

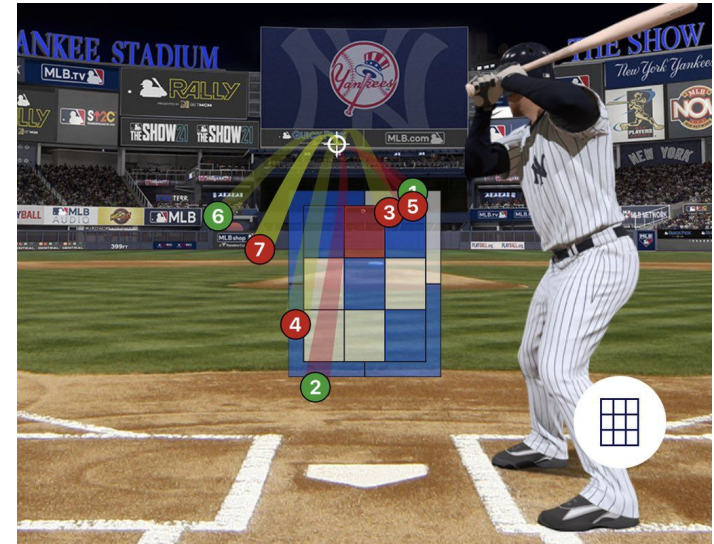
A discussion of:

*Outline Analyses of the Called Strike  
Zone in Major League Baseball*

Ron Yurko

March 31 2023

# What is the strike zone in baseball?



# Umpires call balls and strikes... unfortunately

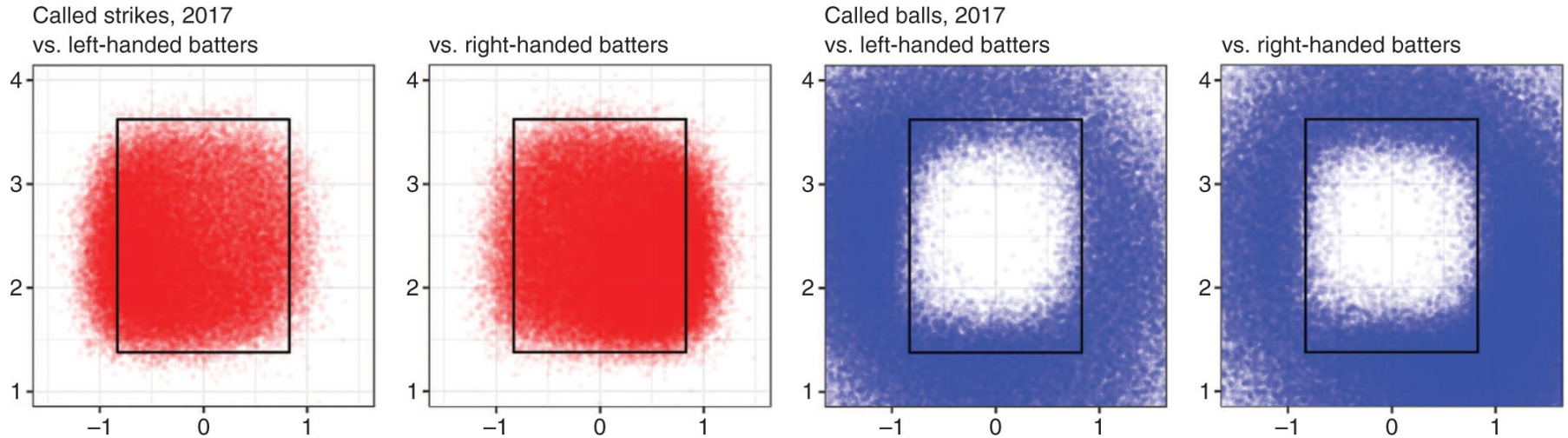


Figure 1 from <https://www.degruyter.com/document/doi/10.1515/jqas-2018-0061/html>

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

# Three stages of outline analyses




1. Use kernel discriminant analysis (KDA) to determine the outline of the called strike zone (CSZ)
2. Fit a model to coordinates of points sampled along outline, producing coefficients describing smoothed CSZs
3. Investigate relationships between outline coefficients and other variables such as year, along with player, game, and umpire factors

# Step 1: Use a model to determine the CSZ

- CSZ for subset of pitches (based on factors of interest like year, handedness) was set of points (x,y) where locally smooth estimate of density of called strikes exceeded similar estimate for called balls
  - No restrictions on shape, just need plenty of pitches (they have over 3.4 million in this paper)
  - Obtained points for outline using some digitization approach with `alphahull` package...
- Other approaches to do this:

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



  - Naive Bayes classifier  
(<https://www.degruyter.com/document/doi/10.1515/jqas-2018-0061/html>)
  - GAM predicting called strike  
(<https://baseballwithr.wordpress.com/2019/02/11/visualizing-the-actual-strike-zone/>)
  - Outline is just set of points corresponding to predicted probability of 0.5



Ball 1

Ball 2

Ball 3



## Step 2: Model the CSZ outline

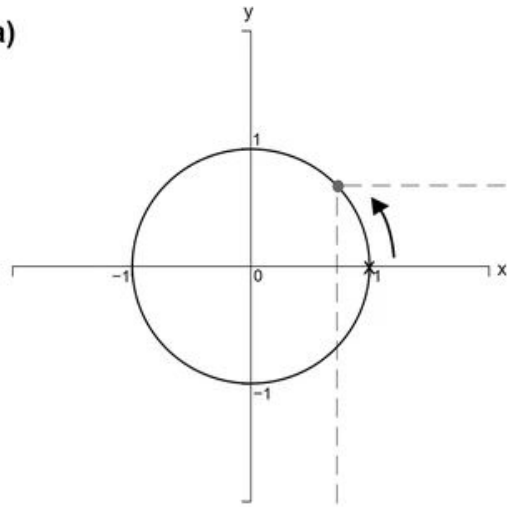
### Approach #1: Elliptic Fourier (EF) models

$$(4.1) \quad x(t) = a_0 + \sum_{n=1}^N \left( a_n \cos \frac{2\pi nt}{T} + b_n \sin \frac{2\pi nt}{T} \right) + \epsilon_x(t),$$

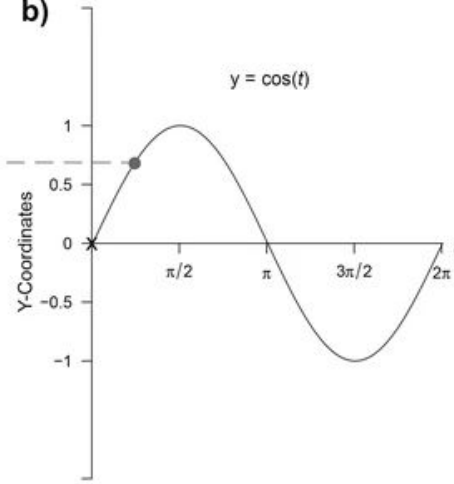
$$(4.2) \quad y(t) = c_0 + \sum_{n=1}^N \left( c_n \cos \frac{2\pi nt}{T} + d_n \sin \frac{2\pi nt}{T} \right) + \epsilon_y(t),$$

where  $N$  is a positive integer less than or equal to  $[I/2]$  (with  $[\cdot]$  being the greatest integer function), and  $\epsilon_x(t)$  and  $\epsilon_y(t)$  are independent Gaussian white noise processes. The leading coefficients,  $a_0$  and  $c_0$ , comprise the outline's centroid, and  $(a_n, b_n, c_n, d_n)$  is known as the  $n$ th *harmonic*. The locus of points  $(x, y)$  on the curve corresponding to the  $n$ th harmonic is an ellipse centered at the origin. Thus, the EF model describes the position of a point travelling (as  $t$  varies) around a series of  $N$  superimposed and successively smaller ellipses, as in a characterization of planetary orbits by Ptolemaic epicycles; for further details and a graphical illustration see [Kuhl and Giardina \(1982\)](#).

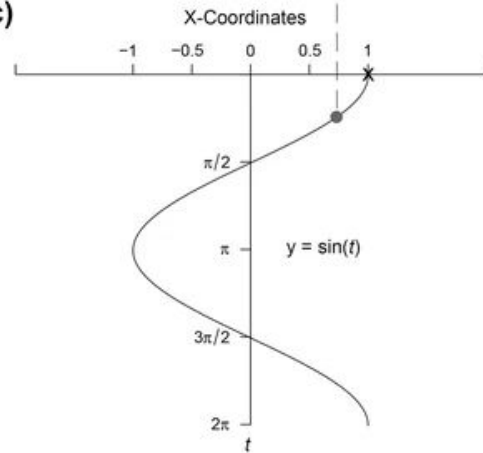
a)



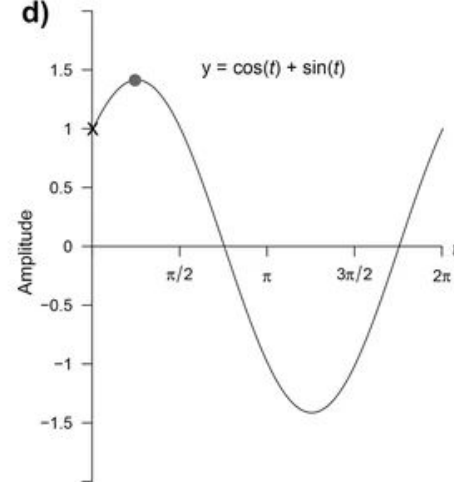
b)



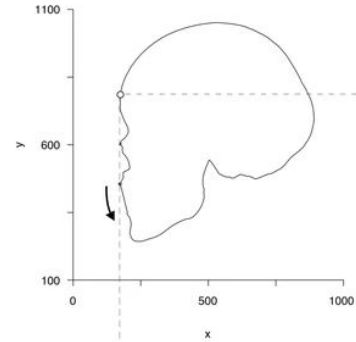
c)



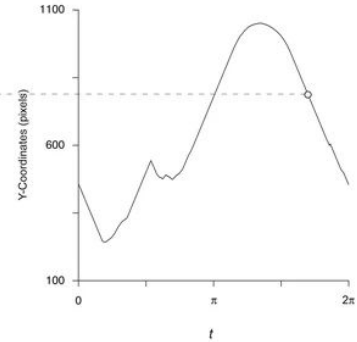
d)



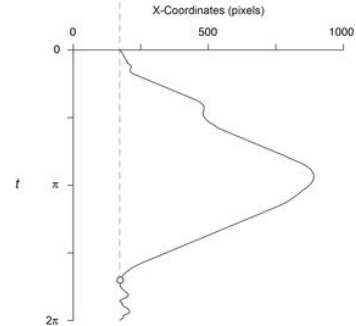
a) Original Shape



b) Y-Coordinates



c) X-Coordinates





# Step 2: Model the CSZ outline

## Approach #1: Elliptic Fourier (EF) models

major axis. The coefficients of the first harmonic may be used to normalize the higher-order coefficients to render them invariant to starting point, size and orientation using the following normalizing transformation (Kuhl and Giardina (1982)):

$$(4.3) \quad \begin{pmatrix} \hat{A}_n & \hat{B}_n \\ \hat{C}_n & \hat{D}_n \end{pmatrix} = \frac{1}{\hat{\lambda}} \begin{pmatrix} \cos \hat{\psi} & \sin \hat{\psi} \\ -\sin \hat{\psi} & \cos \hat{\psi} \end{pmatrix} \begin{pmatrix} \hat{a}_n & \hat{b}_n \\ \hat{c}_n & \hat{d}_n \end{pmatrix} \begin{pmatrix} \cos n\hat{\theta} & -\sin n\hat{\theta} \\ \sin n\hat{\theta} & \cos n\hat{\theta} \end{pmatrix}.$$

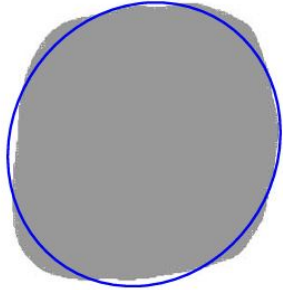
Here,  $\{\hat{A}_n, \hat{B}_n, \hat{C}_n, \hat{D}_n : 1 \leq n \leq N\}$  are the normalized coefficients,  $\hat{\theta} = (1/2) \arctan[2(\hat{a}_1\hat{b}_1 + \hat{c}_1\hat{d}_1)/(\hat{a}_1^2 + \hat{c}_1^2 - \hat{b}_1^2 - \hat{d}_1^2)]$ ,  $\hat{\psi} = \arctan(\hat{c}_1^*/\hat{a}_1^*)$  and  $\hat{\lambda} = (\hat{a}_1^{*2} + \hat{c}_1^{*2})^{1/2}$  where  $\hat{a}_1^* = \hat{a}_1 \cos \hat{\theta} + \hat{b}_1 \sin \hat{\theta}$  and  $\hat{c}_1^* = \hat{c}_1 \cos \hat{\theta} + \hat{d}_1 \sin \hat{\theta}$ . The nor-

$\hat{\psi}$ , and  $\hat{\lambda}$ ). In our context, however, because it is important to compare called strike zones to the RBSZ—which has a well-defined location, size and orientation in addition to a well-defined shape—we retained nearly all of this information. The only quantity we excluded was  $\hat{\theta}$ , which is of no interest. However, we replaced  $\hat{\lambda}$  with the more pertinent size variable  $\hat{\kappa} = \pi \hat{\lambda}^2 |\hat{D}_1|$  which is the area of the best-fitting ellipse. Therefore, our statistical analyses of fitted EF( $N$ ) CSZ outlines utilized the  $4N + 1$  coefficients  $\{\hat{a}_0, \hat{c}_0, \hat{\psi}, \hat{\kappa}, |\hat{D}_1|\} \cup \{\hat{A}_n, \hat{B}_n, \hat{C}_n, \hat{D}_n : 2 \leq n \leq N\}$ . The first five of these correspond directly to discernible geometric features of CSZs, but the remainder do not. R code for fitting the elliptic Fourier models is provided in Section B of the Supplementary Material (Zimmerman, Tang and Huang (2019)).

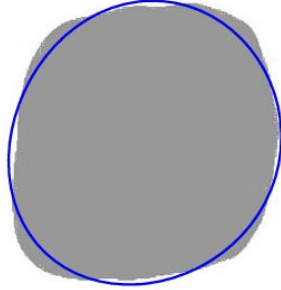
## Step 2: Model the CSZ outline

### Approach #1: Elliptic Fourier (EF) models

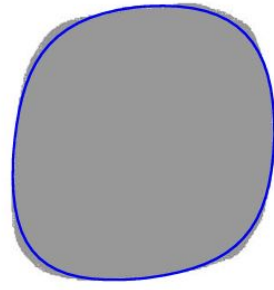
EF(1)



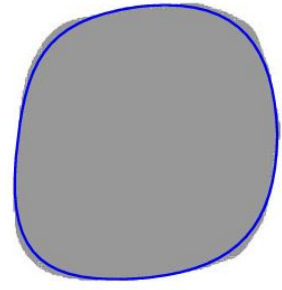
EF(2)



EF(3)



EF(4)



# Step 2: Model the CSZ outline

## Approach #2: Superelliptic models

features of CSZ outlines. We consider only superellipses that are aligned with the  $(x, y)$ -axes. Such an object is a set of points  $(x, y)$  that satisfy the equation

$$\left| \frac{x - x_0}{a} \right|^{2r} + \left| \frac{y - y_0}{b} \right|^{2r} = 1,$$

where  $(x_0, y_0)$  is the center,  $a$  is the half-width,  $b$  is the half-height and  $r > 0$  is the rectangularity index. The value of  $r$  strongly influences the superellipse's shape:  $r < 0.5$  yields a four-armed star with concave sides;  $r = 0.5$  corresponds to a rhombus; and  $r > 0.5$  yields a convex, bilaterally symmetric (with respect to each of the  $x$ - and  $y$ -axes) object, with  $r = 1$  corresponding to an ordinary ellipse.

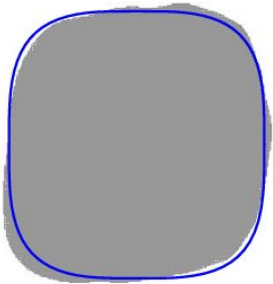
- *Affine-transformed latitudinally asymmetric superellipse (ATLAS)*

$$\begin{aligned} \left| \frac{(x - x_0) + s(y - y_0)}{a} \right|^{2r_1} + \left| \frac{y - y_0}{b} \right|^{2r_1} &= 1 \quad \text{if } y \geq y_0, \\ \left| \frac{(x - x_0) + s(y - y_0)}{a} \right|^{2r_2} + \left| \frac{y - y_0}{b} \right|^{2r_2} &= 1 \\ \text{if } y < y_0, r_1 > 0, r_2 > 0, -\infty < s < \infty. \end{aligned}$$

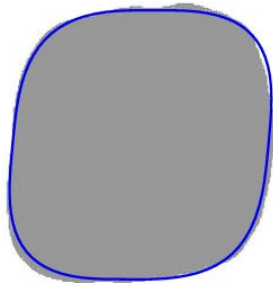
## Step 2: Model the CSZ outline

### Approach #2: Superelliptic models

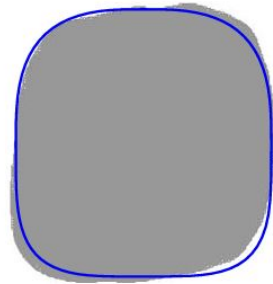
**SE**



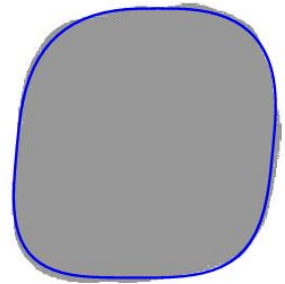
**ATS**



**LAS**



**ATLAS**



## Step 3: Analysis of outlines

### Variability within and between umpires

Let  $y_{ijk}$  denote an arbitrary fitted coefficient from the ATLAS representation of the CSZ outline corresponding to the  $i$ th level of batter handedness,  $j$ th umpire, and  $k$ th replicate. The following two-factor crossed mixed effects model with interaction was fit to each such coefficient:

$$(7.1) \quad y_{ijk} = \mu + \beta_i + u_j + (\beta u)_{ij} + e_{ijk}.$$

Here,  $\mu$  is an overall fixed effect,  $\beta_i$  is the fixed effect of the  $i$ th level of batter handedness,  $u_j$  is the random effect of umpire  $j$ ,  $(\beta u)_{ij}$  is the random interaction effect between the  $i$ th level of batter handedness and umpire  $j$  and  $e_{ijk}$  is the random effect of the  $k$ th replicate. We assumed that the  $u_j$ s,  $(\beta u)_{ij}$ s, and the  $e_{ijk}$ s are mutually independent random variables with  $N(0, \sigma_u^2)$ ,  $N(0, \sigma_{\beta u}^2)$  and  $N(0, \sigma_e^2)$  distributions, respectively. Note that we regard the effects of umpires

# Step 3: Analysis of outlines

## Variability within and between umpires

TABLE 6

*Minimum variance quadratic unbiased estimates ( $\times 10^6$ ) of variance components in the mixed effects model (7.1) for selected ATLAS coefficients (plus area and eccentricity) and the corresponding point estimate and 99% upper confidence interval estimate of the proportion  $\gamma$  of variability attributable to umpires*

ATLAS coefficient	$\hat{\sigma}_u^2$	$\hat{\sigma}_{\beta u}^2$	$\hat{\sigma}_e^2$	$\hat{\gamma}$	Confidence interval
$x_0$	459	1238	987	0.731	(0.634, $\infty$ )
$y_0$	199	360	945	0.372	(0.207, $\infty$ )
$a$	1301	549	987	0.652	(0.526, $\infty$ )
$b$	1122	209	996	0.572	(0.424, $\infty$ )
$A$	14,830	10,510	13,500	0.652	(0.530, $\infty$ )
$E$	2961	824	1920	0.663	(0.538, $\infty$ )