# Measurement Space Partitioning for Estimation and Prediction

Glenn Healey and Shiyuan Zhao
Electrical Engineering and Computer Science
University of California, Irvine, CA 92617

July 23, 2021

## Abstract

An important and challenging problem in the evaluation of baseball players is the quantification of batted-ball talent. This problem has traditionally been addressed using linear regression on the value of a statistic derived from a set of observations. We use large sets of trajectory measurements acquired by in-game sensors to show that the predictive value of a batted ball depends on its physical properties. This knowledge is exploited to estimate batted-ball distributions defined over a multidimensional measurement space from observed distributions by using regression parameters that adapt to batted ball properties. This process is central to a new method for estimating batted-ball talent. The domain of the batted-ball distributions is defined by a partition of measurement space that is selected to optimize the accuracy of the estimates. We present examples illustrating facets of the new approach and use a set of experiments to show that the new method generates estimates that are significantly more accurate than those generated using current methods. The new methodology supports the use of fine-grained contextual adjustments and we show that this process further improves the accuracy of the technique.

# 1 Introduction

The assessment of player skills in baseball is increasingly dependent on data-driven models rather than subjective evaluation. The accuracy of these models is critical to a team's success as executives attempt to maximize performance while abiding by organizational financial constraints. Measuring player skill on batted balls is of particular interest since the majority of Major League Baseball (MLB) matchups result in a batted ball. Estimating batted-ball skill, however, has proven difficult [2] because batted-ball results are subject to a large amount of random variation and are biased by confounding variables such as the atmospheric conditions.

Player talent level on batted balls is defined as the expected value of a statistic which can be estimated from a sample of observations. The utility of an estimate is often evaluated by its ability to predict player performance on unobserved data. An intuitively appealing estimate of talent level is simply the computed value of the statistic over a player's observed sample. But paradoxically this method is less accurate than estimators [8] [13] [21] that are defined by a weighted average of this computed value and the mean of the statistic over a group of players. An example of these estimators is linear regression (LR) for which the weighting depends on the correlation of the value of the statistic across samples. LR estimates have been used by several systems to predict player performance [19][23].

In recent years radar and optical sensors have been used in MLB stadiums to measure characteristics of batted balls such as speed, direction, and spin [11]. We use these sensor measurements to develop a new method for estimating talent level called measurement space partitioning (MSP). After constructing a discrete batted-ball distribution defined over a partition of the multidimensional measurement space for samples from a group of players, we use Cronbach's alpha to show that the expected correlation of distribution values across samples has a strong dependence on location in measurement space. This allows a player's underlying batted-ball distribution and the corresponding talent level to be estimated using regression parameters that adapt to his specific batted-ball distribution. The accuracy of the talent level estimate depends on the partition which leads to the derivation of a method for partition optimization. A set of experiments is used to show that the MSP method improves on the accuracy of linear regression for estimating batted-ball talent level.

Another advantage of the MSP approach is the ability to incorporate fine-grained contextual information into estimates. Contextual information includes a range of variables that can affect batted-ball value. The weather conditions and elevation, for example, will affect how far a batted ball will carry in the air [1]. Batted balls that follow similar trajectories can have different outcomes due to differences in outfield geometry from ballpark to ballpark [9]. A player's running speed [12] and variables that include the height of the infield grass and the composition of the infield surface [3] can affect the value of batted balls hit on the ground. The fate of batted balls also depends on the quality of the defenders in the field. Contextual factors are typically accounted for by a coarse adjustment that compensates for the average effect of the environment [17]. Since the MSP method computes talent level estimates from regressed batted ball distributions defined over physical parameters, contextual adjustments can be employed that depend on the characteristics of individual batted balls. A ball hit in the air at high speed, for example, can be adjusted differently from a ball hit softly on the ground. We will show that the use of fine-grained contextual adjustments improves the accuracy of predictions made by the MSP method.

# 2 Estimation and Prediction

## 2.1 Talent Level

Talent for a skill varies from player to player and can be represented by a statistic that is derived from a set of observations. The computed value of such a statistic equals talent level $T(j)$, which is the expected value of the statistic for player $j$, plus estimation error. In this work, we examine the problem of estimating player talent level on batted balls. Consider a dataset that contains information on $2N$ batted balls for each of $P$ players where the data is arranged so that the first $N$ batted balls for each player are observed and the second $N$ batted balls for each player are unobserved. Let $R(i, j)$ represent the numerical value of batted ball $i$ for player $j$ and define the observed performance statistic for player $j$ as the average over the first $N$ batted balls

$$x(j) = \frac{1}{N} \sum_{i=1}^{N} R(i, j) \tag{1}$$

and define the unobserved performance for player $j$ as the average over the second $N$ batted balls

$$y(j) = \frac{1}{N} \sum_{i=N+1}^{2N} R(i,j). \tag{2}$$

We consider the task of using the observed batted ball data to estimate talent level $T(j)$ for the $x(j)$ statistic for each player $j$. The estimated $T(j)$ can be used to predict the unobserved performance $y(j)$.

## 2.2 Linear Regression

One estimate of $T(j)$ is the observed performance $x(j)$ for player $j$. However, the James-Stein paradox [13] [21] as illustrated by Effron and Morris [8] shows that a more accurate estimate of $T(j)$ is obtained by adjusting the $x(j)$ using an average of the observed $R(i,j)$ values over multiple players. Since an estimate for talent level can be assessed by its ability to predict the unobserved performance $y(j)$, we can define an estimate $\widehat{y}(j)$ for $T(j)$ by minimizing the sum of the square errors

$$E = \sum_{j=1}^{P} (y(j) - \widehat{y}(j))^2 \tag{3}$$

using the linear regression model

$$\widehat{y}(j) = a + bx(j). \tag{4}$$

The values of $a$ and $b$ that minimize $E$ are

$$a = \mu_y - \frac{r\mu_x\sigma_y}{\sigma_x} \tag{5}$$

$$b = \frac{r\sigma_y}{\sigma_x} \tag{6}$$

where $\mu_x$ and $\sigma_x$ are the mean and standard deviation for the $x(j)$, $\mu_y$ and $\sigma_y$ are the mean and standard deviation for the $y(j)$, and $r$ is the correlation coefficient for the set of $P$ points $(x(j), y(j))$ [7].

Since the data used to generate the $y(j)$ are unobserved, the parameters $\mu_y$, $\sigma_y$, and $r$ in equations (5) and (6) cannot be computed directly. The $y(j)$, however, are generated in the same way for the same players as the $x(j)$ which allows us to use the approximations $\mu_y = \mu_x$ and $\sigma_y = \sigma_x$. The remaining unknown parameter, the correlation coefficient $r$, can be approximated from the observed $R(i,j)$ values using Cronbach's alpha [6]

$$\alpha(N) = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_{R_i}^2}{\sigma_{R_T}^2}\right) \tag{7}$$

where $\sigma_{R_i}^2$ is the variance of the observed $R(i,j)$ values over players $j$ for batted ball $i$ and $\sigma_{R_T}^2$ is the variance of

$$R_T(j) = \sum_{i=1}^{N} R(i,j) \tag{8}$$

over players $j$. Using these approximations, equation (4) becomes

$$\widehat{y}(j) = \alpha(N)x(j) + (1 - \alpha(N))\mu_x \tag{9}$$

which can be computed using the observed data. $\widehat{y}(j)$ in equation (9) is consistent with the James-Stein result that an improved estimate for $T(j)$ can be obtained by adjusting $x(j)$ using the overall mean $\mu_x$.

## 2.3 Varying Observed Sample Size

The $\alpha(N)$ that is used to compute the estimate $\widehat{y}(j)$ in equation (9) is derived using a dataset of $N$ observed batted balls for each of $P$ players using equation (7). The utility of the method is enhanced if we can use this dataset to compute the estimate $\widehat{y}(j)$ using a sample of $N'$ batted balls for player $j$ where $N' \neq N$. The value of $\alpha(N)$ tends to increase with $N$ due to a decrease in the variance of the random error in the observed performance $x(j)$ [27]. The Spearman-Brown prophecy formula [4] [20] allows us to predict $\alpha(N')$ from the estimated $\alpha(N)$ using

$$\alpha(N') = \frac{C\alpha(N)}{1 + (C-1)\alpha(N)} \tag{10}$$

where $C = N'/N$. This $\alpha(N')$ can be used in equation (9) to compute $\widehat{y}(j)$ using an observed performance $x(j)$ computed using any number of samples $N'$.

# 3 Exploiting Sensor Measurements

## 3.1 Partitioning the Measurement Space

Sensors allow batted balls to be represented by a point in a measurement space with dimensions defined by properties such as speed, direction, and spin. The measurement space can be partitioned into $B$ disjoint subsets. For the dataset described in Sec. 2.1 let $M(i, j, k)$ be a binary-valued function which is one if batted ball $i$ for player $j$ is in subset $k$ and zero otherwise. Define the observed batted ball distribution for player $j$ over the subsets $k$ by

$$p_x(j, k) = \frac{1}{N} \sum_{i=1}^{N} M(i, j, k) \tag{11}$$

and define the unobserved batted ball distribution for player $j$ over the subsets $k$ by

$$p_y(j, k) = \frac{1}{N} \sum_{i=N+1}^{2N} M(i, j, k). \tag{12}$$

We will show that an estimate for $p_y(j, k)$ can be used to generate an estimate for the talent level $T(j)$.

## 3.2 Estimating Measurement Space Distributions

For a given subset $k$ we can use a linear regression model and approximations similar to those described in Sec. 2 to estimate $p_y(j, k)$ from the observed data according to

$$\widehat{p}_y(j, k) = \alpha(N, k) p_x(j, k) + (1 - \alpha(N, k))\mu(k) \tag{13}$$

where $\mu(k)$ is the average

$$\mu(k) = \frac{1}{P} \sum_{j=1}^{P} p_x(j, k) \tag{14}$$

and $\alpha(N,k)$ is the Cronbach approximation to the correlation coefficient for the set of $P$ points $(p_x(j,k), p_y(j,k))$ for subset $k$. Specifically, $\alpha(N,k)$ is computed using

$$\alpha(N,k) = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_{M_i}^2}{\sigma_{M_T}^2}\right) \tag{15}$$

where $\sigma_{M_i}^2$ is the variance of the observed $M(i,j,k)$ values over players $j$ for batted ball $i$ and subset $k$ and $\sigma_{M_T}^2$ is the variance of

$$M_T(j) = \sum_{i=1}^{N} M(i,j,k) \tag{16}$$

over players $j$ for subset $k$. $\alpha(N,k)$ can then be used in equation (13) to compute the regressed distribution $\widehat{p}_y(j,k)$ using only the observed data. We note that the calculation in equation (15) can yield $\alpha(N,k)$ values that are negative [27] and in these cases $\alpha(N,k)$ is set to zero for the calculation of $\widehat{p}_y(j,k)$.

### 3.3 Estimating Talent Using Measurement Space Partitioning

The batted ball distribution estimate $\widehat{p}_y(j,k)$ for player $j$ can be used to estimate the player's talent level $T(j)$. If $\overline{R}(j,k)$ is an estimate of the expected value of batted balls for player $j$ in subset $k$ then $T(j)$ can be estimated by

$$\widehat{y}_s(j) = \sum_{k=1}^{B} \widehat{p}_y(j,k)\overline{R}(j,k). \tag{17}$$

For cases where we would like to estimate $\widehat{y}_s(j)$ using a sample of $N'$ batted balls for player $j$, the values $\alpha(N',k)$ for each $k$ in equation (13) can be computed using the Spearman-Brown formula as described in Sec. 2.3.

The $\widehat{y}_s(j)$ estimate in equation (17) is equivalent to the linear regression estimate $\widehat{y}(j)$ in equation (9) if $\alpha(N,k)$ has the same value $\alpha(N)$ for all subsets $k$ and the average value of the observed batted balls in any subset $k$ is the same for all players $j$. For this special case, if we let $\overline{R}(j,k)$ equal the overall mean of the observed $R(i,j)$ for subset $k$

$$\overline{R}(k) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{P} M(i,j,k)R(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{P} M(i,j,k)} \tag{18}$$

then equation (17) can be written

$$\begin{aligned} \widehat{y}_s(j) &= \sum_{k=1}^{B} \left[\alpha(N)p_x(j,k) + (1-\alpha(N))\mu(k)\right]\overline{R}(k) \\ &= \left[\alpha(N)\sum_{k=1}^{B} p_x(j,k)\overline{R}(k)\right] + \left[(1-\alpha(N))\sum_{k=1}^{B}\mu(k)\overline{R}(k)\right] \end{aligned} \tag{19}$$

where the first sum in equation (19) equals $x(j)$ and the second sum equals $\mu_x$ which demonstrates the equivalence to equation (9). We will see that by allowing $\alpha(N,k)$ to vary over subsets $k$ and by allowing $\overline{R}(j,k)$ to vary over players $j$, the model in equation (17) can generate estimates that are more accurate than the linear regression estimate in equation (9).

# 4 Experimental Results

## 4.1 Sensor Data

The Trackman (TM) phased-array Doppler radar has been used by MLB's Statcast system [11] since 2017 to track and characterize batted balls. The TM radar operates in the X-band at approximately 10.5 GHz and is positioned high behind home plate. The measured initial speed $s$ and vertical launch angle $v$ (Fig. 1) for batted balls play an important role in determining batted ball value [10]. In particular, batters tend to achieve the best results for batted balls with an initial speed of greater than 90 miles per hour and a vertical launch angle between 10 and 30 degrees.

## 4.2 Representing Batted Ball Value

Many statistics [17] can be used to quantify a batter's performance on batted balls. Batting average, for example, is the fraction of batted balls that result in a hit but has the deficiency that all hits are given equal value. Slugging percentage allocates different weights to different

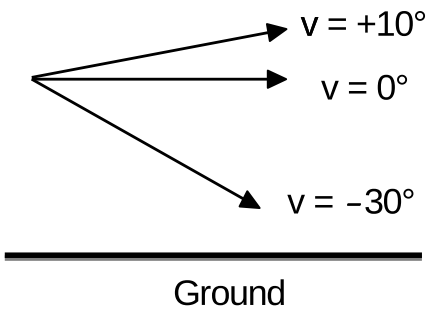


Figure 1: Vertical launch angle $v$

kinds of hits, e.g. single or double, but has been shown to overweight doubles, triples, and home runs. Weighted on base average (wOBA) [23] uses weights for each batted ball outcome that are proportional to run value and, for this reason, we use wOBA to represent batted ball value $R(i, j)$.

## 4.3 Contextual Information

A batted ball with a given set of physical parameters such as $s$ and $v$ occurs in a context that can affect its value. Variation in the outfield geometry across stadiums [9] and variation in the ambient weather conditions [1] can affect the value of a ball hit in the air. The batter's running speed [12] plays a role in determining batted ball value especially for balls hit on the ground. The quality of defenders can also affect the value of a batted ball hit to a given region of the field. These factors cause the batted ball value $\overline{R}(j, k)$ for subset $k$ to vary depending on the distribution of contextual variables for player $j$. We will show later in this section how contextual information can be combined with the batted ball distribution estimates $\widehat{p}_y(j, k)$ to improve the accuracy of the $\widehat{y}_s(j)$ predictions.

## 4.4 Assessing Prediction Accuracy

Statcast data from MLB games in 2019 was employed to evaluate methods for using observed data to predict player performance in unobserved data. After removing bunts from the dataset, each of the $P = 159$ players with at least 300 batted balls during the 2019 season

was considered. Switch-hitters who bat both right-handed and left-handed were regarded as a different batter for each handedness. The first 300 batted balls for each player were divided into an observed set of $N = 150$ batted balls and an unobserved set of $N = 150$ batted balls. The odd batted balls in chronological order for each player defined the observed set and the even batted balls defined the unobserved set. The batted ball value $R(i, j)$ for batted ball $i$ and player $j$ was defined by the wOBA weight for the batted ball result as described in Sec. 4.2. For the 2019 MLB season the wOBA weights are out=0.000, single=0.870, double=1.217, triple=1.529, homerun=1.940, and batter reaches on error= 0.920 [25]. The observed batted ball data was used to generate predictions for the unobserved performance $y(j)$. The accuracy of a set of predictions $\widehat{y}(j)$ is evaluated using the sum of squared errors (SSE)

$$SSE = \sum_{j=1}^{P} (y(j) - \widehat{y}(j))^2 \tag{20}$$

between the unobserved performance and its prediction.

## 4.5 Linear Regression

The linear regression model defined by equation (9) was used to generate the $\widehat{y}(j)$ predictions for the data described in Sec. 4.4. The resulting model is

$$\widehat{y}(j) \;\; = \;\; 0.294x(j) + (1 - 0.294) \cdot 0.402 = 0.294x(j) + 0.284 \tag{21}$$

where the observed batted ball data was used to compute $\alpha(150) = 0.294$ and $\mu_x = 0.402$ as described in Sec. 2.2. This model gives an SSE of 0.647 using equation (20). Two boundary instances of the linear regression model are the naive prediction $\widehat{y}(j) = x(j)$ for $\alpha(N) = 1$ and the baseline prediction $\widehat{y}(j) = \mu_x$ for $\alpha(N) = 0$. For this dataset, the naive prediction gives an SSE of 0.780 and the baseline prediction gives an SSE of 0.743 which are both larger than the SSE obtained using the linear regression model in equation (21). The $\widehat{y}(j)$ prediction lines for the linear regression model and the naive and baseline predictions are shown in Fig. 2 along with the $(x(j), y(j))$ points for each of the 159 players.

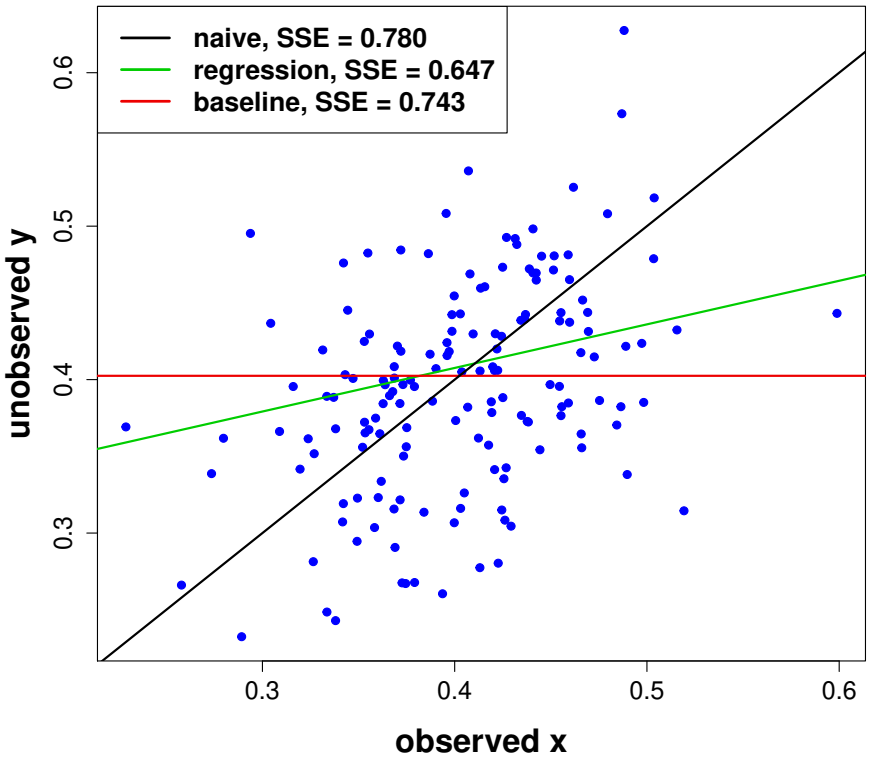


Figure 2: $(x(j), y(j))$ points with naive, regression, and baseline predictions

## 4.6 Measurement Space Partitioning

The measured initial speed and launch angle can be used to represent a batted ball as a point in a two-dimensional $(s, v)$ measurement space. This space can be partitioned into $B$ disjoint subsets as described in Sec. 3.1. In Appendix I, we show that the accuracy of the prediction in equation (17) depends on the partition. In this section we define different ways to partition the $(s, v)$ measurement space and show how training data can be used to optimize partition selection.

### 4.6.1 Partition Definition

The $(s, v)$ space can be divided into an internal region defined by

$$(s_{\min} \leq s < s_{\max}) \quad \text{and} \quad (v_{\min} \leq v < v_{\max})$$

which includes the large majority of batted balls and four boundary regions $B_1, B_2, B_3, B_4$ defined by

$$\begin{aligned} B_1: &\quad s < s_{\min} \\ B_2: &\quad s \geq s_{\max} \\ B_3: &\quad (s_{\min} \leq s < s_{\max}) \text{ and } (v < v_{\min}) \\ B_4: &\quad (s_{\min} \leq s < s_{\max}) \text{ and } (v \geq v_{\max}). \end{aligned}$$

The internal region can be further divided into rectangular subregions $b(i,j)$ of dimension $s_{\text{width}} \times v_{\text{width}}$ which are defined by

$$\begin{aligned} b(i,j): &\quad (s_{\min} + (i-1) * s_{\text{width}}) \leq s < (s_{\min} + i * s_{\text{width}}) \text{ and} \\ &\quad (v_{\min} + (j-1) * v_{\text{width}}) \leq v < (v_{\min} + j * v_{\text{width}}) \end{aligned}$$

so that there are a total of

$$\frac{(s_{\max} - s_{\min})(v_{\max} - v_{\min})}{s_{\text{width}} * v_{\text{width}}}$$

$b(i,j)$ subregions.

We defined the internal and boundary regions for the 2019 data using $s_{\min} = 37.5$ mph, $s_{\max} = 117.5$ mph, $v_{\min} = -75°$, and $v_{\max} = 85°$ which yields an internal region that includes 99.5 percent of all batted balls. The internal region was partitioned into different configurations of fixed-size rectangular subregions $b(i,j)$ where the subregion widths were allowed to vary over the values

$$\begin{aligned} s_{\text{width}} &= 2.5, 5, 10, 20, 40, 80 \text{ mph} \\ v_{\text{width}} &= 2.5, 5, 10, 20, 40, 80, 160 \text{ degrees} \end{aligned}$$

By considering all combinations of the six $s_{\text{width}}$ values and the seven $v_{\text{width}}$ values we can define 42 partitions with each denoted $\mathcal{P}_{s_{\text{width}}, v_{\text{width}}}$ where the boundary regions $B_1, B_2, B_3, B_4$

are the same for each partition. Figure 3, for example, depicts the $\mathcal{P}_{10,40}$ partition with the four boundary regions and thirty-two internal subregions $b(i,j)$ where $b(2,3)$ is explicitly labeled.

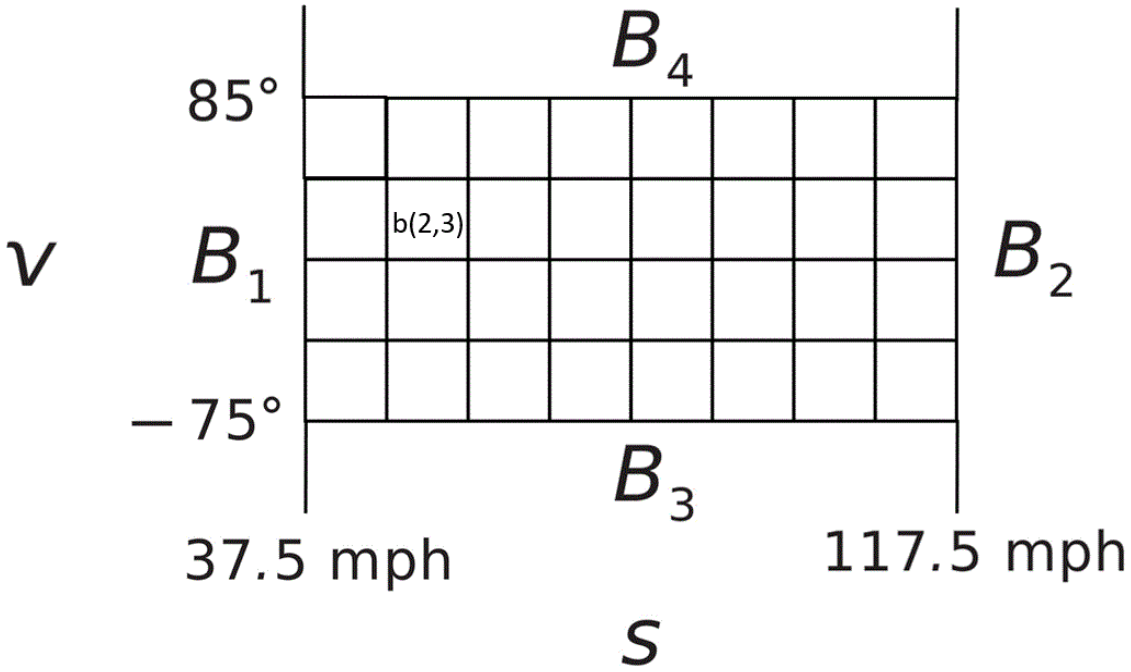


Figure 3: The $\mathcal{P}_{10,40}$ partition of measurement space

The prediction method described in Sec. 3.3 was used to process the observed and unobserved data described in Sec. 4.4 using each of the 42 partitions. For the finer partitions the observed data does not contain enough samples to reliably estimate $\overline{R}(j,k)$ for each $(j,k)$. Therefore, the mean $\overline{R}(k)$ in equation (18) was used to approximate $\overline{R}(j,k)$ for each $j$. The smallest SSE of 0.532 was obtained for $\mathcal{P}_{2.5,40}$ while the largest SSE of 0.743 was obtained for $\mathcal{P}_{80,160}$. If we neglect the effect of the boundary regions, the use of $\mathcal{P}_{80,160}$ is equivalent to the baseline prediction $\widehat{y}(j) = \mu_x$ for which we also reported an SSE of 0.743 in Sec. 4.4.

### 4.6.2 Partition Selection

Partition selection is an important issue since there are large differences in the SSE for different partitions. To address this issue, we examine whether the analysis of previous year data can be used to optimize partition selection for current year data. To this end,

we computed the SSE for each of the 42 partitions defined in Sec. 4.6.1 using 2018 batted ball data arranged as described in Sec. 4.4 for the 2019 data. There were $P = 158$ players with at least 300 batted balls in 2018 that were considered for analysis. Fig. 4 plots the (SSE 2018, SSE 2019) point for each of the 42 partitions and we see that there is a strong correlation between the SSE values for the two years. In particular, the partitions that give the smallest SSE values in 2018 also give the smallest SSE values in 2019. This result suggests that we can use previous year data to select an optimized partition for current year data. The $\mathcal{P}_{5,10}$ partition gives the smallest SSE of .419 on 2018 data. Using this partition for the 2019 data gives an SSE of 0.546 which is close to the smallest value of 0.532 and significantly better than the linear regression SSE of 0.647 reported in Sec. 4.5.

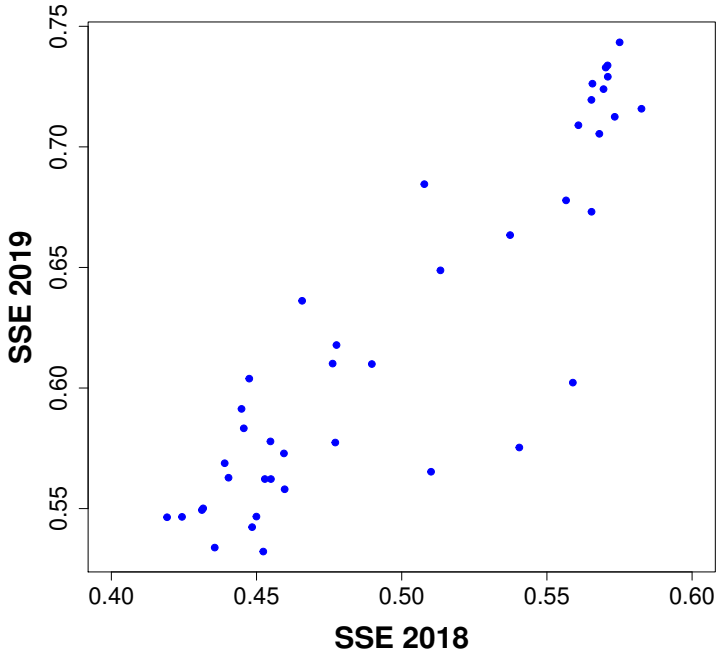


Figure 4: Prediction SSE in 2018 and 2019 for 42 partitions

### 4.6.3 Example

In this section we illustrate the mechanics of the MSP method using the 2019 batted ball data. The example considers the $\mathcal{P}_{5,10}$ partition defined in Sec. 4.6.1 that was selected using

prior year data as described in Sec. 4.6.2. Fig. 5 plots the $\alpha(150, k)$ function and fig. 6 plots the mean $\mu(k)$ function over the subregions $k$ for this partition. The $\alpha(150, k)$ function is approximately in the shape of a rotated V with most of the larger values occurring for $s$ greater than 95 mph. Figures 7 and 8 demonstrate properties of $\alpha(150, k)$ and $\mu(k)$ for specific subregions $S_1$ and $S_2$ of $\mathcal{P}_{5,10}$ defined by

$$\begin{aligned} S_1: &\quad (87.5 \text{ mph} \le s < 92.5 \text{ mph}) \text{ and } (5^\circ \le v < 15^\circ) \\ S_2: &\quad (107.5 \text{ mph} \le s < 112.5 \text{ mph}) \text{ and } (15^\circ \le v < 25^\circ) \end{aligned}$$

which correspond respectively to $b(11, 9)$ and $b(15, 10)$ using the notation in Sec. 4.6.1. The observed data described in Sec. 4.4 gives values of

$$\alpha(150, S_1) = 0.01, \quad \mu(S_1) = 0.017, \quad \alpha(150, S_2) = 0.61, \quad \mu(S_2) = 0.011$$

which predict little correlation between the fraction of batted balls in the observed and unobserved data for $S_1$ and a larger correlation between the fraction of batted balls in the observed and unobserved data for $S_2$. Fig. 7 plots the $P = 159$ points $(p_x(j, S_1), p_y(j, S_1))$ as defined by equations (11) and (12) along with the prediction line from equation (13) where each point in the figure has been moved by a small random amount to increase the visibility of the points. There is little correlation between the $p_x(j, S_1)$ and the $p_y(j, S_1)$ as predicted by the small estimated value of $\alpha(150, S_1)$. Figure 8 is the same plot for $S_2$ where the points have a larger positive correlation as predicted by $\alpha(150, S_2)$. In each figure the red prediction line agrees reasonably well with the structure of the data.

Fig. 9 displays the full observed distribution $p_x(j, k)$ for player $j$ = Jorge Polanco as left-handed batter using $\mathcal{P}_{5,10}$. Figure 10 is the corresponding regressed distribution $\widehat{p}_y(j, k)$ computed using equation (13). We see that the regressed distribution captures the overall structure of $p_x(j, k)$ but is substantially smoother. The regressed distribution results in a talent level estimate $\widehat{y}_s(j)$ in equation (17) of .397. This $\widehat{y}_s(j)$ is much closer to the unobserved performance $y(j)$ of 0.386 than the LR prediction $\widehat{y}(j) = .424$ or the naive prediction of $x(j) = .475$ which corresponds to the observed distribution shown in Fig. 9.

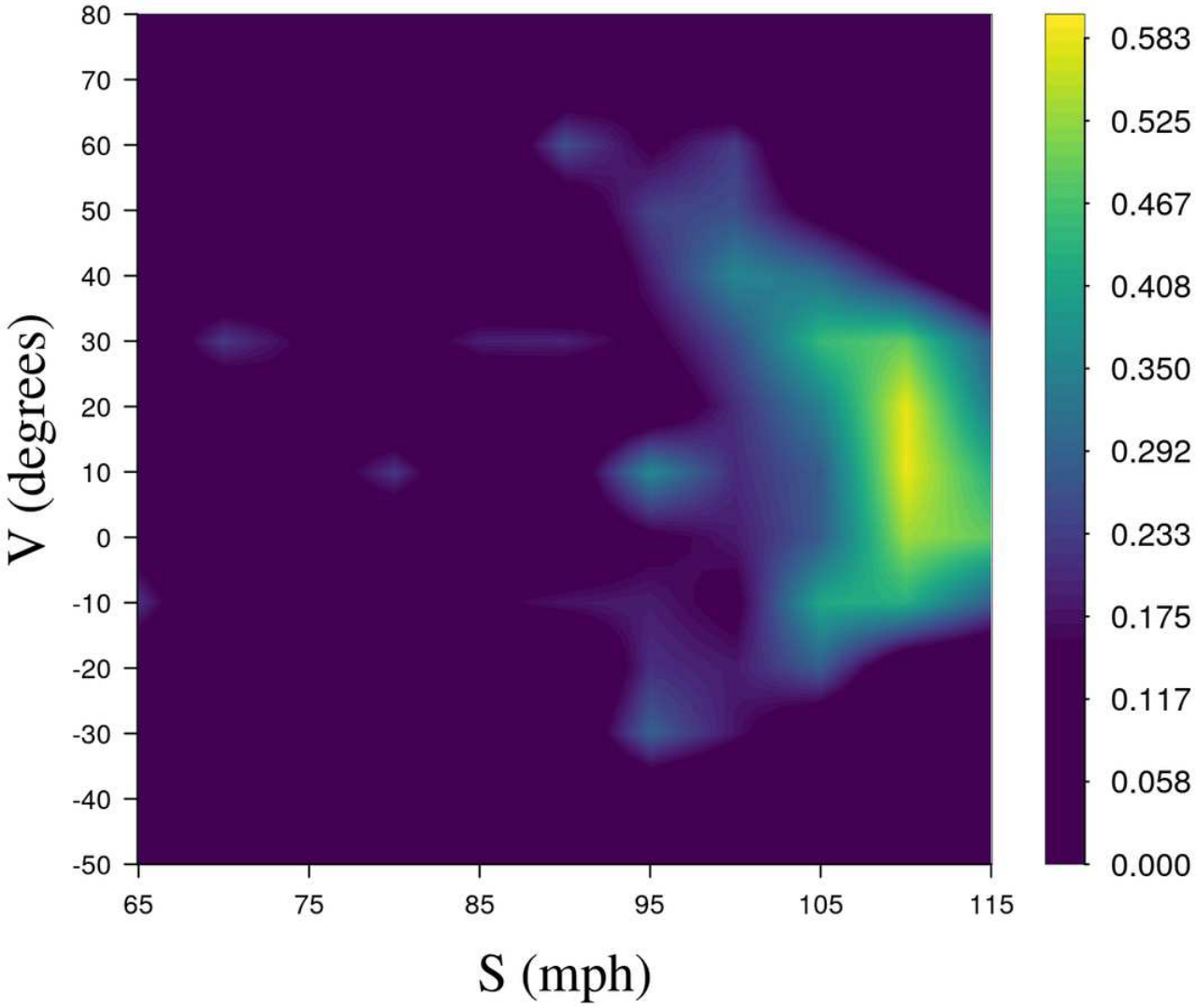


Figure 5: $\alpha(150, k)$ for $\mathcal{P}_{5,10}$ partition

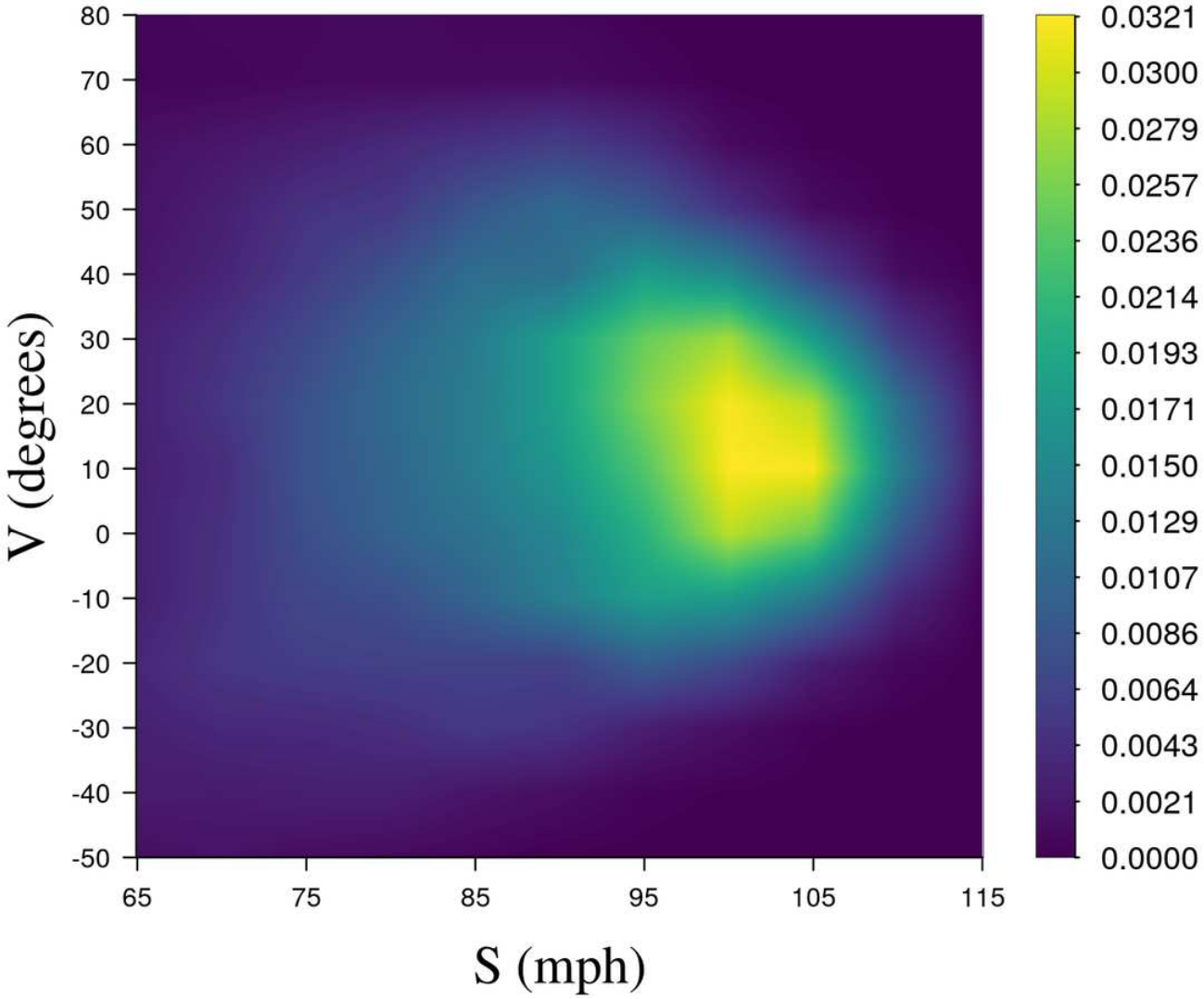


Figure 6: $\mu(k)$ for $\mathcal{P}_{5,10}$ partition

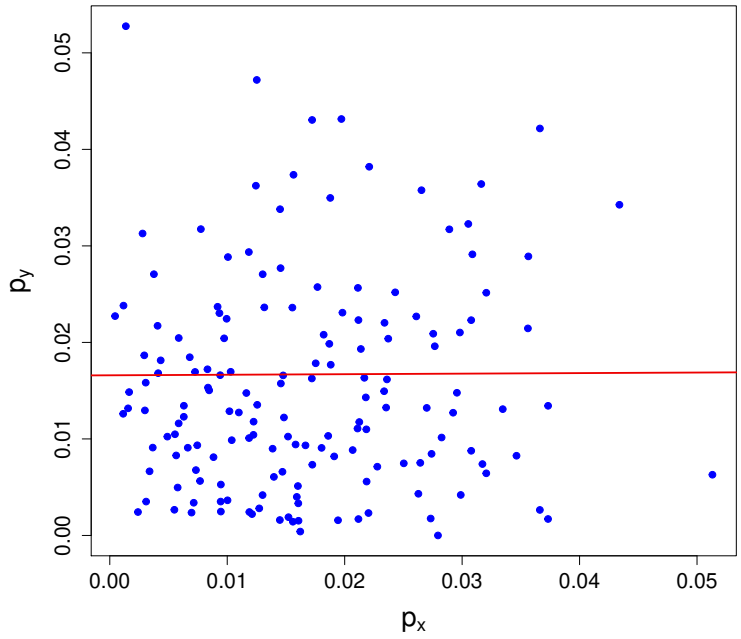


Figure 7: $p_y(j, S_1)$ versus $p_x(j, S_1)$ with $\alpha(150, S_1) = 0.01$

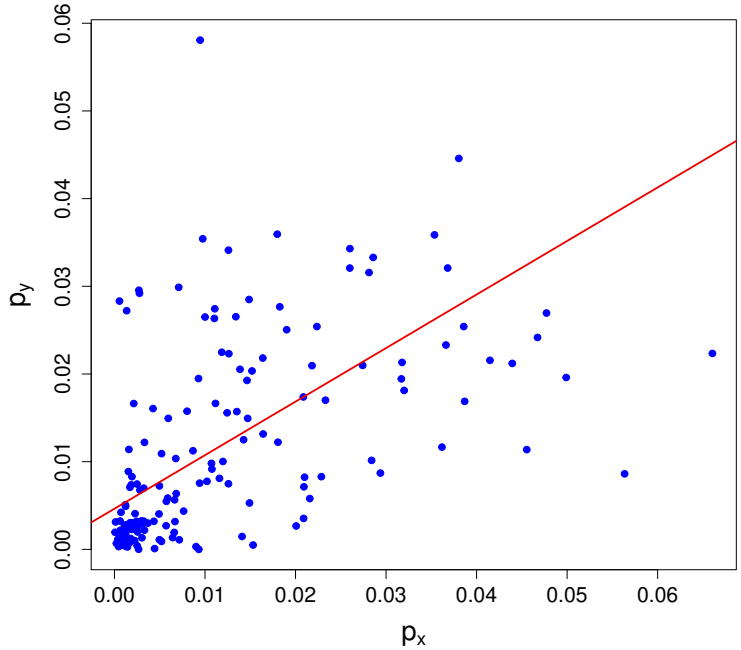


Figure 8: $p_y(j, S_2)$ versus $p_x(j, S_2)$ with $\alpha(150, S_2) = 0.61$

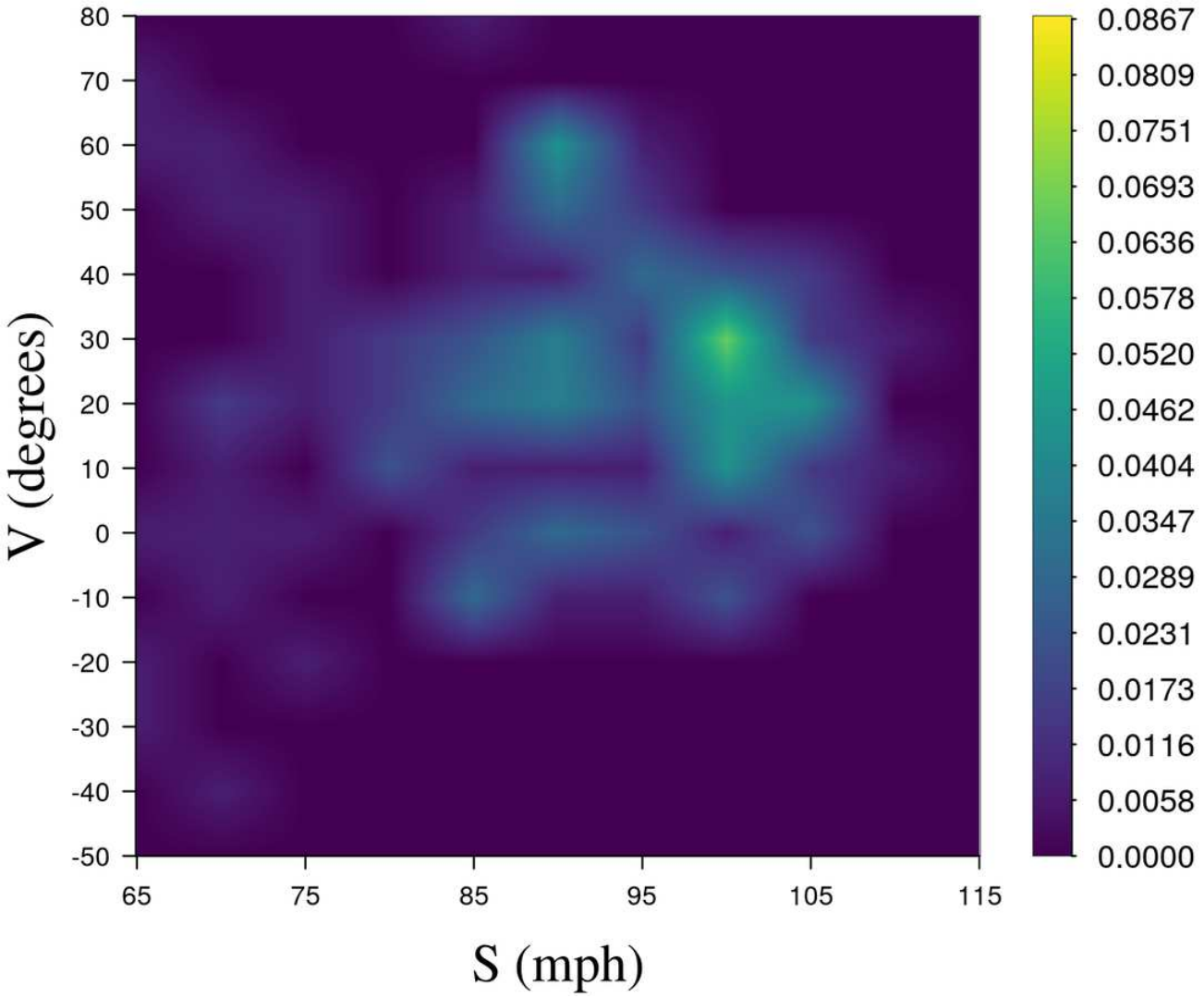


Figure 9: Observed distribution $p_x(j, k)$ for Jorge Polanco as left-handed batter

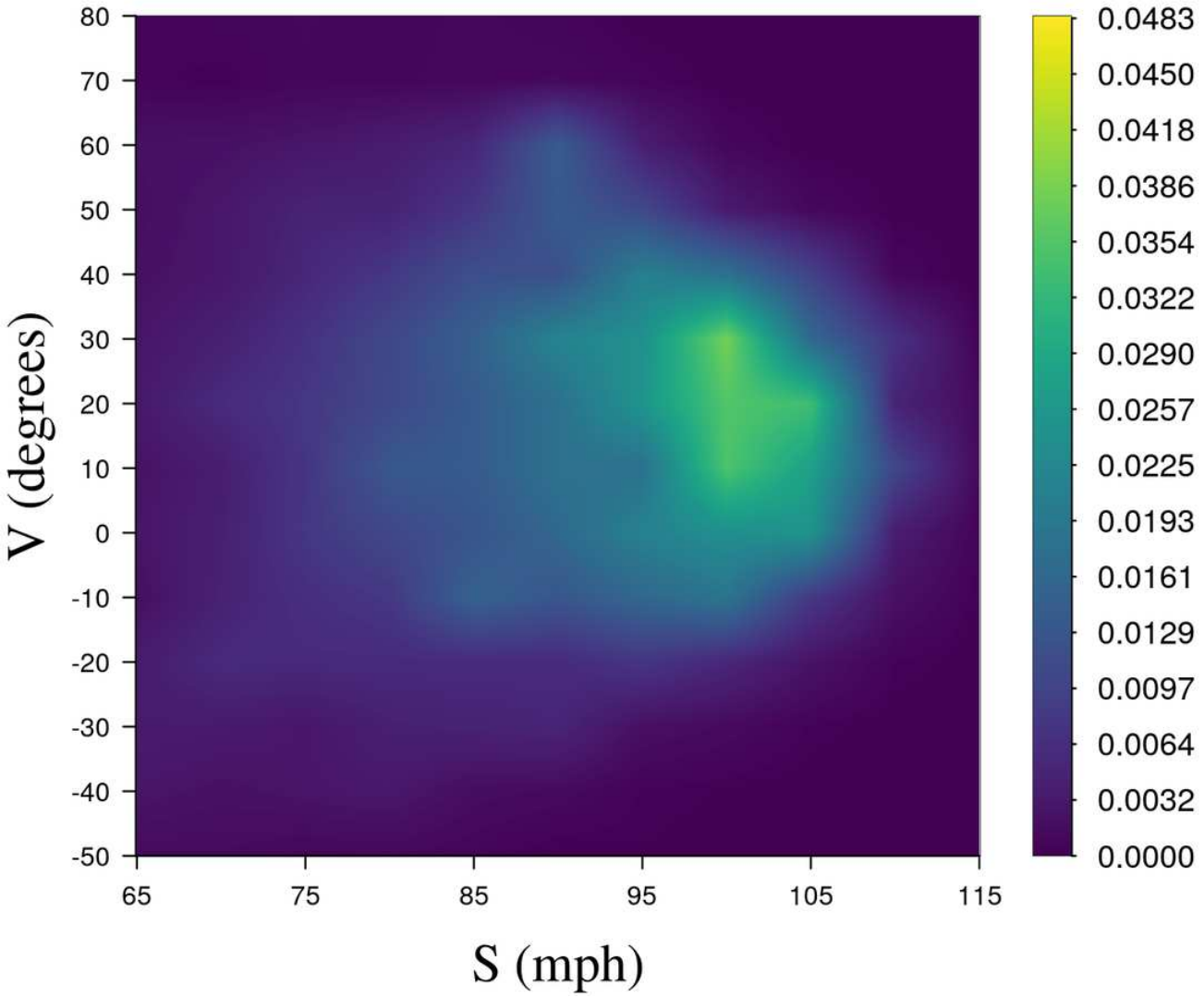


Figure 10: Regressed distribution $\widehat{p}_y(j,k)$ for Jorge Polanco as left-handed batter

### 4.6.4 Comparison with Linear Regression

In this section we compare properties of the LR and MSP predictions. For the data described in Sec. 4.4 the LR prediction is defined by the line (equation (21)) plotted in Fig. 11. This figure also plots the 159 $\widehat{y}_s(j)$ predictions for the same data using the $\mathcal{P}_{5,10}$ partition. We see that players $j_1$ and $j_2$ with the same observed performance $x(j_1) = x(j_2)$ and therefore the same LR prediction $\widehat{y}(j_1) = \widehat{y}(j_2)$ can be assigned different MSP predictions $\widehat{y}_s(j_1) \neq \widehat{y}_s(j_2)$.

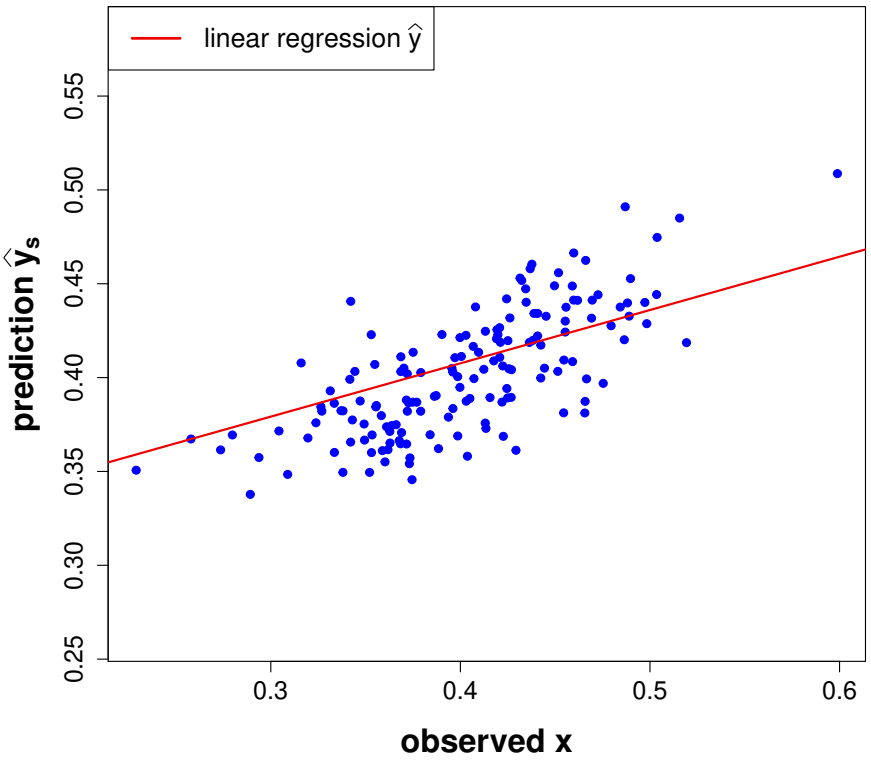


Figure 11: $\widehat{y}_s(j)$ predictions for 159 batters using $\mathcal{P}_{5,10}$ partition and LR line

In Sec. 3.3 we showed that an important difference between $\widehat{y}(j)$ and $\widehat{y}_s(j)$ is that the former is defined using a single $\alpha(N)$ while the latter employs a separate $\alpha(N,k)$ for each subset $k$. Players with an observed batted ball distribution $p_x(j,k)$ that includes a large fraction of batted balls in subsets $k$ with large values of $\alpha(N,k)$ will have less regression to the mean in the calculation of $\widehat{y}_s(j)$ than players with a batted ball distribution that has smaller values of $\alpha(N,k)$. This allows the $\widehat{y}_s(j)$ prediction to adapt the amount of regression

to a player's collection of batted balls. By comparing equations (13) and (17) with the LR model of equation (9) we see that the correlation-weighted expected wOBA

$$C(j) = \sum_{k=1}^{B} \alpha(N,k) p_x(j,k) \overline{R}(k) \tag{22}$$

should capture a large fraction of the variance in the difference $\widehat{y}_s(j) - \widehat{y}(j)$. Fig. 12 is a scatterplot of $\widehat{y}_s(j) - \widehat{y}(j)$ versus $C(j)$ for the $\mathcal{P}_{5,10}$ partition of the 2019 data which shows that the variables have a strong relationship as expressed by a correlation coefficient of 0.87. Thus, $C(j)$ is a batter-controlled component of $\widehat{y}_s(j)$ that measures the combined value and $\alpha$-correlation of a player's batted balls and is strongly related to the deviation of a player's $\widehat{y}_s(j)$ prediction from the LR prediction $\widehat{y}(j)$.

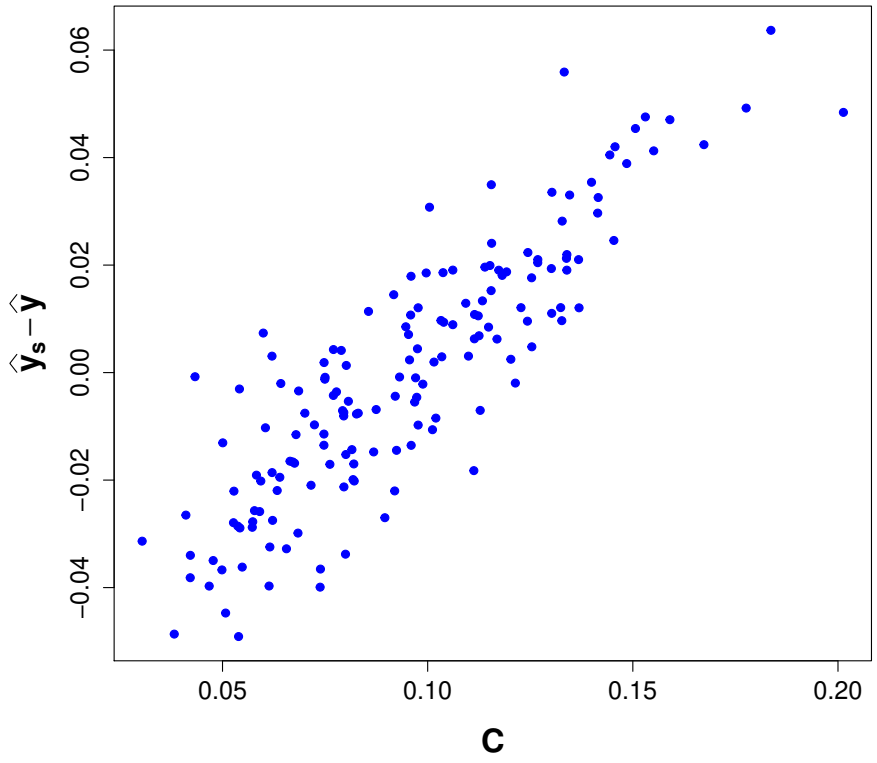


Figure 12: Prediction difference $\widehat{y}_s(j) - \widehat{y}(j)$ versus correlation-weighted expected wOBA $C$

Table 1 considers four players with similar $\widehat{y}(j)$ LR predictions. The table also shows that several of the players have significant differences in correlation-weighted expected wOBA $C$.

The players (Hernandez, DeJong) with below average values of $C$ have negative $\widehat{y}_s - \widehat{y}$ differences while the players (Acuna, Donaldson) with above average values of $C$ have positive $\widehat{y}_s - \widehat{y}$ differences as predicted by Fig. 12. We see from the last two columns of the table that these differences benefit the MSP prediction as the LR prediction error $\widehat{y} - y$ is larger in absolute value than the MSP prediction error $\widehat{y}_s - y$ in each case.

| Player | Hand | $\widehat{y}$ | $C$ | $\widehat{y}_s - \widehat{y}$ | $\widehat{y} - y$ | $\widehat{y}_s - y$ |
|---|---|---|---|---|---|---|
| Cesar Hernandez | Left | .410 | .054 | -.049 | .106 | .057 |
| Paul DeJong | Right | .410 | .082 | -.021 | .067 | .047 |
| Ronald Acuna Jr. | Right | .411 | .144 | .040 | -.077 | -.036 |
| Josh Donaldson | Right | .411 | .146 | .042 | -.081 | -.039 |

Table 1: Players with similar $\widehat{y}$, 2019

#### 4.6.5 Incorporating Contextual Information

In Sec. 4.3 we described several contextual factors that can affect the value of a batted ball with parameters $(s, v)$. Accounting for each of these factors can improve the accuracy of the MSP predictions. In this section we describe a method that can be used to estimate $\overline{R}(j, k)$ in equation (17) to account for the effects of varying outfield geometry and atmospheric conditions across ballparks. Since a player $j$ typically plays about half of his games in a single home park these effects can have a significant impact on $\overline{R}(j, k)$. As an example, Figure 13 plots the outfield boundaries for Fenway Park in Boston and Yankee Stadium in New York where the batter's location is at home plate in the lower left corner. A shorter distance from home plate to the outfield boundary typically improves the batter's likelihood of a home run for a batted ball hit in the air. In addition, the altitude of the ballpark affects the air density which plays an important role in determining how far a batted ball will carry [1]. The outfield geometry can affect players differently depending on whether they bat right-handed or left-handed since right-handed batters tend to hit most of their home runs to left field while left-handed batters tend to hit most of their home runs to right field.

We will learn ballpark-dependent batted ball values from 2018 data and use these values to process the 2019 data described in Sec. 4.4. The value of batted balls in a subset $k$ will

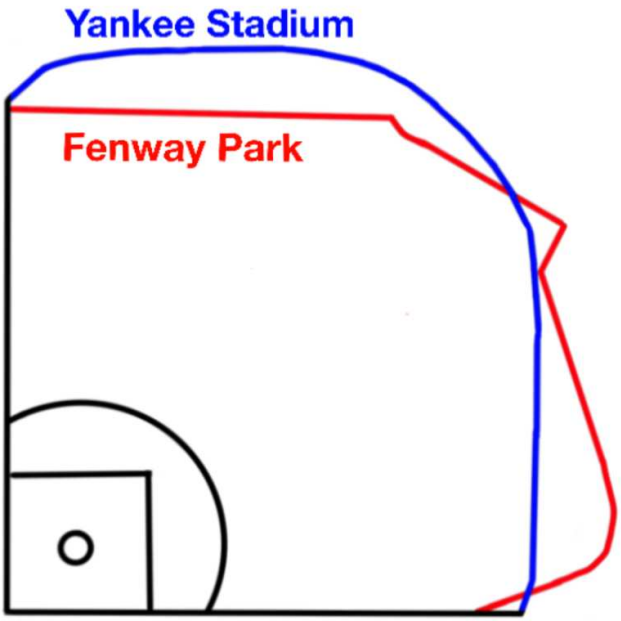


Figure 13: Outfield geometry for Fenway Park and Yankee Stadium

depend on the quality of the fielders that defend against these batted balls. The home team defenders are on the field about half of the time for games played in park $p$ which can cause bias in batted ball values for a given $(k, p)$. Define $R_h(k, p)$ as the average wOBA value for batted balls hit by batters of hand $h$ in subset $k$ and park $p$ with the visiting team on defense in 2018. Let $\overline{R}_h(k)$ be the average wOBA value for batted balls hit by all batters of hand $h$ in subset $k$ in all parks in 2018.

For $(h, k, p)$ groups that correspond to vertical launch angles $v \geq 15^\circ$ and include at least ten batted balls in the calculation of $R_h(k, p)$ we compute the factor

$$F_h(k, p) = \frac{R_h(k, p)}{\overline{R}_h(k)} \tag{23}$$

where otherwise $F_h(k, p)$ is set to 1. For a player $j$ of hand $h$ with home park $p$ in 2019 we define

$$\overline{R}(j, k) = 0.5 \left[\overline{R}(k) + \overline{R}(k) F_h(k, p)\right] \tag{24}$$

where $\overline{R}(k)$ is defined in equation (18) and the 0.5 accounts for the fact that a player plays approximately half of his games in the same home ballpark. The $\overline{R}(j, k)$ can be used to

improve the accuracy of the prediction in equation (17).

To illustrate this process we consider the $b(13, 12)$ subregion of the $\mathcal{P}_{5,10}$ partition which is defined by

$$(97.5 \text{ mph} \leq s < 102.5 \text{ mph}) \text{ and } (35^\circ \leq v < 45^\circ).$$

For this subregion we have the $\overline{R}(j, k)$ values shown in Table 2 which demonstrate that right-handed batters have an advantage in Fenway Park and left-handed batters have an advantage in Yankee Stadium. These observations are consistent with the outfield geometries shown in Fig. 13.

| Hand (h) | Ballpark (p) | $\overline{R}(j, k)$ |
|---|---|---|
| Right | Fenway Park | .557 |
| Left | Fenway Park | .381 |
| Right | Yankee Stadium | .378 |
| Left | Yankee Stadium | .490 |

Table 2: $\overline{R}(j, k)$ for player $j$ of hand $h$ with home park $p$ for $b(13, 12)$

Let $\widehat{y}_{s1}(j)$ be the prediction of equation (17) using $\overline{R}(j, k) = \overline{R}(k)$ and let $\widehat{y}_{s2}(j)$ be the prediction using $\overline{R}(j, k)$ as defined by equation (24). As reported in Sec. 4.6.2, $\widehat{y}_{s1}(j)$ produces an SSE of 0.546 for partition $\mathcal{P}_{5,10}$ on the data described in Sec. 4.4. The use of $\widehat{y}_{s2}(j)$ reduces the SSE to 0.526.

Table 3 presents the five players $j$ with the largest differences $\widehat{y}_{s2}(j) - \widehat{y}_{s1}(j)$ and Table 4 presents the five players with the smallest differences $\widehat{y}_{s2}(j) - \widehat{y}_{s1}(j)$. Thus, the players in Table 3 are expected to benefit from their home ballpark while the players in Table 4 are expected to be hindered by their home ballpark. The parks represented in Table 3 are known to benefit batters. Coors Field in Denver has an altitude of 5197 feet which enables batted balls to carry longer distances and Citizens Bank Park in Philadelphia has an outfield geometry which is beneficial to right-handed batters. Similarly, both Busch Stadium in St. Louis and Marlins Park in Miami which appear in Table 4 have outfield geometries that are detrimental to right-handed batters. The last two columns in each table give the prediction errors $E_1(j) = \widehat{y}_{s1}(j) - y(j)$ and $E_2(j) = \widehat{y}_{s2}(j) - y(j)$ where $y(j)$

is the unobserved performance. $E_1(j)$ is negative for each of the players in Table 3 which is consistent with the expectation that these players should benefit from their home ballpark while $E_1(j)$ is positive for four of the five players in Table 4 which is consistent with the expectation that these players should be hindered by their home ballpark. We see that for nine of the ten players in the two tables we have $|E_1| > |E_2|$ so that the use of home park information reduces the prediction error.

| Player | Hand | Home Ballpark | $\widehat{y}_{s2} - \widehat{y}_{s1}$ | $E_1$ | $E_2$ |
|---|---|---|---|---|---|
| Trevor Story | Right | Coors Field | .020 | -.080 | -.061 |
| Nolan Arenado | Right | Coors Field | .019 | -.071 | -.052 |
| Ian Desmond | Right | Coors Field | .018 | -.016 | .001 |
| Rhys Hoskins | Right | Citizens Bank Park | .017 | -.060 | -.042 |
| Scott Kingery | Right | Citizens Bank Park | .016 | -.029 | -.013 |

Table 3: Players with largest $\widehat{y}_{s2} - \widehat{y}_{s1}$, 2019

| Player | Hand | Home Ballpark | $\widehat{y}_{s2} - \widehat{y}_{s1}$ | $E_1$ | $E_2$ |
|---|---|---|---|---|---|
| Marcell Ozuna | Right | Busch Stadium | -.026 | .088 | .062 |
| Paul Goldschmidt | Right | Busch Stadium | -.023 | -.020 | -.043 |
| Paul DeJong | Right | Busch Stadium | -.021 | .047 | .025 |
| Yadier Molina | Right | Busch Stadium | -.020 | .080 | .060 |
| Brian Anderson | Right | Marlins Park | -.012 | .045 | .033 |

Table 4: Players with smallest $\widehat{y}_{s2} - \widehat{y}_{s1}$, 2019

# 5 Conclusion

Sensor systems that acquire large sets of data have been deployed to document the mechanics of several sports including baseball [11], basketball [26], football [5], and golf [24] at unprecedented levels of detail. Data-driven techniques have been applied to these sensor measurements to discover new skills [15], quantify known skills with greater accuracy [22], and understand biomechanical principles [14] to improve performance and prevent injury. This information has been used by professional sports teams in search of an advantage [16]. While there are large disparities in the financial resources available to teams, the use of

data-driven models has enabled small market franchises to compete successfully against their more affluent opponents [18].

We have used ball-tracking radar data to show that the predictive value of a batted ball in baseball depends on its speed and vertical launch angle. This constraint enables a batted ball distribution to be estimated from a set of observations using a regression process that adapts to a player's particular collection of batted balls. We showed that these estimated distributions can be used to make improved predictions about unobserved data. The methodology can be adapted to include additional sensor measurements for properties such as spin and horizontal angle as they become available. Since the approach is based on estimating distributions defined over a partition of measurement space, fine-grained contextual adjustments can be included to improve the accuracy of the predictions. The measurement space partitioning process can be used for several applications in baseball including performance forecasting and defensive positioning as well as for a range of other estimation and prediction tasks involving large sets of multidimensional sensor data.

# Appendix I: Dependence of Prediction Accuracy on the Partition

The error in a prediction generated using the MSP approach depends on the partition of measurement space. Using equation (17), we can write the unobserved performance for player $j$ as

$$y(j) = \sum_{k=1}^{B} \left(\widehat{p}_y(j,k) + \epsilon_p(j,k)\right) \left(\overline{R}(j,k) + \epsilon_R(j,k)\right). \tag{25}$$

The error terms are defined by $\epsilon_p(j,k) = p_y(j,k) - \widehat{p}_y(j,k)$ and $\epsilon_R(j,k) = \overline{R}_y(j,k) - \overline{R}(j,k)$ where $\overline{R}_y(j,k)$ is the average value of the unobserved batted balls in subset $k$ for player $j$. The prediction error is given by

$$y(j) - \widehat{y}_s(j) = \sum_{k=1}^{B} \left[\widehat{p}_y(j,k)\epsilon_R(j,k) + \overline{R}(j,k)\epsilon_p(j,k) + \epsilon_p(j,k)\epsilon_R(j,k)\right] \tag{26}$$

where each term in the sum depends on the subset $k$.

The error terms have a complex dependence on the group of subsets that define the partition. Reducing the size of the $\epsilon_R(j,k)$ error depends on balancing the competing goals of using subsets $k$ that include enough data to estimate $\overline{R}(j,k)$ accurately but which also allow a single $\overline{R}(j,k)$ to be representative of any particular sample within a subset that might occur in $y(j)$. The variance of the $\epsilon_p(j,k)$ error is given by [7]

$$\text{VAR}\left[\epsilon_p(j,k)\right] = \sigma_p^2(k)\left(1-\alpha^2(N,k)\right) \tag{27}$$

where $\sigma_p^2(k)$ is the variance of $p_x(j,k)$ over batters $j$ for subset $k$. Thus, $\text{VAR}\left[\epsilon_p(j,k)\right]$ depends on both the distribution of the $p_x(j,k)$ and the $\alpha(N,k)$. Since the error terms and the prediction error in equation (26) have a complex dependence on the interaction between the measurement space partition and the structure of the data we use a learning process for partition selection as described in Sec. 4.6.2.

# Appendix II: Applying MSP to First-Half 2021 Data

In this Appendix we apply the MSP approach to batted ball data collected before the All-Star break during the 2021 MLB season. The study considers the 185 batters with at least 150 batted balls during this period. The $\mathcal{P}_{5,10}$ partition was used and the $\alpha(150,k)$ values were estimated using the first 150 batted balls for each of the 185 batters. The subset means $\overline{R}(k)$ were used to approximate $\overline{R}(j,k)$ for each player $j$. The full set of first half batted balls for each player was used to compute $\widehat{y}_s(j)$ where the $\alpha(150,k)$ values were adjusted to $\alpha(N'(j),k)$ using the Spearman-Brown formula to regress each distribution according to each individual player's batted ball distribution and number of batted balls $N'(j)$. The result is an estimate of true talent wOBA on contact (wOBAcon) that has removed all contextual information (ballpark, batter running speed, atmospheric conditions, defense, etc.). The accuracy of wOBAcon predictions can be improved by incorporating context into the $\overline{R}(j,k)$ for each player $j$ as demonstrated in Sec. 4.6.5. Table 5 presents the context-invariant true talent wOBAcon leaders based on first-half 2021 data.

| Player | True Talent wOBAcon |
|---|---|
| Shohei Ohtani | .556 |
| Fernando Tatis Jr. | .518 |
| Giancarlo Stanton | .515 |
| Vladimir Guerrero Jr. | .506 |
| Ronald Acuna Jr. | .506 |
| Aaron Judge | .500 |
| Kyle Schwarber | .493 |
| Joey Gallo | .488 |
| Tyler O'Neill | .480 |
| Nelson Cruz | .474 |
| Yordan Alvarez | .470 |
| Bryce Harper | .470 |
| Rafael Devers | .469 |
| Pete Alonso | .460 |
| Matt Olson | .460 |

Table 5: Context-invariant $\widehat{y}_s(j)$ estimate of wOBAcon using first-half 2021 data

# Acknowledgment

We thank Tom Tango for inspiring this research. The batted ball data was obtained at baseballsavant.com.

# References

[1] A.T. Bahill, D. Baldwin, and J. Ramberg. Effects of altitude and atmospheric conditions on the flight of a baseball. *International Journal of Sports Science and Engineering*, 3(2):109–128, 2009.

[2] B. Baumer and A. Zimbalist. *The Sabermetric Revolution: Assessing the growth of analytics in baseball*. University of Pennsylvania Press, Philadelphia, 2014.

[3] J. Brosnan, A. McNitt, and T. Serensits. Effects of varying surface characteristics on the hardness and traction of baseball field playing surfaces. *International Turfgrass Society Research Journal*, 11:1053–1065, 2009.

[4] W. Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, October 1910.

[5] K. Clark. (Dec. 19, 2018). The NFL's analytics revolution has arrived [Online]. Available: www.theringer.com/nfl/2018/12/19/18148153/nfl-analytics-revolution.

[6] L. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[7] N. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, 3rd edition, 1998.

[8] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

[9] D. Gartland. (March 24, 2021). MLB outfield walls, ranked [Online]. Available: si.com/mlb/2021/03/24/mlb-outfield-walls-ranked-fenway-park-yankee-stadium.

[10] G. Healey. Learning, visualizing, and assessing a model for the intrinsic value of a batted ball. *IEEE Access*, 5:13811–13822, 2017.

[11] G. Healey. The new Moneyball: How ballpark sensors are changing baseball. *Proceedings of the IEEE*, 105(11):1999–2002, 2017.

[12] G. Healey. Combining radar and optical sensor data to measure player value in baseball. *Sensors*, 21(1):64, 2021.

[13] W. James and C. Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379, 1961.

[14] J. Lemire. (Dec. 3, 2019). KinaTrax's magic leap: a new way to see data [Online]. Available: sporttechie.com/mlb-kinatrax-ar-biomechanics-baseball-data.

[15] B. Lindbergh. (May 16, 2013). The art of pitch framing [Online]. Available: grantland.com/features/ studying-art-pitch-framing-catchers-such-francisco-cervelli-chris-stewart-jose-molina-others/.

[16] B. Lindbergh and T. Sawchik. *The MVP machine: How baseball's new nonconformists are using data to build better players*. Basic Books, New York, NY, 2019.

[17] L. Panas. *Beyond Batting Average*. Lulu Press, Morrisville, North Carolina, 2010.

[18] T. Sawchik. *Big data baseball*. Flatiron Books, New York, NY, 2016.

[19] N. Silver. Why was Kevin Maas a bust? In J. Keri, editor, *Baseball between the numbers*, pages 253–271. Basic Books, New York, 2006.

[20] C. Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, October 1910.

[21] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:197–208, 1956.

[22] T. Tango. (Jan. 13, 2020). Introducing infield outs above average [Online]. Available: technology.mlblogs.com/introducing-infield-outs-above-average-6467e61a98dc.

[23] T. Tango, M. Lichtman, and A. Dolphin. *The Book: Playing the Percentages in Baseball*. Potomac Books, Dulles, Virgina, 2007.

[24] S. Wang, Y. Xu, Y. Zheng, M. Zhu, H. Yao, and Z. Xiao. Tracking a golf ball with high-speed stereo vision system. *IEEE Transactions on Instrumentation and Measurement*, 68(8):2742–2754, August 2019.

[25] wOBA and FIP constants [Online]. Available: www.fangraphs.com/guts.aspx?type=cn.

[26] M. Woo. (Dec. 21, 2018). Artificial intelligence in NBA basketball [Online]. Available: www.insidescience.org/news/artificial-intelligence-nba-basketball.

[27] R. Zeller and E. Carmines. *Measurement in the Social Sciences: The Link Between Theory and Data*. Cambridge University Press, 1980.