Including Batter Sprint Speed in the Calculation of the Intrinsic Value of a Batted Ball

William Melville
willmel@byu.edu

Advisor: Sean Warnick
sean@cs.byu.edu

Information and Decision Algorithms Laboratories
(IDeA Labs)
Brigham Young University

July 2019

Abstract: This paper describes the process used to create two different models to define the intrinsic value of a batted ball. The first model, which was originally created by Dr. Glenn Healey, maps a batted ball's speed, vertical angle, and horizontal angle to an intrinsic value. This first model has the property that above average runners tend to be underrated and below average runners tend to be overrated by the intrinsic value. Thus, the second model described in this work attempts to address this property by including the batter's sprint speed in the mapping to an intrinsic value. A visual representation of the first mapping called the wOBA cube is presented as well as a visual representation of the second mapping called the wOBA tesseract. The mean absolute deviation between the intrinsic statistic and the outcomes-based statistic is used to compare the accuracy of both intrinsic values, and it is determined that the sprint speed intrinsic value is at least as accurate as the non-sprint speed intrinsic value. The two intrinsic statistics' reliabilities are compared using Cronbach's alpha, and it is determined that they are similarly reliable and that they are both more reliable than the outcomes-based statistic. Finally, the ten most under and overrated players, in terms of the difference between their outcomes-based statistic and their intrinsic statistic, are identified for both intrinsic statistics, and it is determined that the sprint speed intrinsic value underrates fast runners and overrates slow runners less frequently than the non-sprint speed intrinsic value.

Keywords: baseball, batted balls, Bayes, intrinsic, modeling

1 Introduction

In Glenn Healey's article *Learning, visualizing, and assessing a model for the intrinsic value of a batted ball* [1], he used a Bayesian model to map a batted ball's speed, vertical angle, and horizontal angle to an intrinsic value. He then showed that the intrinsic value statistic he derived has a higher reliability than the outcomes-based statistic. In baseball, there are many confounding variables that can affect the outcome of any given batted ball. These confounding variables include things like the quality of the defenders, the ballpark, and the weather. Thus, the intrinsic value model could potentially improve upon the evaluation of baseball hitters because it ignores these confounding effects and focuses only on what the batter can control.

Healey's intrinsic value model mapped a batted ball vector, $x = (s, v, h)$, to an intrinsic value, where s is the batted ball's launch speed, v is its vertical launch angle, and h is its horizontal angle. In a paper he wrote for *the Hardball Times*, Healey noted that many players with a large difference between their outcomes-based statistic, O, which was given by weighted on base average on contact or wOBAcon, and their intrinsic statistic, I, often had an above average running speed. Likewise, players with a small $O - I$ value had a below average running speed [2][3]. It seems that fast runners are able to exceed the expectation given by the intrinsic value of their batted balls, whereas slow runners have a hard time meeting that expectation. Intuitively, a slow ground ball hit to the third baseman is less valuable to a slow baserunner than it is to a fast baserunner because the fast runner may be able to use his speed to beat the throw to first base, whereas the slow runner is unlikely to beat the throw to first. Likewise, above average runners can sometimes stretch what are normally singles or doubles into doubles or triples using their above average speed. These observations suggest that the intrinsic value of a batted ball as Healey defines it is likely to underrate fast runners and overrate slow runners. Thus, a potential improvement to the intrinsic value model would consider the batter's sprint speed as part of the intrinsic value of the batted ball.

In this paper, an updated version of Healey's intrinsic value model that considers the batter's sprint speed is presented. Healey's intrinsic value statistic will be referred to throughout as I_{ns}, which stands for the intrinsic value statistic with no sprint speed parameter. The updated version will be referred to throughout as I_s, which stands for the intrinsic value statistic with a sprint speed parameter. I will refer to one or both intrinsic value statistics. Be aware of the difference between I and $I(x)$. $I(x)$ will refer to the intrinsic value of a batted ball with batted ball vector x, whereas I will refer to the intrinsic value statistic, which is simply the average value of $I(x)$ for all of a player's batted ball vectors. Visuals that depict $I_{ns}(x)$ and $I_s(x)$ for varying vertical and horizontal angles are also presented in this paper. Finally, I_{ns} and I_s are compared in terms of accuracy, reliability, and their $O - I$ values. Note that O will refer to the outcomes-based statistic, wOBAcon, throughout this paper.

2 Data

Batted ball data from the 2017 MLB season were obtained from Statcast using data scraping functions in the baseballr package of the R programming language [4][5]. This provided the batter, launch speed, launch angle, horizontal angle, and wOBAcon of every batted ball in 2017. The batters' sprint speeds were also obtained separately from Statcast and are defined for each player as "feet per second in a player's fastest one-second window" [4]. FanGraphs provided all the weights used to calculate $I(x)$ except for the weight of reaching base on an error, which was taken from Tom Tango's *The Book* [6][7].

3 Methodology

In Healey's article, he estimated the probability of an outcome, R_j, given a batted ball vector, $x = (s, v, h)$, using Bayes' Theorem, which in this context is defined as:

$$P(R_j|x) = \frac{p(x|R_j)P(R_j)}{p(x)}. \tag{1}$$

For the outcome, R_j, there were six possibilities: out, single, double, triple, home run, and reached base on an error [8]. These outcomes were given by $j = 0, ..., 5$ respectively. A kernel density estimate for $p(x|R_j)$ was obtained using the formula

$$\hat{p}(x|R_j) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i) \tag{2}$$

where x_i is the i^{th} batted ball vector from a set of n observed batted balls with outcome R_j, and K is the zero-mean Gaussian kernel given by

$$K(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp[-0.5x^T\Sigma^{-1}x] \tag{3}$$

where d is the dimension of the probability density function, and Σ is the $d \times d$ covariance matrix. Now, as Healey discussed in his work, Σ gives the amount and orientation of the smoothing and is often chosen to be the identity matrix times a scalar. However, Healey wanted to allow for different amounts of smoothing in different directions, so he adopted a diagonal model for Σ with variance elements given by $(\sigma_s^2, \sigma_v^2, \sigma_h^2)$. Since Healey's vector was given by $x = (s, v, h)$ and was thus three-dimensional, he was able to rewrite (3) as

$$K(x) = \frac{1}{(2\pi)^{3/2}\sigma_s\sigma_v\sigma_h}\exp[-0.5(\frac{s^2}{\sigma_s^2} + \frac{v^2}{\sigma_v^2} + \frac{h^2}{\sigma_h^2})]. \tag{4}$$

Choosing the smoothing bandwidth parameters is the most important part of kernel density estimation. The size of the bandwidth determines the widths of the Gaussian kernels to put at each data point. If the bandwidth parameters are too small, the estimated density is too spikey, which increases the variance in the probability estimates. On the other hand, if the bandwidths are too large, the estimated density is too smooth, which increases the bias in the probability estimates. The optimal bandwidths find an appropriate balance between bias and variance and can often be obtained through maximum likelihood estimation. Healey estimated the three unknown bandwidth parameters, $\sigma = (\sigma_s, \sigma_v, \sigma_h)$, by maximizing the pseudolikelihood,

$$\sigma_k^* = \arg\max_{\sigma} \prod_{x_i \in S_k} \hat{p}(x_i|R_j) \tag{5}$$

where $k = 1, ..., M$ and S_k is one of M disjoint sets of observed batted balls with outcome R_j. The batted balls with outcome R_j that were not in S_k were used for the n observed batted ball vectors with outcome R_j in (2). Then, the optimal bandwidth vector, σ^*, was simply the average of the M different σ_k^* vectors. When Healey did this, he set M equal to two, and he obtained his two disjoint sets by separating batted balls hit on even numbered days from batted balls hit on odd numbered days. He then removed batted balls from the larger of the disjoint sets until both contained n_v batted balls, where n_v is the number of batted balls in the smaller of the two disjoint sets. Then, for each set, S_k, he performed a three-dimensional search in σ_s, σ_v, and σ_h using a step size of 0.1 to find the solution to (5). After finding the optimal bandwidth vector, σ^*, for each outcome, he was able to estimate $\hat{p}(x|R_j)$ using (2). He then estimated $P(R_j)$ by dividing the total number of times that outcome R_j occurred in the batted ball data by the total number of batted balls. Finally, he estimated $p(x)$ using the formula

$$\hat{p}(x) = \sum_{j=0}^{5} \hat{p}(x|R_j)\hat{P}(R_j). \tag{6}$$

This gave him everything he needed to use Bayes' Theorem to estimate $P(R_j|x)$ in (1). From there he calculated the intrinsic value, $I_{ns}(x)$, using the formula

$$I(x) = \sum_{j=0}^{5} w_j P(R_j|x) \tag{7}$$

where w_j is the wOBA value or weight associated with each of the six possible outcomes.

In order to compare the 2017 I_{ns} statistics with the I_s statistics, $I_{ns}(x)$ first had to be calculated for every batted ball in 2017. Since Healey calculated $I_{ns}(x)$ for the 2014 batted balls, new bandwidth vectors, σ^*, had to be calculated for 2017. The method used in calculating the new bandwidth parameters was very similar to Healey's method with just some slight adjustments. Like Healey, the batted balls for each outcome were split into two disjoint sets depending on if they were hit on an odd or an even day. Then for each outcome other than outs, a three-dimensional search with a step size of 0.2 was performed on σ_s, σ_v, and σ_h to find the solution to (5). Then σ^* for each outcome was the average of the two searches. The reason the step size was changed from 0.1, which is the step size Healey used, to 0.2 was simply to save on computational time. For outs, the two disjoint sets were divided into ten. Then, for each of the ten sets of two, equation (5) and the method described above were used to find σ_j^* for j =1,...,10. The optimal bandwidth vector was then the average of all of the σ_j^* vectors. The reason why the outs' bandwidth parameter search was done using this method rather than Healey's method was because the implementation of (2) was slow for the large number of batted balls that resulted in outs. Since (2) gets called on frequently in (5), the outs batted balls were split up even further to reduce the computational time of (2). There were about ten times as many outs as doubles. Since calculating the bandwidth for doubles did not take very long, the outs were split by ten to reduce the problem to essentially calculating ten more doubles' bandwidths, which was much faster than trying to calculate the outs bandwidth without splitting the data any further.

Table 1 gives the bandwidth parameters that were found for the 2017 season. Other than the slight changes made to save time in the calculation of the band-

σ^*	Outs	1B	2B	3B	HR	RBOE
σ_s	3.12	1.9	3	3.1	1.5	5.5
σ_v	2.59	1.9	3.1	4.3	1.6	2.1
σ_h	3.8	4.7	2	2.3	3	8.6

Table 1: 2017 Bandwidth Parameters

width parameters, all other calculations were performed using the same methodology that Healey used. The weights used in (7) were $w_0 = 0$, $w_1 = 0.877$, $w_2 = 1.232$, $w_3 = 1.552$, $w_4 = 1.98$, and $w_5 = 0.92$ [6][7].

To calculate $I_s(x)$ for the batted balls in 2017, the speed of the batter had to be included. Thus, the batted ball vector, $x = (s, v, h)$, became $x = (s, v, h, ss)$, where ss is the sprint speed of the batter as obtained from Statcast [4]. This required making adjustments to (4), and it required estimating new bandwidth parameters given by $\sigma^* = (\sigma_s, \sigma_v, \sigma_h, \sigma_{ss})$. Adjusting (4) was fairly simple. The dimension, d, changed from three to four, and the covariance matrix, Σ, became the diagonal matrix with variance elements given by $(\sigma_s^2, \sigma_v^2, \sigma_h^2, \sigma_{ss}^2)$. This led to the updated equation for K,

$$K(x) = \frac{\exp[-0.5(\frac{s^2}{\sigma_s^2} + \frac{v^2}{\sigma_v^2} + \frac{h^2}{\sigma_h^2} + \frac{ss^2}{\sigma_{ss}^2})]}{(2\pi)^2 \sigma_s \sigma_v \sigma_h \sigma_{ss}}. \tag{8}$$

All of the other equations remained unchanged except that x was (s, v, h, ss) instead of (s, v, h).

New bandwidth parameters needed to be estimated to include sprint speed. Once again, each outcome's batted balls were split by even and odd days. Then a four-dimensional search was performed on σ_s, σ_v, σ_h, and σ_{ss} with a step size of 0.2. As one would expect, increasing the dimension of the search from three to four led to an increase in the computational time. For all outcomes other than singles and outs, σ^*_{even} and σ^*_{odd} were calculated, and then σ^* was simply the average of those two vectors. Since there were a large number of batted balls that resulted in outs or singles, the odd and even sets for singles were split into 10, and the odd and even sets for outs were split into 20. Then, the method described above was used to find σ^*_j for $j = 1, ..., 10$ for the singles and σ^*_k for $k = 1, ..., 20$ for the outs. Then the singles' bandwidth parameters were given by the average of the 10 different σ^*_j vectors, and the outs' bandwidth parameters were given by the average of the 20 different σ^*_k vectors. This was done to save on computational time. There were about ten times as many outs as there were doubles, and there were about five times as many singles as there were doubles. Calculating the bandwidths for doubles didn't take very long, and calculating the bandwidths for doubles using half as many doubles would have taken even less time. Thus, by splitting outs by twenty and singles by ten, the problem was essentially reduced to calculating the doubles' bandwidths using half as many doubles twenty more times and ten more times respectively. This reduced the computational time.

The bandwidth parameters that were obtained for the 2017 season are given in Table 2. Using these bandwidth parameters and the same weights that were used in the calculation of $I_{ns}(x)$, $I_s(x)$ was then calculated using (7).

σ^*	Outs	1B	2B	3B	HR	RBOE
σ_s	3.39	3.97	3.4	2.9	1.9	6
σ_v	4.35	3.79	3.6	4.1	1.9	3.4
σ_h	8.49	6.41	2.1	2.9	3.7	10
σ_{ss}	0.915	0.9	0.8	0.9	0.6	1

Table 2: 2017 Bandwidth Parameters with Sprint Speed

4 Visualizing the Intrinsic Values

In his article [1], Healey created a visual mapping from (s, v, h) to the intrinsic value, $I_{ns}(x)$, called the wOBA cube. Fig. 1 gives a similar wOBA cube that uses the 2017 data rather than the 2014 data that Healey used. The distance from home plate to the fence is typically shortest along the baselines ($h = \pm 45$). Thus, it isn't surprising that Fig. 1 suggests that when a batted ball is hit with a speed of 96 mph, it is most valuable when it is hit down the baselines ($h > 40$ or $h < -40$) with a vertical angle, v, between 25 and 35. The cold spots centered just below $v = 20$ with horizontal angles of -30, 0, and 30 represent balls hit to the left, center, and right fielders which typically result in outs. Likewise, the cold spots below $v = 0$ that are centered near $h = -35, -15, 20$, and 40 represent balls fielded for outs by third basemen, shortstops, second basemen, and first basemen respectively.

A similar visual can be created that maps (s, v, h, ss) to $I_s(x)$ if s and ss are held constant. This visual can no longer be called a wOBA cube though, since there are four inputs instead of three. Instead, it should be called a wOBA tesseract.

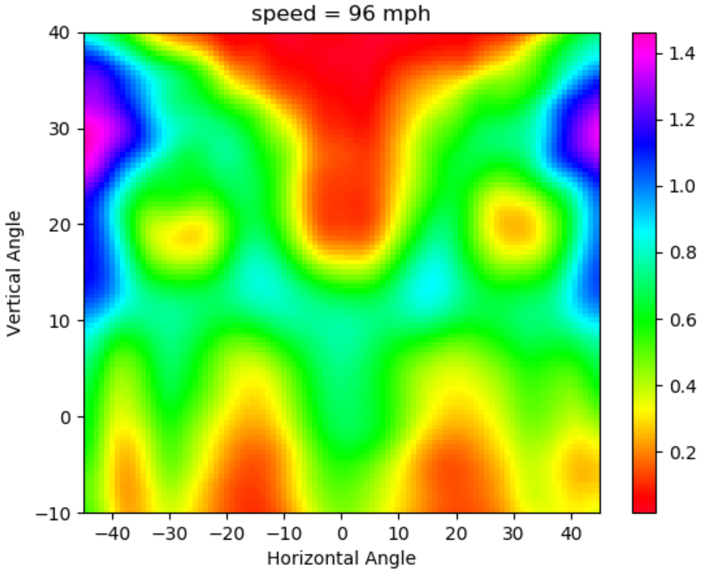

Figure 1: 2017 wOBA Cube with s=96 mph

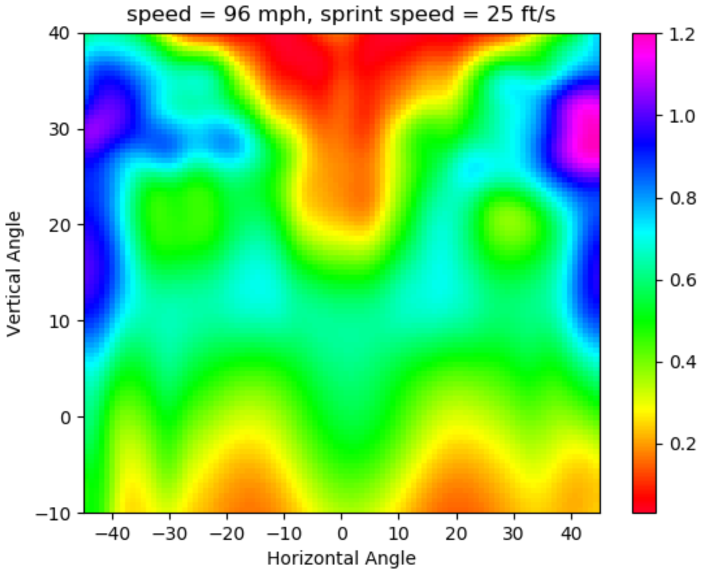

Figure 2: wOBA Tesseract with s = 96 mph and ss = 25 ft/s

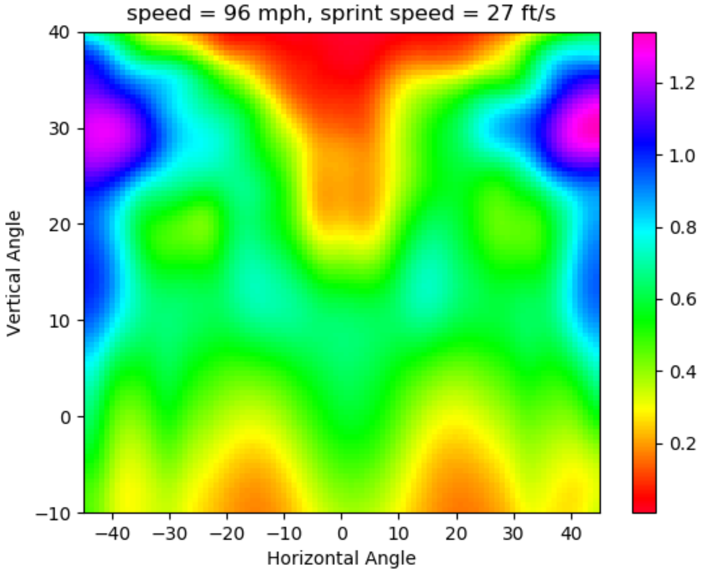

Figure 3: wOBA Tesseract with s = 96 mph and ss = 27 ft/s

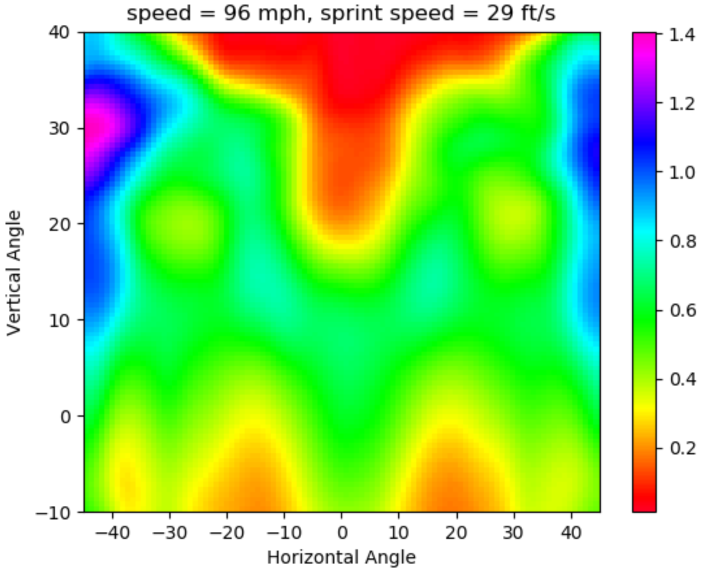

Figure 4: wOBA Tesseract with s = 96 mph and ss = 29 ft/s

Figs. 2, 3, and 4 give the wOBA tesseracts for $s = 96$ mph and $ss = 25$, 27, and 29 ft/s respectively. 27 ft/s is the MLB average sprint speed, and 25 and 29 ft/s are relatively slow and fast MLB sprint speeds. These tesseracts all look pretty similar, but there are some subtle differences. It seems that ground balls down the first base line ($h > 40$, $v \in [-10, 0]$) get increasingly more valuable as the sprint speed increases, as we would expect. Also, balls hit with v close to 30 and h less than -40 seem to get increasingly more valuable as sprint speed increases. This is less obvious because the color scales on each tesseract are a bit different, but for ss equal to 29, balls hit with these characteristics have a value as high as 1.4. For ss equal to 27, these balls have an intrinsic value close to 1.2. Finally, for ss equal to 25, they have an intrinsic value of less than 1.2. Thus, as sprint speed increases, batted balls hit with these characteristics get increasingly valuable in terms of $I_s(x)$. A similar thing happens for $v = 30$ and $h > 40$, but the improvement is less obvious. The intrinsic value improves from about 1.2 to something greater than 1.2 as ss increases from 25 to 27, but then for some reason it goes back down to 1.2 or slightly less than 1.2 for ss equal to 29. This unexpected result could potentially be explained by a smaller sample size of batted balls with similar characteristics. In 2017, there was only one batted ball with $ss = 29$, v between 25 and 35, s between 90 and 100, and h greater than 40. This batted ball resulted in an out, which could explain why the wOBA tesseract for $ss = 29$ has a surprisingly low value for balls with v near 30 and h greater than 40. Contrarily, there were over 300 batted balls in 2017 in the same ranges for s, v, and h with any sprint speed. By adding a sprint speed parameter, the number of observed batted balls with similar characteristics to any given batted ball vector naturally decreases, which could explain why increasing ss isn't guaranteed to increase $I_s(x)$ even though that's what one might expect.

5 Comparing the Sprint Speed Intrinsic Value with the non-Sprint Speed Intrinsic Value

The main purpose of this work was to address the point that Healey brought up in [2] that I_{ns} tends to result in high $O - I$ values for fast runners and small $O - I$ values for slow runners, where O is the outcome statistic wOBA on contact or wOBAcon. By adding a sprint speed parameter to calculate I_s, the goal was to maintain the accuracy and reliability of the I_{ns} statistic and also to avoid the trend of underrating fast runners and overrating slow ones in terms of $O - I$. In this section the accuracy, reliability, and $O - I$ values of I_{ns} and I_s are compared.

5.1 Accuracy

One would expect the intrinsic value of a batted ball to be fairly close to the actual outcome value, or wOBAcon. If there is a significant difference between the wOBAcon and I values of a large number of hitters, then the intrinsic value model is likely inaccurate. To measure this, the mean absolute deviation between O and I, as given by

$$M.A.D = \frac{1}{K}\sum_{j=1}^{K} |I_j - O_j|, \tag{9}$$

where K is the number of batters, I_j is the j^{th} batter's intrinsic value statistic, and O_j is the j^{th} batter's wOBAcon, was calculated using the 2017 batted ball data. The I_{ns} statistics had an M.A.D. of about 0.0367 and the I_s statistics had an M.A.D. of about 0.0365. Thus, the sprint speed intrinsic value seems to be at least as accurate, if not more accurate, than the non-sprint speed intrinsic value. This led to the conclusion that the sprint speed intrinsic value maintained the accuracy of the non-sprint speed intrinsic value, which was one of the goals of this work.

5.2 Reliability

In Healey's work, he used Cronbach's alpha to compare the reliability of the non-sprint speed intrinsic values in 2014 with the wOBAcon values in 2014 [1][9]. He showed that the intrinsic values were more reliable. Another goal for the sprint speed intrinsic value calculations was to maintain or improve upon the level of reliability of the non-sprint speed intrinsic values. Cronbach's alpha was calculated for both intrinsic values and the wOBAcon. The results are given in Figure 5.

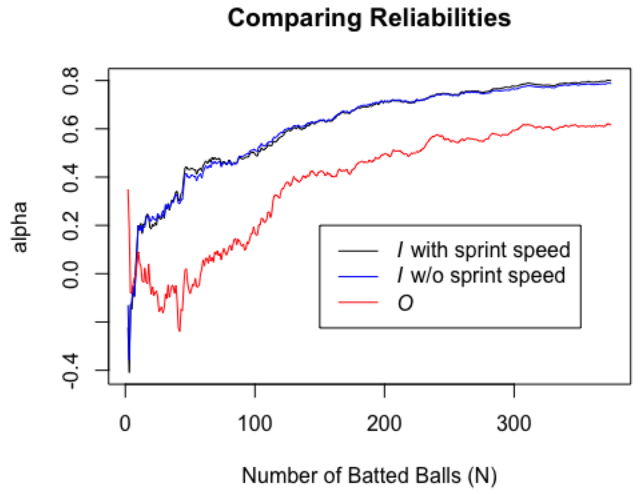

Figure 5: Reliability comparisons using Cronbach's alpha

As expected, the intrinsic value without sprint speed still has a higher reliability than the outcome statistic. Additionally, the intrinsic value with sprint speed is just as reliable as the one without sprint speed. Thus, the goal that the sprint speed intrinsic value be just as reliable as the non-sprint speed intrinsic value was met.

5.3 $O - I$ Comparisons

The final goal of adding sprint speed to the intrinsic value calculation was to not underestimate fast runners and overestimate slow runners. In Healey's 2014 calculations, he noted that most of the top ten highest $O - I$ values came from players with above average running speed. Likewise, the ten smallest $O - I$ values all came from below average runners [2]. A similar trend occurs in the 2017 hitters. Looking at only players with at least 300 batted balls who also had sprint speed data available, the ten largest $O - I_{ns}$ values are given in Table 3. The ten smallest $O - I_{ns}$ values are given in Table 4.

Name	$O - I_{ns}$	Sprint Speed (ft/s)
Trevor Story	0.084	29.3
Charlie Blackmon	0.080	28.2
Javier Baez	0.073	28.5
Scooter Gennett	0.070	26.4
Bryce Harper	0.069	27.7
Zack Cozart	0.069	26
Marwin Gonzalez	0.069	26.7
Didi Gregorius	0.068	28.2
Cory Spangenberg	0.060	28.6
Corey Dickerson	0.060	27.6

Table 3: Largest $O - I_{ns}$

Note, the average sprint speed in 2017 was about 27 ft/s. All but three of the batters with the highest $O - I_{ns}$ values had an above average sprint speed. All of the hitters with small $O - I_{ns}$ values had below average sprint speeds. Just like in 2014, most of the highest $O - I_{ns}$ values were from above average runners, and all of the smallest $O - I_{ns}$ values were from below average runners.

Charlie Blackmon and Trevor Story not only benefitted from good sprint speeds, but they also play for the Colorado Rockies in the most hitter friendly ballpark in all of baseball. This ballpark effect is likely to inflate O, which could be another reason why these players ended up with high $O - I_{ns}$ values in 2017. This provides further evidence that the intrinsic value statistic does a good job of ignoring confounding variables such as ballpark effects.

Name	$O - I_{ns}$	Sprint Speed (ft/s)
Mitch Moreland	-0.051	25.7
Albert Pujols	-0.044	21.8
Todd Frazier	-0.043	26.5
Joe Mauer	-0.037	26.1
Shin-Soo Choo	-0.036	26.3
Kendrys Morales	-0.034	23.5
Yadier Molina	-0.033	23.8
Manny Machado	-0.032	26.3
Maikel Franco	-0.031	26
Hanley Ramirez	-0.031	26.1

Table 4: Smallest $O - I_{ns}$

Now, the primary goal of adding sprint speed to the intrinsic value calculation was to stop overrating slow runners and underrating fast runners. The ten players in 2017 with the largest $O - I_s$ values are given in Table 5. The ten smallest $O - I_s$ values are given in Table 6.

Name	$O - I_s$	Sprint Speed (ft/s)
Zack Cozart	0.081	26
Trevor Story	0.080	29.3
Scooter Gennett	0.076	26.4
Marwin Gonzalez	0.073	26.7
Charlie Blackmon	0.071	28.2
Javier Baez	0.069	28.5
Bryce Harper	0.062	27.7
Didi Gregorius	0.054	28.2
Corey Dickerson	0.054	27.6
Nolan Arenado	0.054	25.7

Table 5: Largest $O - I_s$

The ten players with the highest $O - I_s$ values are similar to the ten players with the highest $O - I_{ns}$ values. The only differences are the ordering and that Nolan Arenado has taken Cory Spangenberg's spot. Nolan Arenado is a below average sprinter, so now there are four below average sprinters in the top ten rather than three like there were in the $O - I_{ns}$ list. Nolan Arenado, like Story and Blackmon, is a Colorado Rockie, so his large value of $O - I_s$ could likely be explained by the fact that he plays his home games in a high scoring ballpark. In addition to all the Rockies on this list, all but Marwin Gonzalez, Javier Baez, and Corey Dickerson played in a ballpark that is slightly more hitter friendly

Name	$O - I_s$	Sprint Speed (ft/s)
Todd Frazier	-0.038	26.5
Ian Kinsler	-0.032	26.9
Manny Machado	-0.032	26.3
Mitch Moreland	-0.031	25.7
Shin-Soo Choo	-0.031	26.3
Ben Zobrist	-0.025	27.1
Robbie Grossman	-0.025	27.9
Dansby Swanson	-0.025	28.7
Hanley Ramirez	-0.025	26.1
Scott Schebler	-0.020	28.3

Table 6: Smallest $O - I_s$

than average. This may have contributed to them having high $O - I_s$ values, but overall it seems that even the sprint speed intrinsic value tends to underrate fast runners. However, it seems to underrate them by less than I_{ns}. The average $O - I_s$ value in the top ten list is 0.0674, whereas the average $O - I_{ns}$ value in the top ten list is 0.0702. Thus, although I_s still seems to have a tendency to underestimate fast runners, it seems to underestimate them by less than I_{ns}, which could be considered a slight improvement.

The top ten smallest $O - I_s$ values list is fairly different from the top ten smallest $O - I_{ns}$ values list. Unlike in the $O - I_{ns}$ list, there are a few players that are not below average sprinters in the $O - I_s$ list. Scott Schebler, Dansby Swanson, Robbie Grossman, and Ben Zobrist are all above average runners, and all of them had small $O - I_s$ values. Thus, it seems we have made an improvement in not overrating slow runners in terms of their I_s. Additionally just as in the top ten list, the $O - I_s$ value list overvalues these players by less on average than the $O - I_{ns}$ list. The $O - I_{ns}$ bottom ten list had an average $O - I_{ns}$ value of -0.0372, whereas the $O - I_s$ list had an average $O - I_s$ value of -0.0284. Thus, not only does I_s not overrate slow runners as frequently as I_{ns} but it also seems to overrate slow runners by less than I_{ns}.

Although the sprint speed intrinsic value still had a slight tendency to overrate slow runners and underrate fast runners in terms of their $O - I$, these top and bottom ten lists suggest that it improved upon the non-sprint speed intrinsic value. The I_s value over and underrated slow and fast runners less frequently than I_{ns}, at least according to these top ten lists. I_s also seemed to under and overrate these top and bottom ten players by less than I_{ns}

6 Alternative Modeling Choices

Although many alternative choices could have been made in the modeling process, this section discusses the decisions to use diagonal covariance matrices in the kernel density estimates and to use Statcast's sprint speed metric.

6.1 Non-Diagonal Σ in Kernel Density Estimation

In both Healey's model and the updated sprint speed model, the covariance matrix, Σ, of the Gaussian kernel is assumed to be diagonal. This makes finding the optimal bandwidths less computationally expensive because most of the entries in Σ are zero. However, assuming Σ is diagonal requires the assumption that the covariances between the entries of the batted ball vector are all zero. This assumption was made to make computations simpler, but it was never justified. It is possible that the covariance between some of these factors is not zero. In particular, the value that sprint speed adds to a batted ball's intrinsic value depends heavily on the type of batted ball. For example, the intrinsic value of a ground ball probably depends a lot on the sprint speed, whereas the intrinsic value of a home run that leaves the ballpark isn't affected at all by the batter's sprint speed. Thus, it is likely that the covariances between sprint speed and the other factors are actually nonzero. It is also possible that the covariances between s, v, and h are nonzero. Therefore, removing the assumption that Σ is diagonal could potentially improve the accuracy and reliability of the intrinsic value models, particularly the sprint speed model, at the cost of increased computational time.

6.2 Alternative Measures of Sprint Speed

The sprint speed metric used in this research is defined on the Statcast website as "feet per second in a player's fastest one-second window" [4]. One of the potentially interesting applications of the intrinsic value statistic is to map an amateur player's batted ball characteristics to an intrinsic value that can be compared to the MLB standard. Many batting cages and even college baseball fields have the ability to track a batted ball's velocity, vertical angle, and horizontal angle. However, the ability to determine a player's sprint speed as defined by Statcast would be difficult without the data collecting technologies inside of MLB stadiums. An alternative measure of sprint speed that is also available on Statcast is a player's home plate to first base time. This data point can easily be collected on amateur players by using a stopwatch. The model described in this work used Statcast's sprint speed metric, but it could likely be adjusted to use the home plate to first base time instead. Making this adjustment would allow for amateur players to more easily compare the intrinsic value of their batted balls with the MLB standard, and it is an interesting alternative to the method used in this work.

7 Conclusion

In [1], Healey used a Bayesian model to map a batted ball's speed, vertical angle, and horizontal angle to an intrinsic value given by the expected wOBA of such a batted ball. He created a visual of this mapping called the wOBA cube, and he showed that the intrinsic value statistic is more reliable than the outcomes-based statistic wOBAcon in terms of Cronbach's alpha. Despite the accuracy and the reliability of this intrinsic value, Healey noted that it had a tendency to underrate fast runners and overrate slow runners [2].

The purpose of this research was to address the intrinsic value's tendency to underrate fast runners and overrate slow runners while maintaining the accuracy and reliability. Sprint speed data for 2017 MLB players were used to update the original model to include the sprint speed of the batter. The mapping was adjusted to map the batted ball's speed, vertical angle, and horizontal angle as well as the batter's sprint speed to an intrinsic value. A new visual of this mapping was created called the wOBA tesseract. Additionally, comparisons between the sprint speed intrinsic value statistic and the non-sprint speed intrinsic value statistic showed that the sprint speed intrinsic statistic maintained the accuracy and reliability of the non-sprint speed statistic. It also seemed to under and overrate fast and slow runners less frequently and by a smaller amount than the non-sprint speed statistic. Thus, the goals of this work were met. The sprint speed intrinsic value not only maintained the accuracy and reliability of the non-sprint speed intrinsic value, but it also seemed to do a better job of not overrating slow runners and underrating fast runners.

The advantage of using intrinsic value statistics over outcomes-based statistics is that they ignore confounding effects like the ballpark, the weather, or the quality of the defense. By ignoring these effects, the intrinsic value statistics are more reliable and represent a purer estimate of a hitter's true talent. MLB teams that improve upon existing intrinsic value models, such as Glenn Healey's or even the sprint speed intrinsic value model of this paper, will be better able to evaluate and sign hitter talent, giving them a competitive advantage in the journey to a World Series.

8 Acknowledgments

The author would like to thank Dr. Sean Warnick and BYU's IDeA labs for paying him to research baseball. He'd also like to thank Dr. Glenn Healey for his help and advice.

9 References

[1] Healey, G. (2017). Learning, visualizing, and assessing a model for the intrinsic value of a batted ball. IEEE Access, 5, 13811-13822. doi:10.1109/ACCESS.2017.2728663

[2] Healey, G. (2016). The intrinsic value of a batted ball. https://tht.fangraphs.com/the-intrinsic-value-of-a-batted-ball/

[3] Slowinski, S. (2010). wOBA. Retrieved from https://library.fangraphs.com/offense/woba/

[4] Baseball savant. Retrieved from https://baseballsavant.mlb.com

[5] Petti, B. (2019). Baseballr.http://billpetti.github.io/baseballr/

[6] wOBA and FIP constants. Retrieved from https://www.fangraphs.com/guts.aspx?type=cn

[7] Tango, T. *The book playing the percentages in baseball.* http://www.insidethebook.com/woba.shtml

[8] What is an error (E)? http://m.mlb.com/glossary/standard-stats/error

[9] Goforth, C. (2015). Using and interpreting cronbach's alpha. Retrieved from https://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/