AN ABSTRACT OF THE DISSERTATION OF

Chris Comiskey for the degree of Doctor of Philosophy in Statistics presented on August 15, 2017.

Title:

Take Me Out to (Analyze) the Ballgame: Visualization and Analysis Techniques for Big Spatial Data

Abstract approved: _____

Alix Gitelman & Charlotte Wickham

For spatial data visualization, we approach two problems and provide solutions: heat map resolution selection, and heat map confidence interval presentation. Analysts often present spatial data in gridded heat maps, at some chosen resolution. However, many data types vary in density across the domain. We propose varying-resolution heat maps to visually accommodate this changing density. Further, heat map confidence intervals (CI) typically consist of two heat maps, one for each CI bound. We propose an interactive heat map CI that changes dynamically as a user moves through the CI.

For spatial data analysis, Bayesian hierarchical models work well for accommodating complex spatial correlation structures. However, with *big* spatial data we face a computational bottleneck on the order of n^3. We discuss three approaches to confronting the "big N" problem with our spatial baseball strike zone data, and present preliminary assessments.

Take Me Out to (Analyze) the Ballgame: Visualization and Analysis Techniques for Big Spatial Data

by

Chris Comiskey

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented August 15, 2017
Commencement June 2017

Doctor of Philosophy dissertation of Chris Comiskey presented on August 15, 2017.

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Chris Comiskey, Author

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

Take Me Out to (Analyze) the Ballgame: Visualization and Analysis Techniques for Big Spatial Data

1 Variable-Resolution Heat Maps

1.1 The Setup

1.1.1 Introduction: "A picture is worth a thousand words."

Statisticians rely on graphical displays—pictures, essentially—to communicate information, analysis, and results. As technology generates more data, and statistical analysis becomes more prevalent, graphical displays become more important; therefore, our graphic-making abilities must improve. The R data visualization package `ggplot2` highlights the importance of graphical displays, and the continued need for graphical innovations; less than ten years old, it is the most downloaded R package (RDocumentation, 2017).

In this chapter we focus on one type of graphical display: heat maps. Current heat maps consist of uniformly sized grid boxes across the domain. This means one resolution must suffice even if data density varies through the domain, which is something of a limitation. We present variable-resolution heat maps which, we believe, provide an improvement over current heat maps in this respect. In addition, variable-resolution heat maps convey spatial data density through the domain, information traditionally not a part of heat maps.

1.1.2 Bernoulli Swings

Baseball revolves around a series of contests between the hitter and the pitcher, comprised of pitches the hitter can swing at. Throughout our research we treat every swing as a Bernoulli trial, and evaluate success or failure of a swing independently from the count in the at bat at the time of the swing. This differs from the norm; all other known research includes only at bat ending pitches (Cross and Sylvan, 2015), (Baumer and Draghicescu, 2010), (Fast, 2011). Therefore, these studies exclude from analysis swinging strikes that do not end at bats and foul balls; but include called strike three pitches (non-swings) that end at bats. We consider the latter event a matter of hitter decision making, not a failed swing attempt.

Accordingly, we define success as trials where the variable `des`, short for description, equals `in play, no out`, and failure as swings where `des` equals `Foul`, `Foul (Runner Going)`, `Foul Tip`, `In play out(s)`, `Swinging Strike`, or
`Swinging Strike (Blocked)`. Next we explain the structure and interpretation of an empirical baseball strike zone heat map.

1.1.3 Empirical Strike Zone Heat Maps

Empirical baseball strike zone heat maps cover the two-dimensional, vertical face of the strike zone with a grid, containing empirical success probabilities in each grid box ($\hat{p}_b$, defined below). We start with PITCHf/x® data on 1,932 right-handed hitters, taking 1,582,581 swings between 2008 and 2015. Noting the heat map below in Figure 2.1, let $b = 1,\ldots,627$ index grid boxes; $i = 1,\ldots,1,582,581$ index swings; and define

$n_b = \sum_i \mathrm{I}_{\{i \in b\}}$ as the total number of swings in box b.

Define Bernoulli random variable, S_i, to equal one for swing success and zero for swing failure, and define $\hat{p}_b = \frac{1}{n_b} \sum_i S_i \cdot \mathrm{I}_{\{i \in b\}}$ as the empirical box b hitter swing success probability. Figure 2.1 displays the empirical heat map for $\hat{p}_b$. The graphic maps $\hat{p}_b$ to a color on a spectrum, for pitches that passed through the space represented by that grid box.

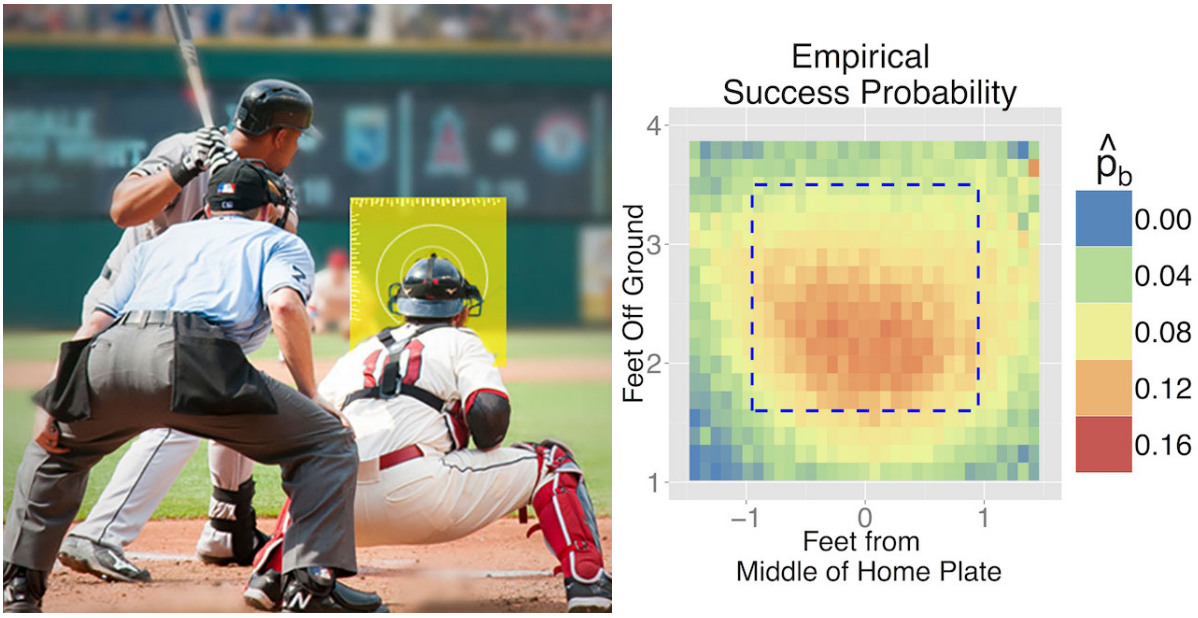

Figure 1.1: The superimposed yellow box in the image on the left shows the plane in space the heat map on the right represents. In particular, the yellow box border approximately coincides with the dashed blue line. The heat map grids the vertical face of the hitting zone with approximately 3/4 inch by 3/4 inch boxes. Each grid box color represents the empirical success probability ($\hat{p}_b$) of hitter swings at pitches passing through that box. The data consists of 1,932 right handed hitters, swinging at 1,582,581 pitches between 2008 and 2015.

The graphic efficiently conveys spatial empirical success probabilities by mapping $\hat{p}_b$ to colors. However, note that the statistician determines the map's resolution.

1.2 Traditional Heat Maps

1.2.1 Resolution

A heat map's creator tacitly chooses a resolution. This determines a uniform grid box size for the entire heat map, which influences heat map quality and appearance markedly. To understand this influence, consider histograms, where bin width selection similarly affects appearance and function. Small box size may yield unreliable estimates at some locations, while larger boxes may fail to convey preferred spatial specificity.

Along these lines, note that heat maps conceal spatially varying data density; in our case, pitch-swings density through the strike zone. Without this information, heat maps also conceal estimate sample sizes and variances.

With these points in mind, consider again the heat map in Figure 2.1. That heat map divides the hitting zone into relatively small boxes, because the data supports it; approximately 1.5 million swings by almost 2000 hitters. By "supports it" we mean that the small, spatially specific boxes at this resolution retain sample sizes large enough to supply reasonable estimates of p_b. "Reasonable," of course, depends on context and objectives. For example, a pitching coach might request estimates within 0.005 points of the true batting average with probability 0.95. This requires a sample size of at least 36 when $p_b = 0.10$.[1]

On the other hand, data for individual hitters varies dramatically in size. In our database, individual hitters range from a single swing to over 10,000 swings. At such

[1] $\text{Var}(\hat{p}_b)$ depends on p_b This creates counterintuitive behavior around the margins of the hitting zone, where $\text{Var}(\hat{p}_b)$ remains very small despite very small box sample sizes. See Dixon et al. (2005) for a discussion of this curious phenomenon.

varying scales, resolution selection becomes more complicated because non-uniform data density implies different regions may support very different resolutions. This is important because, as stated before, the choice of resolution sometimes dramatically affects heat map appearance, but also usefulness. For example, coarse resolution in regions of interest means the parameter estimates lose value.

This important resolution decision usually depends on the size and nature of the data set, and its spatial dispersion through the domain. In the next section we explore this decision in detail, along with its inherent compromises.

1.2.2 Resolution Selection

In this section we use batter 425509, a veteran player named Jhonny Peralta, to explore resolution selection and its implications. The data includes 9,177 Peralta swings, which yields the heat map in Figure 1.2. This map resolution divides the hitting zone into 16 equally sized boxes. Each box maps $\hat{p}_b$ to a color, with box sample sizes, n_b, printed on box centers. We will use box sample sizes to reference boxes. For example, we will call the box in the lower-left "box 22."

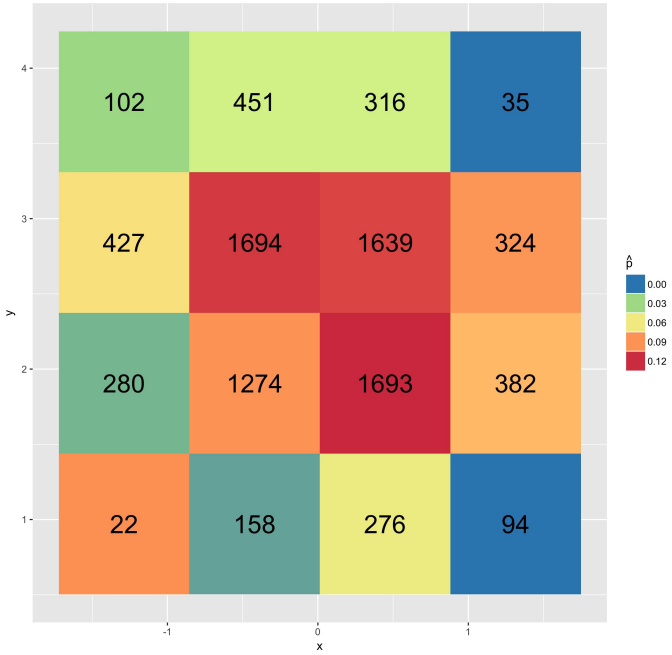

Figure 1.2: This four by four heat map shows the empirical batting average of jhonny Peralta, for pitches passing through the space represented by each of 16 square regions of the hitting zone. Each box maps $\hat{p}_b$ to a color, with box sample sizes, n_b, printed on box centers.

Notice that while the four by four resolution suffices for box 22, further subdivision might yield trivially small sample sizes. On the other hand, the four central boxes, all with sample sizes above 1200, can and should contribute more location-specific estimates. Therefore, the central boxes motivate finer resolution, even though box 22 does not support it. Keeping this trade-off in mind, we increase resolution by dividing each box into four equally sized sub-boxes. Figure 1.3 shows the 8×8 resolution result.

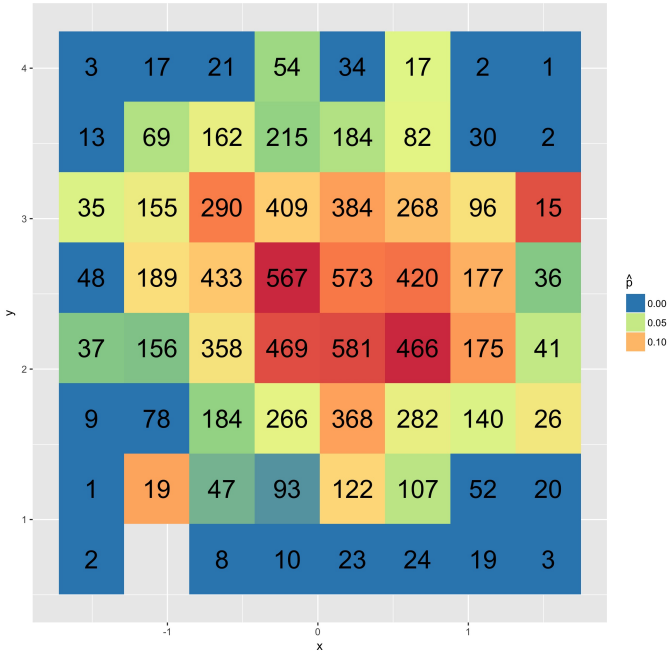

Figure 1.3: This eight by eight heat map shows the empirical hitting success probability of jhonny Peralta, for pitches passing through the space represented by each of 64 square regions of the hitting zone. Each box maps $\hat{p}_b$ to a color, with box sample sizes, n_b, printed on box centers. A grey box indicates no pitches passed through that box. Notice that this resolution imparts additional information in the center of the hitting zone, some box sample sizes toward the margin have dropped uninformatively low.

The centermost 16 boxes still support low variance p estimates; the minimum of these boxes contains 184 swings. Globally, 24 boxes consist of over 150 swings; and 15 boxes still include more than than 250 swings. These boxes could support higher resolution. On the other hand, many boxes, especially edge boxes, now contain sample sizes generally insufficient to support low variance estimates of p_b. Twenty-nine boxes contain fewer than 50 swings, and 17 boxes contain fewer than 20 swings. One box recorded zero swings.

With this range of box sample sizes, the non-extreme resolution choices contain both boxes with exceedingly small sample sizes *and* boxes with unnecessarily large sample sizes. Figure 1.4 shows Peralta's data—the same data—at six resolutions.

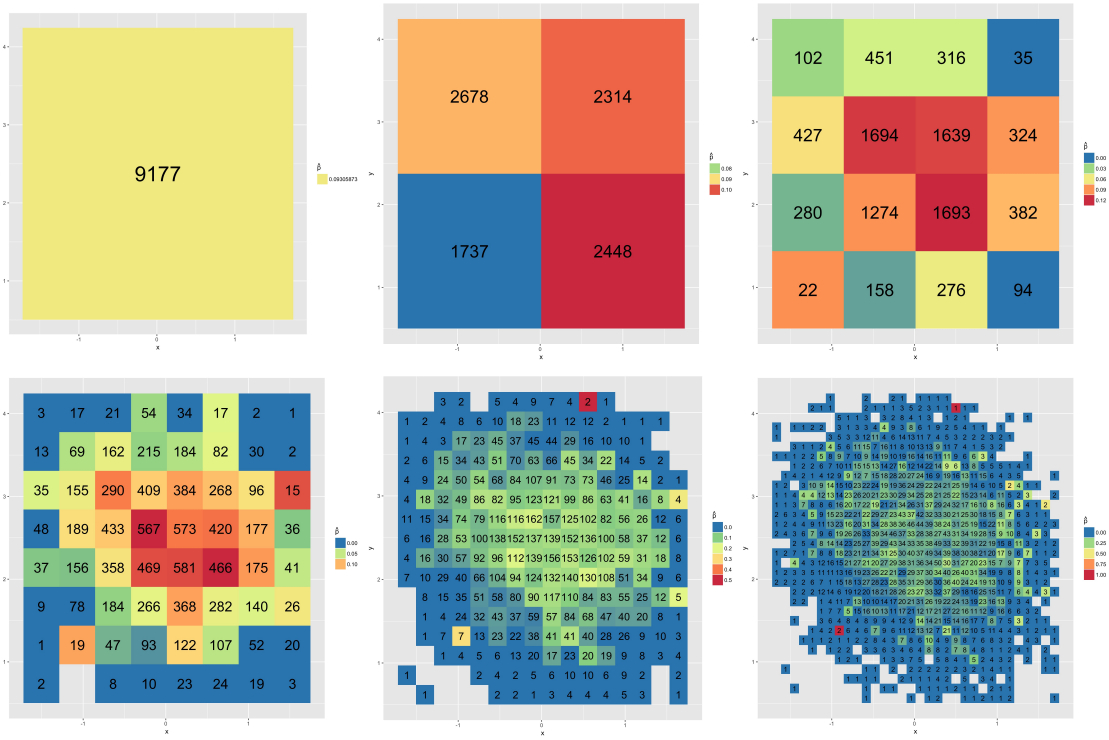

Figure 1.4: These six heat maps show the same data for jhonny Peralta, at increasing resolutions. The maps range from obviously too coarse to perhaps excessively fine. Notice how dramatically the image changes as the resolution increases. Which resolution yields the highest quality heat map?

We start with one box, and subdivided each box into four equally-sized, smaller boxes at each iteration. Which of these six resolutions best balances spatially precise estimates of p with acceptable box sample sizes? The user interested in the center of the strike zone might prefer the bottom-middle map, as the box sample sizes support such

spatially specific estimates. However, the boxes closer to the edges of the strike zone then contain prohibitively small sample sizes, yielding higher variance estimates.

The takeaway from Figure 1.4: all six resolutions involve trade-offs. We propose a new heat map approach that eliminates trade-offs. The solution combines multiple resolutions into one map, according to the data's varying spatial density.

1.3 Variable-Resolution Heat Maps

1.3.1 The Varying Resolution Solution

Consider again the Peralta 4×4 heat map in Figure 1.2. Recall box 22 contains 22 swings, a sample size where subdividing further yields uninformatively small sample sizes. In contrast, box 1694 would support estimates that are more spatially accurate without $\text{Var}(\hat{p}_b)$ increasing beyond acceptable levels. We propose deciding resolution increases box by box algorithmically, with a stopping rule; we christen it the variable-resolution (VR) algorithm. The map's author chooses a stopping rule, giving him/her the flexibility to create the heat map that suits the data.

To demonstrate one iteration in the VR algorithm, let the stopping rule be a maximum box sample size of 200, and recall the 4×4 map in Figure 1.2 (Figure 2.6, left). On this map we divide all boxes where $n_b > 200$, into four smaller, equally sized boxes. Figure 1.5 shows the map before and after. Moving through all boxes of the map to the left, subdividing when $n_b > 200$, yields the heat map to the right.

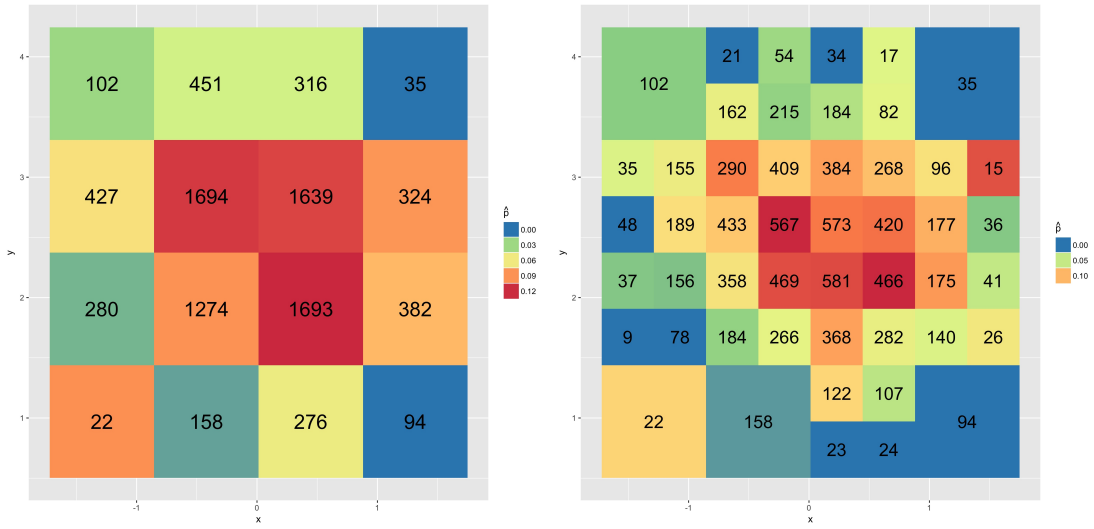

Figure 1.5: One iteration in the variable-resolution algorithm. The algorithm subdivides all boxes with a sample size (printed on box) greater than 200. The iteration shown here yields the map on the right, from the map on the left. Box 22, and others like it, remain intact because further subdivision yields uninformatively low sample sizes.

The algorithm subdivides all boxes that have more than 200 observations, yielding the heat map on the right. Boxes with less than 200 observations, such as box 22, remain intact because further subdivision yields sample sizes the hypothetical author deems uninformative. Sixteen boxes still contain a sample size greater than 200, and 11 still have a sample size greater than 300. The next algorithmic iteration subdivides all 16 boxes that still contain more than 200 observations. The left and middle heat maps in Figure 1.6 duplicate Figure 1.5.

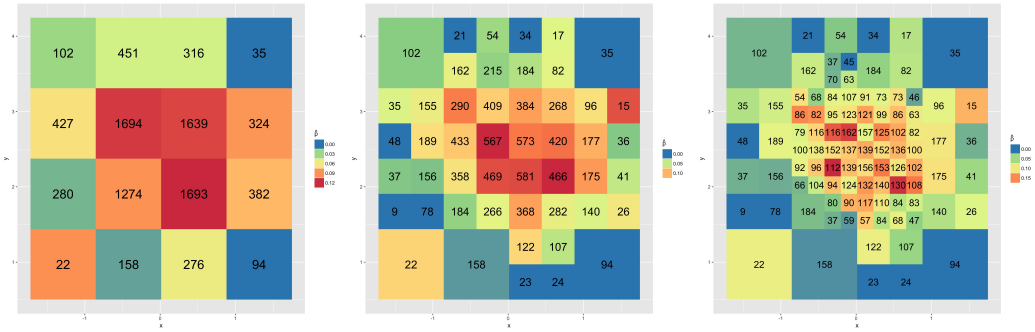

Figure 1.6: Two iterations in the variable-resolution algorithm. The algorithm subdivides all boxes with a sample size (printed on box) greater than 200. The first iteration shown here yields the map in the middle, from the map on the left; the second iteration yields the map on the right, from the middle map.

The new iteration yields the map on the right from the middle map. The new map consists of 97 boxes, with a mean box sample size of 94.57, and median of 94. Box 9 contains the fewest observations, and box 189 contains the most. Box 63 serves as the first quartile, while box 125 serves as the third quartile.

Figure 1.8 shows the VR heat map for every iteration, for stopping rule $n_b < 200$.

Figure 1.7: Variable-resolution heat map sequence. Starting from the top left, the algorithm subdivides all boxes with a sample size (printed on box) greater than 200. The maps convey Jhonny Peralta's empirical success probability by mapping $\hat{p}_b$ to a color.

1.3.2 Data Density Information

Notice how box size corresponds to data density in VR heat maps. Figure 1.7 demonstrates this correspondence by comparing a scatterplot that shows data density, to a VR heat map.

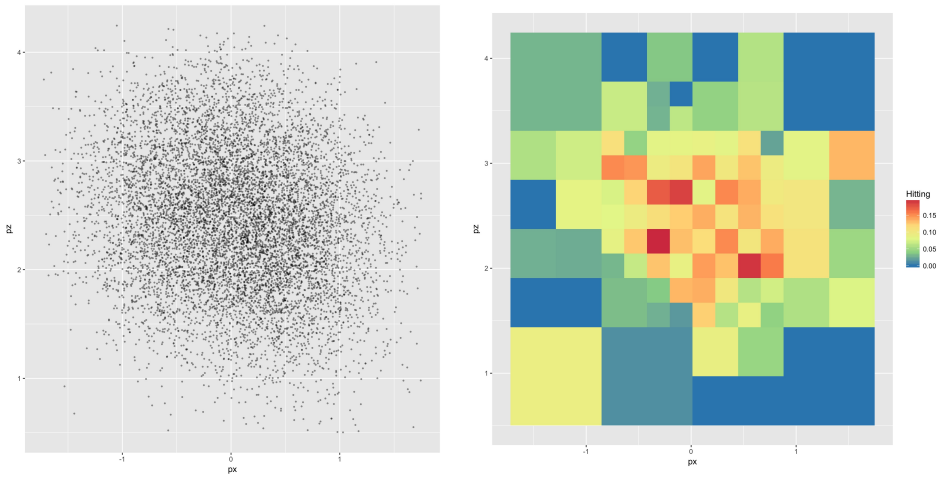

Figure 1.8: A variable-resolution heat maps convey data density. Comparing Jhonny Peralta's scatter plot and variable-resolution heat map shows the correspondence between data density and box size. The finer resolution regions in the heat map correspond to greater data density, whereas biggers boxes indicate lower density. Traditional heat maps omit this information.

Notice how smaller boxes correspond to higher data density, while larger boxes indicate lower density. This contrasts favorably to traditional heat maps, where uniform resolution conceals data density. VR heat maps convey valuable, previously omitted data density information to the viewer.

1.3.3 Stopping Rule Example

In this section we apply the VR algorithm to the same data, but with stopping rule $n_b < 100$, to show how the procedure and outcome differs. Figure 1.9 shows heat maps for all six VR iterations.

Figure 1.9: A variable-resolution heat map sequence. Starting from the top left, the algorithm subdivides all boxes with a sample size (printed on box) greater than 100. The maps convey Jhonny Peralta's empirical success probability by mapping $\hat{p}_b$ to a color.

Compare Figure 1.9 for stopping rule $n_b < 100$, to Figure 1.8 for stopping rule $n_b < 200$; the top rows match exactly. However, notice Box 158 and Box 102 in the 4×4 heat map; both boxes have sample sizes *between* the two stopping rules. Based on this fact, we see diverging paths, where one stopping rule prevents further subdivision, while the other compels it. Figure 1.10 shows the subsequent map for each stopping rule.

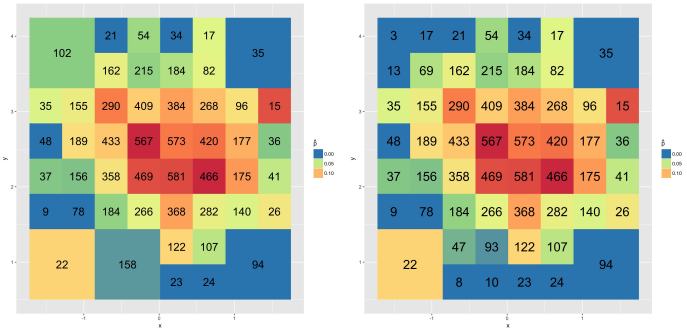

Figure 1.10: Variable-resolution heat maps diverge when box totals fall between sample size stopping rules. Boxes 158 and 102 remain for stopping rule $n_b < 200$ (left), but further subdivide for stopping rule $n_b < 100$ (right).

As expected, the heat maps now differ in the number of boxes of each size, and the total number of boxes. The differences increase at the next iteration, where stopping rule $n_b < 100$ yields 28 box subdivisions (Figure 1.9); and $n_b < 200$ yields 16 box subdivisions (Figure 1.8). Figure 1.11 shows two corresponding iterations for these two stopping rules; $n_b < 100$ in the top row, and $n_b < 200$ in the bottom row.

Figure 1.11: Variable-resolution heat maps diverge when box totals fall between sample size stopping rules. The top row of maps subdivides according to stopping rule $n_b < 200$; the bottom row further subdivides for stopping rule $n_b < 100$. The maps match through three iterations, the last of which we see first in each row. However, the next two iterations produce different maps.

For both stopping rules, notice how corner boxes tend to remain intact. This indicates Peralta swings less often at pitches in the strike zone corners, and/or less often sees such pitches. In the next section we delve into the possible reasons for this pattern, and others, in a baseball context.

1.3.4 Hitter vs. Pitcher

In this section, for illustrative purposes, we interpret features of the best–in this author's opinion–jhonny Peralta variable-resolution heat map. Toward this end, we first

describe basic strategy considerations in the hitter vs. pitcher matchup.

Context

At the basic level, the pitcher wants to get outs and avoid baserunners; and hitters want to avoid outs and get on base. We are concerned with pitch location in this study, so we want to understand the pitch-swing location decisions; where the pitcher decides to throw, and which pitches the hitter decides to swing at.

Three primary factors influence pitch-swing location: pitcher game theoretic strategy, pitch location margin of error (distance by which a pitch misses its intended target), and game state. Game theoretic strategy concerns the pitcher's knowledge of the hitter's strengths and weaknesses, and the hitter's reciprocal knowledge. Margin of error concerns the pitcher's tendency to miss his intended target by some amount. The game state includes the count (number of balls and strikes), the number of outs, and baserunner presence/absence. 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); and the increased penalty for a hit with a runner in "scoring position." Our data includes pitches across all game states, so we will not rely on them for interpretation. Instead, we focus on game theory and margin of error rationale.

Interpretation

Figure 1.13 gives, in the opinion of this author, the best heat map for jhonny Peralta.

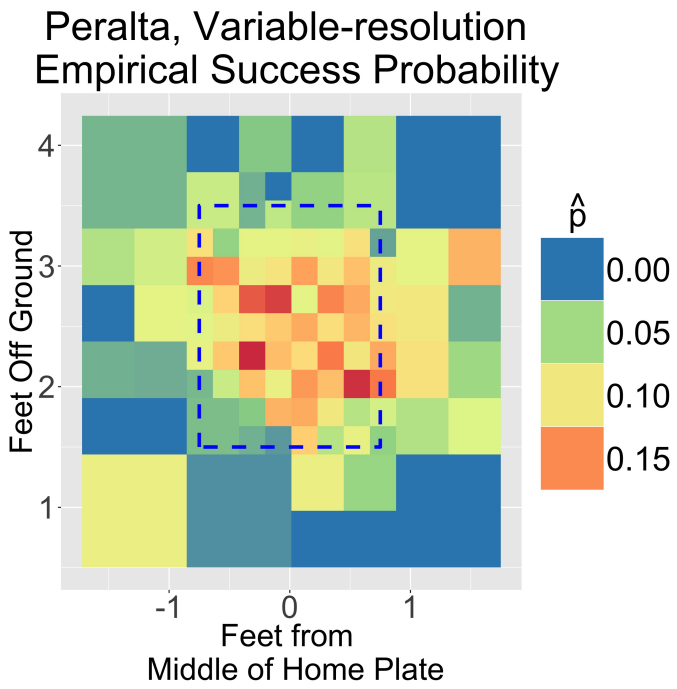

Figure 1.12: This variable-resolution heat map for jhonny Peralta gives spatial, empirical success probabilities; each box maps $\hat{b}_b$ to a color. In addition, the map conveys data density information; box size corresponds to the observation density, because subdivision persisted until box sample sizes dropped below 200.

The map gives spatial, empirical success probabilities for jhonny Peralta. In regions of greater data density, and perhaps greater relevance, the map gives more spatially specific estimates. The size of the box conveys this density information, traditionally concealed in heat maps. This new feature derives from the fact that box subdivisions persist until box sample sizes drop below 200 (in this map).

Interpretable features abound. For example, Peralta generally saw and swing at far more pitches in the strike zone than out. Pitchers throw pitches in the strike zone because each strike is one step closer to an out, and each ball is one step closer to a baserunner. The same logic, with opposite goals/incentives, applies for the hitter. Therefore, many

pitch-swings occur in the strike zone.

Notice the larger boxes in the lower left, upper left, and upper right of the strike zone. This implies Peralta saw and/or swung at fewer pitches in these locations. To explain, consider a pitch aimed for the bottom left of the strike zone. Missing the box below or to the left yields a ball, a result unambiguously in the hitter's favor; one step closer to the fourth ball of an at bat, when the rules award the hitter first base. In contrast, any swing, even those at the most hitter-favoring locations, yield an out with non-zero probability.

Elsewhere, boxes toward the middle of the strike zone contain more observations; finer resolution implies Peralta sees and/or swings at more pitches there. This region also seems to have greater success probabilities $\hat{p}_b$ (warmer colors), indicating Peralta indeed hits pitches in the center of the hitting zone better than pitches toward the edges of the hitting zone. In particular, the warmest boxes seem to be quite vertically centered. This indicates Peralta deals with horizontal variation better than vertical variation–an actionable insight! Peralta will swing at pitches in the middle of the strike zone almost every chance he gets.

Why did pitchers not avoid those locations? They probably tried! However, game state pressures frequently compel pitchers to throw a strike, and aiming for the middle of the strike offers the highest probability of a strike; missing the target by up to ten inches in any direction still yields a strike. Even further, when aiming away from the middle of the strike zone, some pitches miss their target and unintentionally pass through the middle of the strike zone.

Of the four corner locations beyond the strike zone, Peralta enjoyed the most success low and inside, with $\hat{p}_b = 0.10$. However, the size of this box indicates a relatively small

sample size, indicating this estimate may not be particularly reliable. With regard to the smaller sample sizes, Peralta probably swings at these pitches less often because they are, quite simply, difficult to hit successfully. On the other hand, pitchers may *throw* to these spots less often because of their relatively greater associated risk; they are less likely to induce swings, and a non-swing (ball) is one step closer to a baserunner.

1.4 Tornado Example

The National Weather Service, a division of the National Oceanic and Atmospheric Administration (NOAA), maintains seven National Centers for Environmental Protection. One of these, the Storm Prediction Center, collects tornado data, and provides it to the public at their website (NOA, 2017). In this section we use spatial tornado data between 1950 and 2017 to show variable-resolution heat maps in another area of application. The data consists of numerous covariates; we use only the longitude, latitude, and F-Scale rating for approximately 30,000 tornadoes in the last 67 years. The variable-resolution heat map in Figure 1.14 shows the spatially varying intensity and density of tornadoes.

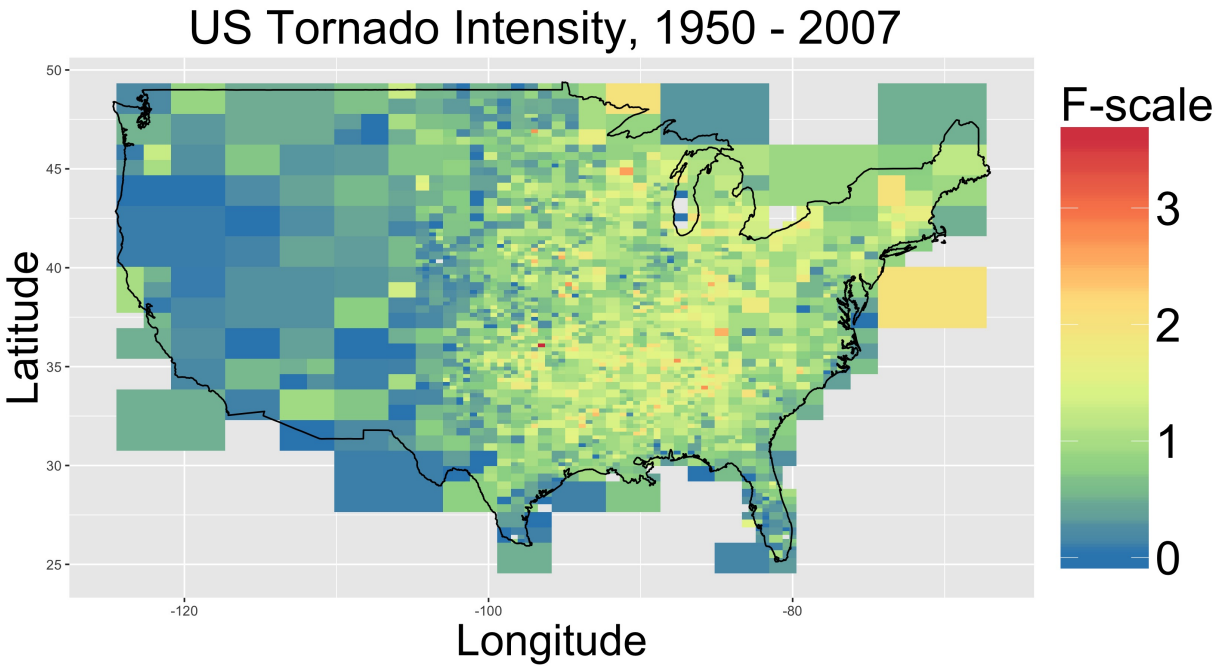

Figure 1.13: This variable-resolution heat map shows tornado F-Scale intensity and density across the U.S. Each box maps average intensity to a color. In addition, the map conveys data density information; box size corresponds to the observation density, because subdivision persisted until box sample sizes dropped below 50 tornadoes.

Notice how the variable-resolution map conveys data density through grid box size. To the East of $-80°$ West, tornadoes are moderately common, and moderately intense. However, between approximately $-80°$ West and $-100°$ West tornadoes are more intense and far more common. Moving further West, tornadoes become less common and less intense than either of the previous two longitudinal ranges.

Some grid boxes remain unnecessarily large, and awkwardly shaped relative to the U.S. borders. This occurs when an early subdivision leaves one protruding quadrant

with at least one observation, but fewer than the cutoff. The early iteration box size then remains throughout. For example, a large horizontal rectangle overlaps Maine, in the North-eastern most corner of the U.S. The same is true of boxes that overlap New Jersey, Southern Florida, Southern Texas, and Southern California. Additional subdivision features could eliminate this issue. For example, another algorithmic layer could examine edge boxes for this quality and subdivide them.

A sequence of images showing the variable-resolution algorithm iterations shows how the image evolves to its final state.

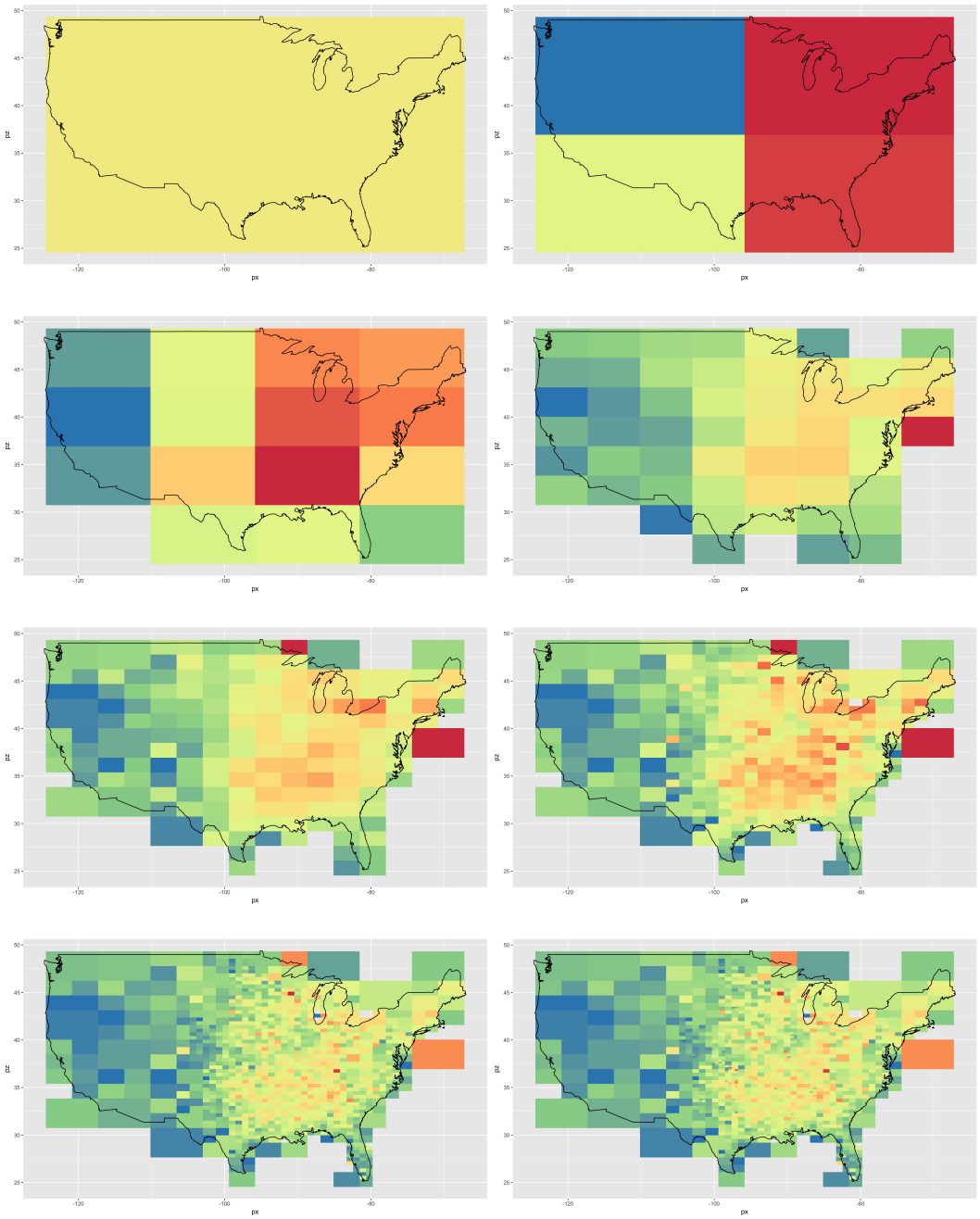

Figure 1.14: This sequence of shows iterations of the variable-resolution subdivision algorithm. The maps show tornado F-Scale intensity across the U.S., and the subdivision sequence indicates regional density. Subdivision persisted until box sample sizes dropped below 50 tornadoes in the final image.

Bibliography

(2017). Noaa's national weather service, storm prediction center.

Baumer, B. and Draghicescu, D. (2010). Mapping batter ability in baseball using spatial statistics techniques. *JSM, American Statistical Association*, pages 3811–3822.

Cross, J. and Sylvan, D. (2015). Modeling spatial batting ability using a known covariance matrix. *Journal of Quantitative Analysis in Sports*, 11(3):155–167.

Dixon, P. M., Ellison, A. M., and Gotelli, N. J. (2005). Improving the precision of estimates of the frequency of rare events. *Ecology*, 86(5):1114–1123.

Fast, M. (2011). Spinning yarn: Can we predict hot and cold zones for hitters?

RDocumentation (2017). Top 5 packages.

APPENDIX

APPENDIX