2

Using Publicly Available Baseball Data to Measure and Evaluate Pitching Performance

Carson Sievert and Brian M. Mills

CONTENTS

2.1 Introduction

In contrast to many other sports, the game of baseball can largely be described in discrete events, and player roles are well defined, making it relatively easy to measure player performance. Perhaps the most important individual role is the pitcher, since they have the most direct control over run prevention. However, directly measuring a pitcher's ability to prevent runs is difficult, since many relevant events depend on outside factors such as the team's ability to field balls hit into play. Sabermetricians have created a number of measures that attempt to isolate pitching from fielding contributions, but many do so by ignoring certain events and/or using subjective/incomplete information to distribute responsibility using probabilistic models.

With the advent of more granular, detailed, and accurate measurements of on-field activity, it is anticipated that our ability to measure and evaluate pitching performance will improve.

Many traditional measures of pitching performance, such as Earned Run Average (ERA) or pitcher Wins (W), confound pitching with other aspects of the game (e.g., fielding, baserunning, hitting, etc.). However, there are certain outcomes in the game that depend only on the skill of the pitcher (and batter): strikeouts, walks, and home runs. Fielding is irrelevant in these plate appearance outcomes since no ball is put in play.* This observation was first noted in Scully (1973) where a strikeout-to-walk ratio is used to measure pitching performance. But in 1999, Voros McCracken introduced and formalized the more robust Defense Independent Pitching Statistics (DIPS), made up of strikeouts, walks, and home runs allowed. While DIPS measures improved the understanding of pitcher contributions to game outcomes, it is still not well understood to what degree pitchers can control balls in play.

For years it was believed that pitchers had little to no control over batting average on balls in play (BABIP), making DIPS a reasonable measure of pitching performance, but this assumption breaks down when considering the type of ball hit into play (e.g., fly ball, ground ball, etc.). This finding should not shock those familiar with Simpson's paradox, as global trends rarely reflect local trends. If the goal is to measure overall performance of a pitcher, we believe it's best to avoid such assumptions and use all the information available. Moreover, pitchers can impact the game in more ways than delivering the ball to the batter (e.g., fielding, batting, baserunning), so an overall measure should respect these aspects of the game as well.

Inspired by James and Henzler (2002) and Jacques (2007), it is now commonplace to measure individual performance by estimating the number of wins contributed relative to "freely available talent" at that player's position. These win above replacement (WAR) measures are becoming increasingly popular since they are relatively easy to interpret and make it easy to compare player value. Unfortunately, most WAR measures use data/methods that are not freely available to the public. However, enough has been written to know that WAR measures are often a simple linear weight of season-level statistics. Moreover, it is common to see WAR measures designed specifically for pitchers use DIPS assumptions and consequently ignore batted ball information. As the public gains access

* We note that recent advances in understanding catcher skill as it relates to pitch framing are relevant here. We will address these in a later section.

to higher-resolution data, we expect to see more measures use at-bat and/or pitch-level data, which can improve our ability to measure individual performance.

openWAR (Baumer et al., 2015) is a completely open and reproducible WAR measure that uses publicly available at-bat level data. Since responses are defined on the at-bat level, openWAR can distribute defensive responsibility amongst pitcher and fielder(s) on batted balls in fashion similar to Jensen et al. (2009). openWAR does currently make the assumption that the value of each run scored counts equally for the offense and against the defense, but the underlying framework could be augmented to incorporate "clutchness" weights. It also assumes the change in expected runs of a plate appearance can be divided amongst four aspects of the game: (1) batting, (2) baserunning, (3) fielding, and (4) pitching.* In order to distribute value among fielding and pitching on balls hit in play, openWAR estimates the probability of an out based on locational data, and uses the estimated residuals to distribute value for good/bad fielding plays.

openWAR uses two-dimensional density estimation, a topic we cover and apply to pitch location data in Section 2.4, to estimate the probability of an out. This is almost certainly an oversimplified model that ignores useful information beyond the terminal location of the batted ball, but at the time, this was the only batted ball data available to the public.† Jensen et al. (2009) use more accurate and detailed (but proprietary) batted ball data to develop a more sophisticated model estimating this same probability. That study found that the location has a large effect on the likelihood of an out, but exit velocity also has a nonnegligible effect.‡ More information (the trajectory of the ball, the starting location of the fielders, etc.), which will soon be available to public, could not only lead to more accurate division of run prevention responsibility on batted balls, but potentially new and better measures of individual performance in general. An overview of the new data that we anticipate, as well as more traditional baseball data sources, is covered in Section 2.2.

In Section 2.3, we provide a more detailed look at traditional measures of pitching performance and propose some ideas for future research directions in this area. In Section 2.4, we discuss research using pitch level (aka PITCHf/x) data. Section 2.4 also covers statistical methodology that has proven useful for research using PITCHf/x data (in particular, generalized additive models, density estimation, and cluster analysis), but these methods could also be useful for at-bat-level data. The chapter is concluded with a discussion of future challenges in the evaluation of pitching performance.

2.2 Data Usage and Availability

2.2.1 Traditional Statistics

Many traditional statistical measures of pitcher performance have been available for more than 100 years. For example, strikeouts, wins, earned run average, walks, batting average against, and other statistics are commonly found on many baseball web sites. Depending

* After estimating individual contributions in each category, these categories are added together, to obtain an overall value measured in runs. The run value of an appropriate replacement player is then subtracted, and the resulting quantity is divided by 10 (empirical evidence suggests that 10 runs roughly equals a win).

† Someone watching the game determines these terminal locations.

‡ The velocities were generated by someone watching the game and were recorded as one of the following: {soft, medium, hard}. Having a more objective and granular measurement of velocity would improve the estimate of that effect.

on the analysis, one might need statistics on a career, seasonal, game, or even a play-by-play level. Serious analysts should consider obtaining play-by-play data from a reputable source such as http://www.retrosheet.org, since this data can always be aggregated to a desired level. Among play-by-play baseball data sources, Retrosheet has the richest set of play-by-play measurements spawning the greatest number of years. Play-by-play data is also available via the PITCHf/x system (discussed in Section 2.2.3), but it is only available for the 2007 season and later.

2.2.2 Balls in Play

The work of McCracken allowed us to separate two main areas of outcomes for pitchers: those pitches put in play, and those not put in play. While initial evaluation of batting average on balls in play (BABIP) indicated that pitchers had no control over BABIP, more recently analysts have argued that ball-in-play outcomes do in fact differ by pitcher. Strikeouts, walks, and home runs have been recorded for a large portion of baseball history, allowing one to make comparisons of the rate at which balls in play result in a hit across the history of professional baseball. For example, we can easily calculate BABIP for players like Cy Young in his 1890 rookie year with the Cleveland Spiders (0.269), as well as the entire careers of players like Nolan Ryan (0.265), Greg Maddux (0.281), and Clayton Kershaw (0.272). These data are readily available at web sites like Fangraphs or Baseball Reference but can also be calculated by anyone willing to delve into the play-by-play files at Retrosheet dating back to 1921.

More detailed data on balls in play, tracked by Baseball Info Solutions (BIS) classified as ground balls, fly balls, or line drives can be found on web sites like Fangraphs but are not available prior to the 2002 season. Information on in-play outcomes such as batted ball velocity, launch angle, and fly ball distance became publicly available in the 2015 season. This information can be obtained from http://www.baseballsavant.com/ or retrieved programmatically (Sievert, 2015a). These data may be particularly useful in developing measures of the quality of contact on pitches put in play, some of which may be attributed to pitcher skill.

2.2.3 PITCHf/x, Trackman, ChyronHego, and Statcast

In 2001, shortly after an umpire labor dispute and years of turmoil, Major League Baseball instituted the QuesTec system to monitor its umpire ball-strike calls. QuesTec revealed the location of each pitch as it crossed the plate, which could be matched with the call made by each umpire to assess the accuracy with which umpires were calling balls and strikes. The system was originally installed in only four MLB parks, but in 2004—after ratification by the umpire union—grew to be installed in half of the stadiums across the league. This data was not publicly available, though the league was able to use it extensively to understand the strike zone that was called by its umpires during this time.

While QuesTec was used through the 2008 season to monitor umpires, the league and its media subsidiary, Major League Baseball Advanced Media (MLBAM), began tracking pitches for display online and on broadcasts. The technology was developed by Sportvision, with three cameras set up to project the trajectory of pitches and where they would cross the front of the plate. This data is known more colloquially as PITCHf/x. In 2009, the league and its umpires agreed to allow the system to replace QuesTec for monitoring and evaluation, calling the new system Zone Evaluation. Fortunately for the baseball

statistician, MLBAM publicly released all play-by-play and pitch location data from part of the 2007 season through to the present (2015 season). This data includes information on pitch trajectories and velocity, where pitches crossed the plate, pitch type, ball-strike count, base-out state, inning, the pitcher and the batter, runners on base, and the umpire(s) working the game. This data has been used extensively in modern baseball analysis and continues to evolve with more advanced integration of other data sources. As we demonstrate in Section 2.4.2.2, it is easy to acquire this data using the R package pitchRx (Sievert, 2014a). It can also be accessed through web sites such as http://www.brooksbaseball.net and http://www.baseballsavant.com.

In 2015, MLB began using what it has branded as Statcast data. This data uses the Trackman system that has been largely unreleased to the public, but used since at least 2011, paired with ChyronHego data. Trackman uses a radar system to reveal similar information to PITCHf/x such as velocity, movement, and where the pitch crossed the plate. It also adds the spin rate, angle of the ball, the angle off the bat, and information on the trajectory of a hit ball. ChyronHego tracks player movements, which has led to calculations of acceleration when stealing a base, measurements of baserunner leads, route efficiency for fielders, and top speed of baserunners and fielders. These can be used to evaluate quality of baserunning and fielding at a level that has not been reached with prior data. However, the large majority of this data is still not consistently made publicly available and is mostly used internally by teams, or for MLB Statcast broadcasts beginning in 2015. Most importantly, the data should lead to more precise measures of fielding ability, which will allow further identification of pitcher talent and performance independent of fielder quality.

2.3 Measures of Pitching Performance

2.3.1 Traditional Measures

Traditional measures of pitcher performance have serious flaws in both their ability to isolate a pitcher's contribution to run prevention and their ability to predict future performance of a pitcher. Traditional measures include Wins, which depend both on defense and the run scoring ability of a team's hitters, and ERA and batting average against (BAA), which are strongly dependent on defense. Other traditional statistics such as walks (BB), strikeouts (K), and home runs allowed (HRA) are more useful (Albert, 2006). As noted in Piette et al. (2010), this is particularly evident when normalized per inning or per nine innings pitched (K/9, BB/9, and HR/9) or as a rate statistic with the denominator being the total number of batters faced (K%, BB%, HR%).

As noted earlier, Voros McCracken and others have shown the importance of isolating pitcher performance from fielders and focusing on aspects of the game under control by the pitcher. For example, errors are dependent both on the quality of fielders behind the pitcher and on the subjective judgment of the scorekeeper for the game. By focusing on other outcomes and assuming BABIP is largely out of the control of the pitcher, the analyst can isolate performance in a more useful way. The advent of McCracken's observation has led to a plethora of measures focused on these other outcomes, though as noted earlier, there has been increased recent interest in balls in play with the advent of new ball tracking technologies.

2.3.2 Strikeout, Walk, and Home Run Rates across Eras

The availability of strikeout, walk, and home run rates makes yearly comparisons convenient for individual pitchers and the league as a whole. Strikeout rates have increased dramatically—part of a 50-plus-year trend. The scoring environment in MLB has decreased dramatically, particularly since the advent of performance-enhancing drug policies in MLB, making each home run allowed more important in the outcome of the game. In any analysis of the performance of pitchers across time, analysts can be misled without knowing the changes in base rates for strikeouts or run scoring. Figure 2.1 displays the time series of the overall strikeout rate, overall walk rate, and overall home run rate for the entire post–World War II era (1946–2015).

It is clear that a comparison of the strikeout rate of an average pitcher in 2015 to one just after WWII does not reveal particularly useful information in the context-specific skill level of a given pitcher. Because of the evolution of the game, standardizing pitcher statistics within season can be helpful to compare two pitchers from different eras. If this standardizing is not performed, it will be unclear whether or not the implied performance in a strikeout rate, for example, is a product of a changing environment in which the game is played, or attributed to innate talent in a given pitcher.

2.3.3 Defense Independent Pitching Statistics (DIPS)

Since McCracken's observation regarding balls in play, a number of statistics have been proposed that attempt measuring pitching performance independent of fielding performance. Basco and Davies (2010) summarize this development in SABR's *Baseball Research Journal*. Most DIPS measures derive from linear weights (Thorn and Palmer, 1985) that can be calculated through a Markov (or other) simulation using the various base-out states and associated transition probabilities. Marchi and Albert (2014) provide R scripts for most of these calculations using play-by-play Retrosheet data.

Fielding Independent Pitching (FIP), probably the most well-known DIPS measure, uses only home runs, walks, and strikeouts by the pitcher.* There are variations of FIP, with the general calculation as

$$FIP = C + \frac{\alpha * HR + \beta * BB - \gamma * K}{IP}.$$

In this notation, C is an arbitrary constant that is often identified as the league average ERA, allowing the measure to be interpretable on an ERA scale. The parameter α identifies the coefficient for home runs allowed (often between 11 and 13), β is the parameter for walks allowed (usually around 3), and γ is the parameter for strikeouts (often approximately 2). The parameters are estimated using relative linear weights run values for each outcome, and the measure is adjusted for the number of innings pitched by the pitcher. Typically the FIP measure is on a nine-inning scale, defined as expected runs allowed per nine innings pitched.†

* There are variations of this that include hit-by-pitches or remove intentional walks as well (Basco and Davies, 2010).

† Sabermetrician Tom Tango (2011) has a useful breakdown of how these parameters are estimated in the FIP formula. The reader is referred to this source for examination of the derivation of the measure: http://www.insidethebook.com/ee/index.php/site/comments/tangos_lab_deconstructing_fip/.

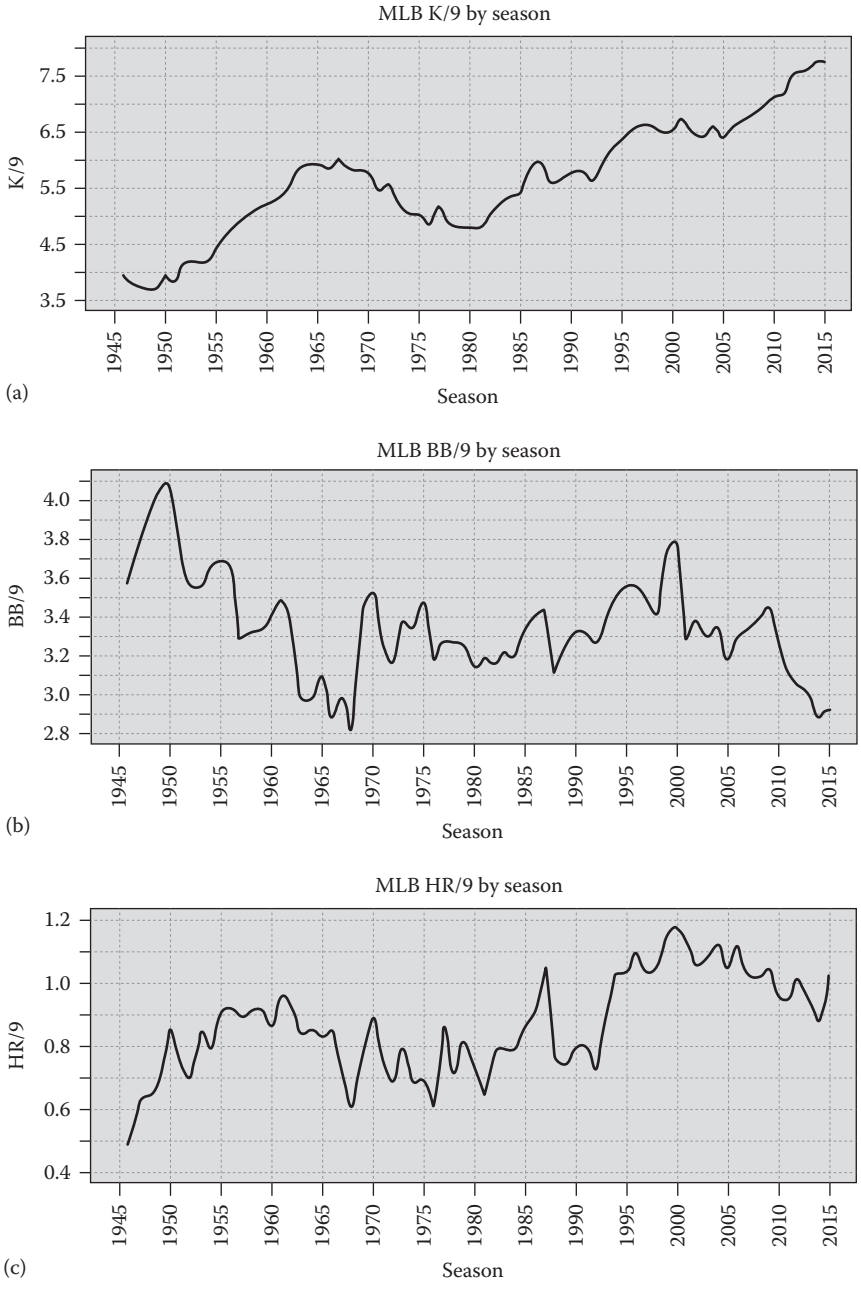

FIGURE 2.1
Plots of strikeouts per nine innings pitched (a), walks per nine innings pitched (b), and home runs per nine innings pitched (c) for the post–World War II era.

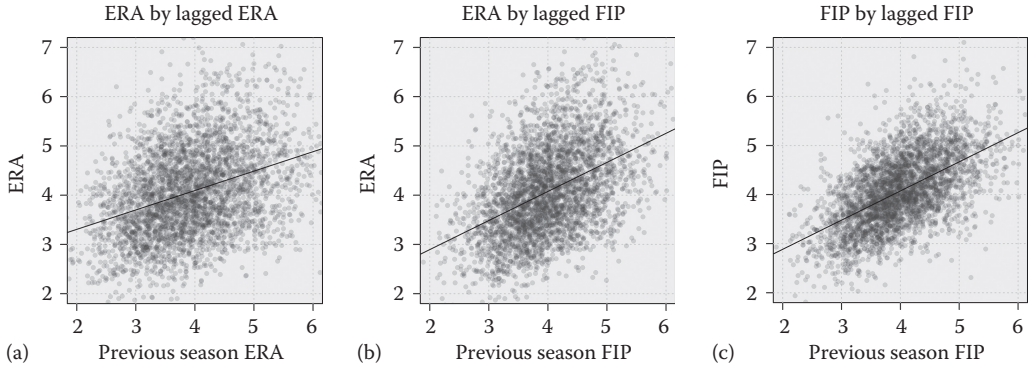

FIGURE 2.2
Plots of individual pitcher ERA in year t as a function of ERA in year $t-1$ (a; $r = 0.364$), ERA in year t as a function of by FIP in year $t-1$ (b; $r = 0.431$), and FIP in year t as a function of FIP in year $t-1$ (c; $r = 0.559$). Notice the stronger correlation for FIP in the rightmost panel. Data are from 1980 through 2015.

The FIP measure avoids the dependence on fielding performance and gives stable measures of pitching performance from year to year. Figure 2.2 constructs scatterplots between FIP in year t and FIP in year $t-1$, ERA in year t and FIP in year $t-1$, and ERA in year t and ERA in year $t-1$. The message from these plots is that the current FIP is a better predictor of the following season's ERA than the current ERA, and the current FIP is a relatively strong predictor of the following season's FIP. This is perhaps not surprising, given that we've already shown that strikeout, walk, and home run rates are more stable than other measures that depend on fielding or randomness of batted ball outcomes. In addition to being more stable, Piette et al. (2010) argue that FIP, HR/9, and BB/9 exhibit the most signal for starting pitchers, while ground ball percentage, fly ball percentage, and K/9 exhibit the most signal for relievers using a random effects model.

Variations of FIP include SIERA (Swartz and Seidman, 2010), xFIP (Studeman, 2005), and Component ERA (James et al., 2000). With the exception of Component ERA, each of these additional measures requires data on batted balls. Unfortunately, batted-ball data is often rather subjective and was not recorded in earlier MLB seasons.

Interestingly, measures such as SIERA and xFIP tend to be more stable over time than the FIP measure due to the repeatability of the ground ball and fly ball rates of individual pitchers, and FIP is influenced by a less stable home run rate. Ultimately, more precise data will improve upon these measures by identifying pitchers that reduce the quality of contact across batted balls. There is substantial room for research in this area, particularly with the advent of new data on balls in play.

2.3.4 Ground Ball and Fly Ball Rates

Identifiers for grounds balls and fly balls on balls in play have been available since 2002 on web sites such as Fangraphs. As with other outcomes like strikeouts, walks, and home runs, ground ball and fly ball rates are best presented as percentages. In this case, they are presented as a total percentage of all batted balls against a given pitcher. League average for

ground ball, fly ball, and line drive rates in 2015 were 45.3%, 33.8%, and 20.9%, respectively. These rates show a little variation across seasons.

The importance of ground ball and fly ball rates relates to the relative value of inducing ground balls and fly balls, though this relationship is currently not well understood. Certain pitchers may be referred to as "ground ball pitchers" or "fly ball pitchers" if they have tendencies to allow pitches of one type. However, having ground ball or fly ball tendencies does not necessarily make one a high or low quality pitcher. For example, Dallas Kuechel found great success as a ground ball pitcher in 2015, with 66.3% of batted balls against him resulting in ground balls, while only 18.5% were fly balls. Alternatively, Max Scherzer was an early season Cy Young candidate in 2015 with only 36% ground balls and 45.4% fly balls. We know that fly balls are more likely to go for extra bases or over the fence for a home run, which might lead one to prefer the ground ball pitcher. Yet, Scherzer shows success can be had in the air. Further, there is some evidence that when ground ball pitchers do allow a fly ball, it is more likely to result in a home run (Swartz, 2011).

The type of pitcher preferred (ground ball or fly ball) depends somewhat on the ballpark in which they play regularly and the quality of infielders and outfielders behind them. For example, if a pitcher plays in a very large park like Citi Field—particularly with players like Juan Lagares roaming the outfield—it may be more advantageous to induce fly balls, as there is more space for outfielders to run down a ball that may otherwise go over the fence at smaller parks. Alternatively, at a stadium such as Coors Field, pitchers may want to avoid too many fly balls that could easily turn into home runs from the thin air.

The type of batted balls contributes to the probability that a given ball put in play will drop in for a hit. While we know that pitchers do not have full control over BABIP, batted ball data has revealed that there are some things that can be controlled by pitchers such as quality of contact (Swartz and Seidman, 2010). Further, it is important to note that ground ball and fly ball percentages include a substantial amount of subjectivity based on the judgment of those entering the data. The advent of more advanced data sets like Statcast will help reduce subjectivity in these classifications and add substantially to the literature on the value of certain types of batted balls. This is discussed in more detail in Section 2.3.6.

2.3.5 Velocity, Angle, and Hard Hit Balls

The availability of some Statcast data, beginning in 2015, has allowed the analyst to learn more about the patterns of balls hit in play. The data currently includes the velocity of balls off the hitters' bat and the angle at which the ball left the bat. This allows more precise and objective definition of the characteristics that define a ground ball, line drive, or fly ball.

One can model the probability of a hit, given the velocity and launch angle of the batted ball, to characterize balls that are "hard hit"—this is likely more accurate than a scorekeeper's judgment regarding what constitutes a line drive or fly ball. As a preliminary look, we model the probability of a hit conditional on batted ball velocity using the 2015 regular season balls in play from the two World Series contenders, the New York Mets and the Kansas City Royals. We use logistic regression to identify changes in hit probability as a function of this batted ball velocity. Here we are ignoring the importance of launch angle, which will play a role in where the ball lands and whether it is a ground ball or fly ball. Two models are fit, one with home runs included and one without.

TABLE 2.1

Logistic Regression of Hit Probability on Batted Ball Velocity

	No HR Included	**HR Included**
Constant	−4.676***	−5.327***
SE	*(0.227)*	*(0.227)*
Velocity (mph)	0.0436***	0.0518***
SE	(0.0025)	(0.0025)

Note: ***Refers to statistical significance at the 99% level.

From Table 2.1 one sees that an increase of 1 mph in batted ball velocity increases the probability of a hit—if the hit is in play—by about 4.5%.* If home runs are included in the model, one sees an increased hit probability of about 5.3%. This is not a surprising result, as we would expect balls that are hit harder would be more likely to drop in for hits.

Pitchers and pitching coaches can use this information to determine the best way to minimize batted ball velocity and therefore decrease hits. Making use of the locational information from PITCHf/x alongside the Statcast batted ball data—and using the methods in the following section—one can build models and visualize how hard batters hit pitches in different locations (Mills, 2015). In particular, if we assume that hard hit balls are more likely to fall in for hits, these visuals are informative about locations where the pitcher can minimize the probability of a hit by minimizing velocity off the bat. Generalized Additive Models (GAMs), discussed in Section 2.4.2, can be useful for this type of analysis.

2.3.6 Expected Run Values and Ball-Strike Count Progression

The ball-strike count, which is independent of fielding influence, can change the expected run value of an at-bat through strategic interaction from the pitcher. Increasing the number of strikes in the count relative to the number of balls could be seen as an independent skill set at a more granular level than even play-by-play data. The measured changes in run values can be used to measure that skill. Using techniques from Albert (2010), one can estimate that the average effect of changing a ball to a strike during the 2007–2010 seasons is about 0.146 runs (Mills, 2014a). Therefore, there is significant value in even individual pitches throughout the game.

Table 2.2 further breaks down values of pitches under various counts. These are likely to change depending on the sample used, as scoring has decreased rather substantially throughout the 2000s and 2010s from peak levels at the turn of the century.

Marchi and Albert (2014) further analyze ball-strike count effects, and how these result in changes in batter swings, noting that much of the data on batting outcomes after reaching a specific count can be found on Baseball Reference, with pitch sequencing information available through Retrosheet. Ultimately, Marchi and Albert show that the progression of the count is important, with a first pitch strike resulting in a decrease of offensive output of approximately 28%, while starting with an 0-2 count results in only 30% of the expected result at the beginning of the at bat. And one knows that the behavior of the batter will change with the count—for example, the batter is more likely to swing when he is behind in the pitch count.

* We exponentiate each coefficient for ease of interpretability.

TABLE 2.2[a]

Changes in Run Value by Pitch Count

Count	Run Value	Δ with Strike Call[a]	Δ with Ball Call	Difference[b]
0-0	−0.038	−0.043	0.038	0.081
0-1	−0.081	−0.052	0.025	0.077
0-2	−0.133	−0.300	0.013	0.313
1-0	0.000	−0.056	0.060	0.004
1-1	−0.056	−0.064	0.058	0.122
1-2	−0.120	−0.300	0.041	0.341
2-0	0.060	−0.058	0.107	0.049
2-1	0.002	−0.081	0.100	0.181
2-2	−0.079	−0.300	0.097	0.397
3-0	0.167	−0.065	0.330	0.265
3-1	0.102	−0.084	0.330	0.414
3-2	0.018	−0.300	0.330	0.630
Weighted Avg.[c]	—	—	—	*0.146*

Source: Modified from Albert, J., *J. Quant. Anal. Sports*, 6(4), 1, 2010; Mills, B.M., *Managerial and Decision Economics*, *35*(6), 387, 2014a.

[a] This—as well as the difference after a ball call—is calculated as, for example, the difference between the run value of an 0-1 count and an 0-0 count: $(-0.081) - (-0.038) = -0.043$.

[b] Take the difference between a successive strike call and successive ball call in order to evaluate the net change from changing a strike to a ball or vice versa.

[c] Proportion of pitches thrown in the given count come from the data in Mills (2014a).

2.4 Tools for Analyzing PITCHf/x Data

In Section 2.4.1, techniques for visualizing pitch locations are described by the use of two-dimensional density estimation. Density estimation allows us to see patterns in a large amount of pitch locations, but does little to help *explain* those patterns or suggest what may have occurred under different circumstances. It is often more useful to have inferential tools to say, for instance, "Conditioned on this pitch location (and some other covariates), the estimated probability of event A is $P(A)$." GAMs are one attractive approach to making such inference, especially when the sample size is large.

GAMs and PITCHf/x data have allowed us to learn more about MLB pitching than just pitcher evaluation. Most noticeably in the academic literature, PITCHf/x has allowed researchers to better understand umpire decision-making with respect to the strike zone (Green and Daniels, 2015; Hamrick and Rasp, 2015; Kim and King, 2014; Mills, 2014, 2014b; Moskowitz and Wertheim, 2011; Parsons et al., 2011; Tainsky et al., 2015). Catchers have also been analyzed using this data (Fast, 2011; Judge et al., 2015; Marchi, 2013; Pavlidis and Brooks, 2014).

Publications on umpire decision-making have modeled the probability of a called strike as a function of pitch location (as it crosses the front of home plate), and other covariates, such as year, ball-strike count, and home field advantage. In Section 2.4.2, we provide a roadmap for building such models in R, from data collection, to model specification, model fitting and diagnosis, as well as visualization. A concise introduction to Generalized Additive Models (GAMs) is provided in Section 2.4.2.1 before diving into an application

in Section 2.4.2.2. These methods are not restricted to umpire decision-making but also helpful in modeling a pitcher's ability to control pitches and make good decisions on the mound.

In Section 2.4.3, we shift focus from estimating densities and event probabilities to classification and clustering methods. The pitch type labels that come with publicly available PITCHf/x data are automatically created using a classification algorithm, which uses variables such as velocity and spin (Fast, 2008), but there is reason to believe the algorithm itself could be improved (Pane et al., 2013). In Section 2.4.4, some of the underlying physics is discussed on how these variables relate to pitch trajectory and classification.

2.4.1 Visualizing Pitch Location Frequencies

PITCHf/x data provides the vertical and horizontal locations of the ball the moment it crosses the front of home plate. There are some known accuracy issues with these locations, but previous work suggests that when the system is properly calibrated, they randomly deviate about a half of an inch away from the exact location (Nathan, 2008). Nevertheless, it is well known that pitch location is tied to pitcher performance (in general, it is more difficult to hit a baseball low in the zone as opposed to high in the zone); so pitch location summaries can offer insight into performance (as well as other things such as umpire decision-making). Sometimes it is useful to collect locations of hundreds or thousands of pitches, but a basic scatterplot of the pitch locations leads to an uninformative graphic, a problem known as overplotting.

There are many proposed graphical solutions to overplotting, each having its own drawbacks. Some approaches simply alter certain visual characteristics of the graphic (Few, 2008), while other methods are based on the estimation of the underlying probability density (so-called density estimation). In Figure 2.3a, a large number of pitches thrown by Yu Darvish from 2009 to 2012 are graphically displayed using a density estimate where the alpha transparency parameter is set such that 100 points would have to be overlaid on the same spot in order to appear fully opaque. This allows us to see where the highest density of pitches occurs, but it is difficult to infer just how many pitches were thrown in a particular location. There is no good rule of thumb for setting the transparency level, so this approach requires some trial and error to generate a decent-looking graphic.

Altering the visual characteristics of a graphic, such as the alpha transparency, can help reveal some structure, but it may not work well when most points are concentrated within a small area and the remaining points are scattered over a large area. A more robust approach is to use a binning and/or smoothing procedure to summarize the data into contiguous regions, thus eliminating the problem of overlaying points.

Two basic examples of binning are presented in Figure 2.3: one with rectangular binning and one with hexagonal binning. Binning algorithms are easy to understand as one simply counts the number of observations that fall within each region (absolute frequency), and (optionally) divide by the total number of observations to obtain a density estimate. In this case, the absolute frequencies are encoded using a linear color scale resulting in a graphic commonly called a heat map or level plot. Absolute frequencies are easier to interpret and are helpful when comparing multiple scenarios since they provide a sense of the total number of pitches.

Binning algorithms are simple, but it can be difficult to select an optimal number of bins. In the case of rectangular binning, traditional solutions depend on characteristics of the unknown underlying distribution and a default rule of thumb assumes a Gaussian

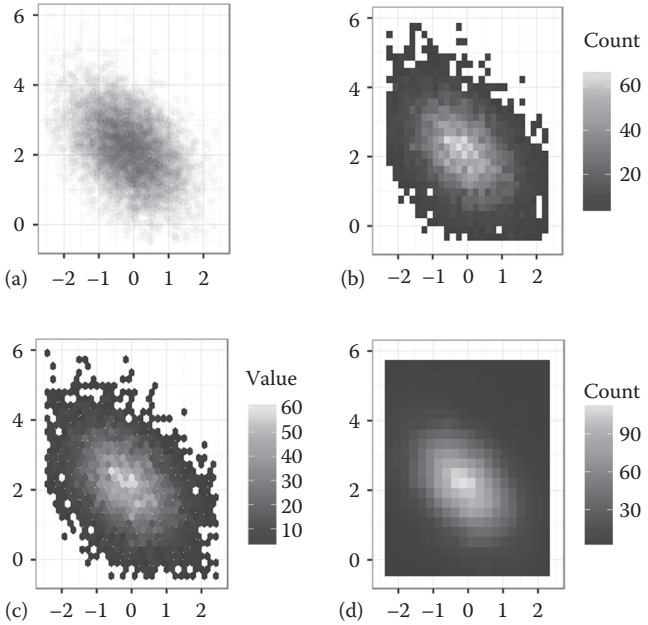

FIGURE 2.3
Four plots of Yu Darvish's pitch locations from 2012 to 2014. These plots take the perspective of the umpire at the pitch crosses the front of home plate. (a) All pitches with alpha transparency, (b) 2D histogram with rectangular binning, (c) 2D histogram with hexagonal binning, and (d) bivariate normal kernel density estimate.

form (Scott, 2015). There are a number reasons to prefer hexagonal to rectangular binning (for one, hexagons are less visually distracting); however, there is relatively little literature on the choice of an optimal or sensible default for bin sizing.

Density estimation via binning can suffer from high variance. The same data and different anchor points for the bins can result in very different values for the binned frequencies. The high variance problem can be avoided by the use of kernel-based density estimation at the expense of higher bias if the bandwidth parameter is misspecified. The **MASS** package in R (Venables and Ripley, 2002) provides the `kde2d()` function for 2D density estimation with a bivariate normal kernel as well as a sensible routine for choosing the bandwidth of the kernel estimate (choosing the bandwidth is essentially the continuous version of choosing the number of bins in a 2D histogram). The density portrayed in Figure 2.3d was generated using `kde2d()`'s default bandwidth selection and the number of bins were set to match the 2D histogram in Figure 2.3b.

One can also apply some sort of smoothing technique to the binned values before visualizing. The `persp()` and `contourLines()` functions in the R package **graphics** both use interpolation techniques to approximate a smooth surface over an *xy*-plane, but the former creates a 3D plot whereas the latter projects the surface into two dimensions using contour lines to encode values of the *z*-axis. We do not recommend 3D surface plots since they can be visually deceptive and obscure/block interesting regions of the surface; however, interactive versions where users can alter the perspective can help alleviate these issues.

It's often useful to compare and contrast a number of density estimates across a variety of scenarios. Generally one should separate pitches thrown to left-handed batters from

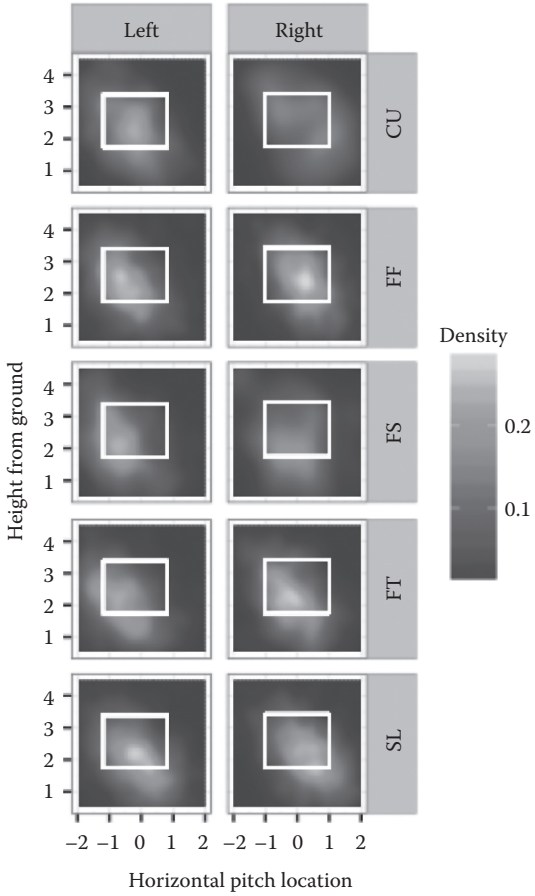

FIGURE 2.4
Bivariate normal kernel density estimates of Yu Darvish's pitches for each combination of batter handedness and pitch type.

right-handed batters since these densities are typically very different. For this reason, the R package **pitchRx** provides the function `strikeFX()` to help quickly create many densities in a small amount of code (Sievert, 2014a). Figure 2.4 shows 10 different bivariate normal kernel density estimates of Yu Darvish's pitches: one for each combination of batter handedness and pitch type. Clearly, Darvish (as with most pitchers) tends to throw down and away to batters. Interestingly, Darvish appears to have thrown two-seam and sinking fastballs down and inside to right-handed batters.

A number of approaches to visualizing pitch locations using various 2D density estimation techniques have been described. For more details on how Figures 2.3 and 2.4 were actually generated, see the code provided at https://gist.github.com/cpsievert/fd83ec5516a07ab59c36. It is important to note that density estimation, in itself, doesn't allow for likelihood comparisons conditioned upon a pitch location. This type of inference is important, especially for studying umpire calls where, for instance, one would like to compare the probability of a called strike for various umpires. The next section covers one particular statistical method that allows us to make this type of inference.

2.4.2 Generalized Additive Models (GAMs)

A detailed overview of the Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1986) framework is beyond the scope of this chapter. However, we introduce this method in Section 2.4.2.1, focusing on important properties, and apply them to PITCHf/x data in Section 2.4.2.2 using R (R Core Team, 2015). Clark (2013) provides a more detailed introduction to additive models in R and Wood (2006) gives a comprehensive introduction to the GAM framework, and its implementation in the R package **mgcv**.

2.4.2.1 A Brief Overview of GAMs

Those familiar with the Generalized Linear Models (GLMs) framework (Nelder and Wedderburn, 1972) will find many connections between GLMs and GAMs. The "generalized" term essentially means that in either framework, one can model the response distribution with any distribution belonging to the exponential family. The exponential family includes a number of common distributions such as normal, binomial, gamma, Poisson, etc. This flexibility is crucial for modeling PITCHf/x data, or any type of data, where the response of interest is a count, percentage, or rate observed over time. In cases where the response does not follow any one distributional form, one can indirectly select a distribution by modeling the relationship between its observed conditional mean and variance.

In addition to a distributional form, both frameworks require a one-to-one function, $g(\cdot)$, which links the conditional expectation, $E(Y|X = x)$, to the predictor space η, so that $g(E(Y|X = x)) = \eta$. This link function is necessary since there is no guarantee the range of $E(Y|X = x)$ will match the range of η. For example, if Y is a Poisson random variable, then $E(Y|X = x)$ is always positive (since $Y > 0$), but there is nothing to restrict η to be positive. In this case, we seek a function such as `log()` that can map the positive real numbers to all real numbers.

The key difference between generalized *linear* models and generalized *additive* models is in the assumed form of η. In the GLM setup, η is assumed to be a *linear* combination of independent variables multiplied by unknown coefficients $\beta_0 + \sum_{i=1}^{p} \beta_i X_i$. In the basic GAM setup, η is a sum of smooth functions of the independent variables $\beta_0 + \sum_{i=1}^{p} f_i(X_i)$, which is referred to as a nonparametric model. In practice, it is common to have a semiparametric model that contains both unknown coefficients and smooth functions; for example, $\eta = \beta_0 + \sum_{i=1}^{r} \beta_i X_i + \sum_{j=r}^{p} f_j(X_j)$.

A natural question at this point is "when are additive terms in a nonparametric model preferred over parametric terms in a linear model?" Parametric terms are generally more interpretable but require strong assumptions about the predictor space, whereas additive terms can automatically detect nonlinear effects. This flexibility is quite useful for modeling strike-zone event probabilities, as it's common to see a nonlinear relationship with pitch location. This flexibility comes at a computational price, however, as the "degree of smoothness" must be estimated using a cross-validation technique. Thankfully, recent advances have produced computationally efficient methods for estimating the level of smoothness (Wood, 2011).

2.4.2.2 Modeling Events over the Strike Zone with GAMs

In this section, we build a GAM to estimate the probability of a called strike as a function of pitch location and year. Although only binary outcomes over the strike zone are covered, remember the GAM framework supports any exponential family, so the response could

even be multinomial. The multinomial case is more suitable for building more general models since a pitch can result in a number of outcomes.

Since the response distribution is binary, a natural choice for the link function is the $\text{logit}(x) = \log(x/(1-x))$, which gives us a mapping from the domain of the response mean [0, 1] to the domain of the predictor space $(-\infty, \infty)$. Now let $Y_i \in \{strike, ball\}$ be the outcome of pitch i, X_{1i} be the corresponding vertical location, X_{2i} the horizontal location, $Z_{h(i)}$ be an indicator variable of whether the batter was left or right handed, and $Z_{t(i)}$ be an indicator of whether the pitch was thrown in 2008 or 2014. Then a sensible GAM model would be

$$\begin{aligned}\text{logit}(EY_i) = \beta_0 + \beta_1 Z_{h(i)} + \beta_2 Z_{t(i)} + \beta_3 Z_{h(i)} Z_{t(i)} + f_0(X_{1i}, X_{2i}) \\ + f_1(X_{1i}, X_{2i}) Z_{h(i)} + f_2(X_{1i}, X_{2i}) Z_{t(i)} + f_3(X_{1i}, X_{2i}) Z_{h(i)} Z_{t(i)}\end{aligned}$$

Note that with this model structure, we have one joint smooth function of horizontal and vertical location for each factor level. The reason why a *joint* function is preferred over marginal functions for each coordinate is that, contrary to the rulebook definition, strike zones are asymmetric Also, there is a different smooth for each factor level since the probability of a called strike on the edge of the rulebook strike zone is known to depend on batter stance and year.

The complete analysis, from data collection to model fitting, diagnostics and visualization, can be performed entirely within R. The R package **pitchRx** makes it easy to collect all available PITCHf/x data and store that data in a database (Sievert, 2014a). Storing PITCHf/x in a database is convenient (but not necessary) since there are many tables available that record information on various levels: pitches, at bats, hits in play, baserunning, games, players, umpires, coaches, etc. **pitchRx** works with any database connection, but it's particularly easy to create a SQLite database from R with **dplyr** which is illustrated here.

```
library(dplyr)
library(pitchRx)
db <- src_sqlite("pitchRx.sqlite3", create = TRUE)
scrape(start = "2008-01-01", end = Sys.Date(), connect =
db$con)
```

Now that a complete PITCHf/x database has been obtained, the package **dplyr** can be used to query variables of interest*:

- px: the horizontal location of the pitch (as it crossed the front of home plate)
- pz: the vertical location of the pitch (as it crossed the front of home plate)
- des: A description of the pitch outcome
- num: The order of at bat (within game). This is used to link pitch info to at-bat info.
- gameday_link: A unique identifier for game. This is used to link pitch info to at-bat info.

```
pitches <-tbl(db, "pitch") %>%
select(px, pz, des, num, gameday_link) %>%
filter(des %in%c("Called Strike", "Ball"))
```

* A more complete list of PITCHf/x variable descriptions can be found here: https://fastballs.wordpress.com/2007/08/02/glossary-of-the-gameday-pitch-fields/.

For this model, one also needs the stance of the batter and year, which is recorded on the at-bat level. For visualization purposes, the height of the batter is also used to draw an average rectangular strike zone for reference.

```
atbats <-tbl(db, "atbat") %>%
mutate(year =substr(date, 5L, -4L)) %>%
select(stand, b_height, num, year, gameday_link)
```

Now these tables are joined together, attention is restricted to 2008 and 2014 seasons, and an indicator variable for called strikes is created.

```
dat <- pitches %>%
left_join(atbats, by = c("gameday_link", "num")) %>%
filter(year %in% c("2008", "2014")) %>%
collect() %>%
mutate(strike = as.numeric(des == "Called Strike"))
```

Last, the model proposed at the start of this section is fit. Here the argument k is specified to increase the dimension of the basis allowing for smoother surfaces but reducing the computational efficiency. This increase in dimension was guided by diagnostics provided by the **gam.check**() function which suggested the initial fit using the default value for k was too smooth.

```
library(mgcv)
m <-bam(strike  ~interaction(stand, year) +
s(px, pz, by =interaction(stand, year), k = 50),
data = dat, family =binomial(link ='logit'))
```

The **strikeFX()** function in the **pitchRx** package is designed to work with **mgcv** for quick visualization of estimated probability surfaces over the strike zone. It also has support for visualizing the differences, as illustrated in Figure 2.5. In addition to visualizing a point estimate of difference, it can also be useful to compute the actual probabilities and associated measures of uncertainty. For quantifying this uncertainty, we recommend the bootstrapping procedure shown in Wood (2006). For an interactive visualization of all this information, see Sievert (2014b).

```
strikeFX(dat, model = m,
density1 =list(year ="2008"),
density2 =list(year ="2014")) +
facet_grid(. ~stand)
```

It is important to note that GAMs are not limited to strike probabilities. There are a number of other outcomes that are important to the pitcher, such as the probability of making contact, the quality of contact, whether balls in play become fly balls or ground balls, which pitches tend to be hit for home runs, and so on. The package used to fit these models in R is also not limited to the logit link function, and other appropriate links can be used. For example, when modeling exit velocity, the Gaussian link function may be used to model this continuous response data. Finally, these models are not restricted to pitching, as they could be used to further model fielding prowess with models of the probability that batted balls fall for hits, conditional on launch angle, batted ball velocity, and landing point.

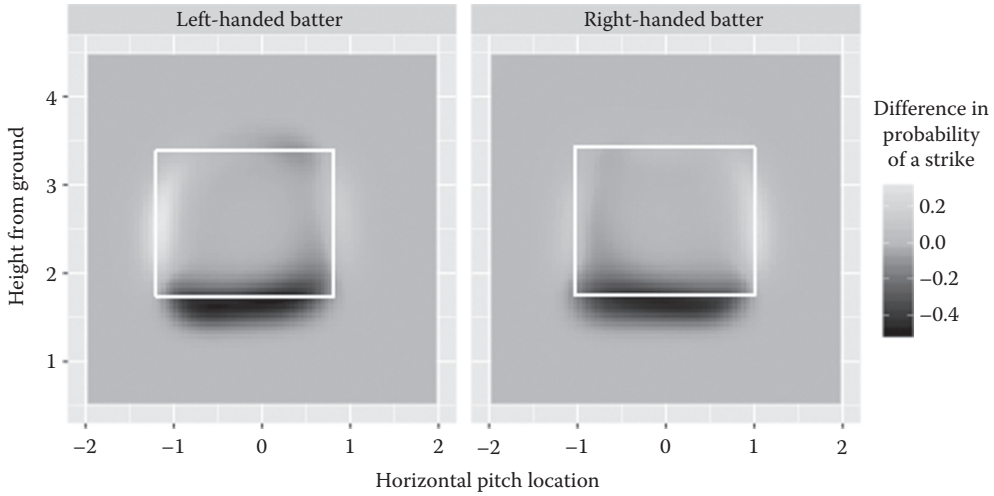

FIGURE 2.5
The difference in the probability of a called strike in 2008 versus 2014 for both left- and right-handed batters. A pitch over the plate at the knees was about 40%–50% more likely to be called a strike in 2014. A pitch on the inner/outer portion of the plate (above the knees) was about 20%–30% more likely to be called a strike in 2008.

2.4.3 Cluster Analysis and Pitch Type Identification

While understanding the location of pitches, called strikes or balls, and which pitches will be hit well or poorly are key components for understanding pitchers, it is also helpful to identify the various types of pitches with different velocities and movements. Cluster analysis can be useful for the analyst in understanding these characteristics.

One of the key goals is the identification of the pitch types so that one may evaluate the success of each pitch in the context of the batting outcome and previous choices made. PITCHf/x data conveniently provide pitch classifications using a machine learning algorithm. However, the method used for this classification is not public, and it may be possible to improve upon this classification with statistical tools (Pane, 2013; Pane et al., 2013). This is ultimately an unsupervised learning problem as there is no training data with known membership to preface group assignment for the data of interest.

One can first think about the pitch type clustering problem by projecting pitches onto a two-dimensional space, inspecting characteristics of points, and look for separate groups. The goal of the clustering methods is to identify these pitches in their respective groups as closely as possible based on the given characteristics. It is desirable to identify every pitch exactly as it was intended with zero misclassifications, though the likelihood of perfect classification of each pitch is unlikely, due to variability around the spatial characteristics of human physical actions. Figure 2.6 shows the clusters of some randomly generated data first without assignment, and then with assignment denoted by color. The goal of the clustering methods is to find these clusters from the data all in black (Figure 2.6a), and identify them as separate groups as shown in red, green, and black (Figure 2.6b). Note that the clusters below are well separated; however, this is not usually the case with all pitch types gleaned from pitch data.

A number of clustering methods are potentially useful for this task, including hierarchical clustering (agglomerative or divisive), *k*-means clustering, and model-based clustering.

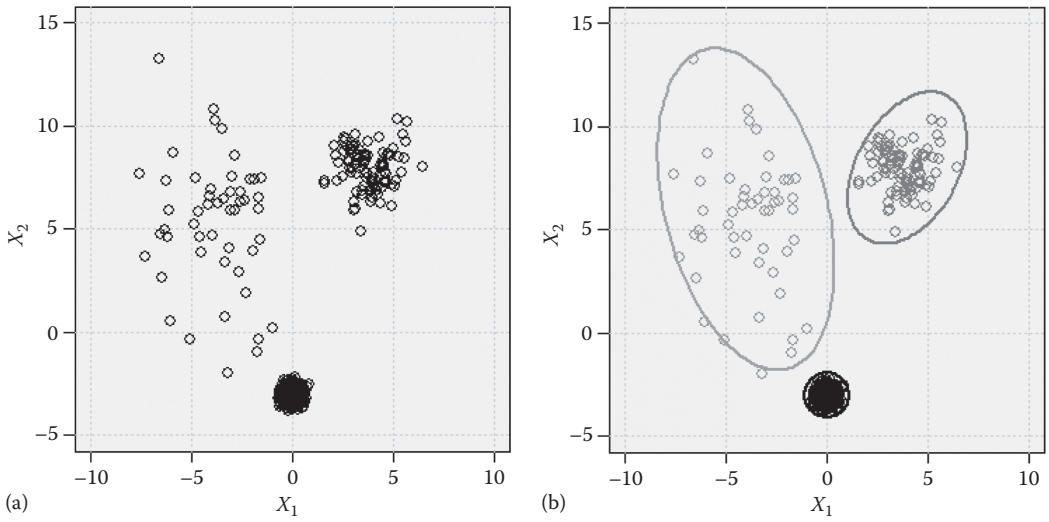

FIGURE 2.6
Exhibition of unclassified data (a) and cluster assignment (b) of generic, randomly generated data.

The choice of clustering method depends largely on the structure and variance of the clusters, and the knowledge of the analyst about the number of different pitches a pitcher throws. If possible, clustering should always be done on the individual pitcher level, since differences across pitchers can confuse the clustering method.*

It is important to note that clustering does not require both a dependent and independent variable, but identifies cluster membership based on a group of variables (e.g., characteristics of each pitch) that the analyst finds appropriate for assigning pitches to different types. These features often include velocity, spin rate, spin direction, and trajectory of the pitch (specifically, trajectory relative to expectations for a four-seam fastball). The *k*-means and model-based clustering methods are described here, but it is worth mentioning that investigations into the appropriateness of hierarchical methods may also be worth examining.

2.4.3.1 k-*Means Cluster Analysis*

k-means clustering is a partitioning method requiring the researcher to choose a priori the number of pitch types, k, as well as the initial location of cluster centroids in a *p*-dimensional feature space. The initial locations are arbitrary and are commonly chosen at random or based on a previously developed hierarchical clustering method. To begin, the k-means algorithm assigns each pitch to its closest centroid in the feature space. The clustering method continues along this path, reassigning centroids and assigning data points to the closest centroids in an iterative fashion according to the Hartigan and Wong (1979) algorithm. The initial placement of these centroids may impact the quality of the clustering solution, and it is recommended that the initial placement of centroids be considered carefully, with possibly multiple starting points attempted. Further, while the *k*-means

* For example, Mark Buehrle's four-seam fastball averaged about 84 mph in 2013, while a change-up from a pitcher like Noah Syndergaard registers at about 88 mph.

iterations reach a stabilized solution relatively quickly, it is worth considering the number of iterations that balance a reasonable solution and computation time.

In the case of pitch clustering, characteristics such as horizontal and vertical movements, angle of the movement, and velocity of the pitch tend to be most useful. Work from physicist Alan Nathan (2008, 2012) discusses how to identify movement of pitches based on the provided variables in PITCHf/x data. But the data itself provide three very useful variables for clustering that are more intuitive before employing the tools of a physicist yourself: *start_speed*, *break_length*, and *break_angle*. The *start_speed* variable simply identifies the velocity of the pitch out of the pitcher's hand in miles per hour (mph). The *break_length* measure—as defined by Cory Schwartz, Vice President of Stats at Major League Baseball Advanced Media (MLBAM)—is "the measurement of the greatest distance between the trajectory of the pitch at any point between the release point and the front of home plate, and the straight line path from the release point and the front of home plate.*" Alternatively, the *break_angle* measure is "the difference between a pitch dropping perfectly perpendicular to the ground, and the actual trajectory of the pitch." The sign of *break_angle* encodes whether the pitch breaks toward the pitcher's handedness (positive if breaking across the pitcher's body, otherwise negative) and the magnitude encodes the actual break angle (0 would indicate the ball drops directly toward the ground).

Together these three PITCHf/x variables give us information about the traditional main characteristics that distinguish pitch types: the amount of movement, the direction of movement, and the velocity of each pitch. It is important to note that scaling variables can be helpful when variables are defined on different scales. This is clearly the case here, where the model includes characteristics of pitches like velocity, measured in miles per hour, break length, measured in inches), and break angle.

Figure 2.7a shows a scatterplot of *start_speed* and *break_angle* for pitches thrown by Mark Buehrle's during the 2013 season. A k-means classifier ($k = 5$) was fit to these pitches using *start_speed*, *break_length*, and *break_angle* for feature variables and the R function **`kmeans()`** in the **stats** package. The result of that fit is shown in Figure 2.7b. In this case, our feature space is 3D, so we *could* produce a 3D graph to view the model fit with respect to the entire feature space, but a method known as touring would be preferable since it allows us to view a high-dimensional feature space, which is a powerful tool for diagnosing statistical models such as k-means (Wickham et al., 2015).†

Understanding the relative quality of clustering solutions can be important in initially developing useful classifications of pitch types. The quality of the clusters using *k*-means methods can be measured using homogeneity (similarity within each cluster) as well as separation (difference across clusters). A popular measure of homogeneity is within-cluster sum of squared error (WSSE), which is defined as the sum of squared distances between each individual pitch and its cluster centroid. A popular measure of separation is between-cluster sum of squares (BSSE), which is defined as the sum of squared distances between each cluster centroid and the grand mean of the data.

The use of a silhouette coefficient and associated silhouette plot combines the homogeneity and separation considerations into a single measure. This silhouette coefficient is measured using information on the average distance a_i of each object i from each other

* From a discussion at The Book Blog (2007) at http://www.insidethebook.com/ee/index.php/site/comments/everything_you_ever_wanted_to_know_about_gameday/.

† Touring is a technique that requires dynamic interactive graphics software. For a simple video explanation and demonstration of touring PITCHf/x data, see Sievert (2015b).

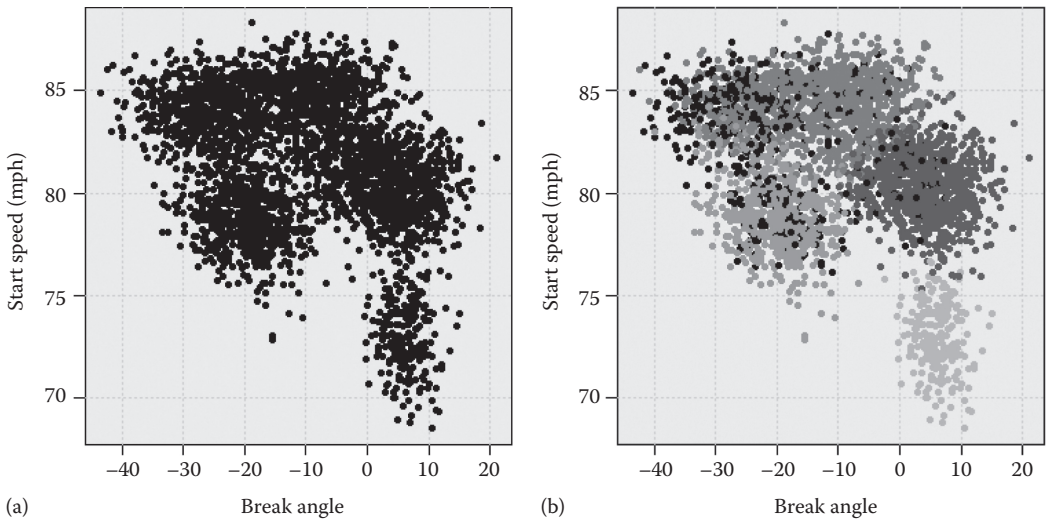

FIGURE 2.7
(See color insert.) Exhibition of unidentified pitches from Mark Buehrle's 2013 regular season (a) and pitch cluster assignment (b) using *k*-means clustering with five clusters.

object in its respective cluster, as well as the minimum average distance b_i of object $i$to objects in other clusters. The silhouette coefficient is defined as

$$s_i = \begin{cases} 1 - \frac{a_i}{b_i} & \text{if } a_i < b_i \\ \frac{a_i}{b_i} - 1 & \text{if } a_i \geq b_i \end{cases}$$

Typically, $s \in [0, 1]$, with values closer to 1 indicating better quality clustering. The average silhouette width can be calculated for a single cluster, as well as an entire clustering output by averaging the silhouette coefficient across objects (pitches). These coefficients can be plotted to visualize cluster assignment quality from the given solution.

There are, however, a number of drawbacks in using the *k*-means clustering method. First, the a priori selection of K requires the analyst to know how many pitches each pitcher intends to throw. Further, this specification may be too restrictive in that even if a pitcher is noted to throw a single curveball, the variability in that pitch may result in two different pitch types as perceived by the batter. Second, iterations of the centroids may result in either empty or very small clusters, in which case some outliers may be removed for better performance of the method. Even in the case where the sparse or empty issue does not arise, it is possible that iterations of cluster placement remain in a local optimum, rather than a global optimum for the clustering problem. Finally, the method does not handle differing size, density, and shape of clusters or other methods, such as model-based clustering.

2.4.3.2 Model-Based Clustering

Model-based clustering (Fraley and Raftery, 2002) is the least restrictive method of clustering pitch types—there is no need for the researcher to a priori choose the number of clusters (pitches) in a given pitcher's arsenal—and exploration of the data reveals that the various cluster densities (most pitchers have a primary pitch, which will occur much more often than others) and variance structures may be well addressed using probabilistic

models. Model-based clustering uses population subgroup (cluster) densities to identify group membership based on maximum likelihood and the EM (Expectation-Minimization) with parameterized Gaussian mixture models (Fraley and Raftery, 1999). Further, the method estimates cluster membership probabilities for each observation (pitch) across each respective cluster (pitch type). Clusters are determined using the Bayesian Information Criterion (BIC), comparing across models with different parameterizations (number of clusters, size/shape, and orientation). Conveniently, R has its own package to implement this method (**mclust**; Fraley and Raftery, 1999, 2007; Fraley et al., 2012).

Fraley and Raftery (2007) exhibit that the clusters in this method come from a mixture density, $f(x) = \sum_{k=1}^{G} \tau_k f_k(x)$, where f_k is the probability density function of observations in group k, and τ_k is the probability that an observation comes from the kth mixture component. Components are largely modeled using a Gaussian distribution, characterized by mean μ_k and covariance matrix Σ_k, with probability density function

$$\phi(x_i; \mu_k, \Sigma_k) = \frac{\exp\left\{-(1/2)(x_i - \mu_k)^{\mathrm{T}} \Sigma_k^{-1} (x_i - \mu_k)\right\}}{\sqrt{\det(2\pi\Sigma_k)}}.$$

The likelihood for the data with n observations, with G components is

$$\prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \phi\,(x_i; \mu_k, \Sigma_k).$$

Fraley and Raftery (2007) further discuss the EM algorithm and geometric constraints on the G components, to which the reader is referred to both for the technical aspects of the clustering method as well as its implementation in R.

Pane et al. (2013) originally identified the usefulness of model-based clustering in the analysis of baseball data, and developed an adjusted Bayesian Information Criterion (BIC) for use with pitch data that avoids identifying too many sparse clusters. The traditional BIC method tends to result in too many identified clusters. The additional penalty from the adjusted BIC measure is shown to perform better with respect to this dimension of the method, and penalizes high intracluster correlations. Specifically, it reduces the number of chosen clusters, k, from a to b, $b < a$, when the intracluster correlation for $k = a$ is much higher than for $k = b$. The adjusted BIC measure is then calculated as (taken directly from Pane, 2013) follows:

$$BIC_{adj} = -2\log(f(Y|\hat{p})) - 2\lambda \sum_{i} \log(f(c_i)) + \left[k \cdot \left(j + \frac{j(j-1)}{2} + (k-1)\right)\right] \cdot \log(n).$$

As noted in Pane (2013), $f(Y|\hat{p})$ identifies the probability of the parameters given the data, $f(c_i)$ is the sum of the upper off diagonal elements of a correlation matrix, Σ_k, and j is the number of clustering variables. Further, the tuning parameter, λ, was determined using a training data set in Pane (2013), with $\lambda = 0.5$ being determined as the optimal tuning parameter choice for reducing pitch misclassifications.

Using this BIC_{adj} measure, we can visualize a solution to the clustering problem posed earlier for Mark Buehrle in Figure 2.8, with five pitch types identified.

Note the improvement in clusters going from k-means to model-based clustering, particularly with the black cluster that is likely to be two-seam fastballs. The k-means solution

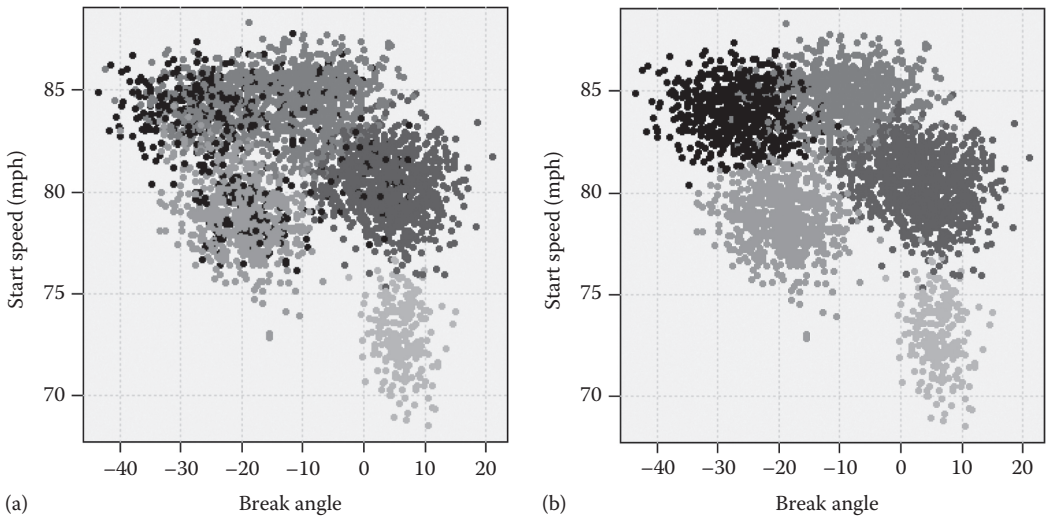

FIGURE 2.8
(See color insert.) Exhibition of Mark Buehrle's 2013 pitch cluster assignment using *k*-means clustering (a) and cluster assignment using model-based clustering (b). The latter shows more clear cluster assignment, particularly for the black-colored cluster.

resulted in a much more scattered cluster, while the model-based solution provides much tighter cluster assignment. The two-seam fastball classification difficulty was noted in Pane (2013), with the model-based solution providing more uniform clusters than other methods. It is likely that the darker blue cluster is a group of cutters, while the red cluster is four-seam fastballs. Buehrle's change-up is made up of the green cluster, while his curveball is the light blue cluster. Note that pitches to the left of the plot indicate "arm side movement," while those to the right of the plot indicate movement across the pitcher's body.

Ultimately, there are important limitations of clustering alone that relate to the inability for an outside researcher to know which pitch was *intended*. With any clustering method, there will exist misclassified pitches that could lead to incorrect conclusions regarding the strategy of the pitcher. Further knowledge of pitcher intention can help to identify shortcomings in approach, or provide a feedback loop for development of pitches.

For example, if there is substantial variation in the movement of a pitcher's curveball, then the use of clustering algorithms can tell us about the variation within this pitch type. Depending on the goals of the pitcher—developing two separate pitches, or increasing consistency in one pitch—the matching of cluster type and intention becomes paramount in this developmental process. Without knowledge of the intention, coaches and staff may not be able to intervene and give feedback on improvement. While this, of course, requires access to the players, analysts should be aware of the limitations of simplistic identification in the context of improving our understanding of pitching and pitch quality. However, a reasonably well-designed clustering method could reveal useful characteristics of strategic decision-making among pitchers, even for the outside analyst.

2.4.4 Pitch Movement and Trajectory

An important aspect of identifying pitch types after clustering includes understanding how the various measures used to cluster the pitches impact the trajectory of the pitch.

Without understanding this trajectory, identifying what each cluster represents in terms of pitch type would be rather difficult. Pitches are often classified casually by their respective trajectory and velocity—fastball, curveball, changeup, knuckleball. In this vein, Nathan (2008, 2012) lays the foundation for identifying pitch trajectories and defining the movement of a pitch as "the amount by which the trajectory deviates from a straight line with the effect of gravity removed." Relating the movement of a pitch—ultimately its trajectory deviation—back to expectations about the perceived movement of certain pitch types reveals the identity of each cluster in the absence of information on the pitcher's original intention with the pitch. While the physics behind pitch trajectories is beyond the scope of this chapter, Pane (2013, p. 18) includes a useful table to evaluate pitch types based on speed, vertical break, and horizontal break.

2.5 Current Extensions

2.5.1 Umpire Influences

There is a now relatively extensive academic literature making use of PITCHf/x data to evaluate umpire decision-making, and the impact that this has on the performance of players and general game play. The first paper to leverage this data in the academic literature is Parsons et al. (2011), who claimed to find evidence of race-based discrimination in umpire ball-strike calls. More recent work has called into question whether this is the case (Hamrick and Rasp, 2015; Tainsky et al., 2015), but the main crux of Parsons et al. (2011) remains: otherwise objective performance measures can be contaminated by subjective judgment of officials, and these must be accounted for when estimating specific skill. This is identical to the lesson of McCracken (1999) with respect to fielders.

Further, Moskowitz and Wertheim (2011), Green and Daniels (2015), and Mills (2014a) found that the size of the strike zone fluctuates depending on recent calls and the current ball-strike count, with calls favoring the pitcher (batter) when he is behind (ahead) in the count. And Mills (2014a) and Kim and King (2014) showed home plate umpires tend to give more favorable calls to both players with a higher status in the league as well as players who remain in close proximity to the umpire during the game. Further, Mills (2014b) and Roegele (2014) found evidence of strike-zone expansion from 2008 to 2014, which is believed to be a large contributor to the decrease in runs scored over the same period. Each of these considerations makes important contributions to the measurement and isolation of pitching performance.

2.5.2 Catcher Influences

The implications of the influence of umpires generalize to the influence of any other actor that could impact the measurement of pitcher performance. This, of course, includes the catcher. Work by a number of analysts (Fast, 2011; Judge et al., 2015; Marchi, 2013; Pavlidis and Brooks, 2014) has revealed not just catcher influences on pitching performance but have managed to use this information to evaluate catchers themselves. Many of these models are hierarchical in nature to identify catcher effects for pitchers when they throw to different catchers across a season or career.

In particular, Pavlidis and Brooks (2014) use PITCHf/x data and a GAM to estimate a probabilistic model of the strike zone—as discussed in Section 2.4—and attribute residuals

to catchers. These are then applied to the run expectancy across each count to evaluate the total (retrospective) framing value of each catcher in the league (Tango et al., 2006). With these estimates in place, the impact of the pitcher's skill level can be isolated from this additional impact.

2.6 Future Challenges

2.6.1 Development and Minor League Projection

MLB teams are currently expanding their investments in technology for use in the development of players both in the minor leagues and when they reach MLB. Development using this technology requires a continuous feedback process, such as noting whether a curveball conforms to the pitch type cluster for an individual pitcher immediately after throwing the pitch. Or, release point data immediately available could result in pitchers developing better command more quickly as they progress. Full integration of data sources and communication with sports scientists should improve development as data availability and analysis techniques advance within franchise operations.

The intersection of statistical analysis using this advanced data and providing direct and actionable feedback is likely to be the most valuable future endeavor for pitching. Previously, pitcher development relied on the naked eye of pitching coaches and catchers to identify specific issues with pitches or approach. However, with new data—including physiological and biomechanical measurements—coaches will be able to provide more objective feedback related to shortcomings in a pitcher's game. While this seems to be a fruitful area of research, it clearly requires an insider's knowledge and access in order to bring it to its full potential.

2.6.2 Quality Pitches and Strategic Considerations

Understanding what makes a successful pitch, and of course how to throw it, has been a question long pondered by pitchers and pitching coaches alike. It is becoming more common to see higher-velocity pitches across the league, and given that this allows less reaction time for the batter, it seems clear that it would be more difficult to make solid contact with higher velocity pitches (all else equal). One might also assume that pitches that have less traditional movement would be harder to hit; however, there have been very few consistently successful knuckleball (perhaps the most nontraditional trajectory) pitchers in MLB. But we also know that these pitches with more movement tend to travel slower than a straight fastball. Contextual clues and recency effects—or the impact of seeing a fastball before a curveball—should play a role in future statistical evaluations of pitching success using the rich data available to the analyst.

While descriptive statistical analysis—as well as projections of future performance—has been widespread and accessible for many practitioners in sports, strategic interaction between pitchers and batters has been sorely under-researched. Surely, the problem at hand is much more complex than a simple description of past output; however, with the advent of PITCHf/x and Trackman/Statcast data, there is opportunity to test strategic theories within the sport. Marchi and Albert (2014) and Alamar et al. (2006) touch on this to some extent, noting some behavioral changes that take place with certain pitches and in

certain counts. And in soccer, for example, economists have investigated the use of minimax and other strategies on penalty kicks (Palacios-Huerta, 2003). There is little reason to believe that this sort of evaluation cannot be extended to pitching, though the task requires substantially more choices among players.

2.7 Conclusion

Measuring pitching performance is difficult, mostly due to the fact that run prevention depends on both pitching and fielding. Compared to traditional statistics like Wins or ERA, DIPS measures (functions of strikeouts, walks, home runs) are a better measure of performance but also ignore important information on balls hit into play. Probabilistic models are a sensible way to distribute run prevention contributions between the pitcher and fielders, and our ability to estimate the "true" distribution will only improve as more granular data becomes available.

Probabilistic models for categorical variables that can incorporate spatial and temporal components as covariates (e.g., Generalized Additive Models) will be important not only for estimating run prevention contribution but also for understanding the influence of confounding factors, such as bias in umpire decision-making or influences of catchers on ball-strike calls. The ability to integrate Statcast with PITCHf/x data will be a crucial first step to such analysis, which is not a trivial data management problem, and will require new tools to provide easy access to the data. There is ample opportunity for extension of these models in this space, which will continue to evolve in the coming years.

References

Alamar, B., Ma, J., Desjardins, G.M., and Ruprecht, L. (2006). Who controls the plate? Isolating the pitcher/batter subgame. *Journal of Quantitative Analysis in Sports*, *2*(3), 1–8.

Albert, J. (2006). Pitching statistics, talent and luck, and the best strikeout seasons of all-time. *Journal of Quantitative Analysis in Sports, 2*(1), 1–30.

Albert, J. (2010). Using the count to measure pitching performance. *Journal of Quantitative Analysis in Sports, 6*(4), 1–28.

Basco, D. and Davies, M. (2010). The many flavors of DIPS: A history and an overview. *The Baseball Research Journal, 39*(2), 41–50.

Baumer, B.S., Jensen, S.T., and Matthews, G.J. (2015). openWAR: An open source system for evaluating overall player performance in Major League Baseball. *Journal of Quantitative Analysis in Sports, 11*(2), 69–84.

Clark, M. (2013). Getting started with additive models in R. Retrieved September 23, 2015 from: http://www3.nd.edu/∼mclark19/learn/GAMS.pdf.

Fast, M. (2008). Drinking from a firehose. Retrieved September 23, 2015 from: http://www.hardballtimes.com/drinking-from-a-fire-hose/.

Fast, M. (2011). Spinning yarn: Removing the mask encore presentation. Baseball Prospectus. Retrieved October 25, 2015 from: http://www.baseballprospectus.com/article.php?articleid=15093.

Few, S. (2008). Solutions to the problem of over-plotting in graphs. Retrieved September 29, 2015 from: https://www.perceptualedge.com/articles/visual_business_intelligence/overplotting_in_graphs.pdf.

Fraley, C. and Raftery, A.E. (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification, 16*, 297–306.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*, 611–631.

Fraley, C. and Raftery, A.E. (2007). Model-based methods of classification: Using the mclust software in chemometrics. *Journal of Statistical Software, 18*, 1–13.

Fraley, C., Raftery, A.E., Murphy, T.B., and Scrucca, L. (2012). mclust Version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report No. 597, Department of Statistics, University of Washington, Seattle, WA.

Green, E. and Daniels, D.P. (2015). Impact aversion in arbitrator decisions. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2391558.

Hamrick, J. and Rasp, J. (2015). The connection between race and called strikes and balls. *Journal of Sports Economics, 16*, 714–734.

Hartigan, J.A. and Wong, M.A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *28*(1), 100–108.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*(3), 297–318.

Jacques, D. (2007). Value over replacement player. Retrieved December 31, 2015 from: http://www.baseballprospectus.com/article.php?articleid=6231.

James, B. and Henzler, J. (2002). *Win Shares.* San Francisco, CA: STATS Publishing.

James, B., Zminda, D., and Munro, N. (2000). *STATS All-Time Major League Handbook*. San Francisco, CA: STATS Publishing.

Jensen, S.T., Shirley, K.E., and Wyner, A.J. (2009). Bayesball: A Bayesian hierarchical model for evaluating fielding in major league baseball. *The Annals of Applied Statistics, 3*(2), 491–520.

Judge, J., Pavlidis, H., and Brooks, D. (2015). Moving beyond WOWY: A mixed approach to measuring catcher framing. Baseball Prospectus. Retrieved October 25, 2015 from: http://www.baseballprospectus.com/article.php?articleid=25514.

Kim, J.W. and King, B.G. (2014). Seeing stars: Matthew effects and status bias in Major League Baseball umpiring. *Management Science*, 60, 2619–2644.

Marchi, M. (2014a). The stats go marching in: Catcher framing before PITCHf/x. Baseball Prospectus. Retrieved October 25, 2015 from: http://www.baseballprospectus.com /article.php?articleid=20596.

Marchi, M. and Albert, J. (2014). *Analyzing Baseball Data with R*. CRC Press, Boca Raton, FL.

McCracken, V. (1999). Defense independent pitching statistics. Retrieved June 14, 2015 from: http://www.futilityinfielder.com/dips.html.

Mills, B.M. (2014a). Social pressure at the plate: Inequality aversion, status, and mere exposure. *Managerial and Decision Economics, 35*(6), 387–403.

Mills, B.M. (2014b). Expert workers, performance standards and on-the-job training: Evaluating major league baseball umpires. Retrieved September 30, 2015 from: http://ssrn.com/abstract=2478447.

Mills, B.M. (2015). Houston Astros whiffs and exit velocity. Exploring Baseball Data with R. Retrieved October 28, 2015 from: https://baseballwithr.wordpress.com/2015/06/30/ houston-astros-whiffs-and-exit-velocity/.

Moskowitz, T.J. and Wertheim, L.J. (2011). *Scorecasting: The Hidden Influences behind How Sports Are Played and Games Are Won.* Crown Archetype, New York.

Nathan, A.M. (2008). A statistical study of PITCHf/x pitched baseball trajectories. Retrieved September 29, 2015 from: http://baseball.physics.illinois.edu/MCAnalysis.pdf.

Nathan, A.M. (2012). Determining pitch movement from PITCHf/x data. Retrieved October 14, 2015 from: http://baseball.physics.illinois.edu/Movement.pdf.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A, 135*, 370–384.

Palacios-Huerta, I. (2003). Professionals play minimax. *Review of Economic Studies, 70*, 395–415.

Pane, M.A. (2013). Trouble with the curve: Identifying clusters of MLB pitchers using improved pitch classification techniques. Dietrich College of Humanities and Social Sciences, Carnegie

Mellon University, Pittsburgh, PA.Retrieved September 1, 2015 from: http://repository.cmu.edu/hsshonors/190/.

Pane, M.A., Ventura, S.L., Steorts, R.C., and Thomas, A.C. (2013). Trouble with the curve: Improving MLB pitch classification. http://arxiv.org/pdf/1304.1756.pdf.

Parsons, C.A., Sulaeman, J., Yates, M.C., and Hamermesh, D.S. (2011). Strike three: Discrimination, incentives, and evaluation. *American Economic Review, 101*, 1410–1435.

Pavlidis, H. and Brooks, D. (2014). Framing and blocking pitches: A regressed, probabilistic model: A new method for measuring catcher defense. Baseball Prospectus. Retrieved October 25, 2015 from: http://www.baseballprospectus.com/article.php?articleid=22934.

Piette, J., Braunstein, A., McShane, B.B., and Jensen, S.T. (2010). A point-mass mixture random effects model for pitching metrics. *Journal of Quantitative Analysis in Sports, 6*(3), 1–15.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Roegele, J. (2014). The strike zone during the PITCHf/x era. The Hardball Times. Retrieved October 1, 2015 from: http://www.hardballtimes.com/the-strike-zone-during-the-pitchfx-era/.

Scott, D.W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken, NJ: John Wiley & Sons.

Scully, G. (1973). Pay and performance in Major League Baseball. *American Economic Review, 64*, 915–930.

Sievert, C. (2014a). Taming PITCHf/x data with pitchRx and XML2R. *The R Journal, 6*(1), 5–19. Retrieved September 22, 2015 from http://journal.r-project.org/archive/2014–1/sievert.pdf.

Sievert, C. (2014b). Interactive visualization of strike-zone expansion. Exploring Baseball Data with R. Retrieved October 20, 2015 from https://baseballwithr.wordpress.com/2014/11/11/interactive-visualization-of-strike-zone-expansion-5/.

Sievert, C. (2015a). Obtaining exit velocity and distance of batted balls. Exploring Baseball Data with R. Retrieved October 20, 2015 from https://baseballwithr.wordpress.com/2015/07/15/obtaining-exit-velocity-and-distance-of-batted-balls/.

Sievert, C. (2015b). The grand tour. Retrieved August 31, 2016 from https://vimeo.com/148050343.

Studeman, D. (2005). I'm batty for baseball stats. Hardball Times. Retrieved October 23, 2015 from: http://www.hardballtimes.com/im-batty-for-baseball-stats/.

Swartz, M. (2011). New SIERA, part two (of five): Unlocking underrated pitching skills. *Fangraphs*. Retrieved October 25, 2015 from: http://www.fangraphs.com/blogs/new-siera-part-two-of-five-unlocking-underrated-pitching-skills/.

Swartz, M. and Seidman, E. (2010). Introducing SIERA: Part 1. Baseball Prospectus. Retrieved October, 23, 2015 from: http://www.baseballprospectus.com/article.php?articleid=10027.

Tainsky, S., Mills, B.M., and Winfree, J.A. (2015). An examination of potential discrimination among MLB umpires. *Journal of Sports Economics, 16*, 353–374.

Tango, T. (2007). Everything you ever wanted to know about Gameday. Retrieved August 10, 2015 from: http://www.insidethebook.com/ee/index.php/site/comments/everything_you_ever_wanted_to_know_about_gameday/.

Tango, T. (2011). Tango's lab: Deconstructing FIP. Retrieved September 15, 2015 from: http://www.insidethebook.com/ee/index.php/site/comments/tangos_lab_deconstructing_fip/.

Tango, T.M., Lichtman, M.G., and Dolphin, A.E. (2006). *The Book: Playing the Percentages in Baseball*. Middletown, DE: TMA Press.

Thorn, J. and Palmer, P. (1985). *The Hidden Game of Baseball*. New York, NY: Doubleday/Dolphin.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. New York, NY: Springer.

Wickham, H., Cook, D., and Hofmann, H. (2015). Visualizing statistical models: Removing the blindfold. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 8*, 203–225.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B* (*Statistical Methodology*), *73*(1), 3–36.