A Survey of Methods for the Statistical Evaluation of Defensive Ability in Major League Baseball

Ben Baumer∗  Andy Galdi†  Robert Sebastian‡

Abstract
We review the advances in evaluating defensive ability among Major League Baseball players, from the early part of the previous century to the most recent developments. Most early attempts to measure individual fielding ability were hampered by a lack of detailed data, but the current availability of such data has led to several interesting evaluation metrics. We provide a generic mathematical model sufficient to describe each of them, and explicate the most prominent traditional, discrete, and continuous metrics. We assess the strengths and weaknesses of each, compare their results, and offer a brief glimpse into the future of defensive evaluation.

Key Words: baseball, fielding, defense, bayesian, hierarchical model, survey

1. Introduction and Motivation

“There is nothing on earth anybody can do with fielding.”
Branch Rickey, Brooklyn Dodgers General Manager, 1954

It was obvious to even the earliest observers that there were real and important differences in fielding ability among Major League players [10]. However, it was less obvious how to quantify those differences. Unlike the confrontation between the batter and the pitcher, for which the outcomes are relatively discrete and the conditions relatively uniform, the interaction between a ball put into play and the defensive players is particularly complicated. The playing surface is two-dimensional, and the distances and heights of the outfield walls are irregular between stadia. Tracking the starting position and movement of the ball and fielders, is difficult and usually requires subjective human observation. But perhaps most importantly (and most intractably), the inherent interaction between the nine fielders makes it particularly challenging to measure an individual player’s ability or performance.

While the statistical evaluation of fielding ability has become more mainstream in recent years [2], the aforementioned complications have led to much frustration. In this paper, we review the most prominent metrics for evaluating the fielding ability of individual players. In Section 2, we present a generic mathematical model capable of describing each of the existing fielding metrics, which we explore in Section 3. These metrics can be grouped into three major categories:

1. Section 3.1: Traditional metrics that make use of only the most basic and commonly recorded fielding statistics

2. Section 3.2: Discrete metrics that use advanced data sets but model the probability of a successful play using discrete bins

∗CUNY Graduate Center, 365 Fifth Avenue, New York, NY 10016. New York Mets, Flushing, NY 11368.
†New York Mets, Flushing, NY 11368.
3. Section 3.3: Continuous metrics that use advanced data sets and model the probability of a successful play as a continuous surface.

In Section 4, we compare these metrics, and briefly discuss the implications of more advanced forthcoming data. Section 5 is an appendix that briefly details the four major sources of play-by-play data.

1.1 Defensive Skills

A player’s defensive ability requires some combination of four largely independent skills:

- Sure-handedness: the ability to field balls cleanly when hit or thrown to you
- Range: a measure of the size of the area over which you can reach balls
- Positioning: how well you anticipate where balls will be hit before the pitch
- Throwing Ability: the ability to make strong, accurate throws

A player’s ability to play different fielding positions will vary based on his particular combination of skills. For example, the fielding positions which are commonly considered to be the most demanding, such as shortstop and centerfield, require an array of skills. Less demanding positions (like first base), may depend heavily on only one of these skills (sure-handedness), while being much more forgiving towards some others (throwing). We will see that traditional metrics primarily measure sure-handedness, while more advanced metrics focus more on range. Jensen & Carruth have outlined a framework for evaluating throwing ability, which as of yet, has not been incorporated into any of the metrics we discuss [1].

2. Generic Mathematical Model

The goal of this section is to present a generic mathematical framework through which we can express each of the relevant defensive metrics, so that they can be compared directly.

Let \( U \) be the set of all batted balls hit into fair territory. If \( k \in \mathbb{Z}_0 \) is a defensive position, then we say that the probability that \( u \in U \) is fielded successfully (i.e. - turned into an out) by the fielder playing position \( k \) is \( \mu_k(u) \). Since a ball can be fielded successfully by only one player, it follows that the overall probability of an out is \( \mu(u) = \sum_k \mu_k(u) \), and thus the probability of a non-out (most likely a hit, but also possibly an error) is \( 1 - \mu(u) \).

If \( I \) is the set of all fielders, then we are interested not only in estimating \( \mu_k(u) \), but also \( \mu_{ik}(u) \), which is the probability that fielder \( i \) will make a successful fielding play on batted ball \( u \) while playing position \( k \). Let \( \chi_k(u) \) be the identity of the fielder playing position \( k \) at the time that batted ball \( u \) was hit, so that the condition \( \chi_k(u) = i \) implies that fielder \( i \) was playing position \( k \) when batted ball \( u \) was hit. It will be convenient to refer to subsets of balls in play that are relevant to a particular fielder, such as \( U_{ik} = \{ u \in U : \chi_k(u) = i \} \). Note that this framework allows us to differentiate between \( \mu_{ik}(u) \) and \( \mu_{ik'}(u) \), for \( k \neq k' \), since many fielders play multiple positions.

---

\[ ^1 \text{By “fielded successfully,” we generally mean handled by that fielder and converted into at least one out.} \]
We assume that each batted ball in our data set represents an observation of a random variable
\[ p_{ik}(u) = \begin{cases} 1 & \text{if } u \text{ is fielded successfully by player } i \text{ while playing position } k \\ 0 & \text{if } u \text{ is not fielded successfully by player } i \text{ while playing position } k \end{cases} \]
We assume furthermore that the observations \( p_{ik}(u) \) come from an underlying Bernoulli distribution with mean \( \mu_{ik}(u) \). There are a number of covariates depending on \( u \) that may affect the probability of a successful fielding play. For example, the location of \( u \) clearly affects the probability that it will be fielded successfully by a specific fielder. We suppose then that there exists a function \( l : U \rightarrow \mathbb{R}^2 \) that gives the location where the batted ball \( u \) was fielded or landed in either rectangular \((x(u), y(u))\) or polar \((\rho(u), \theta(u))\) coordinates. Depending on the data source\(^2\), we may have access to functions that contain any of the following additional information, which may or may not affect \( \mu_k(u) \):

- **Trajectory** \( \tau(u) \): A categorical variable specifying the initial trajectory of the batted ball. Flyballs, grounders, line drives, pop-ups, and bunts are included in all data sets.
- **Velocity** \( v(u) \): An ordinal variable estimating the initial force of the batted ball. Typically, a soft, medium, or hard scale is employed.
- **Ballpark** \( \beta(u) \): The ballpark in which the game was played. The irregularity of outfield wall heights and distances may cause \( \mu_k(u) \) to vary. It is easy to imagine that in a highly irregular park, such as Fenway Park in Boston, locations may exist where the leftfielder has no chance to ever catch a ball, and yet in other parks such plays would be routine.
- **Batter & Pitcher Handedness** \( lhb(u), lhp(u) \): Most hitters are much more likely to “pull” the ball towards their strong side (e.g., right-handed hitters hit the ball to the left side of the field more often), especially on groundballs.
- **Base-Out Situation** \( b(u), o(u) \): The configuration of any baserunners and the number of outs in the inning.

Lastly, we will assume that there exists a function \( r : U \rightarrow \mathbb{R} \) that assigns an expected run value to each batted ball in play.

### 3. Review of Existing Models

#### 3.1 Traditional Metrics

Traditional attempts at measuring the defensive abilities of individual players was severely hampered by a lack of appropriate data. In fact, until fairly recently, only a handful of defensive statistics were commonly recorded:

- **Assists (A)**: Credited to a fielder who throws the ball to another fielder in a process leading to an out
- **Putouts (PO)**: Credited to a fielder who tags or forces a runner out

\(^2\)See the Appendix (Section 5) for a fuller description of the popularly available data sets.
• Errors \((E)\): Credited to a fielder who fails to make a play that should have been made (in the official scorer’s opinion) with “ordinary effort.”

Given these constraints, the concept of Total Chances, set equal to the sum of the three quantities listed above, was the best estimate of the number of fielding opportunities each fielder had.

### 3.1.1 Errors and Fielding Percentage

With these limitations in place, the first and perhaps most natural estimate of fielding ability is Fielding Percentage (FPCT):

\[
\text{FPCT}_k(i) = \frac{A_{ik} + P O_{ik}}{A_{ik} + P O_{ik} + E_{ik}} = 1 - \frac{E_{ik}}{T C_{ik}}
\]

This simple statistic continues to wield significant influence in baseball even today. The benefits of Fielding Percentage are that it can be calculated without detailed play-by-play data, and that it has an obvious interpretation. Unfortunately, that interpretation is frequently misconstrued! In terms of the defensive skills that we identified above, fielding percentage measures sure-handedness and throwing ability, but neither range nor positioning. Furthermore, those measurements are based on a subjective assessment (e.g. - errors).

It would be more accurate to think of Fielding Percentage as the conditional probability of fielding a ball successfully, given an opportunity requiring “ordinary effort.” Thus, Fielding Percentage does not reward, and may even penalize, more athletic fielders who cover more ground. For example, consider a shortstop who always fields balls hit within 10 feet of him successfully. His fielding percentage is a perfect 1. On the other hand, a shortstop who is equally sure-handed on balls hit within 10 feet of him, but who also converts some balls hit more than 10 feet from him into outs has greater range. However, if he makes even one error on these “extra” balls, he will have a fielding percentage of less than 1. It is clear that the latter shortstop is more valuable, since he will convert more balls into outs. However, by virtue of his errors on balls the other shortstop could not even reach, he will have a lower fielding percentage.

### 3.1.2 Range Factor

Dissatisfaction with Fielding Percentage led James to create Range Factor (RF) in 1976. As the name indicates, the idea was to measure range by estimating the number of outs that each player made per 9 innings. Thus,

\[
\text{RF}_k(i) = 9 \cdot \frac{A_{ik} + P O_{ik}}{I P_{ik}}
\]

where \(I P_{ik}\) is the number of innings played by player \(i\) at position \(k\). Unfortunately, Range Factor is confounded by many variables outside of an individual fielder’s control. To name a few, the ball in play rate of the team’s pitching staff, their handedness, ballpark effects, and the range of nearby fielders affect Range Factor. Simple calculation reveals that among teams from 2003-2008, the correlation coefficient between the Range Factor of shortstops and their pitching staff’s ball in play rate was 0.62.
3.1.3 Defensive Efficiency (DER)

Defensive Efficiency Rating is a team level statistic invented by James around the same time that is both meaningful and simple to calculate. It is simply the observed percentage of balls in play that are fielded successfully by that team. Thus, we define for each team $T$:

$$\text{DER}(T) = \overline{p}_T = \frac{1}{|U|} \sum_{u \in U} \sum_{k=1}^9 \sum_{i \in T} p_{ik}(u)$$

Note that Defensive Efficiency implicitly measures all four of the skills that we outlined above, since the ultimate goal of any defense is turning balls in play into outs. Unfortunately, it is not possible to apply this formula to individual players without access to detailed play-by-play data.

3.2 Discrete Metrics

With the availability of detailed play-by-play data in the 1970s and 1980s, it became possible to apply the concept behind Defensive Efficiency to individual players. Each of the metrics detailed in this section works by discretizing the playing surface and using data binning to estimate a baseline probability of each ball $u$ being fielded successfully. While these metrics employ very similar discrete models, they differ in two major aspects: 1) the manner in which their data is binned; 2) the formulas they use to aggregate the results. It will be convenient to assume that for each metric $m$, there exists a binning function $f_m$ that assigns each batted ball $u \in U$ to a unique data bin. We will slightly abuse the language by referring to the number of parameters (or covariates) that go into a bin definition. That is, if for some metric $m$, the binning function $f_m$ uses the two-dimensional location coordinates and trajectory of each batted ball $u$ to determine bin assignment, then we can think of $f_m(u)$ as being $f_m(x(u), y(u), \tau(u))$ and say that $f_m$ is defined by three parameters.

For each metric $m$, we let $[u]_k = \{v \in U : f_m(v) = f_m(u)\}$ be the set of all balls in play that are assigned the same data bin as $u$. In this section, we review the four most prominent discrete models in evolutionary order.

3.2.1 Zone Rating (ZR)

The first popular fielding metric to make use of play-by-play data Zone Rating, was created by Dewan in 1990 [4]. Zone Rating assigns responsibility for a batted ball $u$ to a specific fielding position $k$, if across the league, the fielders at position $k$ fielded at least 50% of the balls that were assigned to $f_{ZR}(u) = [u]$. The bins in Zone Rating are defined by the two-dimensional polar location coordinates and trajectory of each batted ball, and thus there are three parameters $(x(u), y(u), \tau(u))$ used by its binning function. Each fielder’s Zone Rating is simply the percentage of balls he has fielded in his zone of responsibility, with additions made for balls that he has fielded outside of his zone. Note that not every batted ball falls into a zone of responsibility.

Thus, there exist mutually disjoint sets of bins $[u]_k$ for each fielding position $k$. Let

$$U_{ik} = \{u \in U : \chi_k(u) = i \text{ and } (f_{ZR}(u) = [u]_k \text{ or } p_{ik}(u) = 1)\}$$

be the set of all balls put into play while player $i$ was playing position $k$ that were either in his zone of responsibility or were fielded successfully by him. Then player
$i$’s Zone Rating is simply

$$ZR_k(i) = \frac{1}{|U'_{ik}|} \sum_{u \in U'_{ik}} p_{ik}(u)$$

The major drawback of Zone Rating is that balls hit outside of the predefined zones are not counted unless they are fielded successfully. Though the effect is less perverse than it is for Fielding Percentage, a player with sure hands but limited range can still achieve a higher rating than a player with greater range but less sure hands. Secondly, because there is no estimation of how difficult a ball is to field, all successful fielding plays are considered equally difficult.

### 3.2.2 A Probabilistic Model of Range (PMR)

A metric that measures all balls in play, called A Probabilistic Model of Range, was proposed by Pinto [9] in 2003. The idea was simply to apply a Defensive Efficiency Rating to each player, and then compare it to the league average at that position. The binning function for PMR uses six parameters: 1) Direction $\theta(u)$; 2) Trajectory $\tau(u)$; 3) Velocity $v(u)$; 4) Ballpark $\beta(u)$; 5) Pitcher Handedness $lhp(u)$; and 6) Batter Handedness $lhb(u)$. Note that in this case the location coordinates are one-dimensional, and measure only the slice of the field through which the ball traveled (but not how far).

The actual number of balls fielded by player $i$ while playing position $k$ is $\sum_{u \in U_{ik}} p_{ik}(u)$. For any $u \in [u]$, the observed probability of a league-average fielder playing position $k$ fielding $u$ successfully is:

$$\tilde{p}_k(u) = \frac{1}{|[u]|} \sum_{v \in [u]} \sum_{i \in I} p_{ik}(v)$$

This is the most natural estimate of $\mu_{ik}(u)$.

Pinto’s original definition of PMR is then the simple ratio of the actual number of balls fielded to the expected number of balls fielded:

$$PMR_k(i) = \frac{\sum_{u \in U_{ik}} p_{ik}(u)}{\sum_{u \in U_{ik}} \tilde{p}_k(u)}$$

and as such a league average defender will have a PMR of 1.

However, for consistency, we will discuss a popular extension of PMR that is measured on the scale of runs saved relative to a league-average defender. To do this, we need a function $\hat{r}: U \to \mathbb{R}$ that estimates the run value $r$ of each ball in play. With this in tow, we will refer to PMR as:

$$PMR_k(i) = \sum_{u \in U_{ik}} \hat{r}(u) \cdot [p_{ik}(u) - \tilde{p}_k(u)]$$

That is, $PMR_k(i)$ represents the estimated number of runs saved by player $i$ while playing position $k$ relative to a league-average defender. Under the assumption that each fielder is of league-average ability, each player is credited with the residual between his actual performance and his expected performance for each ball in play.

It is easy to verify that PMR has several desirable properties. Namely, for a fixed $k$, the mean of $PMR_k(i)$ over all players is 0, as is the mean $PMR_k(i)$ over all $k$ and $i$ simultaneously. Thus, PMR “zeros out” over several different groupings.
It should be noted that in PMR, the “credit” (or “debit”) that a player receives for each ball in play depends explicitly on $k$. However, it is not immediately obvious that his rating depends on the performance of his teammates. We will see that this dependence is implicit when we explore the “ball-hogging” quandary.

One of the practical limitations of PMR is that due to the large number of parameters, its binning function creates many bins, and thus unless the data set covers many seasons, many of the bins contain very few balls. For example, in a data set with six years worth of data and almost 800,000 balls in play, the average bin size in our implementation of PMR was just 20, and only 23% of the balls fell into bins with at least 100 balls in them.

### 3.2.3 Ultimate Zone Rating (UZR)

Ultimate Zone Rating was developed by Litchman [8] prior to the publication of PMR, but it is most instructive to view it as an extension of that system. In his original model, the bins were defined by each unique combination of six parameters: 1) 2D-Location $(\rho(u), \theta(u))$; 2) Trajectory $\tau(u)$; 3) Velocity $v(u)$; 4) Outs $o(u)$; and 5) Batter Handedness $hbb(u)$. Note that unlike PMR, UZR bins are defined using two-dimensional location coordinates\(^3\). The UZR bin definition function results in substantially fewer bins than PMR. On the same data set mentioned above, 92% of balls fell into bins containing at least 100 balls in play. As such, it is likely that the estimates of $\mu_k(u)$ obtained under $f_{UZR}$ are more robust than those obtained via $f_{PMR}$.

On the other hand, $f_{UZR}$ does not take confounding factors such as ballpark into account, and thus Lichtman makes subsequent adjustments for ballpark, baserunner configuration, and the pitcher’s groundball-to-flyball ratio. For example, let $\epsilon_\beta$ be a global adjustment factor for each ballpark $\beta$. Lichtman estimates the inflation in the run environment at ballpark $\beta$, and then uses this estimate to improve his estimate of $\mu_k(u)$ by dividing it by $1 + \hat{\epsilon}_{\beta(u)}$\(^4\). Adjustments for situation and pitcher groundball/flyball tendencies were executed similarly\(^5\).

However, the estimates of $\mu_k(u)$ are not the only difference between UZR and PMR. Credit for successful plays made in UZR is independent of position, while debit is position-dependent. The result is that a player’s rating may depend quite explicitly on the performance of his teammates. Specifically, player $i$’s UZR at position $k$ is:

$$UZR_k(i) = \sum_{u \in U_{ik}} \hat{r}(u) \cdot \left[ \frac{\text{credit for plays made}}{\text{debit for plays not made}} \right] = \sum_{u \in U_{ik}} \hat{r}(u) \cdot \left[ p_{ik}(u) \cdot (1 - \bar{p}(u)) - (1 - p_i(u)) \cdot \bar{p}_k(u) \right]$$

where $\bar{p}(u) = \sum_k \bar{p}_k(u)$ is the observed probability of a ball in $[u]$ being fielded successfully (by anyone; i.e. - our best estimate of $\mu(u)$), and $p_i(u) = \sum_{k \in Z_9} p_{ik}(u)$ is 1 if $u$ is fielded successfully (by anyone, but with $i$ playing $k$) and 0 otherwise.

\(^3\)In fact, this is only true for flyballs, but we’ll ignore this wrinkle for simplicity.

\(^4\)Lichtman actually splits his ballpark adjustments into regions. For example, the ballpark adjustment for left field in Fenway Park is far more severe that it is for centerfield.

\(^5\)It appears as though the posterior adjustments that Lichtman makes improve the accuracy of his estimates, but it would interesting to examine more rigorously whether the gains in accuracy are justified relative to the increase in measurement error.
Perhaps a more suggestive way of writing UZR is then:

\[ UZR_k(i) = PMR_k(i) + \sum_{u \in U_{ik}} \hat{r}(u) \cdot w_{ik}(u) \]

where we call the quantity \( w_{ik}(u) = p_i(u) \cdot \bar{p}_k(u) - \bar{p}(u) \cdot p_{ik}(u) \) the “ball-hogging correction.”

Note that the units of UZR are runs saved relative to a league-average defender\(^6\).

The Ball-Hogging Correction

The purpose of the ball-hogging correction, which we have isolated above as \( w_{ik}(u) \), is to share credit and blame for plays that likely could have been made by either fielder. Note that for any bin \([u]\):

\[ \sum_{u \in [u]} w_{ik}(u) = 0 \rightarrow \sum_{u \in [u]} \frac{p_{ik}(u)}{\bar{p}(u)} = \frac{\bar{p}_k(u)}{\bar{p}(u)} \]

That is, the ball-hogging correction for player \( i \) playing position \( k \) in bin \([u]\) is 0 if the percentage of plays made by position \( k \) is the same when player \( i \) is on the field as it is in general across the league. Otherwise, the ratings of players who make a disproportionate number of plays on their team (ball-hogs) are adjusted downwards, while the ratings of their teammates are adjusted upwards. The effect of this term is that fielders are not penalized for plays that are made by their teammates, but must share the responsibility for their collective failures (non-outs).

While the ball-hogging correction of UZR is heavily debated, it is straightforward to verify that it has many desirable properties. In particular, no one position is regularly penalized.

**Observation 1.** For any position \( k \), the mean of \( w_{ik}(u) \) is 0.

**Proof.** The sum of \( w_{ik}(u) \) over all players \( i \in I \) is:

\[ \sum_{i \in I} w_{ik}(u) = \bar{p}_k(u) \sum_{i \in I} p_i(u) - \bar{p}(u) \sum_{i \in I} p_{ik}(u) = \bar{p}_k(u) \cdot \bar{p}(u) - \bar{p}(u) \cdot \bar{p}_k(u) = 0 \]

**Observation 2.** The ball-hogging factor for a specific team is always 0.

**Proof.** It suffices to compute this for a given \( u \), which we denote as \( w_T(u) \). We then sum over \( i \in T \), the players on the team, and \( k \). Thus,

\[ w_T(u) = \sum_{i \in T} p_i(u) \left( \sum_{k=1}^9 \bar{p}_k(u) \right) - \bar{p}(u) \sum_{k=1}^9 \sum_{i \in T} p_{ik}(u) = 0 \]

**Corollary 1.** The ball-hogging factor for the entire league is always 0.

However, the addition of the ball-hogging correction also leads to some undesirable consequences.

---

\(^6\)UZR actually includes additional factors to measure the impact of outfield arms and double plays, but such things are beyond our scope.
Observation 3. A defender who makes the expected number of plays at his position can have a UZR below 0 if a teammate fails to make some plays, due to the ball-hogging correction.

Proof. Suppose that for player $i$ playing position $k$, $\frac{1}{|u|} \sum_{u \in [u]} p_k(u) = \bar{p}_k(u)$ for all bins $[u]$. That is, $i$ has made the expected number of plays for a fielder at his position in every bin, and thus $PMR_k(i) = 0$. However, suppose that there exists exactly one non-empty bin $[u]$ for which $\frac{1}{|u|} \sum_{u \in [u]} p_i(u) < \bar{p}(u)$. Then $w_{ik}(u) < 0$ for all $u \in [u]$ and it follows that $UZR_k(i) < 0$.

The intuition behind this quirk is that every player must absorb some of the blame for any ball put in play against his team that is not converted into an out, if it lands in a bin for which he has a non-zero probability of recording an out.

Illustration of the Ball-Hogging Quandary

A brief example will help to illustrate the ball-hogging quandary. Suppose that for some bin $[u]$ lying between the leftfielder and the centerfielder, $\bar{p}_8(u) = 0.5$, $\bar{p}_7(u) = 0.49$, and no other fielders make any plays, so that $\bar{p}(u) = 0.99$. Now suppose that teammates $a, b$ play an entire season next to one another, and their observed out rates on balls in $[u]$ are 0.7 and 0.29, respectively. Then as a pair, they have made exactly as many outs as a league-average tandem, and so their PMR is 0. But $PMR_8(a) = +20\hat{r}(u)$, while $PMR_7(b) = -20\hat{r}(u)$. When the probability of an out is so high, is it fair to penalize $b$ to this extent? Given the high out rate, it is likely that balls hit to this bin are easily catchable by either defender. It may the case that $b$ simply defers to $a$ out of convention, and is not actually a subpar fielder. In this “ball-hogging” example, $UZR_8(a) = +0.051\hat{r}(u)$ and $UZR_7(b) = -0.049\hat{r}(u)$, so that the effect is dampened by a factor of 10.

On the other hand, UZR can also produce counterintuitive results. Suppose that for a different pair of fielders, $c$ is observed to have caught 40% of the balls hit to $[u]$, while $d$ catches 49%. Then $PMR_8(c) = -0.1\hat{r}(u)$ and $PMR_7(d) = 0$, so that they have a combined rating of $-0.1\hat{r}(u)$, since they have made 10% fewer catches than the league-average. UZR gives the pair the same rating, but distributes the blame in proportion to the expected out rate for each position. In this case $UZR_8(c) = -0.051\hat{r}(u)$ and $UZR_7(b) = -0.049\hat{r}(u)$.

3.2.4 Plus/Minus Rating

In 2006, Dewan published a new fielding metric called the Plus/Minus System [3]. After its most recent update, the Plus/Minus System is very similar to UZR, except that it evaluates fielders using the probability of a successful play conditional upon the event that the play was not made by any other fielder. Thus,

$$ PM_k(i) = \sum_{u \in U_{ik}} \left( \frac{1}{1 - \bar{p}(u) + \bar{p}_k(u)} \right) \cdot UZR_k(i, u) $$

The latest update to the Plus/Minus rating also differs from UZR in the way that it adjusts for ballpark, certain configurations of baserunners, and the number of outs.
3.3 Continuous Models

3.3.1 Spatial Aggregate Fielding Evaluation (SAFE)

The latest and most statistically advanced fielding metric has been developed by Jensen, et al. [7]. SAFE is fundamentally different from the models we have discussed heretofore, in that it estimates the probability of a successful fielding play using a fitted continuous surface, rather than an observed average on a discrete bin. An illustration of the difference between the two approaches is shown in Figure 1.

We can outline the SAFE methodology in four steps. Each of the functions that we describe below are continuous for a fixed trajectory $\tau$ over 2D-location $(x,y)$ and velocity $v$. In the hopes of maintaining continuity with our existing mathematical framework, we let $u = (x,y,v)$ be a vector of these parameters.

1. Use probit regression techniques to fit a smooth probability model to the observed data. The model gives, for each position $k$ and each trajectory $\tau$, an estimate of the probability of a successful fielding play, given that no other fielder makes the play. This estimate is based on the three parameters in $u$. Thus, we obtain a smooth function $\hat{q}^{(\tau)}_k(u)$.

2. Perform this same operation for each player $i$, each position $k$, and each trajectory $\tau$, but use a hierarchical Bayesian model with a shared prior distribution. Specifically, use Gibbs sampling techniques to partially pool the results and improve the robustness of each estimate. The result is a function $\hat{q}^{(\tau)}_{ik}(u)$ giving the probability of a successful fielding play by player $i$ while playing position $k$, conditional upon no other fielder having made the play.

3. Estimate three functions for each ball in play from the observed data:

   (a) The relative frequency, $\hat{f}^{(\tau)}(u)$.

   (b) The run value, $\hat{r}^{(\tau)}(u)$.

Figure 1: Probability of a catch by a CF on medium velocity flyballs
Figure 2: Fitted probability surfaces for a league-average centerfielder and Darin Erstad (circa 2002) [7]

(c) The shared responsibility, $\hat{s}^{(}\tau\rangle(u)$. Specifically, let $\hat{s}^{(}\tau\rangle(u)$ be the percentage of all successful fielding plays made by the $k^{th}$ fielder on balls like $u$.

Figure 3: Estimated smooth functions for $\hat{f}$ and $\hat{r}$ [7]

4. For each $\tau$, subtract the result of step one from step two, and multiply that result by the functions in step three. Then integrate numerically over $x$, $y$, and $v$, giving:

$$SAFE_k^{\tau}(i) = \iint \hat{f}^{(\tau)}(u) \cdot \hat{r}^{(\tau)}(u) \cdot \hat{s}^{(\tau)}_k(u) \cdot \left[ q^{(\tau)}_{ik}(u) - q^{(\tau)}_k(u) \right] \, du$$

The final SAFE value is then the sum over all $\tau$.

The result is an estimate of the expected number of runs saved by player $i$ at position $k$ relative to a league-average defender. It is important to recognize that since SAFE
Table 1: A comparison of the major features of selected metrics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>PMR</th>
<th>UZR</th>
<th>+/-</th>
<th>SAFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set</td>
<td>Either</td>
<td>Either</td>
<td>BIS</td>
<td>BIS</td>
</tr>
<tr>
<td>Model</td>
<td>Discrete</td>
<td>Discrete</td>
<td>Discrete</td>
<td>Continuous</td>
</tr>
<tr>
<td>Infield LD &amp; Pop-Ups</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Location</td>
<td>1D</td>
<td>2D</td>
<td>2D</td>
<td>2D</td>
</tr>
<tr>
<td>Trajectory</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Velocity</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Base-Out</td>
<td>No</td>
<td>Yes</td>
<td>Some</td>
<td>No</td>
</tr>
<tr>
<td>Batter Handedness</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Pitcher Handedness</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Ballpark</td>
<td>Bins</td>
<td>Regions</td>
<td>Walls</td>
<td>No</td>
</tr>
<tr>
<td>Result</td>
<td>Plays Made</td>
<td>Runs Saved</td>
<td>Runs Saved</td>
<td>Exp. Runs Saved</td>
</tr>
</tbody>
</table>

uses the expected distribution of batted balls, rather than the actual distribution witnessed by a particular player, the result is an estimate of the expected fielding performance, rather than the observed fielding performance.

It is also worth noting that the regression model for balls hit in the air is different than it is for balls hit on the ground. In particular, the probability of a ball hit in the air being caught is modeled as a function of two-dimensional location \((x(u), y(u))\), velocity \((v(u))\), and direction (in front of or behind the usual starting position), and interaction between these coefficients is allowed. For ground balls, the probability of the ball being successfully fielded is modeled as a function of the angle \((\theta(u))\), velocity \((v(u))\), and direction (left or right of the usual starting position). Again, interaction between the coefficients is allowed. The estimated starting position is fixed for all fielders, and is estimated from the data as the centroid of all successful fielding plays made.

The major limitation of SAFE in its present incarnation is that no corrections are made for significant confounding variables, such as ballpark and situational positioning. A fuller discussion of the details of the SAFE methodology are beyond the scope of this paper, but are made quite explicit in [7].

4. Comparison and Conclusion

4.1 Comparison

A comparison of the major features of PMR, UZR, Plus/Minus, and SAFE is shown in Table 1.

A major obstacle in evaluating fielding metrics is the lack of an objective “gold standard” which we can use as a benchmark. However, a good metric should show consistent over time, and not deviate wildly from firmly-held prior beliefs. In Table 2, we can see that the spread of UZR is considerably narrower than that of PMR\(^7\). This is due to the fairly strong negative correlation between the ball-hogging correction and PMR. Note also that the high correlation between PMR and UZR is not unexpected, given their explicit linear dependence.

Finally, as shown in Table 3, the reliability of UZR, in terms of its year-to-year

\(^7\)PMR and UZR were taken from published results, and rescaled to PMR (UZR) per 4000 BIP. The data covers the years 2004-2007, and was restricted to those players with at least 1000 BIP at a particular position.
<table>
<thead>
<tr>
<th>Pos</th>
<th>N</th>
<th>$\sigma_{PMR}$</th>
<th>$\sigma_{UZR}$</th>
<th>$r_{PMR,UZR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1B</td>
<td>100</td>
<td>28.5</td>
<td>14.5</td>
<td>0.635</td>
</tr>
<tr>
<td>2B</td>
<td>101</td>
<td>38.3</td>
<td>24.8</td>
<td>0.709</td>
</tr>
<tr>
<td>3B</td>
<td>98</td>
<td>36.4</td>
<td>24.7</td>
<td>0.706</td>
</tr>
<tr>
<td>SS</td>
<td>108</td>
<td>34.9</td>
<td>22.4</td>
<td>0.637</td>
</tr>
<tr>
<td>LF</td>
<td>166</td>
<td>15.0</td>
<td>13.1</td>
<td>0.668</td>
</tr>
<tr>
<td>CF</td>
<td>166</td>
<td>17.0</td>
<td>15.1</td>
<td>0.614</td>
</tr>
<tr>
<td>RF</td>
<td>161</td>
<td>15.7</td>
<td>15.4</td>
<td>0.569</td>
</tr>
<tr>
<td>Total</td>
<td>900</td>
<td>26.2</td>
<td>18.3</td>
<td>0.650</td>
</tr>
</tbody>
</table>

**Table 2:** Spread and Correlation between PMR and UZR

<table>
<thead>
<tr>
<th>Pos</th>
<th>N</th>
<th>PMR</th>
<th>UZR</th>
<th>SAFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1B</td>
<td>48</td>
<td>0.429</td>
<td>0.222</td>
<td>0.287</td>
</tr>
<tr>
<td>2B</td>
<td>53</td>
<td>0.431</td>
<td>0.546</td>
<td>0.051</td>
</tr>
<tr>
<td>3B</td>
<td>50</td>
<td>0.327</td>
<td>0.410</td>
<td>0.503</td>
</tr>
<tr>
<td>SS</td>
<td>59</td>
<td>0.453</td>
<td>0.438</td>
<td>-0.030</td>
</tr>
<tr>
<td>LF</td>
<td>66</td>
<td>0.343</td>
<td>0.568</td>
<td>0.594</td>
</tr>
<tr>
<td>CF</td>
<td>76</td>
<td>0.359</td>
<td>0.595</td>
<td>0.525</td>
</tr>
<tr>
<td>RF</td>
<td>68</td>
<td>0.113</td>
<td>0.326</td>
<td>0.444</td>
</tr>
<tr>
<td>Total</td>
<td>420</td>
<td>0.351</td>
<td>0.444</td>
<td>0.372</td>
</tr>
</tbody>
</table>

**Table 3:** Year-to-Year Correlation, 2004-2007 (SAFE 2002-2005)

The correlation, was higher than that of PMR or SAFE\(^8\). However, while the reliability of SAFE seems severely compromised in the infield (especially among middle infielders), it was much higher in the outfield, where it compared favorably to UZR, and far surpassed PMR\(^9\).

### 4.2 The Future: HIT f/x

To this point, the starting positions and movements of the fielders, as well the path of each batted ball, have not been mechanically recorded. Rather, the location coordinates to which we have referred were observations made by human data collectors, based on the spot at which a batted ball landed or was fielded. The forthcoming HIT f/x data promises to change this by using a series of cameras to track the position of all fielders and the ball in three dimensions [11]. This data should fold seamlessly into a continuous model like SAFE. In fact, the SAFE model may actually be simplified, since the location, trajectory and velocity parameters (all of which are simply proxies to determine the difficulty of fielding the ball) could be replaced by three-dimensional location coordinates that were a function of the time $t$ since the crack of the bat. Secondly, the starting position of each fielder, which is currently estimated, could be replaced by its actual value. This would undoubtedly improve the fit of the regression model. However, other than cutting down on the number of parameters, it is not clear how much this new data would improve the discrete models, since they depend fundamentally on data-binning techniques.

\(^8\)PMR and UZR numbers were rescaled to per 4000 BIP, and were taken from the years 2004-2007. However, the SAFE data is from 2002-2005 and was not rescaled. This is the reason for the double line, as this comparison is not apples-to-apples.

\(^9\)Note that since SAFE uses the expected distribution of batted balls, it *should* be more reliable than UZR!
4.3 Summary and Conclusion

Traditional fielding metrics are severely limited by a lack of detailed data and the subjectivity of the error classification. However, several currently prominent discrete fielding models have made use of more complete play-by-play data. These models estimate the observed performance of individual fielders, and among them, Ultimate Zone Rating seems to give the best results. In particular, UZR makes comprehensive adjustments for confounding variables, and shows greater reliability. More recently, a statistically advanced metric (SAFE) based on a Bayesian hierarchical model has been developed, and while this continuous model is promising due to the strength and clarity of its construction, to this point it has not made adjustments for important confounding variables. Going forward, it seems that only this continuous model is poised to fully exploit the forthcoming truly continuous spatial data. It is unclear how much this new data will improve the discrete models.

4.4 Acknowledgements

This work would have impossible without the support of the New York Mets. Personal communications with Mitchel Lichtman, David Pinto, Ben Jedlovec, Shane Jensen, Dave Allen, and Greg Rybarczyk proved to be invaluable. Feedback from many of the above, as well as Tom Tango and Phil Birnbaum, was also helpful.

5. Appendix: Play-by-Play Data Sets

The four major suppliers of play-by-play data are:

5.1 Retrosheet

Retrosheet divides the field into irregularly sized zones based on the normal location and assumed responsibility of the fielders. Thus, the zone labeled 5 is the normal area of responsibility of the third baseman, while the zone labeled 78D lies between the left and center fielders, and deep. By our count there are 68 total zones in the Retrosheet database, of which 55 lie in fair territory.

5.2 STATS, Inc.

STATS, Inc. uses what are essentially polar coordinates, in that fair territory is divided into 22 slices, labeled $C$ through $X^{10}$. The location of each batted ball can then be specified in polar coordinates as $(r, \theta)$, where $\theta \in \{A, ..., Z\}$ is the direction and $r \in Z$ is the distance from home plate. A depiction of STATS, Inc.’s grid is shown in Figure 4b.

STATS, Inc. also includes rectangular $(x, y)$-coordinates for the location of batted balls, where each coordinate records the distance from the baseline in feet.

5.3 BIS

BIS also uses rectangular $(x, y)$-coordinates to describe the location of batted balls, but centers the grid at home plate and orients the field vertically. Thus, second base has the rectangular coordinates $(0, 127)$ in the BIS system, but $(90, 90)$ in the STATS, Inc. system.

---

10There are actually two slices in foul territory on either side, so there are a total of 26 slices labeled $A$ through $Z$. 

368
5.4 MLBAM

Major League Baseball Advanced Media has recently constructed a play-by-play data set that is similar to that of BIS.

References


