Noisy Judgments: A probability surface-based analysis of umpiring variability

Baseball Track
193964

1. Motivation

A measurement is only as good as its instrument. In baseball, umpires are the instrument by which we measure balls and strikes. Umpires are human, which means the measure of balls and strikes involves judgment. In any repeated event involving judgment, we expect variability in the instruments' assessment. To explore this variability, we generated a prior probability surface across the strike zone representing the average strike zone for the umpire corps as it is called within the game. This surface shows the actual behavior of the instruments when making evaluations, providing a novel methodology for assessing trends and variability across Major League Baseball's (MLB) umpire corps.

2. Outline

3. Previous Explorations

In 2007, MLB introduced PITCHf/x, which made pitch-by-pitch location data available to the public. Since then, researchers—using many different methods—have examined the question of the real shape of the strike zone as called during game play.[1]

One of the first analyses to use this data to mathematically describe the strike zone was conducted in 2012 by Matthew Carruth. In this study, Carruth defined the strike zone as the area where a pitch has at least a 50% probability of being called a strike. This research suggests the strike zone is approximately elliptical, as determined by a best-fit through points at the edges of the surface with 50% or greater strike percentage.

Jon Roegele in 2013 expanded our understanding of the shape of the strike zone and the factors that affect that shape by dividing the x-z plane in the front of the plate into a grid of one by one inch squares and binning each pitch into the appropriate grid cell. Roegele defined the strike zone as any grid square where more than 50% of the pitches were strikes. These identified strike zones are more rectangular than Carruth's ellipses. Roegele conducted each analysis within a single season and went on to identify factors including pitch count, out count, pitch type and velocity, batter handedness, and base runner state which affect the surface area of the strike zone.

In 2019, Eli Ben-Porat set out to define the true shape of the strike zone. As Roegele did, Ben-Porat defined the strike zone as the set of points where called pitches have greater than a 50% chance of being a strike. Ben-Porat's model uses pitch data to determine shifts in center location and other descriptive measures of the strike zone to inform the manual construction of a superellipse that visually approximates the strike zone.

The team behind Umpire Scorecards updated their Estimated Umpire Zone (EUZ) calculation for the 2021 MLB season using kernel density estimation and Bayes's Theorem to derive a strike probability function. The EUZ is the 50% contiguous contour line of the probability strike function, or in other words, the boundary at which umpires should change their minds between calling a pitch a ball or a strike.

These methods all provide valuable insights into our collective understanding of the shape of the strike zone.

4. Data

The Statcast database has the precise location of 10.5 million pitches where they cross the front of the plate, covering MLB regular and postseasons between 2008 and 2022[2]. Approximately half of those pitches—or 5,307,386—were called either a ball or a strike by an umpire. These are the pitches of interest for this analysis (Figure 1).

---

[1] MLB changed its pitch tracking system from PITCHf/x to Statcast/TrackMan starting with the 2017 season. For purposes of this analysis, we ignore those measurement effects. Inconsistency in the measurement tool presents a bias to account for in a future iteration of this analysis.

[2] Statcast data is available for the 2023 season. RetroSheet data on umpires for the 2023 season is not yet available as of November 2023.

Statcast data retrieved with the pybaseball python library does not include the name of the umpire for each plate call. For that, we turned to RetroSheet.org. Data available at Retrosheet provided information on the umpiring crew for each game, including who was behind the plate. Joining this umpire data with the pitch data required us to make a key assumption: The umpire does not change within the game. Umpire changes within a game occur only under exceptional circumstances[3], so we are comfortable this assumption does not materially skew the results.

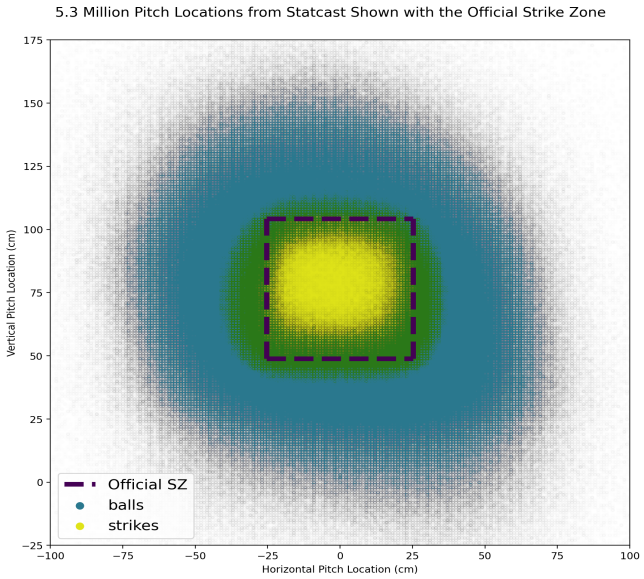


Figure 1: Pitch locations plotted as viewed from the umpire's perspective. Called balls are shown in blue with called strikes in yellow. The official MLB strike zone is overlaid with a dashed line in dark purple. Areas near the edges of the strike zone—the Shadow Zone—show clear variability in ball/strike calls, as evidenced by the wide area of overlapping data shown in green.

For several of our subsequent analyses, we imposed a sample size limit on the number of pitches used to generate a best-fit strike zone. We removed any umpires who called fewer than 1,000 pitches, batters who took fewer than 100 pitches, and pitchers who threw fewer than 100 called pitches.

Our comprehensive data set has the following characteristics[4]:

- Dates: 28 March 2008 to 5 November 2022
- Pitches: 5,307,386
- Umpires: 139
- Pitchers: 1,433 across 30 teams in 30 ballparks
- Batters: 1,688 across 30 teams in 30 ballparks
- Catchers: 228 across 30 teams in 30 ballparks

5. Methods and Assumptions

5.1. The MLB Strike Zone

[3] Umpire replacement occurs in the case of illness or injury, according to MLB Rule 8.02(d).

[4] A small fraction of data records were incomplete, missing pitch position information and/or details about the umpire, pitcher, batter, etc. As a result, the precise count of pitches used in each analysis varies slightly.

The official MLB strike zone is the area over home plate from the midpoint between a batter's shoulders and the top of the uniform pants—when the batter is in their stance and prepared to swing at a pitch—and a point just below the kneecap (see Figure 2).

In order to get a strike call, part of the ball must cross over part of home plate while in the aforementioned area. The official strike zone is therefore a 3-dimensional pentagonal prism. Pitch location data, however, is measured at the point the center mass of the ball crosses the front of home plate, so we flatten our strike zone from a 3-dimensional prism to a rectangle with an x (horizontal) and z (vertical) dimension. The x measurement is constant: The width of the official strike zone–as measured by the center of mass of a baseball–is the width of home plate + 2 * (the radius of a baseball) = 43.2 + (2 * 3.7) = 50.6 cm. The z measurement, however, takes some additional calculation as we'll explore in the next section.

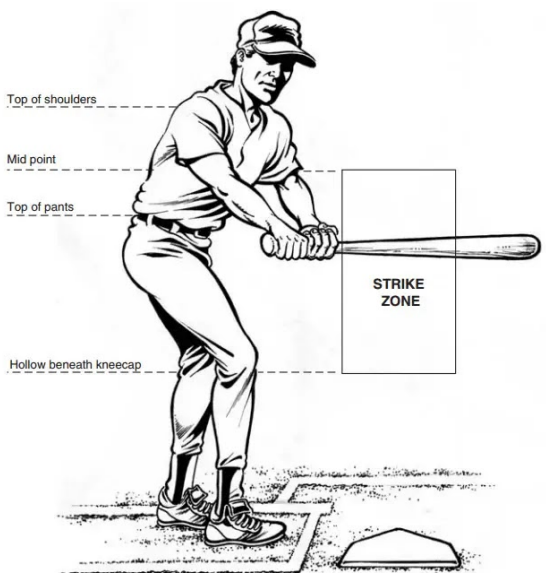


Figure 2: A schematic diagram showing the location of the official MLB strike zone. The width of the strike zone is defined as the width of home plate, and the height is defined as stretching from the midpoint between a batter's shoulders and belt to just below the kneecap. The precise location for the top and bottom of the strike zone varies with batter height and stance.

5.2. Height Normalization

One only has to see mid-2000s highlights of Dustin Pedroia and Richie Sexon to understand that batters have different heights and stances as they approach their plate appearance. Because of the way the strike zone is defined, each batter has a strike zone with a unique height. Further, each batter's strike zone is not necessarily consistent between at bats insofar as they can alter their stance. Thus, defining a single strike zone or comparing strike zones between umpires requires standardizing the height measurement.

For our analysis, we began by using the top and bottom of the strike zone for each pitch as reported by Statcast. From these values, we calculated the average value (arithmetic mean) for the top of the strike zone (104 cm) and the average value for the bottom (49 cm). These values define the vertical extent and location of the strike zone for the average batter in the data set. We then proportionally scaled the data for each individual pitch into the average strike zone height frame of reference. For example, if a pitch was 25% of the height of the reported strike zone above the bottom edge of the zone, then its adjusted z-value in the normalized strike zone would be 25% above the bottom edge as well.

5.3. The Best-Fit Strike Zone Surface

We determined the strike zone as called during games by aggregating umpires' evaluation of the 5.3 million pitches and fitting a well-defined surface to the data. Constructing this probability surface has four phases. First, we gridded the data into 1 cm by 1 cm square cells. Based on the x-z position information from Statcast, we placed each individual pitch into a single grid cell. Once all pitches were assigned to grid locations, the number of called balls and strikes in each cell defined the strike percentage for that location in the grid as shown in Equation 1.

$$Strike\ \% = \frac{N(strikes)}{N(strikes) + N(balls)} \quad (1)$$

Second, we calculated expected errors in our strike percentages based on the sample sizes in each individual grid cell. This is an important consideration for any subsequent fitting or optimization process. Each grid cell is a small region within the larger dataset, and the pitches within a given cell are effectively voting on the likelihood that a new pitch in that location would be called a ball or a strike. Although the overall sample size of 5.3 million pitches is large, the number of pitches present in any given grid cell can be considerably smaller. This is especially true as the data set is further subdivided by batter handedness, individual umpires, etc. Thus, utilizing the heteroscedastic errors present in each cell location during the fitting process is necessary to mitigate noise introduced by small(er) sample sizes and to ensure more consistent results.

Third, we constructed a functional form for the probability surface with the desired properties. We expect umpires to call pitches near the center of the strike zone as strikes almost 100% of the time. Conversely, pitches far from the center will be called strikes 0% of the time. This behavior is true in both the vertical (z) and horizontal (x) dimensions. Sigmoidal curves are a form of off- / on-style function common in many analytic disciplines. They map a continuous, infinite domain onto a finite interval, typically the interval [0-1] when dealing with probabilities. They provide the general behavior needed for fitting the strike zone surface.

The value of a sigmoidal function asymptotically approaches 0 or 1 depending on the direction of travel along the curve; its midpoint is shifted via a single free parameter, and the steepness of the transition from 0 to 1 is adjusted by a second free parameter (Figure 3).

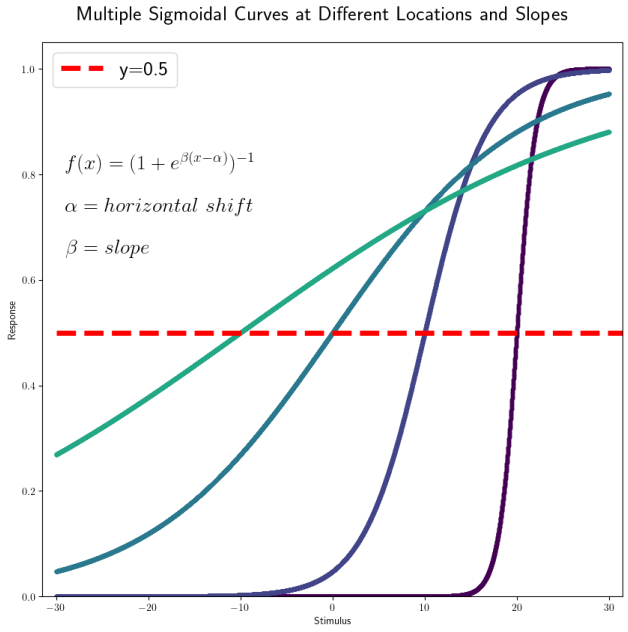


Figure 3: Four different sigmoidal curves are shown plotted along arbitrary x-y axes. A dashed red line highlights the 0.5 threshold on the y-axis, and the equation for a sigmoid is shown overlaid in the plot. As values for α and β are varied, the resulting curve shifts left and right as well as in severity of s-shape.

Our constructed fitting function, continuously defined in x and y, has two sigmoids in the x direction and two in the y direction. The two sigmoids in each dimension account for the rise and then fall of the strike probability as you move across the plate. There are a total of eight free parameters manipulated by the optimizer during the fitting process, two for each of the four sigmoids. The full function describing the general shape used to fit each strike zone is shown in Equation 2.

$$F(x,y) = \frac{\left(1+e^{-\beta_0(x-\alpha_0)}\right)^{-1}}{1+e^{-\beta_1(x-\alpha_1)}} * \frac{\left(1+e^{-\beta_2(y-\alpha_2)}\right)^{-1}}{1+e^{-\beta_3(y-\alpha_3)}} \quad (2)$$

Fourth and finally, we optimized the eight fit parameters using standard gradient descent optimization routines available in the python library, scipy.stats. However, in order to account for heteroscedastic sampling errors, we created a custom cost function for the optimizer that first normalized distances between the data points and a candidate fitted surface before summing them. This additional step in the fitting process accounted for sampling errors in each grid cell when calculating the root mean squared error (RMS) between the data and each successive iteration of the fitted surface.

The eight parameters in Equation 2 are broken into two groups, alpha and beta. The alpha parameters are positional terms that define where the value of the surface is 0.5—or a 50% chance

to be called a strike. These alphas define the edges of the strike zone after the fitting process is completed. The beta terms determine how steeply the probability surface rises from 0 to 1 or falls from 1 back to 0.

The left-hand side of Figure 4 shows the gridded data in the x-z plane with the strike fraction from 0 to 1 on the probability axis. The right-hand side of the plot shows our best-fit surface. It is clear the fitted probability surface is an excellent representation of the underlying data.

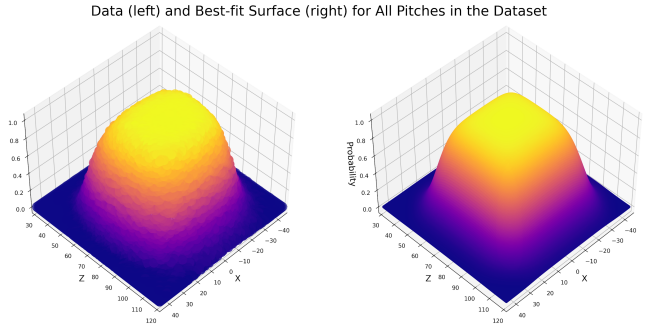


Figure 4: The left panel shows gridded strike fraction data. The x-axis displays horizontal pitch location from the umpire's perspective with pitch height along the z-axis. The third dimension in the plot is the probability axis and varies from 0 to 1. The probability axis tracks the fraction of data called a strike within each grid cell. We see cell-by-cell variability near the edges of the well-defined central strike zone peak. The right panel shows the best-fit model to the data using Equation 2 and the sampling error weighted optimization described in Section 5.3.

Because of the construction of Equation 2, the cross-sections of this surface (constant probability values) are largely rectangular with slightly rounded corners. For the purposes of our analysis, we extracted the alpha terms from each fit and used those positions to define a rectangular strike zone. Umpires are presumably attempting to call pitches according to the official strike zone definitions. Although prior work mentioned in Section 2 has shown the corners of the strike zone as actually called by umpires are somewhat rounded, we chose to measure and extract the height and width of the strike zone to define the rectangle umpires think matches the official zone. This allows for clear comparisons between attempt and execution for each umpire.

6. Results

Looking at the full set of pitches, the best-fit surface gives us a strike zone 57.7 cm wide–7.1 cm wider than the official MLB zone width (Figures 5 and 6). The center of the strike zone is shifted 1.6 cm to the left when viewed from the umpire's perspective. The modeled strike zone is 56.1 cm tall while the average height reported by Statcast is 55.3cm. The shift of the strike zone 1.6 cm to the left is driven by differences between left-handed and right-handed batters and is discussed further below.

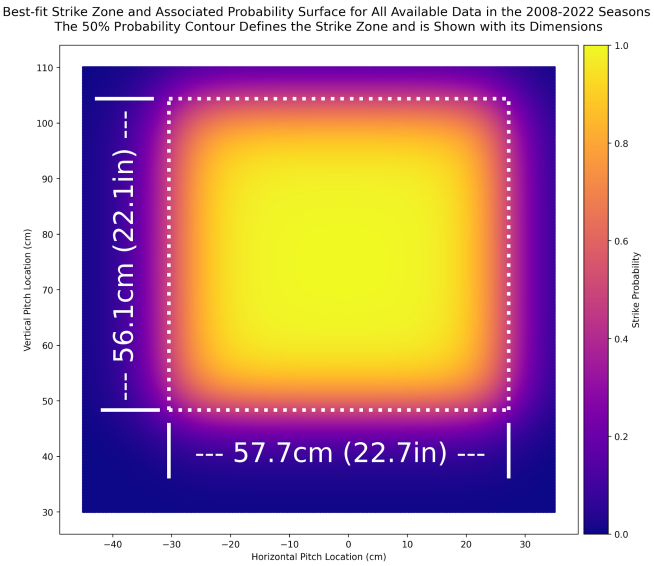


Figure 5: A top-down view of the best-fit strike zone surface shown in the right panel of Figure 4. The probability of pitch being called a strike at each location is shown via the colorbar. The 0.5 contour represents the edges of the strike zone. The rectangle constructed from the four fit parameters defining that 0.5 contour is shown as a dotted white box; its dimensions are overlaid on the plot.

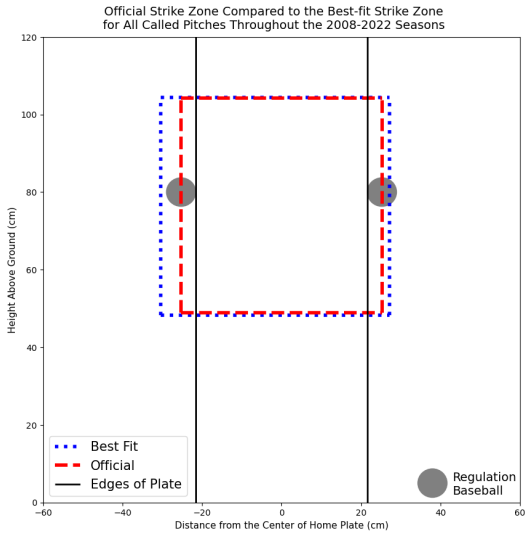


Figure 6: A view of the strike zone from the umpire's perspective. The edges of home plate are marked by solid black vertical lines near ±21.6 cm on the x-axis. The official strike zone as defined in section 5.1 of MLB's rulebook is shown as the dashed red box. Regulation size baseballs are shown as gray circles and indicate center-of-mass locations for pitches that are tangent to home plate. The blue dotted box shows the strike zone derived from the best-fit surface to the full data set used in the analysis. (The blue-dotted box in this plot corresponds to the white-dotted box in Figure 5.) This representation of the data highlights the differences between the official strike zone and what is actually called during play.

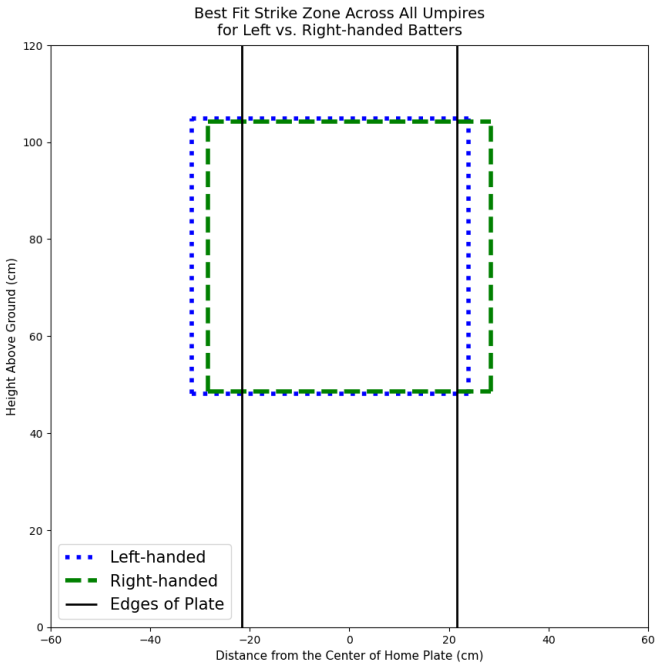
When we derive a best-fit surface for each individual umpire across all seasons, the results confirm what we see in the collective surface. The majority of umpires call strikes wider than the official zone, roughly in line with the top of the zone, and slightly lower than the bottom of the zone. Umpires show greater variability in calling the bottom of the zone than the top. Trends over time and other factors are discussed in the following sections.

6.1. Trusting the Ruler

To test the reliability and consistency of the model, we ran an assortment of tests. The difference in the position of the strike zone between left- (LHB) and right-handed batters (RHB) is well documented in baseball analytics. Our data set has 3 million called pitches to RHB and 2.3 million pitches to LHB. We observe the area of the strike zone is nearly identical between both sets of batters: 3,156 sq. cm for RHB and 3,146 sq. cm for LHB. The righty strike zone measures 56.8 cm

width and is centered over the plate. However, when we look at the strike zone for LHB, it is 55.6 cm wide and the midpoint is shifted 3.9 cm towards the outside of the plate. This shift of our best-fit surface for LHB conforms to prior research, observed evidence of the "lefty strike," and anecdotal complaints from baseball players, thus leading us to trust the model as a ruler for umpire strike zones.

Figure 7: The shapes and positions of the best-fit strike zones for left-handed batters (blue, dotted) and right-handed batters (green, dashed) are shown along with the edges of home plate as vertical black lines at ±21.6 cm. The difference in the centroid and extents along each edge are apparent and discussed in Section 6.1.



The shift in the strike zone for LHB holds true across the umpire corps, i.e. the surface isn't being pulled towards the outside edge because of a few outliers. We also know that the area of umpires' strike zones does not shift between RHB and LHB, but there is a clear shift in aspect ratio (width/height). The only way these truths all hold mathematically is if the strike zone for LHB generally gets a little narrower and a little taller while it shifts. This is exactly what we see in Figure 7: The blue box shifts, but the shift is bigger on the right than on the left—so narrower overall—and the top and bottom extend slightly past the green box.

We can break down the strike zone surface one more time between left-handed pitchers (LHP) and right-handed pitchers (RHP). The lefty-strike is even more pronounced when there is a RHP. Umpires call roughly the same inside edge of the strike zone for LHB regardless of pitcher handedness, but RHP get the outside strike against LHB more frequently than their left-handed counterparts. The magnitude of the strike zone shift between LHB vs. RHB for RHP is roughly double the shift for LHP.

6.2. The Sensitivity of the Ruler

The 5.3 million called balls and strikes used to create the overall best-fit strike zone shown in Figure 5 provide a large sample size for the resulting probability surface fitting process. However, as the number of data points decreases, the influence of any single pitch on the resulting fitted

surface necessarily increases. The use of a well-motivated functional form for the probability surface helps mitigate this effect, but variability in the results caused by small sample sizes are unavoidable and place limits on the strengths of any conclusions based on small subsets of the data.

To help understand the behavior of the method as data counts decrease, we conducted a sensitivity study at a variety of sample sizes. First, we split the data by season and by batter handedness. This resulted in 30 subsets of the data (15 seasons by 2 types of batter). Each data subset had 150-200,000 thousand pitches. From one of these subsets we randomly sampled 50,000 pitches with replacement and created a best-fit strike zone from that sample. We then created a different random sample from the subset and fit a new strike zone. This process of fitting on random samples was repeated 100 times. This resulted in 100 slightly different 'best-fit' strike zones each based on a slightly different sample of 50,000 pitches from a population of ~200 thousand pitches. This was then repeated for progressively smaller sample sizes down to only 1,000 pitches (see Figure 8). We repeated the entire process for each of the season-handedness combinations.

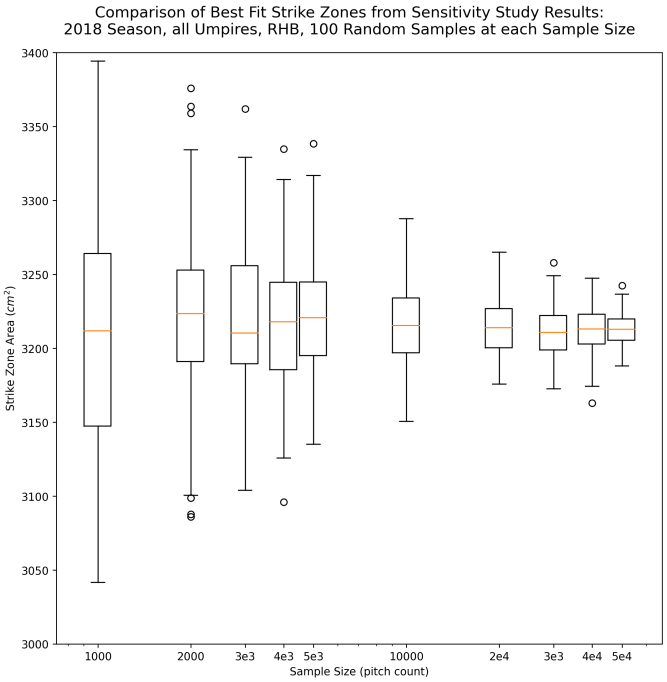


Figure 8: Boxplots showing the variation in the area of the best-fit strike zone for sample sizes ranging from 1,000 pitches to 50,000 pitches. Samples of each size were randomly drawn 100 times from the full population of pitches in the particular season.

The results of our sensitivity study were encouraging. Figure 8 shows data for RHB in the 2018 season. When sample sizes are small the influence of a particular umpire, game, etc. are relatively large. The variation seen in the left-most boxplot in Figure 8 shows this well. Some amount of this variability can be attributed to including all umpires in the data sample. Since they call strikes differently from each other, grouping their calls into small samples increases the spread of the resulting boxplot. The remainder of the variability shown in the boxplots comes from the process of random sampling itself. As the sample size increases (moving to the right in Figure 8), it's clear the variability decreases. Seeing this smooth change in the results along with the unimodality of the individual distributions indicates the method is generally mathematically well-behaved.

The methodology of the sensitivity study also demonstrates a powerful byproduct of the overall method. Although there is no closed-form derivation for the precise expected distribution of a

particular fitted parameter of the best-fit strike zone, we can construct that distribution by repeatedly fitting random sub-samples of the data for a given sample size. This is discussed further in Section 7.1.

6.3. Time Series

Between the 2008 and 2022 seasons, the strike zone narrowed nearly 14%, from 63 to 54 cm. The majority of the change occurred because the left side of the strike zone—as viewed from the umpire's stance, so outside pitches for LHB and inside pitches for RHB—moved towards the plate about 7 cm. The 'lefty strike' largely disappears starting with the 2017 season. As a result, the midpoint of the strike zone moved from more than 3.6 cm to the left of center in 2009 to nearly on center by 2022 (Figure 9).

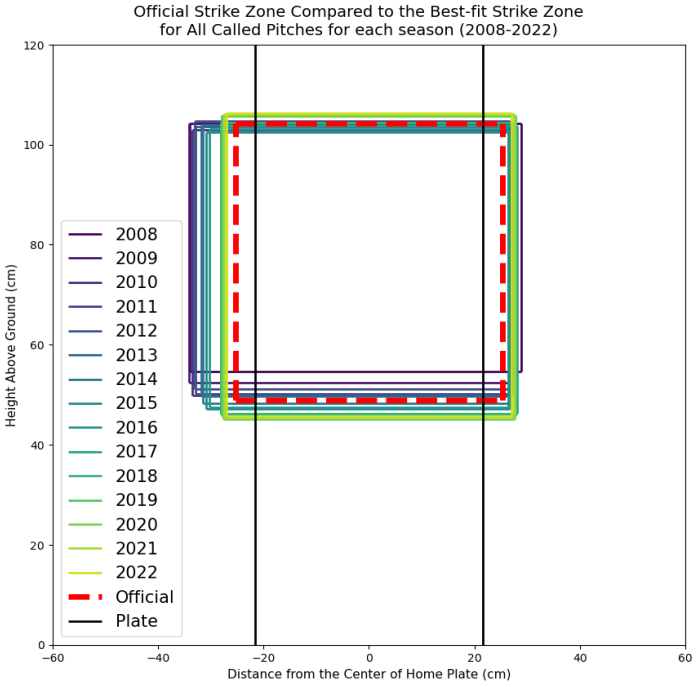


Figure 9: The plot shows each best-fit strike zone for the 2008-2022 seasons as seen from the umpire's perspective. The edges of home plate are the vertical black lines at ±21.6 cm and the official MLB strike zone is the red dashed box. Color-coding for each season reveals the trend over time.

On the z-axis of the strike zone, we see the upper limit of the strike zone has remained relatively consistent across the umpire corps since 2008, rising by 2 cm (2%) during that period. The bottom of the strike zone, however, dropped by nearly 9 cm—more than 16%—between 2008 and 2022. Umpires are increasingly calling the low strike.

Overall, we see an increase in the size of the strike zone (Figure 10) and a 29% drop in the aspect ratio (width/height) for the 15-year time period (Figure 11). In 2008, the strike zone was wider than it was tall. That ratio decreased steadily, flipping by the time we get to the 2018 season, passing through a relatively square strike zone in the 2016-2017 seasons.

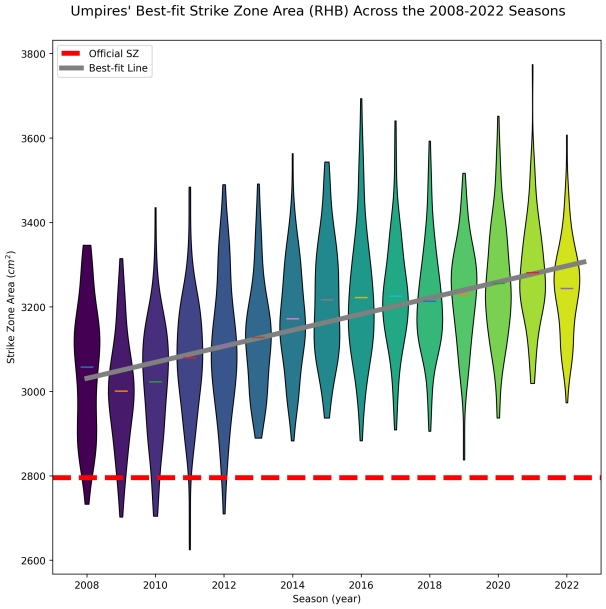


Figure 10: Season-by-season distributions of best-fit strike zone areas for the entire umpiring corps. Mean values for each season are indicated by dashes near the centers of the distributions. The area of the official MLB strike zone is marked by the horizontal red dashed line near 2,800 cm^2. The best-fit line to the season-by-season mean values is shown as the gray line. There is a small, but significant positive slope to the fit.

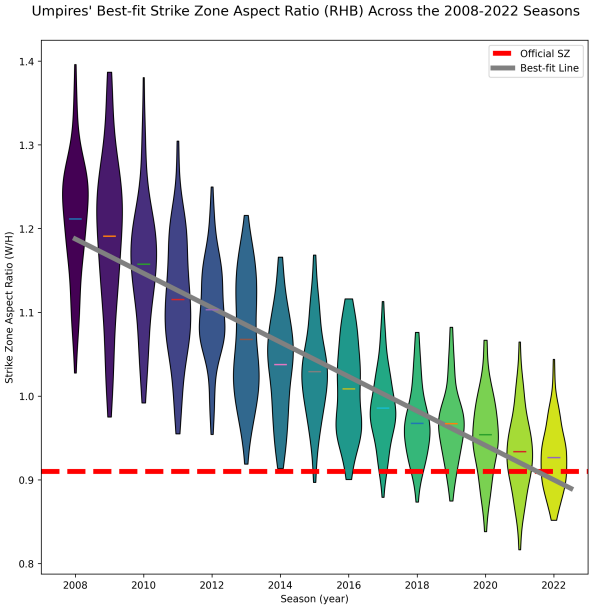


Figure 11: Season-by-season distributions of the best-fit strike zone aspect ratios ($\frac{width}{height}$) for the umpire corps. Mean values for each season are indicated by dashes near the centers of the distributions. The aspect ratio of the official strike zone is marked by the horizontal red dashed line near 0.9. A best-fit line to the season-by-season mean values is shown as the gray line. There is a clear trend toward narrower, slightly taller strike zones.

6.4. Umpire Evaluation

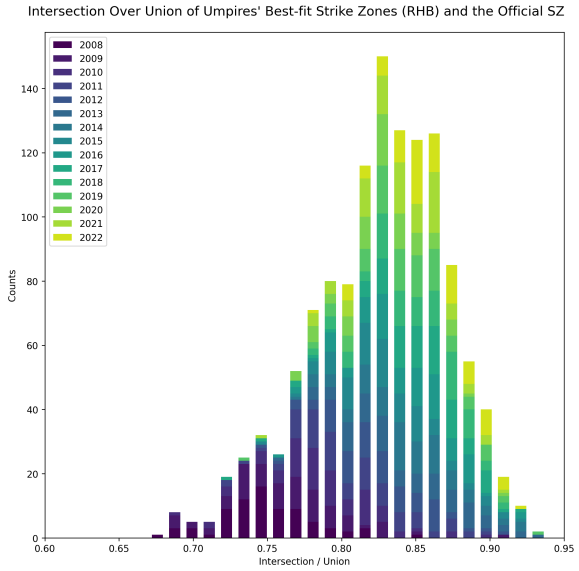
To evaluate umpire accuracy, we overlap the umpire's individual strike with the official MLB zone. We calculate the percent overlap as the intersection-over-union—the IOU—ratio (Equation 3).

$$IOU = \frac{Area\ of\ overlap\ between\ the\ two\ strike\ zones}{Total\ combined\ area\ of\ both\ strike\ zones} \quad (3)$$

Overall, the IOU ratio for the umpire corps ranges from 70 percent to 90 percent, with the majority of umpires clustering around the 84 percent mark (Figure 12). We see a distinct seasonal pattern in the IOU ratio: More recent seasons have a higher IOU, i.e. more overlap with the official strike zone. The tails of the histogram for each season become the middle of the distribution for subsequent

seasons. This mirrors what we see in Figure 9 as the best-fit strike zone moves down and to the right with each successive season.

Figure 12: This histogram shows the IOU between the official MLB strike zone and the best-fit strike zone for each umpire for each season for RHB. The distribution is roughly gaussian with a noticeable skew to lower values. The color scale indicates each season and shows the left skew is the result of changes in the strike zone shape over time.



Do individual umpires change their strike zone over time? Or are the changes seen in the overall strike zone driven by umpire changeover? Figure 13 shows the same information as Figure 12 yet adds in the dimension of tenure along the y-axis. The pattern reveals umpires who started at the same time move together as cohorts, changing the size, shape, and position generally at the same time. We see this as the black dots—lower IOU—are apparent at all tenures in the early seasons and trend to higher IOU as we move toward successive seasons. This indicates umpires are changing their strike zones over time; new additions to the umpire corps are generally adapting to the existing zone rather than driving the changes.

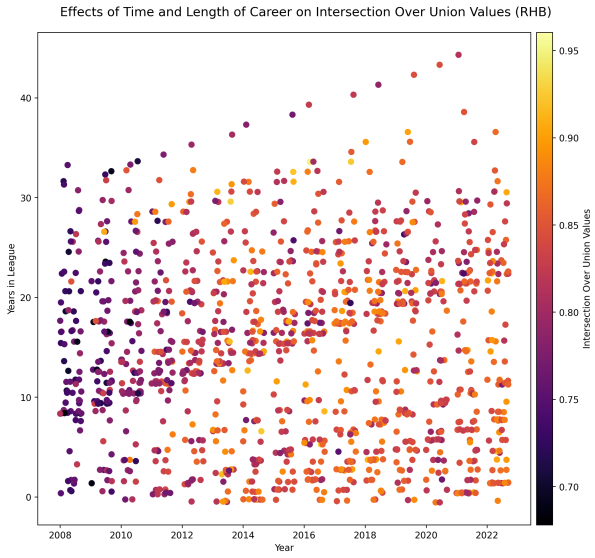


Figure 13: A scatter plot showing IOU for each umpire season and by tenure. Each dot is a specific season for a single umpire. The color scale indicates the IOU ratio from low (black) to high (yellow). The season is indicated on the x-axis and the umpire's tenure on the y-axis. Cohorts of umpires who start together move up and to the right in clusters. Points are plotted with small random variations (left and right) to minimize overlaps.

7. Discussion

Modeling the strike zone's best first surface presents opportunities to examine how players and coaches react to the realities of the umpires' strike zones by slicing the data along multiple dimensions, including time, pitcher, pitch type, batter, handedness, game situation, and count. We can directly evaluate umpires using the best-fit strike zone surface because the definition of the official strike zone—the umpire's presumed target—has not changed within the dataset. We have to normalize by season and batter handedness when we conduct situational and player analysis, however, because these players are adjusting their approach to the changes in the size and position of the strike zone umpires are actually calling. We explore some of these analytic lines here, keeping in mind the height normalizations detailed in Section 5.2 when comparing absolute vertical measurements.

7.1. The 2-Strike Bias

There has long been discussion of a '2-strike bias,' a tendency for umpires to be a little stricter on pitchers for calling strike 3. The method described in this paper provides a natural way to test for the existence of the 2-strike bias and measure of the statistical significance of the result.

As introduced in Section 6.2, we can construct an expected distribution of parameters by repeatedly fitting strike zones to random sub-samples of the data. The resulting distribution of those fits provides a measure of the variation of a given fit that is to be expected from random sampling noise alone. Comparing the strike zone fit for the actual subset of interest to this distribution can determine if any differences between the two are from real effects or are readily explainable by sampling noise.

To test for the 2-strike bias we first split the data by season and batter handedness. Within each of the resulting 30 groups, we counted the number of called pitches on a 2-strike count to determine the sample size. This was 30,000 - 40,000 pitches depending on the group. We then made fits for 1,000 random samples of that sample size from the entire data group (all pitches/counts). Figure 14 shows the distribution of these fits for the 2008 season as the blue histogram. The area of the best-fit strike zones is centered around 3,030 cm^2 with a standard deviation of ~20 cm^2. The area for the 2-strike pitches' fit is shown in red and is more than 400 cm^2 smaller; it is an outlier by any definition of the word.

If you were to fit a gaussian distribution to the blue histogram in Figure 14 and use the result as a probability distribution function, the odds of randomly sampling a group of pitches from the 2008-RHB data that would result in a fitted strike zone with an area as small as the 2-strike strike zone (or smaller) would be roughly 1 chance in 10^{104}. Put plainly, the 2-strike bias is very, very real.

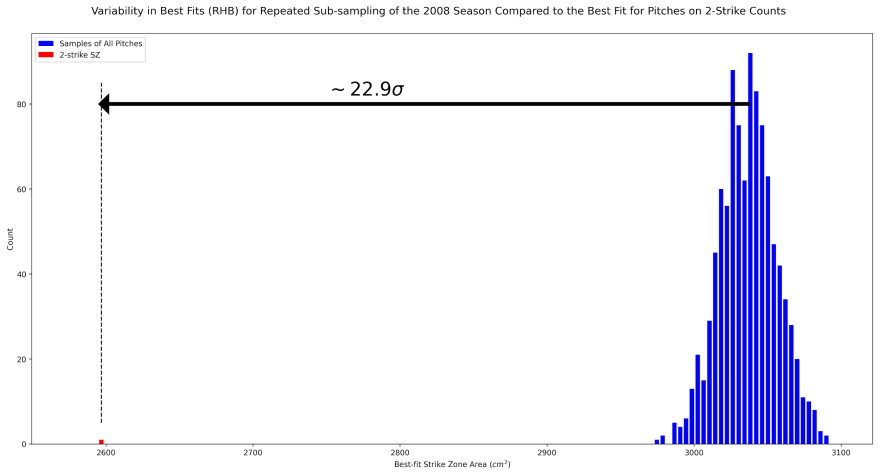


Figure 14: The areas of the best-fit strike zones for 1,000 random samples from the 2008 season are shown in blue. The sample size was chosen to match the number of pitches corresponding to 2-strike counts where the next pitch was taken. The area for the fit to the '2-strike' pitches is shown in red.

To look at the 2-strike bias over time, we repeated the process described above for each season (and RHB/LHB). Figure 15 shows the 2-strike bias has remained broadly consistent through the seasons and has generally tracked the changes seen to the overall strike zone. In addition to the red points showing the area of the 2-strike fit, Figure 15 also shows the areas of the fits made from all other pitches in blue. The 0-or-1-strike zones trend toward higher areas, though to a lesser extent than the decrease seen for 2-strike counts. This is at least partially driven by the relative sizes of the two groups; 2-strike counts made up roughly 1 in 5 called pitches in the data set. The areas of the fits for all pitches combined are shown as the black points seen within the boxplots.

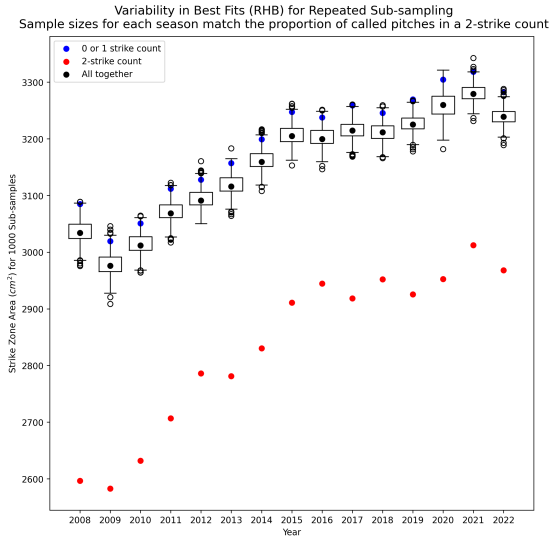


Figure 15: The same information available in Figure 14 is shown here as the boxplot and single red point in 2008. Similar boxplots and 2-strike bias points are shown for the 2009-2022 seasons. It's clear the 2-strike bias is consistent through time and changing strike zones. Although the magnitude of the bias may vary slightly, it is highly statistically significant across all of the available data.

Although the statistical significance of the 2-strike bias appears robust, it's important to also look at the effect's size. It's possible to look at the shifts overtime in each of the individual fit parameters, but it's clearest to simply look at the resulting strike zones. Figure 16 shows the 2-strike bias results in the left panel and the 0-or-1-strike zones on the right. The changes over time are shown by the color coding that matches Figure 9.

The 2-strike bias is primarily a bias against high and low strikes. There are noticeable shifts on the edges of the plate, but the dominant driver of the decrease in size for the 2-strike zone is the tightening of the top and bottom of the zone. This can be seen by comparing each seasons' two zones in Figure 16 (matching colors in the left/right panels).

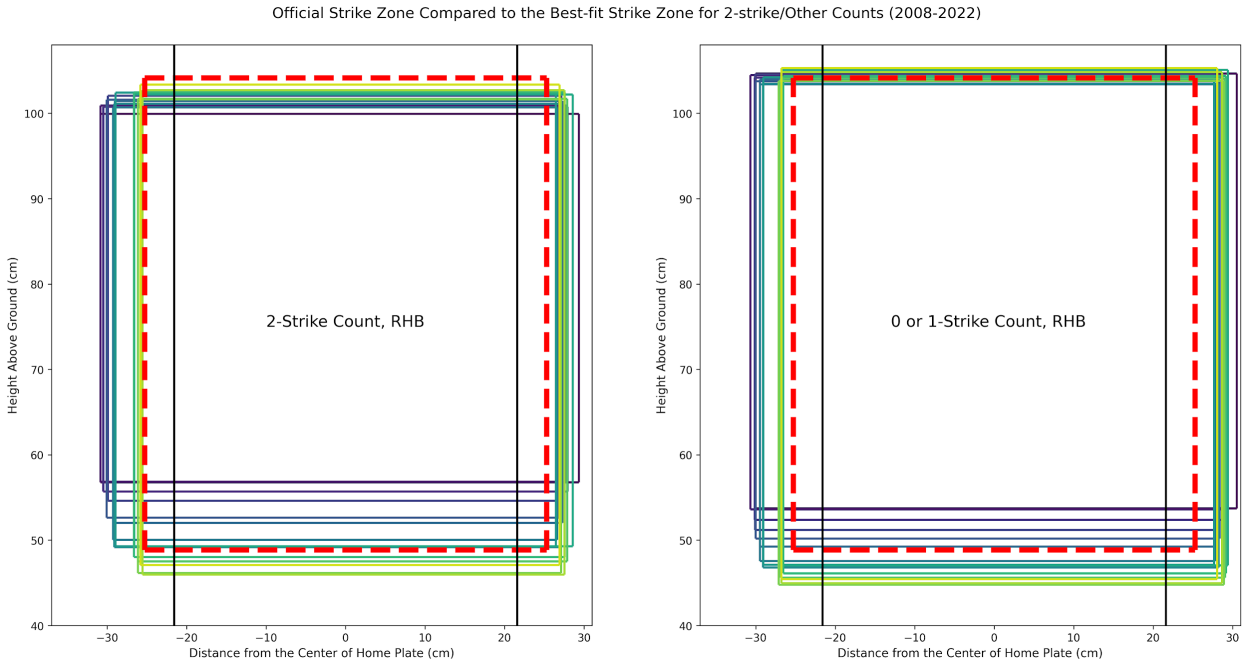


Figure 16: Although the 2-strike bias is robust statistically, the effect-size is not immediately obvious by eye without a comparison. The left panel shows the 2-strike best-fit strike zones for 2008-2022 and follows the same color scheme as Figure 9. The right panel shows the same, but for all other pitches, the 0 or 1-strike counts. Compared to the overall best-fit strike zone for a given season, the 2-strike zone is shorter top to bottom and slightly narrower.

The statistical methodology described in this section can be applied to other similar issues. As long as the question can be framed in terms of disjoint sub-samples in the data, we can test whether the samples are likely to be drawn from the same population or from some distinct subpopulation.

7.2. Pitch Framing

Existing pitch framing metrics focus on the number of called strikes in the Shadow Zone around the official zone. As we've learned, however, the official zone is not what's called during the game, so using the traditional pitch framing methodology means some of these additional strikes that are

being attributed to a catcher's pitch framing ability are probably attributable to the umpire's own zone.[5]

With our surface, we can measure the area of an umpire's strike zone when an individual catcher is behind the plate and compare that to the umpires' average strike zone. All else equal, we would expect to see catchers who are good at pitch framing with a larger area for their strike zone probability surface relative to the umpires' usual strike zone area. We call the difference between the area of the umpires' average strike zone and the strike zone for a particular catcher the catcher's Framing-Induced Strike Zone (FISZ). The FISZ is an improved measure of a catcher's framing ability because it is a more direct measurement of the influence of catchers on umpires' calls.

We measured the FISZ for each catcher for both RHB and LHB. Across all 7 of his seasons in the data set, Jose Molina has the largest FISZ compared to other catchers with 3+ seasons of catching. This holds true for both RHB and LHB. He catches a strike zone area of approximately 3,400 cm^2, more than 300 cm^2 larger than the average strike zone for the season. We see this in Figure 17, where the blue solid box representing Jose Molina's strike zone is larger than the box representing the 2012 strike zone for RHB. We would see similar charts for all of Molina' seasons across both RHB and LHB. Ryan Hanigan, Yasmani Grandal, and David Ross all break the Top 10 in FISZ to both RHB and LHB.

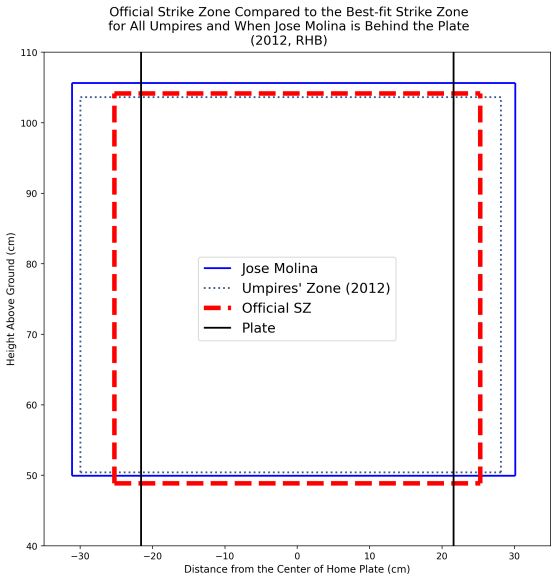


***Figure 17**:* The blue solid box represents the strike zone umpires called when Jose Molina was behind the plate for the 2012 season and a RHB was at the plate. The blue dotted box is what the umpires called with all catchers behind the plate. The larger box indicates Molina was more skilled at framing pitches to look like strikes than his peers.

7.3. Pitchers

We expect the size, shape, and position of the pitchers' best-fit strike zone surfaces to vary more widely than the umpires' because of their different pitch arsenals, throwing mechanics, and game situations. The pitchers' best-fit strike zones are normally distributed when throwing to both LHB and RHB. For pitchers with more than 500 pitches, we see a strike zone area that averages 3,100 cm^2, ranging from ±450 cm^2 larger or smaller than the umpires' zone for that season. (Note: ±450

[5] For modeling pitch framing, we assume the catcher that starts the game plays the entire game. This is a simplifying assumption that can be removed in subsequent iterations of the data set.

cm^2 is about 4 cm on each side of the strike zone rectangle.) Bronson Arroyo, Livan Hernandez, Derek Lowe, Ryan Vogelsong, and Jered Weaver stand out as pitchers consistently earning a strike zone larger than their peers season-to-season.

We can use the model to examine umpires' calls for some of the most elite pitchers of the past 15 seasons—Madison Bumgarner, Zach Greinke, Clayton Kershaw, Max Scherzer, Justin Verlander. These pitchers have long careers through the data set and are likely to have developed reputations for their pitch placement.

- Of this elite group, Bumgarner has the largest strike zone to both RHB and LHB at just under 3,300 cm^2. For both sets of batters, he benefits from getting the calls on 2 cm more of the outside edge of the plate to RHB as compared to his counterparts.
- Max Scherzer has the smallest strike zone of the group, primarily because he's getting fewer calls on the right edge of the plate—outside edge for RHB and inside edge for LHB. He benefits from calls on the left edge of the plate: The midpoint of Scherzer's strikezone averages is 6.4 cm towards the outside edge of the plate to LHB and 2.2 cm towards the inside of the plate to RHB. He's getting that lefty strike call more than his counterparts, even after 2017.

Overall, not surprisingly, the pitchers who we see getting a higher-than-average proportion of strike calls on the outside of the plate to LHB—Bronson Arroyo, Josh Beckett, Doug Fister, Roy Halladay, Jered Weaver—generally have the bulk of their careers in the pre-2017 part of our data set, when the 'lefty strike' was more apparent. We also see some elite closers—Mariano Rivera and Jonathan Papelbon—in this category.

7.4. The Batter's Eye

Joey Votto[6] is a player known for his batting eye, so let's take a look at how umpires call his plate appearances. Overall, we see a slightly smaller strike zone for Votto than the average LHB. A greater number of strike calls on the inside of the plate are being outweighed by fewer strike calls up in the zone. Joey Votto may be earning the umpires' benefit-of-the-doubt at the edges of the strike zone.

Baseball analysts have been marveling at Juan Soto's batting eye since he broke into the big leagues in 2018. Are umpires equally impressed? It doesn't look like it. Soto's average strike zone is more than 200 cm^2 larger than the average LHB for his playing years, and we see no evidence he is getting the "benefit of the doubt" on taken pitches. The differentiating dimension here is the height; Soto's strike zone is nearly 5 cm taller than his LHB contemporaries. This could be a result of Soto's wider/lower stance in his 2-strike approach.

Turn on any YES broadcast during the season and you will hear complaints about Aaron Judge suffering from a particularly low strike zone. Does the evidence bear this out? It does. The bottom of Aaron Judge's strike zone sits at 42.8 cm, nearly 3 cm lower than the best-fit strike zone for all RHB for his playing years. What we don't hear from YES broadcasters, however, is any complaints about the top of the strike zone. The top of Judge's strike zone is consistently 10 cm shorter than the average RHB, which means Aaron Judge has a smaller strike zone than the average RHB. Although

[6] Joey Votto's career and the dataset overlap well, making the model particularly suited to evaluating his performance.

we adjusted the model to account for Judge's height, his 6'7" frame probably distorts the umpire's perception from a crouched position.

7.5. Automated Pitch Calls

There's been a lot of debate about automated pitch calling, or the 'robot umpire.' Many argue an abrupt shift to robot umpires calling the exact MLB defined strike zone would drastically change the game. Our rectangular historical strike zone surface model offers a middle ground between forcing the rectangular 'by the book' strike zone and the existing method of noisy judgment calls. Using the best-fit strike zone from across the umpire corps as the target for a 'robot umpire' would preserve the zone as it's currently called while incorporating an automated system.

8. Conclusions

Overall, the new strike zone surface model improves our understanding of the behavior of umpires calling balls and strikes. By using an entire season to characterize an umpire's typical behavior rather than data from a single game, we're able to construct measures of umpiring variability and consistency across time (e.g. mean absolute error per game). The method's ability to more accurately assess smaller data subsets compared to past efforts means we can ask the model more nuanced, detailed questions while worrying less about sampling errors. Further, the inclusion of repeated sampling and fitting allows for testing of the many patterns and idiosyncrasies fans, players, and coaches experience in the game; some may hold up to scrutiny, while others may not.

References

[1] Arthur, Ron. "Baseball's New Pitch-Tracking System Is Just a Bit Outside." FiveThirtyEight, 28 Apr. 2017, fivethirtyeight.com/features/baseballs-new-pitch-tracking-system-is-just-a-bit-outside/.

[2] Bell, Tanner. "Smart Fantasy Baseball Tools: Smart Fantasy Baseball." Smart Fantasy Baseball - Fantasy Baseball Data, Projections, and Strategy, 2 Mar. 2022, www.smartfantasybaseball.com/tools.

[3] Ben-Porat, Eli. "Rethinking the Strike Zone: It's Not a Square." The Hardball Times, 19 February 2019, tht.fangraphs.com/rethinking-the-strike-zone-its-not-a-square/.

[4] Carruth, Matthew. "The Strike Zone," Lookout Landing, 29 October 2012, www.lookoutlanding.com/2012/ 10/29/3561060/the-strike-zone.

[5] Jedlovec, Ben. "Introducing Statcast 2020: Hawk Eye and Google Cloud." Major League Baseball, 20 July 2020, technology.mlblogs.com/introducing statcast-2020-hawk-eye-and-google-cloud a5f5c20321b8.

[6] Major League Baseball. Official Baseball Rules 2019 Edition, 2019.

[7] Roegele, Jon. "The Living Strike Zone." Baseball Prospectus, 24 July 2013, www.baseballprospectus.com/news/ article/21262/baseball-proguestus-the-living strike-zone/.

[8] Roegele, Jon. "The Strike Zone during the PITCHf/x era." The Hardball Times, 30 January 2014, – https://tht.fangraphs.com/the-strike-zone during-the-pitchf x-era/.

[9] Sullivan, Jeff. "The Changing Reality of the Lefty Strike." Fangraphs, 5 November 2013, – https://blogs.fangraphs.com/the-changing reality-of-the-lefty-strike/.

[10] Sullivan, Jeff. "Tom Glavine's Allegedly Generous Strike Zone." Fangraphs, 14 January 2014, https://blogs.fangraphs.com/tom-glavines allegedly-generous-strike-zone/.

[11] Umpire Scorecards. “Our Estimated Umpire Zone.” Umpire Scorecards, 2021, https://umpscorecards.com/explainers/estimated_umpire_zone.