David J. Hunter*

New metrics for evaluating home plate umpire consistency and accuracy

https://doi.org/10.1515/jqas-2018-0061

Abstract: The availability of pitch-tracking data has led to increased scrutiny of Major League Baseball umpires. While many studies have attempted to rate umpires based on their conformity to the rule book strike zone, players and managers tend to accept deviations from this zone, provided that umpires establish consistent zones within a game. Using tools from computational geometry, we propose new metrics for assessing the consistency and accuracy of an umpire's ball and strike calls over the course of a game. We apply these metrics to pitch-tracking data on all ball and strike calls made during the 2017 MLB regular season, giving some characterizations of the variation in performance of MLB umpires. This analysis demonstrates that measures of consistency can complement current accuracy-based evaluations of umpires.

Keywords: α-convex hull; convex hull; kernel density estimation; principal component analysis.

1 Introduction

Since 2009, Major League Baseball has been using modern pitch-tracking data to evaluate and train its umpires (Mills 2017). While the instant public availability of this data has prompted some calls for electronic automation of ball and strike calls, there is evidence that MLB's Zone Evaluation system has led to an improvement in umpire accuracy (Davis and Lopez 2015).

The Zone Evaluation system focuses on fidelity to the rectangular front of rule book zone, but actual strike calls in practice conform to the patterns shown in Figure 1. These plots suggest that pitches on corners of the rule book zone are likely to be called balls, forming an accepted "consensus" strike zone that is rounded and non-rectangular. In addition, it appears that pitches off the plate away from the batter are more likely to be called strikes than pitches off the plate inside, suggesting that consensus zones differ for left- and right-handed batters. To more accurately assess umpires within the context of accepted practices, measures of strike zone accuracy can be adapted to account for these consensus zones (Roegele 2017).

However, measures of accuracy alone fail to assess umpire consistency within a game. In this paper, we propose several new ways to measure an umpire's consistency, apart from accuracy. We relax the requirement of a rectangular zone, and we allow for variations based on the handedness of the batter. Since factors such as the style of the starting pitcher may influence the shape of an umpire's zone from game to game, we measure consistency within a game and average over all games in the 2017 season, rather than aggregating the call data and taking a single measurement. We also investigate the relationships between consistency, accuracy, and other umpire tendencies.

2 Inconsistency

Over the course of a game, an umpire establishes a region of pitches that are called strikes. Ideally, this *established strike zone* will have a predictable shape, and no pitches that fall inside of it will ever be called balls. In this section, we propose four metrics for assessing the consistency of calls relative to the strike zone that the umpire establishes.

Each of these metrics depends on the chosen geometry of the established strike zone. First, we consider the consequences and limitations of a simple rectangular established strike zone, and propose a refinement to address these limitations. Next, we relax our geometric assumptions to consider non-rectangular established zones, requiring only that these zones be convex.

Throughout this paper, we use publicly-available data from MLB Advanced Media for the 2017 regular season (MLBAM 2018). Ball and strike data are posted as (`px`, `pz`) pairs, indicating the horizontal and vertical position (in feet) of the ball as it crosses the front of home plate, where the center of the plate at ground level corresponds to the point (0, 0). Since the vertical limits of the strike zone depend on the height and and stance of the batter, MLBAM also provides parameters `sz_top` and `sz_bot`, which estimate the top and bottom of the strike

***Corresponding author: David J. Hunter,** Westmont College, Santa Barbara, CA, USA, e-mail: dhunter@westmont.edu

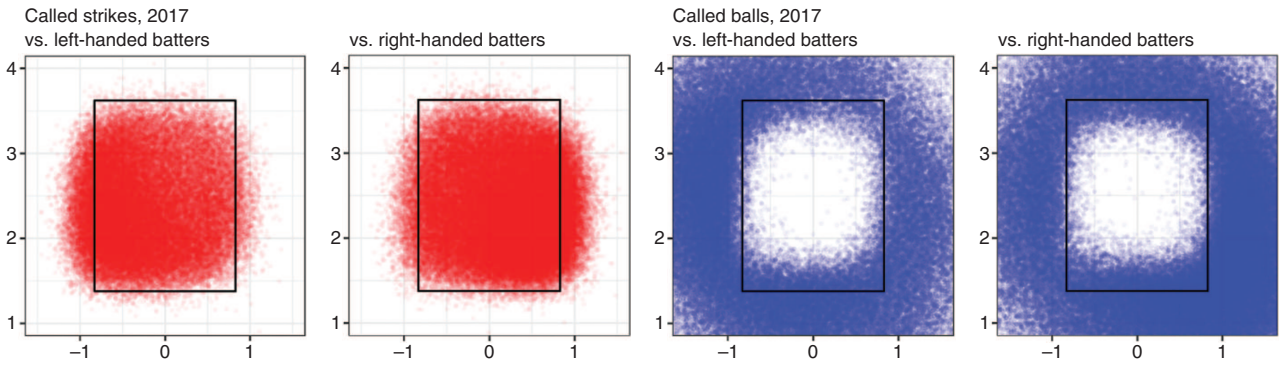

Figure 1: All ball (blue) and strike (red) calls made in the 2017 MLB season, for left- and right-handed batters, from the umpire's perspective. The rectangle indicates the rule book strike zone. Vertical positions have been scaled based on the height and stance of the batter.

zone for each batter. We use these parameters to normalize the vertical positions pz so that the top of the zone corresponds to pz = 3.5 and the bottom of the zone corresponds to pz = 1.5:

$$\text{normalized pz} = \frac{2(\text{pz} - \text{sz\_top})}{\text{sz\_top} - \text{sz\_bot}} + 3.5$$

All of the strike zone plots in this paper show px on the horizontal axis and this normalized value of pz on the vertical axis.

2.1 Rectangular metrics

A natural definition for the established strike zone is the smallest rectangular region containing all the strikes. Figure 2 shows an example of these established strike zones for left- and right-handed batters for a particular MLB game. Any called ball inside these rectangles is *inconsistent*. For a given game, we define the *one-rectangle inconsistency index* I_{R1} as

$$I_{R1} = \frac{\text{number of inconsistent balls}}{\text{total number of called balls}}.$$

In the game shown in Figure 2, out of 110 called balls, 2 were inconsistent to left-handed batters and 13 were inconsistent to right-handed batters, so $I_{R1} = (2 + 13)/110 \approx 0.136$.

While the one-rectangle inconsistency index is natural to define and easy to compute, it is highly sensitive to a single outlying strike call. For example, in the right-handed plot in Figure 2, if we remove the lowest strike at (0.26, 1.27), the lower border of the established strike zone moves up to the next-lowest strike, eliminating three inconsistent balls. Had this single low strike been called a ball, the index I_{R1} would have been $(2 + 10)/110 \approx 0.109$ instead of $(2 + 13)/110 \approx 0.136$.

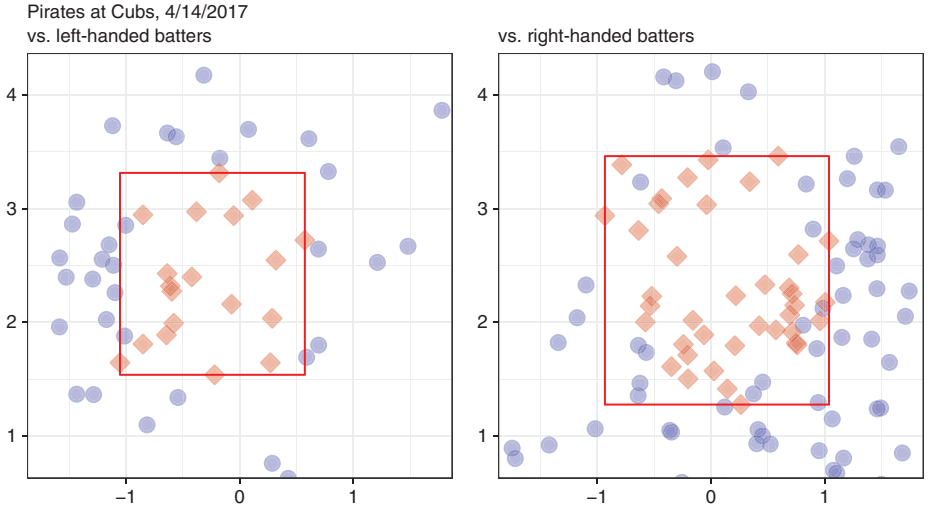

Figure 2: The smallest rectangle containing all the strikes. Balls are drawn as blue circles, and strikes as red diamonds. Note that the center of the circle or diamond must lie on or inside the rectangle to be considered inside the rectangular region.

Another weakness of the one-rectangle index is that it can fail to account for multiple bad strike calls in the same location. Again using the right-handed plot in Figure 2, we see that eliminating the strike at $(-0.93, 2.94)$ has no effect on I_{R1}, because the resulting rectangle will still enclose the same number of called balls. While this strike call seems inconsistent given the five called balls around $\texttt{px} = -0.6$, the measure I_{R1} fails to reflect this inconsistency.

We can mitigate these limitations in the one-rectangle inconsistency index by using more rectangles. As with the definition of I_{R1}, the first rectangular region is the smallest one that contains all the strikes; that is, it is the rectangle determined by the smallest and largest values of both px and pz for the set of called strikes. The second rectangle is then determined by the second-smallest and second-largest values of these coordinates. Continuing in this manner, taking the ith smallest and ith largest coordinates, we can form rectangles $R_1, R_2, \ldots, R_n$ for both left- and right-handed hitters, for some choice of n. (Once i becomes large enough to exhaust all of the called strikes, take R_i to be the empty set.) Let $s(i)$ be the number of called balls inside the two R_i's. Define the *n-rectangle inconsistency index* as

$$I_{Rn} = \frac{s(1) + s(2) + \cdots + s(n)}{\text{total number of called balls}}$$

The rectangles used to calculate I_{R10} are shown in Figure 3. Versus left-handed batters, rectangle R_1 is the only rectangle containing called balls. However, versus right-handed batters, R_1 contains 13 called balls (as above), R_2 contains 11, R_3 contains 7, R_4 contains 1, and the remaining rectangles contain none. Therefore,

$$I_{R10} = \frac{(2 + 13) + (0 + 10) + (0 + 6) + (0 + 1)}{110} \approx 0.29.$$

Notice that, in this example, $I_{R10} = I_{R9} = \cdots = I_{R4}$, illustrating that the value of this index eventually stabilizes, given enough rectangles. In practice, 10 rectangles is plenty for most MLB game data sets.

Since $R_{i+1} \subseteq R_i$, inconsistent balls are weighted according to how many rectangles they are contained in. Balls that are inconsistent due to a single outlying strike will only have weight 1, while egregiously bad ball calls will lie inside several rectangles, increasing their contribution to I_{Rn}. Therefore, compared to the one-rectangle index, the n-rectangle index is less sensitive to a single outlying strike call.

Furthermore, when several bad strike calls lie in the same location, the n-rectangle index will reflect this inconsistency, since the successive rectangles will not shrink until these strike calls are exhausted. In Figure 3, the called balls around $\texttt{px} = -0.6$ are weighted more heavily because of the strike at $(-0.93, 2.94)$, in contrast to the situation in Figure 2, where we saw that this strike had no effect on the one-rectangle index.

2.2 Convex hull metrics

Since rectangles are used to construct the inconsistency index I_{Rn}, it measures inconsistency under the assumption (stipulated in the rules of baseball) that the true strike zone is a rectangle. However, as we will see in Section 3, strike zones in practice tend to be rounded at the corners. In this section we will introduce inconsistency measures that relax the assumption of a rectangular zone. Instead, we assume that a consistent zone will have the property that any pitch landing between two called strikes will also

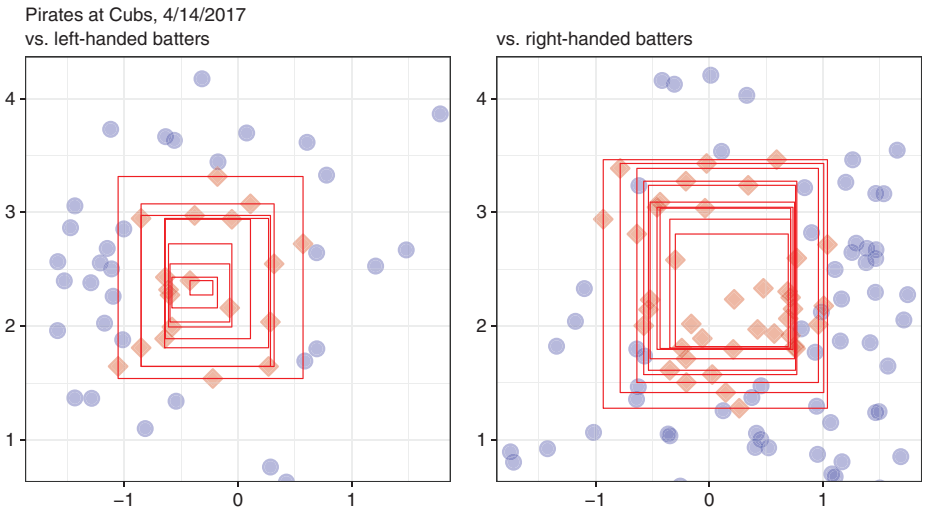

Figure 3: Successive rectangles enclose inconsistent balls. The more inconsistent called balls lie within more rectangles.

be a called strike. In other words, we assume that the established strike zone is convex.

Given a discrete set $P \subseteq \mathbb{R}^2$ representing the locations of called strikes during a game, there is a natural geometric definition for the established strike zone, namely, the *convex hull* of P. We can define the convex hull S as the intersection of all closed half planes that contain P:

$$S = \bigcap_{\{H_l | H_l \cap P = \emptyset\}} H_l^c,$$

where H_l^c denotes the complement of the open half-plane bounded by the line l.

Using the convex hull as our established strike zone, we can define the *convex hull inconsistency index* I_{CH} analogously to the one-rectangle inconsistency index. Now an inconsistent ball is one that lies within the convex hull of strikes, and I_{CH} is given by

$$I_{CH} = \frac{\text{number of inconsistent balls}}{\text{total number of called balls}}.$$

For example, see Figure 4. There were five inconsistent balls versus left-handed batters, and one versus right handed batters, out of a total of 118 called balls. Therefore $I_{CH} = (5 + 1)/118 \approx 0.051$.

Like the one-rectangle index, the convex hull inconsistency index can fail to account for multiple bad strikes in the same location. It can also be unaffected by outlying strikes, depending on their location. For example, in Figure 4 versus right-handed batters, the strike at $(-0.01, 1.31)$ has no effect on I_{CH}; removing this point would shrink the convex hull without changing the number of called balls enclosed. However, this call seems inconsistent, given its proximity to several called balls. The problem is that a vertex of the convex hull can lie in a region populated by called balls, yet fail to enclose any. Creating smaller convex hulls inside the first (as we did to define I_{Rn}) will not address this issue.

To account for this phenomenon, we can use the locations of called balls to define a *called-ball region*. Instead of counting called balls within the established strike zone, we can measure the area of the overlap between the called-ball region and the convex hull of strikes.

Unlike the established strike zone, the called-ball region will typically not be convex, or even simply connected. Given a set $Q \subseteq \mathbb{R}^2$ representing the locations of called balls during a game, and given some radius $\alpha > 0$, define

$$X = \bigcap_{\{B_{x,\alpha} | B_{x,\alpha} \cap P = \emptyset\}} B_{x,\alpha}^c,$$

where $B_{x,\alpha}^c$ denotes the complement in the plane of the open disk of radius α centered at the point x. The region X, which will serve as our called-ball region, is called the *α-convex hull* of Q (Pateiro-López and Rodrıguez-Casal 2010). Note that the α-convex hull is not convex, in general.

Let a_L and a_R be the areas of the intersection of the convex hull of called strikes and the α-convex hull of called balls, for left-handed and right-handed batters, respectively. We define the *α-convex hull inconsistency index* I_{ACH} to be a weighted average of these two areas. Let n_L be the number of called pitches thrown to left-handed batters, and let n_R be the number of called pitches to right-handed batters. Then

$$I_{ACH} = \frac{n_L a_L + n_R a_R}{n_L + n_R}$$

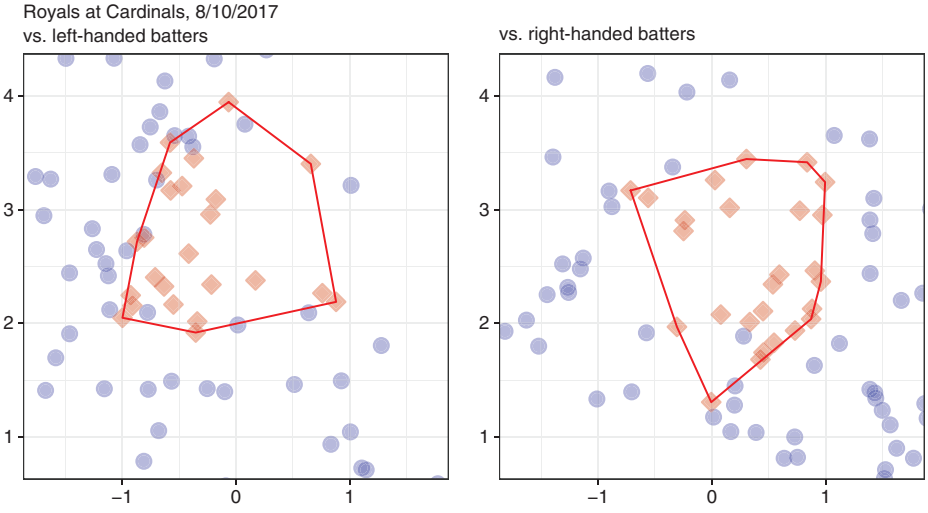

Figure 4: The established strike zone is the convex hull of called strikes.

Figure 5 shows the ball and strike calls for the same game as Figure 4, along with the α-convex hull of called balls, using $\alpha = 0.7$. In this case, $I_{ACH} = 0.127$. Notice that the called strike at $(-0.01, 1.31)$ now has a significant effect on I_{ACH}, since it causes a large region of overlap between the convex hull and the α-convex hull, which contains the nearby called balls.

One limitation to this choice of inconsistency metric is that there is no canonical choice for the constant α. Figure 6 illustrates the issues involved. If the radius α is too small, the α-convex hull will contain isolated points and small disconnected regions. Large values of α (such as $\alpha = 0.9$ in this example) will produce a single, simply-connected α-convex hull, making the called-ball region

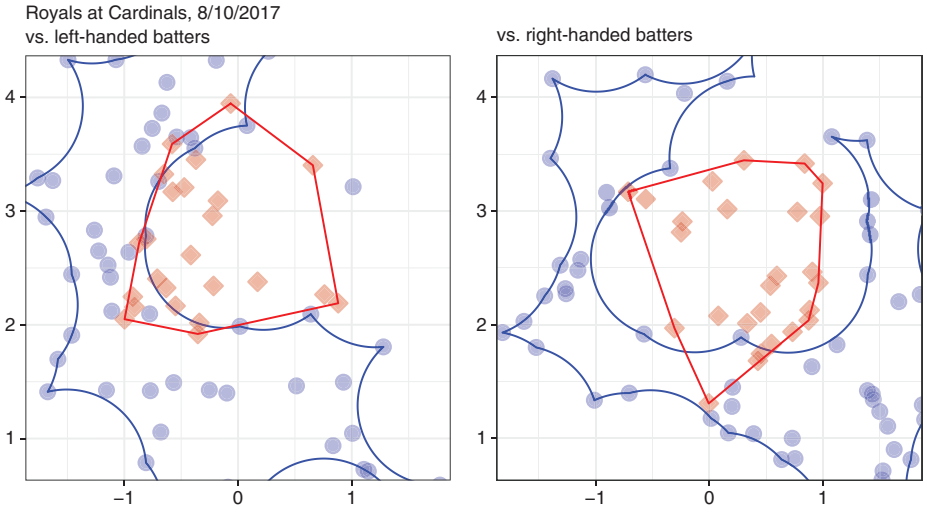

Figure 5: The established strike zone is the convex hull of called strikes (in red), and the called-ball region is the α-convex hull of balls (in blue), where $\alpha = 0.7$.

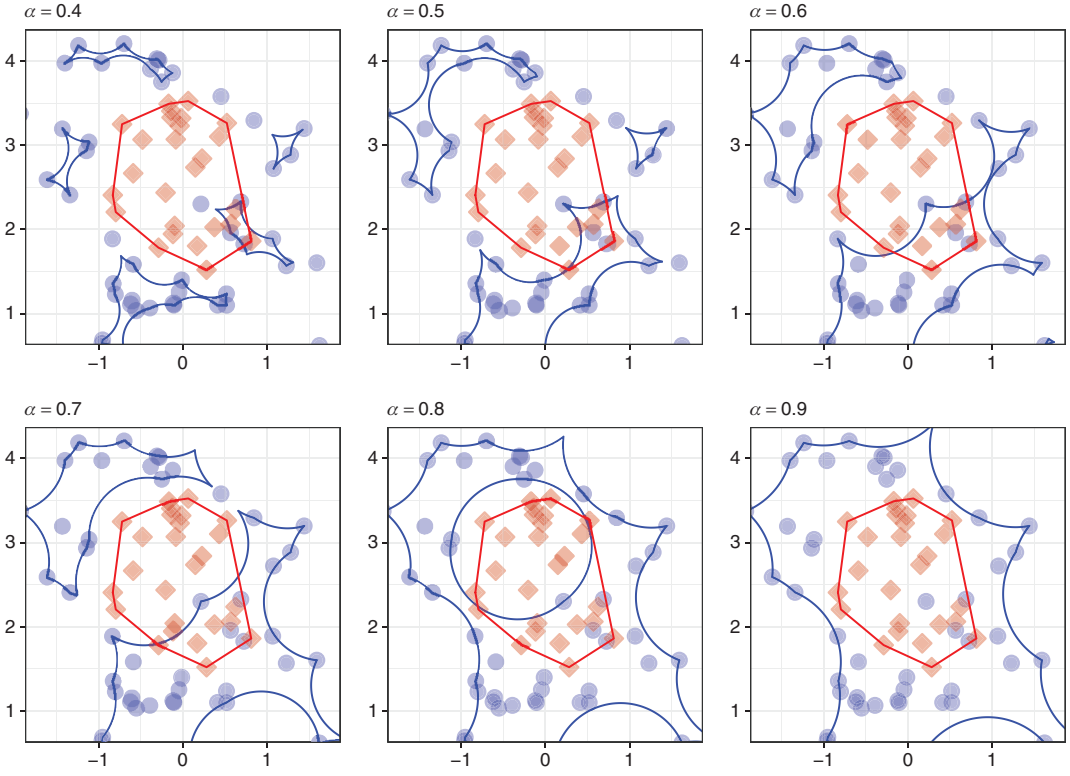

Figure 6: Six different called-ball regions (α-convex hulls) for different choices of α.

completely cover the established strike zone. Generally speaking, the larger the value of α, the tougher the metric I_{ACH} is on the umpires.

In the analysis that follows, we have chosen to use $\alpha = 0.7$, based largely on qualitative inspections of various game examples, as in Figure 6. A correlation analysis can lend some empirical support to this choice. Table 1 gives the pairwise correlations for I_{ACH} computed over all 2017 regular season games using six different values of α between 0.4 and 0.9, in increments of 0.1. For these six values, the greatest correlation is between $\alpha = 0.6$ and $\alpha = 0.7$, and we observe that once α exceeds 0.7, the correlations between adjacent values begin to decrease. These results confirm the observation that choosing α in the range $0.6 \leq \alpha \leq 0.7$ tends to give similar measures of I_{ACH}.

2.3 Statistical properties of the inconsistency metrics

All four inconsistency measures are sensitive to a single outlying called strike, and the n-rectangle and α-convex hull indices are sensitive to an egregiously bad called ball in the middle of the strike zone. This sensitivity is by design, to avoid penalizing umpires who make slightly inconsistent calls as much as those who make clearly bad calls. However, a consequence of this feature is that these metrics are also sensitive to the number of pitches called. As the number of called pitches increases, the chances that an umpire will make an egregious call increases, and once such a call is made, the inconsistency index will remain high.

Figure 7 investigates the association between number of pitches called and inconsistency index. The top row shows scatter plots of the four indices versus number of pitches called in the game, for all regular-season games with between 50 and 300 called pitches. (Only 2 of 2425 games in our sample fall outside this range.) For each index, there is a slight discernible upward trend. The correlation coefficient r is approximately 0.2 in all four cases.

The smoothed density estimates in Figure 8 show that the distributions of I_{R1}, I_{R10}, I_{CH}, and I_{ACH} are all skewed right. Such skewness may be a feature of the metrics, or it may indicate that major league umpires are, on the whole, very good at calling games consistently. To assess sensitivity in the tails of these distributions, the second row of scatter plots in Figure 7 considers only "high-inconsistency games," where $I_{R10} + I_{ACH} > 0.3$. For these games, and for this range of pitches called, there does not appear to be a strong association between the inconsistency measures and the number of called pitches ($|r| < 0.1$ in all cases).

In this section we have considered two simple metrics, I_{R1} and I_{CH}, along with extensions I_{R10} and I_{ACH}, respectively, which attempt to address deficiencies in the simple metrics. Figure 8 shows that the tails of the distributions of the extended metrics are substantially thicker, suggesting that these metrics are better at differentiating between higher levels of inconsistency.

The pairwise correlation coefficients for the four metrics are given in Table 2. The two extended metrics I_{R10} and I_{ACH} are correlated, but not very strongly, indicating that they measure different aspects of inconsistency. Some of the difference may be due to the shape of the strike zone that an umpire tends to call. For example, among the 79 umpires who called at least 20 games behind home plate in 2017, Chad Whitson was the 17th most consistent umpire when measured using I_{ACH}, but ranked 51st when measured using I_{R10}. The methods that we will present in Section 3.1 reveal that Whitson's zone tends to be quite rounded at the corners, rather than rectangular, so it is not surprising that he ranks higher when judged using the convex hull metrics. By comparison, Pat Hoberg, whose zone is somewhat less rounded, has the 6th best I_{R10}, but ranks 34th according to I_{ACH}. See Figure 9.

As a compromise, in some of our analysis, we use the sum $I_{R10} + I_{ACH}$ as a general measure of inconsistency. Notice that I_{R10} typically takes larger values than I_{ACH}, so the 10-rectangle metric is effectively weighted more than the α-convex hull metric in this sum. Over all 2423 games, the mean of the sum $I_{R10} + I_{ACH}$ is 0.191 with standard deviation 0.142, and its median is 0.156.

Table 1: Correlation matrix for I_{ACH} calculated with different values of α over all games in the 2017 season.

	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
$\alpha = 0.4$	1.00	*0.85*	0.73	0.63	0.45	0.35
$\alpha = 0.5$		1.00	*0.91*	0.79	0.60	0.46
$\alpha = 0.6$			1.00	*0.92*	0.74	0.56
$\alpha = 0.7$				1.00	*0.82*	0.62
$\alpha = 0.8$					1.00	*0.67*
$\alpha = 0.9$						1.00

Correlations between adjacent values are shown in italics.

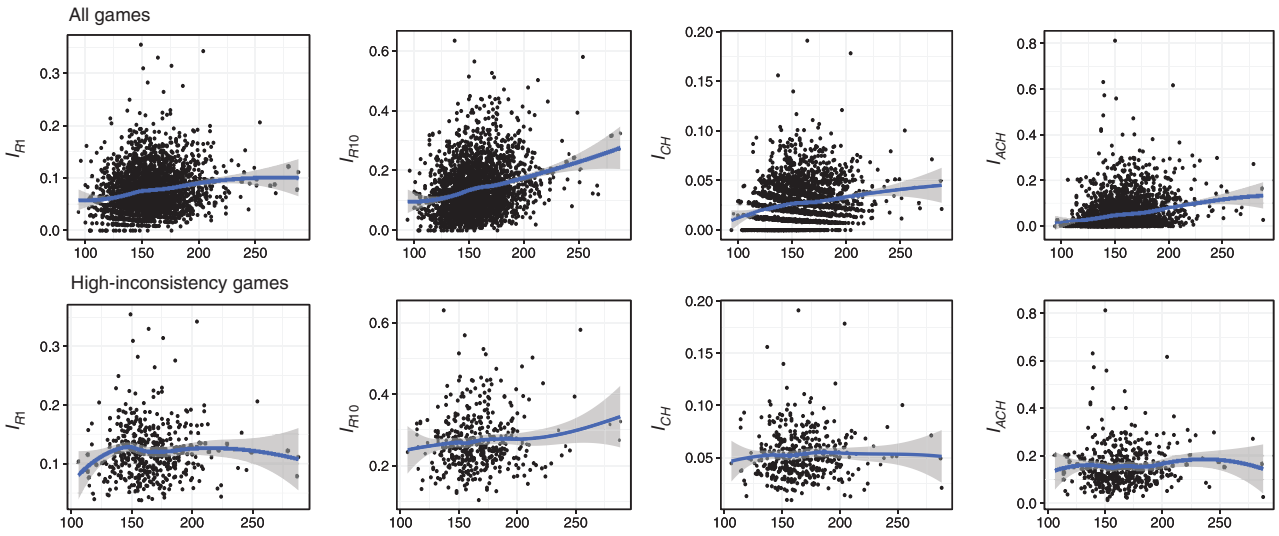

Figure 7: Scatterplots of the four inconsistency indices, I_{R1}, I_{R10}, I_{CH}, and I_{ACH}, versus the number of called pitches in the game. The top row shows the data for all games with between 50 and 300 called pitches, while the bottom row includes only games with high inconsistency indices. The blue curves show the smoothed conditional means.

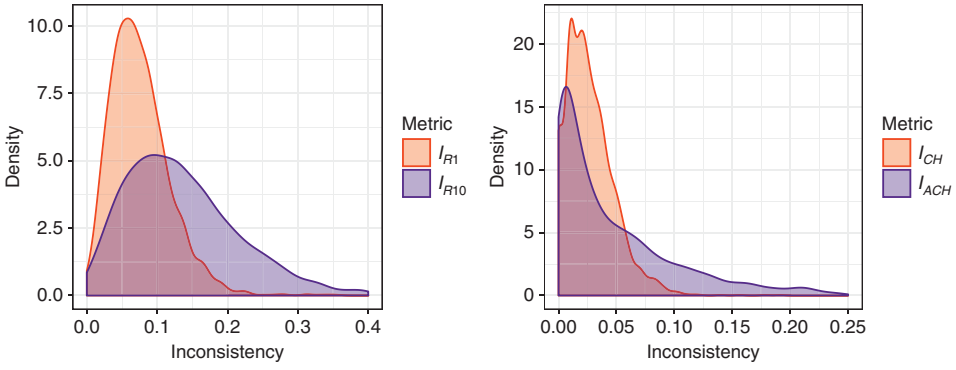

Figure 8: Smoothed density estimates of the four inconsistency indices computed on 2423 games.

Table 2: Correlation matrix for the four inconsistency indices.

	I_{R1}	I_{R10}	I_{CH}	I_{ACH}
I_{R1}	1.00	0.82	0.73	0.48
I_{R10}		1.00	0.78	0.68
I_{CH}			1.00	0.64
I_{ACH}				1.00

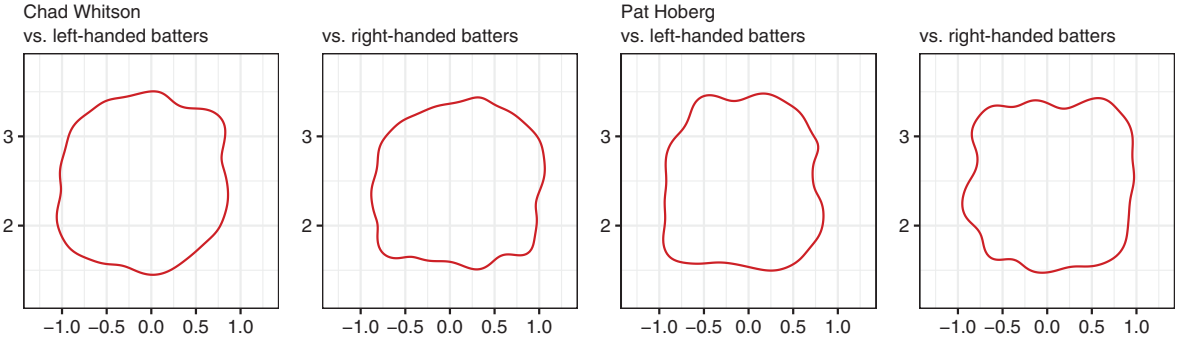

Figure 9: Zone tendencies for Chad Whitson and Pat Hoberg. Whitson is more consistent according to I_{ACH}, while Hoberg is more consistent according to I_{R10}. The method for constructing these contours is discussed in Section 3.1.

3 Zone accuracy

While consistency of ball and strike calls is an important aspect of neutral officiating, it is also expected that home plate umpires conform to established definitions and practices. The official rules of baseball (MLB 2018) define the strike zone as "that area over home plate the upper limit of which is a horizontal line at the midpoint between the top of the shoulders and the top of the uniform pants, and the lower level is a line at the hollow beneath the kneecap." It is often expedient to use the rectangular front of this pentagonal prism as a two-dimensional approximation of the rule book strike zone. The width of the front of home plate is 17 inches, establishing the horizontal limits of this rectangle. Our data has been normalized so that the vertical limits go from 1.5 to 3.5 feet above the ground. Since the rules also state that a pitch should be called a strike if "any part of the ball passes through any part of the strike zone," we add one-half the width of a baseball to each of these limits to obtain the rectangle with opposite corners at $(-0.8308, 1.3775)$ and $(0.8308, 3.6225)$. This rectangle is pictured in Figure 1.

As Figure 1 illustrates, the rectangular rule-book strike zone differs from how the strike zone is officiated in actual games. However, spray charts of called strikes are not appropriate for estimating the borders of the called strike zone, since they show only where called strikes are likely to occur, which is biased according to where pitchers tend to throw. For example, Figure 1 indicates that called strikes occur less frequently on the inside corners than on the outside corners, but this effect could simply be a consequence of pitchers' reluctance to throw inside.

Using a grid of one-inch squares in the plane at the front of the plate, (Roegele 2018) describes a *consensus strike zone* determined by the squares on this grid in which pitches are more likely to be called strikes than balls. In this section, we give a more accurate method for obtaining the borders of the consensus strike zone using kernel density estimation (Venables and Ripley 2010). We can also apply this technique to calls made by individual umpires, giving ways to assess conformity and zone size.

3.1 Kernel density estimation

Let $s(x, y)$ be the two-dimensional probability density function describing the distribution of called strikes in the plane. That is, $s(x, y)$ gives the density of the probability that a called pitch will cross the plate at location (x, y), given that the pitch is called a strike. In order to describe the consensus zone, we would like to compute the reverse conditional probability, that is, the probability density $f(x, y)$ that a called pitch will be called a strike, given that it crosses that plate at location (x, y). Let $\hat{s}(x, y)$ be a two-dimensional kernel density estimate computed on the $(\texttt{px}, \texttt{pz})$ coordinates of called strikes, and let $\hat{c}(x, y)$ be a two-dimensional kernel density estimate computed on the coordinates of all called pitches. Then by Bayes' theorem, an estimate for $f(x, y)$ is given by

$$\hat{f}(x, y) = \frac{\hat{p} \cdot \hat{s}(x, y)}{\hat{c}(x, y)},$$

where $\hat{p}$ is the proportion of called pitches that are strikes. The 50% contour of $\hat{f}(x, y)$ will then be the border of the consensus zone.

Figure 10 shows the smooth contours produced using this method, along with the discrete approximation of the

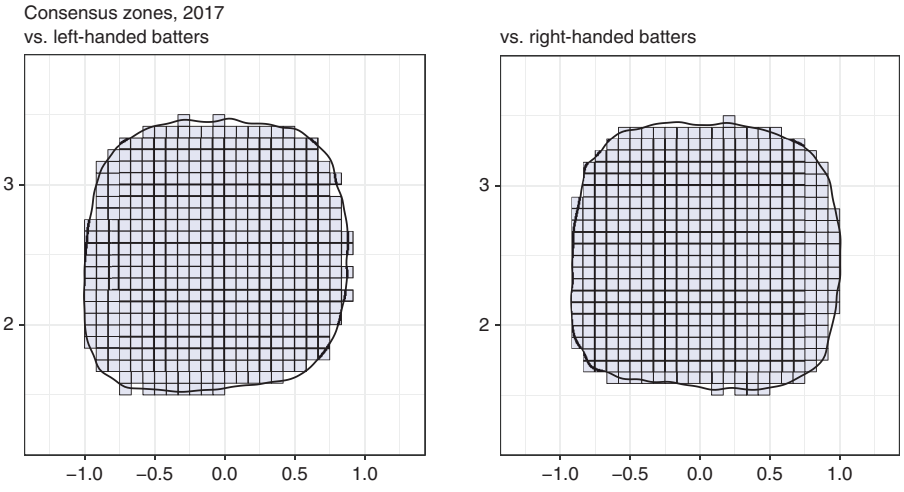

Figure 10: The smooth curves are the boundaries of the consensus zones computed using kernel density estimation. The shaded squares show the result of computing the consensus zone using the discrete method of (Roegele 2018).

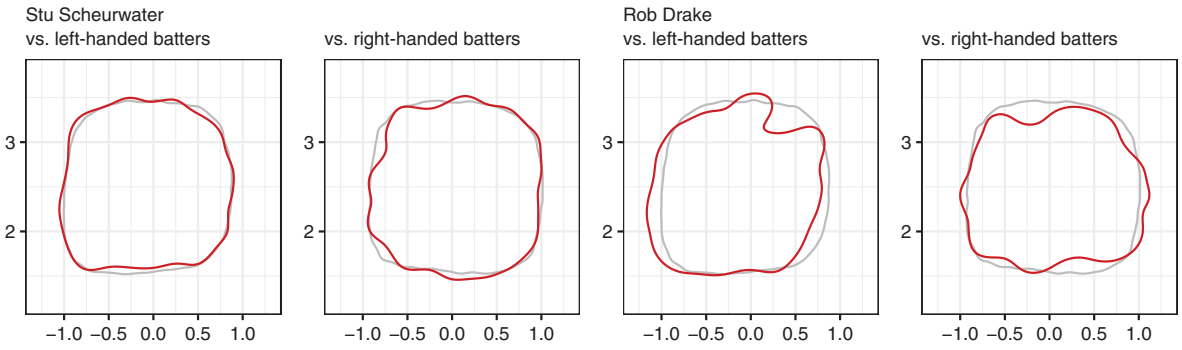

Figure 11: Measured by symmetric difference with the consensus zone (in gray), Stu Scheurwater had the most conforming zone, while Rob Drake's zone was the least conforming.

method described in (Roegele 2018). In addition to improving the resolution of the zone boundary, the kernel density estimation method will work well for smaller samples of pitches. In particular, the kernel density estimation method can produce 50% contours for each MLB umpire's calls over the course of a season, which can be used to describe season-long tendencies.

For example, let X be the region in the plane bounded by the 50% contour for a particular umpire, and let C be the consensus zone described above. The symmetric difference $(X \backslash C) \cup (C \backslash X)$ is the set of all points lying in one zone and not in the other, and its area D_S measures the extent to which the umpire's zone deviates from the consensus zone. To illustrate the extent to which zones can conform to the consensus zone, Figure 11 shows the contour zones for the umpires with the greatest and least values of D_S for the 2017 season, along with the consensus zone.

3.2 Contour-based zone accuracy and size

Given any pair of closed curves $\mathcal{Z}_l, \mathcal{Z}_r$ representing the boundaries of strike zones versus left- and right-handed batters, and any set of called pitches, let $A_{\mathcal{Z}}$ denote the proportion of correctly-called pitches, based on the strike zones $\mathcal{Z}_l$ and $\mathcal{Z}_r$. Let C_l and C_r be the 2017 consensus zones computed using the kernel density estimation method, and let $R = R_l = R_r$ be the rule-book rectangle described above. For each MLB umpire who called at least 20 games behind home plate in 2017, we compute the umpire's *consensus accuracy* A_C and *rule-book accuracy* A_R.

For each umpire, the individual contour zones yield a convenient measure of an umpire's zone size S. It has been suggested (Roegele 2017) that accuracy measurements can function as a proxy for zone size, since inaccurate umpires would tend to have larger strike zones. Our data and measurements do not provide evidence for this assertion. Figure 12 illustrates the associations between these two measures of accuracy, A_C and A_R, along with zone size S. The accuracy measures A_C and A_R are only moderately correlated, and neither is strongly associated with zone size S. The correlation between A_C and S is even weaker if the influential observation with the largest zone (Doug Eddings) is removed.

4 Alternative umpire evaluations

Ball and strike calls have always been subject to the judgment of the home plate umpire. While the MLB rule book offers standards for the extent of the strike zone, Figures 1 and 10 illustrate that variations from the rule-book zone are common, and probably widely accepted. Certainly, it would represent a significant departure from current norms if umpires (perhaps aided by technology) started conforming their zones to the rule-book rectangle. Furthermore, the zone as it is called today could very well be the result of a consensus that has emerged over the years between players and umpires. Therefore, a fair evaluation system for umpires should take history and current accepted practice into account. In this section we consider how the measures of inconsistency and accuracy developed above can complement other measures of umpire performance.

4.1 Correlation and outliers

Ideally, any new metric should evaluate phenomena that previous metrics ignore, since strongly associated measures can yield redundant information. However, it is reasonable to expect that the best umpires are the best at several aspects of the job, so there are likely to be associations among different performance measurements.

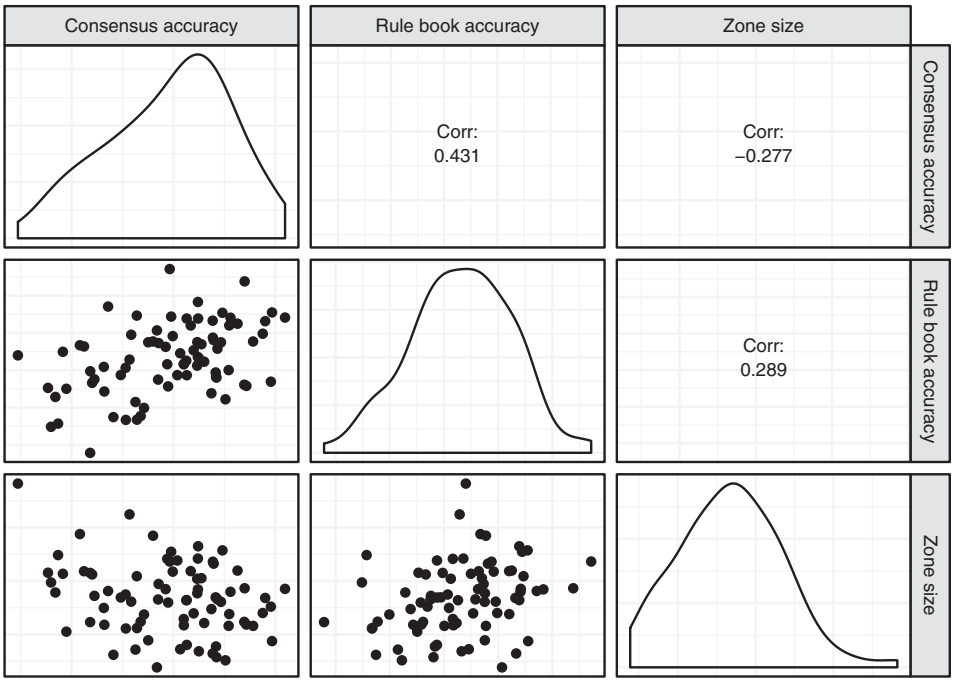

Figure 12: Pairwise correlations of A_C, A_R, and S, for each full-time umpire over the 2017 season. The curves down the diagonal show the distributions of each measure. The scatterplots in the bottom row indicate that measures of accuracy are not strongly associated with zone size.

Figure 13 summarizes the pairwise correlation coefficients for the above measures on all MLB umpires who called at least 20 games behind the plate in the 2017 season. The four inconsistency measures I_{R1}, I_{R10}, I_{CH} and I_{ACH} have been averaged over all the games called by each umpire.

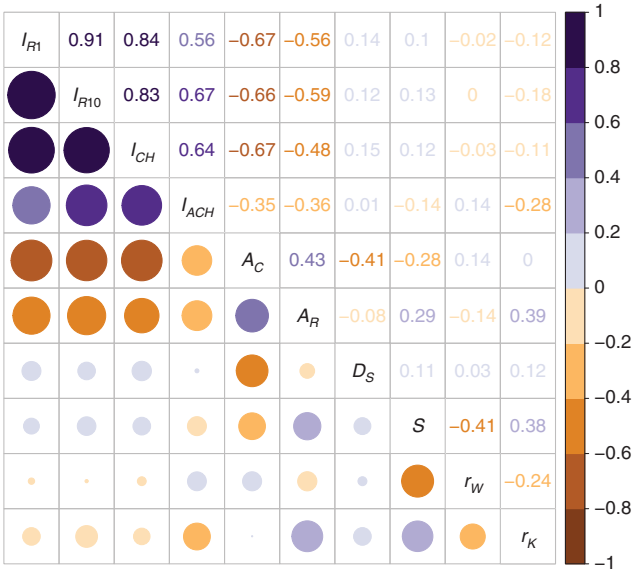

Figure 13: Pairwise correlations for season averages of I_{R1}, I_{R10}, I_{CH}, I_{ACH}, along with A_C, A_R, D_S, S, r_W, r_K, for MLB umpires with at least 20 games called in 2017.

Notice that these inconsistency indices are correlated positively with each other and negatively with the two accuracy measures A_C and A_R. The other measures considered, symmetric-difference nonconformity D_S, zone size S, walk rate r_W, and strikeout rate r_K, generally do not show strong associations with the inconsistency and accuracy measures.

The association between accuracy and inconsistency is not surprising, but it is not strong. In particular, for these umpires, average $I_{R10} + I_{ACH}$ and A_C have a correlation coefficient of $r = -0.58$. The scatterplot in Figure 14 illustrates the negative relationship between consensus zone accuracy A_C and season-average inconsistency, measured as the sum $I_{R10} + I_{ACH}$ of the 10-rectangle and α-convex hull indices. Notable in this graph are the outliers. For example, Carlos Torres, Tim Timmons, and Cory Blaser stand out has having above-average consistency but only average accuracy, suggesting that they would be underrated if evaluated on the basis of accuracy alone.

4.2 Principal component analysis

Any single rating system for umpires can produce a ranking of umpires. When combining several different metrics for umpire performance, we can organize the information

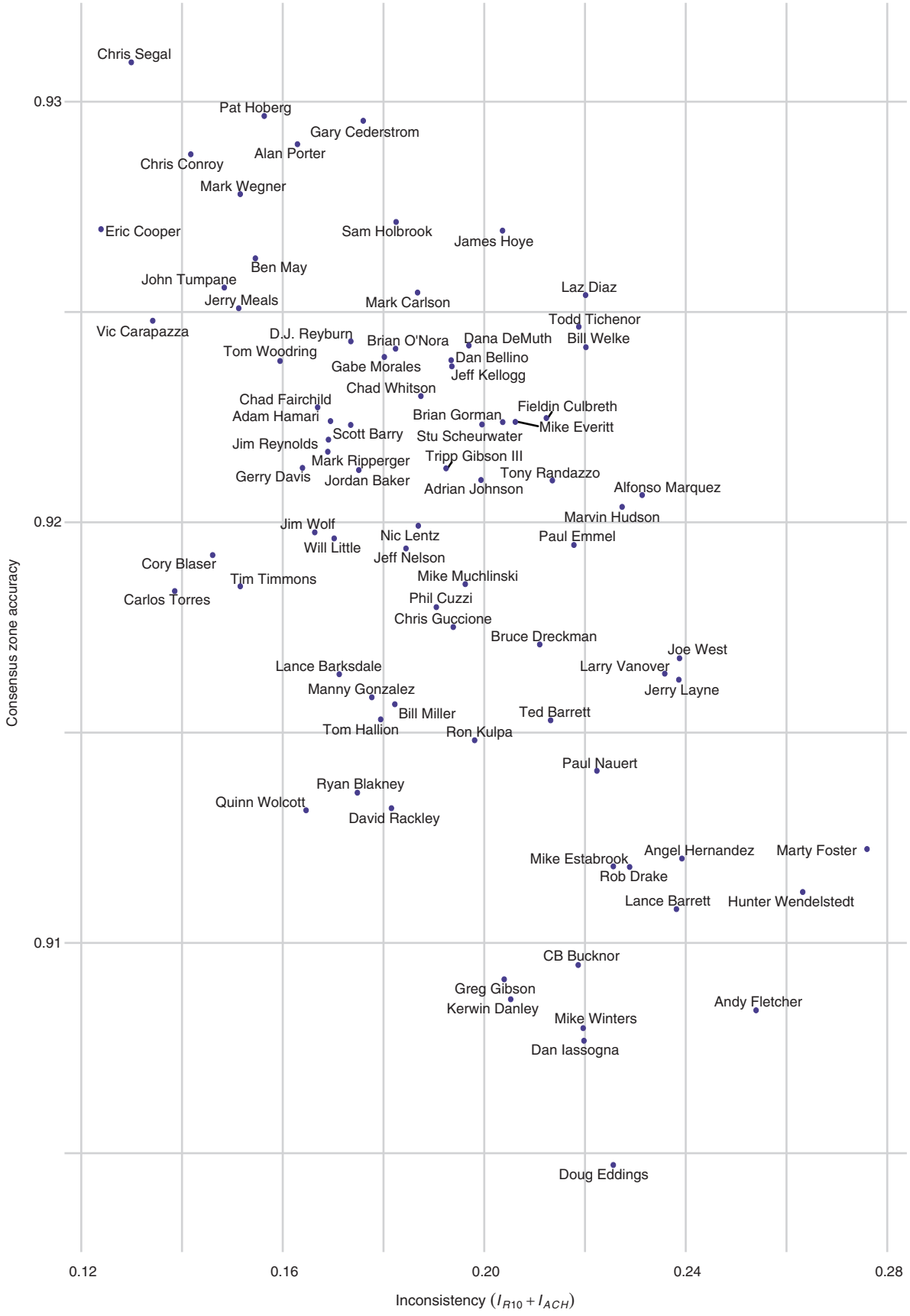

Figure 14: Scatterplot of consensus accuracy A_C versus average inconsistency $I_{R10} + I_{ACH}$.

Table 3: Principal component analysis for four measures of inconsistency, two measures of accuracy, zone size, and walk and strikeout rate.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
I_{R1}	−0.45	0.08	−0.06	−0.09	0.05	−0.24	0.44	0.38	−0.62
I_{R10}	−0.46	0.06	−0.05	0.05	−0.02	−0.29	0.04	0.40	0.73
I_{CH}	−0.44	0.10	−0.14	0.08	0.07	0.00	0.36	−0.79	0.12
I_{ACH}	−0.35	−0.15	−0.28	0.63	0.21	0.15	−0.52	0.04	−0.20
A_R	0.33	0.26	−0.38	0.49	−0.14	0.32	0.51	0.20	0.11
A_C	0.37	−0.24	−0.08	0.33	0.20	−0.79	0.11	−0.13	−0.01
S	−0.02	0.60	−0.04	0.10	−0.62	−0.33	−0.32	−0.12	−0.14
r_W	−0.01	−0.49	−0.69	−0.31	−0.43	−0.03	−0.04	−0.03	−0.01
r_K	0.13	0.48	−0.52	−0.36	0.56	−0.07	−0.18	0.01	0.02

that accounts for most of the variation between umpires by examining principal components.

Table 3 shows the coefficients for a principal component analysis (Mardia, Kent, and Bibby 1980) of the season-average inconsistency indices I_{R1}, I_{R10}, I_{CH}, I_{ACH}, along with consensus accuracy A_C, rule book accuracy A_R, zone size S, walk rate r_W, and strikeout rate r_K, each normalized to have unit variance. The first two components, PC1 and PC2, account for 68% of the variation in the data. Component PC1 is dominated by the accuracy measures (positive) and inconsistency measures (negative), so it seems appropriate to designate it as "strike zone quality." Meanwhile, component PC2 is dominated by walk rate (negative) and strikeout rate and zone size (positive), so this component measures "pitcher friendliness."

The principal components provide a way to summarize the various umpire evaluations developed above. Using the coefficients in each column of Table 3 to form linear combinations of the normalized metrics, we obtain component scores for each umpire. Figure 15 is a scatter plot of the component scores for the first two principal components, PC1 and PC2, labeled by umpire. The most consistent and accurate umpires are those furthest right, while the most neutral arbiters between pitcher and batter are found along the horizontal axis.

5 Conclusions and discussion

The above results illustrate how geometric inconsistency metrics can capture qualities of home plate umpiring not assessed entirely by accuracy, even when measured probabilistically by A_C. We have also shown how the borders of the consensus zone and individual umpire zones can be computed using kernel density estimation, providing efficient methods for comparing umpire tendencies. Such metrics can inform the current debate on the efficacy of human umpires.

While the evidence suggests that MLB umpires are generally quite accurate and consistent, current technology is capable of providing real-time information that would take ball and strike calls out of the hands of umpires and standardize the called strike zone to the rule book rectangle. Doing so, however, would be a significant departure from current practice, and would eliminate facets of the game that arguably contribute to its appeal.

For example, teams value catchers who excel at framing pitches to make them appear as strikes. It is possible that good pitch framers cause umpires to be more inconsistent within a game (Fast 2011a). In addition, accurate pitchers who consistently throw to their catcher's target, even beyond the margins of the zone, can receive favorable strike calls. In such ways, human variation in strike calling influences the way baseball is played, so strike zone analysis can yield insight into aspects of the game beyond umpire performance.

Much of the inconsistency and inaccuracy that we are able to measure could be due to game circumstances. For example, (Walsh 2010) and (Carruth 2012) use the discrete grid method to demonstrate that the called strike zone tends to be larger on 3-0 counts than on 0-2 counts. Other factors, such as the age and experience of the pitcher (Turkenkopf 2008) can influence umpire zones. Presumably, such tendencies vary from umpire to umpire, and could be studied using the tools presented here.

Questions for future investigation include the following.

- Are certain pitch types harder to call consistently?
- What factors contribute to an umpire's strikeout rate? Do more consistent umpires show less variability before and after two strikes have been called on the hitter?
- How do umpire ratings correlate to an umpire's public profile, as measured by press and social media mentions? Are the best umpires those you have never heard of?

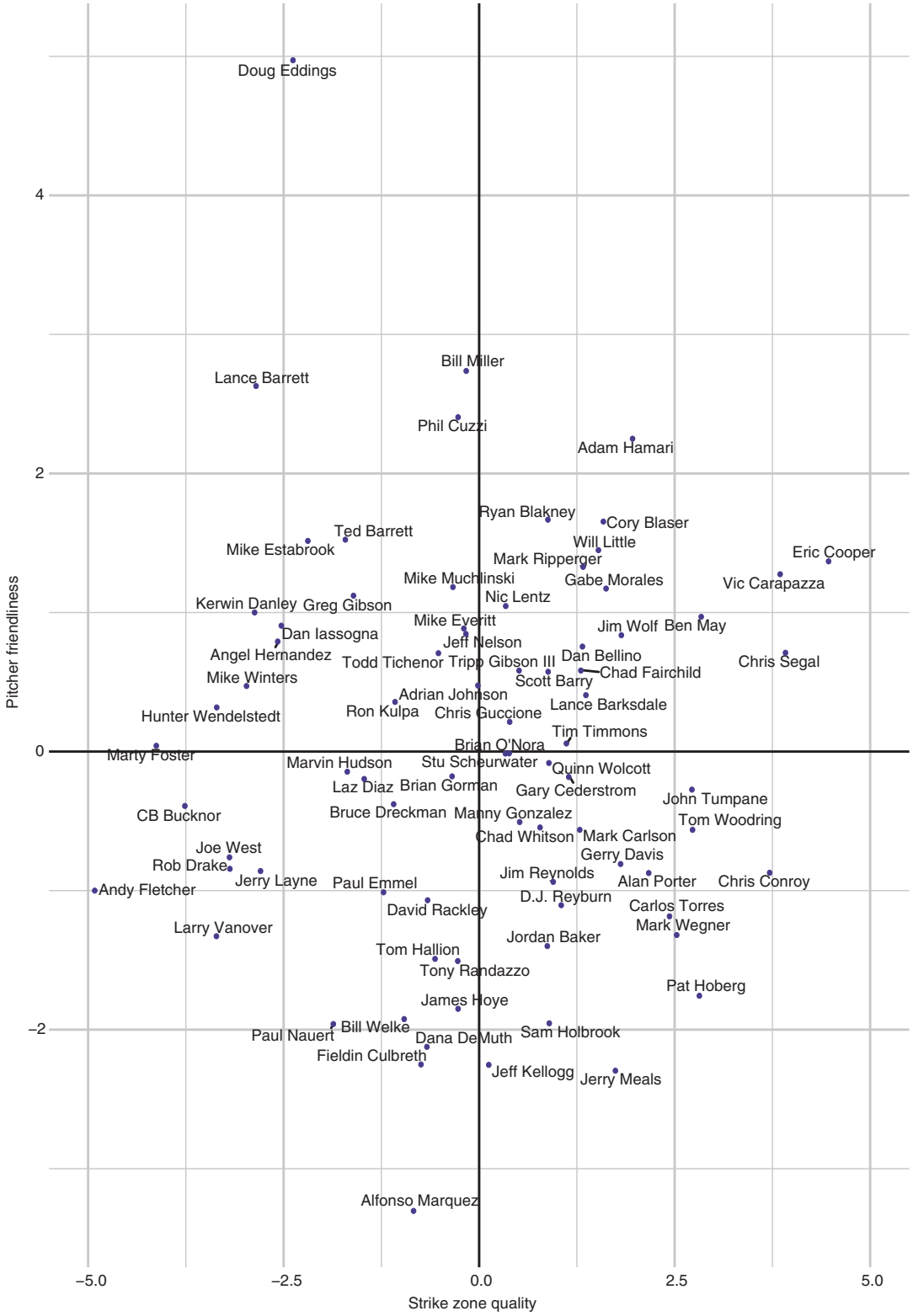

Figure 15: Plot of the first two principal components. The component on the horizontal axis is dominated by accuracy and consistency, designated as "strike zone quality," where the average quality score is zero. The vertical axis component is dominated by walk rate (negative) and strikeout rate and zone size (positive), designated as "pitcher friendliness," with neutrality between pitcher and hitter at zero.

- What has been the effect of the Zone Evaluation system? Have umpires improved in some aspects but not in others? Do the age and years of major league service influence how the strike zone is called?
- Pitch-tracking data has been shown to be noisy (Schifman 2018) (Fast 2011b). In particular, the top and bottom of the zone are estimated by the operator of the pitch-tracking system, based on each batter's stance. Can inconsistency measures be adapted to assess these variations? Can left/right inconsistency be separated from up/down inconsistency?
- The α-convex hull and the n-rectangle indices both generalize to higher dimensions, and the three-dimensional path of a pitch can be approximated using pitch-tracking data. Implement a three-dimensional inconsistency index.
- Are there analogous situations (e.g. in manufacturing), where a geometric measure of inconsistency could be applied?

For reproducibility, data and R code used for the analysis and figures in this paper are available at `https://github.com/djhunter/inconsistency`.

Acknowledgment: The author thanks the anonymous referees for many helpful and constructive comments.

References

Carruth, M. 2012. *The Size of the Strike Zone by Count*. https://www.fangraphs.com/blogs/the-size-of-the-strike-zone-by-count/.

Davis, N. and M. Lopez. 2015. *Umpires Are Less Blind Than They Used To Be*. August. https://fivethirtyeight.com/features/umpires-are-less-blind-than-they-used-to-be/.

Fast, M. 2011a. *Spinning Yarn: The Real Strike Zone*. February. https://www.baseballprospectus.com/news/article/12965/spinning-yarn-the-real-strike-zone/.

Fast, M. 2011b. *Spinning Yarn: The Real Strike Zone, Part 2*, June. https://www.baseballprospectus.com/news/article/14098/spinning-yarn-the-real-strike-zone-part-2/.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1980. *Multivariate Analysis*. Cambridge, MA: Academic Press. ISBN: 0124712525.

Mills, B. 2017. "Technological Innovations in Monitoring and Evaluation: Evidence of Performance Impacts Among Major League Baseball Umpires." *Labour Economics* 46(C):189–99.

MLB. 2018. *Official Baseball Rules*. http://mlb.mlb.com/documents/0/8/0/268272080/2018_Official_Baseball_Rules.pdf.

MLBAM. 2018. *MLB Advanced Media Gameday Data*. http://gd2.mlb.com/components/-game/mlb.

Pateiro-López, B. and A. Rodrıguez-Casal. 2010. "Generalizing the Convex Hull of a Sample: The R Package Alphahull." *Journal of Statistical Software* 34(5):1–28.

Roegele, J. 2017. *Midseason 2017 Strike Zone Review*, July. https://www.fangraphs.com/-tht/midseason-2017-strike-zone-review/.

Roegele, J. 2018. *The 2017 Strike Zone*, March. https://www.fangraphs.com/tht/the-2017-strike-zone/.

Schifman, G. 2018. *The Lurking Error in Statcast Pitch Data*, March. https://www.fangraphs.com/tht/the-lurking-error-in-statcast-pitch-data/.

Turkenkopf, D. 2008. *A Strike Is a Strike, Right?* https://www.beyondtheboxscore.com/2008/4/24/459913/a-strike-is-a-strike-right.

Venables, W. N. and B. D. Ripley. 2010. *Modern applied statistics with S*. New York, NY: Springer.

Walsh, J. 2010. *The Compassionate Umpire*. https://www.fangraphs.com/tht/the-compassionate-umpire/.