The Influence of the Sinking Strike Zone on Major League Baseball's Strikeout Epidemic

Adam Brodie

Abstract

Background. Strikeout rates in Major League Baseball (MLB) are at an all time high and have been growing steadily over the last decade. Concurrently, the *called* strike zone has been evolving over the last nine seasons, expanding downwards. In June of this year (2016), in an effort to curb called strike zone expansion, an MLB competition committee agreed on a rule change that would raise the bottom of the strike zone beginning in the 2017 season. It is currently unknown what role, if any, the downward expansion of the called strike zone has had on the growing frequency of strikeout rates and consequently, how a rule change interrupting the expansion of the called strike zone might influence strikeout rates in the coming season.

Aim. In the present investigation we look to model how the downward expansion of the called strike zone has contributed to surging strikeout rates and predict what effect an intervention on the height of the bottom of the called strike zone might have on strikeout rates in the coming season.

Data. To estimate the influence of the downward expansion of the strike zone on strikeout rates we must first estimate the dimensions of the called strike zone. For this purpose we analyzed pitch tracking data recorded by Major League Baseball Advanced Media's (MLBAM) PITCHf/x system.

Methods. For the purposes of most accurately estimating the relationship between variability in the dimensions of the called strike zone and strikeout rates, we analyze game records at the scale of individual games, a novel scale for called strike zone estimation. This affords a large sample within each season, allowing us to distinguish variability in strikeout rates associated with variability in the dimensions of the called strike zone from variability in strikeout rates due to other factors changing at the season-to-season scale. We estimate the dimensions of the called strike zone by fitting a controlled-complexity classification model to classify regions of called strikes and balls for each game. We then perform a linear regression analysis to estimate the relationship between the estimated called strike zone dimensions and strikeout rates while controlling for confounding factors varying from one season to the next.

Results. We observe that the relationship between the bottom of the called strike zone and strikeout rates is, in fact, largely due to the influence of confounding factors varying from one season to the next as the regression coefficient is reduced from -0.005 to -0.002 when conditioning on season.

Conclusions. We conclude that the causal influence of the height of the bottom of the called strike zone on strikeout rates is small, on the order of an additional strikeout per hundred plate appearances per inch. This effect, itself, appears to be mediated by the area of the called strike zone which exhibits a significantly weaker dependence on the annual strikeout trend, and is nearly independent of the decline in contact rates. We conclude that a rule change inducing a smaller called strike zone will likely have a negligible influence on strikeout rates in the coming season, overshadowed by the unmeasured, but more prominent, factors driving the increase in strikeout rates.

Keywords: Baseball, Strikeouts Rates, Causal Inference, Mediation

DAP Committee:

Richard Scheines (DAP Advisor); Philosophy, Machine Learning (Affiliated Faculty)
Samuel Ventura; Statistics
Peter Spirtes; Philosophy, Machine Learning (Affiliated Faculty)

1 Introduction

Strikeouts in Major League Baseball (MLB) are occurring at unprecedented frequencies and have been continuing to occur more frequently over the past decade. Strikeouts are now occurring so frequently that many writers have begun to refer to the prevalence of strikeouts in today's game as an epidemic (Kurkijan, 2013; Baumbach, 2014; Carleton, 2014). In the midst of this epidemic, while the rulebook definition of the strike zone has been unchanged, pitch tracking records indicate that plate umpires have been exhibiting a growing tendency to call low pitches strikes, resulting in a downward expansion of the *called* strike zone. While many writers continue to monitor changes in the called strike zone and others speculate as to the factors driving surging strikeout rates, no analysis of the relationship between these two trends has been performed. In this work, we seek to provide such an analysis, investigating the causal influence the evolution of the called strike zone has had on Major League Baseball's strikeout epidemic.

2 Background and Related Work

The rules governing the game of baseball have changed little over the past century. The dimensions and manufacture of the game ball, the distance between the bases, between the pitcher's rubber and home plate, the rules for reaching base and scoring; all have remained unchanged. The style of gameplay in Major League Baseball (abbreviated "MLB"; North America's premier professional baseball organization), however, is subject to change from one generation to the next, passing through distinctive phases. Early in the history of the game, gameplay was dominated by sacrifice hits and stolen bases (Halfon, 2014). More recently, home runs were the most prominent event in the game (Vincent, 2007). Baseball's current phase, however, can only be characterized as the strikeout era. A new record for strikeout rates (strikeouts per plate-appearance, abbreviated "SOr") has been set every season since 2008. Since the beginning of the live-ball era in 1921,[1] a pitcher has struck out more than a quarter of the batters he faced over the course of a season 246 times.[2] 103 of those performances occurred between 2008 and 2016, and 58 in just the last three seasons. In the modern strikeout era, hits, walks, and outs on batted balls have all been displaced, to some degree or another, by the strikeout.

The last nine seasons also happen to constitute the interval for which we now have pitch tracking records; records of the velocity, trajectory, and outcome of nearly every pitch thrown in an MLB game. Over the last few seasons a number of baseball writers analyzing this data observed a trend concurrent with the climbing strikeout rates; despite no change in the rulebook definition of the strike zone (see Figure 2), plate umpires are exhibiting an increasing tendency to call low pitches strikes, resulting in a downward expansion of the *called* strike zone.

This phenomenon was first reported by Brian Mills (Mills, 2014) and Jon Roegele (Roegele, 2013) after the 2013 season. Mills used a generalized linear model on data accumulated over full seasons to estimate the dimensions of the called strike zones while Roegele, also using data accumulated over full seasons, binned pitches in inch-by-inch squares, considering those square-inch bins in which more pitches taken by batters were called strikes than called balls to constitute

[1] The *live-ball era*, often called the *modern era*, refers to the time interval spanning from 1921 to the present. In 1921, in response to the death of Ray Chapman after being struck by a pitch the previous season, new regulations requiring the use of fresh, easily visible balls throughout the game were instituted. This rule change had a significant influence on the style of gameplay and is therefore used as a demarcation criterion marking the beginning of modern baseball.

[2] We consider, here, pitcher-seasons of at least 100 innings pitched.

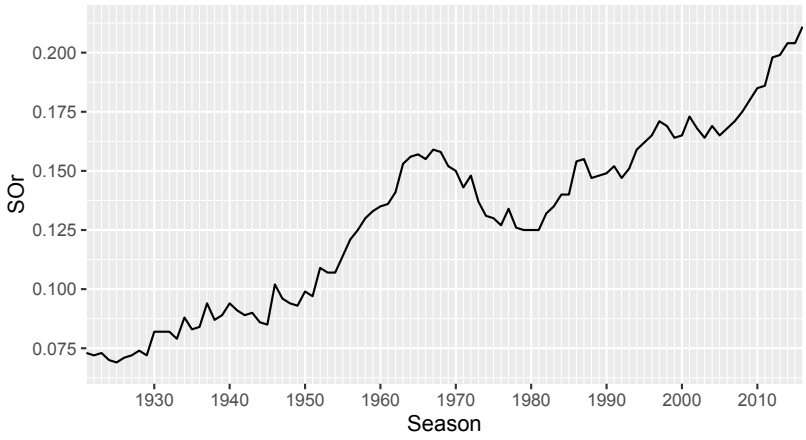

Figure 1: The evolution of strikeout rates during the live-ball era. New records for strikeout rates have been set in the last nine seasons.

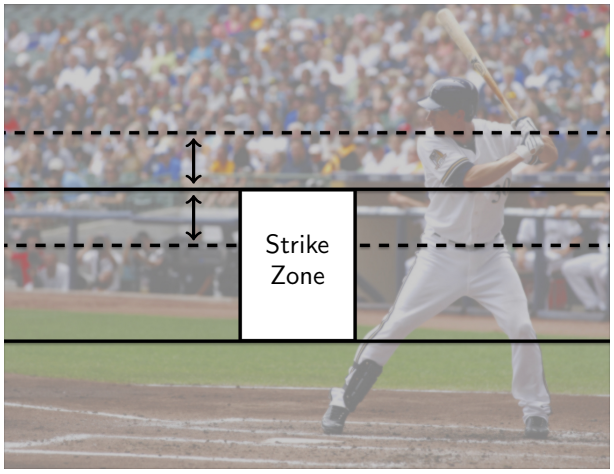

Figure 2: The MLB rulebook defined strike zone since the 1996 season (MLB, 2016a,b). The strike zone is defined as the region over home plate bounded below by the horizontal plane passing through the hollow below the batter's knee caps as he assumes his batting stance and above by the horizontal plane halfway between the top of his shoulders and the top of his pants.

the called strike zone. An illustration of the sinking called strike zone using Roegele's method is presented in Figure 3.

The downward expansion of the called strike zone continued through the following seasons with Roegele and others continuing to monitor the trend, producing headlines such as *The Strike Zone's Still Dropping* and *The Strike Zone Expansion is Out of Control* (Sullivan, 2014; Gaines, 2014; Roegele, 2014, 2015a,b; Speier, 2015). The evolution of the called strike zone ultimately instigated league action. Shortly after the beginning of the 2016 season it was reported that an MLB competition committee agreed to a rule change which would raise the rulebook definition

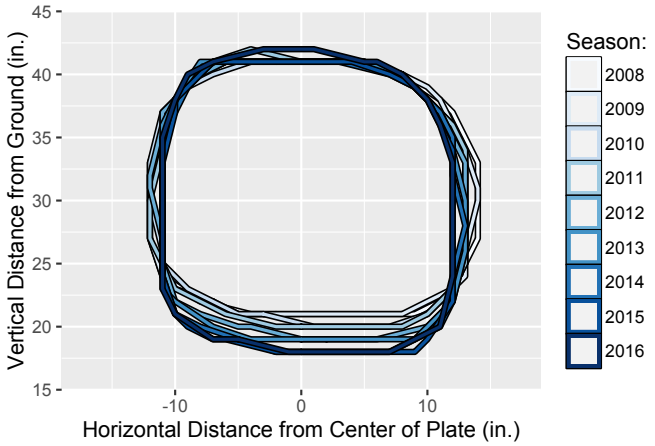

Figure 3: Boundaries of the called strike zones for the 2008 through 2016 seasons estimated using Jon Roegele's method laid over one another. A steady trend of a downward expansion in the called strike zone concurrent with a slower retraction of the sides of the called strike zone is easily visible.

of the bottom of the strike zone (Stark, 2016).

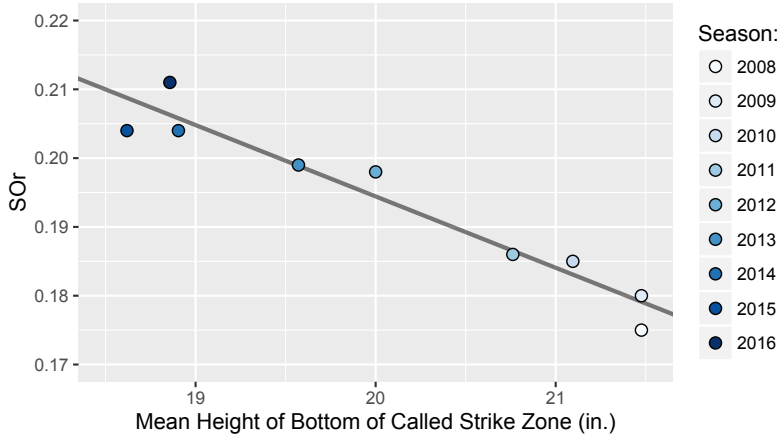

Figure 4: The mean height of the bottom of the called strike zone estimated using Jon Roegele's method plotted against strikeout rates ($\beta = -0.01$, $\rho = -0.97$). It is prima facie quite plausible that the downward expansion of the strike zone has had a causal influence on growing strikeout rates.

An impending change to the way balls and strikes are called in the midst of an epidemic of strikeouts leads us to ask how the evolution of the called strike zone may have contributed to growing strikeout rates over the last nine seasons and what, if any, change in strikeout rates we could expect to see in the coming season with a potentially significant adjustment to the called strike zone. While strikeout rates have received much attention over the last several years, little

progress has been made in explaining the causal influences driving their change. In an article from May of 2014, Russell Carleton enumerated a number of hypothesized factors contributing to the swelling frequency of strikeouts; batters tending to take more pitches, batters swinging and missing more frequently, pitchers becoming more effective at finishing batters in two-strike counts (Carleton, 2014). Carleton fails to find satisfaction in any of these explanations.[3] Next to nothing has been done in investigating the role of changes in the called strike zone in the strikeout epidemic. Cliff Corcoran speculates that the downward expansion of the called strike zone has influenced strikeout rates and anticipates a decline in strikeouts following next year's potential rule change (Corcoran, 2016). Corcoran's claims, however, are based on common intuition and baseball knowledge, citing the unique difficulties of hitting pitches that appear to be approaching the region over which the called strike zone has expanded, and are not informed by any statistical analysis of game records.

3 Approach

There is a clear relationship between the called strike zone and strikeout rates at the scale of events aggregated over the course of whole seasons (see Figure 4). It is not necessarily the case, however, that this is due to an immediate causal relationship; the relationship between the bottom of the called strike zone and strikeout rates may instead be the effect of confounding factors, with the two both evolving over time independently. In order to investigate this question, we need to examine the phenomenon at a finer scale, particularly, at the scale of individual games, to see how variability in the called strike zone within a single season relates to variability in strikeout rates. This is an intuitively reasonable scale since a plate umpire's called strike zone is ideally uniform over the course of a single game, while the existence of variability from one game to the next or from one umpire to another is common knowledge. Examining the dimensions of the called strike zone and strikeout rates at the scale of individual games affords us many instances within each season, allowing us to separate the influence of the called strike zone on strikeout rates from the influences of factors that change more slowly from one season to the next, factors including the distributions over fastball velocities of pitchers in the league and years of Major League experience among batters in the league. Assuming that variability in such factors is independent of the dimensions of the called strike zone within a season we may conclude that any residual dependence between called strike zone dimensions and strikeout rates is due to a causal influence.

Unfortunately, methods previously used to estimate the called strike zone using pitch tracking data aggregated over full seasons are impractical for the comparatively sparse sample sizes at the scale of individual games. Instead, we apply novel strategies for called strike zone estimation, applying low-complexity classification algorithms to estimate a boundary separating called strikes from balls. After acquiring estimates of called strike zone dimensions for every game within our sample, we perform a regression analysis to determine how variability in the dimensions of the called strike zone has influenced variability in strikeout rates while controlling for season, a surrogate for slow-changing confounding factors.

[3] It may be noted that batters' tendency to take more pitches has reversed since Carleton's article was published with strikeouts continuing to occur more frequently.

4 Data

In order to investigate the influence of the expanding called strike zone on strikeout rates we analyze data provided through Major League Baseball Advanced Media (MLBAM). Chief among this data is pitch tracking data recorded by the PITCHf/x system conveying the approximate location relative to the ground that pitches thrown during Major League Baseball games crossed home plate. PITCHf/x uses patented technology developed and maintained by Sportvision, a California-based sports broadcast effects company (White and Alt, 2008). The PITCHf/x system consists of a pair of cameras installed in each Major League ballpark (one in the stands above home plate and another behind first base) and accompanied by a collection of software used to process the photographic data. When a pitch is thrown, the cameras asynchronously record approximately 20 images each, as the pitch travels from the pitcher's hand toward the catcher. The software then processes these images, locates the ball, and derives the ball's real-world position at the time of each photograph based on its position in each image. The mapping from image coordinates to real-world coordinates is computed for each camera as a function of images of markers at known real-world positions in the ballpark and a series of camera parameters such as tilt, pan, etc. (Tsai, 1987). From the sequence of positions of a single pitch derived from these images, the PITCHf/x software then fits a quadratic trajectory to the pitch (Pendleton, 2011), estimates its location and velocity fifty feet from and as it crosses the front of the plate, and the pitch's deflection from a linear path and it's rotation direction and rate. Along with the pitch trajectory data, an estimate of the top and bottom of the strike zone (as described in the MLB rulebook; see Figure 2) is recorded for each batter by the PITCHf/x operator using video from the ballpark's centerfield camera. Finally, an MLBAM employee records additional data annotating the pitch, recording the outcome of the pitch (ball in the dirt, called strike, foul tip, ball put in play, etc.) and the outcome of the at bat (groundout shortstop to first base, single, strikeout, intentional walk, etc.). Additional software operated by MLBAM classifies the pitch type (fastball, sinker, curveball, etc.) for each pitch.

For the purposes of our analysis, we aggregated pitch tracking data from regular season games occurring during the 2008 through 2016 MLB seasons. The data is stored in a relational database publicly available through the MLB (MLBAM, 2016). We scraped and formatted the data automatically using the pitchRx package for R (Sievert, 2014). In total, we processed data from 21,819 games (there were 46 regular season games during this interval for which pitch tracking data was unavailable or corrupted; the identity of games or pitches for which data is missing may be reasonably assumed to be independent of the quantities of interest).

#	gameday_link	num	des	px	pz	sz_top	sz_bot	stand	event
1	gid_2016_04_03_chnmlb_anamlb_1	1	Called Strike	-0.88	2.27	3.61	1.93	L	Strikeout
2	gid_2016_04_03_chnmlb_anamlb_1	1	Ball	-0.70	0.48	3.53	1.69	L	Strikeout
3	gid_2016_04_03_chnmlb_anamlb_1	1	Called Strike	-0.07	2.16	3.65	1.69	L	Strikeout
4	gid_2016_04_03_chnmlb_anamlb_1	1	Swinging Strike	0.21	1.61	3.53	1.69	L	Strikeout
5	gid_2016_04_03_chnmlb_anamlb_1	2	Foul	-0.20	3.01	3.46	1.60	L	Walk
6	gid_2016_04_03_chnmlb_anamlb_1	2	Called Strike	-0.08	1.63	3.46	1.60	L	Walk
7	gid_2016_04_03_chnmlb_anamlb_1	2	Ball	0.22	0.91	3.60	1.60	L	Walk
8	gid_2016_04_03_chnmlb_anamlb_1	2	Ball	0.24	1.21	3.69	1.75	L	Walk
9	gid_2016_04_03_chnmlb_anamlb_1	2	Ball	-0.26	0.86	3.60	1.60	L	Walk
10	gid_2016_04_03_chnmlb_anamlb_1	2	Ball	0.37	1.28	3.60	1.60	L	Walk

Table 1: A sample of PITCHf/x data from the Cubs-Angels game played on April 4 of the 2016 season, restricted to features relevant to our analysis.

A sample of the PITCHf/x data features that we processed is presented in Table 1. Each row of the data corresponds to a single pitch thrown during an MLB game. The first feature, `gameday_link`, is a game id. Every game is assigned a unique game ID in the MLBAM repository

and we use this value to distinguish games for the purpose of computing estimates of the called strike zone and strikeout rates for individual games. The date of the game and home and away teams can be parsed from this id. `num` identifies at bats within individual games; 1 is assigned to the first batter, 2, the second, and so on. This is used primarily for merging the data structures containing the pitch tracking data and game-event annotations. `des` describes the outcome of an individual pitch. This feature is used for identifying called pitches (called strike, ball, etc.) which are used in estimating the called strike zone. `px` and `pz` indicate the lateral and vertical position of the pitch, respectively, as the ball crosses the front of the plate. Both values are given in feet. `px` is centered on the center of the plate with the value increasing from left to right from the plate umpire's or catcher's perspective and `pz` is simply feet from the ground (negative values correspond to the projected height at which pitches that land in front of the plate would have crossed the plate had they been unimpeded by the ground). `sz_top` and `sz_bot` indicate the top and bottom of the rulebook-defined strike zone (in feet from the ground) for the current batter as recorded by the PITCHf/x operator according to the batter's stance as the pitch is released. We use `sz_bot`, in particular, to estimate a called strike zone relativized to the height of the batter's knees. `stand` indicates the handedness of the batter. We use this feature to reflect the lateral position of pitches thrown to left-handed batters about the origin, so that all lateral pitch coordinates may be considered uniformly in terms of the "inside" (towards the batter) and "outside" (away from the batter) parts of the plate without explicitly including batter-handedness in our model, which would increase its complexity. Examination of the called strike zone for left-handed and right-handed batters shows that the called strike zones for batters of either handedness are approximately reflections of one another (Roegele, 2013), so this manipulation is reasonable. Finally, `event` describes the outcome of the at bat (values for `event` include, for example, strikeout, walk, groundout, and home run).

5 Methods

The problem of estimating the influence of the evolution of the called strike zone on strikeout rates can be decomposed into two parts: first, we have to process the pitch tracking records to determine estimates of the called strike zone dimensions for each game and compute strikeout rates, and second, we must analyze the relationship between these quantities. In the present section we describe our method for estimating the called strike zone and discuss our modeling choices for our regression analysis.

Methods for estimating the called strike zone using data accumulated over large portions of a single season are no longer feasible when we wish to estimate the called strike zone for a single game. Over our sample, approximately 150 pitches are taken per game, and only about a third of those for strikes. Neither a generalized linear model making use of features such as batter handedness and dimensions nor a pitch-binning technique works reliably on sample sizes this small with the former technique tending to overfit due to the sparsity of the features and the latter failing due to the sparsity of pitch positions recorded during a single game. Instead, we consider two controlled-complexity classification models well-suited to data of this kind. First, we consider axis-aligned rectangles, and second, support vector machines.

An axis-aligned rectangle classifier is a classifier whose decision boundary is simply a rectangle with edges parallel to the axes. An axis-aligned rectangle model of the called strike zone is the, defined as a quadruple, $(\texttt{inside}, \texttt{outside}, \texttt{top}, \texttt{bottom})$ with a pitch crossing the plate at

coordinates (x, z) classified accordingly;

$$f_{\text{Rect}}(x, z) = \begin{cases} \text{Strike} & \text{if } \texttt{inside} \leq x \leq \texttt{outside} \ \& \ \texttt{bottom} \leq z \leq \texttt{top} \\ \text{Ball} & \text{Otherwise} \end{cases}$$

where here we consider x, the horizontal position of the ball to be in our batter-handedness-relative coordinates. To fit the model we simply find some values of the parameters `inside`, `outside`, `top`, and `bottom` that minimize classification error.

Axis-aligned rectangles constitute a very natural model for the called strike zone. For one, for any given batter, the front surface of the strike zone, is by rule, a rectangle; the inside and outside boundaries of the zone are perpendicular to the ground and the top and bottom boundaries of the zone are parallel to it. It is reasonable to assume that a plate umpire's pitch-calling tendencies should follow this model fairly faithfully with no need for additional complexity. We should also have little fear of overfitting, as axis-aligned rectangles constitute a very restrictive class of classifiers; axis-aligned rectangles have a Vapnik-Chervonenkis dimension of 4 (Bishop, 2006). Finally, solving for parameters of the model is computationally inexpensive using a heuristic greedy search over ordered called strike coordinates (the cost is quadratic in the number of called pitches; note that the brute force examination of all called strike coordinates is quintic in this quantity). This greedy strategy, detailed below, outperforms depth-restricted decision-trees at a comparable cost (results not shown).

Algorithm: Greedy Axis-Aligned Rectangle Algorithm

Data: `Strikes, Balls`
Result: `inside`, `outside`, `top`, `bottom`
Initialize the boundaries; set `inside` and `outside` to the median `Strikes` x coordinate and `top` and `bottom` to the median `Strikes` z coordinate;
repeat
  For each boundary, compute the loss for the next most extremal `Strikes` coordinate value while holding other boundaries fixed;
  If the loss is improved, assign the new value to the boundary and update the loss;
until All `Strikes` coordinates are tested;

Support vector machines (SVMs) are also a natural candidate (Schölkopf and Smola, 2002). One intuitive means of adapting the Roegele-style pitch-binning called strike zone estimation strategy to the sparse pitch samples of single games is to consider some smoothing of the pitch locations. Rather than let the boundary by influenced by every data point, however, which would be sensitive to the concentration of pitches on either side of a suitable boundary, we should prefer to define a boundary as a function of only that subset of the data points that fall closest to data carrying the contrary label (these points are the *support vectors*). This is precisely what the SVM model does. The SVM classifies in a manner similar to a linear classifier, taking sign of the inner product between a weight vector and a data vector to determine a classification value, but is distinguished by the conventional use of a reproducing kernel Hilbert space projection of the data and its definition as a regularized *hinge loss* minimizer. The SVM loss function is defined as,

$$\ell = \frac{1}{2}||\omega||^2 + \frac{C}{n}\sum_{i=1}^{n}\Big[1 - \big(\omega^{\intercal}\phi(x_i) + b\big)y_i\Big]_+$$

where C is a complexity controlling parameter, ω is the weight vector, b an offset parameter, ϕ the kernel feature function, and x_i and y_i are feature and label values for the ith datum (for

this binary classification task $y_i \in \{-1, 1\}$; -1 may be considered to correspond to a called ball and 1 a called strike). The solution to the regularized hinge loss minimization maximizes the margin (in the separable case, the minimum distance between a training point and a point in the contrary label region) subject to a constraint on complexity;

$$\omega^* = \arg\min_{\omega} \frac{1}{2}||\omega||^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

subject to

$$\begin{aligned}\big(\omega^\intercal \phi(x_i) + b\big)y_i &\geq 1 - \xi_i \\ \xi_i &\geq 0\end{aligned}$$

By Lagrangian optimization the dual solution can be expanded as

$$\arg\max_{(\alpha_i)_{i=1}^n} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\phi(x_i)\phi(x_j)y_iy_j$$

which is simply a quadratic programming problem exploiting a kernel product that can be solved using an off-the-shelf quadratic program solver, with b then identified by enforcing satisfaction of our constraints. The resulting classifier is

$$f_{\mathsf{SVM}}(x, z) = \begin{cases}\mathsf{Strike} & \text{if } \omega^{*\intercal}\phi\big([x \quad z]^\intercal\big) + b \geq 0 \\ \mathsf{Ball} & \text{Otherwise}\end{cases}$$

The SVM model is a suitable candidate for modeling the called strike zone because of its flexibility as a nonparametric classifier, the ease with which its complexity can be controlled, and its computational feasibility. In our processing we use the squared exponential feature map with bandwidth and complexity parameters jointly optimized using ten-fold cross-validation. All SVM operations were computed using the kernlab package for R (Karatzoglou et al., 2004).

In addition to the choice of a classification model to represent the called strike zone, there are choices to be made concerning the parameterization of pitch locations. We have already mentioned that we reflect the lateral coordinates of pitches for left-handed batters in order to maintain as large a sample size as possible while allowing for feasible variability in lateral boundaries of the called strike zone. In addition, however, there is a question concerning the representation of the vertical location of each pitch. The rulebook definition of the strike zone is defined in terms of the dimensions of the batter as he assumes his hitting stance (see Figure 2). For this reason it is reasonable to adjust the height of each pitch relative to the batter dimensions. This allows us to avoid the explicit use of batter dimensions in our model which would increase the complexity of our hypothesis space. In order to maintain interpretability of the area of the called strike zone, we consider relativizing the height of the pitch to the bottom of the strike zone for each batter as it is recorded in the PITCHf/x data. Our choice of relativizing with respect to the bottom of the zone rather than the top is motivated by our interest in changes at the bottom of the called strike zone.

Altogether we have four proposed models of the called strike zone, using either the axis-aligned rectangle or SVM classifier and height adjusted or unadjusted pitch location parameterizations. The consistency of these models on our sample is presented in Figure 6. It is apparent from the illustration that the axis-aligned rectangle fits more consistent called strike zones than

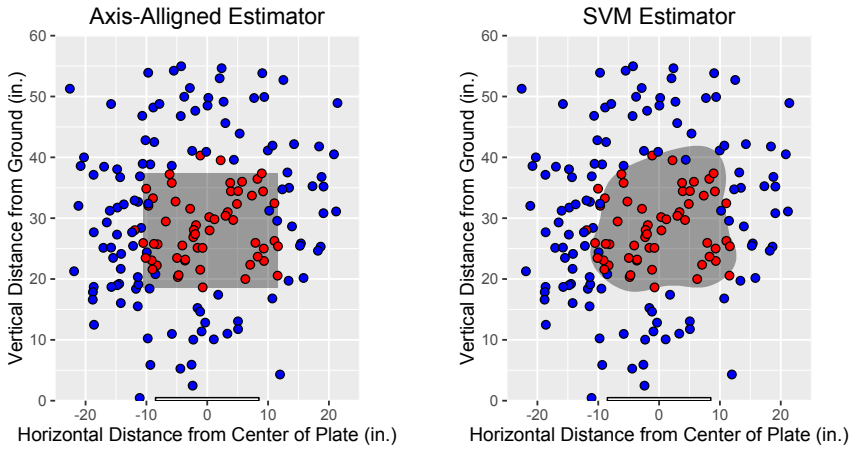

Figure 5: Example of a called strike zone estimated using an axis-aligned rectangle and using an SVM. The called strike zone is represented as the dark grey region. Red points correspond to the locations of pitches that were called strikes and blue points to the locations of pitches that were called balls.

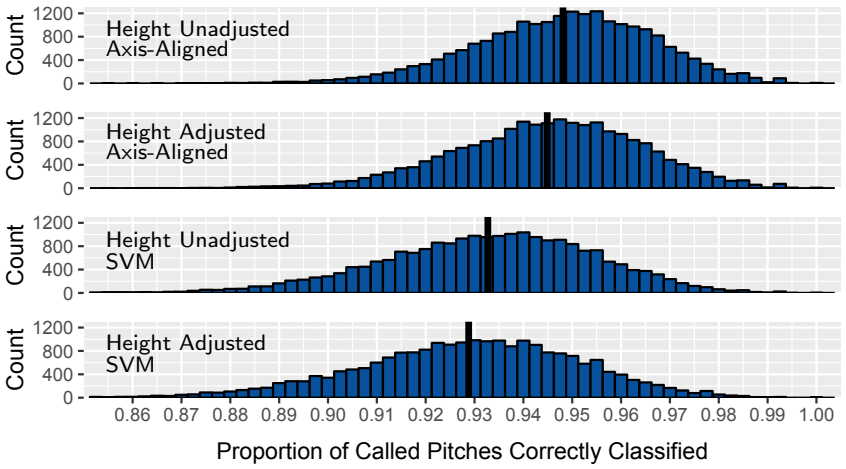

Figure 6: A histogram of empirical classification accuracy for our four candidate called strike zone estimation models. The black vertical bars indicate the mean empirical classification accuracies.

the regularized SVM and that a more consistent called strike zone can be fit to pitch locations with unadjusted rather than adjusted height coordinates. The regularized SVM model is more flexible in the shape it permits the called strike zone to take, but is penalized on sharp corners, which occur frequently in the data. Further, even when the regularized SVM model does separate called balls and strikes effectively, there are games for which the decision boundary "leaks" outside the appropriate region. This occurs when there is a region along the boundary of the strike zone where no called pitches occur. The axis-aligned rectangle is immune to these sorts of

errors due to the fact that its simple boundaries are less sensitive to local patterns in the data.

The fact that adjusting the height of the pitches degraded classification consistency was somewhat surprising, but is plausibly the consequence of greater variability induced at the top boundary of the called strike zone or a genuine bias in plate umpires towards a uniform called strike zone fit to a "typical" batting stance. We should note in support of this latter point that batters often change their stance slightly during the course of a single at bat in response to a change in count or base-runner configuration. It is plausible that umpires are not adjusting their strike-calling behavior in response to these changes in batting stances. Despite the variability in the consistency of these called strike zone models, their relations to strikeout rates are qualitatively similar. For this reason we report results pertaining only to called strike zone estimates acquired using the axis-aligned rectangle model on height-unadjusted pitch location data.

The next step of our analysis is to examine the relationship between called strike zone dimensions and strikeout rates. Our primary interest is in the height of the bottom of the called strike zone but we examine the area of the front surface of the called strike zone, as well. Further, we also look at some auxiliary features such as the proportion of pitches thrown below batters' knees during a game and swing-and-miss rates. We use linear regressions to analyze the relationships among these variables since the linear model is easily interpretable and the data relationships are all, qualitatively, virtually linear. We choose to use linear regression rather than a binomial logistic regression for the dependent variable, strikeout rates, for two reasons. First, for the support of our sample, the relationships with the rates, themselves, are approximately linear; we preserve the interpretability of the regression coefficients and the fit by forgoing the logit transformation. Second, the binomial logistic regression minimizes squared error over individual samples within each count statistic; this would imply weighting the error for games with more at bats more heavily than for games with fewer at bats. Due to the nature of baseball, there is a non-negligible dependence between strikeout rates and the number of at bats in a game, with games with higher strikeout rates typically seeing fewer batters come to bat ($\rho = -0.18$). We weight all rates in our sample equally to avoid biasing our estimates. Finally, we construct a structural equation model to analyze the plausibility of our causal hypothesis (Figure 11).

6 Results

A scatterplot illustrating the relationship between the estimated height of the bottom of the called strike zone and strikeout rates in presented in Figure 7. Notable is the prominent shift in the height of the bottom of the strike zone over the nine seasons in our sample, visible in the color gradient of our point cloud from the right to left side of the plot. Also notable is the fact that the regression coefficient relating these quantities is significantly mitigated relative to the season-aggregated model (Figure 4); the regression coefficient at the season-aggregated scale is -0.01 while the coefficient at the game scale is less than half that. This is the consequence of a high degree of variability in the height of the bottom of the called strike zone from game to game unrelated to strikeout rates.

To visualize the conditional distributions controlling for season, we present a contour plot of kernel density estimates for the conditional densities in Figure 8. Note that the dependence between the height of the bottom of the called strike zone and strikeout rates is again mitigated when we consider each season's data in isolation, much of the overall dependence induced by shifts along both dimensions from one season to the next. Regression and Pearson correlation coefficients detailing this phenomenon are provided in Table 2.

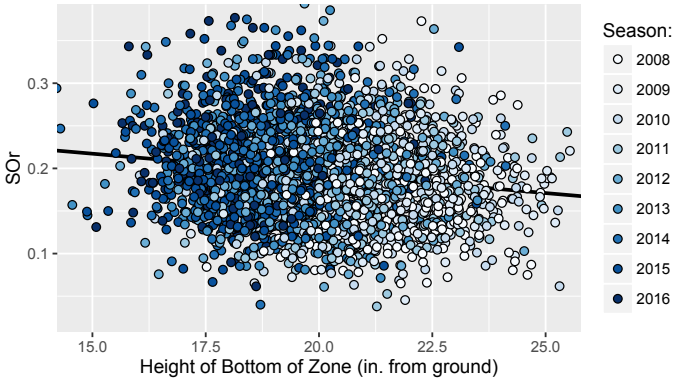

Figure 7: Strikeout rates plotted against the height of the bottom of the called strike zone estimated using axis-aligned rectangles parameterized using our greedy fitting algorithm ($\beta = -0.0046$, $\rho = -0.1538$). A random 3000 game sub-sample is displayed.

Figure 8: The .90 contour lines for kernel density estimates of the joint distribution of the bottom of the called strike zone and strikeout rates by season. The dependency between the height of the bottom of the strike zone and strikeout rates within each season can be seen to be qualitatively less pronounced than the marginal dependency.

Season	β	ρ	p-value
2008	0.0004	0.0121	0.723
2009	−0.0007	−0.0213	0.148
2010	−0.0017	−0.0516	0.006
2011	−0.0028	−0.0836	0.000
2012	−0.0036	−0.1046	0.000
2013	−0.0026	−0.0729	0.000
2014	−0.0018	−0.0474	0.010
2015	−0.0014	−0.0314	0.042
2016	−0.0023	−0.0566	0.003
'08 -'16	−0.0046	−0.1538	0.000

Table 2: Regression and Pearson correlation coefficients for the relationship between the height of the bottom of the strike zone and strikeout rates by season. p-values are for a one-sided t-test with null hypothesis $\beta \geq 0$.

Season	β	ρ	p-value
2008	9.08×10^{-5}	0.1051	0.000
2009	10.65×10^{-5}	0.1226	0.000
2010	12.20×10^{-5}	0.1364	0.000
2011	9.60×10^{-5}	0.1093	0.000
2012	15.37×10^{-5}	0.1746	0.000
2013	9.48×10^{-5}	0.1069	0.000
2014	8.91×10^{-5}	0.0964	0.000
2015	13.32×10^{-5}	0.1411	0.000
2016	12.96×10^{-5}	0.1340	0.000
'08 -'16	14.23×10^{-5}	0.1574	0.000

Table 3: Regression and Pearson correlation coefficients for the relationship between the area of the front surface of the called strike zone and strikeout rates by season. p-values are for a one-sided t-test with null hypothesis $\beta \leq 0$.

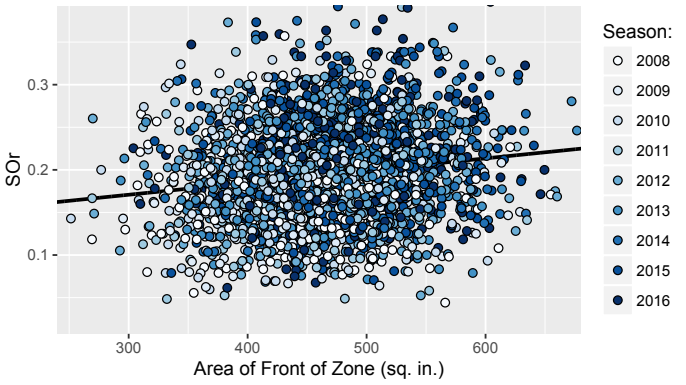

Figure 9: Strikeout rates plotted against the area of the front surface of the called strike zone estimated using axis-aligned rectangles parameterized using our greedy fitting algorithm. A random 3000 game sub-sample is displayed.

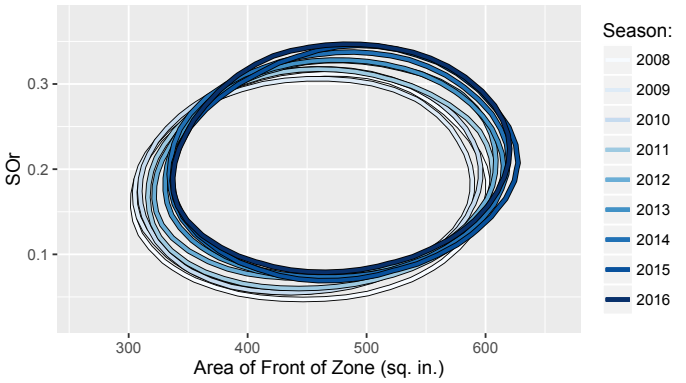

Figure 10: The 0.90 contour lines for kernel density estimates of the joint distribution of the area of the front surface of the called strike zone and strikeout rates by season. Strikeout rates exhibit a significantly more stable relationship from one season to the next with the area of the called strike zone than with the height of the bottom of the called strike zone.

A more robust relationship with strikeout rates is exhibited by the area of the front surface of the called strike zone. A scatterplot is presented in Figure 9 and a contour plot of the season-conditional densities is presented in Figure 10. It is immediately apparent that the changes in the distribution over the area of the front surface of the called strike zone from one season to the next are significantly less pronounced than those changes in the distribution over the height of the bottom of the called strike zone. This is seen quantitatively in the regression and Pearson correlation coefficients collected in Table 3.

Next, we examine how these quantities may be causally related. Table 4 collects the regression and Pearson correlation coefficients relating the height of the bottom of the called strike zone and strikeout rates conditioning on the area of the front surface of the called strike zone. The only season for which this relationship is significant is 2011 (a one-sided t-test for the vanishing regression coefficient rejects at $\alpha = .05$). Overall, the relationship between the height of the bottom of the called strike zone and strikeout rates conditioning on both season and the area of the front surface of the called strike zone is not significant ($\beta = -0.0001$, $p = 0.74$; $\rho = -0.0022$).

This suggests that the influence of the height of the bottom of the called strike zone on strikeout rates is mediated exclusively by the overall area of the strike zone and confounding factors varying with season. A structural equation model representing this hypothesis is depicted in Figure 11. A structural equation model is a probabilistic model representing each variable as a function of its immediate causes and an independent noise term (Bollen, 1989). Goodness-of-fit metrics for structural equation models are often used to evaluate the plausibility of a

Season	β	ρ	p-value
2008	0.0020	0.0570	1.000
2009	0.0008	0.0249	0.980
2010	0.0000	0.0008	0.682
2011	−0.0015	−0.0450	0.099
2012	−0.0010	−0.0297	0.041
2013	−0.0011	−0.0289	0.373
2014	−0.0004	−0.0110	0.703
2015	0.0008	0.0207	0.887
2016	−0.0003	−0.0065	0.470
'08-'16	−0.0026	−0.0883	0.000

Table 4: Regression and partial correlation coefficients for the relationship between the height of the bottom of the called strike zone and the residuals of strikeout rates regressed on area of the front surface of the called strike zone by season. p-values are for a one-sided t-test with null hypothesis $\beta \geq 0$.

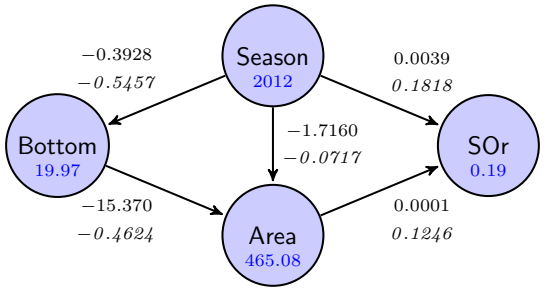

Figure 11: A structural equation model illustrating a proposed causal structure relating the height of the bottom of the called strike zone to strikeout rates fit using the TETRAD causal and structural equation modeling software (Scheines et al., 1998). Offsets are printed in blue, within the nodes. In italics are the coefficients for the standardized model. The model has a single degree of freedom and a χ^2 statistic of 0.1389 ($p = 0.7094$).

hypothesized structure of causal relationships. In our model, all functional relations are linear and noise terms are assumed to be Gaussian. The model has a single degree of freedom and a χ^2 statistic of 0.1389 ($p = 0.7094$), suggesting that this structure is a highly plausible explanation of the relationship between the dimensions of the called strike zone and strikeout rates during the strikeout epidemic. Along with the data-scaled edge coefficients, edge coefficients for the normalized data are indicated in italics. This allows us to easily compare the relative influences of confounders and called strike zone dimensions on strikeout rates in the model. Note that more than three fifths of the correlation between the height of the bottom of the called strike zone and strikeout rates is attributed to confounding factors.

7 Discussion and Conclusion

From our analysis we may conclude that the downward expansion of the called strike zone has had little to do with the strikeout epidemic. Using our structural equation model from Figure 11 we may predict the consequences of an intervention on the height of the bottom of the called strike zone in anticipation of a rule change for the 2017 season using causal modeling techniques (Spirtes et al., 2000; Pearl, 2009). According to our model, if a rule change were to raise the average height of the bottom of called strike zone for 2017 to any height lower than 20.4″, we will still see a new record for strikeout rates in the 2017 season. Despite the fact that the area of the called strike zone would be expected to be smaller than it has been at any point during the last five years in this scenario, the factors contributing to strikeout rates independent of the area of the called strike zone are predicted to continue to grow, resulting in a net increase in strikeout rate. Figure 12 depicts the predicted average strikeout rates per game for the 2017 season as a function of the new height of the bottom of the called strike zone. Note that we predict average strikeout rates per game rather than season averages since our model is fit to rates normalized by the number of at bats for individual games. There is, however, a very robust linear relationship between average strikeout rates per game and season strikeout rates; a predicted season strikeout rate may be recovered by scaling the average strikeout rate per game by 0.994.

Figure 12: Average strikeout rate (per game) for 2017 as a function of average height of the bottom of the called strike zone predicted using our structural equation model. Averages for 2008 through 2016 are plotted for reference. Strikeout rates are predicted to reach a new high in 2017 for any adjustment in the bottom of the called zone less than $1.62''$.

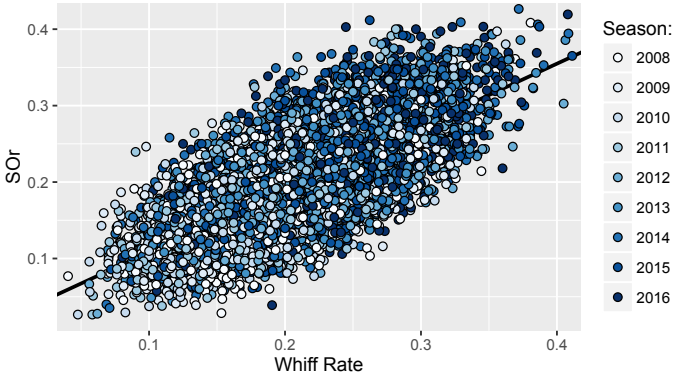

Figure 13: Strikeout rates plotted against swing-and-miss (whiff) rates. While whiffs are obviously a strong predictor of strikeouts, the robustness of this relationship is still surprising ($\beta = 0.8216$, $\rho = 0.7077$).

It is hard to say how confident we should be in this prediction. The model achieves an impressive fit for the data in our sample, but there are many unmeasured factors that may have exhibited stable trends throughout this period that may change dramatically. We saw this happen with home run numbers over the past couple of seasons, for instance; after several years of home runs trending downward they returned in dramatic fashion in the second half of the 2015 season and throughout the 2016 season. Here the question concerns the primary factors driving surging strikeout rates. We don't believe that the downward expansion of the called strike zone has been one of them, but we don't know precisely what they are, either, and therefore our prediction that they should continue to trend in the same fashion next year should be taken with a grain of salt.

A little more can be said in support of our hypothesis that the downward expansion of the called strike zone has not been a primary factor in the strikeout epidemic. In examining the pitch tracking records we see that the primary mechanism by which strikeout rates are rising is through a decline in the rate at which batters are successfully making contact on swings. A scatterplot of swing-and-miss (or *whiff*) rates, swing-and-misses per swing, is presented in Figure 13. Whiff rates by season are collected in Table 5. What's notable is that the height of the bottom of the called strike zone is very weakly related to whiff rates over the last few seasons (Table 6).

The weak relationship between the height of the bottom of the called strike zone and whiff rates over the last few seasons immediately challenges Cliff Corcoran's argument for the influence of the downward expansion of the called strike zone on strikeout rates due to the difficulties

Season	Whiff Rate
2008	0.1908
2009	0.1928
2010	0.1968
2011	0.1973
2012	0.2077
2013	0.2093
2014	0.2109
2015	0.2161
2016	0.2226
'08-'16	0.2051

Table 5: Swing-and-miss (whiff) rates by season. The primary means by which strikeouts are becoming increasingly frequent is through the decline in the rate at which batters successfully make contact on swings.

Season	β	ρ	p-value
2008	0.0004	0.0125	0.730
2009	−0.0008	−0.0265	0.099
2010	−0.0017	−0.0596	0.002
2011	−0.0029	−0.1025	0.000
2012	−0.0017	−0.0570	0.003
2013	−0.0007	−0.0242	0.117
2014	−0.0003	−0.0091	0.327
2015	0.0000	0.0011	0.522
2016	−0.0005	−0.0158	0.218
'08-'16	−0.0037	−0.1423	0.000

Table 6: Regression and Pearson correlation coefficients for the relationship between the height of the bottom of the called strike zone and whiff rates by season. p-values are for a one-sided t-test with null hypothesis $\beta \geq 0$.

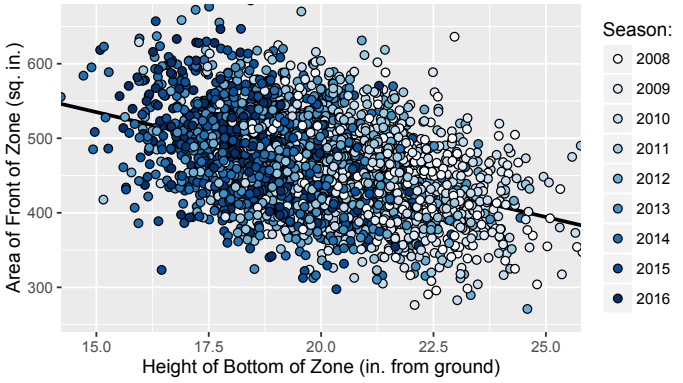

Figure 14: The area of the from of the called strike zone plotted against the height of the bottom of the called strike zone ($\beta = -14.07$, $\rho = -0.4233$). A random 3000 game sub-sample is displayed.

associated with hitting low pitches. It certainly may be the case that batters are struggling to make contact with modern pitchers' low-zone pitches, but in recent years whiffs are increasingly common regardless of how low pitches are being called.

Something should also be said about the mediation of the influence of the height of the bottom of the called strike zone on strikeout rates by the area of the front of the called strike zone. This finding may appear trivial due to the logical relationship between the height of the bottom of the called strike zone and its area, but this not the case. First, there is a large amount variability in the area of the called strike zone independent of the height of the bottom of the strike zone due to variability in the other boundaries (see Figure 14). While the downward expansion of the called strike zone constitutes a robust trend over the last nine seasons, the overall expansion of the called strike zone has been notably lesser due to this variability and the concurrent retraction of the sides of the called strike zone partially mitigating the effect of the expansion at the bottom (see Figure 3). While the area of the called strike zone does appear to be a relevant factor contributing to strikeout rates, in general, it does not account for a large proportion of the variability shared between season and strikeout rates.

Over the last nine seasons, strikeout rates have surged. Simultaneously, the called strike zone has evolved, expanding downward. Upon modeling the relationship between the dimensions of the called strike zone and strikeout rates we conclude that the downward expansion of the called

strike zone has not been one of the primary factors driving the strikeout epidemic. In the coming year we may see a rule change, possibly raising the height of the bottom of the called strike zone by more than an inch. We predict that despite such a change reducing the size of the called strike zone, the strikeout epidemic will persist.

8 Acknowledgements

I'd like to thank participants of the Cascadia Symposium on Statistics in Sports (Vancouver, BC; September 24, 2016) where I first presented this material, and the symposium organizers Luke Bornn and Tim Swartz (Statistics, Simon Fraser University). I'd also like to thank the Tartan Sports Analytics Club at Carnegie Mellon University for allowing me to workshop this material, as well, and Richard Scheines and Peter Spirtes (Philosophy, Carnegie Mellon University) for their support as I pursued this project in parallel with my dissertation research. I also want to recognize Roy Maxion (Machine Learning, Carnegie Mellon University) for his instruction and guidance in my preparation of this project to meet the Machine Learning Department's Data Analysis Project requirements.

Finally, I'd like to give special thanks to Sam Ventura (Statistics, Carnegie Mellon University) for his guidance on this project from the beginning, and his support and encouragement as I've pursued involvement in sports analytics.

References

Baumbach, J. (2014). Striking out has become an epidemic, even for non-sluggers. *Newsday;* `http://www.newsday.com/sports/baseball/striking-out-has-become-an-epidemic-even-for-non-sluggers-1.7911756`.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

Carleton, R. (2014). There's gotta be a reason for the strikeout epidemic. . . right? *FOX Sports.* `http://www.foxsports.com/mlb/story/there-s-gotta-be-a-reason-for-the-strikeout-epidemic-right-052714`.

Corcoran, C. (2016). Pending rule change could fix MLB's increasingly high strikeout rate. *Sports Illustrated.* `http://www.si.com/mlb/2016/05/21/mlb-rule-changes-strikeouts-strike-zone`.

Gaines, C. (2014). What an mlb strike zone really looks like and why players are always so mad about it. *Business Insider.* `http://www.businessinsider.com/mlb-strike-zone-2014-9`.

Halfon, M. S. (2014). *Tales from the Deadball Era: Ty Cobb, Home Run Baker, Shoeless Joe Jackson, and the Wildest Times in Baseball History*. Potomac Books.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kurkijan, T. (2013). MLB's universal chorus: strike three! *ESPN;* `http://www.espn.com/mlb/story/_/id/9404316/mlb-players-striking-record-pace`.

Mills, B. M. (2014). Expert workers, performance standards, and on-the-job training: Evaluating major league baseball umpires. Available at SSRN: http://ssrn.com/abstract=2478447.

MLB (2016a). *Official Baseball Rules*. Office of the Commisioner of Baseball, 2016 edition.

MLB (2016b). The strike zone: A historical timeline. http://mlb.mlb.com/mlb/official_info/umpires/strike_zone.jsp.

MLBAM (2016). Repository. http://gd2.mlb.com/components/game/mlb/.

Moyer, S. (2014). Baseball's 'shift': Does it work? *The Wallstreet Journal*. *http://www.wsj.com/articles/baseballs-shift-does-it-work-1410304648*.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, second edition.

Pendleton, R. (2011). PITCHf/x nine parameter trajectory model fitting. Technical report, Sportvision, LLC.

Roegele, J. (2013). The strike zone during the PITCHf/x era. In *The Hardball Times Baseball Annual 2014*. FanGraphs.

Roegele, J. (2014). The strike zone expansion is out of control. *The Hardball Times*. http://www.hardballtimes.com/the-strike-zone-expansion-is-out-of-control/.

Roegele, J. (2015a). The 2015 strike zone. *The Hardball Times*. http://www.hardballtimes.com/the-2015-strike-zone/.

Roegele, J. (2015b). The expanded strike zone: It's baaaack.... *The Hardball Times*. http://www.hardballtimes.com/the-expanded-strike-zone-its-baaaack/.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. S. (1998). The TETRAD Project: Constraint Based Aids to Causal Model Specification. *Multivariate Behavioral Research*, 33(1):65–117.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Sievert, C. (2014). Taming PITCHf/x data with XML2R and pitchRx. *A peer-reviewed, open-access publication of the R Foundation for Statistical Computing*, page 5.

Speier, A. (2015). Baseball's strike zone has expanded, and hitters aren't happy. *The Boston Globe*. https://www.bostonglobe.com/sports/2015/07/16/baseball-strike-zone-expands-offense-shrinking/FenP9Yj0MLEgBlsELMDCfM/story .html.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. The MIT Press, Cambridge, MA, second edition.

Stark, J. (2016). Sources: Competition committee agrees to change strike zone, intentional walks. *ESPN*. http://www.espn.com/mlb/story/_/id/15633876/mlb-competition-committee-agrees-changes-strike-zone-intentional- walks.

Sullivan, J. (2014). The strike zone's still dropping. *FanGraphs*. http://www.fangraphs.com/blogs/the-strike-zones-still-dropping/.

Tsai, R. (1987). A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal on Robotics and Automation*, 3(4):323–344.

Vincent, D. (2007). *Home Run: The Definitive History of Baseball's Ultimate Weapon*. Potomac Books.

White, M. and Alt, A. (2008). Tracking an object with multiple asynchronous cameras. US Patent App. 11/688,149.

9 Appendix I: Limitations

There are a few limitations to the conclusions of this project that should be reviewed. First, as we discussed in section 3, the interpretation of our model requires that we consider season as a reasonable surrogate for confounding factors changing much more slowly than called strike zone dimensions. These factors may include a number of relevant factors that have evolved over the past nine seasons, factors such as fastball velocity and spin rate distributions, the proportion of batters with "all-or-nothing" hitting strategies, and the frequencies with which batters face new pitchers for the first time, to name just a few. In order for this interpretation to be reasonable it must be the case that these factors are independent of called strike zone dimensions. If, for instance, plate umpires had a tendency to call a more restrictive strike zone for pitchers with a stronger propensity for striking out batters then this strategy would fail to properly separate the influence of the dimensions of the called strike zone from influences of these other factors. An informal analysis of called strike zone dimensions for individual starting pitchers suggests this is highly unlikely, but the assumption should be emphasized.

Second, we note again that the predictions based on our structural equation model incorporate the assumption that certain trends exhibited over the last nine seasons will persist. This includes the assumption that the unmeasured factors that we consider to be the primary drivers of the strikeout epidemic continue to evolve monotonically. Presumably this trend will not persist indefinitely, but since the study of these features is not the primary purpose of this project we take a conservative position, assuming that trends that have been apparently stable over the last nine seasons will continue on this trajectory in the coming year. Another trend incorporated in the model is the retraction of the called strike zone along borders other than its bottom. The tightening of the sides of the called strike zone has been a firm trend, as well, but this appears to be the product of feedback from pitch tracking reports and has resulted in more accurate lateral boundaries of the called strike zone. At some point we should expect the accuracy of the sides of the called strike zone to converge to an optimum and see this neutralizing influence of season on called strike zone area diminish. This is of less concern, however, due to the relatively small magnitude of this effect. Finally, in the 2016 season a new phenomenon concerning the dimensions of the called strike zone was observed; the top of the called strike zone expanded upwards, at some locations by more than an inch. The top of the called strike zone has been its most stable boundary over our sample and therefore this phenomenon is not a strong contributor to our model fit. However, if this trend persists or is exaggerated in the coming season we should expect even less of an influence from raising the bottom of the called strike zone in mitigating strikeout rates.

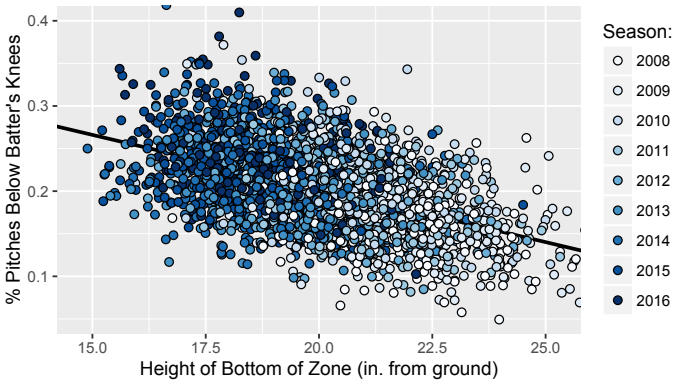

Figure 15: The proportion of pitches thrown below the batter's knees plotted against the height of the bottom of the called strike zone ($\beta = -0.0126$, $\rho = -0.4849$). A random 3000 game sub-sample is displayed.

10 Appendix II: Future Work

While we have focussed here on the influence (or absence thereof) of the downward expansion of the called strike zone on strikeout rates, there are related questions that could be investigated from the estimates of the called strike zone at the scale of individual games collected in this project. For one, despite the fact that the downward expansion of the called strike zone has had little influence on strikeout rates, it is possible that it has had an influence on other aspects of the game. Perhaps the downward expansion of the called strike zone contributed more directly to the decline in run scoring, for instance.

Another question that may arise in consideration of the phenomena described in this project concerns the origins of the downward expansion of the called strike zone. The trend has been enormously robust; even though the average height of the bottom of the called strike zone did not get any lower in the 2016 season than that of the 2015 season, the general nine year trend persists across both plate umpires and starting pitchers. While the retraction at the sides of the called strike zone may be attributed to umpires' responding to feedback from pitch tracking records, reducing the frequency of pitches called strikes off the plate (Mills, 2014), the downward expansion of the called strike zone has resulted in a larger proportion of pitches called strikes below the suggested lower boundary of the strike zone. So why was the called strike zone expanding downward?

A possible explanation may be seen in examination of pitch location over the last nine seasons. Figure 15 displays the proportion of pitches crossing the plate below the batter's knees per game plotted against the height of the bottom of the called strike zone. The dependence between the height of the bottom of the called strike zone and the frequency with which pitches are thrown below the batter's knees is stronger than we would have anticipated ($\rho = -0.4849$). It is obvious that pitchers would target the bottom of the called strike zone when it has been exaggerated, exploiting the availability of called strikes at an extreme region of the zone, but it is also possible that there is some influence in the opposite direction, as well. It is possible that pitchers have been targeting the bottom of the zone for other reasons, and that umpires have been adjusting their strike zone downward in response to the new distribution of pitch locations. Why would pitchers target the bottom of the strike zone? One plausible answer is to induce more ground balls. Batters are more likely to hit the ball on the ground on pitches that cross the plate at a lower location, with more of the top surface of the ball exposed to the batter's swing. Ground balls have become increasingly valuable with the growing implementation of defensive *shifts*. A shift involves an arrangement of fielders (typically infielders) in an unconventional configuration.

For batters who exhibit a strong tendency to pull the ball (for a right-handed batter to hit the ball to the left side of the field or for a left-handed batter, to the right side) the defensive team may shift the configuration of their infielders to that side of the field. The increased implementation of defensive shifts has been estimated to have significantly influenced offensive production, reducing league batting averages by several points (Moyer, 2014). Shifting the infielders is only effective at reducing the number of hits on the infield, however, predominantly on ground balls. We therefore hypothesize that the downward expansion of the called strike zone has been driven by a downward shift in the distribution of pitch locations which, in turn, has been the product of a strategy to maximize the effectiveness of defensive shifting. The investigation of this hypothesis would make for interesting future work, as well.