Mapping Batter Ability in Baseball Using Spatial Statistics Techniques

Ben Baumer* Dana Draghicescu†

Abstract
In baseball, an area in or around the strike zone in which batters are more likely to hit the ball is called a hot zone. Scouting reports are often based on maps displaying a player’s batting average in a discretized area of the strike zone. These reports are then used by both batting and pitching coaches to devise game strategies. This paper is motivated by the Sportvision PITCHf/x data, which provides accurate continuous location coordinates for individual pitches using high-speed cameras. Extended exploratory analyses show a number of interesting and challenging spatial features that we exploit in order to produce improved hot zone maps based on both parametric (kriging) and nonparametric (smoothing) techniques.

Key Words: baseball, hot zone, kernel smoothing, kriging, PITCHf/x data, spatial prediction

1. Motivation

In baseball, the concept of a hot zone (an area in or around the strike zone in which batters are more likely to hit the ball well) is common, going at least as far back the publication of Ted Williams’s book, The Science of Hitting, in 1986 (Williams and Underwood 1986). Major television broadcasts commonly display a graphic in which the strike zone is divided into a 3×3 grid, and a batter’s batting average in each cell is shown. Within the industry, several vendors produce more elaborate scouting reports based on this same principle of displaying a batter’s batting average in a discretized area of the strike zone. These reports are then used by batting coaches to help batters identify and correct weaknesses in their swings, but more often by pitching coaches to help pitchers devise a strategy to minimize the performance of the opposing batters. These reports typically suffer from three major drawbacks:

1. They model batting ability over a discretized grid rather than a smooth surface;

2. They convey little or no information for players with few observations;

3. They provide no uncertainty estimates.

In this paper, we explore the PITCHf/x data set in detail using both parametric and nonparametric spatial techniques. We seek a robust and automatic method for producing smooth, accurate, easily understandable hot zone maps. We hope to address each of the drawbacks listed above.

Examples of typical hot zone maps using binning and filled contour techniques are shown in Figure 1 for Josh Thole, with a sample of just 23 observations, and

---

* CUNY Graduate Center, 365 Fifth Avenue, New York, NY 10016. New York Mets, Flushing, NY 11368.
† Hunter College, CUNY, Department of Mathematics and Statistics, 695 Park Ave, New York, NY 10065.
in Figure 2 for David Wright with a more typical 554 observations. It can be seen that the map for Thole is completely non-informative, while the maps for Wright remain quite rough. Allen has produced a series of similar maps using nonparametric techniques (see Allen 2009).

To summarize, the problem of interest is to identify spatial patterns for each player, and to develop a comprehensive exploratory tool that should automatically and accurately map the batter ability for all players, based on pitch locations around the strike zone. In particular, our goal is to improve upon the maps shown above. The paper is organized as follows. Section 2 presents the baseball setup and the PITCHf/x data, followed by statistical modeling in Section 3. Applications of the methodology described in Section 3 are shown in Section 4. A brief discussion is given in last section.
2. The PITCHf/x data

Sportvision has produced PITCHf/x data since 2006, and has licensed it to Major League Baseball Advanced Media for distribution. All such data is available online via MLB GameDay (see Sportvision and MLBAM 2010). The data is collected by two high-speed sensor cameras, which are mounted high above the field in each major league ballpark. Each camera takes approximately 60 images per second while the pitch is in flight, and records the location of the ball in three dimensions. Proprietary software takes those measurements and solves for the equations of motion, using a coordinate system outlined below. The cameras begin tracking when the ball is 50 feet from home plate\(^1\). Let \(v = v(x, y, z)\) and \(a = a(x, y, z)\) be the velocity and acceleration vectors of the ball, respectively. The position \(s\) of the ball at time \(t\) is then given by \(s(t) = s_0 + v_0 t + \frac{1}{2} a_0 t^2\), where \(t\) is the time since the ball is released by the pitcher. Here \(s_0\) is the initial position of the ball, \(v_0\) and \(a_0\) are the initial velocity and acceleration vectors, respectively. The PITCHf/x data set contains explicit values for \(s_0, v_0, a_0\), as well as the position of the ball when it crosses the front of home plate, \(s_f\). The margin of error of these measurements is claimed to be around 0.4 inches. Note that downward acceleration due to gravity is implicit in the \(z\)-coordinate of \(a_0\), and that \(y(s_0) = 50, y(s_f) = 0\) for every pitch. Finally, the initial speed \(v_0\), as well as the speed at which the pitch crosses the plate \(v_f\), are also recorded in miles per hour.

Let \(D \subset \mathbb{R}^2\) represent a subset of the plane that is perpendicular to both the ground and a line from home plate to the pitcher’s mound. More specifically, we choose the coordinate axes so that \(x\) runs across the back tip of home plate, \(y\) goes from the pitcher’s mound to home plate, and \(z\) measures the vertical distance off the ground. The origin of this coordinate system is thus the back tip of home plate, on the ground. In this case \(D\) is a plane in the \(x, z\)-directions defined by \(y = 0\). Let \(B\) be the set of all batters. Then for any \(b \in B\), let \(\alpha(b)\) be the height in inches of the hollow beneath the batter’s kneecap, and let \(\beta(b)\) be the height in inches of the midpoint between the tops of the batter’s shoulders and the top of his uniform pants. Batter \(b\)’s strike zone is then defined as the set \(\{(x, z) \in D : |x| \leq 8.5, \alpha(b) \leq z \leq \beta(b)\}\), with \(x\) and \(z\) measured in inches. Recall that home plate is 17 inches wide, as well as being 17 inches from front to back (for details see Commissioner 2008).

In this paper, we use \(s_f\) as our locations \(s_i \in D\), and model the hitting ability as a function of \(s \in D\). We restrict our data set to include only fastballs that were put into play. The fastball is the only pitch that every pitcher throws, and tends to be thrown with much better accuracy. That is, for most pitchers, his ability to throw a pitch in a specific location is profoundly greater if that pitch is a fastball. While our methodology could be extended to model the rate at which a batter makes contact, we chose to focus on the expected run value of each pitch as a function of pitch location, and excluded pitches that were not put into play.

Regarding quantification of batting ability, there has been a great deal written about which metric is most appropriate (see Albert and Bennett 2003). Summarizing this body of literature is beyond the scope of this paper, but in our research we considered three prominent metrics, each having its own advantages and disadvantages:

\(^1\)Since the pitcher must maintain contact with the rubber (which is 60.5 feet from home plate) until the ball is released, every pitch will have certainly been released by the time it reaches 50 feet from home plate.
Event Code | Event Type | Run Value
---|---|---
HR | Home Run | 1.44
3B | Triple | 1.04
2B | Double | 0.72
1B | Single | 0.50
SF | Sacrifice Fly | 0.37
SH | Sacrifice Hit | 0.04
Out | Field Out | -0.09
GIDP | Grounded Into Double Play | -0.37

Table 1: Run Values for eXtrapolated Runs

1. Batting Average (AVG): measures hit frequency, but not the magnitude of the hit. While batting average is not considered to be an accurate measure of overall offensive production, it may be appropriate for some purposes. For example, there may be certain situations near the end of a close game in which any hit must be avoided, regardless of the type of hit.

2. Slugging Percentage (SLG): measures power, but not on the scale of runs. Slugging percentage conveys more information than batting average, since extra-base hits (such as doubles, triples, and home runs) are given greater weight. Indeed, slugging percentage can be viewed as a weighted batting average. Again, slugging percentage may be appropriate in certain situations, but in general the weights ascribed to the different types of hits have no direct interpretation in terms of runs.

3. eXtrapolated Runs (XR): measures expected runs. XR belongs to a larger family of linear weights formulas, all of which map batting events to average run values (see Albert 2003 for a fuller description of linear weights formulas). The weights in XR are derived from data analysis, rather than posited (as with SLG), and are measured on the scale of runs, which has an inherent meaning.

For our purposes, we chose eXtrapolated Runs (XR) to be an appropriate metric, as it translates the result of each batted ball into a number representing the expected run value of that event (see Furtado 1999). The resulting mapping is shown in Table 1. In our work, the spatial field $Z$ of a batter’s hitting ability is measured on the scale of runs, which is the central currency in baseball. A sample of PITCHf/x data showing the resulting mapping made by XR is listed in Table 2.

### 3. Statistical Modeling

For a fixed batter, let $Z(s)$, $s \in D \subset \mathbb{R}^2$ be the random field of the batter’s hitting ability, measured in runs. We observe $Z(s_1), \ldots, Z(s_n)$, where $s_i = (x_i, z_i)$, $i = 1, \ldots, n$ are spatial locations around the strike zone, and want to estimate $Z_0 = Z(s_0)$ at locations $s_0 \in D$ where there are no observations. Suppose that $Z$ has mean $\mu(s) = E[Z(s)]$ and covariance $C(s_i, s_j) = \text{cov}(Z(s_i), Z(s_j))$, $i, j = 1, \ldots, n$. Denote by $\Sigma$ the $n \times n$ matrix whose $(i,j)^{th}$ element is $C(s_i, s_j)$, and let $\sigma_0 = (C(s_i, s_0))$, $i = 1, \ldots, n$, be the $n$-vector of covariances between the data and the target. A natural estimator of $Z_0$ can be obtained by taking a linear combination of the observations,
### Table 2: A sample of typical PITCHf/x data for Derek Jeter

<table>
<thead>
<tr>
<th>i</th>
<th>x(s)</th>
<th>z(s)</th>
<th>Result</th>
<th>Z(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−0.709</td>
<td>2.341</td>
<td>single</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>0.204</td>
<td>2.439</td>
<td>single</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>0.755</td>
<td>2.225</td>
<td>field out</td>
<td>−0.09</td>
</tr>
<tr>
<td>4</td>
<td>2.870</td>
<td>3.047</td>
<td>field out</td>
<td>−0.09</td>
</tr>
<tr>
<td>5</td>
<td>0.326</td>
<td>3.183</td>
<td>triple</td>
<td>1.04</td>
</tr>
<tr>
<td>6</td>
<td>2.701</td>
<td>2.911</td>
<td>field out</td>
<td>−0.09</td>
</tr>
<tr>
<td>7</td>
<td>0.121</td>
<td>2.570</td>
<td>field out</td>
<td>−0.09</td>
</tr>
<tr>
<td>8</td>
<td>0.442</td>
<td>1.696</td>
<td>field out</td>
<td>−0.09</td>
</tr>
<tr>
<td>9</td>
<td>1.363</td>
<td>2.886</td>
<td>field out</td>
<td>−0.09</td>
</tr>
</tbody>
</table>

The problem then becomes how to choose the weights \( \lambda_i \) in (1) optimally, and under which assumptions on the observed process. In geostatistics, it is usually assumed that the field of interest is second-order (or weakly) stationary, involving only assumptions on the first two moments: constant mean \( \mathbb{E}[Z(s)] = \mu \), and covariance function only depending on the spatial lag (separation) \( s_1 - s_2 \) and not on direction, \( \text{Cov}(Z(s_1), Z(s_2)) = C(s_1 - s_2) \). In this setting, it is customary to express the covariance function in terms of the lag vector, \( C(h) = \text{Cov}(Z(s), Z(s+h)) \). In many practical applications, it is more realistic to consider intrinsic stationary random fields, which is a larger class of processes (Cressie 1993, Section 2.5.2). It is only assumed that the increments \( Z(s) - Z(s+h) \) are second-order stationary. For intrinsically stationary processes, the variogram \( \gamma : \mathbb{R}^2 \rightarrow \mathbb{R} \) is defined by \( \gamma(h) = \frac{1}{2} \text{Var}[Z(s)-Z(s+h)] \), and is connected to the covariance through the relation \( \gamma(h) = C(0) - C(h) \). When the covariance function or the semivariogram depends only on the absolute distance between points, the function is termed isotropic. In this case \( C(h) = C(||h||) \), and \( \gamma(h) = \gamma(||h||) \), where \( || \cdot || \) denotes the Euclidean norm. The weights \( \lambda_i \) in (1) are uniquely determined by either \( \Sigma \) and \( \sigma_0 \), or their variogram counterparts, yielding the best linear unbiased predictor (BLUP) for \( Z(s_0) \). For the derivation of the solution of universal kriging system we refer to Chilès and Delfiner 1999, Section 3.4. The variance of this BLUP can be also determined in closed form, and it depends on the parametrization of the process as well.

In practice, the covariance function \( C \) or the variogram \( \gamma \) are chosen based on the observed patterns in the data. Figure 3 displays several empirical variograms (Matheron 1962), \( \hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} |Z(s_i) - Z(s_j)|^2 \), where \( |N(h)| \) denotes the number of spatial locations with \( ||s_i - s_j|| \leq h \). It can be seen that different players display different variogram patterns, and thus we expect that this approach might not produce reliable maps for all players. In addition, it would be quite hard to give a reasonable justification for all the aforementioned assumptions, in particular for constant mean and isotropy.

In contrast to the geostatistical approach, a nonparametric setting is more flexible, and relies on fewer assumptions. In particular, we do not need to specify the mean and covariance functions. Commonly used methods in nonparametric mod-
Figure 3: Empirical variograms for Josh Thole (a), David Wright (b), Derek Jeter (c), Carlos Beltran - R (d), Fernando Tatis (e), Marcus Thames (f), Mark Teixeira - L (g), and Jose Reyes - R (h).
eling are based on local least squares, smoothing splines, orthogonal series, and kernel smoothing. All of these techniques produce localized weighted averages of the data, and different estimators differ only with respect to the weight functions, but asymptotically have similar properties.

Among these many smoothing techniques, one of the most popular is that of kernel smoothing, introduced by Rosenblatt (1956) for estimating density functions of independent, identically distributed data, and widely applied since, more recently to dependent data as well. A kernel estimator can be viewed as the convolution of a smooth, known function (the kernel) with a rough empirical estimator, so as to produce a smooth estimator. Thus, we can still predict $Z(s_0)$ as a linear combination of the observations (1), only assign the weights through a smooth kernel function

$$\lambda_i = \frac{K \left( \frac{s_0 - s_i}{b} \right)}{\sum_{j=1}^{n} K \left( \frac{s_0 - s_j}{b} \right)}.$$ 

This is an adaptation of the well-known Nadaraya-Watson estimator (Nadaraya 1964, Watson 1964) to the two-dimensional case. Here the kernel $K$ is a symmetric bivariate density function with compact support, such that $\lim_{||u|| \to \infty} ||u||^2 K(u) = 0$, and $\int_{\mathbb{R}^2} ||u||^2 K(u) < \infty$. The parameter $b = b(n)$, called the bandwidth, is a sequence of positive real numbers satisfying $b \to 0$ as $n \to \infty$. It measures the size of the window around the target point, thus controlling the amount of smoothing. The question becomes then under which assumptions this smoothed estimator has minimum mean squared error and optimal rate of convergence. For details on kernel smoothing we refer to Simonoff (1996). In the next section we show applications of both methods described above, followed by comments and open questions in Section 5.

4. Applications

For the applications in this section we used the fields package in R (Furrer et al. 2009) to compute estimates for the batting ability field $Z(s)$. For the kriged maps we used the isotropic exponential covariance model, $C(h) = \sigma e^{-\alpha ||h||}$, where $\sigma$ is the variance of the field $Z$, and the parameter $\alpha$ measures how fast the covariance decays with distance.

In Figure 4, we see side-by-side hot zone maps for Derek Jeter$^2$. On the left is a hot zone generated via kriging, while on the right is a hot zone generated via kernel smoothing. In both maps, we can see that Jeter hits the ball best when it is up and out over the plate. The small red circles on the plot indicate home runs hit by Jeter, and it is informative to note that nearly all of them occur in the upper third of the strike zone. What is interesting to note, and seems to be typical of the differences between the two types of maps, is that the kernel smoothed image tends to reveal circular regions of similar ability (like a topographic map), while the kriging image tends to show a more diagonal pattern. We attribute this difference to the stationarity restriction in the kriging model.

The maps for Carlos Beltran as a right-handed hitter, displayed in Figure 5, show this in stark contrast, and further, may be considered a failure of the kriging method. The kernel smoothed image conveys a nuanced pattern, wherein Beltran has much greater success on pitches on the inside part of the strike zone, but not nearly as much with pitches on the outside part of the plate. While the kriging

$^2$All maps shown in this paper are from the pitcher’s perspective.
image does show that same general trend, it captures none of the nuances. Instead, in suggests a uniformity that is clearly unrealistic. It is exactly in this sense that the kriging method lacks robustness.

In Fernando Tatis’s hot zone maps, shown in Figure 6, we see a different type of failure. Note how the kernel smoothed map shows a pocket of concentrated ability on the outside corner of the strike zone, while the kriging image interprets this as a more general drift towards the upper left hand corner of the plot. It is clearly unrealistic to infer from the data that Tatis hits the ball best when it is a foot outside and at his shoulders, as the deep red corner in the kriging image would indicate.

Figure 7 shows that there are players for which both maps largely agree. For areas within the strike zone, in particular, there is considerable agreement between the two maps for Marcus Thames. At this point, it is also worth reminding the reader of what neither map conveys. Thames’s map shows that he is a dangerous hitter, particularly on pitches up and away (Thames bats right-handed). However,
Figure 6: Fernando Tatis: kriging (left), kernel smoothed (right) hotzone maps.

Figure 7: Marcus Thames: kriging (left), kernel smoothed (right) hotzone maps.

since these maps only show balls that were put into play, they do not convey the frequency with which Thames swings and misses, which is very high. We show only Thames’s production, given that he put the ball in play. A more comprehensive scouting report should also include a measure of how often this occurs.

Figure 8 shows that the problem of extrapolating edge effects is not limited to kriging. Even though almost all of Mark Teixeira’s home runs while batting left-handed occur on pitches above the top half of the strike zone, both hot zones show area of deep red bleeding into the lower left-hand corner. It seems unlikely that Teixeira hits the ball best near his shoetops (though we concede it is possible).

Lastly, it does not seem to be the case that the kernel smoothed maps are always preferable. To the contrary, consider the maps for Jose Reyes as a right-handed hitter shown in Figure 9. While both show that Reyes is most dangerous on pitches on the inside edge of the strike zone, the kernel smoothed image shows three additional small pockets of great hitting ability, while the kriging image does not. It seems much more likely that the kernel smoothing technique is overly sensitive
to what were most likely aberrations, than to think that Reyes has tiny pockets of deep power. Clearly, our choice of bandwidth for the kernel smoother may affect the sensitivity to this type of aberration.

5. Concluding remarks

In this paper we use both parametric and nonparametric statistical techniques to model the hitting ability of batters in Major League Baseball, based on the spatial location of the pitches each batter puts into play. Our methods yield hot zones that are more natural and easier to interpret than the typical hot zones constructed using discrete data binning techniques. This is a preliminary study, intended to devise a fast and reliable exploratory tool to be used for advance scouting purposes.

There are a number of statistical issues that need to be further addressed, and which will make the subject of a follow-up research paper. In the geostatistical (parametric) approach, we mention in Section 3 that the variance of the BLUP
depends on the parametrization of the process. In practice, however, these spatial predictors are in fact EBLUP’s (empirical or estimated BLUP’s) since the mean and covariance parameters are estimated from the same data. To account for this added uncertainty, their standard errors need to be adjusted. This can be done by using conditional simulation (Chilès and Delfiner 1999, Chapter 7, Stein 1999, Chapter 6), or resampling (Lahiri 2003). While many contributions have been made for independent, identically distributed data, not much is understood about bootstrap schemes in the presence of complex dependence structures, even in the one-dimensional case, and thus new methodology needs to be developed for the spatial setting.

As for the kernel smoothing approach, an important practical aspect is the choice of tuning parameter (bandwidth). While cross-validation is a common way to select the size of the smoothing window, it may not be appropriate in the presence of spatial dependence (by using the leave-one-out principle, the dependence structure of the process may change, thus leading to inaccurate results). Another selection criterion could be to minimize the mean squared error of the spatial predictor. However, a closed form for the optimal smoothing parameter is in general very hard, if not impossible to obtain while making only minimal assumptions on the field of interest. No distributional specifications are needed, only assumptions on the smoothness of the field (continuous second derivatives), based on which Taylor expansions are used to approximate the bias and variance of the resulting predictor. As mentioned in the previous section, another practical point has to do with edge effects. The non-parametric approach could better address this aspect, one common way is to use modified boundary kernels with asymmetric support (Müller 1991). However, adaptation to the multidimensional case is needed as well.

References


