Take Me Out to (Analyze) the Ballgame

Chris Comiskey

February 14, 2017

Chapter 1

1 Introduction

Graphical displays are an irreplaceable tool for the statistician, and the statistical community. They allow us to quickly communicate information about a data set to another statistician, or, perhaps even more importantly, to non-statisticians. As technology generates more data the importance of this area of statistics expands, statistical analysis becomes more widespread, and our graphic-making abilities improve. The R data visualization package `ggplot2` highlights (i) the importance of graphical diplays, because it is among the most downloaded R packages; and (ii) the need for innovation in this area, because its popularity explosion signals it met a need.

In this chapter we focus on one type of graphical display, the heat map. We show the way it fails to adequately communicate spatial data dispersion and distribution attributes, and we innovate a solution. This innovation improves our ability to communicate spatial data density and dispersion attributes to the viewer using a heat map.

2 Heat Maps

We start with a baseball heat map example, as baseball data motivates this research. Consider the empirical heat map in Figure 1, of the two dimensioned vertical face of the strike zone, where grid box colors represent empirical batting averages at pitch locations. In Figure 1 we show PITCHf/x® data on 1,932 right-handed hitters, taking 1,582,581 swings between 2008 and 2015. Let $b = 1, \ldots, 627$ index grid boxes. Let $i = 1, \ldots, 1,582,581$ index swings, and define $n_b = \sum_i \mathrm{I}_{\{i \in b\}}$ as the total number of swings in box b. Define

a Bernoulli random variable, S_i, that equals one for swing success and zero for swing failure, and let $\hat{p}_b = \frac{1}{n_b}\sum_i S_i \cdot \mathrm{I}_{\{i \in b\}}$ be the empirical box b success probability. Figure 1 displays the resulting empirical heat map for 1,932 right-handed hitters, taking 1,582,581 swings between 2008 and 2015. The heat map graphic maps the empirical success probability, $\hat{p}_b$, of hitters swinging at pitches that passed through the space represented by that grid box, to a color on a spectrum.

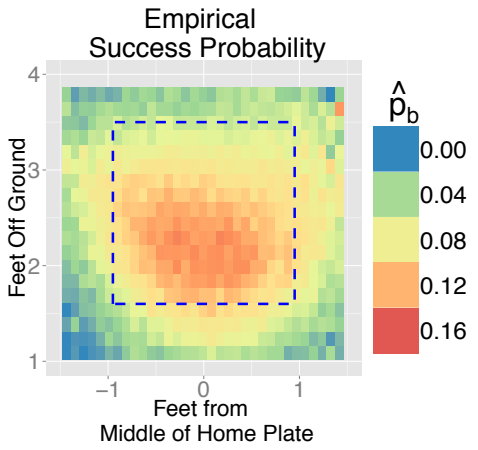

Figure 1: The gridded hitting zone with 3/4 inch by 3/4 inch boxes, from the catcher's point of view. The color of the box represents the empirical batting average ($\hat{p}_b$) for right handed hitters, swinging at pitches in that location. Calculations based on 1,932 right handed hitters, swinging at 1,582,581 pitches between 2008 and 2015. The dashed line marks the boundaries of the called strike zone for a 6'2" hitter.

While not sophisticated statistically, the graphic efficiently conveys empirical spatial success probabilities; it maps the statistic $\hat{p}_b$ to a color on a spectrum. Though it can easily go unnoticed, a heat map's creater *chooses* the grid box size that, hopefully, best communicates the data's information content; here, average spatial hitting ability estimates. The data's varying spatial density through the strike zone is not part of the communicated information, no matter the grid box size selected. This missing information is important, because in general more data means better estimates, and the viewer gets no such indication in Figure 1. Traditional heat maps do not communicate this information. We propose an innovation that addresses this shortcoming. To illustrate, consider a heat map for an individual hitter.

The heat map above divided the strike zone into relatively small boxes, because the data supported it. By "supported it" we mean the small, spatially specific boxes retain a sample size large enough to keep the variance of $\hat{p}_b$ acceptably [1] This is important because individual hitter datasets vary dramatically in size,

[1]Defining "acceptable" variance ranges and thresholds will depend on context and analysis objectives. For example, a pitching coach may be satisfied with estimates accurate 95% of the time to within 20 batting average points. This margin of error, 0.02, requires a sample size of 32 when $p_b = 0.09$. Note that the variance depends on the mean for a Bernoulli random

with swing totals ranging from a single swing to over 10,000 swings. Similar to bin width selection for a histogram, the choice of heat map resolution can dramatically affect how the data is represented, and the usefulness of the parameter estimates of interest. The resolution decision depends on the size and nature of the data set in question, and its spatial dispersion through the domain. To explore this decision in detail we look at batter 425509, a veteran player named Jhonny Peralta, whose data set contains 9,177 swings.

3 Empirical Heat Maps - Resolution Selection

The heat map in Figure 2 divides the central region of the strike zone into 16 equally sized boxes. Each box maps $\hat{p}_b$ to a color, and the box sample size, n_b, is printed on the box center. For convenient referencing in this chapter, we adopt the convention of numbering boxes with a horizontal component (left to right) and a vertical component (bottom to top [**Alix: "WHY?"**]). For example, we call the top-left box (1,4), and the bottom-right box (4,1).

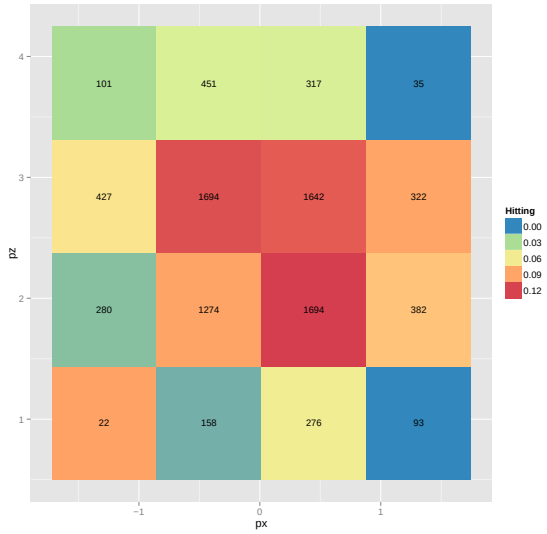

Figure 2: This four by four heat map conveys the empirical batting average of batter 425509, Johnny Peralta, in each of 16 square regions of the hitting zone. Each box maps $\hat{p}_b$ to a color. The number printed on each box represents the number of pitches the hitter swung at that passed through that box.

Peralta swang at 22 pitches in the region of space represented by box (1,1), and he swung at 1694 pitches in box (2,3). We interpret these two box sample sizes to contextualize. Three primary factors influence pitch location: pitcher game theoretic strategy, pitch-swing location margin of error (distance by which a pitch

variable.

misses its intended target), and the game state. Game theoretic strategy concerns the pitcher's knowledge of the hitter's strenths and weaknesses, and the hitter's reciprocal knowledge. Margin of error concerns the pitcher's usual outcome of not exactly hitting his target.[2] Game state characteristics include the at-bat count, the number of outs, and if runners occupy bases.[3] Peralta probably swung at only 22 pitches in box (1,1) because he did not see many pitches there. We can speculate this is because pitches there are mostly out of the strike zone, both in the horizontal and vertical directions. Therefore, it is less likely to induce a swing at a bad pitch to hit, and unlikely to be called a strike despite being out of the strike zone. With $n_{(1,1)} = 22$, the four by four resolution is sufficiently fine to present $\hat{p}_{(1,1)}$. Box (2,3), with $n_{(2,3)} = 1694$ pitches, can support more location specific, but still reliable estimates of p. This motivates finer resolution in that region of space. Peralta has relatively high success in Box (2,3), and he undoubtedly swings at as many pitches in that box as possible. The pitcher knows this, so will seldom aim there. However, by virtue of being closer to the center of the strike zone, this location collects more pitch location mistakes.

Because, as mentioned, a finer resolution is justified for box (2,3), we subdivide all boxes further. For simplicity, without implying this is the only or best way to increase resolution, we divide each box into four equally sized sub-boxes. Figure 3 shows the 16 by 16 result.

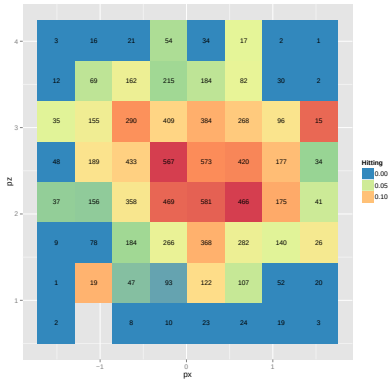

Figure 3: This 16 by 16 heat map conveys the empirical batting average of batter 425509, Johnny Peralta, in each of 64 square regions of the hitting zone. Each box maps $\hat{p}_b$ to a color. The number printed on each box gives the number of pitches the hitter swung at that passed through that box. A grey box indicates no pitches passed through that box.

Boxes (3,5), (3,6), (4,5), and (4,6)—the boxes created by dividing box (2,3) at the four by four resolution—

[2]If you visualize a 12 inch diameter archery target where the pitcher aims, he will usually hit the target somewhere, but not the bull's-eye.

[3]Two example game state pressures include the increased penalty for throwing a pitch outside the strike zone on a three ball count (the runner gets on base at four balls); the increased penalty for a hit with a runner in scoring position (runner on second or third base).

still contain sample sizes sufficient to support low variance p estimates. More generally, 24 boxes still have a sample size greater than 150; and 15 boxes still have a sample size of greater than 250. These boxes could support further subdivision. On the other hand, numerous boxes—corner and edge boxes in particular—now contain sample sizes generally insufficient to support low variance estimates of p_b. Twenty-nine boxes have a sample size of less than 50, and 17 boxes have a sample size of less than 20. At this resolution one box recorded zero swings.

In this way, due to the particular dispersion of the data, a heat map at any resolution will contain boxes of exceedingly small sample sizes (high variance), and/or boxes of unnecessarily large sample size (unnecessarily low variance). Figure 3 shows six different heat map resolutions, constructed with the same data from Peralta. We started with one box, and subdivided each box into four at each iteration. We chose this simple resolution increasing algorithm to illustrate the resolution selection challenge, and to provide a foundation for our innovation in the next section.

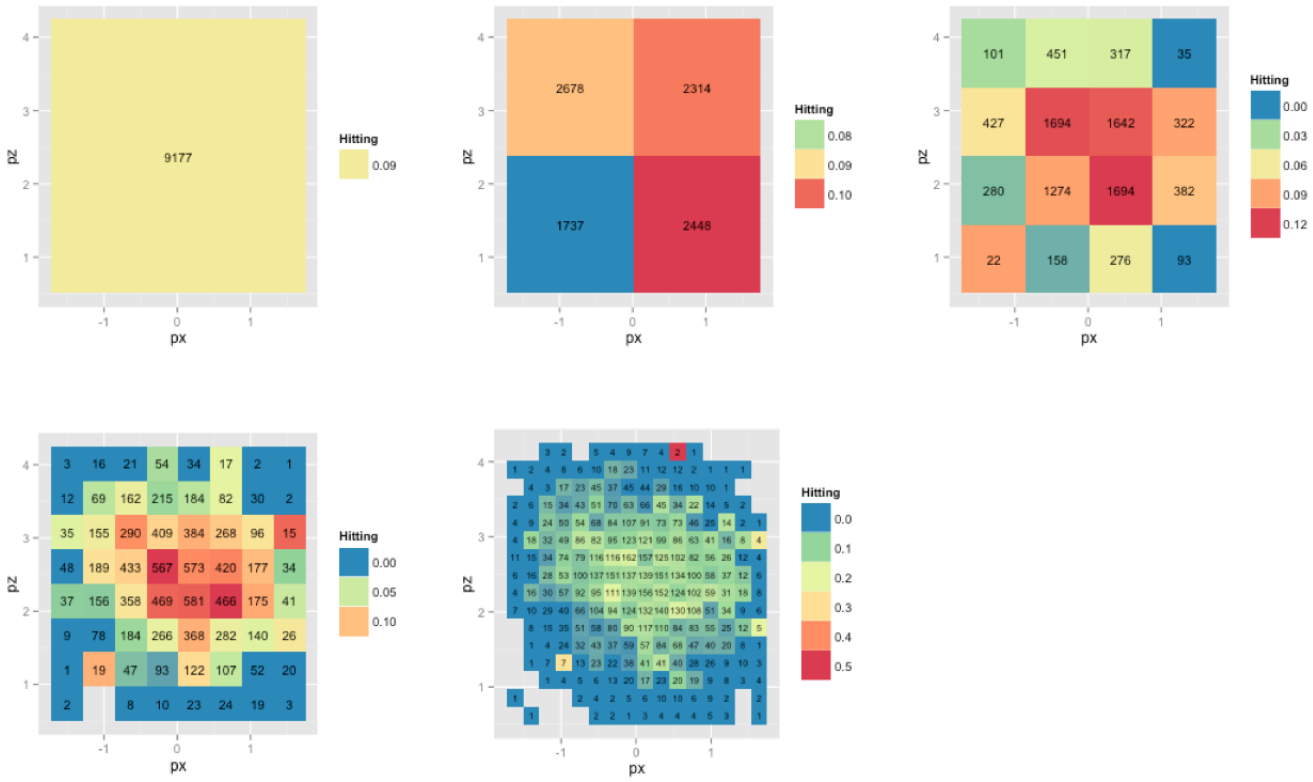

Figure 4: These **six (?)** heat maps show the same data, 9177 swings by batter 425509, Johnny Peralta, at increasing resolutions. Heat map one is unnecessarily coarse, while heat map six is excessively fine. Note how dramatically the visual imact and impression varies as the resolution increases. Which resolution best conveys the data?

It is unclear which of these five resolutions best combines spatially precise estimates of p where possible, and box sample sizes with $\text{Var}(\hat{p}_b)$ in a desirable range. The viewer interested in the center of the strike zone should prefer the last (**need labels** heat map, as the box sample sizes are sufficient to provide such spatially specific low variance estimates. The boxes closer to the edges of the strike zone contain higher variance, and thus less reliable estimates, due to prohibitively small sample sizes. We propose a new heat map approach that combines resolutions according to the data's varying spatial density.

3.1 Variable Resolution Empirical Heat Maps

Consider again the heat map in Figure 2. Notice box (1,1) contains data on 22 swings, a sample size where subdividing would yield sample sizes uselessly small, and thus estimate variances prohibitively high. Box (2,3), in contrast, contains data on 1694 swings, which would support estimates that are more spatially accurate without $\text{Var}(\hat{p}_b)$ increasing past acceptable levels. We propose defining a stopping rule and a subdividing method, and subdividing boxes further accordingly. For example, in Figure 4 we subdivide, into four equally sized boxes, all boxes where $n_b > 200$, . One iteration through all boxes at their current size, subdividing according to this rule, converts the heat map on the left to the heat map on the right.

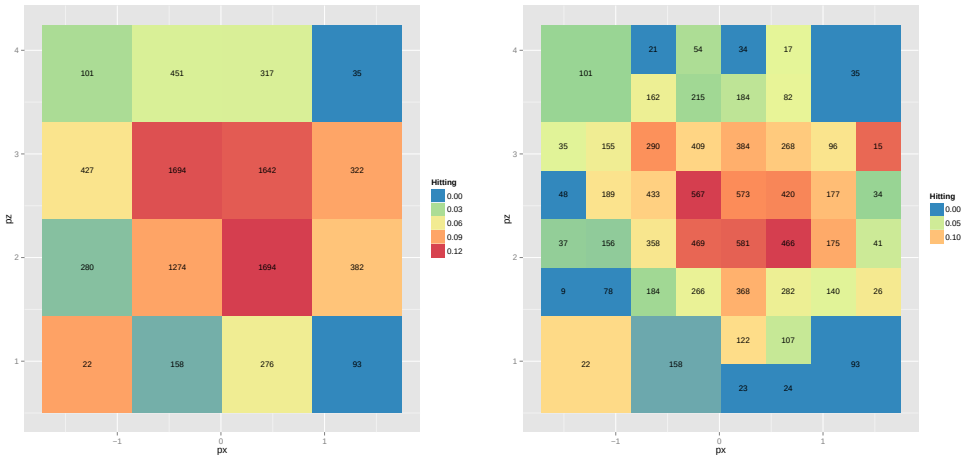

Figure 5: These heat maps convey the empirical batting average of Johnny Peralta in each square region of the hitting zone. Each box maps $\hat{p}_b$ to a color. The number printed on each box represents the number of pitches Peralta swung at that passed through that box. Notice that all boxes with a sample size greater than 200 in the heat map on the left, have been subdivided in the heat map on the right.

Notice all four corner boxes have not subdivided, indicating Peralta seldom sees and swings at pitches in these locations. The boxes toward the middle of the map tend to have larger sample sizes, and higher $\hat{p}_b$. Pitches pass through the middle of the hitting zone more frequently because many pitch target margin of

error circles overlap there; and it is the region where pitch target margin of error circles are entirely inside the strike zone. Sixteen boxes still have a sample size greater than 200, and 11 still have a sample size greater than 300. We iterate again, and further subdivide 16 boxes where $n_b > 200$.

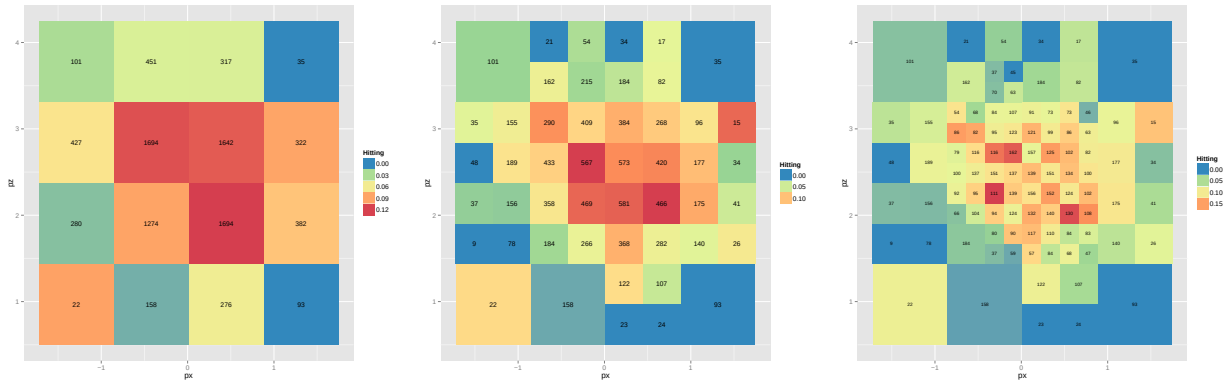

Figure 6: These heat maps convey the empirical batting average of batter 425509, Johnny Peralta, in each boxed region of the hitting zone. Each box maps $\hat{p}_b$ to a color. The number printed on each box represents the number of pitches the hitter swung at that passed through that box. All boxes with a sample size greater than 200 in the heat map on the left, have been subdivided in the heat map in the middle. All boxes with a sample size greater than 200 in the heat map in the middle, have been subdivided in the heat map on the right.

In Figure 6, the middle heat map has 16 boxes with $n_b > 200$. In the heat map on the right these 16 boxes have been subdivided into four boxes each. After this iteration, the heat map on the far right consists of 97 boxes, with a mean box sample size of 94.57, and median of 94. The minimum box sample size is 9, and the maximum is 189. The first quartile box sample size is 63, and the third quartile is 125. Regions with a higher density of pitch-swings necessarily have smaller boxes, which acts to convey additional information to the reader, compared to a heat map on a uniform grid. Note that the stopping rule and subdivision algorithm can be defined by the map's creator, offering flexibility to create the heat map structure that suits the data.

Figure 8 gives the full sequence of heat maps that result from applying the stopping rule $n_b < 100$, starting with a single box.

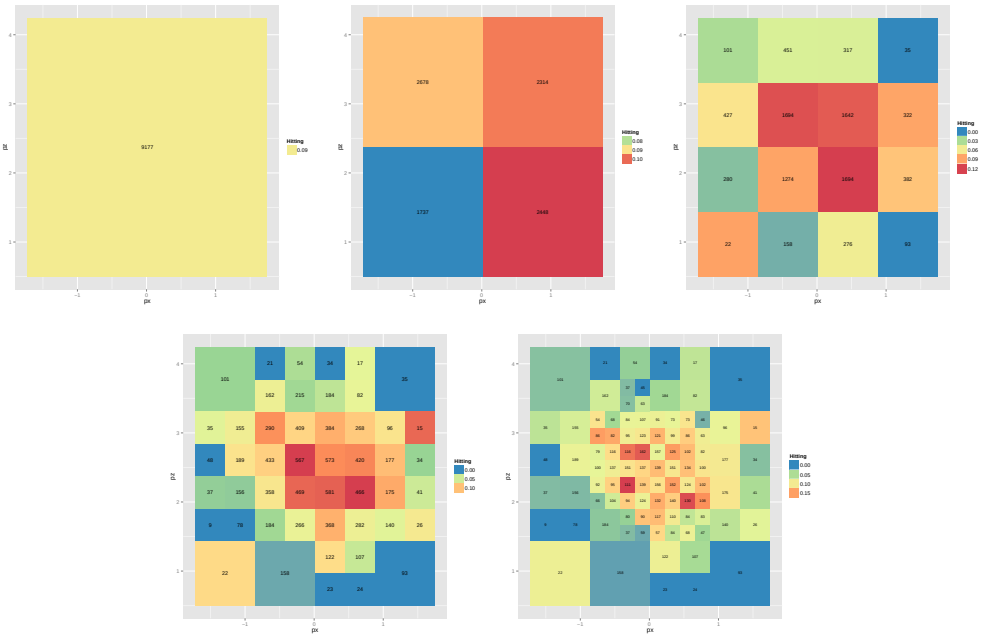

Figure 7: These heat maps convey the empirical batting average of batter 425509, Johnny Peralta, in each boxed region of the hitting zone. Each box maps $\hat{p}_b$ to a color. The number printed on each box represents the number of pitches the hitter swung at that passed through that box. All boxes with a sample size greater than 200 in each heat map have been subdivided in the subsequent heat map.

To demonstrate the flexibility, consider a different stopping rule, $n_b < 100$. Figure 8 gives the sequence of heat maps that result from applying this stopping rule, with the same subdividing algorithm (***need to delineate what this algorithm is).

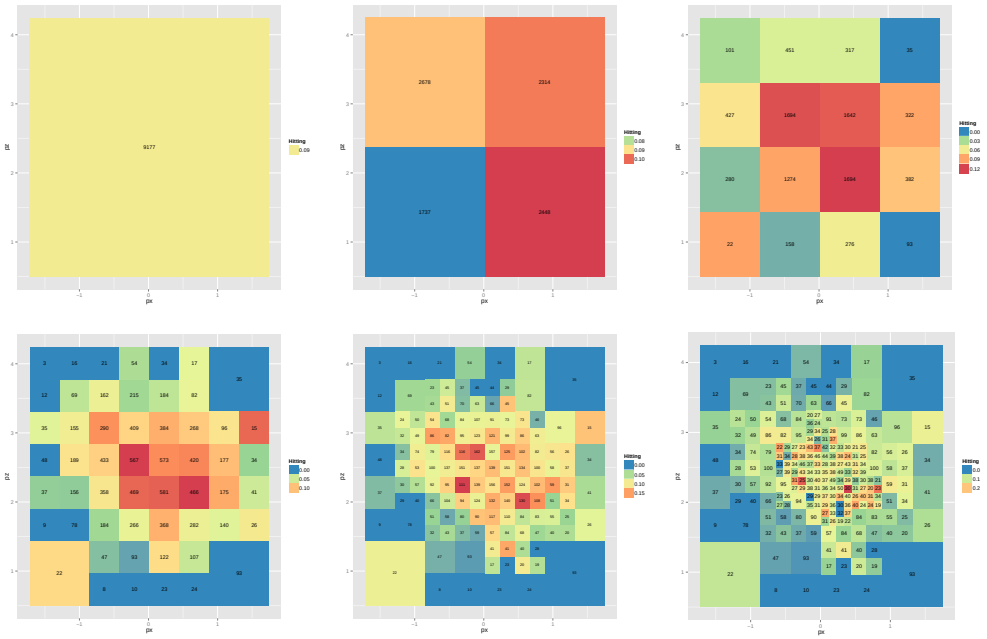

Figure 8: These heat maps convey the empirical batting average of batter 425509, Johnny Peralta, in each boxed region of the hitting zone. Each box maps $\hat{p}_b$ to a color. The number printed on each box represents the number of pitches the hitter swung at that passed through that box. All boxes with a sample size greater than 100 in each heat map have been subdivided in the subsequent heat map.

Compare this sequence to Figure 7, where the stopping rule was $n_b < 100$. The top row of heat maps in Figure 7 and Figure 8 are identical, but notice in the four by four heat map that $100 < n_{(2,1)} < 200$, and $100 < n_{(1,4)} < 200$. This implies one stopping rule applies, but the other does not.

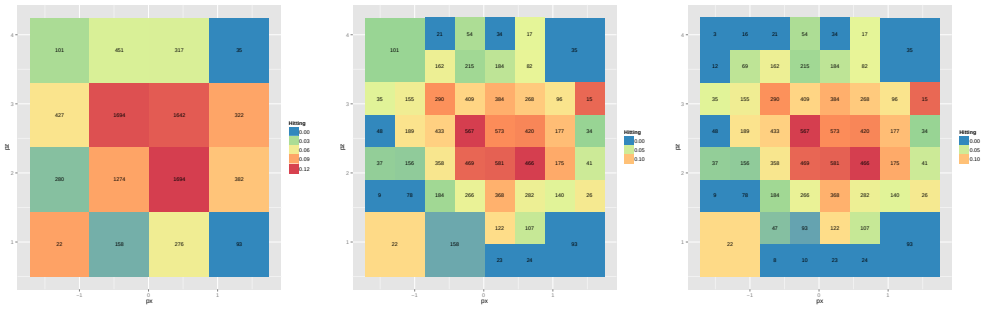

Figure 9: ...(these images, and others, need labels: (A) (B) (C) etc)

For this reason the bottom left heat maps in Figures 7 and 8, shown in Figure 9, differ in the number of boxes of each size, and the total number of boxes. This divergence continues at the next iteration, where the stopping rule $n_b < 100$ requires 28 box subdivisions in Figure 8, map three; and $n_b < 200$ gives 16 box

subdivisions in Figure 7, map three.

Chapter 2 - Modeling Hitter Success Probabilities

4 Generalized Linear Model

Our goal is to create a statistical model for the heat map of success probabilities. Nonparametric methods, while straightforward, sacrifice interpretability; while they achieve a modeled heat map, there may not be interpretable components of the model. Nonparametric models cannot relate spatially varying hitter success probabilities to hitter attributes. We propose a parametric approach using biomechanically interpretable covariates. Existing research analyzes the biomechanics of the baseball swing [Welch et al., 1995], but no research integrates those results with spatial swing outcomes in a statistical model.

Consider again batter 425509, veteran Jhonny Peralta, whose data set contains 9,177 swings. Let success indicator variable, Y_{iklm}, be a Bernoulli random variable with spatially varying mean [Sheldon et al., 2002]. Subscript $i = 1, \ldots, n_{klm}$ indexes Peralta's swings in at bat k against pitcher l in year m. Subscript $k = 1, \ldots, n_{lm}$ indexes Peralta's at bats against pitcher l in year m. Subscript $l = 1, \ldots, n_m$ indexes pitchers Peralta faced, where n_m is the total number of pitchers; and $m = 2007, \ldots, 2016$ indexes year. In this study we make the simplifying assumption that location success probabilities depend on only location and hitter. This means we dispense with subscripts k, l, and m. We also assume that, given a pitch location $\boldsymbol{s}_i$, swings are independent Bernoulli trials; $Y_i|\boldsymbol{s}_i \sim \text{Bernoulli}(p_i)$, where $\text{E}[Y_i|\boldsymbol{s}_i] = p_i$

Accordingly, let $i = 1, \ldots, 9177$ index Peralta's 9,177 swings on record. Let $\boldsymbol{s}_i = (px_i, pz_i)$ be the horizontal and vertical locations, respectively, of pitch i as it passes through the two dimensioned vertical face of the hitting zone. The origin, $\boldsymbol{s} = (0, 0)$, is the midpoint of the front edge of home plate, at ground level. Let $\boldsymbol{X}(\boldsymbol{s}_i)$ be covariates specific to Johnny Peralta and location $\boldsymbol{s}_i$. A Bernoulli random variable suggests, as a starting point, a generalized linear model with logit link function for relating success probability to covariate information:

$$\text{logit}(p_i|\boldsymbol{X}(\boldsymbol{s}_i), \boldsymbol{s}_i) = \boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of coefficient parameters [Myers et al., 2012]. The next step is to develop covariates.

4.1 Biomechanically Interpretable Covariates

Why does Peralta, and why do hitters in general, hit pitches in some locations better than others? We submit biomechanics as potentially part of the answer. Biomechanics underpin why hitters prefer some pitch locations more than others. Given the choice, athletes select a specific place for the ball before swinging. Consider golf, a sport where the ball is stationary, and the acting athlete chooses where to stand in relation to the ball. In fact, golfers position themselves very precisely in relation to the ball to achieve impact at the optimal point in their swing [Cochran and Stobbs, 2005]. If the impact point deviates from the ideal location, performance suffers. Consider tennis, a step closer to baseball, in that the ball approaches, but the player has time to position himself relative to the incoming ball. Once again, tennis players strive to hit the ball at a specific point in their forehand, a precise distance from the ground and from their body [Elliott, 2006]. As with golf, if the point of impact deviates from this location, performance suffers. Note that in both sports the ideal player to ball positioning depends on, at the very least, anatomy, biomechanics, and equipment. We submit the same dynamics affect baseball hitting. However, in baseball the hitter cannot predtermine ball location, nor does he have time to reposition himself in response to the location and trajectory of the incoming pitch. For these reasons, meaningful measurements of hitter to ball distance and angle are reasonable covariates. Polar coordinate pitch locations would inherently provide this type of meaningful covariate for use and interpretation in our models.

To illustrate, in Figure 3 we shift the origin to a hitter's approximate center of gravity in his stance, where the extended bat line intersects his axis of rotation at the moment of contact [Welch et al., 1995].

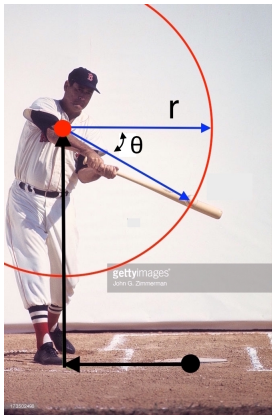

Figure 10: The ground level black dot represents the origin (0,0) in the rectangular coordinate system. The translated origin (red dot) coincides with the hitter's approximate center of gravity, and thus the polar origin. The bold arrows show the origin shift. The length of the arrows moving out from his center of gravity to specific grid locations represent r, and the angle between the same two arrows represent θ.

Referring to Figure 3, let r measure the distance from the hitter's center of gravity to the ball at impact, and let θ be the angle below horizontal of the line segment connecting the center of gravity and the ball at impact. As in golf and tennis, ball location—too far/close to the hitter, or above/below the ideal point of impact— affects hitting performance. Letting $\boldsymbol{X}(\boldsymbol{s}_i)$ in (1) be comprised of r and θ terms provides an exploratory starting point.

4.2 Generalized Linear Model with Biomechanically Interpretable Covariates

Let covariate vector $\boldsymbol{X}(\boldsymbol{s}_i)$ in (1) be defined as $\boldsymbol{X}(\boldsymbol{s}_i) = \{r_i, \theta_i, r_i\theta_i, r_i^2, \theta_i^2, r_i^2\theta_i^2\}$. Substituting into (1) yields:

$$\text{logit}(p_i|\boldsymbol{s}_i, r_i, \theta_i) = \beta_0 + \beta_1 r_i + \beta_2\theta_i + \beta_3 r_i\theta_i + \beta_4 r_i^2 + \beta_5\theta_i^2 + \beta_6 r_i^2\theta_i^2 \tag{2}$$

We fit this model and find maximum likelihood estimates of $\boldsymbol{\beta}$ using an iteratively reweighted least squares algorithm [Myers et al., 2012].

(IT WOULD BE NICE: EMPIRICAL, ESTIMATES, FITTTED HEAT MAP)

Covariate	Parameter	MLE	SE	p
N/A	β_0	-4.08	0.70	< 0.001
r	β_1	1.19	0.51	0.018
θ	β_2	-1.93	1.90	0.311
$r * \theta$	β_3	-1.64	0.70	0.064
r^2	β_4	-0.32	0.09	< 0.001
θ^2	β_5	-3.92	1.10	< 0.001
$r^2 * \theta^2$	β_6	-0.46	0.21	0.025

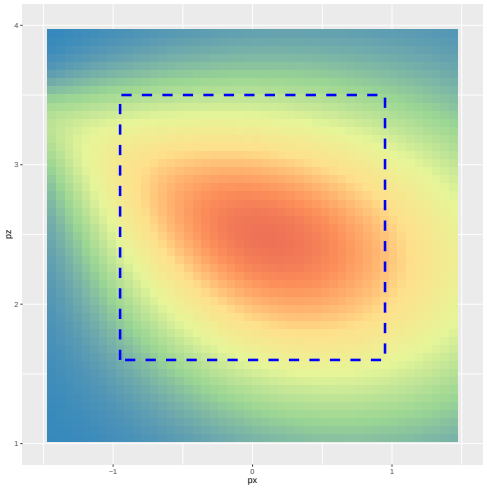

Figure 11:

4.3 Model Evaluation/Validation Here?

4.3.1 Hosmer-Lemeshow Goodness of Fit Test

Generalized Linear Models [Myers et al., 2012]

- (pg 147) Hosmer-Lemeshow test = Logistic regression Goodness of Fit test.
- Like Pearson Chi-Sq, but for continuous covariates.
- Order all responses according to fitted $\hat{p}$, then group into deciles.
- Then basically $\chi^2 = \Sigma\Sigma\frac{(O-E)^2}{E}$
- p-value = 0.1513

5 Spatial Generalized Linear Mixed Model

Independent of the performace of the previously fit model, we expect there to be unexplained spatial variation in the mean. The covariates are limited in scope and depth, and Tobler's First Law of Geography tells us that things close together in space tend to behave more similarly than things further apart [Tobler, 1970]. For these reasons, among others, we enhance the model by adding a spatially correlated random effect.

To explicitly define a spatial model, let $\boldsymbol{s} \in \boldsymbol{D} \subseteq \boldsymbol{R}^2$ be the set of all possible pitch locations $\boldsymbol{s}_i = (px_i, pz_i)$, on $\boldsymbol{D}$, the two dimensional vertical face of the hitting zone that the pitcher throws his pitches through.

5.1 Gaussian Random Field

Let random variable Z_i

5.2 Spatial Logistic Regresion Mixed Model

References

Christian M Welch, Scott A Banks, Frank F Cook, and Pete Draovitch. Hitting a baseball: A biomechanical description. Journal of Orthopaedic & Sports Physical Therapy, 22(5):193–201, 1995.

Ross Sheldon et al. A first course in probability. Pearson Education India, 2002.

Raymond H Myers, Douglas C Montgomery, G Geoffrey Vining, and Timothy J Robinson. Generalized linear models: with applications in engineering and the sciences, volume 791. John Wiley & Sons, 2012.

Alastair J Cochran and John Stobbs. Search for the perfect swing. Triumph, 2005.

B Elliott. Biomechanics and tennis. British Journal of Sports Medicine, 40(5):392–396, 2006.

Waldo R Tobler. A computer movie simulating urban growth in the detroit region. Economic geography, 46 (sup1):234–240, 1970.