Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Hillary Superak Date

Analyzing Batting Patterns of Major League Baseball Players
For Advance Scouting Reports:

Using R to Generate High-Level Spatial Plots of PITCHf/x Data

By

Hillary M. Superak
Master of Science in Public Health

Biostatistics & Bioinformatics

_____

Patrick D. Kilgo, MS
Committee Chair

_____

Julie A. Clennon, PhD
Committee Member

Analyzing Batting Patterns of Major League Baseball Players For Advance Scouting Reports:

Using R to Generate High-Level Spatial Plots of PITCHf/x Data

By

Hillary M. Superak

B.S., Washington University in St. Louis, 2009

Thesis Committee Chair: Patrick D. Kilgo, MS

An abstract of
a thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2011

Abstract

Analyzing Batting Patterns of Major League Baseball Players
For Advance Scouting Reports:

Using R to Generate High-Level Spatial Plots of PITCHf/x Data

By Hillary M. Superak

Baseball, regarded as "America's National Pastime," has been a constantly evolving sport in many respects. In particular, the concept of data analytics has increasingly been applied to baseball in recent years. Statistical and graphical analyses have enhanced two main areas of Major League Baseball: player selection and game strategy. In 2006, Sportvision developed PITCHf/x, a system of high-speed cameras installed in every MLB stadium that records every pitch. PITCHf/x data include the three-dimensional spatial coordinates of the ball's trajectory, along with several other pitch characteristics. This technology provided the first opportunity to evaluate individual pitches based on information not contained in box scores or other statistics.

Graphical analyses have been conducted on the PITCHf/x data, but there is room for improvement. Specifically, new techniques were needed that could display spatial data for large data sets – for example, an entire season's worth of pitches received by a batter. Also desirable was a method of plotting two-dimensional spatial data that could be categorized according to the levels of a third, discrete variable. The figures needed to be informative, yet concise and easily interpretable.

A user-friendly graphical analysis tool for the advance scouting of MLB batters was developed. A function was programmed in R to generate three types of plots: (1) heat map density plots, (2) heat map contour plots, and (3) hexbin pie charts. Plots were created according to the hierarchical four-level "pitch universe": (1) all pitches, (2) swings only, (3) balls in play only, and (4) base hits only. Classification by pitch type was also examined.

Results for two players, Jason Heyward and Dustin Pedroia, were evaluated. As expected, the "hotspot" for Jason Heyward was consistently in the lower outside area of the strike zone; for Pedroia, it was in the center of the strike zone. Due to Pedroia's small stature, his power hits were located primarily on the inside of the strike zone; Heyward, a much taller player, was more successful on the outside. The readily apparent sensitivity of these plots to different batting tendencies speaks to the legitimacy of these graphical analyses as a potentially valuable advance scouting tool.

Analyzing Batting Patterns of Major League Baseball Players For Advance Scouting Reports:

Using R to Generate High-Level Spatial Plots of PITCHf/x Data

By

Hillary M. Superak

B.S., Washington University in St. Louis, 2009

Thesis Committee Chair: Patrick D. Kilgo, MS

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2011

TABLE OF CONTENTS

TABLES & FIGURES

INTRODUCTION

Baseball: America's National Pastime

For well over a century, baseball has been a sport played and watched by millions of people around the world. A game similar to baseball is thought to have been played in France as early as the 14th century, but modern baseball most likely descended from a popular 18th century British game called rounders (Block, 2006). By the early 1800s, bat-and-ball games resembling baseball were being played in Canada and the Northeastern United States; the first formal baseball game occurred on June 19, 1846 in Hoboken, New Jersey (Block, 2006). Over the decades that ensued, there was marked progression and growth in the world of baseball – the game was continually advancing. The popularity of the sport came to thrive – not only in terms of participation in organized leagues, but also in terms of fandom.

Despite the constant evolution and development of the game of baseball, it has maintained one particular consistency of note: in 1856, the title of "America's National Pastime" was bestowed upon baseball, and it has retained that designation ever since (Tygiel, 2001). The inherent beauty of baseball is that nearly anyone can take interest in some aspect of the game, whether it is the competitive spirit, the rich history, or the endless opportunities for strategic and statistical analyses. In the words of the esteemed baseball historian Bill James:

> *"The essential definition of baseball is that baseball is a thing which welcomes and sustains our interest. Whoever we are, however we think, however old we are, wherever we live, whatever we like to do, baseball wants us – and this is what makes baseball what it is."*

Major League Baseball

Major League Baseball (MLB), founded in 1876, is the top tier of professional baseball in the United States and Canada. The organization is comprised of 30 teams, with 14 in the American League and 16 in the National League. Since 1994, each league has consisted of three divisions: East, Central, and West. Each team plays 162 regular-season games. Four teams in each league – the three division champions, plus a wild card team (which has the next best record) – earn playoff berths. The season culminates with the World Series, a best-of-seven series between the National League and American League champions, as determined by the two earlier rounds of playoff series. Winning a World Series can have remarkable impacts on a team's city, ranging from economic benefits to improved morale (Cuellar, 2008; Delaney & Madigan, 2009).

Scouting is an important aspect of MLB. In addition to scouts who assess the talent of potential draft picks, there are also advance scouts. The role of an advance scout is to watch a team's upcoming opponents prior to the games. Advance scouts seek information not conveyed in a box score or game summary that might provide an edge to their team. Advance scouting techniques have evolved over the years, but it is still a largely subjective practice. Assessment of players' body language, tendencies, and hot or cold streaks have been of primary interest, as well as measured outcomes such as pitch speed, direction, and location. Because the art of advance scouting is so subjective, it can be a demanding and tedious job. According to Wayne Krivsky, a former advance scout and the current Special Assistant to the General Manager of the New York Mets, having a good advance scout "might be the difference in winning the division. It's one of the toughest jobs in baseball" (White, 2006).

A Numbers Game

Among sports, baseball is arguably at the forefront with respect to the potential for improving both winning percentages and talent level of rosters through the applications of statistical analysis. Each play in a baseball game is discrete, and thus can be examined in isolation from the rest of the game. Only a limited number of outcomes can occur on each play. These

outcomes can often be grouped or sub-divided into broad or specific classifications. Furthermore, when it comes to officiating, there is a lesser degree of subjectivity than is present in many other sports. The judgment of well-trained umpires, though never perfect, can be expected to have minimal impact on the outcome of a game.

Perhaps the factor of paramount significance that causes baseball to lend itself so ideally to statistics is the box score. The concept of a box score originated in 1861, when Henry Chadwick sought to summarize baseball games in a database format (Dickson, 2007). Early box scores were fairly crude and minimally informative, providing only the number of times each player got out and the number of runs he scored. Shortly thereafter, box scores were expanded to include sections for both offensive and defensive figures. In addition to runs, each player's at-bats and hits were tallied, as were defensive put-outs, assists, and fielding errors (Dickson, 2007). Chadwick later introduced the concept of a batting average, as well as the statistics Runs Batted In (RBI) and Earned Run Average (ERA) (National Baseball Hall of Fame and Museum, 2010).

Modern box scores include columns for number of: at-bats, runs, hits, RBI, walks, strikeouts, and total pitches received. There are also three calculated columns: batting average, on-base percentage (OBP), and slugging percentage (SLG). Box score data are most valuable in electronic database format, where they can be analyzed with statistical software. Information contained in box scores allows baseball to be analyzed in a hierarchical manner; statistics can be tabulated for an entire season, a single game, a particular inning of a game, or a specific at-bat.

SABR: Society for American Baseball Research

The Society for American Baseball Research (SABR) was founded in 1971. It is an organization of 7,000 members worldwide who share a passion for baseball statistics, research, and history (SABR, 2011). One of the core disciplines in which SABR members partake is known as *sabermetrics*. Popularized by Bill James, sabermetrics is "the search for objective knowledge about baseball" (SABR, 2011). Viewing baseball from an economic perspective has recently become popularized among some MLB teams. They want to use their payrolls efficiently

by trading for, and drafting, players that will be effective in helping generate wins for the least possible cost. This goal is being optimized by the performance measurement facet of sabermetrics. Not only do sabermetricians strive to assess players' value in past seasons, but they generate models and statistics that predict future value as well (Lewis, 2004).

The rise of sabermetrics and electronically recorded baseball statistics has brought about the emergence of fantasy baseball leagues. Fantasy league participants draft MLB players, and score points based on the performances of their players in several statistical categories over the duration of the regular season. Another spin on fantasy baseball is computerized simulation games. In this variation, computer programs generate results of hypothetical games between teams comprised of MLB players that are drafted by participants – called "owners" – with the intent of imitating the role of a real general manager. With approximately 11 million fantasy baseball league participants and simulation game users, it is evident that there is a great deal of interest in baseball's statistical attributes (ESPN, 2011). This is an area with ample room for development, especially as it continues to captivate more and more people. As sports columnist Thomas Boswell put it in *Inside Sports*, "More than any other American sport, baseball creates the magnetic, addictive illusion that it can almost be understood" (Quote Garden, 2010).

A New Graphical Analysis Tool for Advance Scouting

The goal of the following graphical analyses is to create an advance scouting tool to help maximize the probability of winning. In baseball, there are two main areas that can be optimized with the help of statistical analysis: player selection and game strategy. A tactical, well thought-out approach to player selection is crucial for successfully drafting the best players; however, these analyses focus on game strategy. A pitcher who is thoroughly prepared – by knowing the tendencies, strengths, and weaknesses of the batters he is slated to face – will have both physical and psychological advantages over a pitcher whose team has not scouted the opponents adequately. By providing a series of informative and user-friendly graphical analyses, the aim of this project is to enhance the existing MLB advance scouting reports that are available to pitchers.

CURRENT METHODS & TECHNOLOGY

PITCHf/x

In 2006, Sportvision, a company that specializes in enhanced viewing of professional sporting events, developed a system called PITCHf/x. Every Major League Baseball stadium is now equipped with two high-speed cameras that track the ball over its entire trajectory from the pitcher's hand to home plate. Every pitch that is thrown is recorded by the PITCHf/x system. The system's ability to capture 60 measurements per second lends itself to a high degree of precision (Nathan, 2008). The PITCHf/x data include the three-dimensional spatial coordinates of the ball's trajectory, so the location of each pitch can be pinpointed as the ball crosses home plate. Also captured by the system are the upper and lower boundaries of the strike zone for each at-bat. The strike zone – the imaginary box encompassing the area from the batter's knees to his shoulders, in between the side boundaries of home plate – varies by player due to differences in stature and batting stance. Other parameters measured by the PITCHf/x system include the ball's initial and final velocities, spin direction and rate, and break angle and direction. Based on calculations involving these parameters, the system is able to classify the type of each pitch – although there is some inherent degree of uncertainty associated with this algorithm (Fast, 2010).

The abundance of data provided by PITCHf/x has opened up new pathways for data analysis at a higher level than what was previously possible. For the first time, PITCHf/x allowed for examination of the characteristics of each individual pitch. This breakdown added another level to the hierarchy of baseball data available for analysis, more refined than analyses at the season-long, single-game, single-inning, or single-at-bat levels. So extensive is the range of possible applications for these data that new relevance continues to be realized, even several years after the PITCHf/x data were first released (Boudway, 2011).

Graphics

To a certain degree, analysts of baseball data have been able to utilize the PITCHf/x data in graphical analyses. However, the surface has barely been scratched considering the amount of information available and the realm of conceivable possibilities. Thus far, most graphical displays of the PITCHf/x data have focused on the actions of the pitcher, as opposed to the batter.

Many of the existing graphics simply display frequencies of interest. For example, a team may be interested in predicting the type of pitch that will be received from the opposing pitcher in a certain game situation. This concept can be displayed in a relative frequency display of, for example, curveballs, changeups, two-seam fastballs, and four-seam fastballs thrown by a single pitcher, categorized by pitch count (*Fig. 1*) (Fast, 2010).

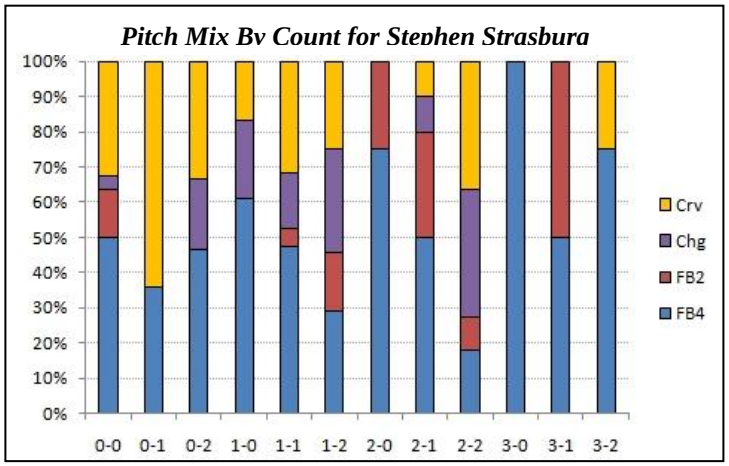

Figure 1. *Example of a relative frequency display using PITCHf/x data.*

Some plots have been created that utilize two continuous axes. The data can further be classified into discrete categories. These plots have been created in both polar and rectangular coordinate systems. Polar coordinates are desirable in cases where one of the continuous variables is measured on a cyclical scale. For example, a polar plot could show the data for one pitcher's pitch speed vs. spin axis angle (*Fig. 2*) (Fast, 2010).

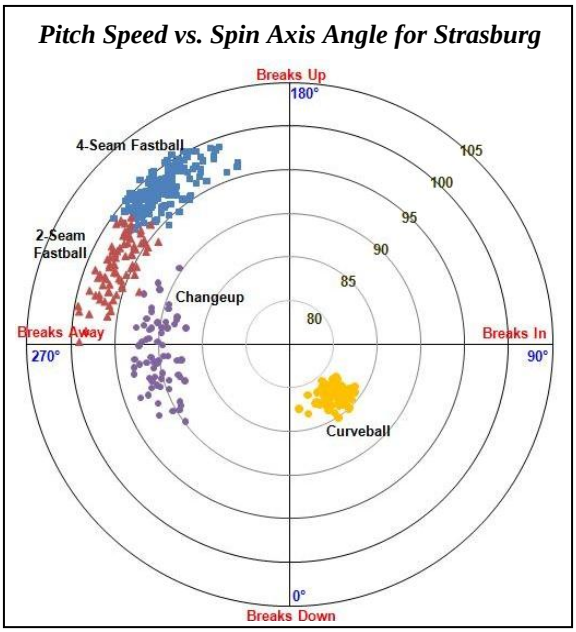

Figure 2. *Example of a polar plot using PITCHf/x data.*

More complex PITCHf/x graphics have taken advantage of the spatial aspect of the data. Plots of pitches according to their horizontal and vertical coordinates as they cross home plate are most valuable to MLB data analysts. This information can be readily extracted from the three-dimensional coordinate vectors: since the center of home plate represents the origin of the three-dimensional coordinate system, the coordinate corresponding to the distance from home plate towards the pitcher is simply set to zero. Some plots further categorize the plotted data points to probe more deeply into the information that is available. Useful categorizations include – but are not limited to – pitch type, pitch outcome, and number of bases attained by the batter.

In particular, two types of plots have been developed that can serve as analytical tools regarding batters' actions. A hit chart is a concise, yet descriptive, way of plotting the pitch locations corresponding to a batter's base hits (*Fig. 3*). Classification of the base hits according to the number of bases attained can help show spatial trends in a batter's power. A strike chart can provide information about where a batter is most likely to swing at, and miss, a pitch (*Fig. 4*). Taken together, these two charts reveal a batter's past actions for all at-bats (Kalk, 2007).

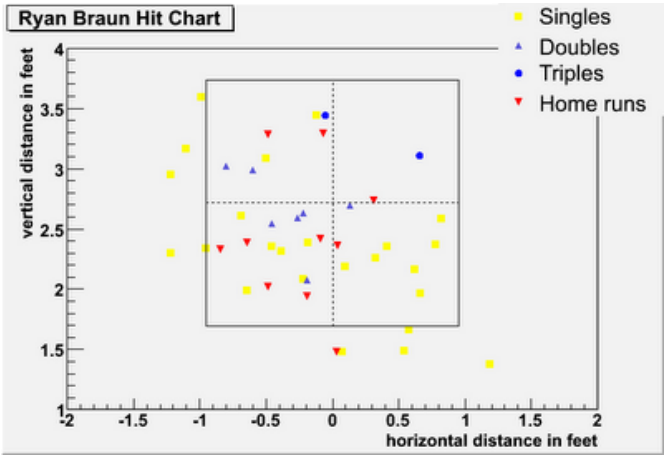

Figure 3. *Example of a hit chart using PITCHf/x data.*

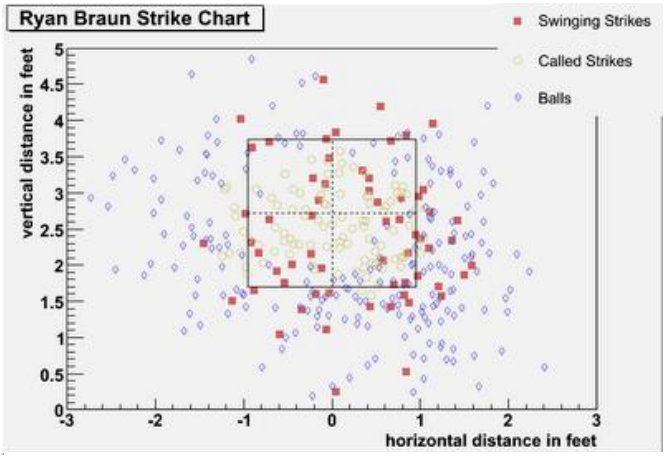

Figure 4. *Example of a strike chart using PITCHf/x data.*

Similar spatial plots have also been constructed for the actions of pitchers. For example, it may be of interest where a certain pitcher tends to throw pitches that result in favorable (or unfavorable) outcomes. *Figure 5* illustrates the outcomes of all four-seam fastballs thrown by a particular pitcher, stratified by the handedness of the batters. As a reference point, a box representing the strike zone is overlain (Fast, 2010).

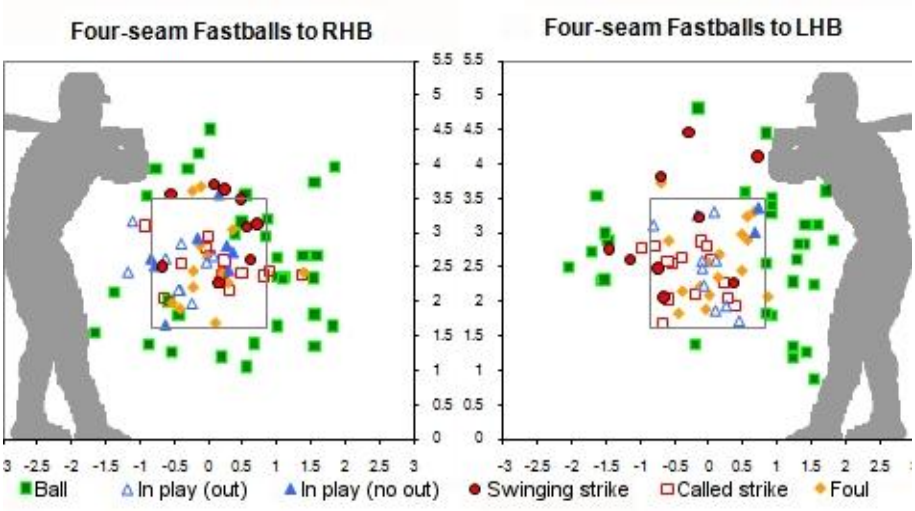

Figure 5. *Example of a spatial plot of the pitcher's actions using PITCHf/x data.*

Areas for Improvement

While there has been a substantial amount of graphical analysis performed on the PITCHf/x data, there are still further steps that can be taken to enhance these analyses. Advances in graphics packages of statistical software programs continue to add to the growing list of options that are available. Additionally, new perspectives on the use of baseball statistics are contributing innovative ideas that had not been previously conceived.

The sample plots previously shown can be effective for small sample sizes. However, over the duration of an entire season, many batters receive several hundred pitches; likewise, pitchers may throw several hundred pitches. Season-long graphical analyses could be desirable for comparison of a player's tendencies from year to year. These graphics could reveal either consistency in batting or pitching patterns, or change from season to season according to some sort of trend. Using an entire season's worth of data is also advantageous in that it provides a

large sample size. Generally, the more observations included in a data set, the more precise and reliable the statistical analyses will be. Attempting to plot several hundred data points, representing an entire season's worth of data, in one figure would produce results that are hard to interpret and visualize. Many points would likely lie so close together that they would be indistinguishable from one another, and different shapes representing the levels of a categorical variable could become indistinct. Furthermore, color-coding would be ineffective if the points overlaid one another and the earliest plotted colors did not show through. A spatial plotting method that also displays categorical data in a more user-friendly manner would help mitigate this issue.

Another way to enhance existing plotting methods would be to use a series of contour plots or density plots. These are both three-dimensional spatial plots. Each data point is plotted according to its horizontal and vertical coordinates, and the third dimension is expressed as the density of data points falling within certain areas. As opposed to the discrete scatterplots previously cited, contour plots and density plots display data in a continuous fashion. Thus, blurring among points that are located near each other is not a concern for these types of plots. Points overlaying one another are not problematic either, since the purpose of these plots is to express the density. It could be beneficial to view these plots in a series if there were different inclusion criteria for the data points in each plot (e.g., compare a contour or density plot of all pitches received by a player with an analogous plot of all pitches swung at by the same player).

NEW DEVELOPMENTS

Advance Scouting of Batters

The objective of this project is to develop a tool for the advance scouting of Major League Baseball batters. This aim can be accomplished through various graphical analyses of PITCHf/x data, providing information beyond what is currently available. The PITCHf/x data, along with secondary information that can be derived from the data, are quite comprehensive; a systematic algorithm is the best approach to these analyses.

Ideally, the information contained in these graphical analyses will relieve some of the pressure that burdens advance scouts. Statistics can provide objective measures of batting tendencies, as well as reveal areas of strength and weakness – information that is not readily conveyed in a box score or easily conceptualized and memorized by the human mind. Part of the advance scout's job that is especially rigorous is the constant travel. As the field of sabermetrics and the associated graphical analyses continue to develop, this aspect of advance scouting may diminish or even disappear in the future.

Hexbin Pie Charts

Hexbin pie charts are a method of graphically displaying two-dimensional spatial data points that are categorized by a discrete variable. The process of generating a hexbin pie chart involves first creating a simple x-y scatterplot of the data. Then a grid of equilateral hexagons, all of the same size, is overlain. Within each hexagon that contains data points, a pie chart is drawn. The radius of each pie chart is proportional to the number of data points that lie within the given hexagon. In other words, larger pie charts represent areas of greater density, and vice versa. Each pie chart shows the relative frequencies of the points inside that hexagon, as classified by the levels of the categorical variable of interest. In some cases, the densest areas of a plot will have multiple hexagons containing pie charts that fill the entire area of the hexagon. When this

phenomenon occurs, the number of data points contained in the hexagons with maximal-radius pie charts is displayed at the center of the pie chart.

Hexbin pie charts are a relatively new graphical technique, and are not included in the standard graphics packages that come with most statistical software programs. R does not contain a built-in command for generating hexbin pie charts; rather, they are created with a user-defined function. This function was programmed in 2006 by Romain François, but it has not been widely utilized. The lack of use can be partially attributed to the fact that the function is programmed to use simulated data from known distributions – the function must be modified in order for the user to input real data (François, 2006).

One feature that makes hexbin pie charts an optimal plotting method for the PITCHf/x data is that they can handle very large data sets. In contrast to the scatterplots that are currently used, hexbin pie charts do not present any issues in dealing with overlapping or hidden data points. Furthermore, hexbin pie charts make it easy to spot spatial trends in the observed frequencies and relative frequencies of the categorical variable values – large images and distinct color contrasts tend to be most eye-catching. As such, they tend to be more visually appealing than simple scatterplots. Although deciphering François's hexbin pie function is a relatively complex task, it can be adapted to increase the degree of user-friendliness. For example, the hexagon size can be adjusted so that more or fewer data points are contained in each hexagon. This adjustment can help produce a favorable ratio of filled hexagons to blank space. The benefit of working with user-defined functions is that any number of parameters can be adjusted, or variables added, to customize the output.

Contour Plots

Contour plots display two-dimensional spatial data that are associated with a third dimension. The classic use of contour plots is in topographical maps, which show elevation as a third dimension yet remain planar in form. Each line on a contour plot connects points where the function representing the third dimension has the same value. The numbers associated with the

contour lines can be interpreted in terms of relative risk. For example, a data point would be 5 times more likely to fall inside a contour line labeled 0.50 as it would be to fall within a contour line labeled 0.10 (but outside of the next contour line). Contour lines that are very close together signify more tightly clustered (higher density) data points than contour lines that are more spread out. This plotting method is useful for PITCHf/x data because it can generate separate plots to show, upon quick glance, the pitch locations corresponding to where the batter is most likely to swing, put a ball into play, or get a base hit.

Density Plots

Density plots, like contour plots, create two-dimensional representations of three variables – two of which are comprised of vectors of spatial coordinates. A density plot illustrates an intensity function, which is the expected number of random points per unit area, based on the underlying data set. Instead of contour lines, a density plot uses a continuous color spectrum to represent the third dimension. Red typically indicates areas of the highest intensity – the "hotspots" – while blue indicates low intensity. For the PITCHf/x data, the density plots can be used much like the contour plots. Again, the main advantage of these plots is that their content can be readily interpreted based on a quick visual inspection; they are not complicated, nor do they involve the use of numerous symbols.

METHODOLOGY

The complete SAS and R programs are provided in Appendix A and Appendix B, respectively.

To aid in reading this section, color coding has been used as follows:

- R libraries (green)
- R built-in functions (blue)
- R user-defined functions & variables (red)

The Data Set

The data set was comprised of the PITCHf/x system output for all pitches thrown during the 2010 Major League Baseball regular season. There were 710,329 observations (pitches), and 41 variables, many of which were not utilized for these analyses.

Data Cleaning and Variable Manipulations

Data cleaning and variable manipulations were conducted using SAS® V 9.2 Software. The positional data (horizontal and vertical spatial coordinates) were missing for 4,171 of the pitches. These observations were excluded from the analyses, leaving 706,158 complete records. The original PITCHf/x coordinates are recorded from a vantage point facing the pitcher. Thus, the horizontal coordinate of each pitch was multiplied by -1 to make the plotted data viewable from the pitcher's perspective (since the primary objective is to scout the batters). The data in their original format were de-identified, with each player represented by a unique six-digit ID number. The PITCHf/x data were merged with another data set that identified each player by first and last name. The first and last names, originally stored as separate variables, were concatenated into a single variable containing the players' full names.

Table 1 presents a list of the PITCHf/x variables that were retained for analyses; variables that were deemed to be unnecessary were dropped[1]. Several new variables, presented in *Table 2*, were created based on the existing data. For categorical variables, the classifications provided by the PITCHf/x system are very specific. However, visual representation of categorical data often conveys information most effectively with as few groups as are necessary. Creation of the new variables helped classify the pitch and at-bat outcomes into categories that were more useful for plotting purposes. Furthermore, the pitch types were re-grouped into four broader categories. The goal is for the pitcher to have a general sense of the optimal pitch speed, spin, and location; he can customize the pitch into a more particular type according to personal preferences. For example, the pitch types *cut fastball, four-seam fastball, two-seam fastball,* and *sinker* were all classified into a single pitch category called *fastball* due to the similarities in their typical speeds, spin rates, break angles, and trajectories. Different pitchers may individualize their pitching techniques within the broad category of fastballs, but the general term "fastball" will universally convey the same meaning to all pitchers.

As a quality method of ensuring the proper coding of new variables, frequencies were tabulated for the different levels of the categorical variables. These frequencies were compared with the official statistics provided by an ESPN Major League Baseball database (ESPN, 2011). In some cases, there were slight discrepancies between the true frequencies and those obtained from the newly coded variables. These differences can be attributed to the missing data in the PITCHf/x records. The final data set was exported from SAS and saved as a Microsoft Office Excel 2007 CSV file, a format that can be read into other software packages.

[1] Unnecessary variables for these analyses included the pitcher's name, umpire's name, game site, inning number, pitch number, and certain characteristics of the pitch (ball's spin direction, spin rate, break angle, and break length).

Table 1. *PITCHf/x Variables Retained for Analyses.*

Variable Name	Variable Definition	Values	
atbat_result	At-bat result	Batter Interference Bunt Groundout Bunt Pop Out Catcher Interference Double Double Play Fan Interference Field Error Fielders Choice Fielders Choice Out Flyout Forceout Grounded Into DP Groundout Hit By Pitch	Home Run Intent Walk Lineout Pop Out Runner Out Sac Bunt Sac Fly Sac Fly DP Sacrifice Bunt DP Single Strikeout Strikeout – DP Triple Triple Play Walk
batter	6-digit batter ID number		
pitch_result	Pitch result	Automatic Ball Ball Ball in Dirt Called Strike Foul Foul (Runner Going) Foul Bunt Foul Tip Hit By Pitch In play- no out	In play- out(s) In play- run(s) Intent Ball Missed Bunt Pitchout Swinging Pitchout Swinging Strike Swinging Strike (Bloc Unknown Strike
pitch_type	Pitch type	CH (Changeup) CU (Curveball) EP (Eephus Pitch) FA (Fastball) FC (Cut Fastball) FF (Four-Seam Fastball) FO (Forkball) FS (Split-Finger Fastball)	FT (Two-Seam Fastball) IN (Intentional Walk) KC (Knuckle Curve) KN (Knuckleball) PO (Pitchout) SC (Screwball) SI (Sinker) SL (Slider)
px	Horizontal coordinate of the ball (ft) as it crosses home plate	0 is at the center of home plate. Negative values are to the left of home plate, as seen from the batter's perspective.	
pz	Vertical coordinate of the ball (ft) as it crosses home plate	0 is at ground level	
sz_bot	Bottom of strike zone	Feet above ground level	
sz_top	Top of strike zone	Feet above ground level	

Table 2. *New Variables Created for Plotting Purposes.*

Variable Name	Variable Definition	Variable Creation
BIP	1 = Ball put into play 0 = Ball not put into play	*pitch_result* = Otherwise → *BIP* = 0
swing	1 = Batter swung at pitch 0 = Batter did not swing at pitch	*pitch_result* = Otherwise → *swing* = 0
good	1 = Good outcome for pitching team 0 = Bad outcome for pitching team	*pitch_result* = Otherwise → *good* = 1
basehit	1 = Batter got on base 0 = Batter did not get on base	*BIP* = 1 AND *good* = 0 → *basehit* = 1 Otherwise → *basehit* = 0
basehit_type	Type of base hit	*pitch_result* = - AND *atbat_result* = Single → *basehit_type* = Single *atbat_result* = Double → *basehit_type* = Double *atbat_result* = Triple → *basehit_type* = Triple *atbat_result* = Home Run → *basehit_type* = Home Run Otherwise → *basehit_type* = No Base Hit
pitch_cat	Pitch category	*pitch_type* = FC, FF, FT, SI → *pitch_cat* = fastball *pitch_type* = CU, KC, SL → *pitch_cat* = breaking ball *pitch_type* = CH → *pitch_cat* = changeup Otherwise → *pitch_cat* = other

Preparation for Plotting

All plots were generated using R 2.11.1 Software, and the program was composed using Tinn-R. Prior to generating high-level spatial plots, several R add-on packages had to be installed:

- *dichromat*
- *graphics*
- *KernSmooth*
- *sp*
- *spatial*
- *spatstat*
- *spdep*
- *splancs*

The data set was cropped before any plots were created so that pitches included for analysis were limited to those that crossed home plate within the 4-foot by 4-foot area extending 2 feet on either side of home plate and ranging from 1 foot to 5 feet above the ground. These specifications were chosen because they provide a sufficient margin surrounding any player's strike zone, such that the majority of all pitches to any player fell within the designated area. A batter would not be expected to swing at, or hit, a pitch *not* falling within this area; thus, these pitches are not of primary interest. The data were sorted according to batter name. It was necessary to order them in this manner for later sub-setting of the data to examine only observations for a single batter of interest at a time.

Plotting Function

The aim of this project was to create a user-friendly tool that will generate spatial plots to be utilized for batter scouting purposes. As such, it was ideal to write a function requiring minimal input from the user. The function that was created, baseball.plot, requires only the batter's first and last name to be entered when it is called. This input automatically restricts the data set to only pitches thrown to the batter of interest.

Also to promote user-friendliness, the program was written to automatically export all of the plots from R and save them as pdf files. Therefore, the plots can be opened one at a time, or multiple plots can be compared side by side. Furthermore, the plots can be viewed at a later time, even after the R session has been terminated. A new folder called "graphs" was created in the user's "My Documents" directory. Every time a figure is saved, it goes to this folder. Each

different type of plot has its own descriptive file name of the format "*PLOT TYPE universe level – classification – batter name*". So, for example, a hexbin pie chart of all pitches thrown to Jason Heyward, classified by whether or not he swings, would have the file name "*HEXBIN All Pitches – Classify by Swing or No Swing – Jason Heyward*". The R paste function was used to insert the batter's name into each file name. Thus, the plots are easily identifiable once they are saved. As a further benefit, data from multiple players can be plotted in the same R session, and the saved pdf files will not be overwritten to only store information on the most recent batter called by the baseball.plot function.

The function generates three different types of spatial plots:

1. Hexbin Pie Charts
2. Heat Map Contour Plots
3. Heat Map Density Plots

There is a four-level "universe" of outcomes for the pitches (*Fig. 6*). The broadest category includes *all pitches*. The next smallest category is *swings only*, followed by *balls in play (BIP) only*, and finally *base hits (BH) only*.

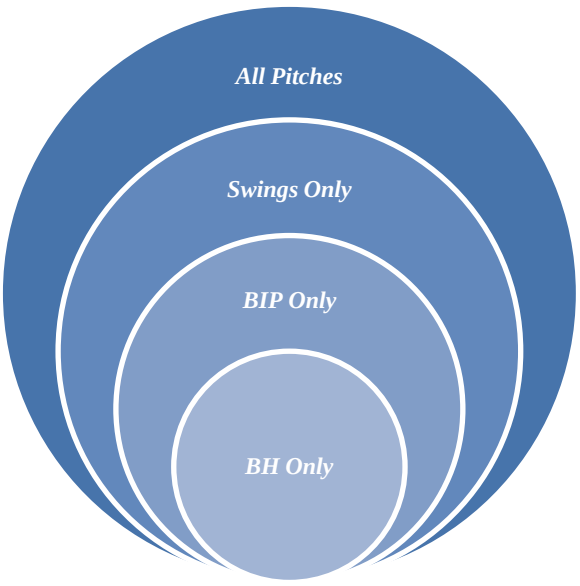

Figure 6*. Diagram of the four-level pitch universe.*
(Figure not proportional to actual frequency distribution.)

Notably, there is a hierarchical structure to the universe of pitch outcomes: in order to be put into play, a pitch must be swung at; in order to achieve a base hit, the batter must put the ball into play. For the three broadest categories, the pitches can be further classified by whether or not the outcome is good for the pitching team. (The outcome of a base hit is never good for the pitching team.) The baseball.plot function was designed to subset the data according to the levels of the pitch universe so that the three types of plots can be generated at each level.

For all plots, the strike zone was overlain by adding line segments to form a rectangle. A pitch should be classified as a strike if *any* part of the ball is inside the strike zone. The side boundaries of the strike zone correspond to the side coordinates of the 17-inch wide home plate. A standard baseball is 3 inches in diameter, and the PITCHf/x cameras track the center of the ball. Therefore, a 1.5-inch margin was added to either side. This margin, added to the 8.5 inches on either side of the center of home plate, results in a 10-inch distance to the left and the right. Consequently, the strike zone extends 0.833 feet to either side of the center of home plate. The top and bottom boundaries of the strike zone were different for each player due to differences in height. These coordinates were calculated as the means of the measurements for the top and bottom of the strike zone for the player of interest. The 1.5-inch margin of error was then accounted for on both the top and the bottom.

Hexbin Pie Charts

The existing hexbin pie function was programmed to use simulated data from a normal distribution as the horizontal and vertical coordinates, and simulated data from a binomial distribution as the categorical variable. The function was modified to be able to read in an actual data set containing horizontal and vertical spatial coordinates. Furthermore, the categorical variable input was altered to allow for more than two levels to be represented in the pie charts.

An additional variable, xbin, was initiated and inserted into the function, allowing the hexagon size to be manipulated. By default, the function produced 20 hexagons spanning the width of the horizontal axis. However, the data points were less dense in the upper levels of the

pitch outcome universe. Instead of using the same hexagon size for all charts, it was more visually appealing to double the hexagon width for plots of *base hits only* and *balls in play only*. In other words, there were 20 hexagons spanning the horizontal axis for plots of *all pitches* and *swings only*, but only 10 for plots of *balls in play only* and *base hits only*. Manipulation of this parameter helped eliminate much of the blank space that could potentially have distorted some of the plots.

The modified hexbin pie function partitions the plot area into hexagons – called bins – according to the number specified by the variable xbin. A hexagon outline is drawn for any bin into which one or more data points fall. If no data points are located in a bin, that area is left blank and no hexagon is drawn. Each hexagon contains a circular pie chart inside, with the radius of the circle proportional to the number of data points falling inside that hexagon. If a bin contains so many data points that the pie chart fills the entire hexagon, then the number of data points is displayed. This feature was designed to help the viewer distinguish among the spatial densities while maximizing the information conveyed within the given plotting area.

Heat Map Contour Plots

As a preliminary step towards generating contour plots, the as.points function from the splancs library was used to combine into a single vector the two vectors containing the horizontal and vertical coordinates of the pitches. Next the bkde2d function in the KernSmooth library was used to generate a matrix of density estimates for each area of a 40x40-unit mesh grid. The grid size, in this case 40 points in each direction, can be specified by the user prior to plotting in the variable grid.number. The bkde2d function also employed a kernel smoothing estimation technique in order to produce continuous contour lines. The contour function from the graphics library generated the actual three-dimensional plots, based on the spatial coordinates and the density estimates.

By default, the contour lines were all black. Because the objective was to illustrate the batter's "hotspot," the lines were re-colored according to the rev(heat.colors) function. The

heat.colors function, also in the graphics library, generates colors along the spectrum of red, orange, and yellow shades. The default setting is for the function to be applied starting at the outermost contour lines (i.e., red at the lowest densities and yellow at the highest densities). The function was called in reverse so that the highest densities (corresponding to the batter's hottest spots) would appear in red, and the lowest densities would appear in yellow.

Heat Map Density Plots

The heat map density plots were created in a similar manner as the heat map contour plots. Analogous to the first step previously described in creating the contour plots, the two coordinate vectors first had to be combined. For the density plots, this step was accomplished by using the ppp function from the spatstat library. This function provides a single object in the two-dimensional plane that represents the two vectors of a point pattern data set. The density.ppp function, also from the spatstat library, then added a third dimension by computing a kernel smoothed intensity function based on the two-dimensional point pattern.

The built-in R function heat.colors that was used for the heat map contour plots did not provide adequate coloring when used in the heat map density plots generated by these particular data. The coloring on the outer areas of the plot was very faint, and it was difficult to distinguish some of the orange and yellow shades. To resolve this problem, a user-defined color function was added to the program. The new function, hs.color, employs the colorRampPalette R function contained in the dichromat library. The colorRampPalette function allows the user to specify, by color name, any number of discrete colors in a certain order. Consequently, a continuous set of interpolated colors spanning the input discrete spectrum is generated. The hs.color function can be called any time a color is needed. The user specifies the number of shades to be utilized, and that many shades, evenly spaced, are taken from the continuous interpolated color spectrum. For these heat map density plots, 100 shades were used, spanning the spectrum of colors interpolated between blue, yellow, orange, and red. Blue represented the lowest density areas, and red represented the highest density areas, with yellow and orange falling in between. The use of 100

shades was sufficient to achieve an appropriate level of blending among the density levels. Too few shades would result in rough, boxy borders between the levels; too many shades would create a blurred pattern.

The density plots were generated by the R plot function. Alongside each plot is a legend that identifies, on a continuous spectrum, the estimated data point frequencies per unit of plot area, corresponding to each color in the plot. Since the addition of the legend skewed the plot area, the default locations of the plot title and horizontal axis label were off center. These labels were deleted from the plot command, and re-generated with the text command. The text command allows the exact coordinates of the labels to be specified, so this method was used as a way of re-centering the plot labels.

Validation

Due to variation in batting patterns among different players, it was deemed important to test the plotting function on several players. This validation process helped ensure that the selected values of the plotting parameters – axis limits, line widths, color spectra, legend placement, and grid size, among others – would result in visually appealing figures regardless of the batting pattern. The players used for test plotting were intentionally selected to have a variety of batting patterns. These patterns can differ in two major regards: location and precision.

First, different locations of batting hotspots were checked. The data were plotted for many batters, both right- and left-handed. For example, right-handed batter Derek Jeter of the New York Yankees tends to swing at, and hit, pitches that are centrally located with respect to the strike zone. In contrast, the Atlanta Braves' Jason Heyward, a left-handed batter, has the greatest swing and hit densities in the lower outside corner of the strike zone. Of primary interest was that no part of the figure fell outside of the plotting region; it was necessary to avoid inadvertently cropping out any data points.

In terms of batting precision (i.e., swinging at, and hitting, pitches in the same location), Jeter's data points tend to be tightly clustered and fall largely within the strike zone. This pattern

differs from the batting trend exhibited by Vladimir Guerrero, a right-handed batter formerly of the Texas Rangers. Guerrero does not have a certain spot where he can be predicted to swing or hit; his data points are scattered around a relatively large area encompassing the strike zone. Validation of the plots with respect to batting precision enabled the contour levels and hexbin widths to be set to appropriate values for each level of the previously described pitch universe. This process also enhanced the coloring of the plots, since the number of shades to be used in each plot depends upon the range of densities that are to be examined.

RESULTS & INTERPRETATION

This plotting tool can be used to analyze the batting tendencies of any non-rookie Major League Baseball player. It performs best for players who bat often, since larger sample sizes work better with the smoothing techniques that underlie the heat map contour and density plots. In all, the baseball.plot function produces 22 different plots each time it is called. Following are the results for two players, along with analyses and interpretations. These two players were intentionally chosen because the author and committee members working on this paper are very familiar with their batting tendencies. Thus, instead of producing novel analyses with a "shot-in-the-dark" approach, this section is designed to be a verification of trends that were already hypothesized. The face validity of the plotting tool can be confirmed by comparing the plot results with the authors' expectations. In terms of an actual MLB advance scouting application, this approach is the most realistic. Before utilizing this plotting tool, advance scouts will already know something about the players of interest – whether it be from box score data, observation during past games, or word of mouth.

Jason Heyward and Dustin Pedroia were selected for the following sample graphics because they are both well-known players who had several hundred at-bats during the 2010 regular season (520 for Heyward; 302 for Pedroia). In many respects, they have different batting tendencies, which makes for interesting comparisons between the two players. The figures are presented in an alternating fashion for the two players, allowing for easily accessible visual comparisons. The degree of divergence between the players' results for each plot type will illustrate the sensitivity of the graphical analyses.

Jason Heyward

Jason Heyward, drafted in 2007, made his major league debut for the Atlanta Braves in 2010. He is a 6'5" right fielder who hits left-handed. Heyward's career batting average is 0.276. In his rookie season, he hit 18 home runs, including one in his first at-bat. He was named to the National League (NL) All-Star team as a reserve, and he ultimately finished in second place in the NL Rookie of the Year voting (Sports Reference LLC, 2011).

Dustin Pedroia

Dustin Pedroia is a 5'9" right-handed second baseman on the Boston Red Sox. He was drafted in 2004 and first saw limited major league action at the end of the 2006 regular season. In 2007, his official rookie season, he won the title of American League (AL) Rookie of the Year. The next year, 2008, he earned the prestigious title of AL Most Valuable Player (MVP) – receiving more than twice as many votes as the runner-up. Pedroia has hit 55 career home runs, with a career batting average of 0.306. He has been named to the AL All-Star team, either as a starter or a reserve, for each of the past three seasons (Sports Reference LLC, 2011).

Note for Plot Analysis

The vertical span of the strike zone corresponds to the distance between the batter's knees and shoulders. The 8" height discrepancy between these two players causes their strike zones to be located at different vertical coordinates. Furthermore, Heyward's strike zone is more elongated than Pedroia's since he is so much taller. The plots are viewed from the pitcher's perspective, so the left-handed Heyward would be standing on the left side, while the right-handed Pedroia would be standing on the right side.

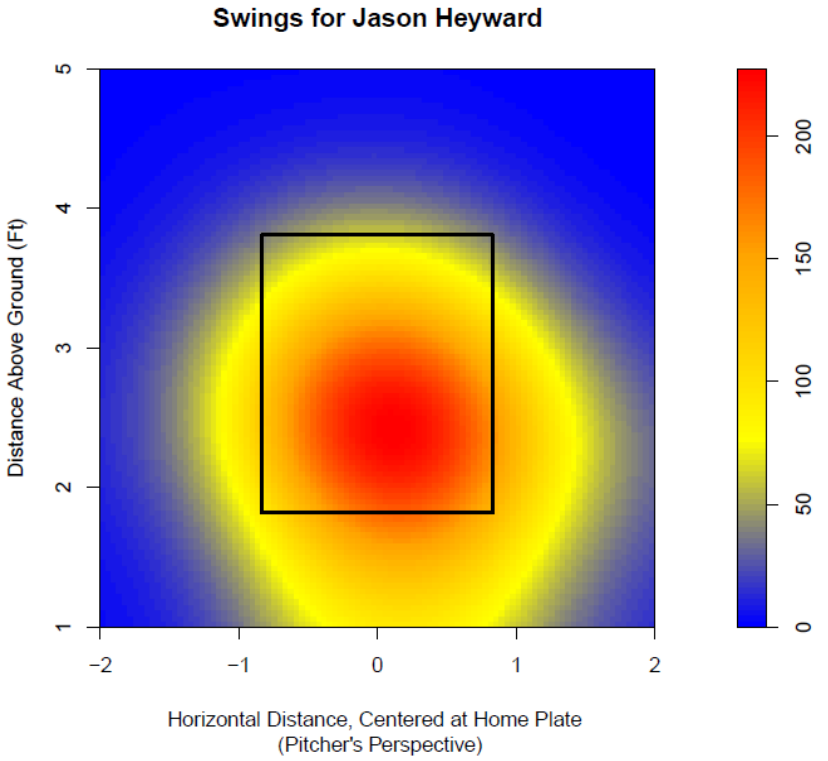

Figure 7. *Heat map density plot for Heyward's swings.*

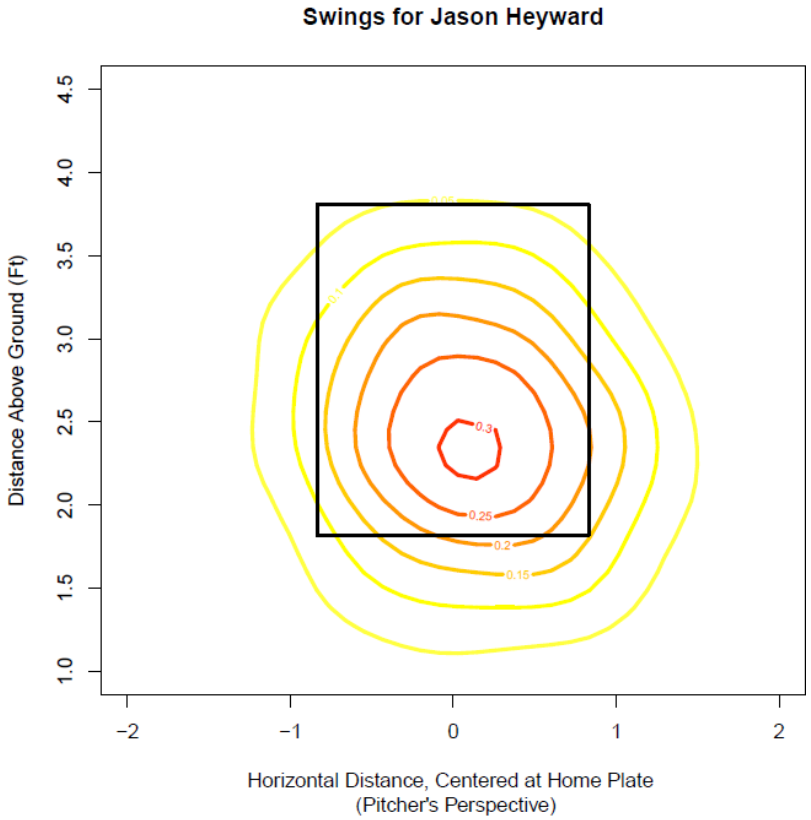

Figure 8. *Heat map contour plot for Heyward's swings.*

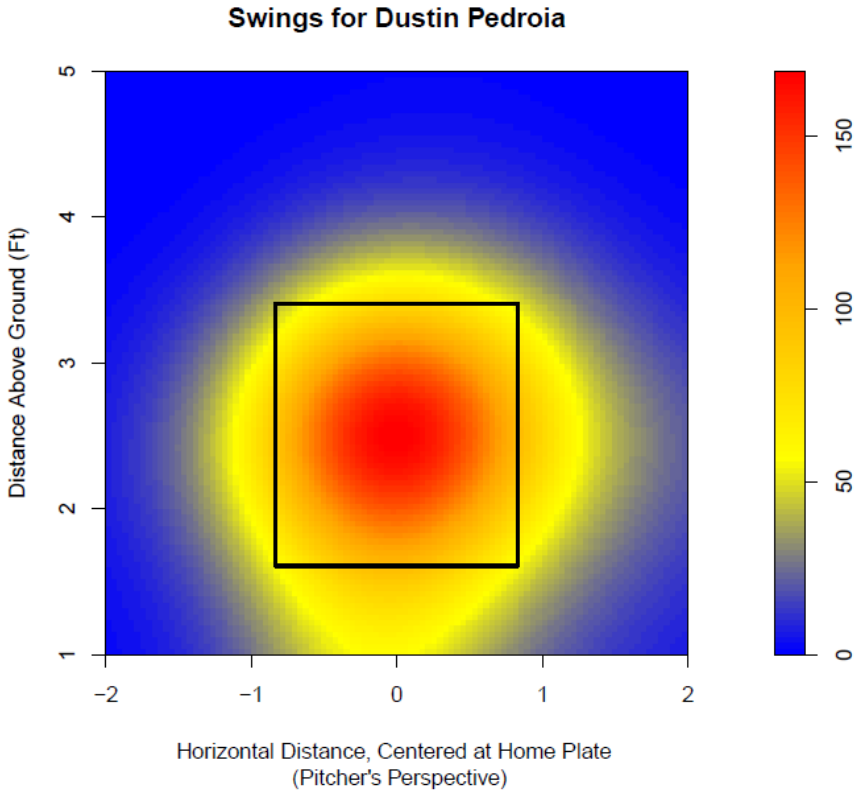

Figure 9. *Heat map density plot for Pedroia's swings.*

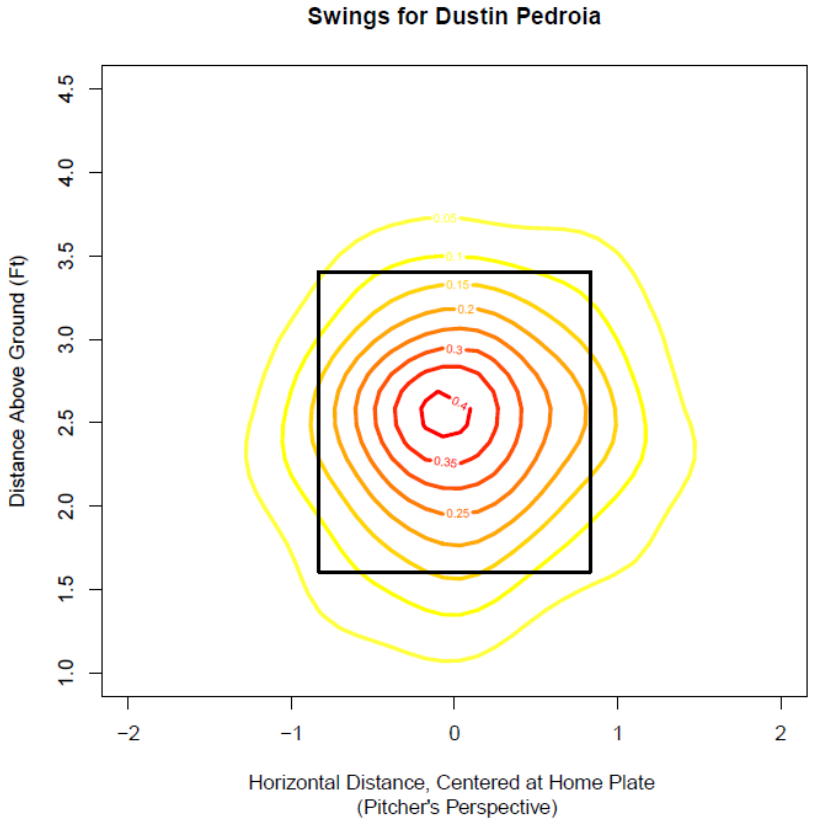

Figure 10. *Heat map contour plot for Pedroia's swings.*

As indicated in Heyward's heat map density and contour plots, he tends to swing at the most pitches in the lower outside corner of the strike zone. While he will swing often at low pitches, it is very rare for him to swing at anything above shoulder level (i.e., above the strike zone). In fact, Heyward is roughly twice as likely to swing at a pitch in the lower outside quadrant of the strike zone, compared with the upper inside quadrant (*Figs. 7 & 8*).

The heat map density and contour plots of Pedroia's swings show a hot spot that is essentially in the center of the strike zone. Pedroia can be expected to swing with relatively equal frequency at pitches that are wide of the strike zone on either side, although slightly more on the near side. He swings slightly more often at pitches that are lower in the strike zone compared with pitches that are high of the strike zone (*Figs. 9 & 10*).

Comparison of the heat map density and contour plots for Heyward and Pedroia reveals a stark difference in the two players' preferred swinging areas. A pitcher seeking a called strike against Heyward should aim for the upper area of the strike zone since Heyward is least likely to swing in this area. Pedroia's swings are more evenly distributed throughout the strike zone, so it may be more difficult to attain a called strike against him. However, he swings at a lot of pitches below the strike zone, so this may be a good area for a pitcher to target. In order to know whether these swings tend to result in favorable outcomes for the pitcher or not, additional plots must also be examined. The contour lines produced by Pedroia's swinging habits are closer together than Heyward's. This detail indicates that Pedroia can be considered a more disciplined swinger; the locations of the pitches at which he swings are more tightly clustered.

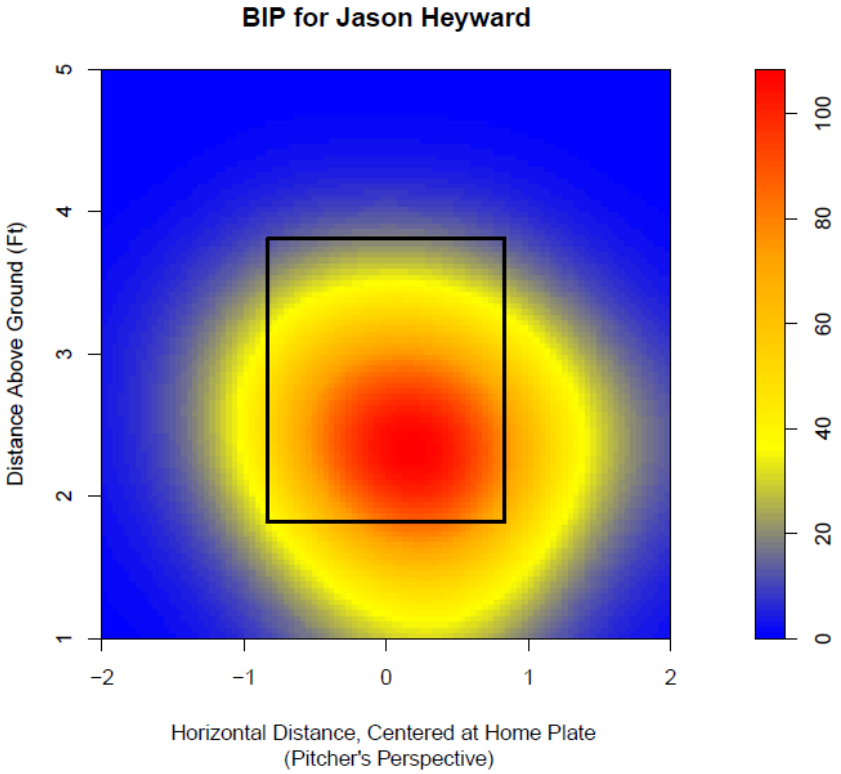

Figure 11. *Heat map density plot for Heyward's balls in play.*

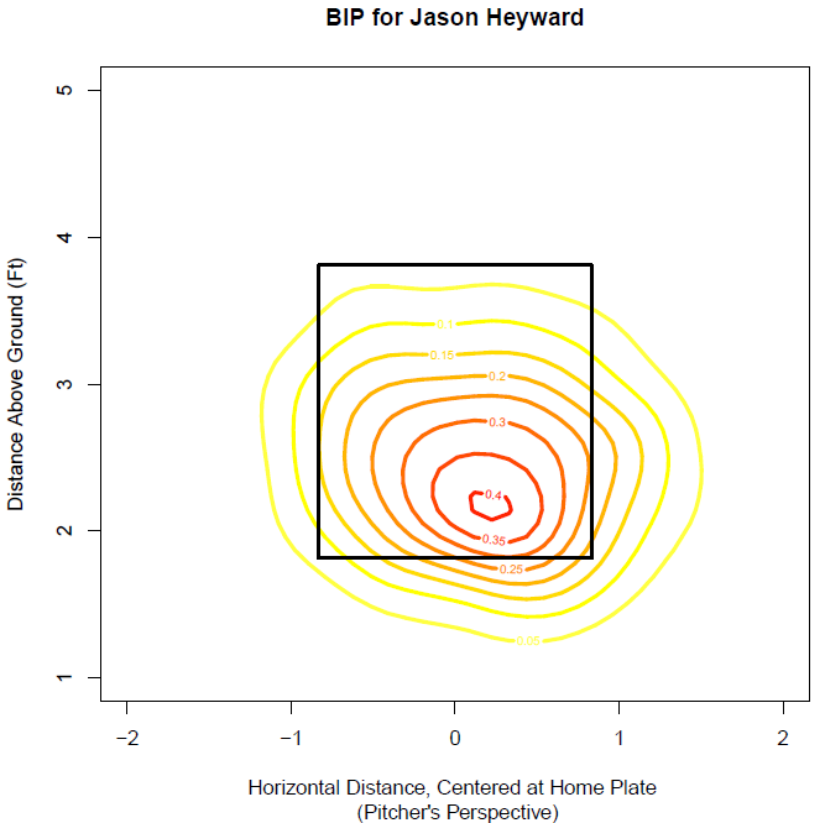

Figure 12. *Heat map contour plot for Heyward's balls in play.*

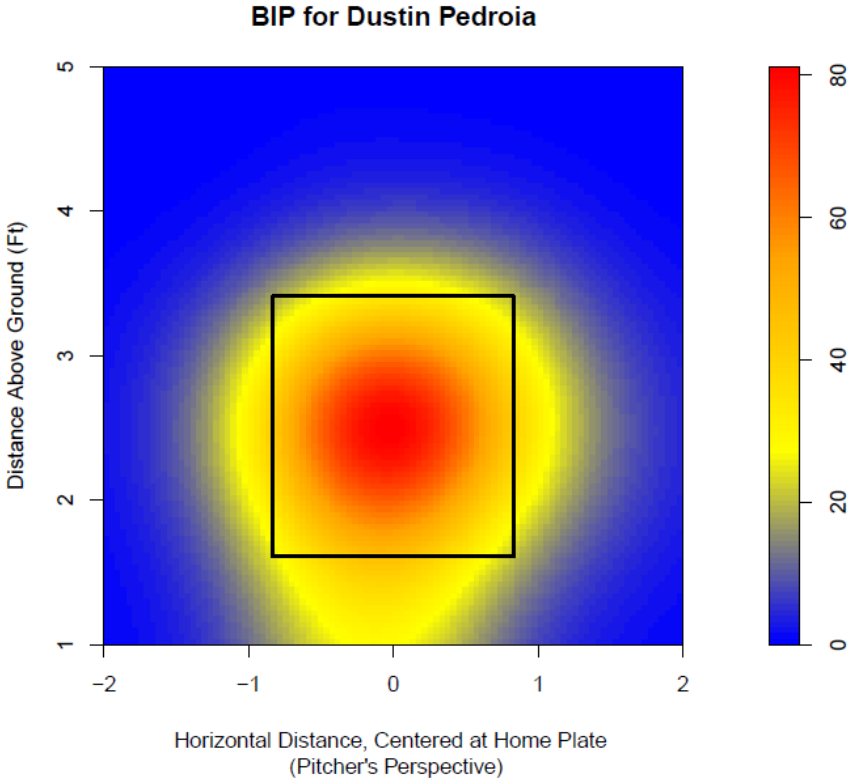

Figure 13. *Heat map density plot for Pedroia's balls in play.*

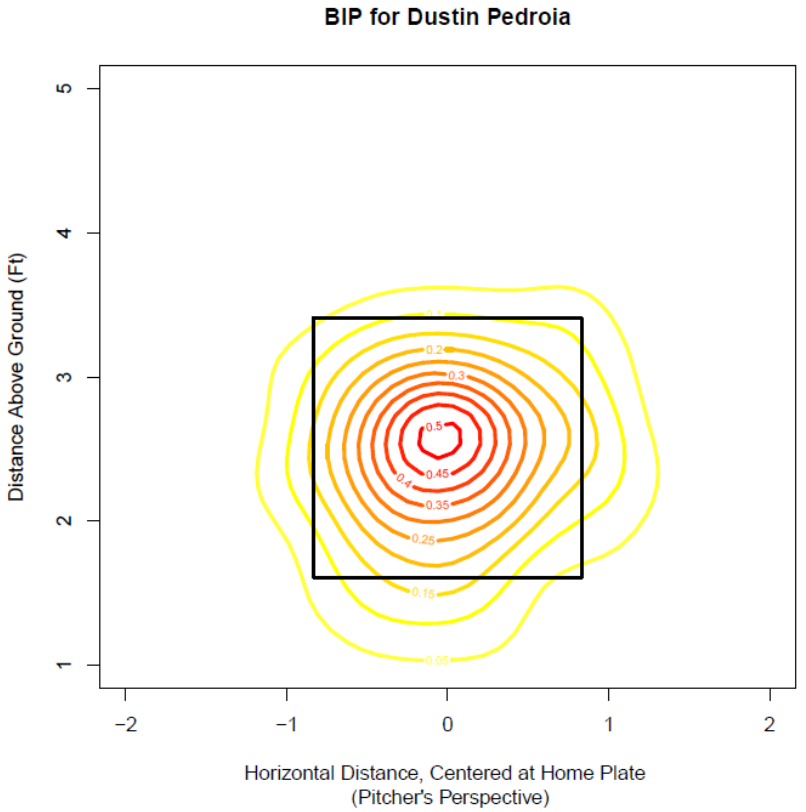

Figure 14. *Heat map contour plot for Pedroia's balls in play.*

As expected, the locations of the pitches resulting in balls put into play by Heyward generally correspond with the locations of the pitches at which he swings. Thus, the hotspot in the density and contour plots is, once again, in the lower outside corner of the strike zone. Notably, this trend is even more marked for balls put into play by Heyward than it is for his swings. On the density plot, essentially the entire area above the strike zone is blue, indicating very few pitches in this region were put into play (*Fig. 11*). The density plot of Heyward's swings has slightly more yellow area above the strike zone, meaning that while Heyward is unlikely to swing at high pitches, he is even less likely to put a ball into play that crosses home plate above the strike zone (*Fig. 7*). The contour plot of balls put into play by Heyward reveals that this event is about four times as likely for a pitch in the lower outside quadrant as for a pitch in the upper inside quadrant of the strike zone (*Fig. 12*). Compared with the contour plot for Heyward's swings (*Fig. 8*), the innermost contour for his balls put into play is much smaller. Furthermore, the contour lines are packed more closely together for balls put into play. The implication is that balls put into play by Heyward tend to be pitched relatively consistently in the lower outside area of the strike zone. The spatial trend mirrors that of the pitches at which he swings, but to a higher degree of consistency.

The prime location of pitches resulting in balls put into play by Pedroia is more or less exactly in the center of the strike zone. The density plot of Pedroia's balls put into play is nearly identical to the density plot of his swings (*Figs. 9 & 13*). There is a similar outcome for the two contour plots, except the contour lines for Pedroia's balls put into play are more closely spaced (*Figs. 10 & 14*). Therefore, from the pitcher's perspective, the most dangerous place to throw the ball to Pedroia is in the center of the strike zone. The density and contour plots are roughly symmetric with respect to a vertical axis through the center of the strike zone, so Pedroia cannot be predicted to put balls into play with greater frequency on one side of the strike zone or the other.

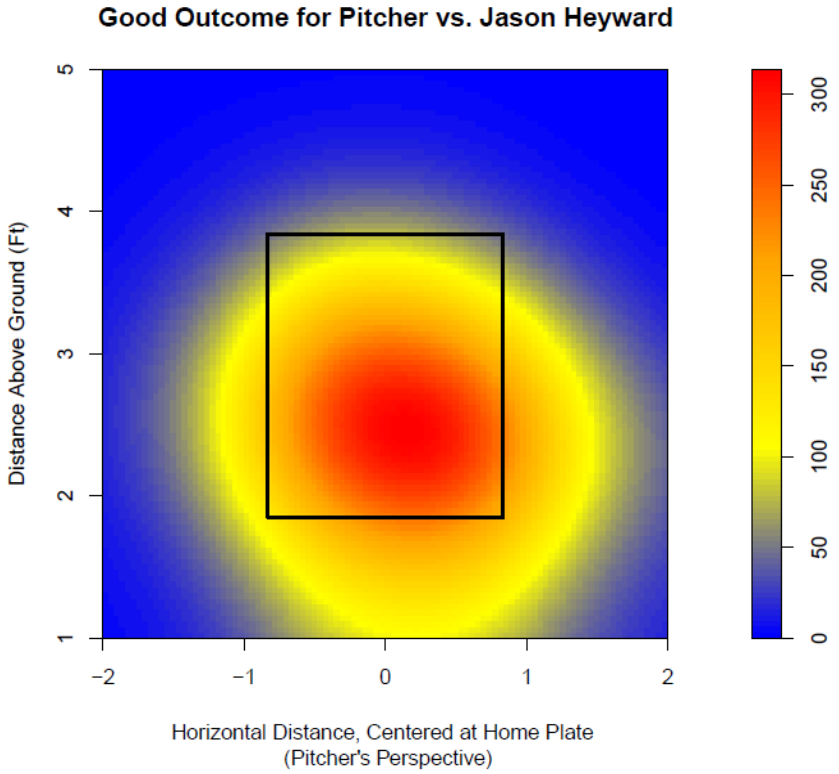

Figure 15. *Heat map density plot for good outcome vs. Heyward.*

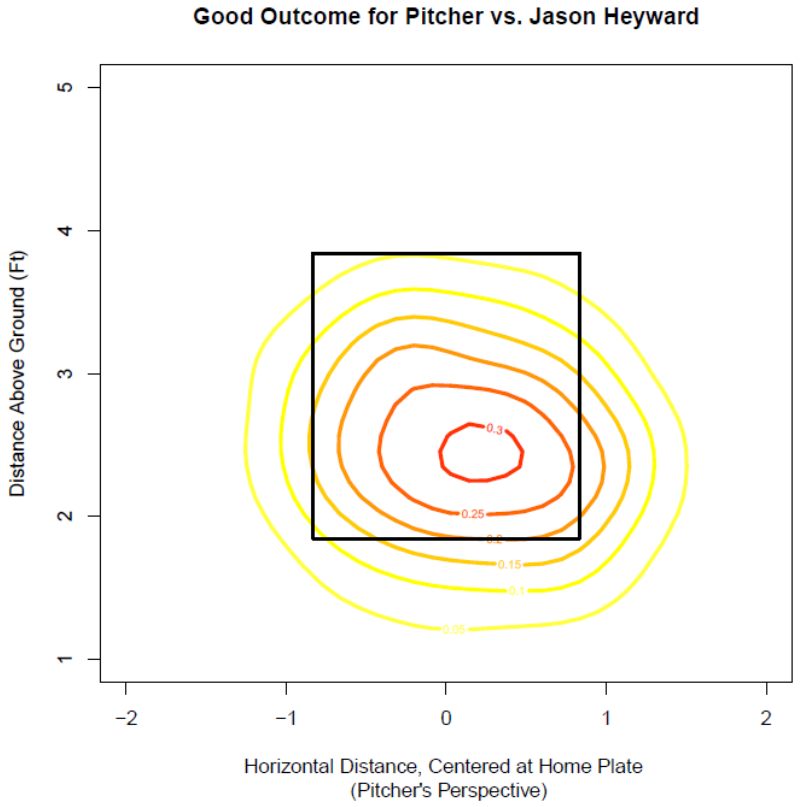

Figure 16. *Heat map contour plot for good outcome vs. Heyward.*

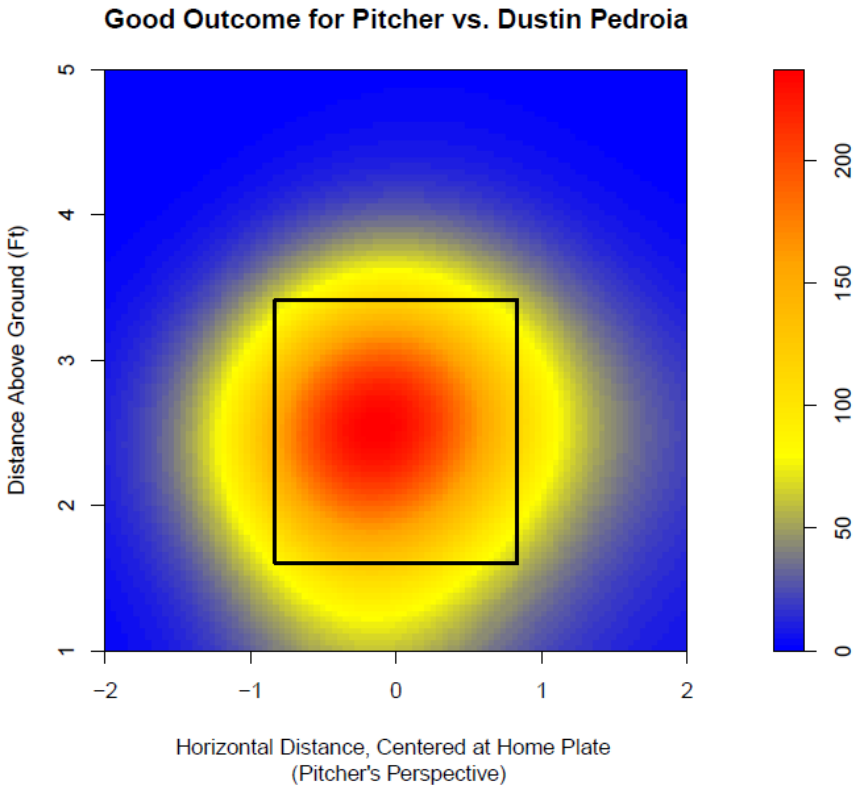

Figure 17. *Heat map density plot for good outcome vs. Pedroia.*

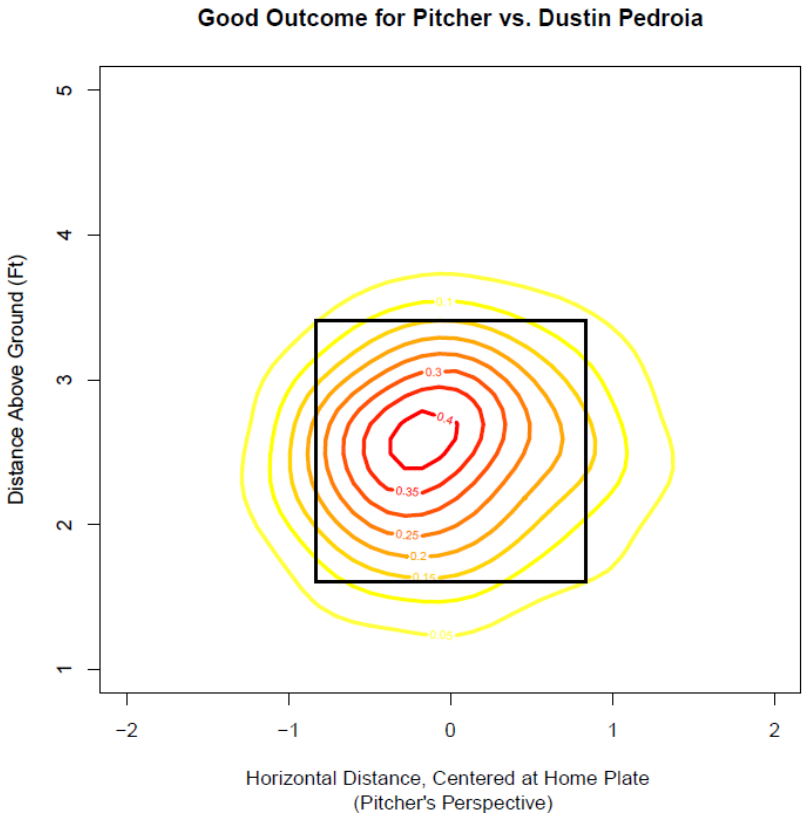

Figure 18. *Heat map contour plot for good outcome vs. Pedroia.*

For both players, the density and contour plots that show pitches resulting in favorable outcomes for the pitcher largely resemble the corresponding plots showing the batters' swings. This is a logical relationship to observe, considering the possible pathways to a favorable outcome for the pitcher. The majority of the time, a favorable outcome results from a swinging strike, a called strike, a pop-up fly ball that is caught by a field player, or a tagged out after a hit. An unfavorable outcome could result from a ball, a base hit, or a pitch that hits the batter. (There are a few other rare scenarios that could be classified as either favorable or unfavorable, but those occur so infrequently that they do not influence these plots.) A swing could result in either a good or bad outcome for the pitcher, depending on how the play unfolds. The same is true for a ball in play, but the pitcher is more likely to attain a favorable outcome from the collection of all swings versus the collection of all balls put into play. This pattern emerges because the batter gains an advantage by putting the ball into play.

Notably, the centers of mass in the plots of favorable outcomes for pitchers facing Pedroia lie left of the center of home plate. Compared to the centers of mass in the plots of his swings, these pitching hotspots are farther away from his body. Due to his relatively small stature and short arms, this outer area of the strike zone is beyond Pedroia's comfortable reach. Thus, pitches thrown to this area are more likely to favor the pitcher.

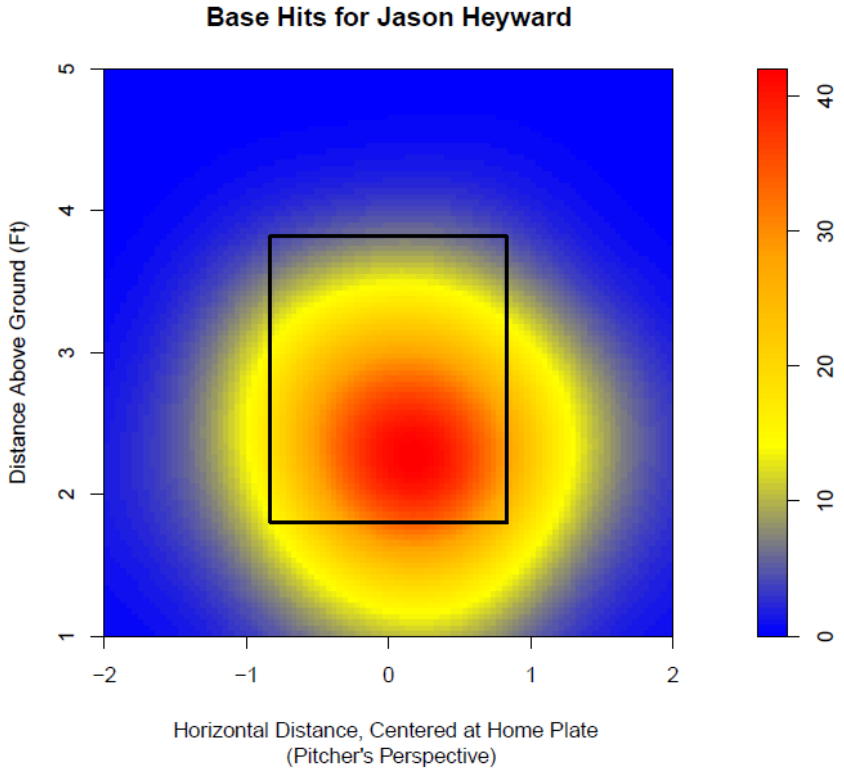

Figure 19. *Heat map density plot of Heyward's base hits.*

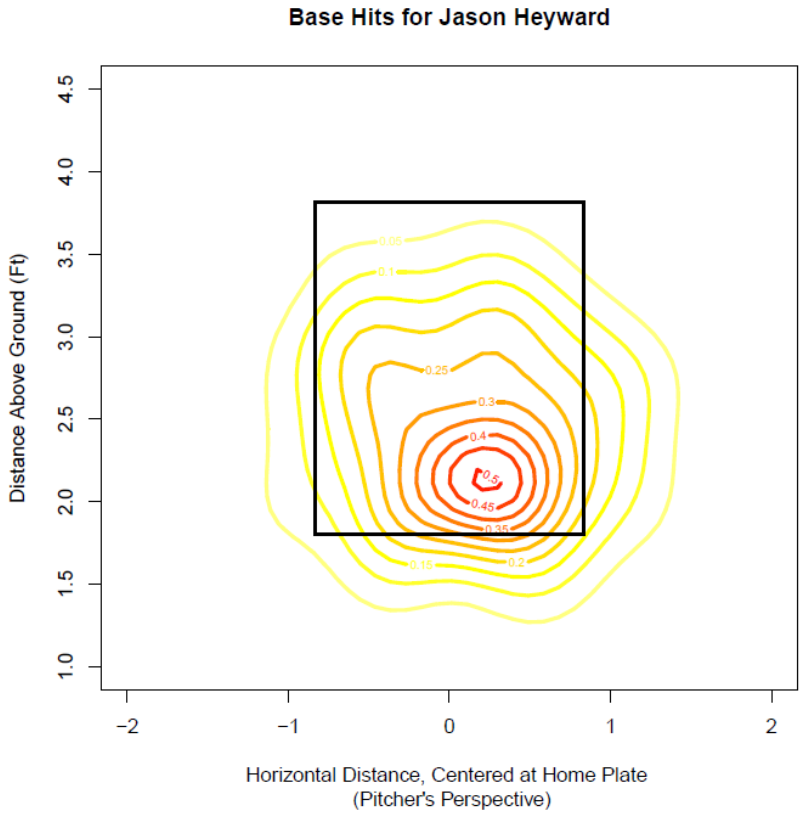

Figure 20. *Heat map contour plot of Heyward's base hits.*

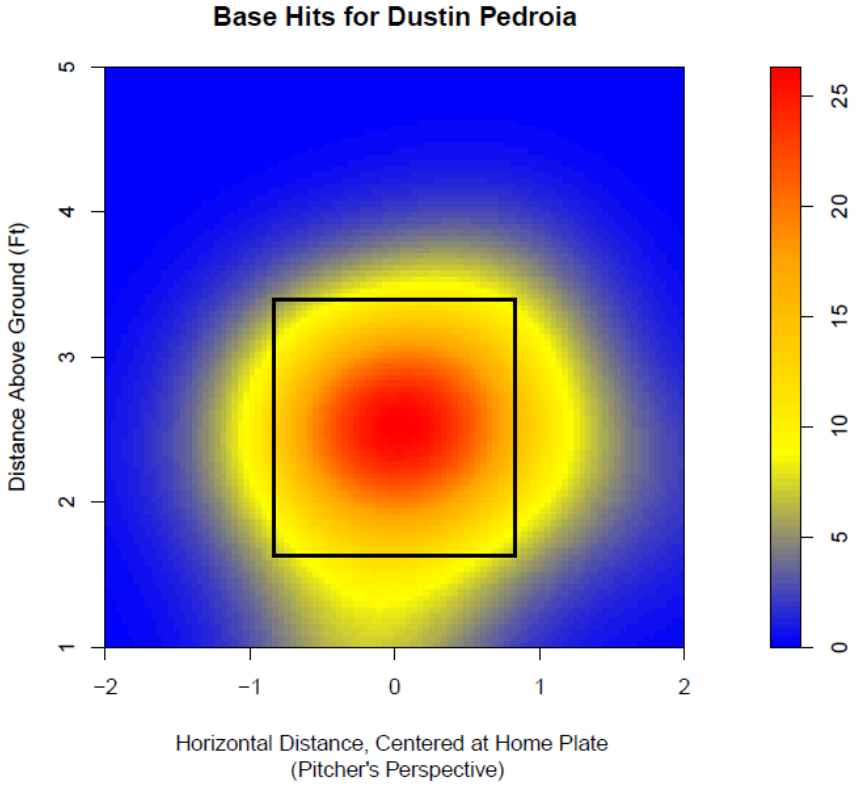

Figure 21. *Heat map density plot of Pedroia's base hits.*

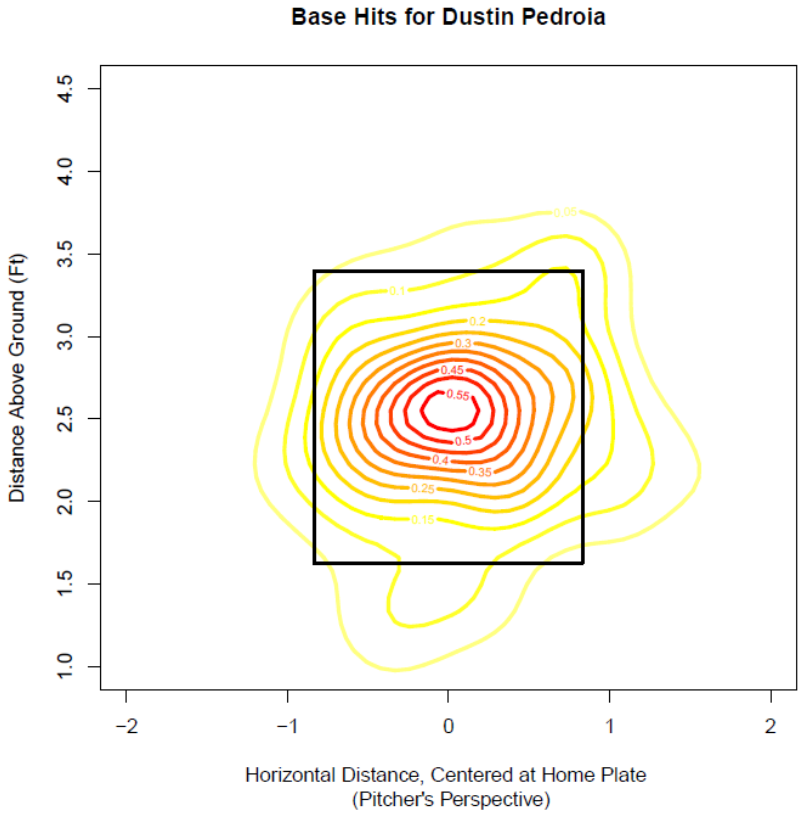

Figure 22. *Heat map contour plot of Pedroia's base hits.*

Base hits produced by Heyward are, once again, clustered in the lower outside corner of the strike zone. Since the density plot of Heyward's balls put into play showed extremely low intensity above the strike zone, by necessity there is even less intensity in the same area of the density plot showing his base hits (*Fig. 19*). The contour plot of Heyward's base hits shows a very tight inner contour in the lower outside quadrant of the strike zone (*Fig. 20*). He is approximately five times as likely to attain a base hit from a pitch in this hotspot as he is from a pitch in the upper third of the strike zone. There is a prominent trend in the spacing of the contour lines here that was present in the previously shown contour plots for Heyward, but not to such a great extent. The lines are much more tightly spaced in the lower outside corner of the strike zone than they are in the other areas of the plot. The fact that this trend was not so pervasive in Heyward's other contour plots means that Heyward's base hits are even more likely than his balls put into play to be located in the lower outside area of the strike zone. In turn, it was previously seen that Heyward's balls put into play were more likely than his swings to be located in this area. This trend signifies that Heyward's batting is optimized when he receives pitches that are located low and away from his body.

Pedroia's tends to attain base hits most frequently for pitches located directly in the center of the strike zone. The density plot shows a margin of moderate intensity bordering each side of the strike zone (*Fig. 21*). The lines on the contour plot appear to be conventional towards the center, but there are some odd distortions of the outermost contour lines (*Fig. 22*). These blips are showing areas where Pedroia earned base hits, although they were infrequent occurrences. Nevertheless, a pitcher should be aware that Pedroia is capable of attaining a base hit in an irregular fashion. He has the ability to swing at, and hit, pitches that are well below or inside of the strike zone. Thus, Pedroia can be labeled as a mostly predictable hitter who can sometimes have success unexpectedly.

In terms of game strategy for the pitcher, it is more straightforward to figure out how to approach Heyward as the opposing batter. It is clear that the pitcher should avoid the lower

outside corner of the strike zone and aim for a higher area, particularly on the inside. Pedroia, however, presents more of a challenge. Since his hotspot is in the center of the strike zone, the pitcher has a diminished area within the strike zone where Pedroia is less likely to get a base hit. The pitcher can aim for either the top or the bottom of the strike zone, but there is less margin for error. Furthermore, Pedroia's ability to achieve a base hit from pitches thrown to such a variety of unusual locations presents an additional complication.

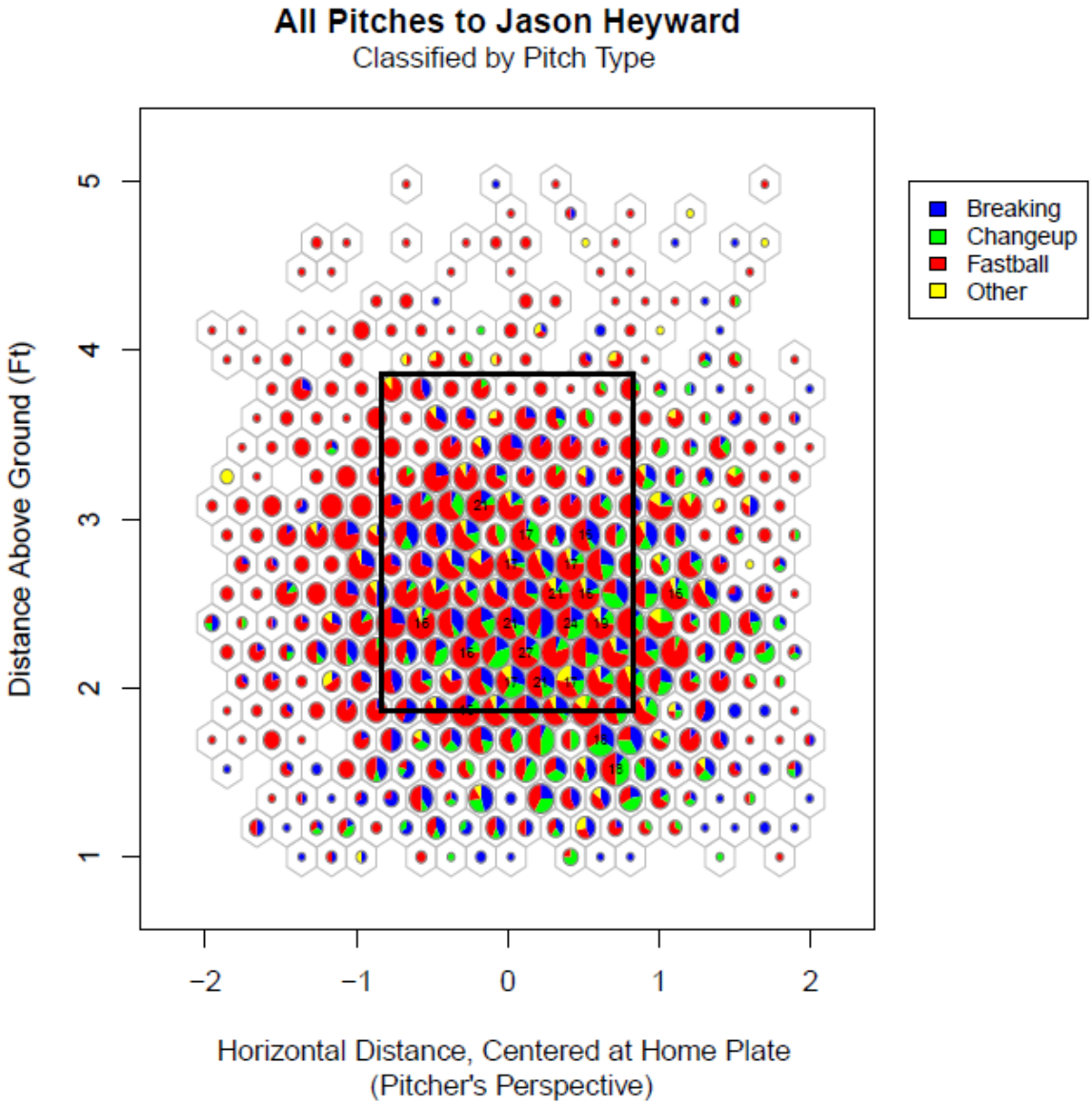

Figure 23. *Hexbin pie chart of all pitches to Heyward, classified by pitch type.*

Hexbin pie charts of all pitches thrown to each player, classified by pitch type, are provided first. These plots are more informative with regard to the opposing pitchers than the actions of the batter, but they will be valuable baselines for comparison with subsequent hexbin plots. The majority of the pitches thrown to Heyward are located in the lower portion of the strike zone, most likely because he is a tall player. Heyward receives changeups more frequently than most other players, especially in the lower outside quadrant of the strike zone (*Fig. 23*).

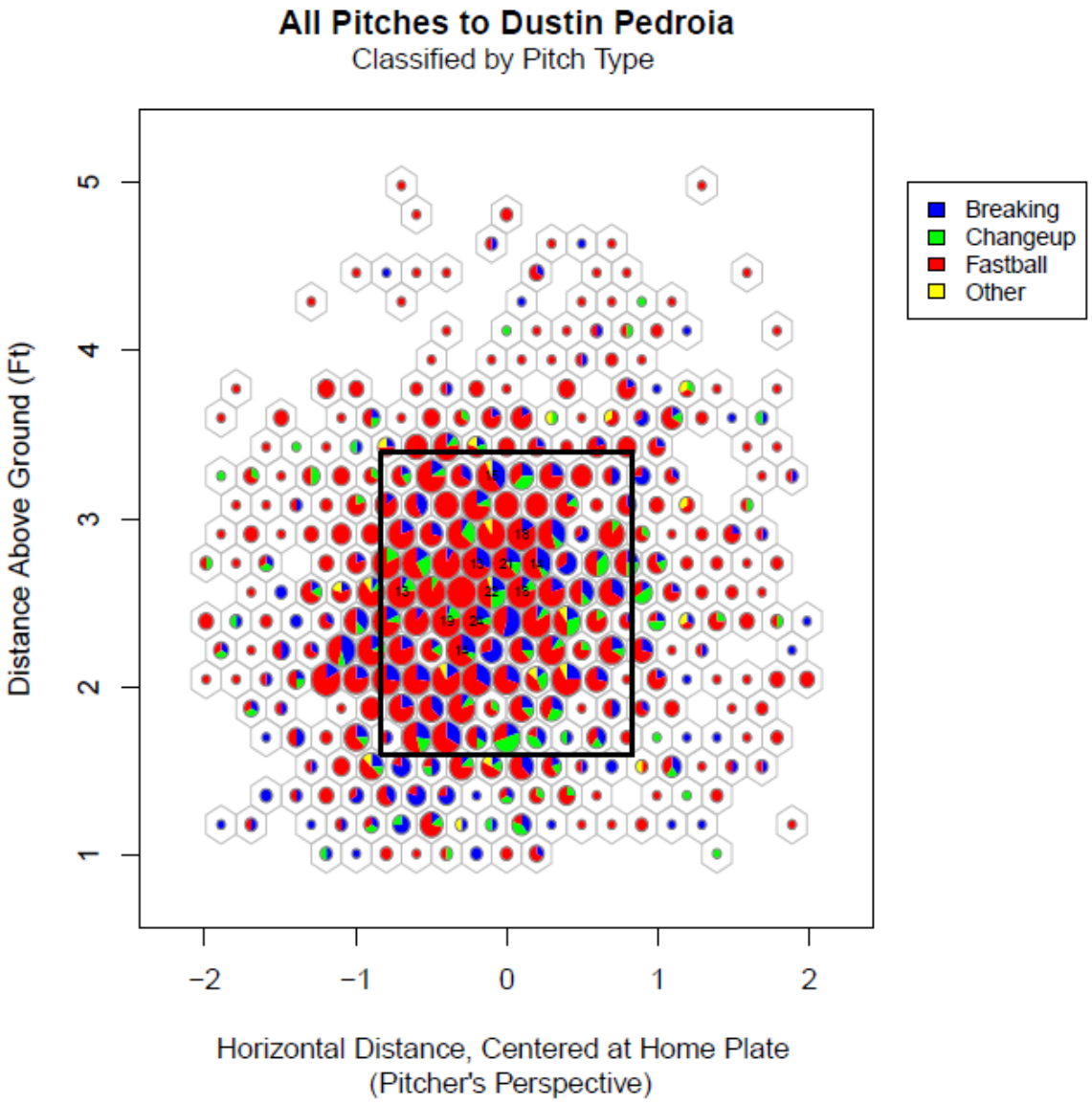

Figure 24. *Hexbin pie chart of all pitches to Pedroia, classified by pitch type.*

Pedroia receives pitches that are predominantly classified as fastballs. They are placed most prevalently on the far side of the strike zone, distributed fairly uniformly from top to bottom (*Fig. 24*). The far side of the strike zone is an optimal pitch location against Pedroia since he is one of the smallest players in MLB – his short arms make it more difficult for him to accurately swing at pitches that are farther from his body.

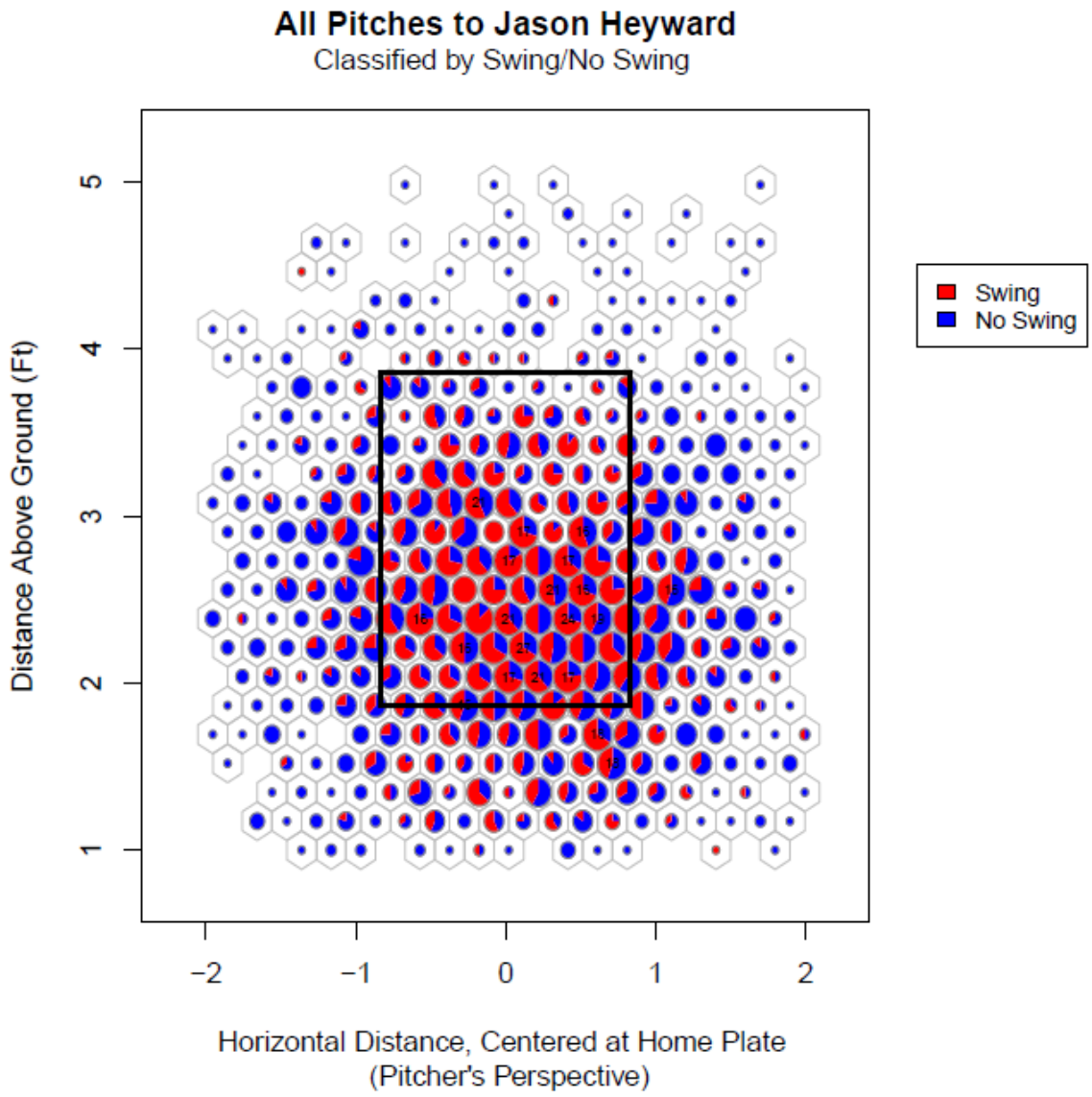

Figure 25. *Hexbin pie chart of all pitches to Heyward, classified by swinging or not.*

For hexbin pie charts of all pitches classified by whether or not the batter swings, the pitcher would be most interested in two characteristics: (1) where the batter swings outside of the strike zone, and (2) where the batter does *not* swing inside the strike zone. Heyward tends to swing at pitches located outside of the strike zone if they are placed on the lower or far side peripheries. He almost never swings at pitches located above the strike zone. The best pitch placement for a called strike against Heyward is the upper inside quadrant of the strike zone (*Fig. 25*).

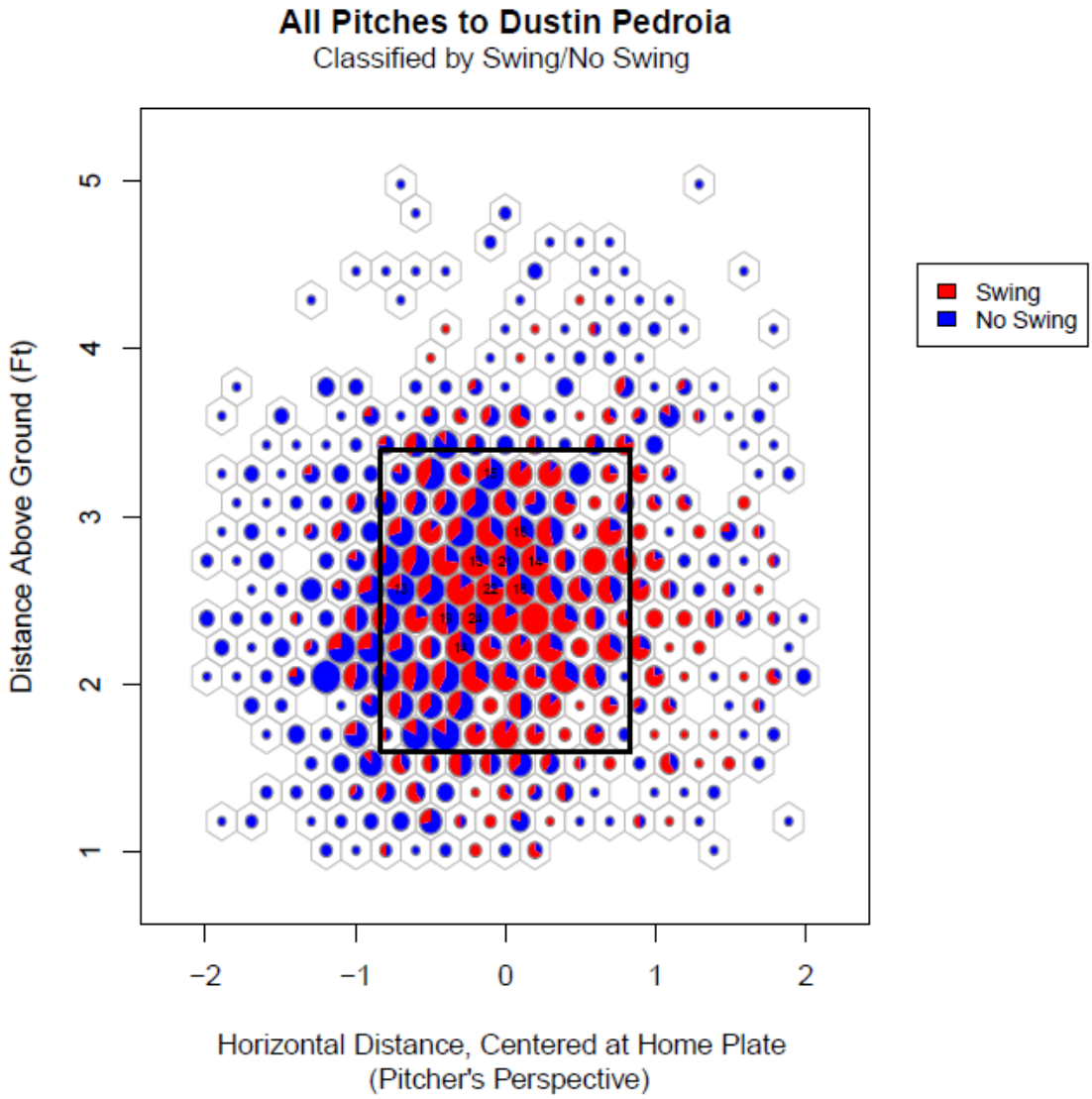

Figure 26. *Hexbin pie chart of all pitches to Pedroia, classified by swinging or not.*

Pedroia has a more readily apparent swinging trend than Heyward. He swings at nearly everything in the center of the strike zone and rightward. He has a fair number of swings scattered outside of the strike zone on the top, bottom, and far side. Most conspicuously though, is the high frequency of swings out of the strike zone on the side closest to his body. Despite the fact that these pitches lie outside of the strike zone, Pedroia consistently swings at them anyway (*Fig. 26*). This pattern is not typical of most players. As previously mentioned, Pedroia's inside swinging habit can most likely be attributed to his small body size and short arms; due to his physical stature, it is most comfortable for him to swing at pitches in this area.

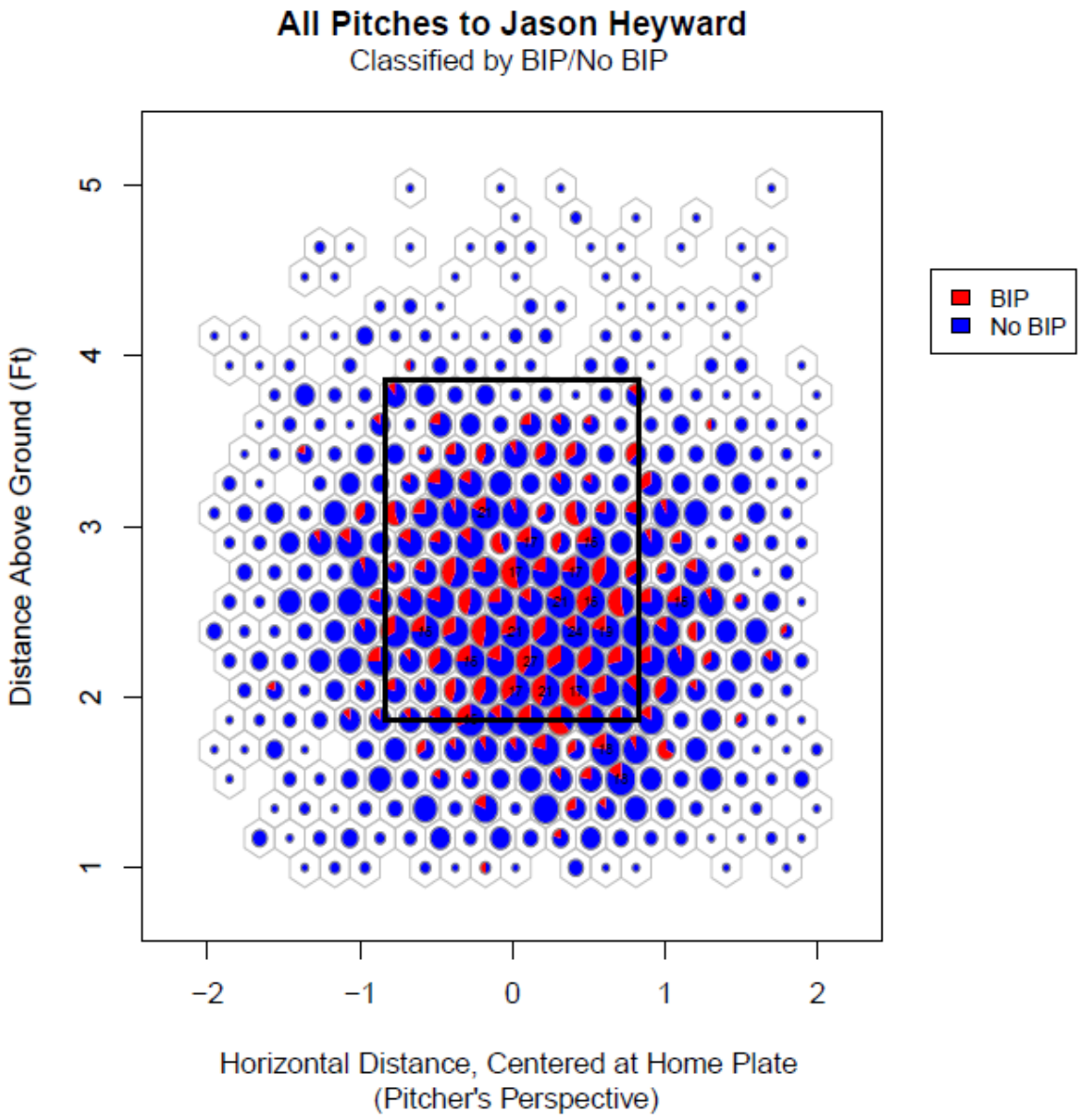

Figure 27*. Hexbin pie chart of all pitches to Heyward, classified by BIP or not.*

Confirming the trends that were previously noted in the heat map density and contour plots, the hexbin pie chart of all pitches thrown to Heyward, classified by whether or not the ball was put into play, shows that Heyward favors the lower outside quadrant of the strike zone. He rarely puts balls into play that are pitched out of the strike zone on the side closer to his body or above the strike zone. However, he relatively often puts balls into play when they are pitched below the strike zone or out of the strike zone on the side opposite his body (*Fig. 27*).

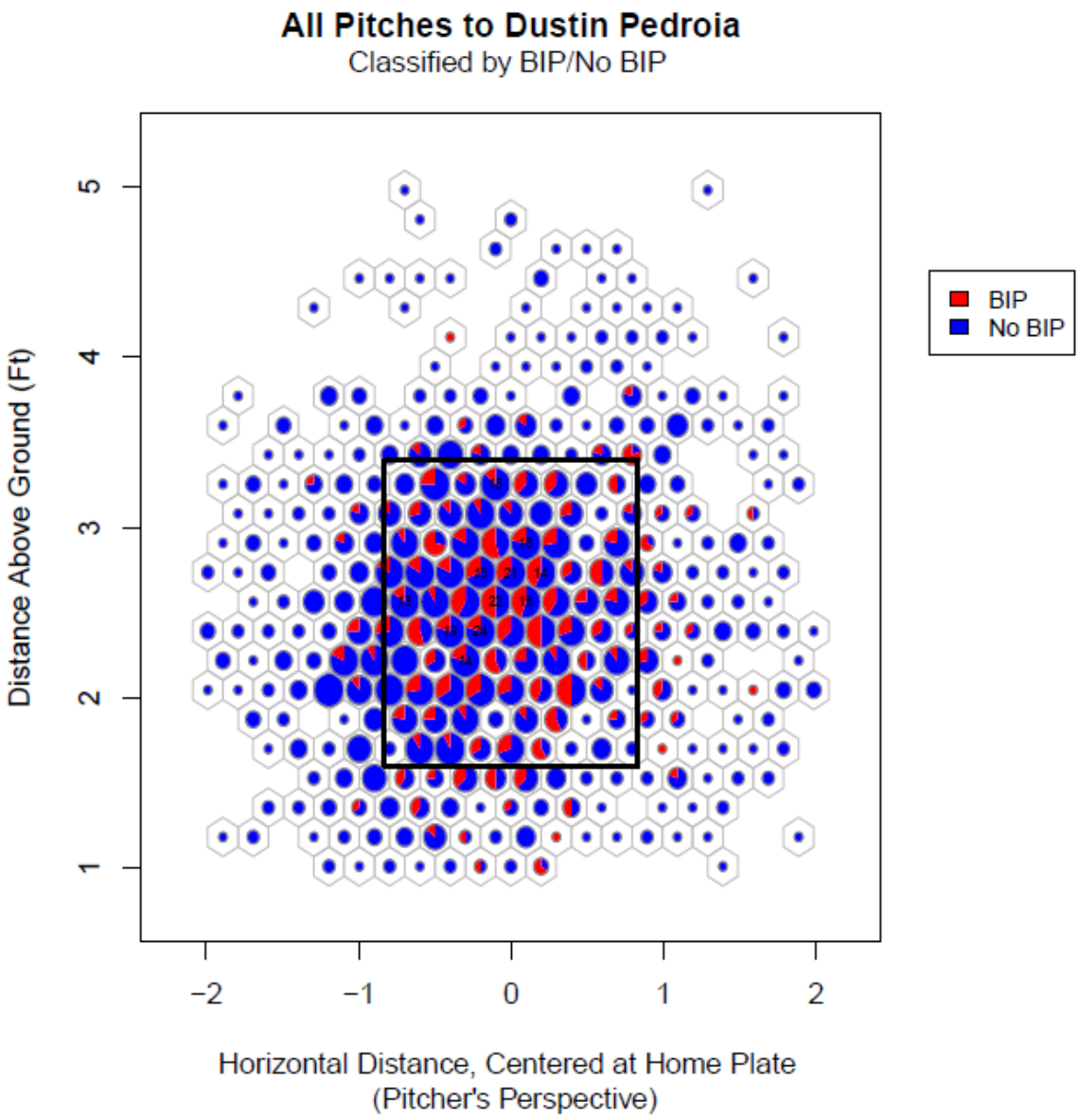

Figure 28. *Hexbin pie chart of all pitches to Pedroia, classified by BIP or not.*

Perhaps the most prominent feature of Pedroia's hexbin pie chart of all pitches classified by whether or not the ball was put into play is the relative frequency of balls *not* put into play that were pitched out of the strike zone on the side away from his body. Otherwise, the balls put into play appear to be distributed fairly uniformly throughout the strike zone with a hotspot in the center. Pedroia can put balls into play that are pitched out of the strike zone on the three remaining sides (*Fig. 28*).

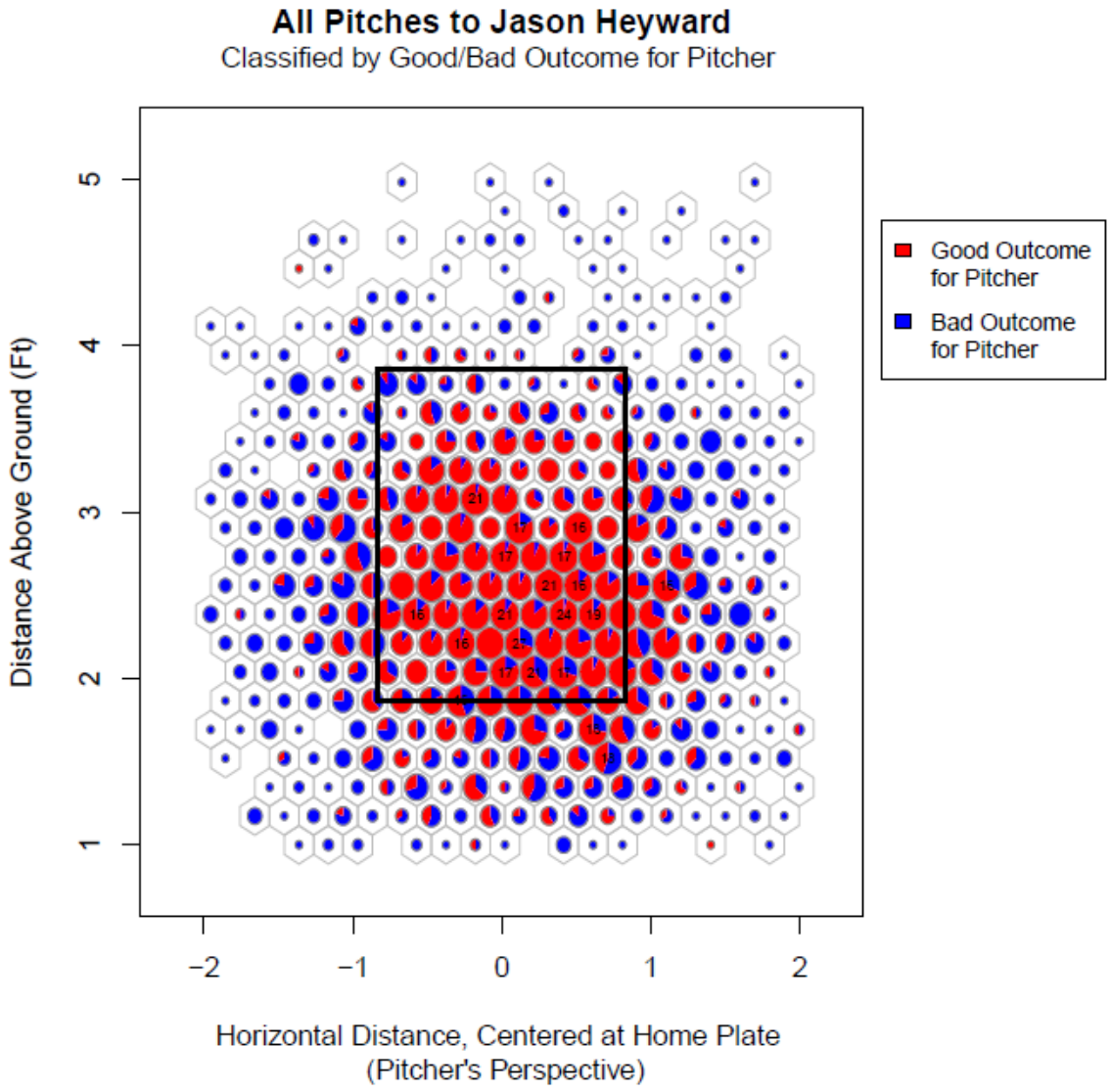

Figure 29. *Hexbin pie chart of all pitches to Heyward, classified by outcome for pitcher.*

There are two primary areas within Heyward's strike zone where unfavorable outcomes for the pitcher tend occur more frequently (*Fig. 29*). One of these areas is the lower outside corner, which was previously shown to be Heyward's hotspot for base hits. The other area is the upper inside corner. According to the preceding density and contour plots, the upper inside corner of the strike zone is *not* a batting hotspot for Heyward. The fact that so many outcomes of pitches in this area benefit Heyward can probably be attributed to bad calls by the umpires. In these scenarios, the pitches would have been ruled as balls even though they were located within the

strike zone and should have been called strikes. The opposite trend can be observed below the strike zone: there is a high density of outcomes benefitting the pitcher. A possible explanation is that since Heyward is such a tall player, the umpires estimate his strike zone to be lower than it actually is. Thus, pitches high in the strike zone are often ruled balls, and pitches below the strike zone are often called strikes. A final observation on this hexbin pie chart is that there is a high density of good outcomes for the pitcher outside of the strike zone on the side opposite Heyward's body. These areas predominantly represent pitches at which Heyward swings and misses. In summary, from a strategic standpoint, a pitcher should avoid the upper inside and lower outside corners of Heyward's strike zone, and aim for either the center of the strike zone or just wide on the far side.

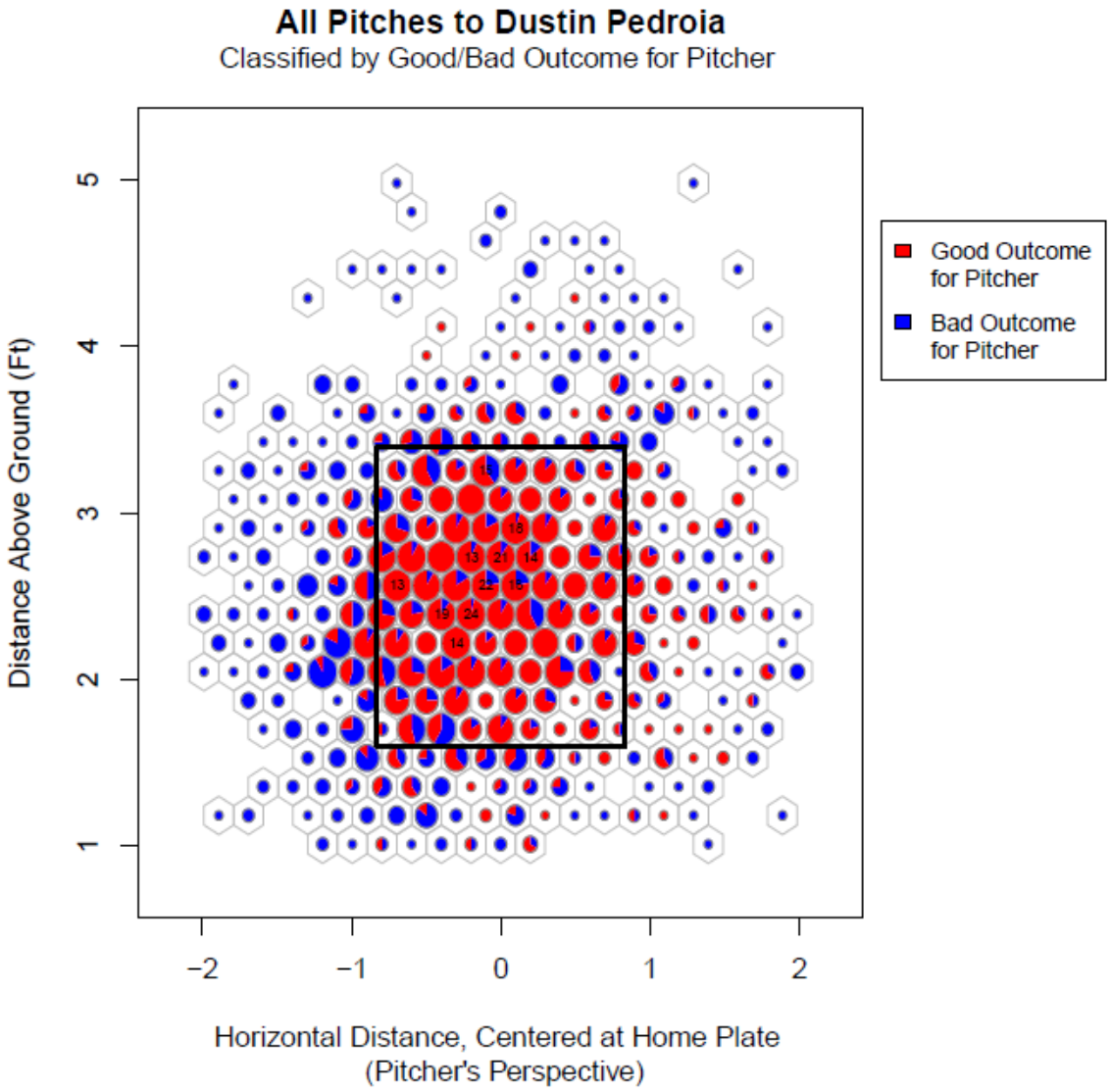

Figure 30. *Hexbin pie chart of all pitches to Pedroia, classified by outcome for pitcher.*

Of all pitches thrown to Pedroia, unfavorable outcomes for the pitcher tend to appear in the center of the strike zone, his hotspot for base hits. There is also a high density of adverse outcomes for the pitcher for pitches located outside of the strike zone on the side away from Pedroia (*Fig. 30*). This pattern likely occurs because Pedroia rarely swings at pitches that are far from his body; he has short arms, so it is less comfortable for him to swing at outside pitches. Therefore, pitchers can expect to acquire a lot of balls if they aim for the outer edge of Pedroia's strike zone and end up missing wide.

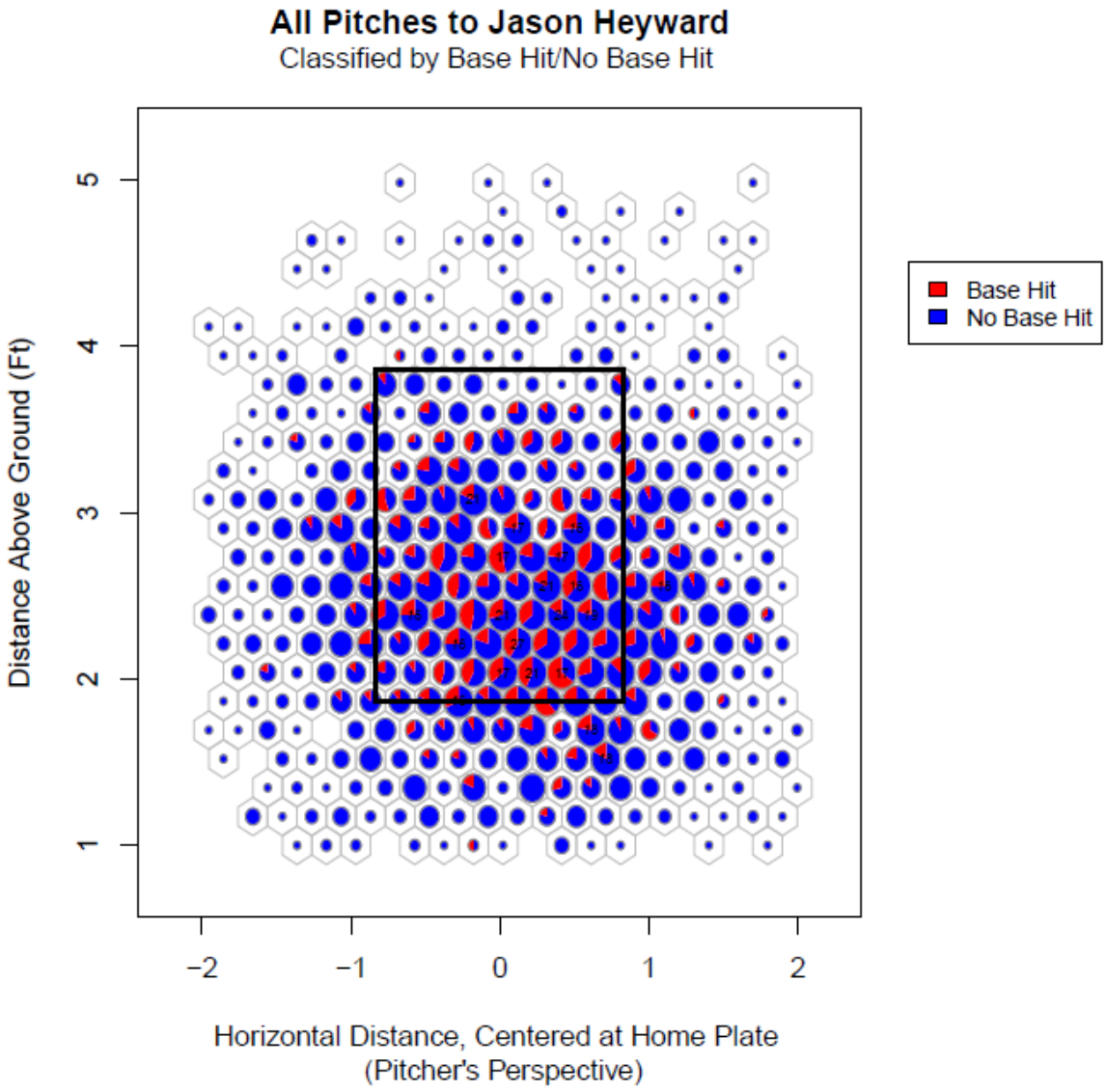

Figure 31. *Hexbin pie chart of all pitches to Heyward, classified by base hit or not.*

Confirming the trend that was readily apparent in the heat map density and contour plots, Heyward's base hits are mainly clustered around the lower outside corner of the strike zone. Heyward rarely achieves base hits when the ball is not pitched inside the strike zone, but when he does, they occur most frequently on pitches below or on the far periphery of the strike zone (*Fig. 31*).

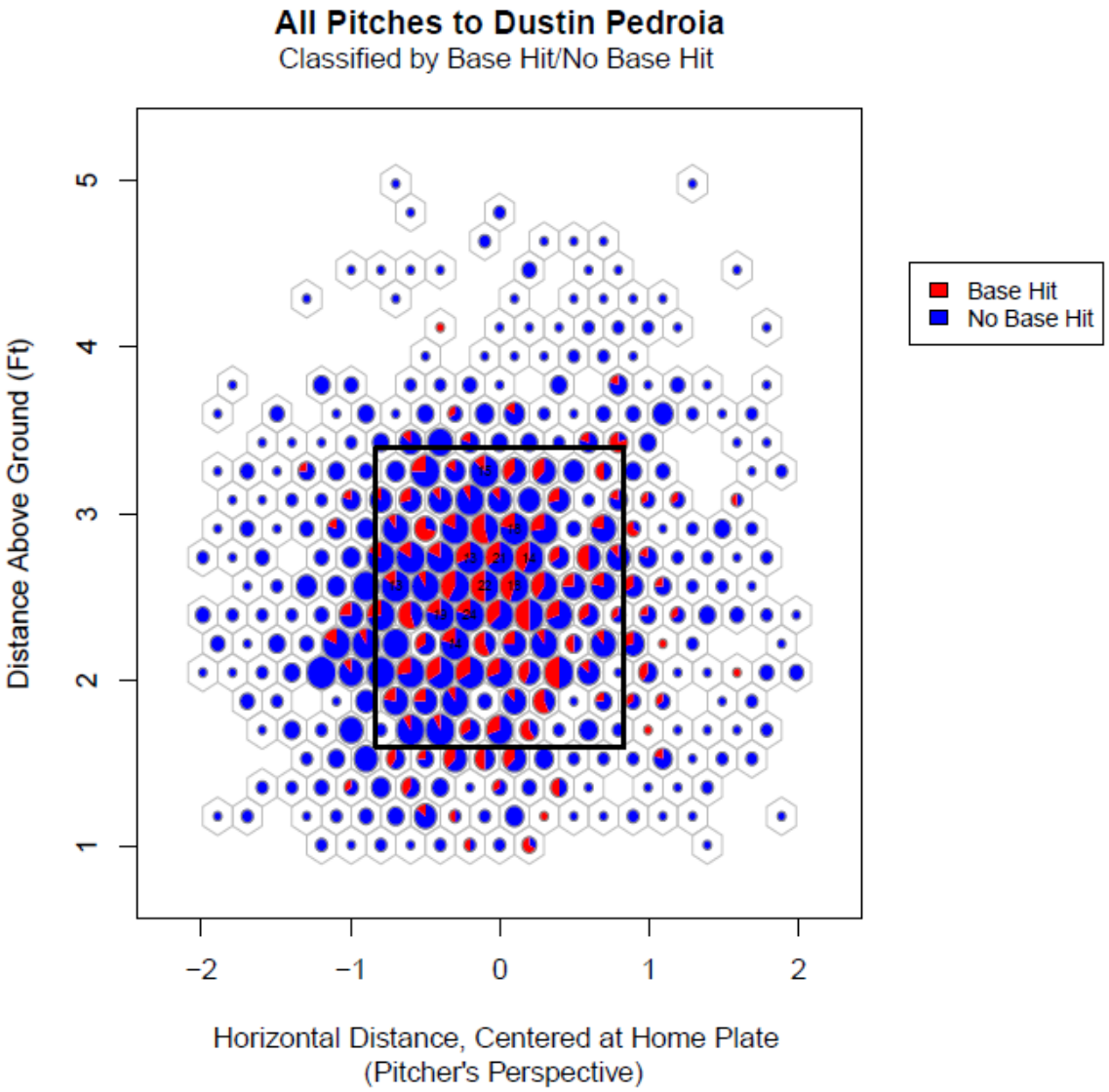

Figure 32. *Hexbin pie chart of all pitches to Pedroia, classified by base hit or not.*

In concordance with the results of the heat map density and contour plots, Pedroia achieves base hits primarily on pitches thrown to the center of his strike zone (*Fig. 32*). Unlike Heyward, Pedroia has relatively equal occurrences of base hits off of pitches located outside of the strike zone on all four sides (although slightly less so on the far side).

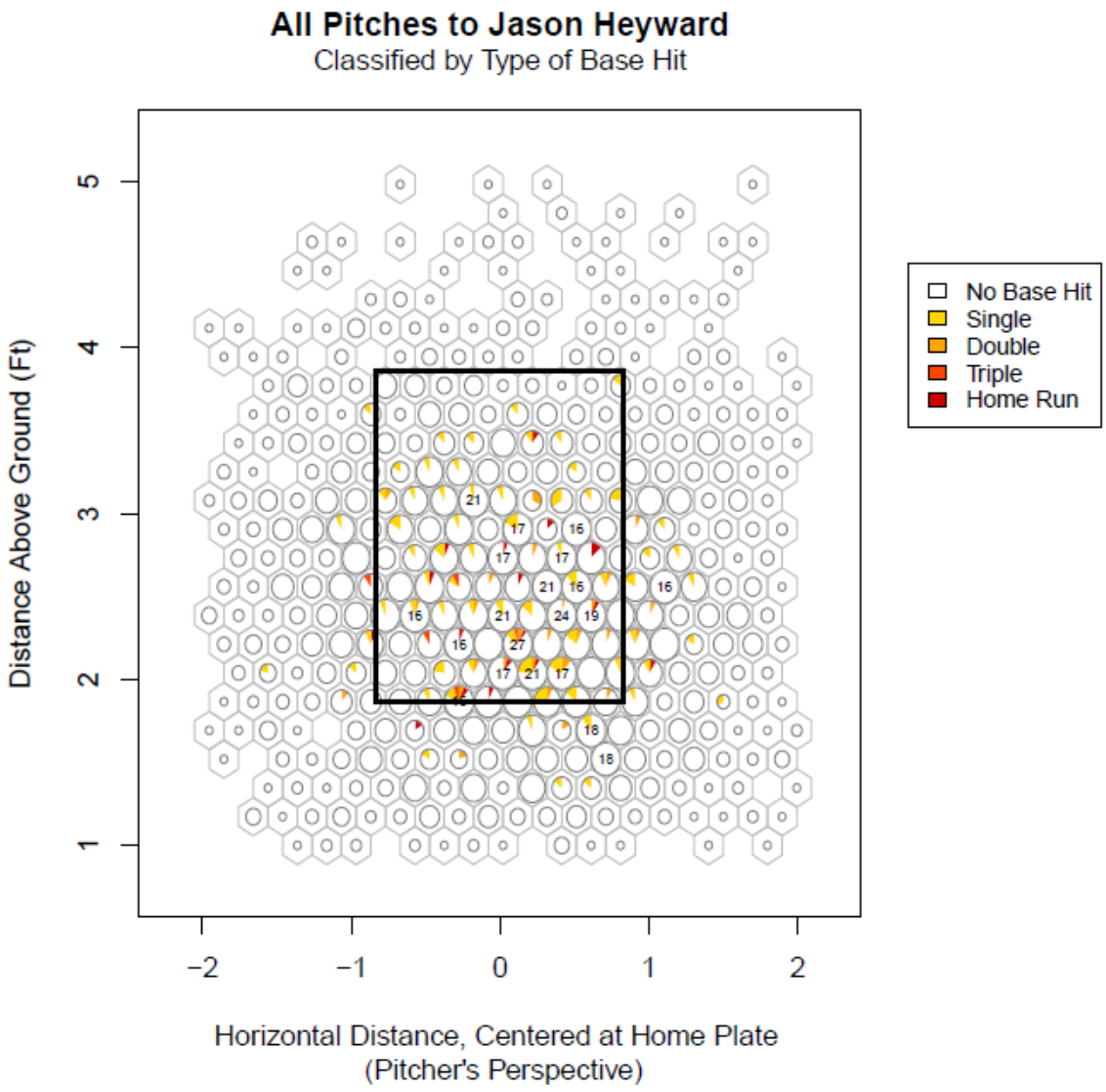

Figure 33. *Hexbin pie chart of all pitches to Heyward, classified by type of base hit.*

In addition to whether or not a batter gets a base hit, it is also important to consider the power of the hits – how many bases are attained. Heyward's power hits, represented by the darker shades in the hexbin pie chart, arise from pitches located in the lower part of the strike zone. Notably, although Heyward does occasionally get base hits off of pitches that do not fall within the strike zone, these are almost all singles; Heyward's most powerful hitting area is clearly contained within the strike zone (*Fig. 33*).

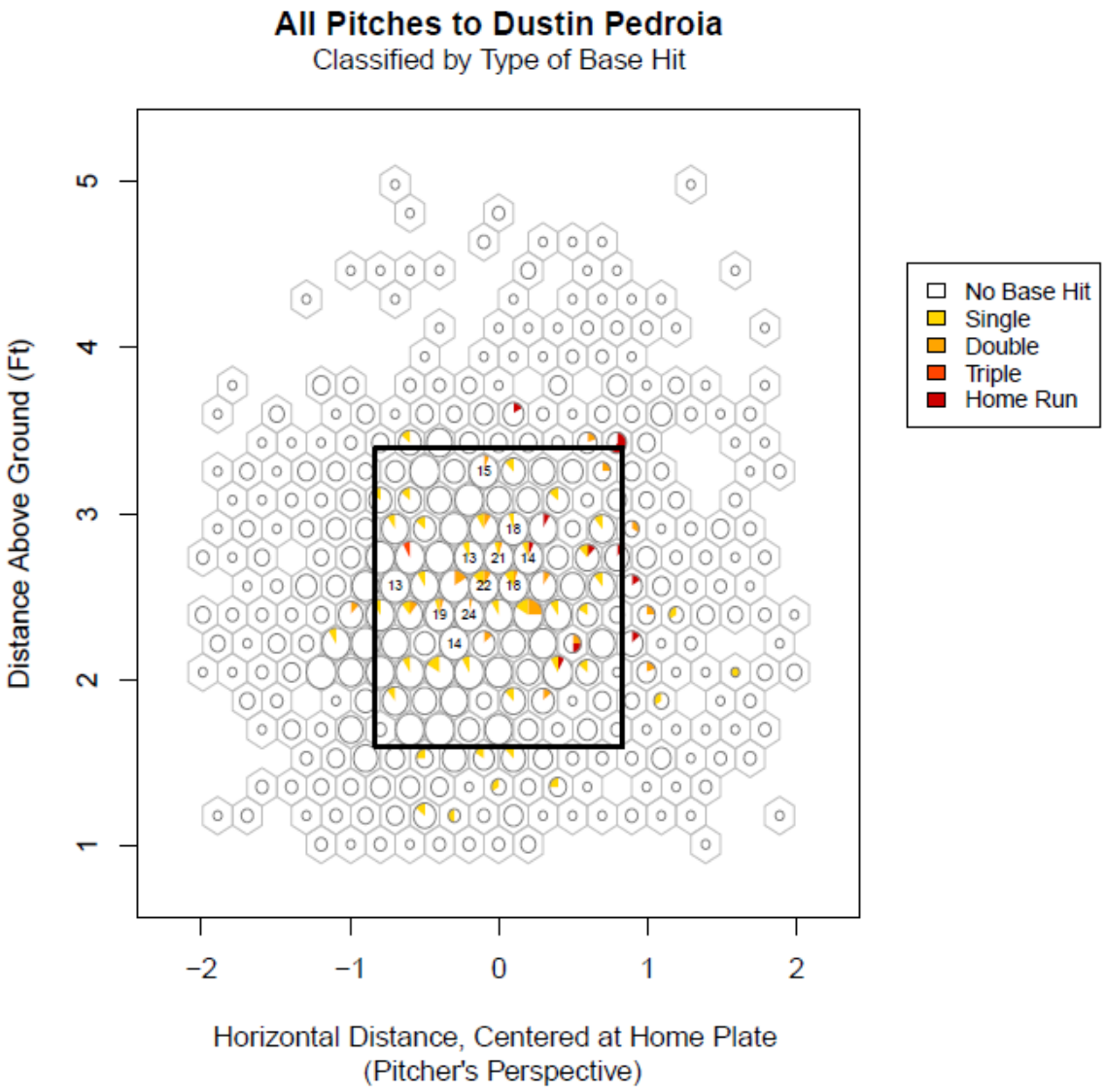

Figure 34. *Hexbin pie chart of all pitches to Pedroia, classified by type of base hit.*

Pedroia's power hits display an interesting trend. He gets base hits most frequently in his hotspot, the center of the strike zone. However, these base hits tend to be singles or doubles. Many of Pedroia's more powerful hits are actually gained from pitches that are placed slightly beyond the strike zone boundary, on the side closest to his body (*Fig. 34*). This pattern would be expected of Pedroia, and any other batter of a similar stature, since smaller players are able to swing more naturally at pitches closer to their bodies.

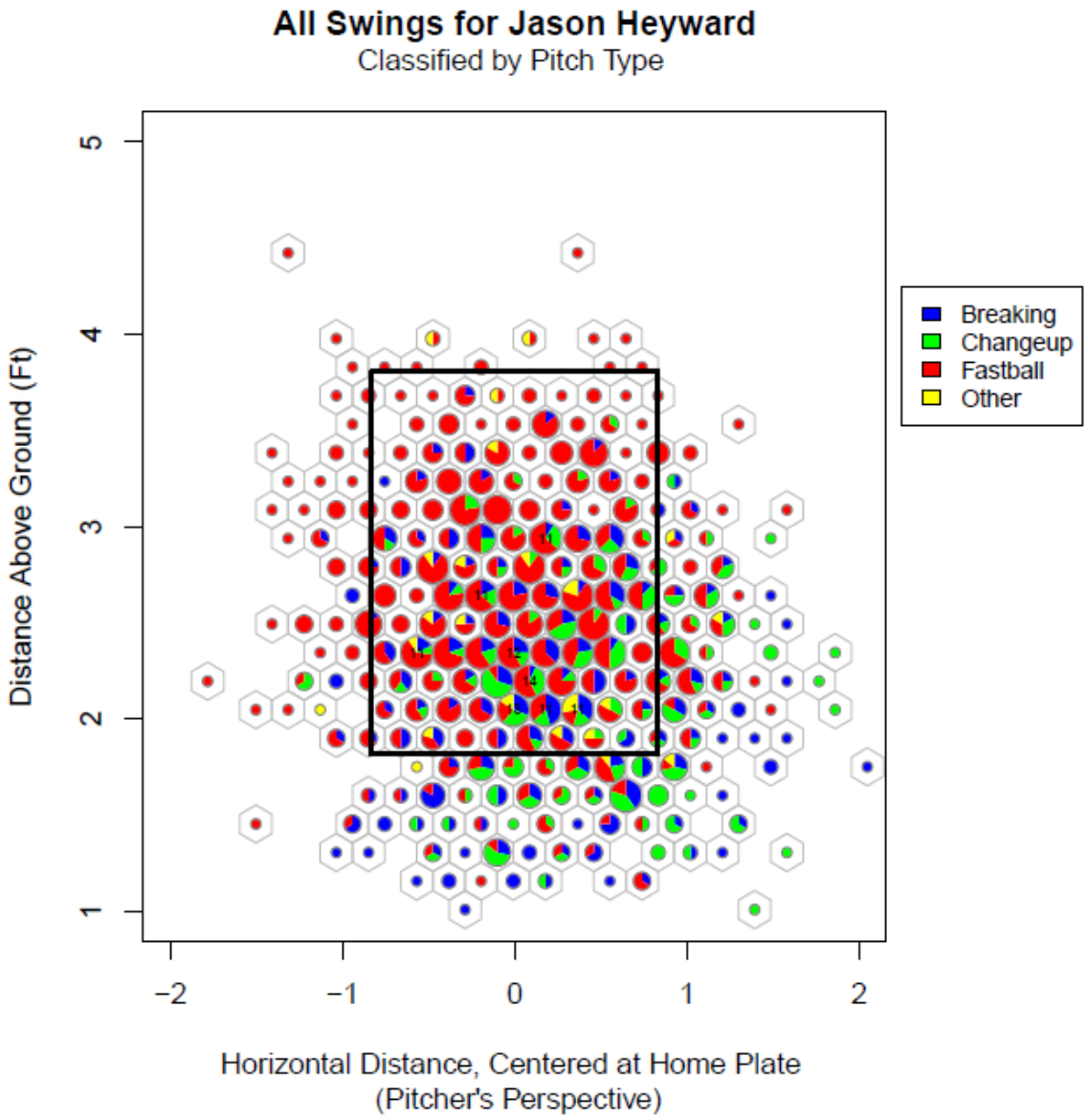

Figure 35. *Hexbin pie chart of all swings for Heyward, classified by pitch type.*

A notable trend emerges from the comparison of the hexbin pie charts, categorized by pitch type, of all pitches thrown to Heyward and all swings for Heyward (*Figs. 23 & 35*). Relative to the overall number of changeups he receives, Heyward swings at a very large proportion of these pitches. This tendency is especially true for pitches thrown below the strike zone – the overall distribution of pitches to Heyward below the strike zone is approximately evenly distributed among fastballs, changeups, and breaking balls; however, the vast majority of his swings below the strike zone are directed at changeups.

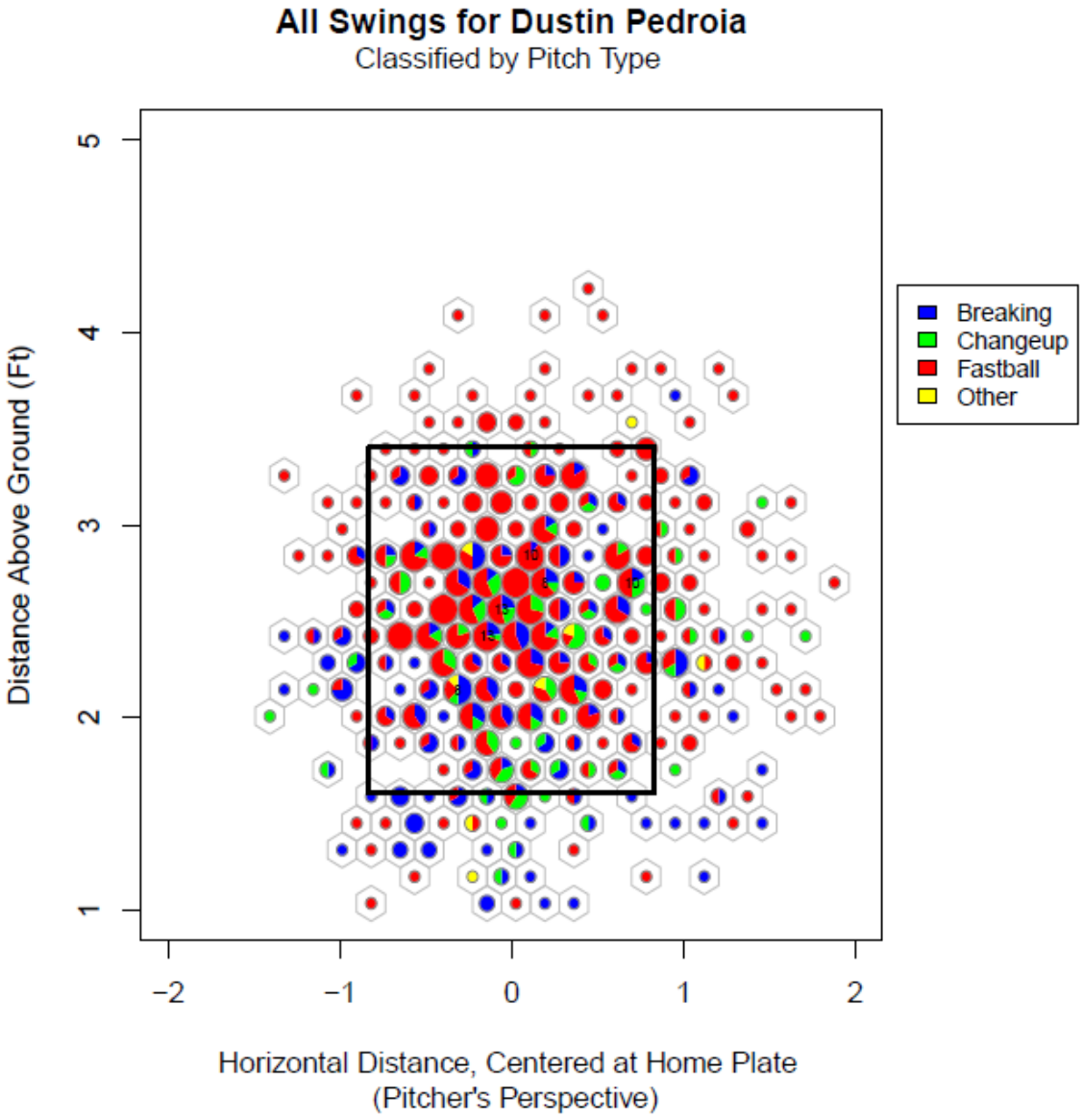

Figure 36. *Hexbin pie chart of all swings for Pedroia, classified by pitch type.*

A trend that stands out in Pedroia's swinging habits is that he almost never swings at any pitch other than a fastball above the strike zone. Comparing the hexbin pie charts, classified by pitch type, of all pitches to Pedroia and all swings for Pedroia, it is apparent that he does indeed receive non-fastball pitches above the strike zone but does not swing at them (*Figs. 24 & 36*). Also notably, Pedroia swings disproportionately often at breaking balls below the strike zone, especially on the outer edge (*Fig. 36*).

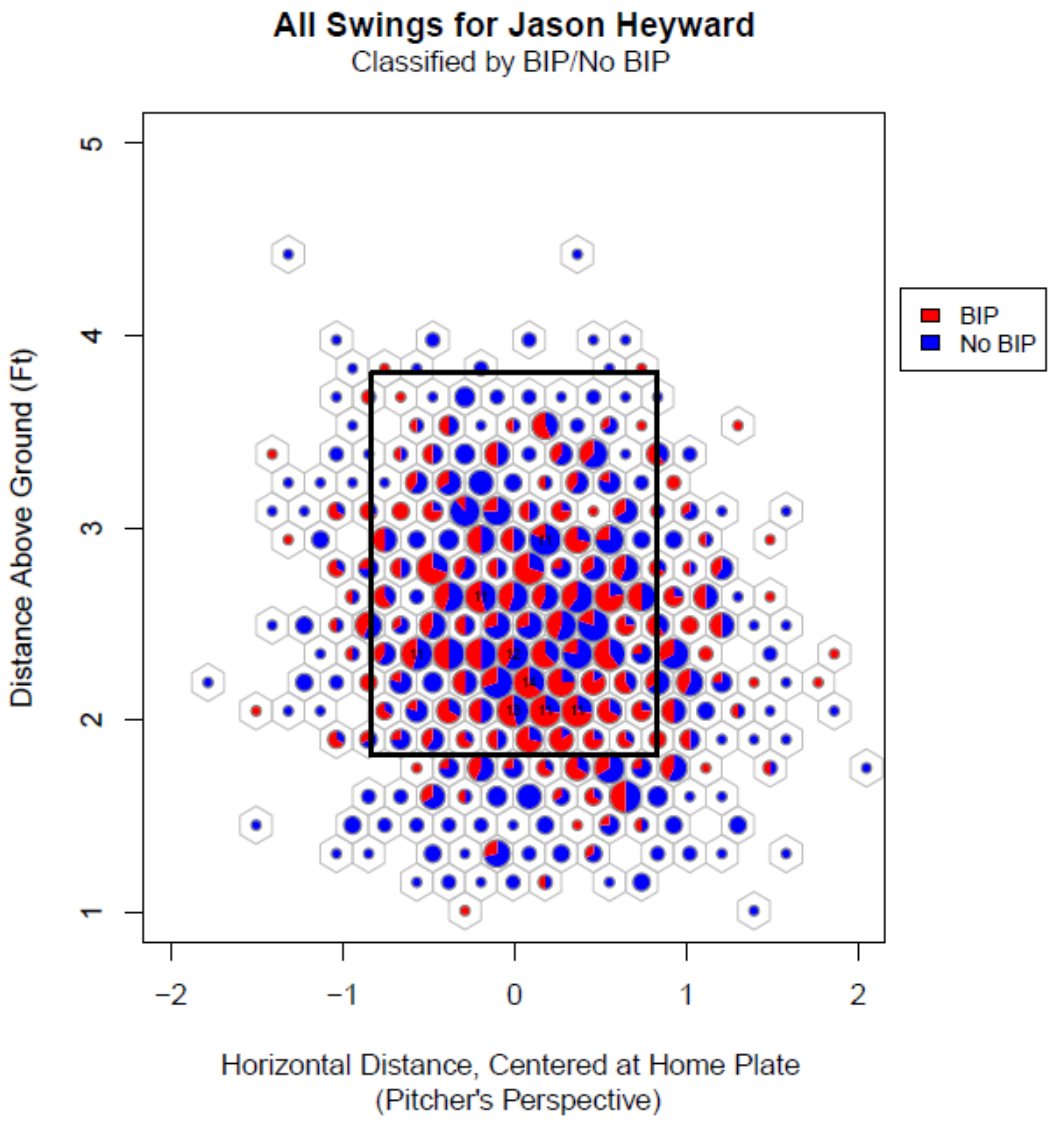

Figure 37. *Hexbin pie chart of all swings for Heyward, classified by BIP or not.*

When Heyward swings, he clearly has the most success at putting the ball into play for pitches located in the lower outside quadrant of the strike zone. He rarely ever swings at pitches above the strike zone; when he does, he usually does not make beneficial contact with the ball. Within the strike zone, Heyward's weak spot appears to lie in the central area; there are many hexagons in which he swings and misses more often than he swings and puts the ball into play. Another weak spot for Heyward lies below the strike zone on the side closest to his body. He swings relatively often at pitches in this region, but rarely puts these pitches into play (*Fig. 37*).

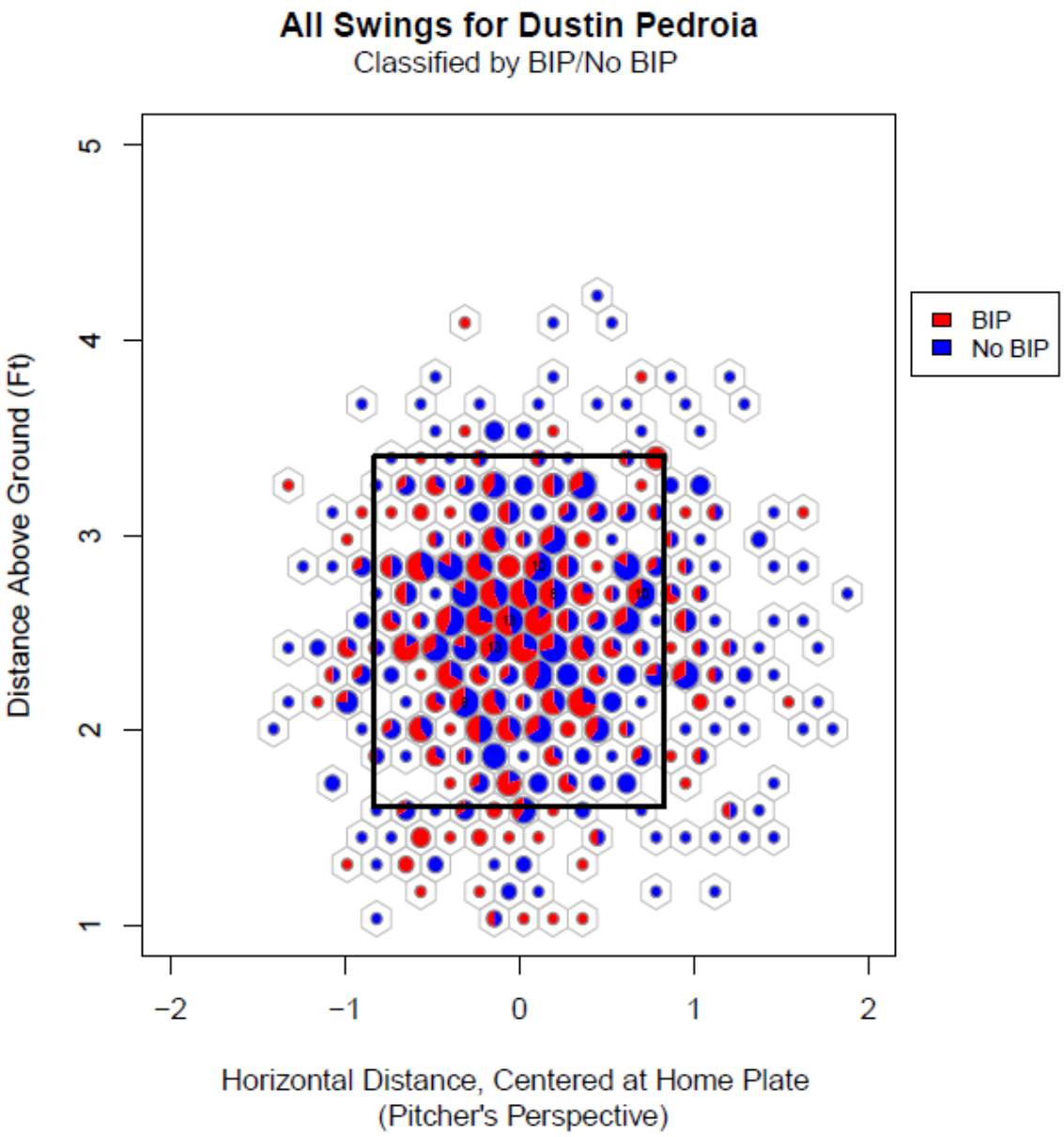

Figure 38. *Hexbin pie chart of all swings for Pedroia, classified by BIP or not.*

Pedroia's swinging successes are less predictable than Heyward's. His hotspot with regard to putting the ball into play is clearly located in the center of the strike zone. However, unlike Heyward, there does not appear to be much of a trend to the locations of pitches he puts into play in the corners of the strike zone or the area surrounding the strike zone (*Fig. 38*).

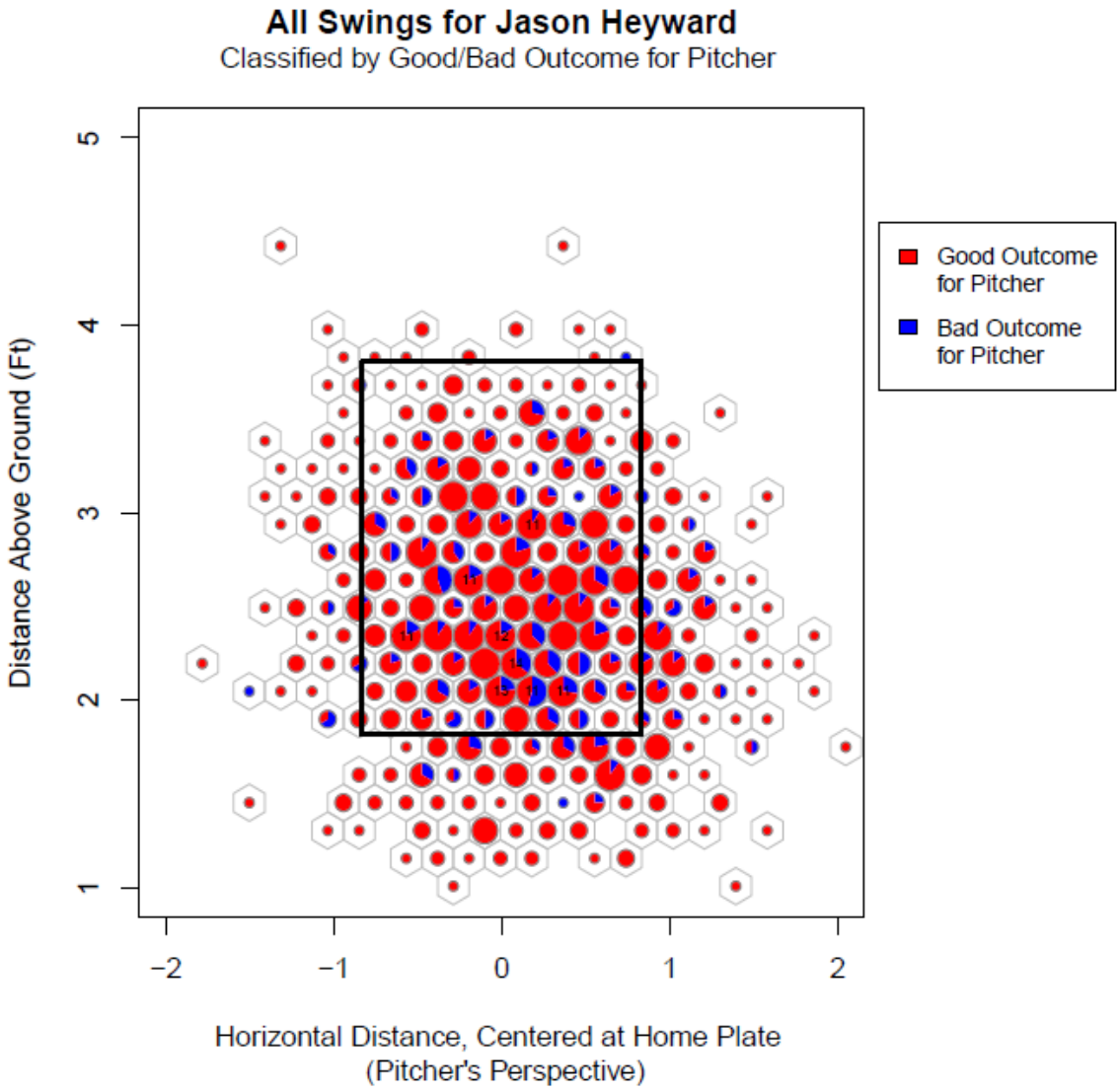

Figure 39. *Hexbin pie chart of all swings for Heyward, classified by outcome for pitcher.*

Examination of the distribution of Heyward's swings, classified by whether or not the outcome is favorable for the pitcher, reveals that Heyward achieves the best outcomes in the lower outside quadrant of the strike zone (*Fig. 39*). Thus, the pitcher should avoid this area. An optimal strategy for the pitcher would be to target the top inside quadrant of the strike zone, or the area located just below the strike zone on the inside. These findings support the previously suggested hypothesis concerning the umpires' misjudgment of Heyward's strike zone. When Heyward *does* swing at high inside pitches he is rarely successful, implying that the pitchers' successes on high inside pitches are occurring on non-swings.

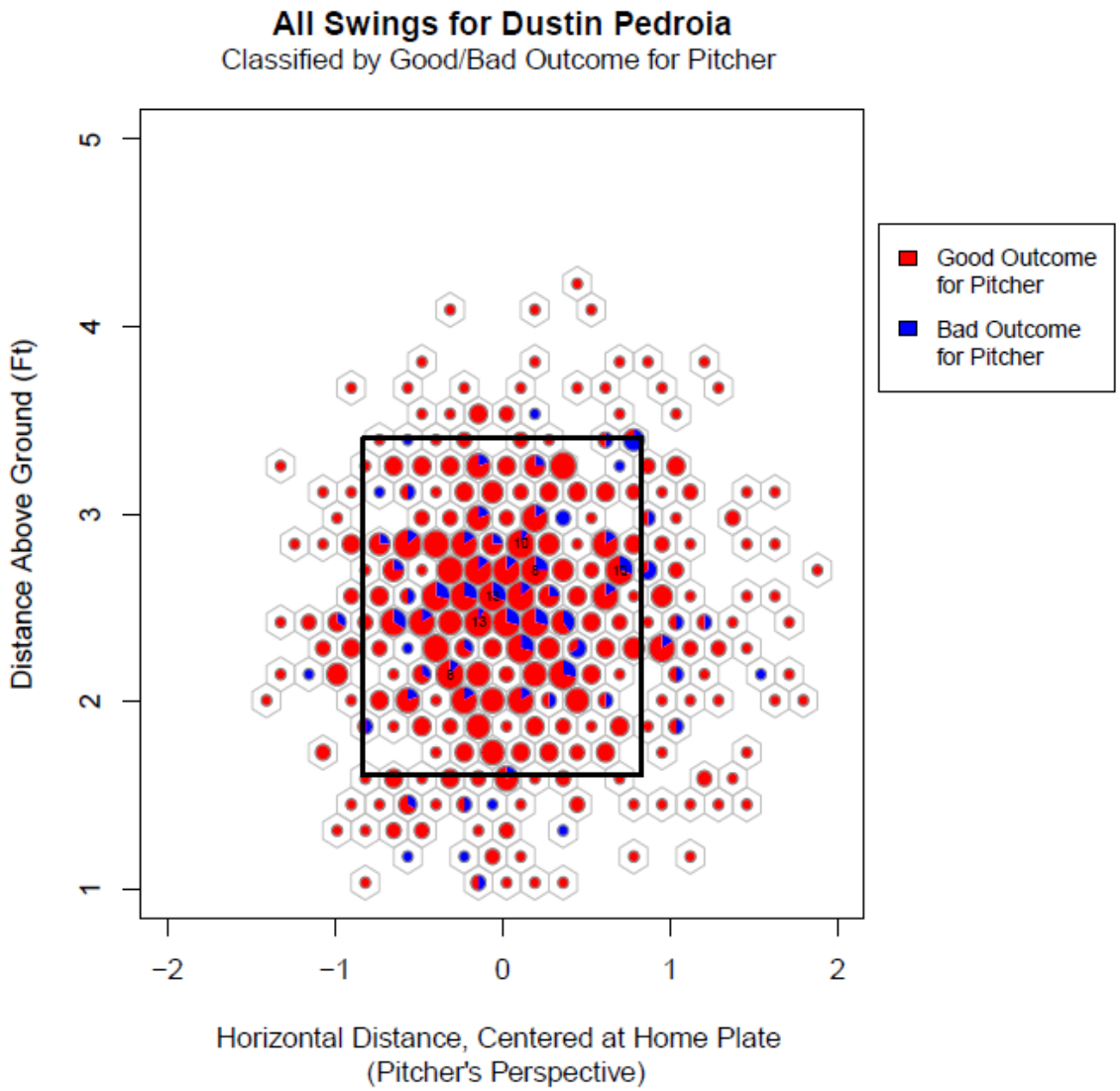

Figure 40. *Hexbin pie chart of all swings for Pedroia, classified by outcome for pitcher.*

Pedroia, once again, is less predictable than Heyward when it pretains to the areas in which his swings are successful. The pitcher should generally avoid the central area of the strike zone. Otherwise, though, the unfavorable outcomes for the pitcher appear to be fairly scattered (*Fig. 40*).

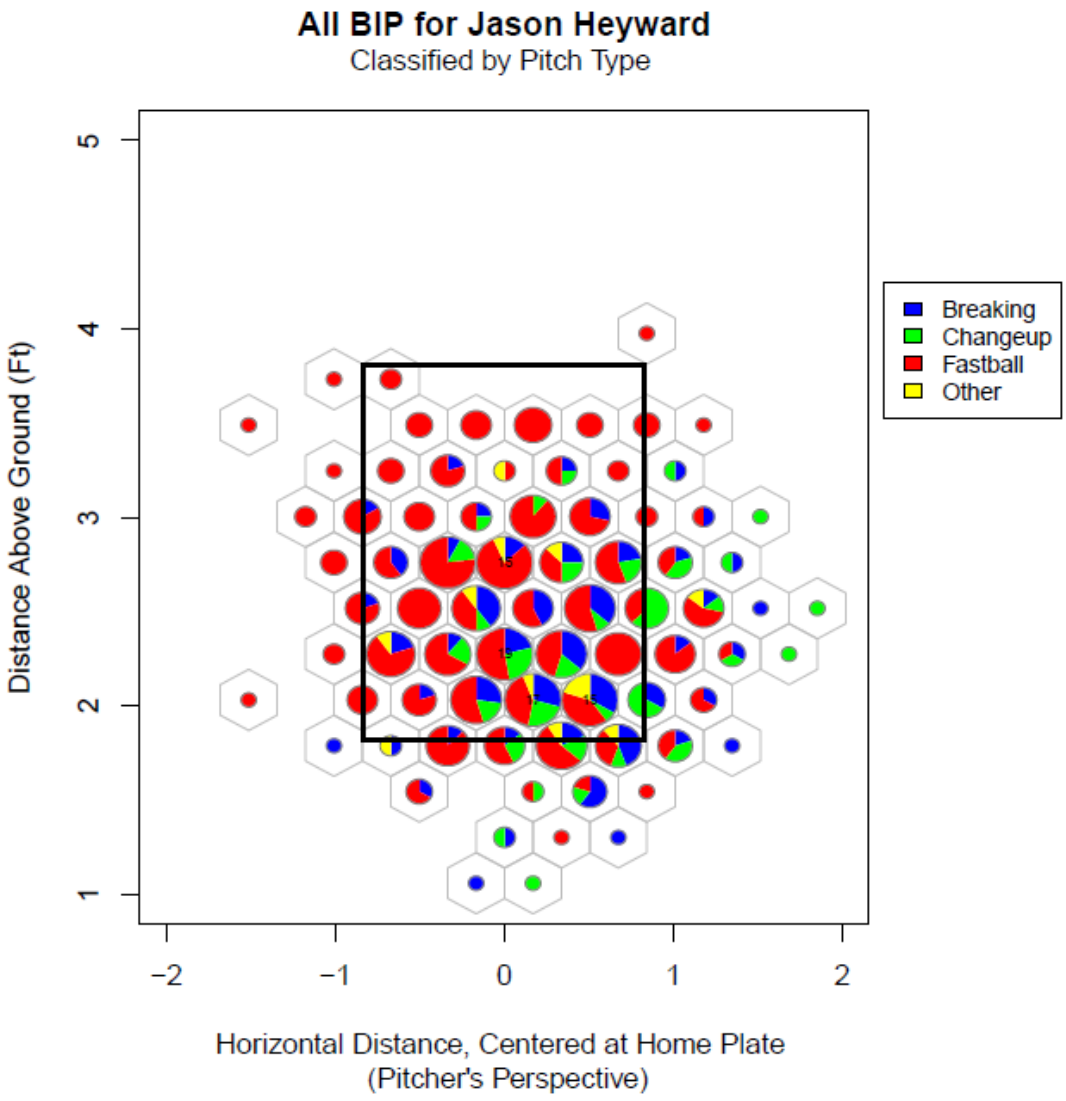

Figure 41. *Hexbin pie chart of all BIP for Heyward, classified by pitch type.*

Of the balls put into play by Heyward, the ones resulting from pitches located high in the strike zone are nearly all fastballs. This finding is notable because it was previously illustrated that Heyward also frequently receives breaking balls in this area; however, those pitches tend not to be put into play. Furthermore, a prior hexbin pie chart (*Fig. 35*) showed that Heyward swings very often at changeups below the strike zone; however, he has relatively little success at putting those pitches into play. The pitch type distribution for balls put into play on the outside periphery of the strike zone, where Heyward is known to swing often, more or less mirrors the pitch type distributions of the pitches received, and swung at, by Heyward (*Fig. 41*).

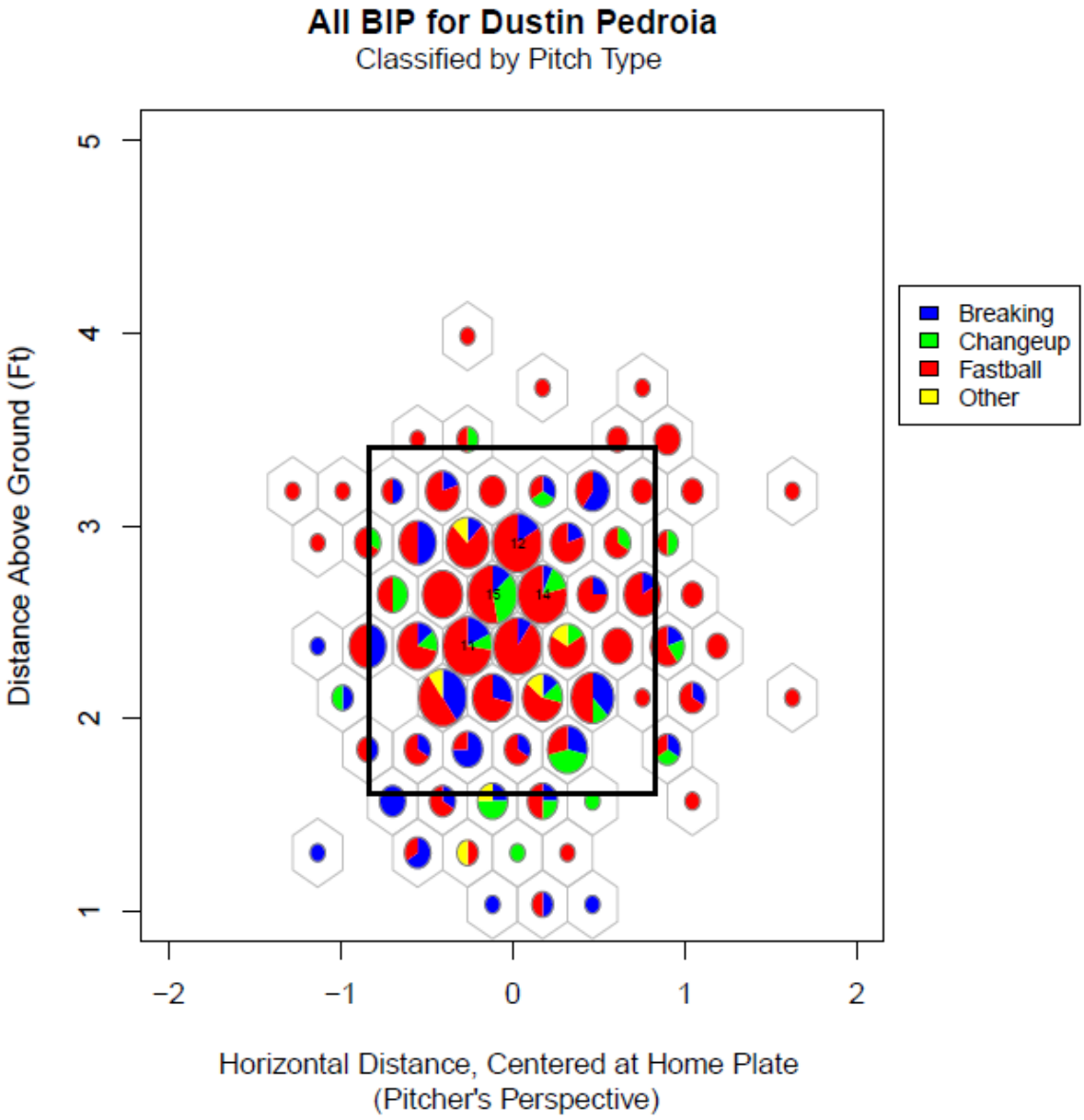

Figure 42. *Hexbin pie chart of all BIP for Pedroia, classified by pitch type.*

It was previously observed that Pedroia tends to swing often at breaking balls below the strike zone (*Fig. 36*). However, he does not put these pitches into play as often as would be expected. Pedroia's spatial pitch type distribution of balls put into play is otherwise fairly representative of the pitch type distribution of his swings (*Figs. 36 & 42*).

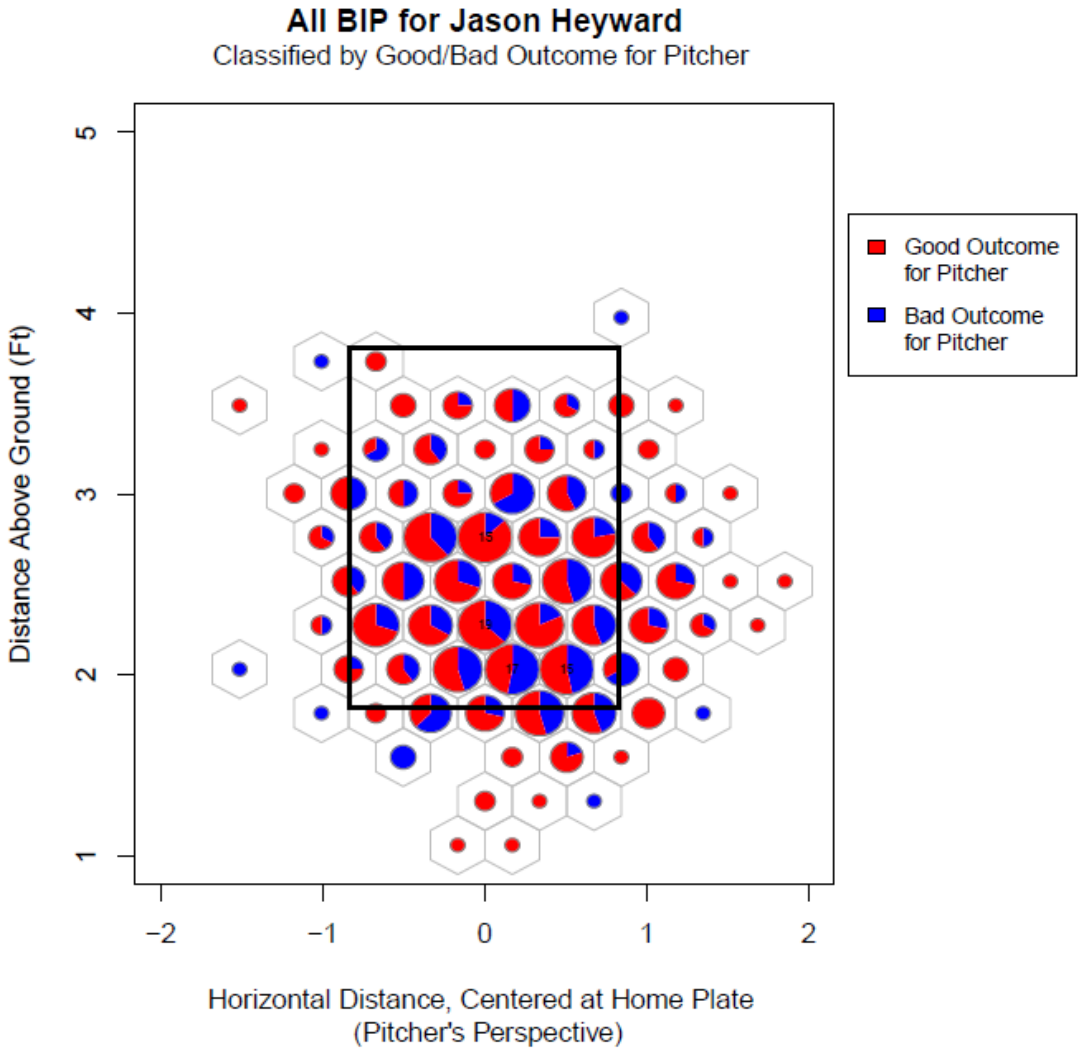

Figure 43. *Hexbin pie chart of all BIP for Heyward, classified by outcome for pitcher.*

The spatial distribution of balls put into play with unfavorable outcomes for the pitcher reflects the previously noted trend in Heyward's power hitting. Specifically, Heyward's most powerful base hits resulted from pitches in the lower outside quadrant of the strike zone (*Fig. 33).* This is the same region where the pitcher can expect to be most likely to suffer an unfavorable outcome. When Heyward hits pitches that are placed too far away from him outside of the strike zone, he seems to not be able to get enough force behind the ball; hence, this area has more favorable outcomes for the pitcher (whose fielding teammates would get Heyward out) despite the fact that the balls are put into play (*Fig. 43*).

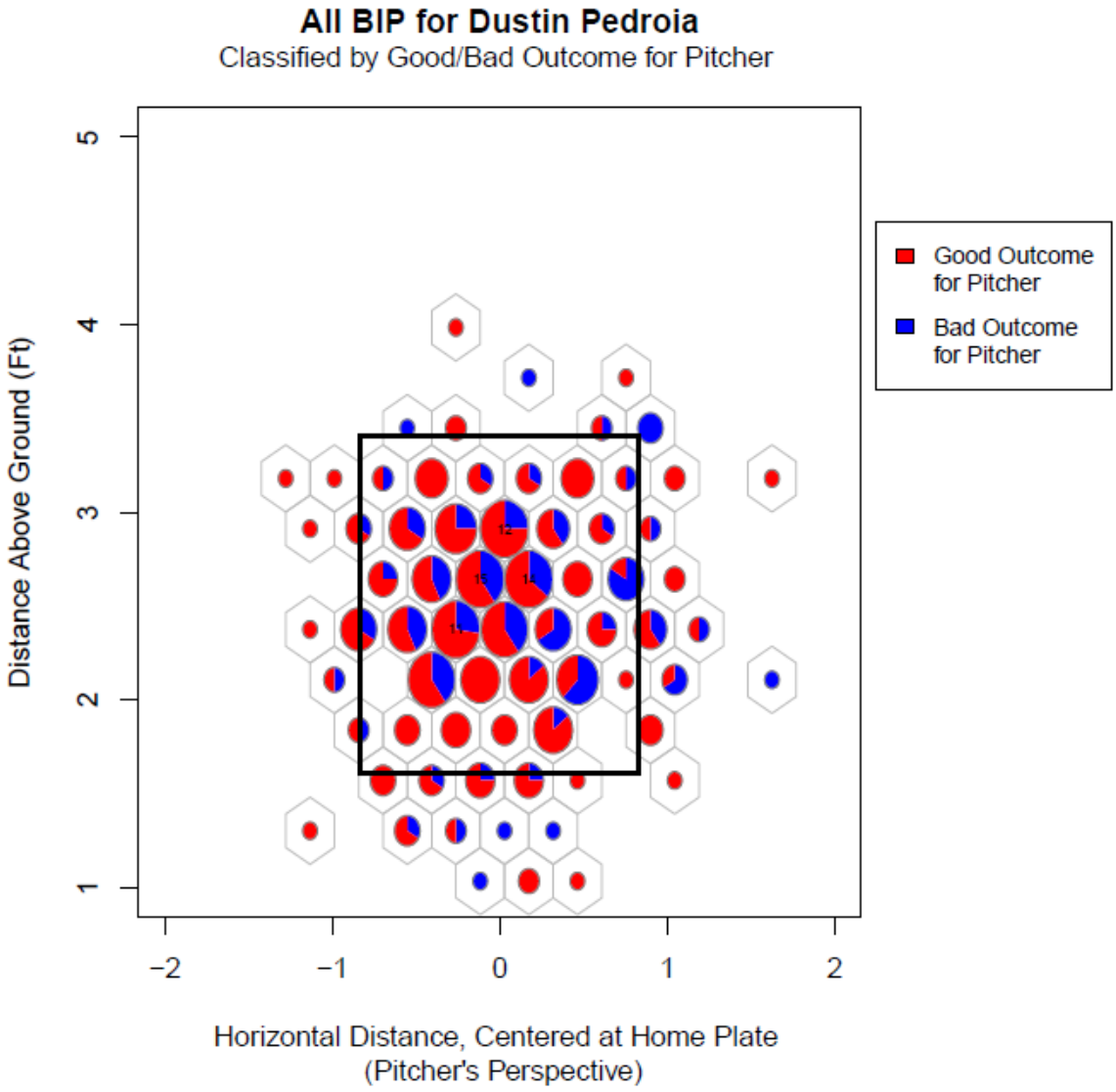

Figure 44. *Hexbin pie chart of all BIP for Pedroia, classified by outcome for pitcher.*

Pedroia's successes on balls put into play also reflect the spatial distribution of his power hits, which are derived mostly from pitches placed in the center of the strike zone and outside of the strike zone border on the side closest to his body (*Figs. 34 & 44*).

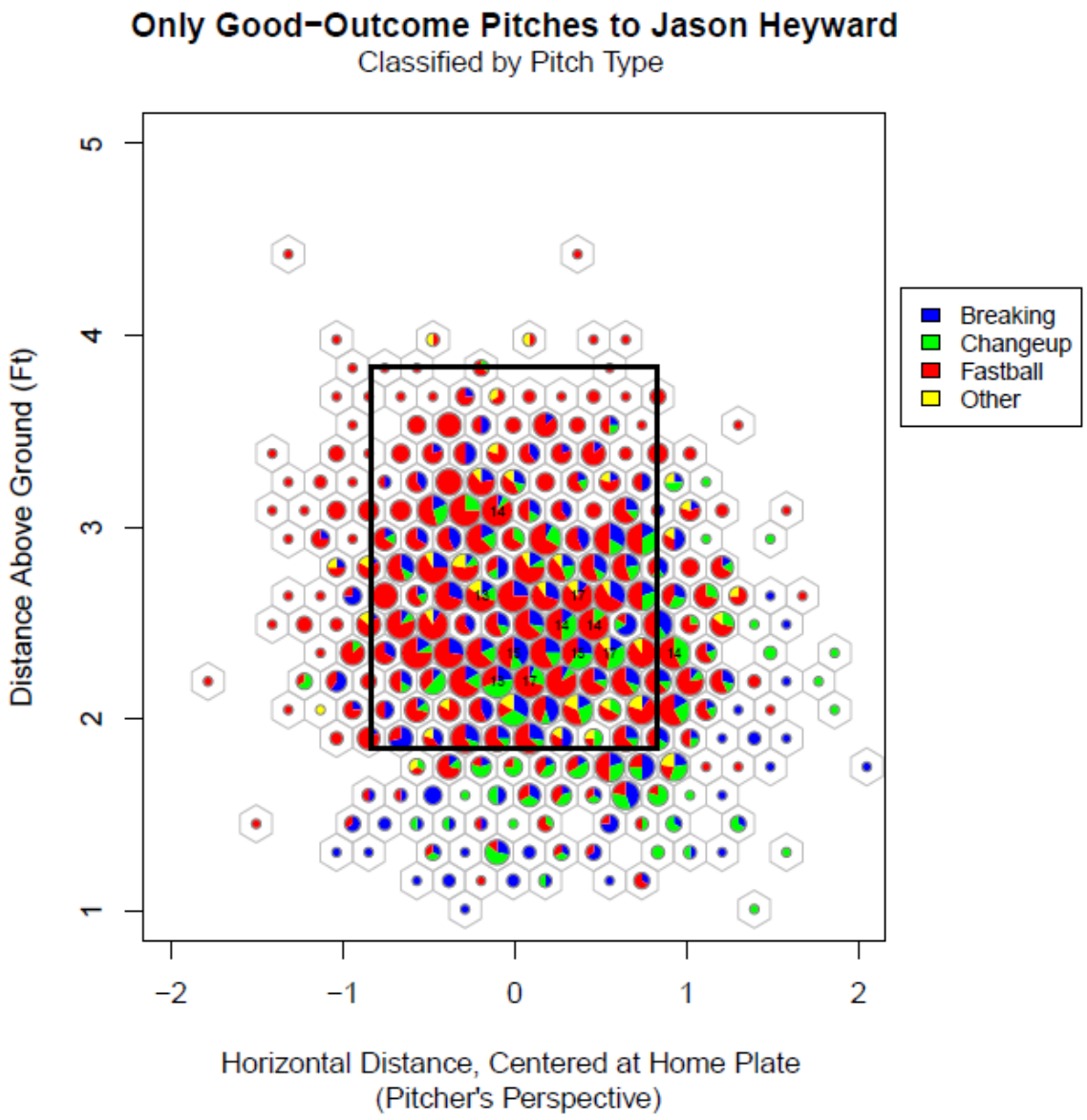

Figure 45. *Hexbin pie chart of all good-outcome pitches to Heyward, classified by pitch type.*

The most striking feature of the spatial pitch type distribution of pitches to Heyward that result in favorable outcomes for the pitcher is the relatively high frequency of changeups below the strike zone and outside of the strike zone on the far side (*Fig. 45*). Compared with the overall pitch type distribution (*Fig. 23*), pitchers have far more success against Heyward with changeups than would be expected if the two distributions mirrored one another.

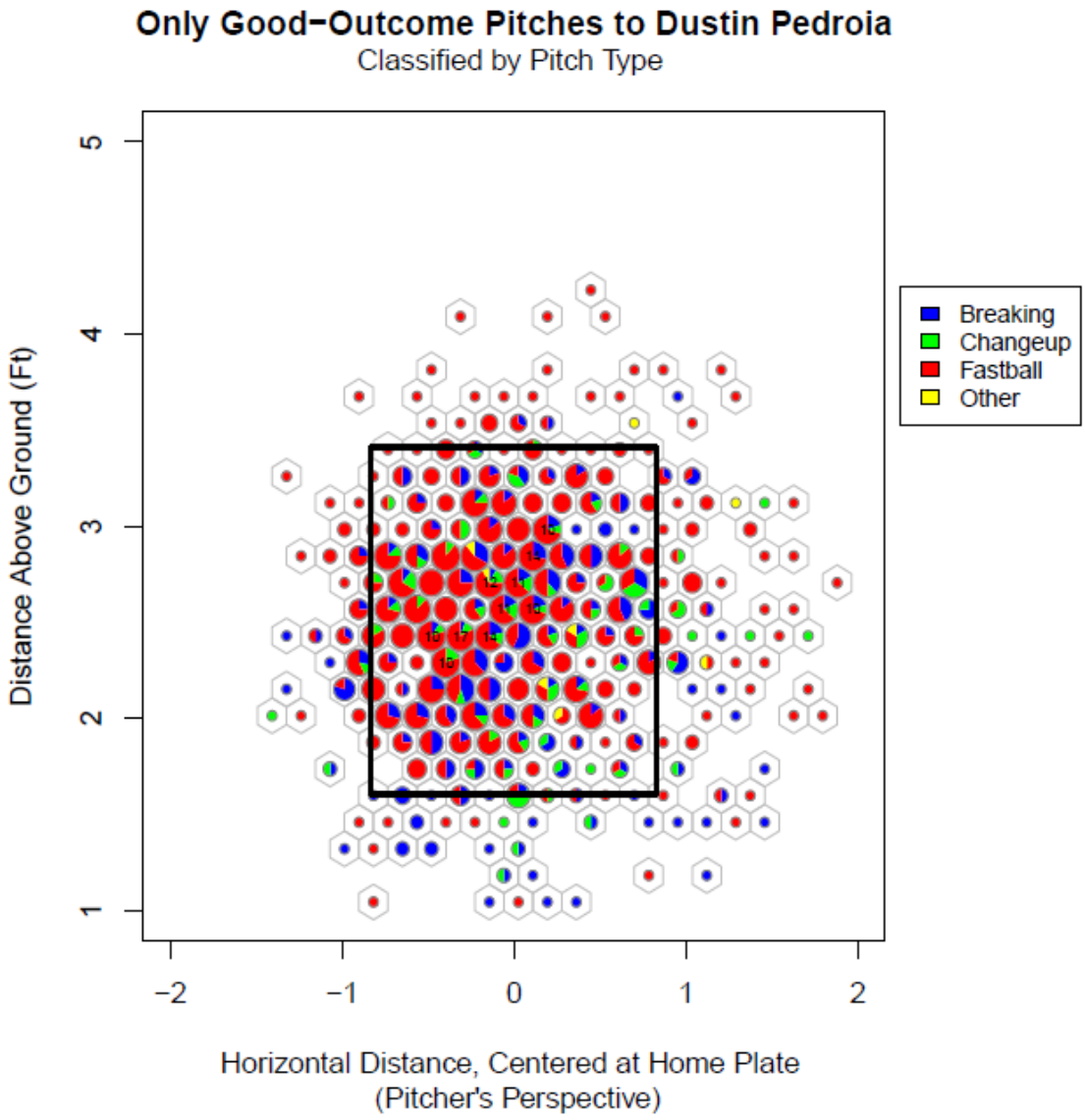

Figure 46. *Hexbin pie chart of all good-outcome pitches to Pedroia, classified by pitch type.*

Fastballs pitched to Pedroia on the inside of the strike zone, and on the inner periphery, tend to be an unsuccessful strategy. Based on the spatial pitch type distribution for all pitches to Pedroia, there would be far more favorable outcomes expected on inside fastballs than were actually observed (*Figs. 24 & 46*). Otherwise, there are no notable trends among the pitch types with favorable outcomes when pitching to Pedroia. There appears to be a high density of fastballs inside the strike zone with good outcomes for the pitcher, but since Pedroia also receives fastballs inside the strike zone more often than any other type of pitch, this finding is trivial.

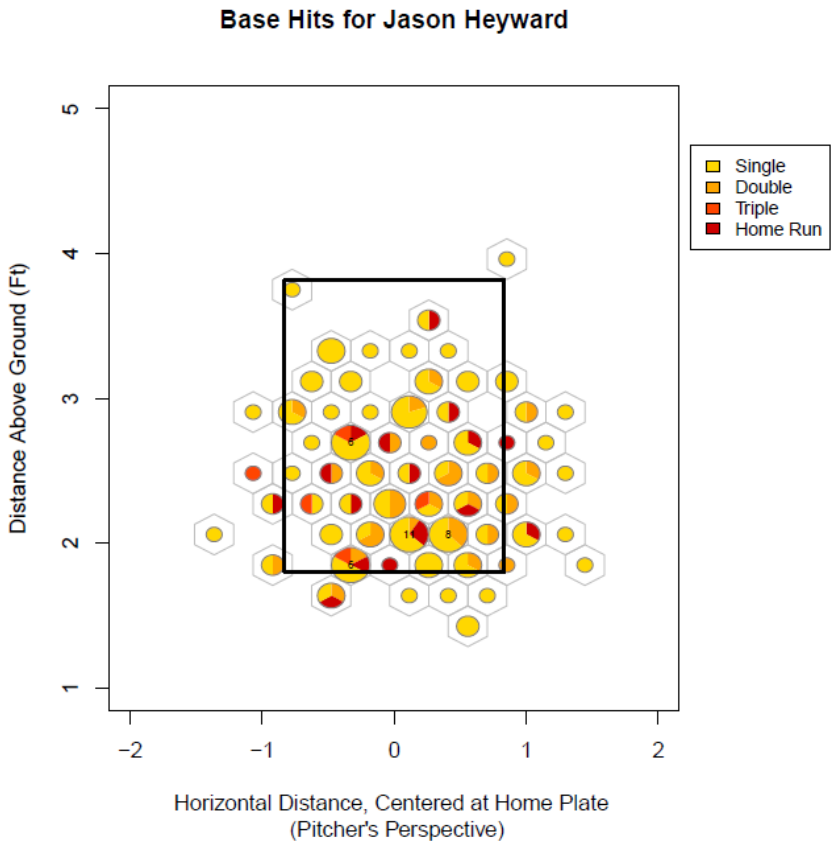

Figure 47. *Hexbin pie chart of all base hits for Heyward, classified by base hit type.*

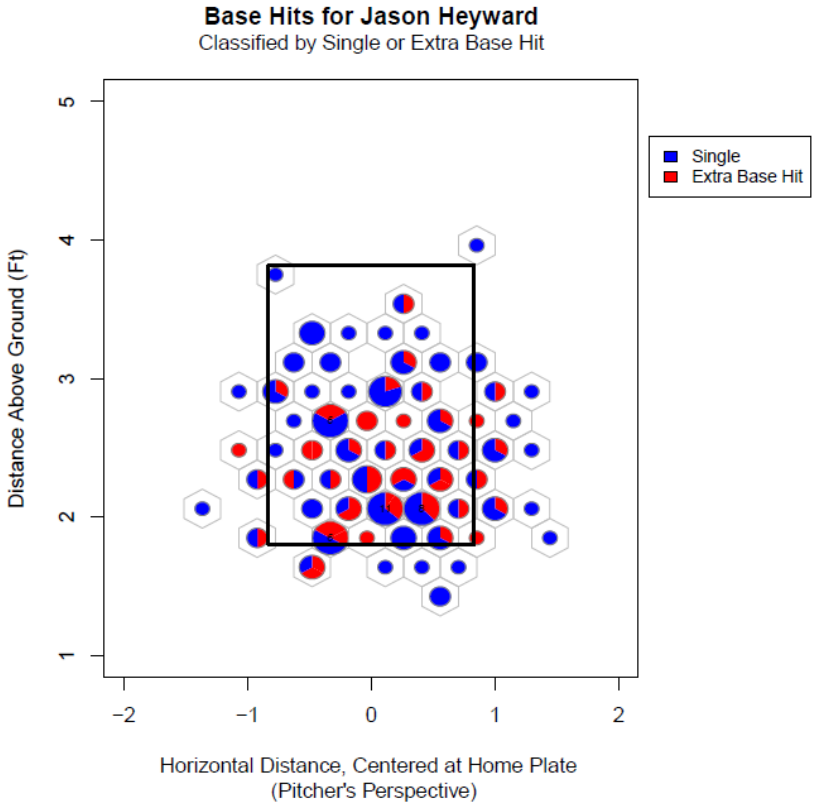

Figure 48. *Hexbin pie chart of all base hits for Heyward, classified by power.*

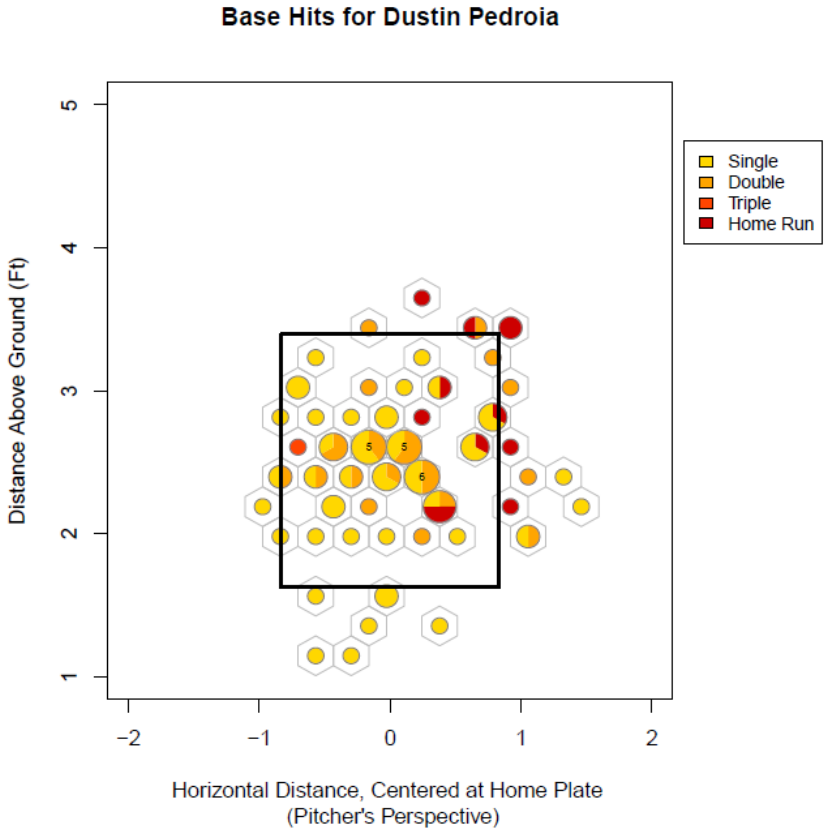

Figure 49. *Hexbin pie chart of all base hits for Pedroia, classified by base hit type.*

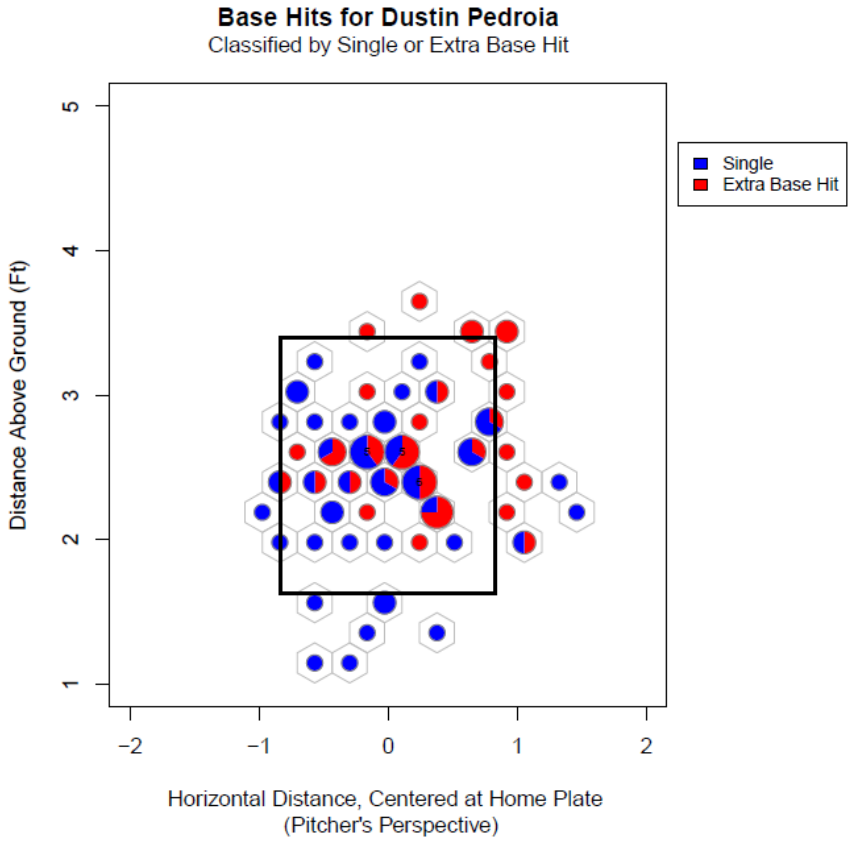

Figure 50. *Hexbin pie chart of all base hits for Pedroia, classified by power.*

Consideration of only base hits reveals that Heyward's power hits, as previously mentioned, tend to take advantage of pitches that are located low in the strike zone. He has a fair number of base hits off of pitches outside of the strike zone, mostly low and away from his body. However, these base hits are primarily singles, with a few doubles, as opposed to extra base hits (*Figs. 47 & 48*).

Pedroia earns base hits most frequently when he receives pitches in the center of the strike zone. However, his power hits, particularly home runs, result from pitches on the inside – often so close to his body that they do not even fall into the strike zone (*Figs. 49 & 50*). The fact that Pedroia's base hits come from pitches that are clustered towards the inside is noteworthy, because he tends to receive more pitches towards the outside. As previously mentioned, this trend is most likely due to Pedroia's small stature.

DISCUSSION & POSSIBILITIES FOR FUTURE RESEARCH

Value of Graphical Analysis to Advance Scouting

The baseball.plot function produces 22 plots that can be used to graphically analyze the batting tendencies of any non-rookie Major League Baseball player. The value of this tool resides in the fact that the plots reveal information that cannot be discerned from a box score or any other calculated statistics. By accounting for spatial distributions, this tool can help pitchers mentally simulate facing a particular batter in a real-game situation, as well as visualize the optimal locations at which to aim certain types of pitches. These plots can reduce the degree of subjectivity that is associated with advance scouting, and also provide more concrete evaluations on which to base game strategy.

Graphical analysis has previously been utilized in advance scouting applications, but there have been missing pieces to the puzzle. This new tool is practical for large data sets, which is not necessarily true of the previously existing plotting methods. In order to examine the pitches received by one player during an entire season, which may amount to several hundred or thousand data points, it is essential to have a tool that can handle sizeable amounts of data. The plots generated by the baseball.plot function are not subject to the detriments of cluttering due to too many data points, nor obscuring of early-plotted points that may be covered up. These plots are able to effectively display the points while also accounting for the spatial aspect of the data. Whereas before the information contained in *multiple* plots would have to be synthesized and combined in order to draw a conclusion, now there exist three-dimensional plots that are able to consolidate that information into a *single* plot in a concise yet enlightening manner.

The sensitivity of the heat map density plots, heat map contour plots, and hexbin pie charts was substantiated in the examples of Heyward and Pedroia. The plots generated for these two players, who were known beforehand to have different batting tendencies, showed marked

discrepancies. Thus, the information gleaned from the plots is helpful in compiling advance scouting reports. If the plots had instead turned out to be invariable between two players known to have different batting patterns, then they would be essentially worthless; however, the plots' face validity has been proven legitimate. This scouting tool can be used to examine spatial trends in the swinging and hitting tendencies of one individual player. Moreover the plots can be compared and contrasted for multiple players in succession, as was done in the example concerning Heyward and Pedroia. The benefit of scouting the entire opposing team in order of batting lineup is that a pitcher can anticipate when he will need to alter the speed, location, or type of his pitch in specific manners.

Limitations

As with any statistical or graphical analysis technique, there were some limitations that were encountered in this project. First, there were some missing data: over 4,000 pitches were missing the spatial coordinates, so these observations had to be excluded from the analyses. The missing observations accounted for less than 0.6% of the data, which is a relatively low rate of missingness. However, to ensure the accuracy of any analysis, it is always ideal to minimize the extent of the missing data. The missing spatial coordinates were caused by malfunctions of the PITCHf/x system. Other variables are included in the data set for these observations, indicating that the system was in use at the time; the pitch trajectories were simply not recorded properly. Electronic devices are often imperfect, so minor glitches would be expected. Furthermore, the PITCHf/x system has only been in existence for a few years, so it may take some additional time before the technology is fully reliable. The key issue to check is whether or not there was any sort of systematic pattern underlying the absence of these data points. If, for example, there were a flaw in the data recording at one particular stadium, then the missing data points would be largely associated with that stadium's home team. An issue such as this would be problematic because it would introduce bias into the data collection procedure. Examination of the observations with missing spatial coordinates did not reveal any substantial trends in the missingness. Therefore,

while it was undesirable to have missing data due to the loss of information, it did not introduce any problems into the analysis.

Going along with the missing data, another possible limitation could be inaccuracies in the spatial coordinate measurements. The PITCHf/x technology is assumed to be both accurate and precise, but it is unrealistic to expect it to function perfectly every single time. Although the PITCHf/x cameras at each stadium are periodically re-calibrated by Sportvision, there will never be perfection within a single stadium, or exact duplication among different stadiums. The plate locations of pitches are "usually accurate to within an inch or so" – not a huge margin of error, but enough to warrant consideration (Fast, 2010). Slight errors in the coordinate measurements would not drastically alter the graphical analyses, but the user should still bear in mind the possibility of data recording errors or imperfections.

There are a couple of issues associated with the pitch type classifications. First of all, from the pitcher's perspective, pitch type is a subjective label. There are so many different pitch types, several of which are very similar to one another, that two different pitchers may classify the same pitch differently; the distinguishing features can be quite subtle (Fast, 2010). In other words, if two different pitchers were both told to throw a changeup, they might throw different variations. Furthermore, the PITCHf/x system assigns pitch type labels based on several measurements, including the ball's initial and final velocities, spin rate, and break angle and direction. As previously discussed, the measurements recorded by PITCHf/x are not always completely accurate. The issue of fine lines between pitch classifications compounds this problem. If a PITCHf/x measurement is slightly off, the pitch may be incorrectly classified. The fact that the automatic classification depends on multiple measured PITCHf/x parameters increases the likelihood of at least one measurement imperfection. The creators of the PITCHf/x data management system recognize this flaw, so the PITCHf/x data set includes a variable that calculates the confidence of the pitch type classification (Fast, 2010).

Human error could be another source of inaccuracy underlying the data. These graphical analyses rely heavily upon the outcomes of the pitches, as determined by the umpires. MLB umpires are highly experienced and skilled, but nevertheless they do make mistakes. In a separate analysis conducted by Patrick Kilgo, it was calculated that the calls were correct approximately 88% of the time. Considering that the umpires must remain alert at all times in order to make split-second decisions, 88% accuracy is relatively high. The incorrect calls will influence the analyses, though not to a detrimental extent. Incorrect calls would be expected to occur mostly on pitches located near the borders of the strike zone, since those judgments are the least obvious. Anyone utilizing these plots should keep that notion in mind, and realize that the figures are most error-prone at the strike zone borders.

Finally, due to the smoothing techniques applied to the data points in the density and contour plots, these figures are most appropriate for large data sets (i.e., players who bat often). For a player who did not receive many pitches, a small number of outliers would have a profound effect on the spacing and curvature of the contour lines, or the shading of the intensity function underlying a density plot. For example, a player with few at-bats could erroneously appear to have multiple batting hotspots if he coincidentally attained base hits in multiple locations; there would not be sufficient "background noise" to smooth out the data as intended.

Ideas for Further Analysis

Due to the revolutionary nature of PITCHf/x technology, the system is able to generate an extensive data set. There are numerous parameters that are measured, and even more that can be derived or calculated based on the information provided. Many of the PITCHf/x variables were not even considered in these analyses. By taking into account *all* of the information contained in the data set, there is a wealth of additional analyses that could be conceived. The major challenge is to comb through the available data to determine which parameters can provide valuable information without over-complicating the results.

One idea is to stratify by the handedness of the pitcher, sub-dividing each plot into one for left-handed pitches and one for right-handed pitches. The handedness of the pitcher can impact several characteristics of the pitch, including the spin and the break angle (Fast, 2010). It is hypothesized that batters may react differently based on the pitcher's handedness. These differences can be accounted for in terms of the types of pitches at which the batter swings, the locations of the pitches, and the success rate. For a left-handed pitcher, the batter's tendencies in response to right-handed pitches are irrelevant, and vice versa. Therefore, this classification would increase the significance of the graphical analyses for individual pitchers.

Another possible stratification scheme is based on pitch count. Specifically, batters might have different tendencies when there is a full count (i.e., three balls and two strikes). In this situation, a batter may swing less aggressively for fear of striking out. It would therefore be beneficial for the pitcher to have a separate strategy for a full-count scenario.

Future Technological Developments

Sportvision is developing two additional data collection systems: FIELDf/x and HITf/x. HITf/x, which tracks the batter's swinging mechanisms and the ball post-hit, has not been released yet. FIELDf/x, however, was installed this year at AT&T Park in San Francisco. By 2012, if the plan runs according to schedule, FIELDf/x will be in all 30 MLB stadiums. FIELDf/x is even more powerful than PITCHf/x. The system involves four cameras that track every player on the field, as well as the ball, for a total of over 2.5 million records per game. The measurements are said to be accurate to within one foot (Boudway, 2011).

The FIELDf/x technology is still being refined. The major issue is that it provides *too much* data. The system is so sensitive that it even picks up motion of birds flying across the field – information that is not useful for any sort of data analysis. There is also extraneous information that is baseball-related, such as players moving through warm-up drills. Removal of these data records would facilitate easier use of the true game data. The vast amount of data presents problems with electronic storage of the records as well. The PITCHf/x data set from the entire

2010 season consisted of approximately 710,000 records; FIELDf/x generates over 2.5 million records *per game*. Accounting for the 162 regular-season games of the 30 MLB teams, the overall number of records quickly grows to be astronomical. Ryan Zander, the general manager of Sportvision's baseball division, declared, "It's almost overwhelming how much data we're creating" (Boudway, 2011).

Overall, baseball data analytics is a fascinating field that is wide open for the implementation of innovative ideas. As technological advancements and analytical techniques continue to develop, more and more data become available. In turn, the range of possible applications of the data becomes essentially endless. Statistical and graphical analyses are assuming larger roles in the game strategies of MLB teams, at increasingly complex levels. The value of sabermetrics to the game of baseball may someday become so influential that baseball evolves into a battle of tactics rather than physical power and ability.

APPENDIX A: SAS CODE

```
                        **********************
                        *   Hillary Superak   *
                        *       THESIS        *
                        **********************;
libname hs '...';
** Import batter identification Excel file **;
PROC IMPORT OUT= WORK.batters
            DATAFILE= "C:\Documents and Settings\Hillary Superak\My
                  Documents\Grad School\THESIS\elias_to_lahman(1).xls"
            DBMS=EXCEL REPLACE;
     RANGE="'elias_to_lahman(1)$'";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

** Change PITCHf/x batter variable to numeric to match identification
batter variable -- Need to do this for merging the two data sets **;
data hs.baseball;
   set hs.pkbaseball;
   batter2 = input(batter, 6.0);
run;

** Match names of batter variables for merging **;
data work.batters;
   set work.batters;
   rename eliasid = batter;
run;

** Sort by batter before merging **;
proc sort data=work.batters;
   by batter;
run;

data hs.baseball;
   set hs.baseball;
   drop batter;
   rename batter2 = batter;
run;

proc sort data=hs.baseball;
   by batter;
run;

** Merge PITCHf/x data with batter identification data **;
data hs.baseball;
   merge hs.baseball work.batters;
   by batter;
run;
```

```
** Remove observations with missing pitch coordinates **;
data hs.baseball;
   set hs.baseball;
   where px ne .;
   batter_name = compress(first||last);
run;

** Data cleaning & new variable creation **;
data hs.baseball;
   set hs.baseball;

   if pitch_result = "In play- run(s)'"
   then pitch_result = "In play- run(s)";

   if pitch_result = "In play- run(s)" or pitch_result = "In play- no
      out" or pitch_result = "In play- out(s)"
   then BIP = 1;
   else BIP = 0;

   if pitch_result = "Foul" or pitch_result = "Foul (Runner Going)" or
      pitch_result = "Foul Bunt" or pitch_result = "Foul Tip" or
      pitch_result = "In play- no out" or pitch_result = "In play-
      out(s)" or pitch_result = "In play- run(s)" or pitch_result =
      "Missed Bunt" or pitch_result = "Swinging Pitchout" or
      pitch_result = "Swinging Strike" or pitch_result = "Swinging
      Strike (Bloc"
   then swing = 1;
   else swing = 0;

   if abs(px) < 0.833  and sz_bot <= pz <= sz_top
   then realstrike=1;
   else realstrike=0;

   if pitch_result = 'In play- run(s)' and atbat_result = 'Fan
      interference' or atbat_result = 'Home Run' or atbat_result =
      'Single' or atbat_result = 'Double' or atbat_result = 'Triple'
   then good = 0;

   else if pitch_result = 'In play- out(s)' and atbat_result = 'Fan
      interference' or atbat_result = 'Home Run'
      or atbat_result = 'Single' or atbat_result = 'Double' or
      atbat_result = 'Triple'
   then good = 0;

   else if pitch_result = 'In play- no out' and atbat_result = 'Fan
      interference' or atbat_result = 'Home Run' or atbat_result =
      'Single' or atbat_result = 'Double' or atbat_result = 'Triple'
   then good = 0;
   else good = 1;

   if pitch_result_type = 'B' then good = 0;
   if pitch_result_type = 'S' then good = 1;

   ** Type 1: Ball in Play
      Type 2: Swing
      Type 3: Base Hit
      Type 4: Swing and Hit **;
```

```
   if BIP = 1
   then type1 = 1;
   else type1 = 0;

   if BIP = 1 or swing = 1
   then type2 = 1;
   else type2 = 0;

   if BIP = 1 and good = 0
   then type3 = 1;
   else type3 = 0;

   if type2 = 1 and realstrike = 0
   then type4 = 1;
   else type4 = 0;

   px_pitcher = px*-1;

   if pitch_type = 'FC' or pitch_type = 'FF' or pitch_type = 'FT' or
      pitch_type = 'SI' then pitch = 'fastball';
   else if pitch_type = 'CU' or pitch_type = 'KC' or pitch_type = 'SL'
   then pitch = 'breaking ball';
   else if pitch_type = 'CH' then pitch = 'changeup';
   else pitch = 'other';

   if (pitch_result='In play- no out' or pitch_result='In play-
      run(s)') and atbat_result='Home Run' then basehit='Home Run   ';
   else if (pitch_result='In play- no out' or pitch_result='In play-
      run(s)') and atbat_result='Single' then basehit='Single  ';
   else if (pitch_result='In play- no out' or pitch_result='In play-
      run(s)') and atbat_result='Double' then basehit='Double  ';
   else if (pitch_result='In play- no out' or pitch_result='In play-
      run(s)') and atbat_result='Triple' then basehit='Triple  ';
   else basehit='No Base Hit';

   if (pitch_result='In play- out(s)' or pitch_result='In play- no out'
      or pitch_result='In play- run(s)') then BIP_outcome='Out';
   else BIP_outcome='';
   if (pitch_result='In play- out(s)' or pitch_result='In play- no out'
      or pitch_result='In play- run(s)') and (atbat_result='Triple' or
      atbat_result='Double' or atbat_result='Home Run')
   then BIP_outcome='Extra Base Hit';
   if (pitch_result='In play- out(s)' or pitch_result='In play- no out'
      or pitch_result='In play- run(s)') and atbat_result='Single'
   then BIP_outcome='Single';
run;

** Export to CSV file **;
PROC EXPORT DATA= HS.BASEBALL
            OUTFILE= "C:\Documents and Settings\Hillary Superak\My
   Documents\Grad School\THESIS\pitches.csv"
            DBMS=CSV LABEL REPLACE;
     PUTNAMES=YES;
RUN;
```

APPENDIX B: R CODE

```
                    ################################
                    #             THESIS           #
                    #  PITCHf/x Plotting Function  #
                    #        Hillary Superak       #
                    ################################

library(sp)
library(splancs)
library(spatial)
library(KernSmooth)
library(spatstat)
library(spdep)
library(dichromat)
library(graphics)

require(hexbin)

## Read in data ##
pitchfx <- read.csv("C:/Documents and Settings/Hillary Superak/My
   Documents/Grad School/THESIS/pitchesR.csv", header=T)

## Zoom in on relevant pitches ##
pitches3 <- subset(pitchfx, px_pitcher>-2 & px_pitcher<2 & pz>1 & pz<5)

## Create variable including batter first & last name ##
pitches2 <- data.frame(pitches3[,1:34], batter=paste(pitches3[,22],
   pitches3[,23]))

## Sort by batter ##
pitches2 <- pitches2[order(pitches2$batter),]

## Hexbin Pie Function ##
hexbinpie <- function(x, y, kat, numbin, xbnds=range(x), bnds=range(y),
   pal=terrain.colors(length(levels(as.factor(kat)))),
   hex="gray", circ=NA, cnt="black", ...) {

   hb <- hexbin(x, y, xbnds=xbnds, ybnds=ybnds, IDs=T, xbin = numbin)
   hbc <- hcell2xy(hb)
   rx <- diff(hb@xbnds) / (2 * hb@xbins)
   ry <- diff(hb@ybnds) / (2 * hb@xbins*hb@shape)
   hexC <- hexcoords(dx=rx, dy=ry/sqrt(3), n=1)
   nl <- length(levels(as.factor(kat)))
   zbnds <- quantile(hb@count,prob=c(.05,.95), na.rm=TRUE )
   maxhb <- max(hb@count)
   zz <- pmax(pmin(sqrt(hb@count/zbnds[2]),1),0.2)
   tt <- unclass(table(kat,hb@cID))
   for (i in seq(along=zz)) {
      if (!is.na(hex)) polygon(hbc$x[i]+hexC$x, hbc$y[i]+hexC$y,
      col=NA, border=hex)
      tp <- pi/2 - 2*pi*c(0,cumsum(tt[,i])/sum(tt[,i]))
```

```
      for (j in 1:nl) {
         if (tp[j+1]==tp[j]) next
         pp <- seq(tp[j], tp[j+1], length=floor((tp[j]-tp[j+1])*4)+2)
         xi <- hbc$x[i]+c(0,zz[i]*rx*cos(pp))
         yi <- hbc$y[i]+c(0,zz[i]*ry*sin(pp))
         polygon(xi,yi, col=pal[j], border=NA,...)
      }
   if (!is.na(circ)) polygon(hbc$x[i]+rx*zz[i]*cos((1:18)*pi/9),
       hbc$y[i]+ry*zz[i]*sin((1:18)*pi/9), col=NA, border=circ)
   }
   for (i in seq(along=zz)) {
      if ((!is.na(cnt))&(hb@count[i]>zbnds[2]))
      text(hbc$x[i],hbc$y[i],hb@count[i],col=cnt,cex=.5)
   }
}

################################
## Function to Generate Plots ##
################################

baseball.plot <- function(playername) {

## Sub-set Data ##
pitches <- subset(pitches2, batter==playername)
pitches.p <- as.points(pitches$px_pitcher, pitches$pz)
BIPdata <- subset(pitches, BIP==1)
BIP.p <- as.points(BIPdata$px_pitcher, BIPdata$pz)
swingdata <- subset(pitches, swing==1)
swing.p <- as.points(swingdata$px_pitcher, swingdata$pz)
BHdata <- subset(pitches, basehit != 'No Base Hit')
BH.p <- as.points(BHdata$px_pitcher, BHdata$pz)
gooddata <- subset(pitches, good==1)
good.p <- as.points(gooddata$px_pitcher, gooddata$pz)
bad.p <- as.points(pitches$px_pitcher[pitches$good==0],
   pitches$pz[pitches$good==0])

mypolyr <- sbox(pitches.p, xfrac=0.1, yfrac=0.1)
bandw <- 0.25
grid.number <- 40
hs.color <- colorRampPalette(c("blue", "yellow", "orange", "red"))

######################
## Hexbin Pie Charts ##
######################

## All pitches - classify by good/bad outcome for pitcher ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN All Pitches - Classify by Good or Bad
   Outcome for Pitcher vs. ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(pitches$px_pitcher, pitches$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2.25,2.25), ylim=c(0.75,5.25))
title(paste("All Pitches to ", pitches[1,35], sep=""))
mtext("Classified by Good/Bad Outcome for Pitcher \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
```

```
legend(2.5, 5, c("\nGood Outcome \nfor Pitcher", "\nBad Outcome \nfor
   Pitcher"), fill=c("red","blue"), cex=0.8, bty="n")
hexbinpie(pitches$px_pitcher, pitches$pz, kat=pitches$good, 20,
   hex = "gray", circ = "gray50", pal = c("blue","red"))
segments(2.55, 4.75, 2.55, 3.8, lty=1, lwd=1, col='black')
segments(2.55, 4.75, 4.05, 4.75, lty=1, lwd=1, col='black')
segments(4.05, 4.75, 4.05, 3.8, lty=1, lwd=1, col='black')
segments(4.05, 3.8, 2.55, 3.8, lty=1, lwd=1, col='black')
## Add strike zone ##
SZtop <- mean(pitches$sz_top)
SZbot <- mean(pitches$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## All pitches - classify by swing/no swing (type 2) ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN All Pitches - Classify by Swing or No
   Swing - ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(pitches$px_pitcher, pitches$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2.25,2.25), ylim=c(0.75,5.25))
title(paste("All Pitches to ", pitches[1,35], sep=""))
mtext("Classified by Swing/No Swing \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.75, 4.5, c("Swing", "No Swing"), fill=c("red","blue"),
   cex=0.8)
hexbinpie(pitches$px_pitcher, pitches$pz, kat=pitches$type2, 20,
   hex = "gray", circ = "gray50", pal = c("blue","red"))
SZtop <- mean(pitches$sz_top)
SZbot <- mean(pitches$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## All pitches - classify by BIP/No BIP (type 1) ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN All Pitches - Classify by BIP or No BIP
   - ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(pitches$px_pitcher, pitches$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2.25,2.25), ylim=c(0.75,5.25))
title(paste("All Pitches to ", pitches[1,35], sep=""))
mtext("Classified by BIP/No BIP \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.75, 4.5, c("BIP", "No BIP"), fill=c("red","blue"), cex=0.8)
hexbinpie(pitches$px_pitcher, pitches$pz, kat=pitches$BIP, 20,
   hex = "gray", circ = "gray50", pal = c("blue","red"))
SZtop <- mean(pitches$sz_top)
SZbot <- mean(pitches$sz_bot)
```

```
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## All pitches - classify by base hit/no base hit (type 3) ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN All Pitches - Classify by Base Hit or No
   Base Hit - ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(pitches$px_pitcher, pitches$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2.25,2.25), ylim=c(0.75,5.25))
title(paste("All Pitches to ", pitches[1,35], sep=""))
mtext("Classified by Base Hit/No Base Hit \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.75, 4.5, c("Base Hit", "No Base Hit"), fill=c("red","blue"),
   cex=0.8)
hexbinpie(pitches$px_pitcher, pitches$pz, kat=pitches$BIP, 20,
   hex = "gray", circ = "gray50", pal = c("blue","red"))
SZtop <- mean(pitches$sz_top)
SZbot <- mean(pitches$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## All pitches - classify by base hit type ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN All Pitches - Classify by Base Hit Type
   - ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(pitches$px_pitcher, pitches$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2.25,2.25), ylim=c(0.75,5.25))
title(paste("All Pitches to ", pitches[1,35], sep=""))
mtext("Classified by Type of Base Hit\n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.75, 4.5, c("No Base Hit", "Single", "Double", "Triple", "Home
   Run"), fill=c("white","gold", "orange", "orangered", "red3"),
   cex=0.8)
hexbinpie(pitches$px_pitcher, pitches$pz, kat=pitches$basehit, 20,
   hex = "gray", circ = "gray50", pal = c("orange","red3", "white",
   "gold", "orangered"))
SZtop <- mean(pitches$sz_top)
SZbot <- mean(pitches$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()
```

```
## Only BH - classify by BH type ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN Only BH - Classify by BH Type - ",
   pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, xaxp=c(-2,2,5), yaxp=c(1,5,5),
   mar=par()$mar+c(1,1,1,6))
plot(BHdata$px_pitcher, BHdata$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2,2), ylim=c(1,5))
title(paste("Base Hits for ", pitches[1,35], sep=""))
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.25, 4.75, c("Single", "Double", "Triple", "Home Run"),
   fill=c("gold", "orange", "orangered", "red3"), cex=0.8)
hexbinpie(BHdata$px_pitcher, BHdata$pz, kat=BHdata$basehit, 10,
   hex = "gray", circ = "gray50", pal = c("orange","red3", "white",
   "gold", "orangered"))
SZtop <- mean(BHdata$sz_top)
SZbot <- mean(BHdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Only BH - classify by single vs other ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN Only BH - Classify by Single vs Other -
   ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(BHdata$px_pitcher, BHdata$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2,2), ylim=c(1,5))
title(paste("Base Hits for ", pitches[1,35], sep=""))
mtext("Classified by Single or Extra Base Hit\n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.25, 4.75, c("Single", "Extra Base Hit"), fill=c("blue",
   "red"), cex=0.8)
hexbinpie(BHdata$px_pitcher, BHdata$pz, kat=BHdata$basehit, 10,
   hex = "gray", circ = "gray50", pal = c("red", "red", "white",
   "blue", "red"))
SZtop <- mean(BHdata$sz_top)
SZbot <- mean(BHdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Only BIP - classify by good/bad outcome for pitcher ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN Only BIP - Classify by Good or Bad
   Outcome for Pitcher vs. ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(BIPdata$px_pitcher, BIPdata$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2,2), ylim=c(1,5))
title(paste("All BIP for ", pitches[1,35], sep=""))
mtext("Classified by Good/Bad Outcome for Pitcher \n")
```

```
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.25, 4.75, c("\nGood Outcome \nfor Pitcher", "\nBad Outcome
   \nfor Pitcher"), fill=c("red","blue"), cex=0.8, bty="n")
hexbinpie(BIPdata$px_pitcher, BIPdata$pz, kat=BIPdata$good, 10,
   hex = "gray", circ = "gray50", pal = c("blue","red"))
segments(2.25, 4.55, 2.25, 3.65, lty=1, lwd=1, col='black')
segments(2.25, 4.55, 3.65, 4.55, lty=1, lwd=1, col='black')
segments(3.65, 4.55, 3.65, 3.65, lty=1, lwd=1, col='black')
segments(3.65, 3.65, 2.25, 3.65, lty=1, lwd=1, col='black')
SZtop <- mean(BIPdata$sz_top)
SZbot <- mean(BIPdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Only BIP - classify by pitch type ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN Only BIP - Classify by Pitch Type - ",
   pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(BIPdata$px_pitcher, BIPdata$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2,2), ylim=c(1,5))
title(paste("All BIP for ", pitches[1,35], sep=""))
mtext("Classified by Pitch Type \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.25, 4.25, c("Breaking", "Changeup", "Fastball", "Other"),
   fill=c("blue","green","red","yellow"), cex=0.8)
hexbinpie(BIPdata$px_pitcher, BIPdata$pz, kat=BIPdata$pitch, 10,
   hex = "gray", circ = "gray50", pal=c("blue","green","red","yellow"))
SZtop <- mean(BIPdata$sz_top)
SZbot <- mean(BIPdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Only swings - classify by pitch type ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN Only Swings - Classify by Pitch Type -
   ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(swingdata$px_pitcher, swingdata$pz, type="n", xlab="",
   ylab="Distance Above Ground (Ft)", xlim=c(-2,2), ylim=c(1,5))
title(paste("All Swings for ", pitches[1,35], sep=""))
mtext("Classified by Pitch Type \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.25, 4.25, c("Breaking", "Changeup", "Fastball", "Other"),
   fill=c("blue","green","red","yellow"), cex=0.8)
hexbinpie(swingdata$px_pitcher, swingdata$pz, kat=swingdata$pitch, 20,
   hex = "gray", circ = "gray50", pal=c("blue","green","red","yellow"))
```

```
SZtop <- mean(swingdata$sz_top)
SZbot <- mean(swingdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## All pitches - classify by pitch type (for comparison) ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN All Pitches - Classify by Pitch Type -
   ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(pitches$px_pitcher, pitches$pz, type="n", xlab="", ylab="Distance
   Above Ground (Ft)", xlim=c(-2.25,2.25), ylim=c(0.75,5.25))
title(paste("All Pitches to", BIPdata[1,35]))
mtext("Classified by Pitch Type \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.65, 5, c("Breaking", "Changeup", "Fastball", "Other"),
   fill=c("blue","green","red","yellow"), cex=0.8)
hexbinpie(pitches$px_pitcher, pitches$pz, kat=pitches$pitch, 20,
   hex = "gray", circ = "gray50", pal=c("blue","green","red","yellow"))
SZtop <- mean(pitches$sz_top)
SZbot <- mean(pitches$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Only swings - classify by BIP/no BIP ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN Only Swings - Classify by BIP or No BIP
   - ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(swingdata$px_pitcher, swingdata$pz, type="n", xlab="",
   ylab="Distance Above Ground (Ft)", xlim=c(-2,2), ylim=c(1,5))
title(paste("All Swings for", BIPdata[1,35]))
mtext("Classified by BIP/No BIP \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.25, 4.25, c("BIP", "No BIP"), fill=c("red","blue"), cex=0.8)
hexbinpie(swingdata$px_pitcher, swingdata$pz, kat=swingdata$BIP, 20,
   hex = "gray", circ = "gray50", pal = c("blue", "red"))
SZtop <- mean(swingdata$sz_top)
SZbot <- mean(swingdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()
```

```
## Only swings - classify by good/bad outcome for pitcher ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN Only Swings - Classify by Good or Bad
   Outcome for Pitcher vs. ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(swingdata$px_pitcher, swingdata$pz, type="n", xlab="",
   ylab="Distance Above Ground (Ft)", xlim=c(-2,2), ylim=c(1,5))
title(paste("All Swings for", BIPdata[1,35]))
mtext("Classified by Good/Bad Outcome for Pitcher \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.25, 4.75, c("\nGood Outcome \nfor Pitcher", "\nBad Outcome
   \nfor Pitcher"), fill=c("red","blue"), cex=0.8, bty="n")
hexbinpie(swingdata$px_pitcher, swingdata$pz, kat=swingdata$good, 20,
   hex = "gray", circ = "gray50", pal = c("blue","red"))
segments(2.25, 4.55, 2.25, 3.65, lty=1, lwd=1, col='black')
segments(2.25, 4.55, 3.65, 4.55, lty=1, lwd=1, col='black')
segments(3.65, 4.55, 3.65, 3.65, lty=1, lwd=1, col='black')
segments(3.65, 3.65, 2.25, 3.65, lty=1, lwd=1, col='black')
SZtop <- mean(swingdata$sz_top)
SZbot <- mean(swingdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Only good outcome - classify by pitch type ##
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/HEXBIN Only Good Outcome for Pitcher - Classify
   by Pitch Type - ", pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T, mar=par()$mar+c(1,1,1,6.5))
plot(gooddata$px_pitcher, gooddata$pz, type="n", xlab="",
   ylab="Distance Above Ground (Ft)", xlim=c(-2,2), ylim=c(1,5))
title(paste("Only Good-Outcome Pitches to", gooddata[1,35]))
mtext("Classified by Pitch Type \n")
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
legend(2.25, 4.25, c("Breaking", "Changeup", "Fastball", "Other"),
   fill=c("blue","green","red","yellow"), cex=0.8)
hexbinpie(gooddata$px_pitcher, gooddata$pz, kat=gooddata$pitch, 20,
   hex = "gray", circ = "gray50", pal=c("blue","green","red","yellow"))
SZtop <- mean(gooddata$sz_top)
SZbot <- mean(gooddata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()
```

```
##################
## Density Plots ##
##################

## BIP Density Plot ##
BIPdens <- ppp(BIPdata$px_pitcher, BIPdata$pz, c(-2,2), c(1,5))
BIPdens.z <- density.ppp(BIPdens)
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/DENSITY BIP - ", pitches[1,35], ".pdf",
   sep=""))
par(pty="m", xpd=T)
plot(BIPdens.z, col=hs.color(100), main="", axes=T, xlab="", ylab="")
text(0, 5.35, paste("BIP for", gooddata[1,35]), cex=1.25, font=2)
text(0, 0.25, "Horizontal Distance, Centered at Home Plate \n
   (Pitcher's Perspective)")
text(-2.6, 3, "Distance Above Ground (Ft)", srt=90)
SZtop <- mean(BIPdata$sz_top)
SZbot <- mean(BIPdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Swing Density Plot ##
swing.dens <- ppp(swingdata$px_pitcher, swingdata$pz, c(-2,2), c(1,5))
swing.dens.z <- density.ppp(swing.dens)
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/DENSITY swings - ", pitches[1,35], ".pdf",
   sep=""))
par(pty="m", xpd=T)
plot(swing.dens.z, col=hs.color(100), main="", axes=T, xlab="",
   ylab="")
text(0, 5.35, paste("Swings for", gooddata[1,35]), cex=1.25, font=2)
text(0, 0.25, "Horizontal Distance, Centered at Home Plate \n
   (Pitcher's Perspective)")
text(-2.6, 3, "Distance Above Ground (Ft)", srt=90)
SZtop <- mean(swingdata$sz_top)
SZbot <- mean(swingdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Base Hit Density Plot ##
BH.dens <- ppp(BHdata$px_pitcher, BHdata$pz, c(-2,2), c(1,5))
BH.dens.z <- density.ppp(BH.dens)
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/DENSITY Base Hits - ", pitches[1,35], ".pdf",
   sep=""))
par(pty="m", xpd=T)
plot(BH.dens.z, col=hs.color(100), main="", axes=T, xlab="", ylab="")
text(0, 5.35, paste("Base Hits for", gooddata[1,35]), cex=1.25, font=2)
text(0, 0.25, "Horizontal Distance, Centered at Home Plate \n
   (Pitcher's Perspective)")
text(-2.6, 3, "Distance Above Ground (Ft)", srt=90)
```

```
SZtop <- mean(BHdata$sz_top)
SZbot <- mean(BHdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Good Outcome for Pitcher Density Plot ##
good.dens <- ppp(gooddata$px_pitcher, gooddata$pz, c(-2,2), c(1,5))
good.dens.z <- density.ppp(good.dens)
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/DENSITY Good Outcome for Pitcher - ",
   pitches[1,35], ".pdf", sep=""))
par(pty="m", xpd=T)
plot(good.dens.z, col=hs.color(100), main="", axes=T, xlab="", ylab="")
text(0, 5.35, paste("Good Outcome for Pitcher vs.", gooddata[1,35]),
   cex=1.25, font=2)
text(0, 0.25, "Horizontal Distance, Centered at Home Plate \n
   (Pitcher's Perspective)")
text(-2.6, 3, "Distance Above Ground (Ft)", srt=90)
SZtop <- mean(gooddata$sz_top)
SZbot <- mean(gooddata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()


###################
## Contour Plots ##
###################

## Base Hit Contour Plot ##
type3.p <- as.points(BHdata$px_pitcher, BHdata$pz)
type3 <- bkde2D(type3.p, bandw, gridsize=c(grid.number, grid.number),
   truncate=TRUE, range.x=list(c(min(type3.p[,1], type3.p[,1])-
   1.5*bandw, max(type3.p[,1], type3.p[,1])+1.5*bandw),
   c(min(type3.p[,2], type3.p[,2])-1.5*bandw, max(type3.p[,2],
   type3.p[,2])+1.5*bandw)))

gridlocs.y <- rep(type3$x2[1], grid.number)
for (i in 2:grid.number)
  {gridlocs.y <- c(gridlocs.y, rep(type3$x2[i], grid.number))}
  gridlocs.x <- rep(type3$x1, grid.number)
gridlocs <- as.points(gridlocs.x, gridlocs.y)

par(pty="m")
plot.new()
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/CONTOUR Base Hits - ", pitches[1,35], ".pdf",
   sep=""))
contour(type3$x1, type3$x2, type3$fhat, nlevels=10,
   col=rev(heat.colors(12)), lwd=3, xlim=c(-2,2), ylim=c(1,4.5),
   xlab="", ylab="Distance Above Ground (Ft)")
title(paste("Base Hits for", gooddata[1,35]))
```

```
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
SZtop <- mean(BHdata$sz_top)
SZbot <- mean(BHdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Swing Contour Plot ##
swing.p <- as.points(swingdata$px_pitcher, swingdata$pz)
swings <- bkde2D(swing.p, bandw, gridsize=c(grid.number, grid.number),
   truncate=TRUE, range.x=list(c(min(swing.p[,1],
   swing.p[,1])-1.5*bandw, max(swing.p[,1], swing.p[,1])+1.5*bandw),
  c(min(swing.p[,2], swing.p[,2])-1.5*bandw, max(swing.p[,2],
   swing.p[,2])+1.5*bandw)))

gridlocs.y <- rep(swings$x2[1], grid.number)
for (i in 2:grid.number)
  {gridlocs.y <- c(gridlocs.y, rep(swings$x2[i], grid.number))}
  gridlocs.x <- rep(swings$x1, grid.number)
gridlocs <- as.points(gridlocs.x, gridlocs.y)

par(pty="m")
plot.new()
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/CONTOUR Swings - ", pitches[1,35], ".pdf",
   sep=""))
contour(swings$x1, swings$x2, swings$fhat, nlevels=8,
   col=rev(heat.colors(9)), lwd=3, xlim=c(-2,2), ylim=c(1,4.5),
   xlab="", ylab="Distance Above Ground (Ft)")
title(paste("Swings for", gooddata[1,35]))
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
SZtop <- mean(swingdata$sz_top)
SZbot <- mean(swingdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## BIP Contour Plot ##
BIP.p <- as.points(BIPdata$px_pitcher, BIPdata$pz)
BIP <- bkde2D(BIP.p, bandw, gridsize=c(grid.number, grid.number),
   truncate=TRUE,
  range.x=list(c(min(type3.p[,1], BIP.p[,1])-1.5*bandw,
   max(type3.p[,1], BIP.p[,1])+1.5*bandw),
   c(min(BIP.p[,2], BIP.p[,2])-1.5*bandw, max(BIP.p[,2],
   BIP.p[,2])+1.5*bandw)))

gridlocs.y <- rep(BIP$x2[1], grid.number)
for (i in 2:grid.number)
  {gridlocs.y <- c(gridlocs.y, rep(BIP$x2[i], grid.number))}
  gridlocs.x <- rep(BIP$x1, grid.number)
gridlocs <- as.points(gridlocs.x, gridlocs.y)
```

```
par(pty="m")
plot.new()
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/CONTOUR BIP - ", pitches[1,35], ".pdf",
   sep=""))
contour(BIP$x1, BIP$x2, BIP$fhat, nlevels=10, col=rev(heat.colors(10)),
   lwd=3, xlim=c(-2,2), ylim=c(1,5), xlab="", ylab="Distance Above
   Ground (Ft)")
title(paste("BIP for", gooddata[1,35]))
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
SZtop <- mean(BIPdata$sz_top)
SZbot <- mean(BIPdata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

## Good Outcome for Pitcher Contour Plot ##
good.p <- as.points(gooddata$px_pitcher, gooddata$pz)
good <- bkde2D(good.p, bandw, gridsize=c(grid.number, grid.number),
   truncate=TRUE, range.x=list(c(min(good.p[,1], good.p[,1])-1.5*bandw,
   max(good.p[,1], type3.p[,1])+1.5*bandw),
   c(min(good.p[,2], good.p[,2])-1.5*bandw, max(good.p[,2],
   good.p[,2])+1.5*bandw)))

gridlocs.y <- rep(good$x2[1], grid.number)
for (i in 2:grid.number)
  {gridlocs.y <- c(gridlocs.y, rep(good$x2[i], grid.number))}
  gridlocs.x <- rep(good$x1, grid.number)
gridlocs <- as.points(gridlocs.x, gridlocs.y)

par(pty="m")
plot.new()
pdf(paste("C:/Documents and Settings/Hillary Superak/My Documents/Grad
   School/THESIS/graphs/CONTOUR Good Outcome for Pitcher vs. ",
   pitches[1,35], ".pdf", sep=""))
contour(good$x1, good$x2, good$fhat, nlevels=8,
   col=rev(heat.colors(8)), lwd=3, xlim=c(-2,2), ylim=c(1,5), xlab="",
   ylab="Distance Above Ground (Ft)")
title(paste("Good Outcome for Pitcher vs.", gooddata[1,35]))
mtext("Horizontal Distance, Centered at Home Plate \n (Pitcher's
   Perspective)", SOUTH<-1, line=4)
SZtop <- mean(gooddata$sz_top)
SZbot <- mean(gooddata$sz_bot)
segments(-0.833, SZbot, -0.833, SZtop, lty=1, lwd=2, col='black')
segments(0.833, SZbot, 0.833, SZtop, lty=1, lwd=2, col='black')
segments(-0.833, SZbot, 0.833, SZbot, lty=1, lwd=2, col='black')
segments(-0.833, SZtop, 0.833, SZtop, lty=1, lwd=2, col='black')
dev.off()

}
```

WORKS CITED

Block, D. (2006). *Baseball Before We Knew It: A Search for the Roots of the Game.* Winnipeg, Manitoba: Bison Books.

Boudway, I. (2011, March 31). Baseball Set for Data Deluge as Player Monitoring Goes Hi-Tech. *Bloomberg Businessweek.*

Cuellar, D. (2008, October 17). *Economic Impact of World Series – Philadelphia News*. Retrieved April 11, 2011 from ABC Action News: http://abclocal.go.com/wpvi/story?section=news/sports&id=6454404

Delaney, T., & Madigan, T., (2009). *The Sociology of Sports: An Introduction*. Jefferson, NC: McFarland Publishing.

Dickson, P. (2007). *The Joy of Keeping Score: How Scoring the Game Has Influenced and Enhanced the History of Baseball.* New York: Walker & Company.

ESPN. (2011). *2010 MLB Team and Player Stats*. Retrieved March 30, 2011, from ESPN MLB: http://espn.go.com/mlb/statistics/_/year/2010

Fast, M. (2010, June 23). *Four Plus Pitches*. Retrieved April 4, 2011, from The Hardball Times: http://www.hardballtimes.com/main/article/four-plus-pitches/

François, R. (2006, April 11). *hexbin pie*. Retrieved March 10, 2011, from R Graph Gallery: http://addictedtor.free.fr/graphiques/RGraphGallery.php?graph=143

Kalk, J. (2007, August 31). *More fun with pitchFX*. Retrieved March 30, 2011, from from small ball to the long ball: http://www.baseball.bornbybits.com/blog/2007/08/more-fun-with-pitchfx.html

Lewis, M. (2004). *Moneyball: The Art of Winning an Unfair Game*. New York: W. W. Norton & Company.

Nathan, A. M. (2008, February 8). *A Statistical Study of PITCHf/x Pitched Baseball Trajectories*. Retrieved March 8, 2011, from Department of Physics, University of Illinois: http://webusers.npl.illinois.edu/~a-nathan/pob/MCAnalysis.pdf

National Baseball Hall of Fame and Museum. (2010). *Henry Chadwick*. Retrieved April 4, 2011, from Hall of Famers: http://baseballhall.org/hof/chadwick-henry

Quote Garden. (2010). *Quotations About Baseball*. Retrieved April 4, 2011, from Quote Garden: http://www.quotegarden.com/baseball.html

Society for American Baseball Research. (2011). *The SABR Story*. Retrieved April 4, 2011, from About SABR: http://sabr.org/about

Sports Reference LLC. (2011). *Baseball Encyclopedia of MLB Players.* Retrieved April 10, 2011, from Baseball-Reference.com: http://www.baseball-reference.com/players/

Tygiel, J. (2001). *Past Time: Baseball as History*. New York: Oxford University Press.

White, P. (2006, April 3). Art of Advance Scouting Advances. *USA Today*.