Measuring Baseball Defensive Value using Statcast Data

by

Drew Jordan

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Colin Rundel, Supervisor

_____
Sayan Mukherjee

_____
Brian Hare

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2017

Abstract

Measuring Baseball Defensive Value using Statcast Data

by

Drew Jordan

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Colin Rundel, Supervisor

_____
Sayan Mukherjee

_____
Brian Hare

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in the Department of Statistical Science in the Graduate School of Duke University
2017

Abstract

Multiple methods of measuring the defensive value of baseball players have been developed. These methods commonly rely on human batted ball charters, which inherently introduces the possibility of measurement error and lack of objectivity to these metrics. Using newly available Statcast data, we construct a new metric, SAFE 2.0, that utilizes Bayesian hierarchical logistic regression to calculate the probability that a given batted ball will be caught by a fielder. We use kernel density estimation to approximate the relative frequency of each batted ball in our data. We also incorporate the run consequence of each batted ball. Combining the catch probability, the relative frequency, and the run consequence of batted balls over a grid, we arrive at our new metric, SAFE 2.0. We apply our method to all batted balls hit to centerfield in the 2016 Major League Baseball season, and rank all centerfielders according to their relative performance for the 2016 season as measured by SAFE 2.0. We then compare these rankings to the rankings of the most commonly used measure of defensive value, Ultimate Zone Rating.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

AUC	Area Under Curve
BIP	Ball-in-play.
BIS	Baseball Info Solutions.
DRS	Defensive Runs Saved.
MLB	Major League Baseball.
ROC	Receiver Operating Characteristic
SAFE	Spatial Aggregate Fielding Evaluation.
UZR	Ultimate Zone Rating.

Acknowledgements

I'd like to thank my thesis supervisor Professor Colin Rundel. I would also like to thank Professor Sayan Mukherjee and Professor Brian Hare for being on my thesis committee. Without their guidance, I would not have been able to produce a paper of anything near the same quality.

1

Introduction

Since the early days of baseball history, baseball has been a game of numbers. Almost every facet of the game is meticulously recorded and tracked via a wide range of statistics. Players are largely valued by their statistics and teams are built with the statistics of their players in mind. Beginning around 1977 when Bill James released his first Baseball Abstract, the practice of evaluating player performance via statistics has become increasingly scientific and widespread throughout the sport. The scientific analysis of baseball statistics is referred to as "Sabermetrics". Sabermetrics entered the mainstream eye with the release of Michael Lewis' book *Moneyball* in 2003 and the subsequent production of the movie *Moneyball* in 2011 based on the book. Recently, Sabermetric research has expanded beyond the analysis of more traditional baseball statistics, such as batting average and on-base percentage, as the technology for collecting baseball data has improved. With the invention of the PITCHf/x system in 2006, Sabermetricians have been able to examine the exact trajectories of pitches, allowing for more granular analysis of the abilities of pitchers and hitters alike. Most recently, beginning in 2015, the Statcast system was implemented across Major League Baseball (MLB). The Statcast system is essentially an extension

of the PITCHf/x system to the entire field of play that tracks the movement of every player and the ball for every play. This system allows for even more granular analysis of baseball statistics and is a big area of focus in current Sabermetric research.

Historically, much of Sabermetric analysis has focused on the evaluation of pitchers and hitters and less on the evaluation of a player's performance defensively. This focus has been due in large part to the availability of data to evaluate both hitting and pitching. The nature of the one-on-one matchups between hitters and pitchers lends itself more conveniently to statistical analysis due to the concrete nature of the matchup outcomes. Defensive value, on the other hand, has been less simple to objectively quantify because of the difficulty in measuring the relative difficulty of defensive plays. Defensive performance has always been measured, but not until the inventions of Ultimate Zone Rating by Mitchel Lichtman and Defensive Runs Saved by John Dewan (Basco and Zimmerman, 2010) have more objective measures of players' defensive contributions been available. These metrics benefit from being measured in units of runs, which allows for direct comparison of a player's defensive performance to his offensive performance.

This paper aims to use the newly available Statcast data to measure the defensive value of fielders and compare these values to the existing metrics.

1.1 Existing Methods

As previously mentioned, the two prevailing metrics evaluating the defensive value of baseball players are Ultimate Zone Rating and Defensive Runs Saved. This section provides background information regarding the calculation of each metric, their similarities and differences, and their shortcomings. Additionally, we will discuss Jensen, Shirley, and Wyner's SAFE metric, which serves as the primary inspiration for the methodology used herein.

All three of these existing measures are derived using data collected by Baseball

Info Solutions (BIS), an independent baseball data collection firm. BIS collects ball-in-play (BIP) data via human charters who estimate the BIP location and hit speed through observation. The human charters classify each BIP as a ground ball, line drive, fliner (between a line drive and a fly ball), or fly ball. They also classify the hit speed as soft, medium, or hard contact.

1.1.1 Ultimate Zone Rating (UZR)

UZR is measured in runs prevented above or below an average fielder at a player's position in that player's league and year. The BIP data from BIS is gathered into distinct zones, that cover the entire baseball field, based on location, BIP type (groundballs, line drives, and fly balls), and estimated hit speed. An example of the distribution of zones on a baseball field is shown in Figure 1.1 (FederalBaseball.com, 2010). Using the previous six years of BIP data, the average value of the BIPs in which no outs are recorded in each zone is calculated. When a fielder makes a play, he gets credit for the average run value of a BIP in the appropriate zone adjusted by the difficulty of the play. Therefore, more difficult plays credit more value to the fielder. Likewise, failing to make an easy play debits value from the fielder. These calculations are made context neutrally, meaning that the number of outs in an inning and the number of baserunners on base are not used in the calculations (with the exception of a few specific situations). The general calculations for successfully or unsuccessfully fielded BIPs are

Value of a successful play

$$= (1 - P(\text{Out is recorded}))(\text{Run value of a hit in that zone}) \quad (1.1)$$

Value of an unsuccessful play

$$= P(\text{Out is recorded})(\text{Run value of a hit in that zone}). \quad (1.2)$$

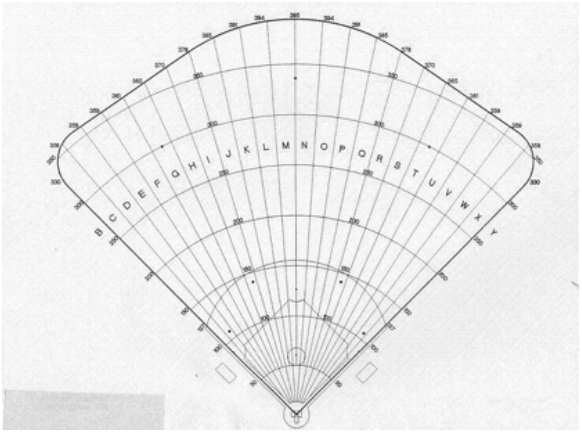

FIGURE 1.1: Distribution of UZR zones on a baseball field.

These values are then adjusted for particular base runner configurations, number of outs in an inning, specific infielder and outfielder positioning, and the park in which the play is made. To arrive at a UZR score for an individual player for a season, you simply add the values of all individual BIPs hit to that player for the entire season. UZR is commonly converted into a rate statistic by scaling it to 150 games played. UZR's year-to-year correlation is approximately 0.5. For a more in-depth explanation of UZR, see the "UZR Primer" on Fangraphs.com (FanGraphs.com, 2017a).

1.1.2 Defensive Runs Saved (DRS)

Like UZR, DRS is measured in runs saved above or below an average fielder at a player's position for a given year and is derived using BIP data provided by BIS. The data is then gathered into distinct groups based on the trajectory and velocity of the BIP and the angle in which the ball is hit with respect to home plate. The average run value of balls that are not fielded for each combination is then calculated using one season of BIP data (as opposed to UZR's six seasons). If at least one player made a play on a particular category of batted ball, the fielder will get credited or debited depending on whether or not they made that play. The calculations for the value assigned to the fielder on a play are identical to the calculations for

Table 1.1: UZR and DRS approximate ability tiers.

Defensive Ability	UZR/DRS
Gold Glove Caliber	+15
Great	+10
Above Average	+5
Average	0
Below Average	-5
Poor	-10
Awful	-15

UZR (Equations 1.1 and 1.2). DRS also makes adjustments for particular base runner configurations, number of outs in an inning, and defensive positioning. The biggest difference between UZR and DRS lies in the adjustments each makes in their calculations. Just like UZR, a player's DRS for a season is the sum of fielding outcome calculations for all batted balls hit in the area of the field that that player is responsible for defending. Table 1.1 provides a reference to classify a player's UZR or DRS score (FanGraphs.com, 2017c). For a more in-depth explanation of DRS see John Dewan's "Frequently Asked Questions about Plus/Minus and Runs Saved" (Dewan, 2012).

1.1.3 SAFE

SAFE, or Spatial Aggregate Fielding Evaluation, is a modeling based approach to measuring the defensive value of baseball players (Jensen et al., 2009). SAFE is interpretted in the same way as UZR and DRS as a measure of runs saved compared to the average player at the same position. It is also measured on a similar scale to UZR and DRS and utilizes the same data collected by BIS.

The backbone of SAFE is a Bayesian hierarchical probit regression model that models the probability that an out is recorded on a given BIP. Models are fit for each combination of BIP type (ground balls, line drives, fly balls), player position, and

season. The covariates in each model are the distance that the fielder runs to the BIP, the speed in which the BIP is hit, an indicator if the fielder is moving forwards or backwards to the BIP, and interactions between these covariates. The hierarchical structure of the models allows each player his own set of regression coefficients, while also sharing information across players. Players with fewer opportunities in a given season have their regression coefficients shrunk to a greater degree to the mean levels across all players.

For each player, SAFE is calculated by predicting the probabilities of successful plays over a grid of covariate values using that player's derived regression coefficients from the hierarchical model. Over the same grid of values, the probabilities of successful plays using the average population coefficients are predicted. The difference between these two sets of values is then calculated for each location on the covariates grid. This process computes a player's catch probability above or below average for each location on the grid. These differences are then multiplied by the relative frequency of each BIP on the grid as approximated via kernel density estimation. These values are then multiplied once more by the average run consequence of each BIP on the grid. Finally, the proportion of times that each given BIP is caught by the given player's position is estimated and multiplied with the previous values. SAFE is then the sum of these calculated values over the grid for each player. Players can then be arranged from highest SAFE to lowest SAFE for each position to arrive at a ranking of players at each position.

1.1.4 Limitations of Existing Methods

The major limitation of UZR, DRS, and SAFE is their reliance on human charters to record the nature of each BIP to determine the difficulty of each play made by a fielder. Additionally, the BIP data from BIS is not particularly granular in that hit velocity is only classified into three distinct bins (soft, medium, or hard). An addi-

tional shortcoming to UZR and DRS is the length of time it takes for them to become predictive of a player's true defensive talent level. According to Fangraphs.com's UZR page (FanGraphs.com, 2017c), both metrics stabilize after around three baseball seasons' worth of data for each player has been collected. In baseball, this is considered a long time and can prove to be prohibitive to accurately evaluate players. Finally, UZR and DRS are limited in that each player is only measured on BIPs that are hit to that player over the course of the season. This presents the possibility of bias due to lack of equality of opportunity across players. Some players may have more opportunities to make high-valued plays than others, which could potentially artificially inflate those players' UZR and DRS ratings. SAFE does not suffer from this same issue because of its modeling approach to measuring defensive value.

1.2 Proposed Method

We propose a new metric SAFE 2.0 that is calculated very similarly to SAFE. There are three major differences between our methodology and the methodology used to calculate SAFE. First, instead of using BIS's BIP data, we use the recently released Statcast data. The Statcast data is much more granular than BIS's data, measuring the exact BIP location, hit speed, and vertical hit angle on continous scales. Statcast data also does not rely on human charters.

Second, SAFE 2.0 is calculated using a Bayesian hierarchical logistic regression instead of a probit regression. This is done to improve the interprettability of the regression coefficients. Because Statcast data is very granular, we do not fit separate models for line drives and fly balls, but rather fit one model encompassing both BIP types. Because we have a different dataset than the dataset used to originally calculate SAFE, our model also differs in its inputs.

Third, because our data only covers balls that are hit to centerfield, we do not incorporate the estimated proportion of times that the ball is caught by other fielders

in our calculations.

Besides these three differences, SAFE 2.0 is calculated practically identically to the procedure described by Jensen, Shirley, and Wyner. Our goal is to apply the framework of their methodology to the much more precise and granular Statcast data. Our hope is that we will receive more accurate and objective rankings of the defensive contributions of MLB players.

2

Statcast Data

2.1 Background

The data we use in our analysis is collected by the Statcast tracking system. The Statcast tracking system has been in place in all 30 MLB stadiums since the beginning of the 2015 MLB season. According to MLB.com, the Statcast system collects tracking data of all of the players on the field and the baseball for every play of every game using a combination of high-resolution cameras and radar technology (MLB.com, 2017). The particular subset of Statcast data that we will analyze is all 17,578 line drives and fly balls that were hit to centerfield during the 2016 MLB season.

2.2 Variables Used in Analysis

There are four variables that are collected from the raw Statcast data and five variables that we derive from the raw Statcast variables used in our analysis.

2.2.1 Raw Statcast Variables

The four variables from the raw Statcast data are as follows:

1. **Hit Speed** - The velocity in miles per hour in which a ball is hit by a batter.

2. **Hit Angle** - The vertical angle with the horizontal in which a ball is hit by a batter.

3. **X and Y Coordinates** - The coordinates where a BIP is caught or where it lands.

4. **Fielder** - The player who fields the BIP.

2.2.2 Derived Statcast Variables

The five variables derived from Statcast data are as follows:

1. **RunValue** - The average run value associated with each BIP outcome for the 2016 MLB season. These values are derived using play-by-play data from all MLB games for the 2016 season. The basis for these values comes from the expected number of runs that will be scored in the remainder of a half inning given the current base/out state. We will refer to the average number of runs that will be scored in the remainder of a half inning as the Run Expectancy of that half inning. "Base/out state" refers to the location of runners on base and the number of outs in a half inning. There are 24 possible base/out states in which the batting team can continue batting in a half inning. The values in the Table 2.1 are calculated by taking the Run Expectancy of the half inning given the base/out state for the 2010-2015 MLB season as displayed on Fangraphs.com (FanGraphs.com, 2017b).

   For example, with runners on first base and second base with 1 out, the Run Expectancy from that base/out state to the end of the half inning is .884, meaning that we expect .884 runs to score in the remainder of the half inning. To calculate the run value of a BIP, we take the Run Expectancy of that half

Table 2.1: Run Expectancy for each base/out state from 2010-2015.

	0 Outs	1 Outs	2 Outs
_,_,_	0.481	0.254	0.098
1B,_,_	0.859	0.509	0.224
_,2B,_	1.100	0.664	0.319
1B,2B,_	1.437	0.884	0.429
_,_,3B	1.350	0.950	0.353
1B,_,3B	1.784	1.130	0.478
_,2B,3B	1.964	1.376	0.580
1B,2B,3B	2.292	1.541	0.752

Table 2.2: Run Values for BIP Outcomes

Outcome	Run Value
Out	0.00
Single	0.72
Double	1.02
Triple	1.29

inning before the BIP event and subtract it from the Run Expectancy after the BIP event. To find the run value of a BIP event over the course of a season, we take the average of these differences for each class of BIP event. We then adjust the run values of the different BIP events to the value of an out. This sets the baseline for the BIP events to zero, the adjusted run value of an out. In Sabermetric analysis, these run values are referred to as Linear Weights (FanGraphs.com, 2017b). The run values of the BIP events that occur in our analysis are presented in Table 2.2.

To derive the value of fielder errors in the dataset, we assign the value of the BIP outcome associated with the base in which the batter finishes the play. For example, if a fielder makes an error on a play and the batter ends up on second base, the play is valued the same as if the batter hit a clean double.

This procedure penalizes fielders appropriately when they make errors.

2. **Air Time** - Hangtime of each BIP from the point it makes contact with the bat to the point it hits the ground. This variable was created using simple projectile motion equations inputting the Hit Speed and Hit Angle variables for each BIP.

3. **Distance Ran** - The approximate distance the fielder had to travel to make a play on the BIP. Because the exact starting location of the fielder is not contained within the data, the starting location has to be approximated. For every play, the starting location is the most central point of all BIPs in the dataset that were turned into outs. Distance Ran is the Euclidean distance between this constant starting location and the coordinate of the location the play was made by the fielder.

4. **BIP Distance** - The Euclidean distance from home plate to the landing spot of each BIP.

5. **Play Made** - Indicator variable marking whether a BIP was turned into an out or not. The value 1 indicates that an out was recorded and 0 indicates otherwise. Figure 2.1 shows the the plays made by all centerfielders during the 2016 MLB season.

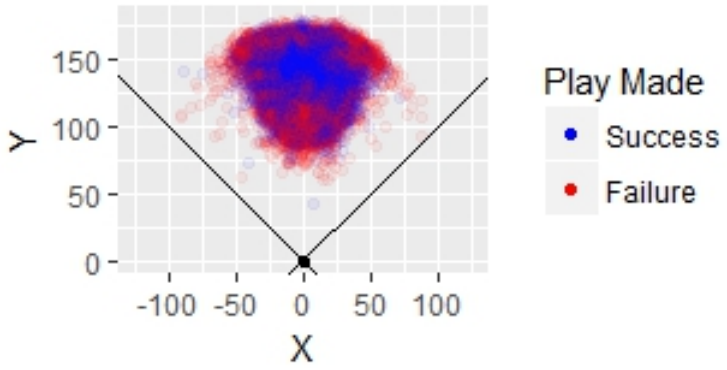

FIGURE 2.1: Distribution of BIPs to centerfield during the 2016 season.

3

Bayesian Hierarchical Logistic Regression Model

In this chapter, we will discuss our implementation of a Bayesian hierarchical logistic regression model to predict whether or not a centerfielder makes a successful play on a given BIP. We will cover the model specification, along with our use of weakly informative priors to influence the shrinkage of player specific regression coefficients to the population means. We will assess the model fit by examining the coefficients for the model's main effects, the model's residuals, and the model's ROC curve. Finally, we provide a posterior predictive check by examining the estimated differences between the best and worst centerfielders for the 2016 season.

3.1 Model Specification

In specifiying our model, we borrow notation from (Jensen et al., 2009). We begin by specifying our model for a particular fielder i, for j BIPs hit to fielder i. We denote the outcome variable Y:

$$Y_{ij} = \begin{cases} 1 & \text{if the jth BIP to the ith player is caught} \\ 0 & \text{if the jth BIP to the ith player is not caught} \end{cases} \tag{3.1}$$

It then follows that Y is Bernoulli distributed:

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}) \tag{3.2}$$

The Bernoulli probabilities for each BIP are then modeled as a function of how far the fielder had to travel to field each BIP (Distance Ran = DR), how far each BIP is hit from home plate (BIP Distance = BIPD), how long each BIP is in the air (Air Time = A), and how hard each BIP is hit (Hit Speed = HS). We standardize each of these covariates to Normal(0, 1) distributions when fitting our model to compare their regression coefficients on the same scale. The model we use to fit the Bernoulli probability for BIP j hit to fielder i is therefore,

$$\begin{aligned} p_{ij} = logit^{-1}(B_{i0} + B_{i1}(DR_{ij}) + B_{i2}(DR_{ij}^2) + B_{i3}(A_{ij}) \\ + B_{i4}(A_{ij}^2) + B_{i5}(HS_{ij}) + B_{i6}(HS_{ij}^2) + B_{i7}(BIPD_{ij})) \end{aligned} \tag{3.3}$$

The B coefficients in our model make intuitive sense when you consider the nature of BIPs. B_{i0} is the average prediction when all other covariates are at their standardized mean levels of 0. B_{i1} through B_{i7} are all included to capture the complex quadratic relationship between how hard and far a BIP is hit and the probability that a centerfielder will catch it. BIPs that are hit relatively softly will be challenging for centerfielders to catch because they will tend to land in front of them. BIPs that are hit very hard and far, will also be challenging to catch because they will tend to fly over the centerfielders' heads. BIPs that fall in between these two extremes will tend to be easier to catch for centerfielders because they will be hit closer to directly at them. We also tried including $BIPD^2$ in our model, but this term did not add anything substantial to the model fit as its effect was captured by the other squared terms.

3.1.1 Prior Specification

We model p_{ij} using a hierarchical model so that we can share information across players. This will result in player specific regression coefficients that can be used to make player specific predictions. This also enforces shrinkage towards the population means for each regression coefficient, which will be greater for players with relatively fewer BIPs hit to them during the 2016 MLB season. Our prior distributions are specified as follows:

$$B_i \sim \text{Normal}(\mu, \Sigma) \tag{3.4}$$

where μ is an 8×1 vector of population regression coefficient means distributed $\mu \sim \text{Normal}(0, 100)$ and Σ is an 8×8 prior covariance matrix. We assume a priori that there is no covariance between our regression coefficients, so the off-diagonal elements of Σ are all zero. On the diagonal of Σ we place Cauchy(0, 1) priors on σ_k^2 for k = 1, ..., 8 at the suggestion of (Gelman, 2006). We choose Cauchy(0, 1) priors because each covariate is standardized to be distributed Normal(0, 1). Doing so places a prior on the variance that is on the same scale as the data. We choose a Cauchy distribution because the Cauchy distribution's fat tails allow for relatively extreme player specific regression coefficients to remain extreme given a large number of BIPs. Figure 3.1 shows the increasing amount of shrinkage to the mean for the main effects in our model as the number of BIPs hit to each player decreases. This model specification allows for information to be shared across players, but also allows each player to have their own specific set of regression coefficients.

3.1.2 Posterior Inference

We sample posterior draws of B, μ, and σ^2 from the following joint distribution:

$$p(B, \mu, \sigma^2|Y, X) \propto p(Y|X)p(B|\mu, \sigma^2)p(\mu, \sigma^2), \tag{3.5}$$

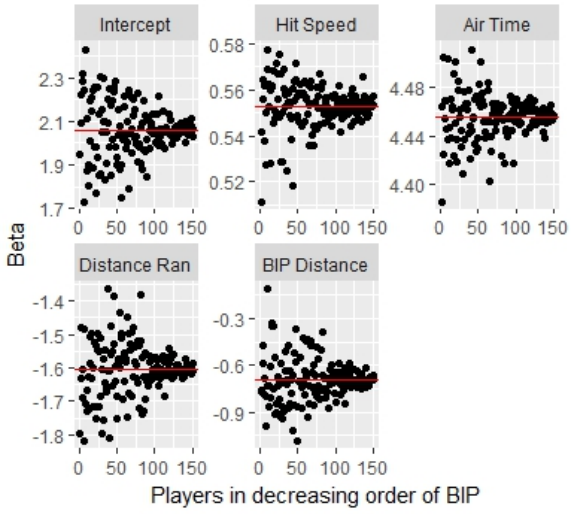

FIGURE 3.1: Shrinkage of main effects caused by the hierarchical nature of our model. The red lines indicate the population mean for each regression coefficient.

where, $p(Y|X)$ is the Bernoulli data likelihood given the covariates X, $p(B|\mu, \sigma^2)$ is the prior distribution placed on the regression coefficients B, and $p(\mu, \sigma^2)$ are the hyperpriors placed on μ and σ^2. Samples of the parameters in our model are drawn from this posterior distribution via a Stan implementation of our model specification (Stan, 2016). The outputs from our model are μ, an 8×1 vector of population regression coefficients, σ_k^2 for k = 1, ..., 8, an 8×1 vector of population regression coefficient variances, and, B, an $N \times 8$ matrix of player specific regression coefficients, where N = 151, the number of players represented in our data.

3.2 Model Fit

Now that we have specified our model for predicting the probability that a center-fielder will catch a given BIP, we will examine the fit of our model via a series of informative visualizations. First, in Figure 3.2, we see the posterior means of each of the main effects from our model. The strongest effect belongs to Air Time. This result is intuitive; the longer the ball hangs in the air, the longer the fielder has to

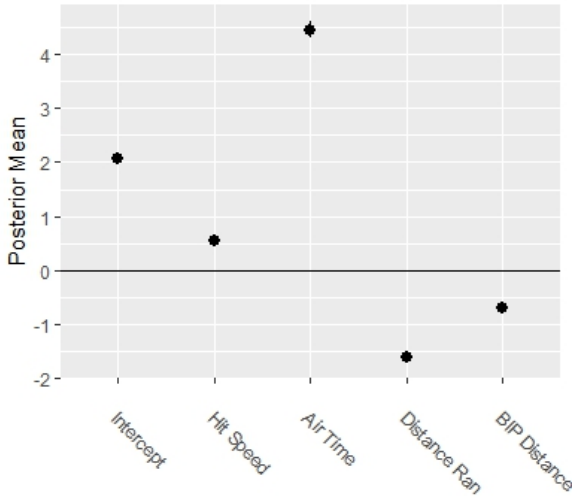

FIGURE 3.2: Main effects of hierarchical logisitic regression model.

get in position to catch it. We also see negative effects for Distance Ran and BIP Distance. These also match intuition. The further a fielder has to run to catch a BIP, the more difficult the play. Additionally, BIPs hit over the fielder's head are commonly known to be more difficult to catch.

Figure 3.3 displays a histogram of the residuals from our model. We can see that approximately 8,000 of our 17,578 observations are pretty accurately predicted by our model. The extreme residuals near the value of -1 indicate very poorly played BIPs and possibly even errors. The residuals near the value of 1 indicate incredible plays that are very rarely made.

Next, we bin the residuals into 100 equally-sized bins according to their fitted probabilities from our model. We take the average of the residuals in each bin and plot them against their predicted values in Figure 3.4. We do this to examine the possible existence of structure in our residuals. A lack of structure indicates that our model is not failing to account for some feature inherent in the data. Our model seems well-calibrated due to lack of obvious structure in this plot.

We repeat this process for each of the covariates in our model, with the exception

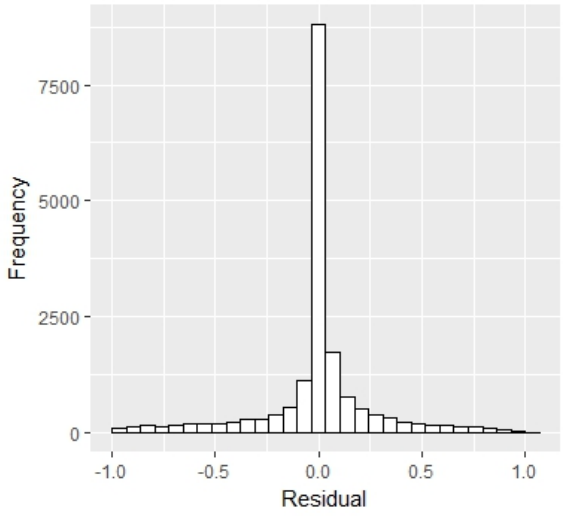

FIGURE 3.3: Histogram of posterior residuals from our fitted model.

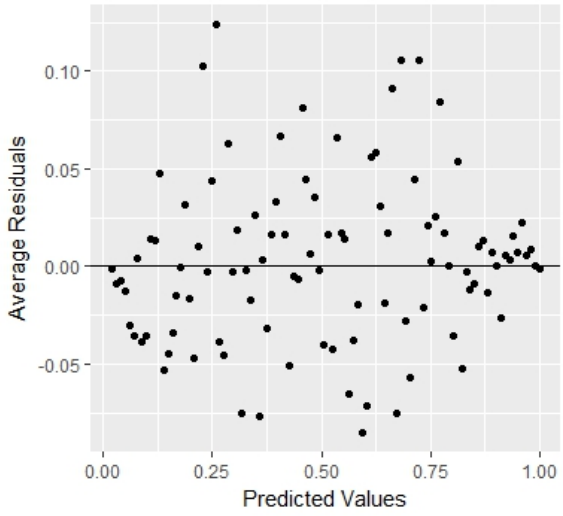

FIGURE 3.4: Binned residuals vs. predicted probabilities plot. There is no apparent trend in this plot.

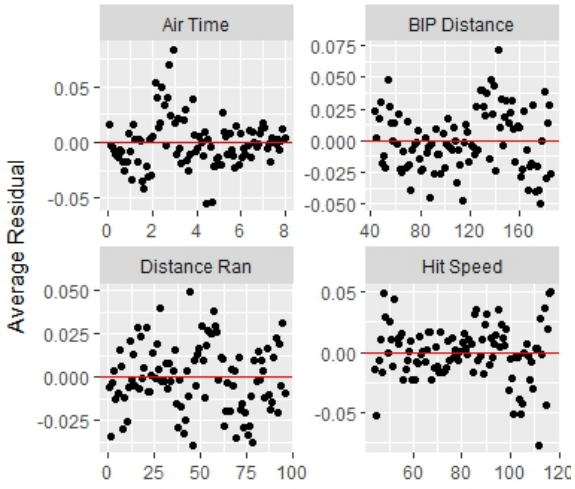

FIGURE 3.5: Binned residuals vs. continuous covariates plot. There are no obvious trends in these plots.

that instead of binning the residuals by the associated predicted values, we bin by the values of the covariates. Once again, we are looking for any trends across the values of the covariates. Figure 3.5 displays these plots. It was by examining these plots that we decided to add the squared terms to our model. Before correcting our model, there were quadratic trends across Distance Ran, Air Time, and Hit Speed. While there is a slight spike in the residuals in the Air Time plot between 2 and 3 seconds of Air Time and perhaps increased variance at the extreme values of Hit Speed, these minor issues do not significantly influence the output of our model.

Next, we bin the data by our model's fitted values and find the proportion of times that the BIPs within each bin were actually caught. We plot this in Figure 3.6. A well-calibrated model will generally have points that follow a 45 degree line through the plot. In our case, our points follow the line closely.

Finally, we examine the ROC (Receiver Operating Characteristic) curve and its corresponding AUC (Area Under Curve) value. Figure 3.7 plots the ROC curve with an associated AUC of .9642. In this application, the AUC measures the probability

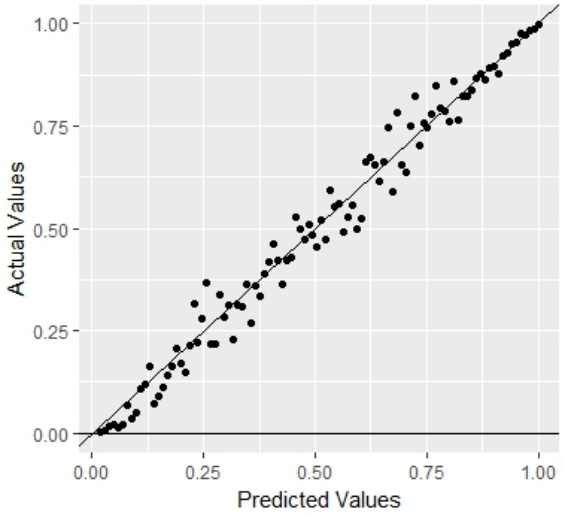

FIGURE 3.6: Predicted Values vs. Actual Values plot. We expect the points to follow a 45 degree line through the plot.

that a randomly chosen BIP that is caught will have a higher predicted probability of being caught than a randomly chosen BIP that is not caught. Our AUC of .9642 is substantially high to conclude that our model classifies BIPs sufficiently well.

3.3 Posterior Predictive Check

Once again following the example of (Jensen et al., 2009), we perform a posterior predictive check to ensure that our model not only fits the data well at a macro level, but also performs well in capturing the heterogeneity between individual players. To check for this, we find the player with the highest proportion of caught BIPs among the 15 players with the most BIPs hit to them. Likewise, we find the player with the lowest proportion of caught BIPs. The difference between these two proportions is the value against which we are measuring the accuracy of our model.

To test our model against this difference, we find the difference in proportions of caught BIPs between the best (Joc Pederson) and worst (Ian Desmond) centerfielders from the data across 500 posterior predictive samples for each player. This give us

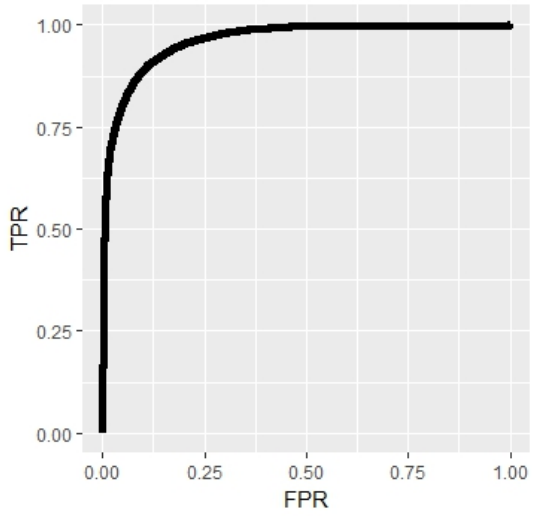

FIGURE 3.7: ROC curve for our logistic regression model. The AUC for this curve is .9642.

500 posterior samples of the difference between the two players. Figure 3.8 plots the posterior distribution of the differences. The vertical line marks the actual difference from the data. We can see that the mode of our posterior distribution is slightly smaller than the actual difference. This is due to the inherent shrinkage to the mean that our model specification implies. Overall, this posterior predictive check indicates that the model is well calibrated to capture the heterogeneity between individual players.

To visualize this difference over the surface of centerfield, we make predictions for both the best and worst fielders over a grid of covariate values. We then plot this difference for four different Air Times in Figure 3.9. The Air Times listed are 1, .66, .33, and 0 standard deviations below the mean for Air Time. We can see that as Air Time increases, the differences between the two players decrease. This matches intuition because once a BIP reaches a certain level of Air Time, we expect almost all centerfielders to catch the BIP.

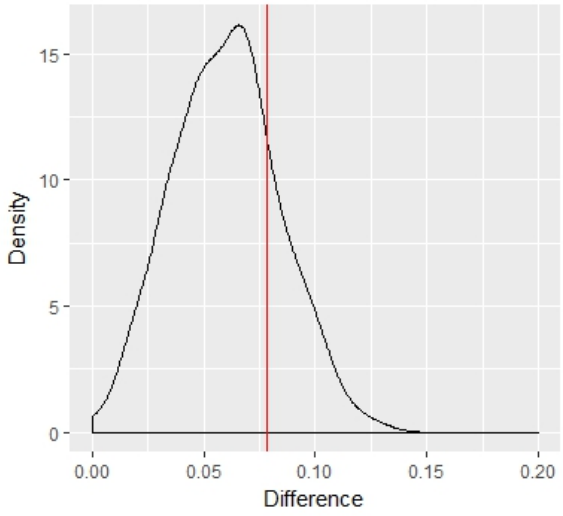

FIGURE 3.8: Posterior predictive check between the actual difference between the best and worst centerfielders from the data (red) and the posterior distribution of the same difference.

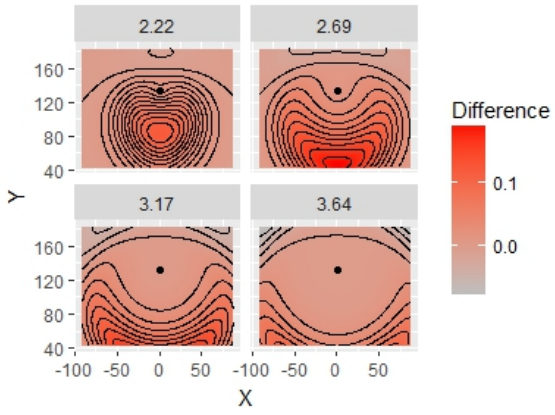

FIGURE 3.9: Plot of the differences between the best (Joc Pederson) and worst (Ian Desmond) fielders.

4

SAFE 2.0

Now that we have fit and validated our Bayesian hierarchical logistic regression model to predict the probability that a given BIP will be caught, we will now use this model as a component of our new metric SAFE 2.0. In this chapter, we will discuss in detail each component used in the calculation of SAFE 2.0. We will then calculate SAFE 2.0 scores for every centerfielder for the 2016 MLB season and compare these rankings to UZR's rankings. Finally, we will discuss the potential shortcomings of SAFE 2.0 and future SAFE 2.0 possibilities. In this chapter, we borrow notation from and follow the techniques of (Jensen et al., 2009).

4.1 SAFE 2.0 Derivation

In this section we derive each component of SAFE 2.0 and explain the process for calculating SAFE 2.0 for each player.

4.1.1 Predicted Catch Probability Relative to an Average Centerfielder

In calculating SAFE 2.0, we are interested in creating a metric that compares each fielder to the league average fielder. For a single BIP we can compare the difference

between the i^{th} fielder and an average fielder by taking the difference between their respective predicted probabilities from our logistic regression model. For a single BIP this is the calculation:

$$\hat{p_i}(X) - \bar{p}(X), \tag{4.1}$$

where $\hat{p_i}(X)$ is the predicted probability for the BIP to the i^{th} player using player i's individual regression coefficients and $\bar{p}(X)$ is the predicted probability for the average player using the population regression coefficients. X is the set of covariates of the BIP. We can extend Equation 4.1 to make predictions for player i over the average player to all possible BIP profiles by numerically integrating Equation 4.1 over a fine grid of BIP covariate values as follows:

$$\int \hat{p_i}(X) - \bar{p}(X) dX, \tag{4.2}$$

where X is now the grid of BIP covariate values. This numeric integration gives us the total difference in predicted probabilities between player i and the average player.

4.1.2 BIP Relative Frequency Adjustment

Due to the non-uniform distribution of BIPs across the field, as seen in Figure 4.1, Equation 4.2 is insufficient. We do not want to weight each BIP equally over our grid. Rather, we would like to weight each BIP in our grid by the relative frequency in which they occurred during the 2016 MLB season. This will ensure that we weight the value of each BIP appropriately when calculating SAFE 2.0. We do not want to reward or punish centerfielders too heavily for rare BIP occurrences relative to commonly occurring BIPs. We suggest the following addition to Equation 4.2:

$$\int [\hat{p_i}(X) - \bar{p}(X)] \times \hat{f}(X) dX, \tag{4.3}$$

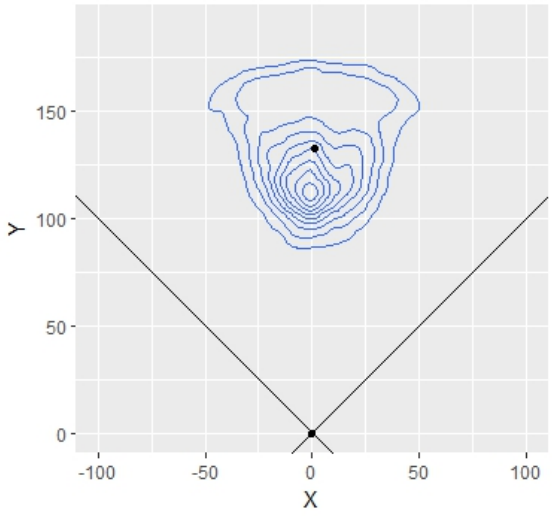

FIGURE 4.1: Distribution of BIPs hit to centerfield. The distribution is non-uniform, so an adjustment for the relative frequency of BIPs in SAFE 2.0's calculation is necessary.

where $\hat{f}(X)$ is the relative frequency of each BIP given covariates X as estimated using kernel density estimation.

4.1.3 Run Value Adjustment

Once again, we find that Equation 4.3 is insufficient. While Equation 4.3 provides appropriately-weighted differences between player i and the average player, a measure of value is not attached to these differences. We update Equation 4.3 by finding the average Run Value for every proposed BIP in our grid when it is not caught. Once again, we use kernel density estimation to estimate the relative frequencies for the occurrence of singles, doubles, and triples for each BIP in our grid. We then find the weighted average Run Value according to these relative frequencies and the average Run Value for the 2016 season of singles, doubles, and triples respectively:

$$\hat{r}(X) = 0.72\hat{f}_{Singles}(X) + 1.02\hat{f}_{Doubles}(X) + 1.29\hat{f}_{Triples}(X), \tag{4.4}$$

where each $\hat{f}(X)$ term is the relative frequency of each outcome for each set of BIP covariates X. The constants 0.72, 1.02, and 1.29 are the average Run Values for singles, doubles, and triples for the 2016 season. We incorporate these values into our final SAFE 2.0 equation:

$$\text{SAFE 2.0} = \int [\hat{p}_i(X) - \bar{p}(X)] \times \hat{f}(X) \times \hat{r}(X) dX, \tag{4.5}$$

Now, SAFE 2.0 measures the difference between player i and an average centerfielder while accounting for the relative frequency and the Run Value of BIPs over a fine grid of proposed BIPs. In the next section, we calculate the SAFE 2.0 values for every CF for the 2016 season with more than 200 fielding opportunities.

4.2 SAFE 2.0 Results

For each centerfielder who had at least 200 BIP opportunities over the course of the 2016 season, we calculate the posterior means and 95% credible intervals of their SAFE 2.0 scores. We arrive at these values by calculating the differences in predicted catch probability between each player and the average player for each BIP over our grid of proposed BIPs at each iteration of our Bayesian sampler. Each player's individual scores can be seen in Table 4.1, which also includes each player's UZR/150 score for reference (FanGraphs.com, 2017d).

4.3 SAFE 2.0 Analysis

There are a few aspects to point out regarding our SAFE 2.0 results. Obviously, the scale of the results is very small and the differences between individual players are not very substantial. Additionally, only 2 players at each extreme of the measurement have credible intervals that do not include 0. This indicates that for a large majority of players we cannot conclude with certainty that they are substantially different

Table 4.1: SAFE 2.0 calculations for every centerfielder with more than 200 BIP opportunities during the 2016 MLB season. We also display each player's UZR/150 for each

	Name	Lower	Mean	Upper	SAFE Rank	UZR150	UZR Rank
1	Adam Jones	0.0048	0.0279	0.0509	1	-9.9000	30
2	Leonys Martin	0.0021	0.0248	0.0475	2	4.2000	13
3	Keon Broxton	-0.0113	0.0176	0.0514	3	23.2000	3
4	Ender Inciarte	-0.0043	0.0164	0.0389	4	14.9000	7
5	Billy Hamilton	-0.0075	0.0161	0.0407	5	17.2000	6
6	Dexter Fowler	-0.0116	0.0158	0.0436	6	1.0000	17
7	Jacoby Ellsbury	-0.0084	0.0151	0.0382	7	2.1000	16
8	Jake Marisnick	-0.0107	0.0141	0.0388	8	4.3000	12
9	Kevin Pillar	-0.0115	0.0115	0.0359	9	26.3000	1
10	Travis Jankowski	-0.0147	0.0110	0.0379	10	20.9000	4
11	Charlie Blackmon	-0.0109	0.0104	0.0316	11	-11.9000	32
12	Kevin Kiermaier	-0.0135	0.0096	0.0307	12	24.2000	2
13	Joc Pederson	-0.0155	0.0090	0.0330	13	0.4000	18
14	Jon Jay	-0.0242	0.0064	0.0350	14	-7.4000	27
15	Ben Revere	-0.0238	0.0041	0.0323	15	-1.1000	20
16	Jarrod Dyson	-0.0213	0.0039	0.0325	16	19.6000	5
17	Odubel Herrera	-0.0199	0.0037	0.0271	17	3.3000	15
18	Lorenzo Cain	-0.0264	0.0036	0.0313	18	13.6000	9
19	Mike Trout	-0.0243	-0.0019	0.0207	19	-0.4000	19
20	Byron Buxton	-0.0283	-0.0038	0.0192	20	6.4000	10
21	Ian Desmond	-0.0276	-0.0040	0.0169	21	-5.7000	26
22	Billy Burns	-0.0295	-0.0044	0.0204	22	-5.6000	25
23	Jackie Bradley Jr	-0.0258	-0.0058	0.0137	23	5.1000	11
24	Denard Span	-0.0274	-0.0067	0.0191	24	-10.0000	31
25	Rajai Davis	-0.0374	-0.0087	0.0196	25	3.6000	14
26	Carlos Gomez	-0.0344	-0.0091	0.0154	26	-1.7000	21
27	Michael Bourn	-0.0367	-0.0106	0.0127	27	-2.8000	23
28	Cameron Maybin	-0.0366	-0.0117	0.0115	28	-9.8000	29
29	Kirk Nieuwenhuis	-0.0430	-0.0155	0.0119	29	13.7000	8
30	Randal Grichuk	-0.0431	-0.0177	0.0091	30	-1.8000	22
31	Yoenis Cespedes	-0.0514	-0.0199	0.0047	31	-20.6000	33
32	Marcell Ozuna	-0.0446	-0.0207	0.0034	32	-3.2000	24
33	Tyler Naquin	-0.0590	-0.0289	-0.0017	33	-9.4000	28
34	Andrew McCutchen	-0.0556	-0.0341	-0.0130	34	-23.2000	34

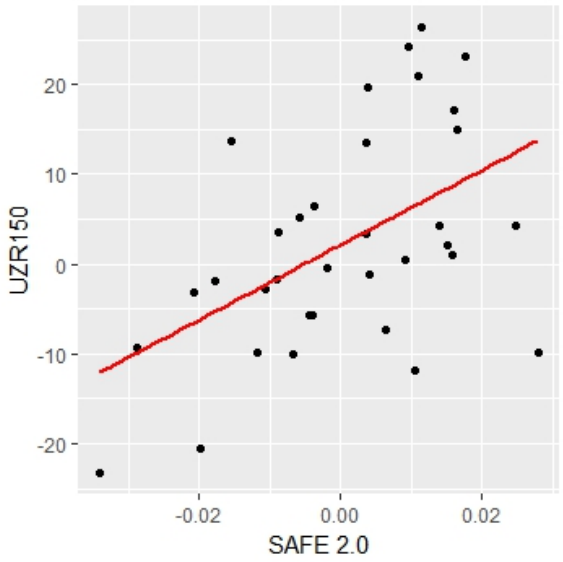

FIGURE 4.2: Scatterplot of SAFE 2.0 vs. UZR150 for centerfielders for the 2016 MLB season. The red line indicates the best fit linear regression line between the two metrics. The correlation between the two metrics is 0.503.

than league average defenders. This is an issue because we would like to be better able to differentiate between players. This may, however, indicate that there are not substantial differences in the fielding ability of MLB centerfielders when controlling for their individual BIP opportunities. Differences between players' UZR scores may be driven by the nature of their BIP opportunities in a given season and not substantial differences in their inherent defensive abilities. Figure 4.2 displays the relationship between SAFE 2.0 and UZR. The two metrics generally agree with each other with a correlation of 0.503. This indicates that SAFE 2.0 is generally capturing what the commonly used UZR is communicating. This is a promising finding because we do not expect these two metrics to agree or disagree to a great extent.

4.4 SAFE 2.0 Limitations and Future Possibilities

4.4.1 Limitations

Despite SAFE 2.0's seemingly decent performance, there are a few shortcomings to consider regarding the metric. First, while SAFE 2.0 generally agrees with UZR/150, they do disagree substantially on several players, most notably Adam Jones, who is ranked first by SAFE 2.0 but is in the bottom five of UZR/150. This behavior requires further research. Second, the scale of SAFE 2.0 is not useful when determining the value of a player's performance. In baseball, value is predominantly measured in terms of runs saved or created, and SAFE 2.0 is not currently measured on that scale. Further research should focus on converting SAFE 2.0's scale to runs in order for it to be more compatible with current definitions of baseball value. Third, SAFE 2.0 does not consider any aspects of defense besides catching line drives and fly balls. UZR/150 measures performance on all aspects of defense including performance on ground balls and a fielder's throwing arm.

4.4.2 Future Possibilities

With the basis laid out in (Jensen et al., 2009)'s original SAFE calculations and continued here with the calculation of SAFE 2.0, there is potential that SAFE can be further improved with more research. Perhaps the greatest enhancement would come in the form of predicting the probability of catching each BIP using a Spatial Point Process model as opposed to the hierarchical logistic regression proposed here. This would allow for the exact locations of BIPs to provide more information to the predictions than they do currently. Another future possibility would be incorporating centerfielders' performance on aspects of their defense beyond just catching BIPs such as their throwing ability and their ability to limit extra base hits. Additionally, the methodology for the calculation of SAFE 2.0 can be extended to every position on

the baseball field. This would allow comparisons beyond just the one position we examine here. Finally, calculating SAFE 2.0 over multiple seasons and testing its consistency would be very valuable in determining how well it captures each player's true defensive talent.

4.5 Conclusion

In this paper, we introduced the methodology behind three prevailing measures of a player's defensive value: UZR, DRS, and SAFE. We introduced the exciting newly available Statcast data and explained how it could be used to help measure baseball player's defensive value. Next, we described the Bayesian hierarchical logistic regression model that we used to model the probability that a given BIP would be caught. We then combined this model with the relative frequency and Run Value of BIPs to calculate our metric SAFE 2.0 for each centerfielder for the 2016 MLB season. Finally, we compared SAFE 2.0 to UZR/150 and found that they generally agree with each other. Overall, SAFE 2.0 is a decent metric that might be a useful way to rank baseball player's defensive value. As more Statcast data becomes available, the value of SAFE 2.0 will become more apparent as its ability to measure the true talent of players becomes more well-defined.

Bibliography

Basco, D. and Zimmerman, J. (2010), "Measuring Defense: Entering the Zones of Fielding Statistics," *Baseball Research Journal*, 39.

Dewan, J. (2012), "Frequently Asked Questions about Plus/Minus and Runs Saved," .

FanGraphs.com (2017a), "The FanGraphs UZR PRimer," .

FanGraphs.com (2017b), "Linear Weights," .

FanGraphs.com (2017c), "UZR," .

FanGraphs.com (2017d), "UZR Leaderboard," .

FederalBaseball.com (2010), "Washington Nationals: Adam Dunn, UZR and Defensive Value," .

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *International Society for Bayesian Analysis*.

Jensen, S. T., Shirley, K. E., and Wyner, A. J. (2009), "Bayesball: A Bayesian Hierarchical Model for Evaluating Fielding in Major League Baseball," *Institute of Mathematical Statistics*.

MLB.com (2017), "Statcast," .

Stan (2016), *Stan Modeling Language: Users Guide and Reference Manual*, mc-stan.org.