The Quality of Pitches in Major League Baseball

Philippa Swartz, Mike Grosskopf, Derek Bingham and Tim B. Swartz *

Abstract

This paper considers the quality of pitches in Major League Baseball (MLB). Based on approximately 2.2 million pitches taken from the 2013, 2014 and 2015 MLB seasons, the quality of a particular pitch is evaluated as the expected number of bases conceded. Quality is expressed as a function of various covariates including pitch count, pitch location, pitch type and pitch speed. The estimation of pitch quality is obtained through the use of random forest methodology to accommodate the inherent complexity of the relationship between pitch quality and the associated covariates. With the fitted model, various applications are considered which provide new insights on pitching and batting.

Keywords: Machine learning, PITCHf/x data, Random forests.

*Philippa Swartz is an undergraduate student, Mike Grosskopf is a PhD candidate, and Derek Bingham and Tim Swartz are Professors, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. P. Swartz was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a USRA (Undergraduate Student Research Assistantship). Bingham and Swartz have also been supported by NSERC. The authors thank two anonymous reviewers whose comments have helped improve the manuscript.

1 INTRODUCTION

The purpose of this investigation is the assessment of the quality of pitches thrown in Major League Baseball (MLB). An initial reaction may be that this is a straightforward problem. For example, if a pitch results in a *home run*, then it was a bad pitch. Alternatively, if a pitch leads directly to an *out*, then it was a good pitch.

However, sometimes home runs occur on good pitches and sometimes outs occur on bad pitches. We therefore want to assess the quality of a pitch without referring to its direct outcome.

If we are able to assess the quality of pitches, then a host of practical questions may be addressed. For example, can we detect when a pitcher's performance begins to deteriorate prior to bad results? Can we determine *pitch count* rules for removing pitchers from a game? There seems to a general rule of thumb of a limit of 100 pitches; is this sensible? Should pitch count rules be pitcher dependent? Does warming up between innings contribute to tiredness and pitch count limits? In a typical game, a manager monitors a pitcher's performance. When the pitcher's *control* deteriorates or when the speed of his *fastball* diminishes according to the assessment of the manager, these are signs that the pitcher is tiring and might be pulled. Hopefully, the decision is made before a rash of runs occurs.

Most of the research related to the assessment of pitch quality has not been published in peer reviewed journals. Rather, some work has been carried out confidentially by analytics companies or by MLB teams who hope to gain a competitive edge. Other work has been disseminated through websites and blogs. For example, Wilson (2015) used experts to assess the quality of pitches QOP on a scale of -10 to 10. A fitted linear regression equation was obtained to estimate QOP for future pitches based on observed covariates. Wilson (2015) used QOP to assess pitchers and compared the approach with conventional pitching measures. Rosales and Spratt (2015) developed a methodology known as Strike Zone Plus/Minus which measures a catcher's ability to "frame" a pitch as a *strike* that would otherwise been called a *ball*. Roegele (2013a, 2013b) divided the strike zone into 9 regions and averaged the outcomes of strike zone pitches stratified according to handedness and pitch type. Using these averages as benchmarks for pitch quality, various pitchers were compared together with their fastball velocities. The methods proposed by Moore (2009) also attempt to measure pitch quality. Although the details are not fully explained, Moore (2009) looked specifically at fastballs thrown by right-handed pitchers to right-handed batters and used regression methods to investigate the quality of various types of fastballs thrown by various pitchers. On a scale of 0-100, a variable referred to as the "Nasty Factor" has been introduced (www.mlb.com/mlb/gameday/y2010) that attempts to assess the

difficulty of hitting a pitch of a given type. Although the nasty variable depends on pitch velocity, pitch sequence, pitch location and movement, the exact details of its calculation do not appear to be publicly available.

This paper makes use of PITCHf/x technology which provides detailed information on pitches thrown in MLB. Although research based on PITCHf/x has been mainly confined to blogs and online sites such as "The Hardball Times" (www.hardballtimes.com), there is little doubt that the analysis of PITCHf/x data will soon become standard fare in academic sports journals (Albert, 2002 and Baumer and Zimbalist, 2014).

In Section 2, we describe the data as obtained through PITCHf/x technology. The PITCHf/x database is large but is manageable due to various software initiatives. In Section 3, we propose a simple statistical model which relates the quality of a pitch to various covariates. Pitch quality (the dependent variable) is defined as the number of bases conceded from a given pitch. The pitch is characterized according to one of 12 states corresponding to the count and a vector referred to as the pitch descriptor. An expression is provided for the expected number of bases conceded by a pitch. In Section 4, random forest methodology is utilized to predict the expected number of bases by a future pitch. An advantage of random forests over some of the regression type procedures previously mentioned is that it allows us to entertain many possible covariates without specifying a specific functional relationship between the response and the covariates. As will be seen, there are many potential covariates that are made available through PITCHf/x. For posthoc analyses, the methodology can be used to assign an expected number of bases to pitches that already form part of the database. In Section 5, various applications are considered. We investigate how a typical pitcher's performance changes over the course of a game. We also produce some heatmaps that investigate the importance of pitch location. Finally, we use pitch quality as a metric for evaluating pitchers. We conclude with a short discussion in Section 6.

For those not particularly interested in the game of baseball, our work can be viewed as a case study with various lessons. For example, in this study and with the advent of recording technologies in sport generally, sporting datasets are often extensive and complex, and have unique features that are worthy of the attention of statisticians. In terms of statistical practice, this paper demonstrates how more advanced statistical tools (i.e. random forests) can be utilized in sports analytics, an area which relies heavily on simple descriptive statistics. Finally, a major challenge in any statistical application is to convey results clearly to practitioners. We believe that this study is successful in this regard by introducing an accessible response variable (total bases) and displaying it via simple plots, heat maps and ranked lists.

2 DATA

The data that we use in this investigation arise from PITCHf/x technology. The PITCHf/x service is a product of Sportvision and is comprised of three cameras installed in every MLB stadium that are able to track every pitch. Through these cameras, PITCHf/x is able to gather a massive amount of information. The information is organized into five tables: pitch, at-bat, runner, action and po (pickoffs). Within each of these tables are variables that provide information on the pitch, batsman, runners, and more. Some variables overlap between tables, where the pitch and at-bat tables are by far the largest. The at-bat and pitch tables contain all of the variables relevant to our investigation. All of the data is available on the Gameday website (http://www.mlb.com/mlb/gameday/) through XML files. At the Gameday website, information on MLB games are conveniently displayed including popular information such as boxscores. In addition, the Gameday server stores XML files that are accessed by entering specific urls. Since we need huge amounts of data for our investigation, we chose to download the data and create a database to store our PITCHf/x data. Fast (2009) provides more detailed information on PITCHf/x technology.

To download this data, we utilized the R package `pitchRx` (Sievert 2014, 2015) which facilitates the use of PITCHf/x data within the statistical software package R. Within `pitchRx` there is a function *scrape* which pulls data (within a timeframe of your choosing) from Gameday onto your computer. We utilized the *scrape* function to pull all PITCHf/x data from the 2013, 2014 and 2015 MLB seasons, and then stored the data in an SQL database. To access the stored data, it was also necessary to install `RSQLite` (Wickham, James and Falcon 2014) and `dplyr` (Wickham and Francois 2015). Once this data was downloaded and stored in a database, we merged the at-bat table with the pitch table by matching the "url" and "num" variables so that all of the data were stored in one large matrix. We then ordered each pitch by the variables "url", "num", and "id" so that the pitches were in chronological order. The last step was to create a new variable "Pitchcount" so that we could keep track of the number of pitches thrown by a pitcher within a single game. The pitch count is distinct from the count a batter faces within an at-bat. We retain the terminology to be consistent with common nomenclature. The resulting data is a matrix comprised of 2,170,035 observations (pitches) and 74 corresponding variables which describe aspects of the pitch such as count, strike zone, pitch speed, batter name, pitcher name, game url, inning, spin, pitch type, whether the batter is right or left handed, outcome of the at-bat (e.g. out, single), outcome of the pitch (e.g. ball, in play, run), etc.

3 APPROACH

It is clear that the quality of a pitch depends on various underlying conditions. For example, when the count is three balls and two strikes (i.e. a 3-2 count), throwing a ball on the next pitch has a greater negative consequence to the pitcher than throwing a ball on a 0-0 count. In the former situation, if the batter does not swing, then a *walk* is awarded to the batter. In assessing the quality of a pitch, we therefore introduce the count variable C which has the following 12 possible states preceding the pitch

$$\begin{array}{lllll} C: & \text{0-0} & \text{1-0} & \text{2-0} & \text{3-0} \\ & \text{0-1} & \text{1-1} & \text{2-1} & \text{3-1} \\ & \text{0-2} & \text{1-2} & \text{2-2} & \text{3-2} \end{array} .$$

The quality of a pitch also depends on the variable

$$D: \quad \text{pitch descriptor}$$

where D may include measurements related to the pitch location, pitch speed, pitch type, handedness of pitcher, handedness of batter, etc. One may also hypothesize that the batter himself should be a component of the pitch descriptor. However, we take the view that a good pitch is a good pitch regardless of the batter. This belief is supported by Cross and Sylvan (2014, 2015) who argue that sweet/hot zones (areas where batters bat well) look similar across batters. This somewhat counterintuitive result is also supported by Fast (2011) who argues that it is difficult to identify hot and cold zones for individual batters. The introduction of hitter-specific covariates is problematic due to small sample sizes but may be something that could be considered in future work.

We now define T_{CD} as the number of bases yielded corresponding to a pitch with count C and pitch descriptor D. Therefore

$$T_{CD} = \begin{cases} 0 & \text{out} \\ 1 & \text{nonintentional walk or single} \\ 2 & \text{double} \\ 3 & \text{triple} \\ 4 & \text{home run} \\ a(C) & \text{ball but not walk} \\ b(C) & \text{strike but not out} \end{cases} \tag{1}$$

where we ignore all other possibilities such as sacrifice flies, errors, intentional walks, etc. When the count of an at-bat is C, the number of effective bases corresponding to throwing a ball $a(C)$ and the number of effective bases corresponding to throwing a strike $b(C)$ are discussed and estimated in the next section. The quantities $a(C)$ and $b(C)$ are contextual; the merits of throwing a ball or a strike depend on the count C. Also, although there is no immediate batting outcome corresponding to a ball or a strike, a batting outcomes eventually occurs, and this is the basis for evaluating $a(C)$ and $b(C)$.

As we have previously argued, pitch quality should not be measured directly on the results defined in (1). Rather, pitch quality is an expected value that is dependent on both the count C and the pitch descriptor D. Therefore, the expected number of bases yielded is given by

$$\begin{aligned} \mathrm{E}(T_{CD}) &= 1 * \mathrm{Prob}(T_{CD} = 1) \\ &+ 2 * \mathrm{Prob}(T_{CD} = 2) \\ &+ 3 * \mathrm{Prob}(T_{CD} = 3) \\ &+ 4 * \mathrm{Prob}(T_{CD} = 4) \\ &+ a(C) * \mathrm{Prob}(T_{CD} = a(C)) \\ &+ b(C) * \mathrm{Prob}(T_{CD} = b(C)) \end{aligned} \tag{2}$$

when the count is C corresponding to a pitch with descriptor D. We have $\mathrm{E}(T_{CD}) \geq 0$ and we note that larger values of $\mathrm{E}(T_{CD})$ denote pitches of lesser quality. We emphasize that there is a discrete probability distribution corresponding to the random variable T_{CD} that depends on both the pitch count C and the pitch descriptor D. In the next section, our goal is to estimate $\mathrm{E}(T_{CD})$ in (2).

Before proceeding, it is worth considering the response variable T_{CD} in (1). Rather than using total bases, there is considerable interest in the sabermetrics community to consider runs. Runs scored are obviously important in baseball as scoring runs leads directly to winning. For example, the commonly cited WAR statistic (wins above replacement) is based on run calculations for both position players and for pitchers. An issue with measuring runs directly is that the measurement requires more context and a longer dependence structure. For example, a pitch leading to a triple yields 0 runs when no runners are on base but yields x runs when x runners are on base. An alternative approach to measuring runs directly involves the modification of the values 0, 1, 2, 3, 4, $a(C)$ and $b(C)$ in (1) with weights that better reflect the contribution of the batting outcome in terms of runs. We note that the wOBA (weighted on-base average) batting statistic contains weights assigned to batting outcomes where the weights correspond to run production. The

statistic wOBA was developed by Tango, Lichtman and Dolphin (2007).

4 ESTIMATION OF EXPECTED BASES

Following (2), the estimation of the expected number of bases $\mathrm{E}(T_{CD})$ first requires the determination of $a(C)$ and $b(C)$.

4.1 Estimation of $a(C)$ and $b(C)$

Recall that $a(C)$ is the effective number of bases yielded by throwing a ball when the count is C. We do not consider the counts 3-0, 3-1 and 3-2 since throwing a ball in any of these three states leads to a nonintentional walk which corresponds to $T_{CD} = 1$.

For illustration, consider the estimation of a(0-0). We use the entire history of MLB pitches thrown in the 2013-2015 seasons. We then restrict our attention to the N at-bats where the count went from 0-0 to 1-0. In these at-bats, we obtain the total number of bases B that are eventually yielded (i.e. from subsequent pitches in the at-bat). The estimate of a(0-0) is B/N.

In Table 1, we provide the estimates of $a(C)$. Note that there are physical constraints on the estimates such that entries above or to the left of a given estimate in the table should be larger than the estimate itself. For example, it is logical that a(1-0) $> a$(1-1) since the resultant count 2-0 is more in the batter's favor than the resultant count 2-1. Similarly, a(2-1) $> a$(1-1). We observe that all of the estimates in Table 1 satisfy the logical constraints. This is due to the enormity of the dataset.

a(2-0) = 0.79 (22669)	a(1-0) = 0.61 (74564)	a(0-0) = 0.51 (222363)
a(2-1) = 0.69 (32469)	a(1-1) = 0.51 (76907)	a(0-1) = 0.42 (116315)
a(2-2) = 0.55 (50985)	a(1-2) = 0.39 (77336)	a(0-2) = 0.32 (63841)

Table 1: Estimates of $a(C)$ where the entries in parentheses are the numbers of at-bats used in the calculation.

The estimates of $b(C)$ are calculated in a similar fashion to the estimates of $a(C)$. We note that the counts 3-2, 2-2, 1-2 and 0-2 remain the same when a fouled strike occurs. We disregard these states since they are redundant with other states. For example, the effect of throwing a strike on a 3-1 count is the same as the effect of a fouled strike on a 3-2 count. Therefore the

estimate of b(3-1) is used as the estimate of b(3-2). Table 2 provides the estimates of $b(C)$ where again we have the same constraint structure where entries above or to the left of a given estimate in the table should be larger than the estimate. We observe that all of the estimates in Table 2 satisfy the logical constraints.

b(3-0) = 0.68 (14303)	b(2-0) = 0.51 (37640)	b(1-0) = 0.42 (108469)	b(0-0) = 0.37 (281951)
b(3-1) = 0.54 (22018)	b(2-1) = 0.38 (52948)	b(1-1) = 0.31 (98522)	b(0-1) = 0.28 (114678)

Table 2: Estimates of $b(C)$ where the entries in parentheses are the numbers of at-bats used in the calculation.

There is one further set of constraints involving the estimates of $a(C)$ and $b(C)$. Logically, the effect of throwing a ball on a 0-1 count should be the same as the effect of throwing a strike on a 1-0 count since in both cases, the count advances to 1-1. Considering all of the states, we should have six equalities a(0-1) = b(1-0), a(1-1) = b(2-0), a(2-1) = b(3-0), a(0-2) = b(1-1), a(1-2) = b(2-1) and a(2-2) = b(3-1). Although these equalities do not hold exactly between Table 1 and Table 2, the pairs of estimates are all within 0.01 tolerance. For our final set of estimates of $a(C)$ and $b(C)$, we use weighted averages to combine the relevant pairs. Therefore, instead of the $9 + 8 = 17$ estimates that are presented in Table 1 and Table 2, we have only $17 - 6 = 11$ estimates in total.

4.2 Estimation of $\mathrm{E}(T_{CD})$ using Random Forests

Having estimated $a(C)$ and $b(C)$, perhaps the first inclination is to estimate $\mathrm{E}(T_{CD})$ using standard regression methodology. In regression, the observed values T_{CD} are regressed against the covariates C and D. In this application, the form of the parametric relationship between the response T_{CD} and the covariates is unknown. For example, when considering the horizontal pitch location variable, it is intuitive that $\mathrm{E}(T_{CD})$ is maximized for pitches in the middle of the plate. However, it is unclear how the relationship changes as pitches extend from the central region and beyond the strike zone. Pitch count is another variable which is likely nonlinear. Another challenge with regression methodology is the huge number of potential covariates and possible interaction terms. For example, suppose that we have the pitch descriptor D coarsely discretized on the following subset of covariates:

- 13 regions for pitch location (the strike zone divided into 9 rectangles and four regions corresponding to pitches that are outside, inside, high and low)

- six intervals for pitch speed
- 8 types of pitches
- four types of handedness (both batter and pitcher) .

Together with the 12 states corresponding to the count C, this leads to $13(6)(8)(4)(12) \approx 30,000$ parameters. With roughly 2.2 million pitches during the 2013-2015 MLB seasons, this leaves on average only $2200000/30000 \approx 73$ data values per parameter. Such sparsity often leads to unreliable parameter estimation. On the other hand, one of the attractive features of standard regression models is interpretability. With a fitted regression model, one can readily ascertain the change in $\mathrm{E}(T_{CD})$ by modifying a covariate value.

Due to some of the above difficulties with standard regression procedures, we have taken a regression tree approach where "important" covariates are determined by the splits in the tree. The approach is attractive when we do not know in advance which variables (e.g. pitch location, pitch speed, pitch type, handedness) are predictive. With regression trees, there is also no need to arbitrarily discretize variables that are naturally continuous (e.g. the x and y coordinates corresponding to pitch location). Regression tree methodology consists of recursively partitioning the covariate space by finding the value of a single covariate that will optimally reduce the training MSE when the split is applied and the sample means are calculated for each partition. The terminal nodes, also called the leaves of a resultant tree provide estimates of $\mathrm{E}(T_{CD})$ whereby future pitches with the same characteristics (i.e. splits) as a particular leaf are predicted to have the corresponding value. Regression tree methodology is discussed in the texts by Hastie, Tibshirani and Friedman (2009) and James et al. (2014). The partitioning procedure allows for complex response surfaces, including those with discontinuities, to be estimated.

Single regression trees are known to be very sensitive to small differences in covariate importance, especially when making splits early in the tree. Therefore, we do not base our predictions on a single tree but instead use random forest methodology where many trees are utilized and predictions consist of averages over the individual trees. The details of the random forest methodology can be found in the Hastie, Tibshirani, and Friedman (2009) and James et al. (2014) references as well. Random forests are known to provide improved predictions compared to single trees. In our implementation, we use the *randomForest* function from the `randomForest` package in R (Liaw and Wiener 2002). The random forest procedure has various parameters for tuning the method to optimize predictive performance. For the application discussed later in this paper, we trained the random forest on a simple random sample of $n = 100,000$ pitches from the data set. We then used the remaining observations as a validation set for choosing the

tuning parameters. The optimal predictive performance was found using 5,000 trees, each with a maximum of 250 nodes. At each node, the best split was found by searching over a random size 4 subset of the total covariates. This random subsetting causes the individual trees in the random forest to be less correlated and allows the model to identify subtle effects of covariates that might otherwise be missed by individual regression trees.

One of the challenges of random forests is that their complexity does not lead to straightforward interpretation. In particular, it is difficult to view the entire random forest. In this investigation, it is instructive to know which covariates are the most important in determining pitch quality. We measure the importance of covariates using a variance decomposition approach from Saltelli et al. (2008) that is standard in the sensitivity analysis literature. The intuition is that the predicted pitch quality from the random forest will vary more for inputs that are important. The total variability of predicted pitch quality can be broken down based on the changes due to individual inputs (main effects), changes in pairs of inputs (two-way interaction effects), and so on. We can also obtain total effects, which indicate the importance of a variable through it's main effect and all interactions involving that variable. The variability for each effect is divided by the total variability to give a percentage attributed to that input. The total effect percentage can exceed 100% because the variability due to interactions is shared between inputs. In Table 3, we provide a list of the covariates in the random forest ordered from the most important to the least important based on their main effects. We see that the count variable C is the most important. This corresponds to baseball tradition where pitchers are careful when they are ahead of hitters and need to throw strikes when they are behind in the count. Also, corresponding to our intuition, the location variable for the pitch (both horizontal and vertical) plays an important role. Although the covariates speed, Pitchcount and pitch type show weaker effects, they are instrumental in the overall predictive capability of the random forest. The two-way interactions between count and both horizontal and vertical locations each contributed about 9.2% to the prediction variance, indicating that the relationship between location and pitch quality depends strongly on the count. No other two-way interactions were as important.

5 APPLICATIONS

In this section, we investigate various applications that arise from an assessment of expected bases $\mathrm{E}(T_{CD})$. We remind ourselves that for any pitch that has occurred in the dataset, its characteristics (C, D) yield a value $\mathrm{E}(T_{CD})$ from the random forest.

Covariate	Main Effect	Total Effect
count C	69.5%	93.5%
horizontal location	2.8%	15.1%
vertical location	1.3%	13.5%
speed	0.5%	5.4%
Pitchcount	0.5%	2.4%
pitch type	0.5%	2.0%

Table 3: Relative importance of the six covariates in determining pitch quality listed from the most important to the least important.

5.1 Change in Pitching Performance Throughout a Game

Over the years, there has been a stunning decrease in the number of complete games in MLB (i.e. games where a starting pitcher is not relieved). For example, the percentage of complete games in the 1984, 1994, 2004 and 2014 MLB seasons has decreased according to 15.0%, 8.0%, 3.1% and 2.4%, respectively. Part of this phenomenom can be attributed to the fear and prevention of injury to pitchers, and the subsequent adherence to pitch count maxima. Another reason is that there has been a growing specialization with respect to relief pitchers where the notion is that a relief pitcher can provide better performance than an overworked or overexposed starting pitcher.

To investigate pitching performance over a game, we considered the 12,760 occasions during the 2013-2015 MLB seasons where starting pitchers threw at least 80 pitches in a game. From these pitching performances, a subset of 6,033 starting pitchers continued in their games and threw at least 100 pitches. In Figure 1, we provide a plot of $\mathrm{E}(T_{CD})$ for each pitch number in a game averaged over all pitching occasions. For example, the average number of expected bases eventually yielded in the at-bat from the first pitch of a game is 0.412. From Figure 1, we observe an interesting trough where the best pitching tends to occur roughly between the 20th and the 70th pitches. Pitching performance tends to deteriorate beyond the 70th pitch, and even more quickly beyond the 100th pitch. However, what is also interesting is that pitchers do not seem to start games quite as sharply as one might expect. It may be the case that pitchers need to "find their groove" early in games before settling into a natural rhythm. Warming up in the *bullpen* prior to a game is not the same as pitching in an actual game. We also note that it is typically good batters (i.e. at the top of the batting order) that face the first 20 pitches. Perhaps these batters cause pitchers to be overly cautious and they do not pitch as well in these circumstances.

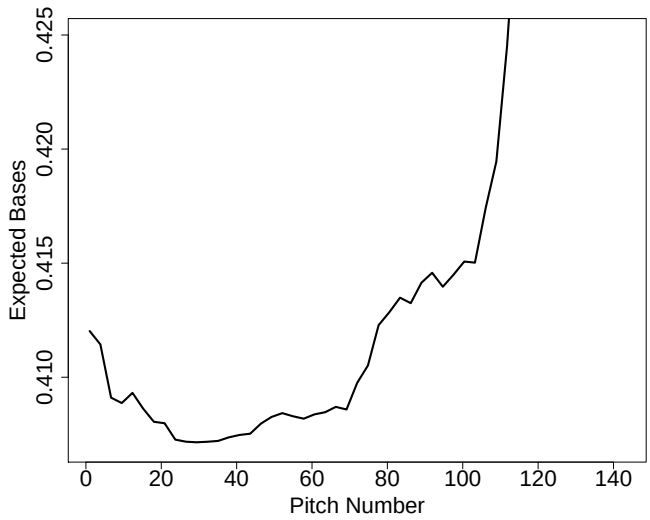

Figure 1: Expected bases versus pitch number based on all pitching occasions during the 2013-2015 seasons where pitchers threw at least 80 pitches in a game.

5.2 Heatmaps based on Pitch Location

In Section 1, we mentioned that the variable "nasty" has been proposed to assess the quality of pitches and that nasty is included as part of the PITCHf/x dataset. In Figure 2, we provide a heatmap of the nasty variable with respect to pitch location where only right-handed batters were considered. The data forms 32.6% of the observations from the 2013 MLB dataset. The heatmap (quiltplot) and the subsequent figures were produced using the `fields` package in R (Nycha, Furrer and Sain 2015). What is most prominent in Figure 2 is the symmetry corresponding to nasty. Clearly, the developers of nasty imposed constraints on the calculation of nasty and did not allow the data to fully inform pitch quality. For example, there is a strong delineation of the strike zone. Nasty makes it impossible for a pitch that is only slightly outside the strike zone to be considered a good pitch. Pitches that are inside but near the four corners of the strike zone are deemed the most difficult for batters to hit.

By contrast, in Figure 3, we provide a plot of expected bases $\mathrm{E}(T_{CD})$ for right-handed batters. Average values are obtained for each pixel using pitches from the 2013-2015 seasons. The colors in Figure 3 do not directly correspond to those in Figure 2 because of the different scales in which $\mathrm{E}(T_{CD})$ and nasty are calibrated. Here, the data tell us something different. The most difficult pitches to hit are those that are low and away. Also, outside pitches are generally more difficult to hit than inside pitches. We note that similar features were observed when restricting the data to left-handed batters. As expected, pitches within the strike zone (the inner rectangle) are generally better than pitches outside the strike zone. High pitches (especially high and inside)

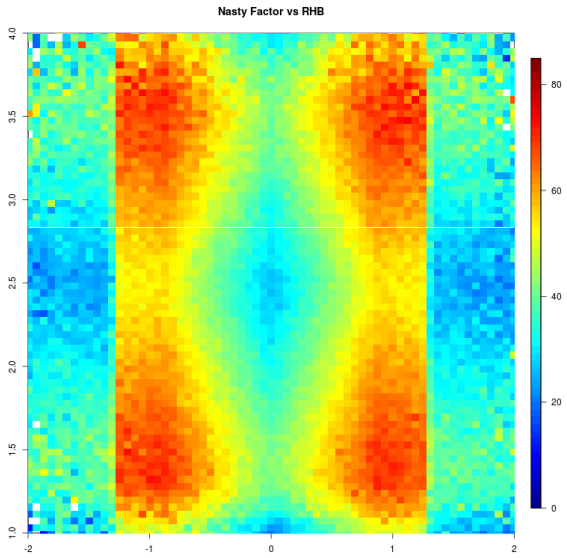

Figure 2: Heatmap of the nasty variable for right-handed batters as viewed behind the batter. The axes are labelled in feet.

are not very good from the pitcher's point of view; typically batters lay off these pitches and they are called balls.

In Figure 4, we stratify the data further by considering right-handed batters who faced left-handed pitchers. What is most interesting here is that pitches that are high and inside (but in the strike zone) are also difficult to hit. This may have something to do with the ball motion of a left-handed pitcher but a physical explanation is not obvious to us. In any case, this seems to be something that is not widely known. We remark that symmetric results where obtained when considering the matchup between left-handed batters who faced right-handed pitchers. We experimented further by stratifying the data on the type of pitch but did not find any noteworthy results.

5.3 Evaluation of Pitchers

The evaluation of pitchers is not straightforward. Perhaps the most commonly reported pitching statistic is the win-loss record. The win-loss record for a pitcher is simply the tabulation of the number of games that are credited to a pitcher as wins throughout the season versus the number of games that are credited as losses. However, it is well-known that the win-loss record is highly dependent on the performance of a pitcher's teammates. If a pitcher's team does not score many runs while he is pitching, it is unlikely that he will be credited with a win even if he is pitching well.

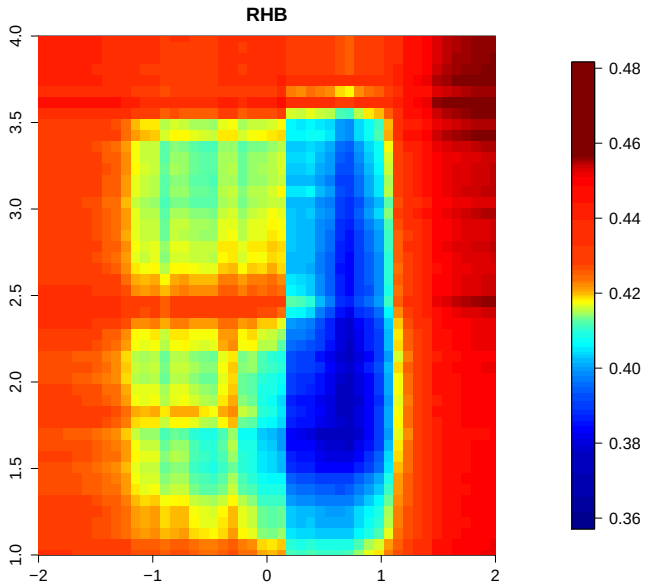

Figure 3: Heatmap of E(T_{CD}) for right-handed batters as viewed behind the batter.

Another commonly reported pitching statistic is the earned run average (ERA). The ERA is the number of earned runs that a pitcher allows over a season relative to 9 innings. Although ERA is a better pitching performance metric than the win-loss record, ERA is subject to considerable variability (Basco and Davies, 2010).

In an attempt to further evaluate pitching, the statistic "fielding independent pitching" (FIP) has been proposed which is a composition of variables that are directly controllable by the pitcher (i.e. home runs, walks, strikeouts and the number of times that batters are hit by pitches). FIP has gained some traction in the analytics literature and is believed to be a better measure of pitching ability (http://www.fangraphs.com/library/pitching/fip).

In this exercise, we compare E(T_{CD}) against some of the above pitching metrics for starting pitchers. We only considered the 248 starting pitchers who pitched at least 1000 pitches over the 2013-2015 MLB seasons. We then averaged E(T_{CD}) for each pitcher over all his pitches. For ERA and FIP, these values were also calculated over the entire 2013-2015 seasons and were obtained from www.fangraphs.com. The correlation between E(T_{CD}) and ERA for these pitchers is 0.33. The correlation between E(T_{CD}) and FIP for these pitchers is 0.48. The correlations are statistically significant and provide some evidence that E(T_{CD}) is a viable measure of pitching quality.

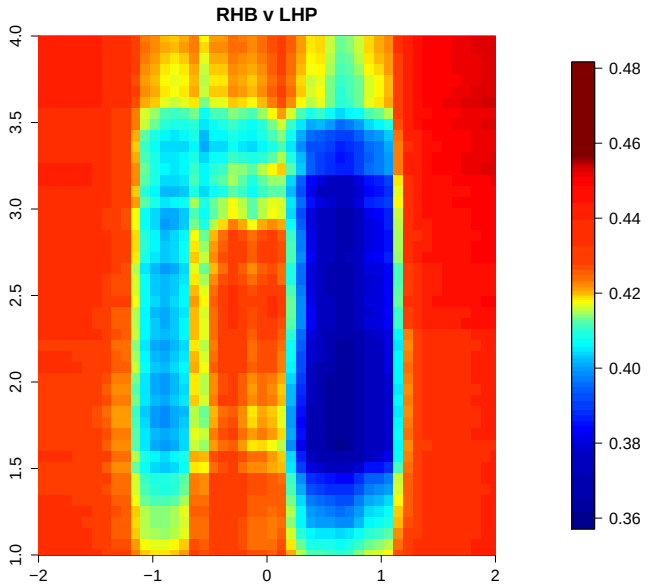

Figure 4: Heatmap of $\mathrm{E}(T_{CD})$ for right-handed batters who faced left-handed pitchers as viewed behind the batter.

From the pitchers considered above, we list the top 15 pitchers according to their average of expected bases allowed $\mathrm{E}(T_{CD})$ in Table 4. We then make a comparison with their ERA calculated over the three seasons 2013-2015. Many of the expected names appear near the top of the list. What is most interesting are the discrepancies between $\mathrm{E}(T_{CD})$ and ERA. For example, we note that Phil Hughes, Vidal Nuño, Kevin Slowey and Brandon McCarthy are perhaps better pitchers than indicated by their ERAs. It is worth emphasizing the interpretation of the $\mathrm{E}(T_{CD})$ values in Table 4. In the case of Phil Hughes, he gives up on average 0.392 bases during an at-bat based on a typical Phil Hughes pitch.

Whereas we do not claim that $\mathrm{E}(T_{CD})$ is the preferred and definitive measure for the evaluation of pitchers, we believe that it offers insights that may not be present in the other pitching statistics. $\mathrm{E}(T_{CD})$ solely judges pitch quality whereas other measures are based on run scoring which is subject to greater variability.

	Pitcher	E(T_{CD})	ERA (rank)
1.	Phil Hughes	0.392	4.27 (143)
2.	Jordan Zimmermann	0.395	3.19 (29)
3.	Vidal Nuño	0.396	4.12 (134)
4.	Max Scherzer	0.396	2.94 (14)
5.	Michael Pineda	0.397	3.57 (62)
6.	Kevin Slowey	0.397	4.45 (158)
7.	Clayton Kershaw	0.399	1.92 (1)
8.	Hisashi Iwakuma	0.399	3.17 (26)
9.	Brandon McCarthy	0.400	4.35 (148)
10.	Cliff Lee	0.400	3.08 (20)
11.	Carlos Carrasco	0.400	3.63 (68)
12.	David Price	0.400	3.01 (16)
13.	Chris Sale	0.400	2.92 (12)
14.	Madison Bumgarner	0.401	2.90 (11)
15.	Michael Wacha	0.402	3.21 (35)

Table 4: Top 15 average values of E(T_{CD}) for starting pitchers during the 2013-2015 seasons compared to ERA. The ERA rank (in parentheses) gives further context for the pitchers success relative to others in the league.

6 DISCUSSION

Although baseball analytics (also known has sabermetrics) has been active for a long time, PITCHf/x technology has now brought big data to the forefront. In this paper, we have used machine learning techniques on a massive dataset to learn about the quality of pitches. One of the attractive features of the approach is that the outcome variable total bases yielded is a true measure of what is important with respect to pitching. Also, we have made no assumptions regarding the functional relationship between total bases and the myriad of potential covariates.

This paper only begins to explore what may be possible when the quality of pitches is quantified. We have seen that pitch count is relevant to pitch quality. What may not be so well known is that the optimal period of pitch quality for starters lies roughly between the 20th pitch and the 70th pitch. We have also investigated heatmaps that suggest that low outside pitches within the strike zone are generally the best pitches from a pitcher's point of view. Also, we have taken a cursory look at the evaluation of pitchers.

In future work, it may be possible to look at the importance of pitch sequencing and potential game theory issues that come to bear on pitch selection. It may also be possible to develop pitch count rules (when to relieve a pitcher) that are pitcher dependent. There seems to be no shortage of interesting questions that may be addressed once the quality of pitches has been quantified.

7 REFERENCES

Albert, J. (2012). Looking ahead - a bright future for JQAS, *Journal of Quantitative Analysis in Sports*, 8, 1-1.

Baumer, B. and Zimbalist, A. (2014). *The Sabermetrics Revolution: Assessing the Growth of Analytics in Baseball*. The University of Pennsylvania Press: Philadelphia.

Basco, D. and Davies, M. (2010). The many flavors of DIPS: A history and a overview. *Society for American Baseball Research, Baseball Research Journal*, 32, Article 9.

Cross, J. and Sylvan, D. (2014). Bringing the heat. Presentation in contributed session #573, *2014 Joint Statistical Meetings*, Boston, MA, August 2-7.

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

Fast, M. (2009). What the heck is PITCHf/x? In *The Hardball Times Baseball Annual 2010*, Editors: J. Distelheim, B. and B. Jacobs, ACTA Sports: Chicago, 153-158.

Fast, M. (2011). Spinning yarn: can we predict hot and cold zones for hitters? In *Baseball Prospectus*, http://www.baseballprospectus.com/article.php?articleid=15363.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction, Second Edition*, Springer: New York.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18-22.

Moore, C. (2009). Best fastballs in baseball. In *Baseball Analysts: Rich Lederer & Friends*, http://baseballanalysts.com/archives/2009/09/.

Nycha, D., Furrer, R. and Sain, S. (2015). fields: tools for spatial data. *R package version 8.2-1*, http://CRAN.R-project.org/package=fields.

Rosales, J. and Spratt, S. (2015). Who is responsible for a called strike? Presentation at *MIT Sloan Sports Analytics Conference*, Boston, MA, February 27-28.

Roegele, J. (2013a). Strike zone pitch quality, part I: location. In *SB Nation: Beyond the Box Score*, http://www.beyondtheboxscore.com/2013/3/29/4153440/strike-zone-pitch-quality-part-i-location-pitchfx-sabermetrics.

Roegele, J. (2013b). 2013 Closers: pitch location vs velocity. In *SB Nation: Beyond the Box Score*, http://www.beyondtheboxscore.com/2013/4/5/4171286/2013-closers-pitch-location-vs-velocity-strike-zone-quality-pitchfx-sabermetrics.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, New York.

Sievert, C. (2014). Taming PITCHf/x data with {pitchRx} and {XML2R}. *The R Journal*, 6(1), http://journal.r-project.org/archive/2014-1/sievert.pdf.

Sievert, C. (2015). pitchRx: tools for harnessing MLBAM Gameday data and visualizing PITCHf/x. *R package version 1.7*. http://CRAN.R-project.org/package=pitchRx.

Tango, T., Lichtman, M. and Dolphin, A. (2007). *The Book: Playing the Percentages in Baseball*, Potomac Books, Washington, D.C.

Wickham, H. and Francois, R. (2015). dplyr: a grammar of data manipulation. *R package version 0.4.1*, http://CRAN.R-project.org/package=dplyr.

Wickham, H., James, D.A. and Falcon, S. (2014). RSQLite: SQLite interface for R. *R package version 1.0.0*, http://CRAN.R-project.org/package=RSQLite.

Wilson, J. (2015). Pitch quantification part 1: between-pitcher comparisons of QOP with conventional statistics, http://www.qopbaseball.com.