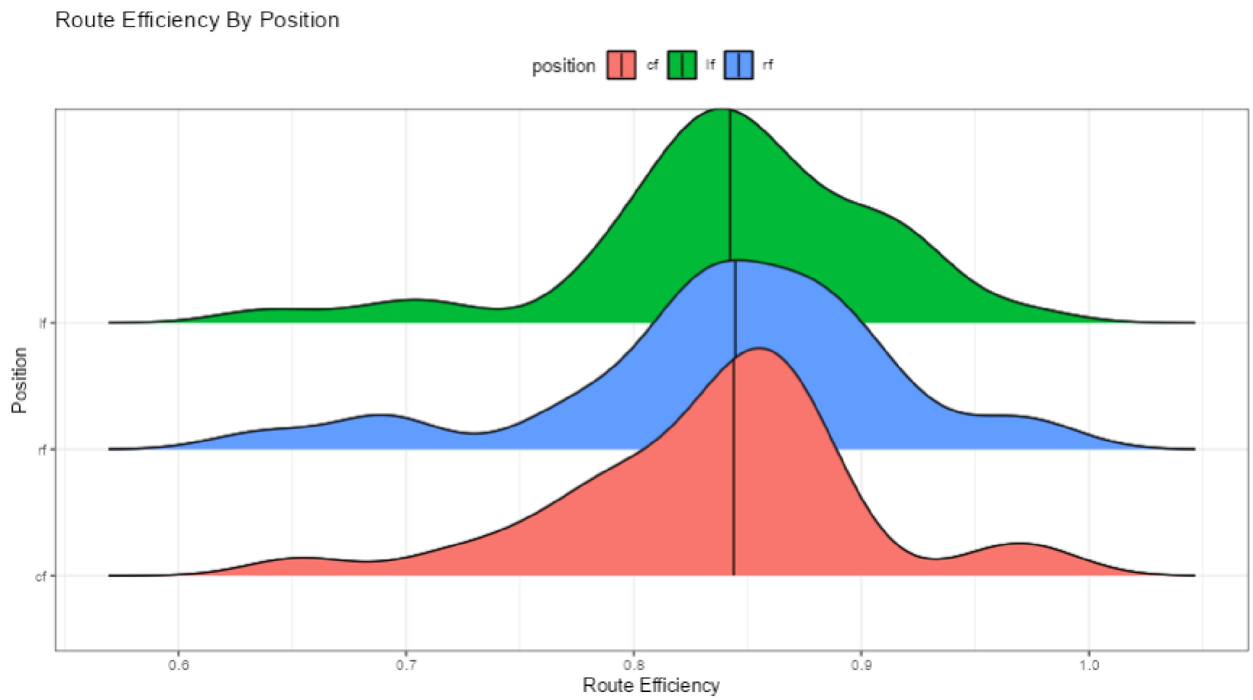
As seen above, players 1008 and 2488 put up outstanding routes consistently, whereas players 2837 and 2006 lag behind by a fair bit. This data provides great use for players, as it could easily validate concerns they have about wasting time on their routes or confirm their defensive prowess. It also can be analyzed on a team-level basis to see how teams perform overall, possibly due to specific coaching techniques or just having skilled players.

The chart below demonstrates the route efficiency of the four main teams included: A1, A2, A3, and B.



On a team-level scope, we see there is not too much variance in route performance. Team B slightly outpaces the other three, with Team A3 leaving a little to be desired. Again, this data provides value in terms of recognizing defensive strengths or flaws. Focusing on the flaws, this would signal to organizations like Team A3 that slight changes may be needed, whether in training regimen or coaching input.

One last, yet very important, analysis was conducted to see if route efficiency possibly biases towards center fielders. A typical trope mentioned is that your center fielder is usually your best outfielder, so it was worth analyzing to see if that was really true for this set of data. As seen in the chart below, there really is no stark difference between the three outfield positions when examining route efficiency.



If anything, we do see slightly more consistent performance from center fielders with more variance from right fielders, but all three positions are largely similar in terms of performance.

Route efficiency allows for succinct yet meaningful analysis across a player and team level. Discerning which players take worse routes and whether that's a team-wide trend or simply a player-specific issue provides great value for teams. On its own, though, it can only do so much. It is best utilized with metrics that can consider other aspects of fielding performance, including arm strength and initial jumps. Nonetheless, by finding route efficiency on outs and hits, we get a solid glimpse into which fielders outperform others.

IV. Outfielder Outs Above Average

Two key factors affect an outfielder's ability to make a play on a ball hit towards them in the air: their distance from the ball's landing point, and the amount of time the ball spends in the air before hitting the ground.

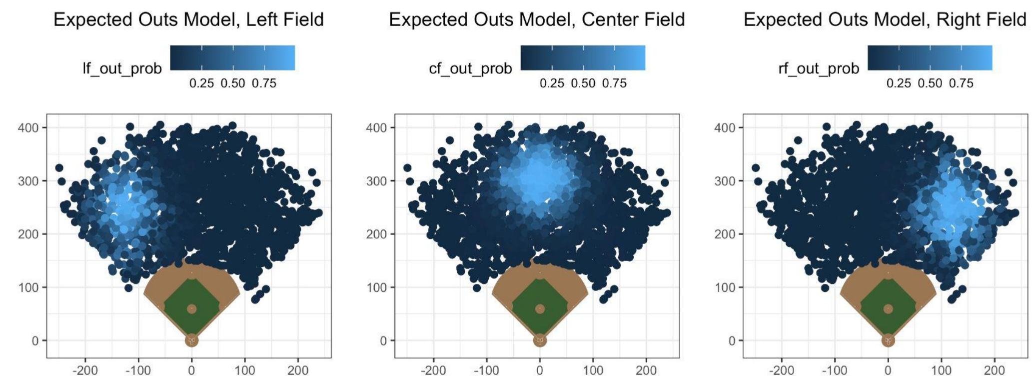
Hits with low distance and high time — for example, a can-of-corn popup directly to the center fielder — should be expected to result in outs almost 100% of the time. On the other hand, hits with high distance and low time — such as a line drive ripped into the center-right field gap — should be expected to result in outs almost 0% of the time. Where fielders can add value is in the area between those two extremes. Players who can turn "50/50 balls" into outs 75% of the time, for example, are incredibly valuable to their team.

This is where the concept of "Outs Above Average" comes into play. For each ball in play, a statistical model taking into account those two key constraints of distance and time can be used to create an expected outs value between 0 and 1; the higher the ball's expected outs value, the easier the play is expected to be for an outfielder.

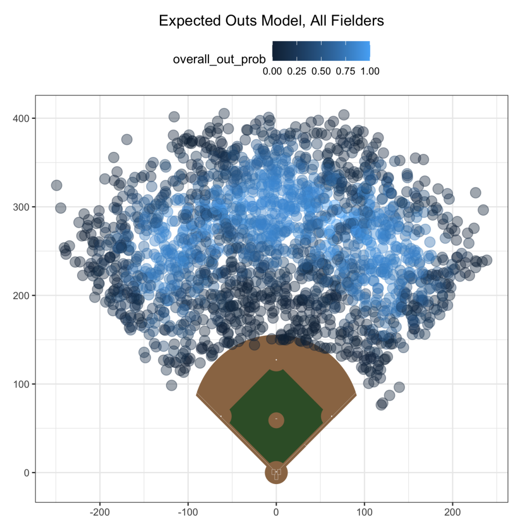
An outfielder's total expected outs can be calculated by taking the sum of their expected outs on every play they are deemed to be responsible for, which we defined as every play where they started closer to the ball than any other outfielder. By comparing a player's actual outs to their expected outs, their fielding play can be evaluated: good fielders record more outs than expected, while poor fielders record fewer outs than expected.

To create the expected outs model, three logistic regression models were fit using the dataset outlined in Section II. These models projected the probability of the left fielder, center fielder, and right fielder recording an out on each play. Four terms were considered in each of the final models: the ball's horizontal distance from the average starting location of the fielder, the ball's vertical distance from the average starting location of the fielder, an interaction term between those two distances, and an indicator variable for the type of hit as classified by its launch angle: line drive, fly ball, or pop-up.

The chart below displays the results of the expected outs model for each position for every play in the dataset. Lighter areas of the charts indicate a higher probability of an out being recorded. These charts show that most batted balls for a given fielder have an out probability very close to 1 (a guaranteed out) or 0 (no chance of making a play on the ball), with a small area in between where the best and worst fielders can differentiate themselves. They also confirm the hypothesis that distance is a key factor in out probability: unsurprisingly, balls hit closer to an outfielder are more likely to result in outs.

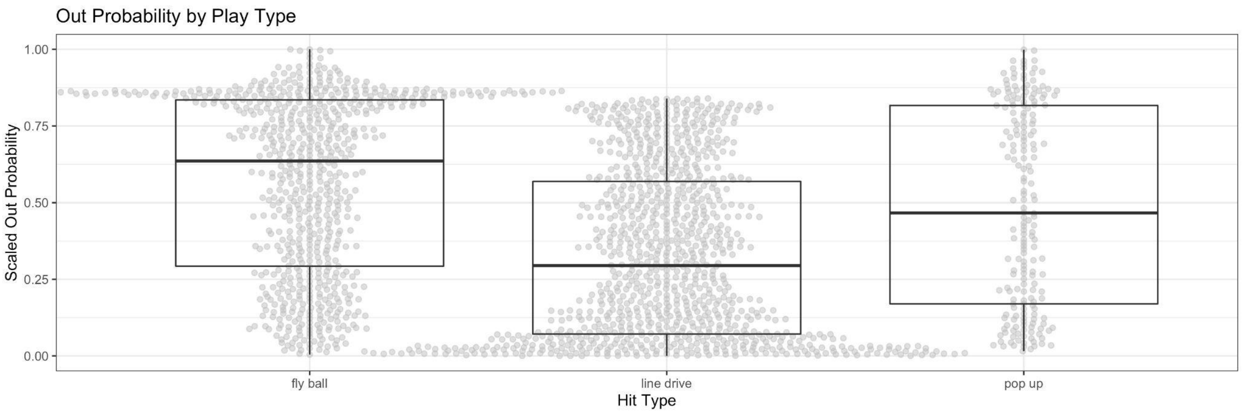


These three probabilities can also be added together and scaled between 0 and 1 to get an approximation of "overall out probability." This isn't a mathematically sound statistic in the absolute for calculating expected outs, but it does allow for relative comparison between the plays in the dataset. The chart below shows the estimated likelihood of an out being recorded on each ball in play: the balls with the lowest overall out probability seem to be balls hit into the right-center and left-center gaps, balls hit either very shallow or very deep, or balls hit down the left and right field lines.

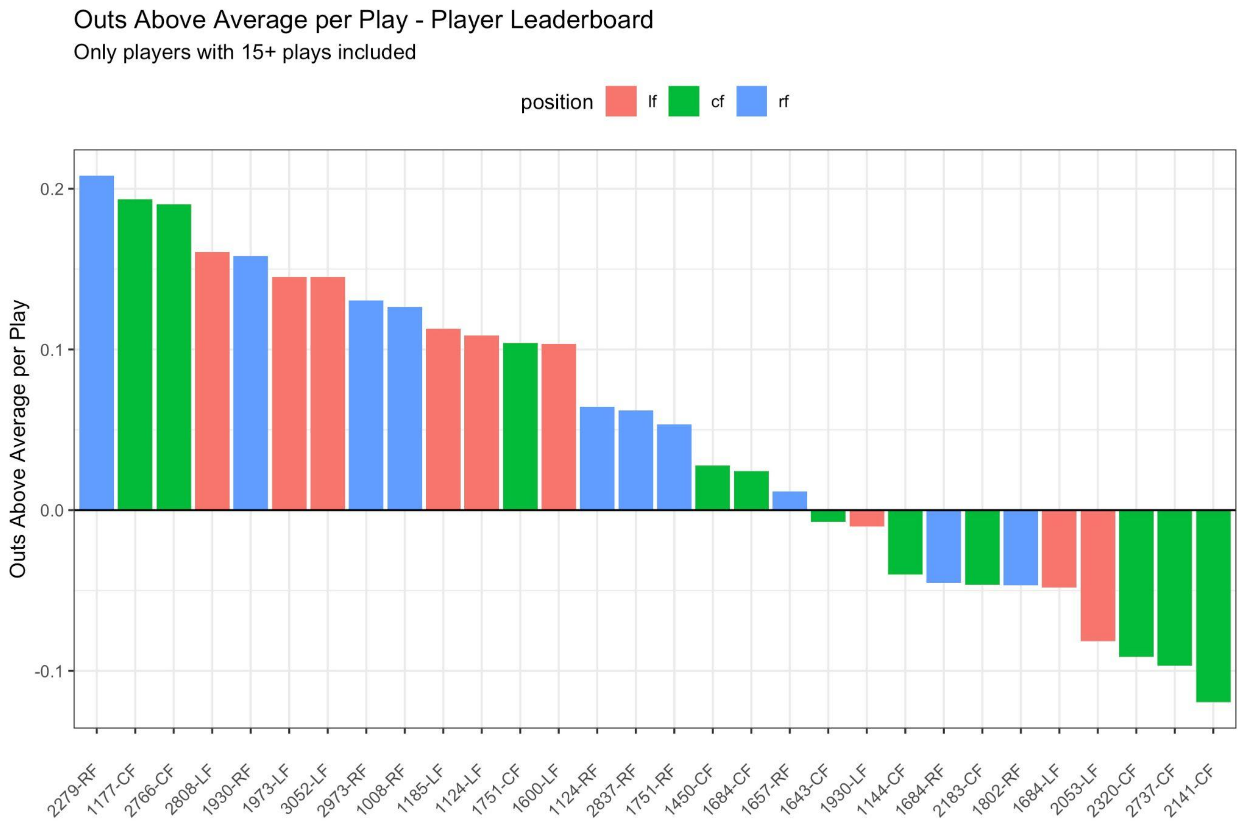


While the previous charts display out probability as a function of the location of the hit, there is another important variable to consider in the model: hit type. The boxplots below show the differences between line drives, fly balls, and pop-ups in scaled out probability, allowing for relative comparison between the different categories.

Of the three hit types, line drives have the lowest mean out probability, and their large cluster of points around zero indicates that a line drive hit to a challenging area of the field for outfielders to cover is practically a guaranteed hit. Fly balls were generally the most likely plays to result in outs, especially when hit in the area of an outfielder. Pop-ups, surprisingly, were not the most likely plays to result in outs; this is because pop-ups were generally hit to the shallow outfield and forced fielders to cover a greater distance to reach the ball.



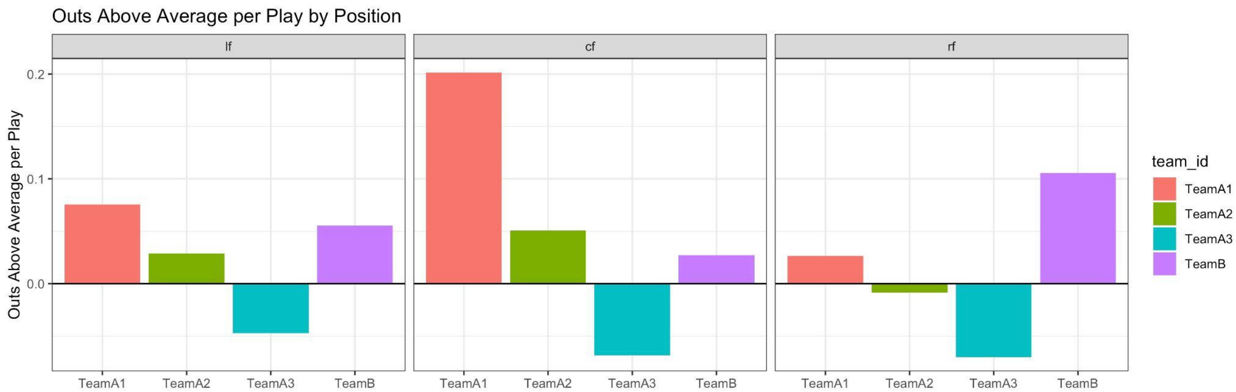
One useful interpretation of outs above average is the evaluation of individual fielders. The bar chart below showcases the outs above average per play for every fielder-position combination with at least 15 plays in the dataset.



This data is extremely useful for a few reasons. First, it allows for easy evaluation and comparison between outfielders. Second, it creates a very straightforward delineation between "good" and "bad" outfielders: players above zero recorded more outs than expected, and players below zero recorded fewer outs than expected.

It can also be used to evaluate an outfielder's play and relative skill at different positions. Sometimes, outfielders are consistently good or bad no matter where they play: for example, fielder 1124 recorded 0.11 OAA per play in left field and 0.06 OAA per play in right field. However, some players excel at one position but might struggle at another spot. Fielder 1930 dominated in right field, recording an elite 0.16 OAA per play, but struggled in left field where he recorded -0.01 OAA per play. However, he spent more time in left field than right field. Using OAA to evaluate outfielders can ensure that a team's outfield corps is placed in optimal positions to succeed.

The usefulness of this statistic goes beyond just the player level. The bar graphs below display the performance of each of the four main teams included in the data: A1, A2, A3, and B.



These charts give an at-a-glance summary of a team's outfield defensive performance during the games included in the dataset. Team A1 had solid outfielders across the board, bolstered by its elite center field play. Teams B and A2 appear to be average, with the exception of Team B's right fielder. Team A3 had below average fielding at all three positions.

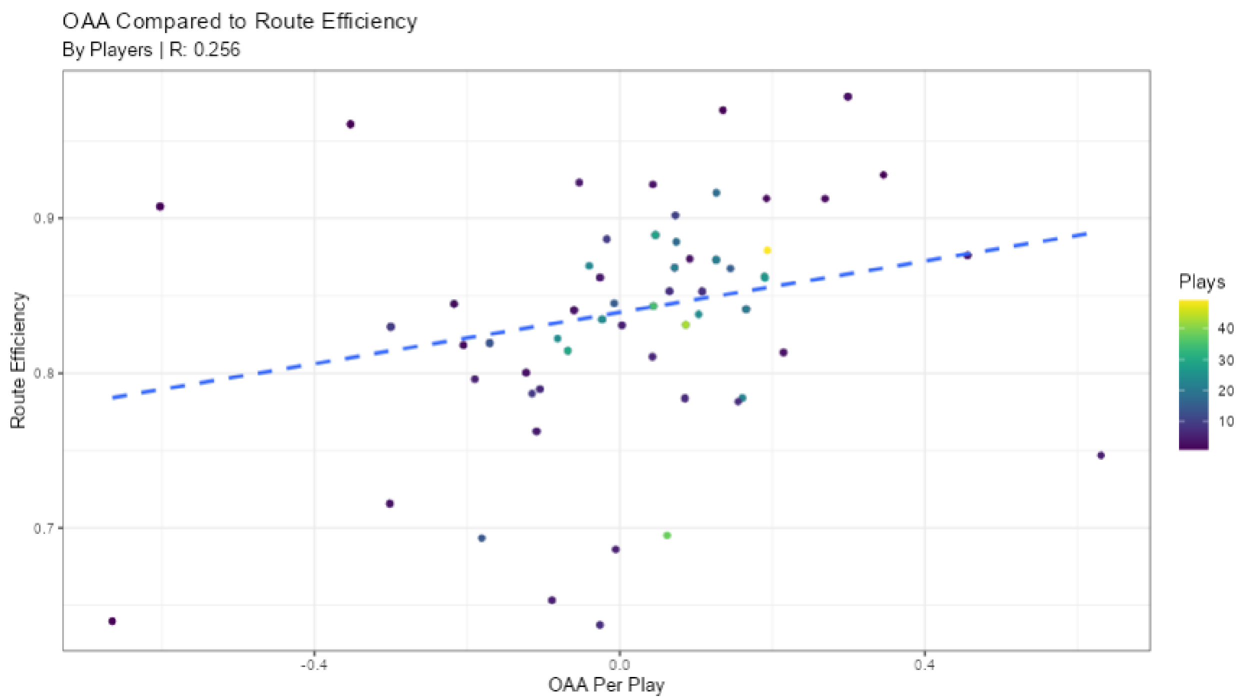
Analyzing data at the team level also demonstrates which teams are ahead of or behind the curve in positioning their fielders. The model assumes that each fielder starts at their standard position on every play; a team that routinely shifts their fielders to areas of the outfield closer to where the ball is hit will consistently outperform outs above average. Outs above average is a composition of two key fielding factors: an outfielder's post-contact ability to react to and break on a batted ball, and a team's pre-contact ability to put their fielders in optimal positions to record outs before the pitch is thrown.

For purely analyzing individual performance, an outs above average model considering the fielder's starting position instead of the average starting positions of each fielder would be more accurate to individual skill, and this model would be easily producible. However, at the team level, this version of outs above average allows decision-makers to compare their team to other teams in terms of how effective their pre-pitch outfield shifts are. The flexibility of the outs above average model as a tool for both player-level and team-level evaluation makes it uniquely valuable for decision making.

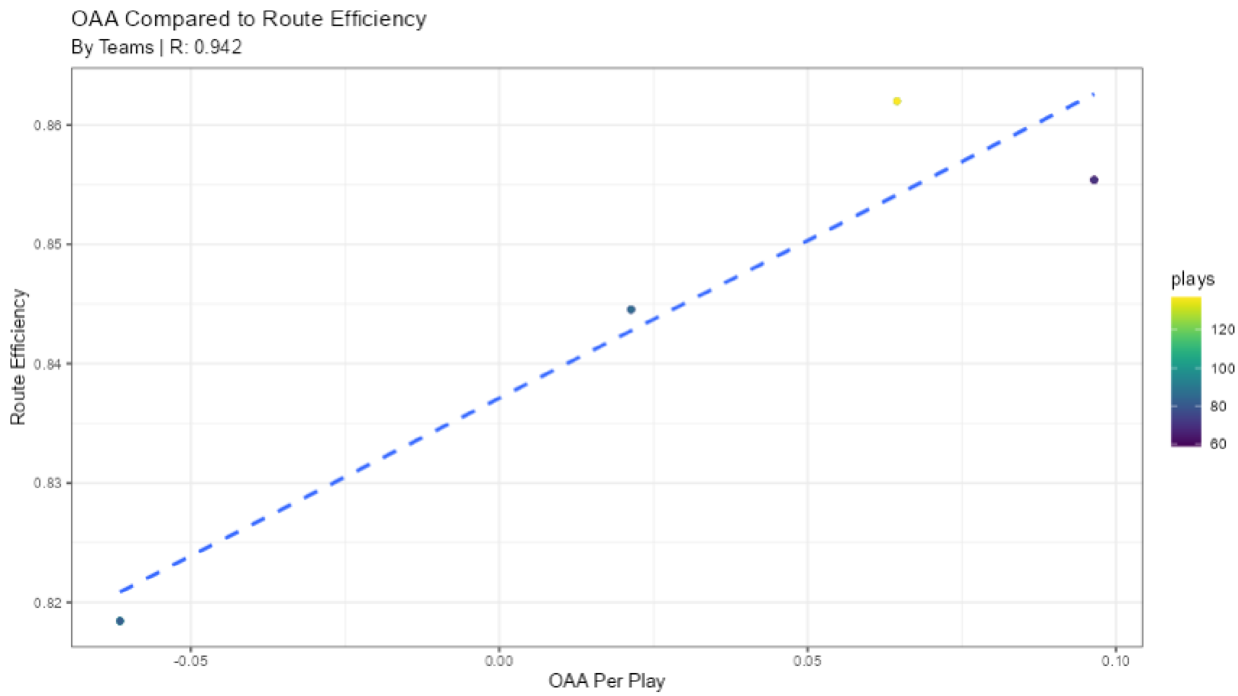
V. Comparisons

Route efficiency and outs above average are two separate measures of outfielder skill, but intuitively they should be positively correlated: outfielders who take more efficient routes to the ball should record more outs than less efficient players. At both the player and team level, the two statistics demonstrate this correlation.

At the player level, among players with 10 or more outfield chances in the dataset, the data indicated a moderate positive relationship between the two variables.



At the team level, outs above average and route efficiency demonstrate an even stronger correlation. Though the sample size is quite small (only 4 teams played in more than 3 games included in the dataset, and only those 4 teams are included in the chart), the two metrics appear to be related.



A strong relationship is evidenced between OAA and route efficiency for these four teams as seen above. Of course, more data would be necessary to truly prove the relationship but we get a good idea of how route efficiency and OAA relate to each other.

The two stats provide a valuable framework for evaluating fielding performance on their own, but when used in conjunction a deep analysis of fielding performance can be created. What cannot be explained by route efficiency can usually be handled by OAA, and route efficiency serves as a great validator of OAA, given how important it is in creating outs. While there are still small gaps in capturing overall fielder performance such as arm strength, these two metrics do most of what is necessary, especially for minor leaguers. Not only do we get a concrete look at how outfielders perform, but we are also able to analyze performance on various batted ball types for all fielders. In an ideal world, more steps could be taken but this framework is extremely important as our understanding of fielding evaluation continues.