

CREATING A HITTER'S APPROACH:  
ANALYZING AT-BAT  
DATA

by

Dominic C. Ramos, BBA

A thesis submitted to the Graduate Council of  
Texas State University in partial fulfillment  
of the requirements for the degree of  
Master of Business Administration  
with a Major in Business  
December 2017

Committee Members:

Francis A. Méndez Mediavilla, Chair

David Wierschem

Tahir Ekin

**COPYRIGHT**

by

Dominic C. Ramos

2017

## **FAIR USE AND AUTHOR'S PERMISSION STATEMENT**

### **Fair Use**

This work is protected by the Copyright Laws of the United States (Public Law 94-553, section 107). Consistent with fair use as defined in the Copyright Laws, brief quotations from this material are allowed with proper acknowledgement. Use of this material for financial gain without the author's express written permission is not allowed.

### **Duplication Permission**

As the copyright holder of this work I, Dominic Ramos, authorize duplication of this work, in whole or in part, for educational or scholarly purposes only.

## **DEDICATION**

I dedicate this work to my loving wife and young family in hopes that they will always believe in making the desires of their hearts come true, and to ballplayers out there looking to understand themselves and the game of baseball more completely.

## **ACKNOWLEDGEMENTS**

I would like to thank my mother and father for supporting my efforts and decisions throughout my baseball career. They respected my decision to continue playing after being released from the Red Sox and found a way to support me in every circumstance. Thank you to all the coaches and players I have had a chance to compete with and against. You all have motivated me to become greater each day and I have learned from you all. I would like to thank the host families and administrators throughout the country and world who brought me in and cared for me throughout my ten-year playing career. Also, thanks to Matt Mader and his team for making it possible for me and other players to record at-bat data on an easy online platform. I would also like to thank Dr. Mendez and Dr. Wierschem for leading by example through the standard of excellence they hold themselves up to each day. They have worked hard to explain clearly the concepts they teach and push the students in a loving way to become successful in the classroom. All the MBA professors have taught with compassion and a true interest in each student here at Texas State. My classmates have been understanding and easy to work with the last two years. My wife and daughter have sacrificed time and energy to help me continue pursuing this paper, thank you. Lastly, I would like to thank God, our Father, for giving me the gift of baseball, the experiences, and the heart for understanding the game. I am happy to share the gift of baseball to the world. I could not have done this work without the loving grace Jesus gives so freely. Thank you all.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS .....	v
LIST OF TABLES .....	viii
LIST OF FIGURES .....	ix
LIST OF ABBREVIATIONS .....	x
ABSTRACT .....	xii
CHAPTER	
I. INTRODUCTION .....	1
Pitch Tracking Tools .....	1
Predictors .....	2
Hitter's Perspective .....	4
Research Hypothesis .....	7
II. METHODOLOGY .....	8
Data Collection .....	8
Data Discrepancies .....	10
Strike Zone .....	11
Analysis .....	12
Contour Maps .....	12
Clustering .....	13
Gaussian Mixture Model Cluster Method .....	13

Scree Plot .....	15
Multinomial Logistic Regression.....	16
III. RESULTS .....	18
Contour Maps.....	18
Clusters .....	21
Regression.....	24
Validation.....	28
IV. DISCUSSION.....	29
Benefits of Contour Maps.....	29
Improving Approach using GMM .....	30
Importance of Regression Analysis .....	31
Overall Model Performance.....	33
Improving Model .....	34
Other's Models.....	35
Collection and Implementation.....	36
Conclusion .....	37
APPENDIX SECTION .....	38
LITERATURE CITED .....	41

## LIST OF TABLES

Table	Page
1. Online At-Bat Database Variables.....	8
2. Pitch Result Number of Pitches .....	19
3. Clusters by Pitch Result Percentage .....	22
4. Cluster 1 Probability .....	24
5. Cluster 3 Probability .....	25
6. Cluster 4 Probability .....	26
7. Cluster 5 Probability .....	27
8. Confusion Matrix .....	28



## LIST OF FIGURES

Figure	Page
1. Ted Williams Heat Map.....	5
2. Eric Thames Heat Map .....	6
3. Online At- Bat Strike-Zone Platform.....	12
4. Scree Plot of Clusters BIC .....	16
5. Contour Map of All Pitches .....	18
6. Contour Maps of Pitch Result.....	19
7. Contour Map with 5 Clusters.....	21

## LIST OF ABBREVIATIONS

Abbreviation	Description
PA	<i>Plate Appearance</i> - Each time a player completes a turn batting, regardless of the result.
AB	<i>At-Bat</i> - An official at-bat comes when a batter reaches base via fielder's choice, hit or an error or when a batter is put out on a non-sacrifice.
HZ	<i>Hitting-Zone</i> - The area around Homeplate where the ball is thrown including balls in the strike zone and balls outside of the strike zone.
SZ	<i>Strike-Zone</i> - Pitches thrown to a hitter that pass over the plate and are beneath the hitter's midpoint between a batter's shoulders and the top of the uniform pants and a point just below the kneecap.
H	<i>Hit</i> - When a batter strikes the baseball into fair territory and reaches base without doing so via an error or a fielder's choice.
BB	<i>Walk</i> - When four pitches are thrown out of the strike zone, none of which are swung at by the hitter. After refraining from swinging at four pitches out of the zone, the batter is awarded first base.
FC	<i>Fielder's Choice</i> - The act of a fielder who handles a fair grounder and, instead of throwing to first base to put out the batter-runner, throws to another base in an attempt to put out a preceding runner.
SAC	<i>Sacrifice</i> - When a hitter bunts to advance a runner or when a batter hits a fly-ball out to the outfield or foul territory that allows the runner to score.
HR	<i>Homerun</i> - When a batter hits a fair ball and scores on the play without being put out or without the benefit of an error. In almost every instance of a home run, a batter hits the ball in the air over the outfield fence in fair territory.
AVG	<i>Batting Average</i> - One of the oldest and most universal tools to measure a hitter's success at the plate, batting average is determined by dividing a player's hits by his total at-bats for a number between zero and one. In recent years, the league-wide batting average has typically hovered around .260.

O-Swing	<i>Out-of-Zone Swing Rate</i> - The percentage of swings at pitches outside the strike zone divided by the total number of pitches outside of the strike zone.
K, SO	<i>Strike Out</i> - When a pitcher throws any combination of three swinging or looking strikes to a hitter.
K%	<i>Strike Out Rate</i> - The frequency with which a pitcher strikes out hitters, as determined by total strikeouts divided by total batters faced.
OPS	<i>On-Base Plus Slugging</i> - OPS adds on-base percentage and slugging percentage to get one number that unites the two.
ERA	<i>Earned Run Average</i> - The number of earned runs a pitcher allows per nine innings – with earned runs being any runs that scored without the aid of an error or a passed ball.
SLG	<i>Slugging Percentage</i> - The total number of bases a player records per at-bat after a hit.

## **ABSTRACT**

In this paper, we begin the process of creating an approach for a hitter in baseball dependent on predicting the pitch location. Predicting pitch location has not been introduced in academics up to this point. Most predictions have been done on pitch type or on the decision of the hitter to swing or wait on a pitch. In this study, we find a way to help the hitter create five locations in the hitting-zone to look for pitches to cross the plate and how to maximize the hitter's chance of predicting correctly in different scenarios of a game. Creating an approach dependent on predicting the pitch location was found by using contour maps, the Gaussian mixture model to create five areas for the hitter to set as approaches, and a multinomial logistic regression to explain when to change from one location to another during different situations in a game. The hitter learns to create approaches to handle at-bats when facing a right-handed or left-handed pitcher, when the hitter has strikes on the count, when the game increases in innings played, when outs increase, and when baserunners are on second and third. The future is bright for predicting pitch location. Teams who can implement these strategies will have an advantage over teams who do not.

## I. INTRODUCTION

A good hitter in baseball must have two things, good mechanics, and a good approach. The life of a hitter is mostly spent learning the mechanics of their swing, but not as much time is devoted to developing an approach to maximize the hitter's chances at having successful at-bats. In this paper, we will move away from the mechanics of hitting and talk about creating an approach on how to predict pitch location. Predicting pitch location has not been introduced in academics up to this point, most prediction has been done on pitch type or on the decision whether to swing or wait on a pitch.

We have analyzed strike zone data of a single player, *the batter*, who collected strike-zone data over the course of five seasons of independent baseball. We will use this strike-zone data to gain a better understanding of the pitch selection and the results of the hitter's at-bats per pitch, eventually coming up with a method to predict pitches. By minimizing the hitting-zone to five locations we help the hitter create an approach to maximize the hitter's chance of predicting correctly in different scenarios of a game. Creating an approach during different scenarios in a game give the hitter the best chance to find success.

In the following section, we will discuss the literature that is relevant in terms of pitch tracking tools, variables used in prediction, and important factors to help understand strategy from the hitter's perspective.

### **Pitch Tracking Tools**

In 2006, Major League Baseball introduced Pitchf/x, a system to capture and keep track of pitch speed, pitch trajectory, pitch type, and pitch location for every pitch thrown in a major league baseball game (Albert, 2010). The data became available not only to major league baseball teams but also to the public. The pitch-by-pitch dataset contains

general information about the pitcher and batter to identify the players in all circumstances of a game and the outcome of the plate appearance, including pre-pitch and post-pitch count (Albert, 2010). In 2015, Major League Baseball upgraded the pitch tracking system from Pitchf/x to Statcast. Statcast captures the same strike zone data that Pitchf/x captured, but records much more data, including the trajectory and exit velocity of the ball off the bat and the ability to track players located on the field (Law, 2016).

Because Pitchf/x data has become public there have been many articles published on many different topics ranging from showing users how to access and manipulate the data (Fast, 2010) to understanding pitching strategies (Cox, 2015), to predicting pitch type (Hamilton, 2014), and even grading umpire accuracy (Flanagan, 2016).

We must keep in mind that the data being captured is not perfect but does give the user enough clean information to gain new perspectives on the game of baseball. When recording the characteristics of each pitch, Pitchf/x is correct much more often than it is wrong, but it is far from perfect (Fast, 2010). The same is true for Statcast, as the implementation of the new system had difficulty distinguishing all the movements on the field, as it was attempting to track pitches (Law, 2016). Each of the variables recorded in at-bat data have been used differently in studies to predict the outcome of a pitch or performance of a player.

## **Predictors**

When predicting pitches, it is important to note the difference between pitch classification and pitch prediction. Pitch classification uses post-pitch information about a pitch to determine which type of pitch is thrown, whereas pitch prediction uses pre-pitch variables (Hamilton, 2014). In a study to predict pitch type, Hamilton uses around 80

pitch features to predict pitch type with the variable *pitcher tendency* being grouped into 4 to 12 features (Hamilton, 2014). Hamilton also uses the count as a basis for decisive strategy (Hamilton, 2014). In this study, the model achieves accuracy as high as 90 percent. They go on to say one could consider the prediction problem of determining where the pitch will be thrown (Hamilton, 2014).

Predicting the next pitch type has also been studied using machine learning techniques. In this study, the most useful predictors for pitch type were the previous batter, the count, the previous pitch, and the score (Ganishapillai and Guttag, 2012). It was found that pitchers are more predictable at less favorable counts to the pitcher, where the hitter has more balls in their advantage as opposed to strikes in the current count (Ganishapillai and Guttag, 2012). Therefore, a hitter could predict the pitch type with higher accuracy when the count is in the hitter's favor.

In a study to determine the streakiness of hitters, Albright in 2009 and Quintana in 2012 both used the number of outs at the time of plate appearance, the number of runners on base, the game location at batter's home field, and the earned run average of the opposing pitcher in logistic models to explain the success of the hitter (Quintana 2012).

In a more recent study on the matching principle of the strike zone, the position of the runners on the base paths may also be important in influencing pitching strategy (Cox, 2017).

The variables used to help predict an outcome through the different studies mentioned above are, pitcher tendency, number of outs, the number of runners on base, game location, earned run average of the opposing pitcher, the previous batter, the count, the previous pitch, the score, and the position of the runners on the base paths. Pre-pitch

variables strengthen the likelihood of pitch prediction in studies, but a hitter focuses on different aspects of the game to find success.

### **Hitter's Perspective**

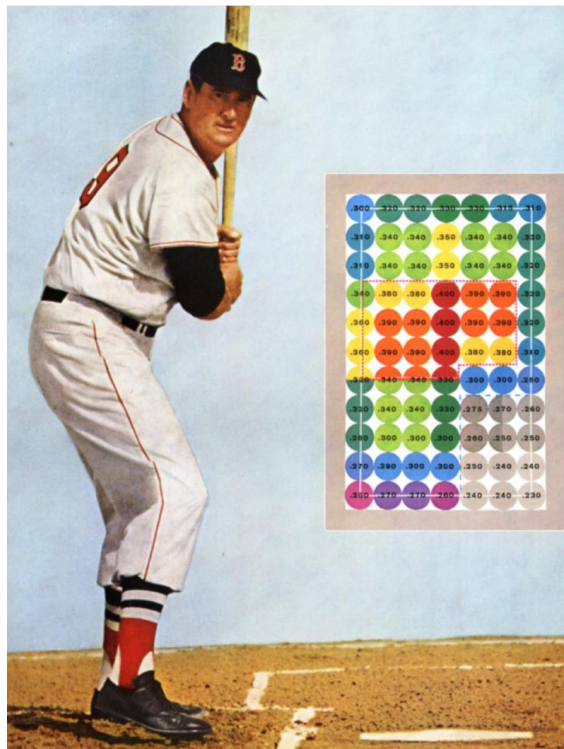
A major league average 86 mile-per-hour fastball takes 450 milliseconds to reach the batter. Most batters take about 200 milliseconds to swing the bat. Therefore, the batter has around 250 milliseconds to decide if and where to swing his bat (Müller, 2016). Because of this limited amount of time to decide where to swing the bat, hitters anticipate what type of pitch might be thrown and predict the eventual location. It has been found that expert hitters track the ball from the pitcher's release point up to a third of the ball's flight to home plate, where nonexpert hitters are not as good at picking up the ball at the release point (Faddie, 2006).

Hitters are not only faced with the challenge of limited time to react to a pitch but are now facing new bullpen strategies from opposing teams. Today hitters are seeing more relief pitchers on average and have a lower OPS+ when facing relief pitchers compared to starting pitchers (Harrison and Salmon, 2017). As hitters tend to do better the third time facing a pitcher in a game, this study suggests making it more difficult for hitters to gain an advantage after each at-bat by changing pitchers more often (Harrison and Salmon, 2017).

As teams begin to change pitchers more often hitters can still use a strategy that was introduced by Ted Williams more than fifty years ago. Ted Williams in his book *The Science of Hitting* explains his *selectively aggressive* approach which helped him become the last man to have a batting average over .400. In this book, Ted Williams shows a picture of the strike-zone with balls of different colors with the resulting batting average



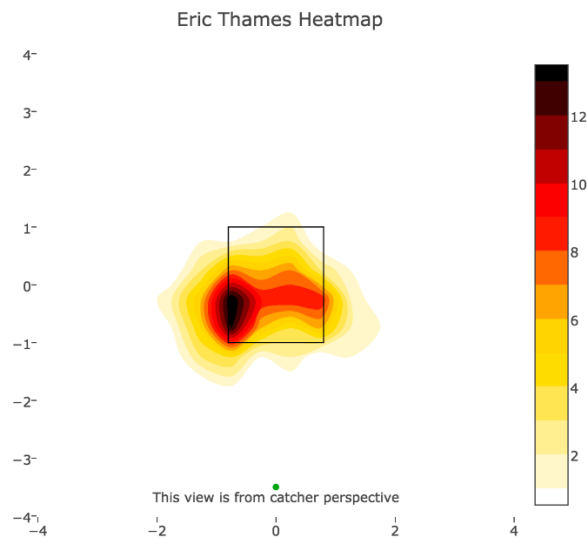
for each location in the strike zone, seen in Figure 1. Ted explains, “My first rule of hitting was to get a good ball to hit.” He goes on to explain a hitter should choose pitches in the “happy zone” those balls with the highest batting average, and go after lower percentage pitches only when the situation demands it.” (Williams and Underwood, 1970). This is one of the first visual analysis of the strike zone which major league teams today can create in a matter of seconds using strike-zone data.



**Figure 1. Ted Williams Heat Map** – color strike zone with batting averages for each pitch location using red, orange, yellow, green, blue, indigo, violet, and grey in descending order to describe the batting average of Ted Williams (Williams & Underwood, 1970).

In 2017, a visual heat map of Eric Thames out-of-zone swing rate was published in an article to show his ability to swing at pitches located in the strike-zone. Thames trained himself using visualization, imagining different pitch types in a certain location to make him one of the most consistent hitters at swinging at pitches in the strike zone early in the 2017 season (Shawchik, 2017). Eric Thames talks about the power of narrowing

his focus to a ‘three-and-a-half-inch zone’ after researching Barry Bonds approach (Shawchik, 2017). He went from striking out at 30% rate in 2012, having a 35.6% out-of-zone swing rate, to having the lowest out-of-zone swing rate in major league baseball in the first month of the 2017 baseball season at 17.6% (Shawchik, 2017). Having a strategy focused on the location of the pitch has allowed Thames to eliminate looking for pitch type and react to any pitch that comes into his “happy zone”.



**Figure 2. Eric Thames Heat Map** – out-of-zone swing rate displayed in a descending black, red, orange, and yellow scale to indicate the frequency of swings by Eric Thames from April 3, 2017, through May 4, 2017 (Shawchik, 2017).

As the hitter limits their strategy to pitch location the hitter now has the decision to swing or wait. In a study conducted to determine if the hitter should swing or take a pitch, batters do not seem to take 2-0 (2 balls and 0 strikes) and especially 3-1 with enough frequency (Bickel, 2009). It is suggested hitters take pitches in advantage counts, 2-0, 3-0, and 3-1 to increase the probability of reaching base instead of swinging to get a hit (Bickel, 2009). If the team needs a hit, the hitter should not take on any counts, especially 3-0 and 3-1 (Bickel, 2009). In a more recent study on whether to swing or wait

dependent on the pitch resulting in a ball or strike the findings suggest hitters should use an unpredictable mixed strategy (Che and Kim, 2016). The results show that professional baseball players do not fully optimize their strategies and have predictable patterns in their mixed strategies (Chloe and Kim, 2016).

The study done by Choo and Kim on whether to swing or wait also shows the difficulty in predicting pitch location stating the strategy set will become infinite (Choe and Kim, 2016). Some of the best hitters in the baseball choose to look for pitch location, but studies find it difficult to predict pitch location. This idea will lead us to our research in predicting pitch location.

### **Research Hypothesis**

Based on the available at-bat data with variables used to help predict outcomes in baseball, and the improved results of a hitter's strategy predicting pitch location, we believe using pre-pitch variables will help us start the discussion on predicting pitch location. In this study, it is conjectured that the location of a pitch depends on the handedness of the pitcher, inning, number of outs, number and location of baserunners, the number of balls on the count, and the number of strikes on the count. The belief is that these variables are related to the area of the strike-zone where any given pitch will be located, ultimately leading the hitter to have a better ability to predict where the pitch will end up in the strike zone.

## II. METHODOLOGY

### Data Collection

The data used in the at-bat analysis is cumulative input from a single player, over the course of five seasons. The player input the data into an online platform after each game, making for routine and easy entry. The experimental subject is five feet ten inches tall, 180 pounds, right-handed, and was 27 to 31 years old in the seasons during data collection. Over six hundred plate appearances and over two thousand pitches were recorded and analyzed. Table 1 below shows the different variable inputs the player could choose to enter into the database.

**Table 1. Online At-Bat Database Variables**

<b>Title</b>	<b>Description</b>
Player	Person at bat
Date	Game date
DayNight	Tells whether the game was played during the day or night
Opponent	Opposing team
PitcherName	Name of person pitching the ball
PitcherNumber	Number of person pitching the ball
PitcherR/L	Tells whether the pitcher is right or left-handed
PitcherStyle	Tells what style the pitcher uses to pitch the ball. Measured as over the top, side-arm, 3 quarters, underneath.
Inning	Tells what inning the pitch was thrown
Outs	Tells the number of outs the opposing team has during that inning
Baserunners	Tells which bases have runners
TotalBaserunners	Tells total number of runners on base
Scoring Position	Tells whether there are runners in scoring position (2nd or 3rd base). Runners in scoring position = 0. No runners in scoring position = 1.
Reach	Tells if the hitter reached or did not reach base

**Table 1. Continued. Online At-Bat Database Variables**

Result	Tells what was the result of the whole at-bat. 1B, 2B, 3B, HR (home run), BB (walk), K (strikeout), K Looking (strikeout looking), SAC (sacrifice bunt or sacrifice fly), E (error made by opponent), HBP (hit by pitch), fly-out, ground-out, line-out, other-reached, other-out, did not finish.
DefenseRating	Tells how difficult it was to field the ball. Measured as none, easy out, hard out, error, or legit.
Quality	Tells if the at-bat was a quality at-bat or not. Measured as yes or no.
RBI	Number of runs batted in on this hit
PitchNumber	Tells which pitch of the at-bat this pitch is
PitchHeight	This measure the height of the pitch: the higher the number, the lower the pitch. Min=0 and Max=20
PitchWidth	This measures how far away the pitch is from the batter from right to left: the higher the number, the farther away the pitch. Min=0 and Max=15
PitchResult	6 options: ball, watched strike, swinging strike, strike foul, base hit, out. These are noted for each pitch during a single at-bat. I.e. Pitch1_Result.
PitchType	10 options: 2 seam FB, 4 seam FB, change, split, slider, curve, cutter, knuckle, other, unknown. These are noted for each pitch during a single at-bat. I.e. Pitch2_Type.
Balls	Tells how many balls were on the count prior to this pitch
Strikes	Tells how many strikes were on the count prior to this pitch
Hit Height	Tells where the ball was hit on the field from 0-25: the higher the number, the shorter the ball was hit.
Hit Width	Tells where the ball was hit on the field from 0-35: the higher the number, the further the ball was hit to the right side of the field.
Hit_Type	6 options: fly out, line out, ground out, bloop hit, ground hit, line hit
Thoughts	Hitter's thoughts on the at-bat

Pre-pitch variables are those which are recorded before the at-bat, where post-pitch variables are those which are recorded after the pitch is thrown. Pre-pitch variables in this data set are Player, Date, Day/Night, Opponent, Pitcher Number, PitcherR/L, Pitcher Style, Inning, Outs, Baserunners, Total Baserunners, Scoring Position, Pitch

Number, Balls, and Strikes. Post-pitch variables are Reach, Result, Defense Rating, Quality, RBI, Pitch Height, Pitch Width, Pitch Result, Pitch Type, Hit Height, Hit Width, Hit Type, and Thoughts.

The main difference in this data from Pitchf/x data or Statcast data is that the player himself inputs the data, whereas in major league baseball the computer system and others input at-bat data for each pitch, play, and player. Inputting data manually has been attempted before, but the data was found unreliable (Fast, 2016).

### **Data Discrepancies**

For this research, we assume that each player has recorded the pitch location and sequence of pitches to the best of their ability. We have extracted incomplete recordings of at-bats to clean the data. The player inputting data after an at-bat could incorrectly label the area in the strike-zone where the ball passes through the hitting area and the sequence of pitch type, pitch result, and the result of the at-bat. The two leagues the player played in during the time of collecting the data has online statistics with the pitch result for each pitch and at-bat results in all games but does not keep track of pitch location. Therefore, pitch result and at-bat result can be verified, whereas pitch location and pitch type have no way of being tested for accuracy.

In each of these at-bats, the player used a wooden bat. The size length and shape of each bat is unique to each player and can change during the season. For most of the at-bats in this data set the player used a 33-inch bat with a weight of 30 ounces. The hitter used a bat type with a medium to small barrel and a thin handle. For the majority of 2012, 2013, and 2014 seasons the player used the model DB-159, made by DBat.

## **Strike Zone**

The strike zone is defined, in major league baseball, by height and width. The lowest point in the strike zone is the bottom of the knees, where the top of the strike zone is the midpoint between the shoulders and the waist. Each strike zone varies for each player because of the difference in height. The width is the same for each player, which is simply the width of home plate. If the baseball crosses the hitting zone over home plate and inside the top and bottom boundaries of the strike zone, the pitch is considered a strike. If the ball is thrown outside of the strike zone boundary, the pitch is considered a ball.

Home plate measures 17 inches across, which fits five balls completely, almost six baseballs, but up to seven given only part of the ball needs to cross over the plate (Williams and Underwood, 1970). The height of the player in this study is five feet ten inches tall, so his strike zone is approximately 27 inches high, from the mid-point between his shoulders and waist down to the hollow of his knees.

The strike zone used to collect the data is a grid of 20x15 with the top left box counted as 0x0. Each pitch is input by the user by clicking the box they believe describes the location of the pitch. Each click in the located box represents a different color for the type of pitch result (ball, watched strike, swinging strike, hit foul, hit reached, or hit out). The drop-down menu allows the user to change the pitch type (2 seam FB, 4 seam FB, curveball, slider, change-up, cutter, and other). Please reference Figure 1 below. The hitter records each pitch for the entire at-bat. If a pitch is in the same quadrant as a previous pitch in the at-bat the user chooses a box closest to the location of the pitch to

indicate the pitch location. When the final pitch is recorded the player submits the at-bat and the at-bat is recorded in the data set.

**Figure 3. Online At-Bat Strike-Zone Platform** –visual display of the strike-zone platform to input pitch data for each pitch in an at-bat.

## Analysis

To manipulate the data into predicting pitch location we first chose to create a contour map displaying the frequency of pitches. Once the visual display was captured, we used the Gaussian Mixture Model to create five clusters within the hitting zone to use as the best five locations for the hitter to look for the ball to cross the hitting-zone. After the five clusters were created, we ran a multinomial logistic regression using pre-pitch variables to predict when pitches were thrown in the five different clusters.

## Contour Maps

The first step in analyzing the data was to create a contour map of the pitch height and pitch width for all pitches thrown in the data set. A contour map is a three-dimensional figure of all pitches thrown in the hitting area displayed in a two-dimensional varying grey scale. The lighter the grey, the less frequent pitches were thrown in that part of the hitting area, and the darker the grey, the more frequent pitches



were thrown in that part of the hitting area. Using a contour map allows the user to easily understand where pitchers tend to throw hitters in the hitting area. These visual displays allow statisticians to analyze individual pitches based on information not contained in box scores or other statistics (Superak 2009).

After analyzing the contour map for all pitches, we then can filter using pitch result or pitch location. Stratifying by pitch result and location will come in handy once we start analyzing the material.

## **Clustering**

Clustering consists of grouping data records or cases in a set into subsets called clusters. The grouping assigns data records to the same cluster that are very similar among themselves while being very dissimilar to data points in other clusters.

Hierarchical and k-means clustering methods are recommended when clusters are well separated. However, in this study, clusters overlap. Therefore, assigning each record to one cluster is problematic because there are pitches from several clusters sharing the same space. A method called Gaussian mixtures is recommended, instead of k-means clustering for an accurate estimate of the total population in each group. The Gaussian mixtures method is based on membership probabilities, instead of the arbitrary cluster assignments based on distances (Friedman, 2001).

## **Gaussian Mixture Model Cluster Method**

The Gaussian mixture model (GMM) is an iterative technique, which relies on an estimation method to characterize the cluster groups. Rather than classifying each row into a cluster, it estimates the probability that a row is in each cluster (McLachlan and Krishnan, 1997). A GMM is used to represent normally distributed subpopulations.

GMMs are parametrized by two types of values: the mixture components weights; and, the mean, variance or covariances for the components.

The Expectation Maximization (EM) algorithm is the commonly used technique to estimate the values for these parameters. Each iteration is a two-step procedure: 1) perform the expectation (i.e., estimation) of the component assignments for each data record, given the model parameters (i.e., weights, means, variance/covariance); and 2) choose to maximize the likelihood (i.e., maximize) (Dellaert, 2002; Dempster, 1977; Hartley, 1958). Since the maximum likelihood of the data strictly increases with each iteration, the result is guaranteed to approach a local maximum (Dellaert, 2002).

GMM can be used for “soft” or “hard” clustering. GMM clustering may be used in such a way that any given data record can be assigned to more than one cluster or “soft” clustering (Friedman, 2001). GMM does the clustering by assigning data records to the multivariate normal component that maximizes the component posterior probability, given the data (Friedman, 2001). For instance, given two probability density functions:

$$g_0(x)$$

and

$$g_1(x)$$

; the terms

$$\frac{g_0(x)}{g_0(x) + g_1(x)}$$

and

$$\frac{g_1(x)}{g_0(x) + g_1(x)}$$

, are called “responsibilities” for the corresponding clusters (i.e., 0 and 1) (Han, 2011).

Assigning a data record can be done in such a way that each of them is assigned to exactly one cluster or “hard” clustering (Friedman, 2001). In GMM the number of clusters, which specifies the number of components, must be specified prior to the fitting (Friedman, 2001)). For this paper, the GMM Cluster Method was applied to identify clusters of pitch locations using pitch height and pitch width.

### **Scree Plot**

First, we ran the analysis to find the best number of clusters to use in the analysis. Testing the possibility of three clusters through seven, using 300 iterations, we checked the variability of the clustering at each number of clusters using Bayesian Information Criteria (BIC). BIC is defined as  $BIC = n \ln(RSS/n) + k \ln(n)$ , where RSS is the residual sum of squares, and where k is the number of estimated parameters, clusters, in the model and n is the number of pitches in the data set. The lower the BIC the better the fit of the model. By creating a scree plot based on the BIC we found the appropriate number of clusters, using the elbow rule to determine five clusters as the most relevant number of clusters, shown in Figure 4.



**Figure 4. Scree Plot of Clusters BIC** - BIC for clusters 2 through 7, showing the elbow at 5 clusters.

Once the correct number of clusters was identified, we took the results for the five clusters and plotted the located means on the Contour Map of All Pitches. Once each pitch is identified by a cluster with even distributions of pitches, we were able to run the multinomial logistic regression.

### **Multinomial Logistic Regression**

We then wanted to know the predictability of a pitch location dependent on nominal and categorical pre-pitch variables of the game using one of the five clusters to describe each pitch. The multinomial regression is good for predicting outcomes using certain variables to explain the outcome of each cluster. In equation 1, a linear model is fit to each cluster using maximum likelihood. The regression shows the equation of the likelihood of a pitch determined by pre-pitch variables: PitcherR/L, baserunners, outs, inning, pitch number, balls, and strikes, where  $P_i$  is the probability of the pitch being in cluster 2.

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_P X_P \quad (1)$$

Once the regression results were obtained only the significant variables in each cluster were used to predict the likelihood of a pitch passing through any one of the clusters, which is calculated using equation 2. To find the exact probabilities of each pitch cluster using the values of each significant cluster variable (circumstances of the game) we used the following equation.

$$P(X) = \left( \frac{e^y}{1 + e^y} \right) \quad (2)$$

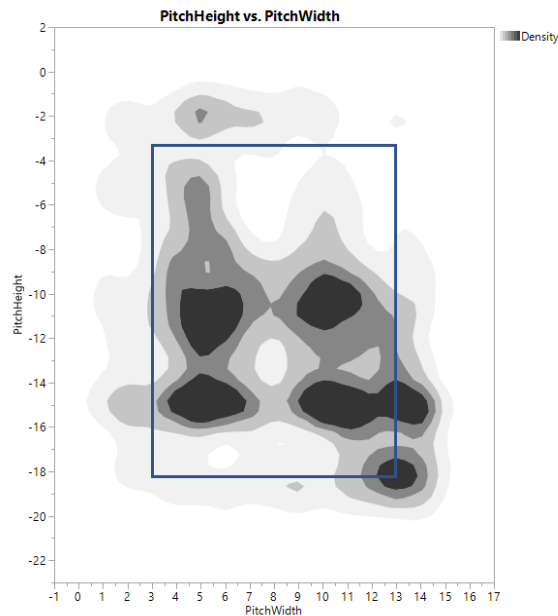
where y is

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_P X_P \quad (3)$$

### III. RESULTS

#### Contour Maps

After constructing the method in which we would analyze the data, the first visual result of the data was a contour map. The contour map in Figure 5 shows the distribution of all pitches in the data set. This view is from the umpire's perspective, seeing pitches as they came into the strike zone. The batter stood in the right-handed batter's box which is on the left side of the contour map. The dark rectangle inside the hitting zone represents the strike zone border. All pitches inside the strike zone are located inside the dark border and all pitches outside of the dark border are pitches thrown out of the strike zone.



**Figure 5. Contour Map of All Pitches** – visual display of frequency of all pitches in the hitting-zone and strike-zone described by shades of grey with the darkest grey indicating the most frequent area where pitches crossed.

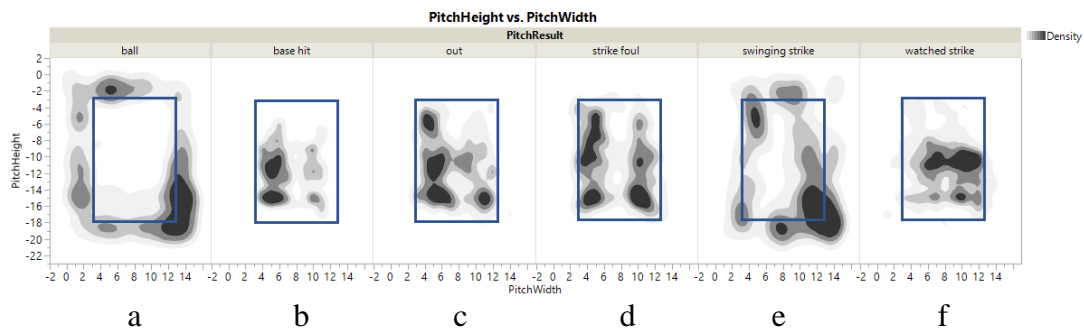
In the contour map there are five to seven areas of the strike zone where pitches were thrown most frequently which appear darkest in color. Most of the pitches look to be thrown in the middle and bottom half of the strike zone. The middle of the strike zone

looks to be avoided, where the inside and outside half of the strike-zone describe pitches equally. The pitches thrown inside tend to be higher in the strike-zone, whereas pitches thrown on the outside seem to be thrown lower. High frequencies on the bottom right corner seem to flow continuously from a strike to a ball, showing how often pitchers in this part of the strike-zone try to sneak a strike into the strike zone while also trying to get hitters to swing at pitches that are considered a ball.

After seeing the original contour map with all pitches in the strike zone the next step in analysis was to look at the contour map dependent on the result of the pitch (i.e., ball, base hit, out, strike foul, swinging strike, and watched strike), which can be seen in Figure 6. Table 2 gives the number of pitches described by each pitch result with the percentage of pitches in each pitch result.

**Table 2. Pitch Result Number of Pitches**

Pitch Result					
Ball	Base Hit	Out	Strike Foul	Swinging Strike	Watched Strike
629	168	299	339	179	401
31.22%	8.34%	14.84%	16.82%	8.88%	19.90%



**Figure 6. Contour Maps of Pitch Result** – visual display showing the frequency of pitches displayed in a varying grey scale dependent on pitch results of a ball, base hit, out, strike foul, swinging strike, or watched strike.

In the Figure 6a, all pitches that were called balls by the umpire are shown mostly outside of the strike zone. Though most of the pitches located around the dark bordered strike zone were called balls, we also see a few pitches inside the strike zone being called a ball.

The most important contour map in the figure below is the contour map of hits, seen in Figure 6b. We see the inside part of the plate and a little of the outside were the location of pitches where the at-bat resulted in a hit. This is important for the hitter to know because this helps create an area of the strike-zone to look for pitches. This can be referred to as a “hot zone” for a hitter, describing the area with most success.

In Figure 6c shows the contour map of outs, where pitches were thrown in the hitting-zone that resulted in an out. Most of these balls were on the inside part of the plate or low and away. When comparing the out map (6b) to the base hit map (6c) we see similar areas of the hitting-zone being worked. From these two maps, we see the area where the hitter has put the ball in play. More outs happen at higher and lower extremes within the strike zone compared to hits.

In Figure 6d, the map shows the fouled off balls by the hitter. The foul balls are usually in or away, not much in the middle of the plate, indicating the pitcher rarely throws in the middle of the plate or that the hitter does not miss hit the pitch in the middle of the plate.

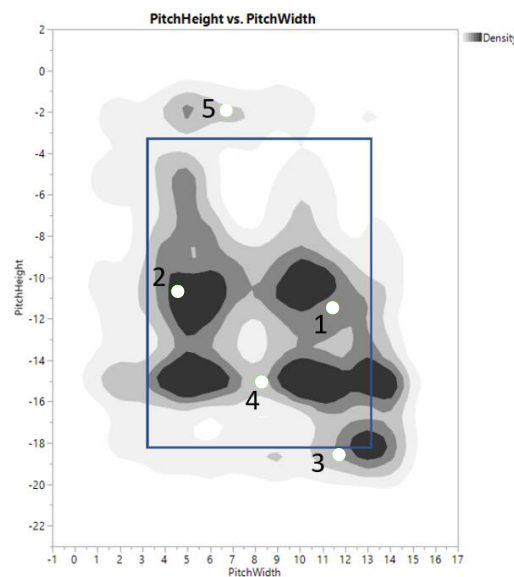
Another important map is the map of swinging strikes, seen in Figure 6e. Looking at the bottom right corner of the map, we see more swing and misses in this part of the strike-zone than any other area. Most of the swing and miss strikes are located on the outer limits of the strike-zone, sometimes even outside of the strike-zone, indicating a player swinging at a pitch that would be considered a ball.



Lastly, Figure 6e is a map displaying watched strikes, showing most pitches that are throw away and in the middle part of the strike zone to be the most watched strikes. We will see a possible explanation of why most of these good strikes were let go by the hitter once we run the regression results.

## Clusters

After running the results for five clusters using the GMM the five clusters are described by the number of pitches in each cluster and by the mean location within the hitting-zone. The results of the clusters can be seen in the Appendix under Cluster Results. In Figure 7, the five clusters are shown on the map by a white dot representing the mean for each cluster accompanied by a number describing which cluster it is, 1-5.



**Figure 7. Contour Map with 5 Clusters** – visual display of numbered cluster means located in the Contour Map of All Pitches

Immediately, we see clusters 3 and 5 are located outside of the strike-zone with cluster 5 located up and inside and cluster 3 located down and away. With these clusters being located outside of the strike-zone, the hitter will limit the amount of focus on these pitches and focus more on the pitches inside the strike-zone. Clusters 1, 2, and 4 are

located inside the strike- zone, mostly covering the bottom half of the strike-zone.

Clusters 1 and 2 are mid-height with cluster 2 covering the inside part of the strike zone and cluster 1 covering the outside part of the strike zone. Cluster 4 is located on the bottom of the strike zone and in the middle of the plate.

According to the results, most pitches are thrown into cluster 2 at 37.75% of the time, followed by cluster 1 at 30.99% (see Appendix 2). Cluster 4 is thrown into close to half as much as cluster 1 and 2 at 15.34%. These three clusters inside of the strike-zone account for 84.08% of all pitches. Cluster 3 accounts for 8.54% of pitches and pitches go into cluster 5 only 7.38%. If we consider clusters 1, 2, and 4 as locations for the hitter to look for pitches coming into the strike-zone, the hitter can limit the strike-zone to only three different locations most of the time, as opposed to covering locations over the entire strike zone.

Another helpful way to understand each cluster is by comparing the percentage of each pitch result to other clusters. Table 3 shows the clusters described by the percentage of each pitch result.

**Table 3. Clusters by Pitch Result Percentage**

	Normal Cluster				
	1	3	4	5	2
PitchResult	Column %	Column %	Column %	Column %	Column %
ball	21.96%	70.00%	32.45%	71.26%	19.51%
base hit	7.32%	0.00%	11.50%	0.00%	11.68%
out	14.14%	2.22%	15.63%	2.40%	21.02%
strike foul	18.97%	2.78%	16.52%	7.19%	20.88%
swinging strike	8.32%	22.78%	4.72%	14.97%	6.46%
watched strike	29.28%	2.22%	19.17%	4.19%	20.47%

In Table 3, we see clusters 3 and 5 described by balls 70.00% and 71.26% of the time, respectively. In these clusters, seven out of ten pitches were called balls. The highest percentage of swinging strikes also happen in clusters 3 and 5 at 22.78% in cluster 3 and 14.97% in cluster 5. Base hit percentage in cluster 3 and 5 are 0.00%. If the

pitch was thrown into cluster 3 or 5 the hitter had the highest chance of a positive outcome by taking the pitch, but if the hitter swung at a pitch in these two clusters the result ended in either a strike or an out, which shows the power of taking the pitches that cross the plate through clusters 3 and 5.

Cluster 2 and cluster 4 have the highest percentage of hits occurring as a result when compared to other clusters. The hitter got a hit 11.68% of the time when the pitch was in cluster 2 and 11.50% of the time when the pitch was in cluster 4. The highest percentage of outs happen in cluster 2 at 21.02% and cluster 4 at 15.63%. This shows the hitter puts the ball in play the most when pitches are thrown into cluster 2 or 4. The hitter is smart to swing at pitches in these two clusters if he is looking to get a hit or advance a runner on base.

Cluster 1 has the highest percentage of watched strikes at 29.28% when compared to the other four clusters. Pitches going into cluster 1 also result in a hit 7.32% of the time. Cluster 1 also has the highest percentage of strikes. By adding strike foul, swinging strike, and watched strike, the percentage of pitches resulting in a strike is 56.57%. Over half of the pitches thrown into cluster 1 resulted in a strike. If the percentage of outs is added to the percentage of strikes, cluster 1 has the highest percentage of negative results at 70.71%. Because these are negative results for the hitter, the pitcher had the highest likelihood of getting a hitter out if they threw the ball into cluster 1.

Now that the clusters located within the hitting-zone and the percentages of pitch result are understood, the next step in the process is the multinomial logistic regression to see if variables can help us distinguish when balls will pass into a certain cluster. The

goal of the regression is to find the best cluster to look for a pitch in different situations of a game.

## Regression

Considering each cluster, the pre-pitch variables used in the regression are PitcherR/L, baserunners, outs, inning, pitch number, balls, and strikes. After running the multinomial logistic regression, each cluster came back with a different combination of significant variables, which can be seen in Appendix 4. Significant variables for the whole model have a p-value less than 0.05. The p-value for Pitcher R/L is 0.00001, for Outs is 0.001, and for Baserunners is 0.003, making all three of these variables significant in helping predict a cluster. We are weighing all clusters against cluster 2 because this is the cluster the hitter has the highest percentage of hits. Through the regression, the hitter will learn when to look for pitches in cluster 2 or the other clusters.

When comparing cluster 1 to cluster 2, pitcherR/L is the only significant variable at 0.007 significance with a negative slope of -0.163. If the pitcher is left-handed the pitcher is less likely to throw the ball into cluster 1. Once the probability of the pitch occurring in this part of the plate was calculated for a lefty, there is a 45.24% likelihood that the pitcher will throw the pitch in cluster 1 as opposed to cluster 2. A right-handed pitcher is more likely to throw the ball into this cluster, at 57.81%. It is smarter for the hitter to look in cluster 2 when a left-handed pitcher is throwing and smarter to look in cluster 1 when a right-handed pitcher is throwing.

**Table 4. Cluster 1 Probability**

Cluster 1	
Lefty	Righty
45.24%	57.81%

This is the only cluster with only one significant prediction variable, showing it is the least predictable pitch in the strike-zone based on the variables used. It is also the only cluster that does not have a significant intercept, meaning when cluster 1 is compared to cluster 2 they are non-distinguishable. Cluster 1 is also the location where most pitches were watched strikes, seen in Figure 6f. This shows when the pitch is less predictable the hitter takes the pitch more often suggesting the hitter may have an innate sense in predicting the location of a pitch and an inability to predict correctly pitches coming into cluster 1.

Cluster 3 has significant p-values when compared to cluster 2 at the intercept at 0.006, pitcherR//L at 0.002, and strikes at 0.0146. The probability of the pitch being thrown into cluster 3 can be seen in Table 5.

**Table 5. Cluster 3 Probability**

Cluster 3		
Strikes	Lefty	Righty
0	20.17%	25.46%
1	36.22%	43.44%
2	56.17%	63.32%

The pitcher is less likely to throw the ball into cluster 3 when the pitcher is left-handed (Lefty), as the slope is -0.301, and more likely to be thrown into cluster 3 by a right-handed pitcher (Righty). The more strikes on the count the higher likelihood the pitch will be thrown into cluster 3 with a slope of 0.810. Cluster 3 is the only cluster described by strikes, showing an increase in the likelihood that a pitcher will throw the ball into cluster three as they increase their advantage in a count. The likelihood of the pitch being thrown in cluster 3 by a lefty with zero strikes on the count is 20.16%. The likelihood of the pitch being thrown by a lefty one strike on the count is 36.21%. The

likelihood of the pitch being thrown by a lefty with two strikes on the count is 56.07%. The likelihood of a righty throwing the ball into cluster 3 is even higher at 25.46% with zero strikes, 43.44% with one strike, and 63.32% with two strikes. In general, the probability of a pitcher throwing the ball into cluster 3 with two strikes is above fifty percent, therefore cluster 3 is the best cluster to look for a pitch when a hitter has 2 strikes on the count. When the hitter has less than two strikes it is smarter for the hitter to look for a pitch in cluster 2.

Cluster 4 has the most variables to describe the predictability with significant p-values at the intercept at 0.0025, pitcherR/L at 0.00, outs at 0.0001, and baserunners on second and third at 0.0057. The probability of a pitch crossing into this part of the strike-zone can be seen in table 6.

**Table 6. Cluster 4 Probability**

<b>Cluster 4</b>					
No BR			BR 2+3		
Outs	Lefty	Righty	Outs	Lefty	Righty
0	21.59%	28.29%	0	49.25%	58.18%
1	27.69%	35.43%	1	57.45%	65.93%
2	34.75%	43.29%	2	65.25%	72.91%

With no baserunners on second and third, the probability of a left-handed pitcher as well as a right-handed pitcher rises as outs increase in the inning. With zero outs the probability for a lefty is 21.59%, with one out 27.69%, with two outs 34.75%. With a righty on the mound, without runners on second and third, the probability rises to 28.29% with zero outs, 35.43% with one out, and 43.29% with two outs. With runners on second and third the probability of the pitch being thrown into the cluster 4 rises even higher. With a left-handed pitcher on the mound and zero outs the probability is 49.25%, with one out 57.44%, and with two outs 65.93%. With a right-handed pitcher on the mound

with runners on second and third the probability is the highest with 58.17% with zero outs, 65.93% with one out, and 72.91% with two outs.

The hitter's best option is to look for pitches in cluster 4 when baserunners are on second and third with a lefty on the mound when there are one or two outs. When runners are on second and third and a righty is pitching the hitter should also look for pitches in cluster 4. With no baserunners the hitter has a higher likelihood of receiving a pitch in cluster 2 as opposed to cluster 4, therefore the hitter should look in cluster 2.

Moving to cluster 5, the probability is predicted by the intercept, the inning and baserunners on second and third, with p-values of .0001, 0.0435, 0.0421 and slopes of -1.7524, .0682, and 1.1956 respectively. The later in the game the higher likelihood the pitch will be thrown to cluster 5. It is highly unlikely early in the game, where, with runners on second and third, late in the game the probability rises to 26% with no baserunners on and 54% with runners on second and third. The only time the hitter should look for a pitch in cluster 5 when compared to cluster 2 is when there are runners on second and third in the eighth, ninth, or tenth inning.

**Table 7. Cluster 5 Probability**

<b>Cluster 5</b>		
<b>Inning</b>	<b>No BR</b>	<b>BR 2+3</b>
1	15.71%	38.23%
2	16.67%	39.89%
3	17.65%	41.58%
4	18.69%	43.29%
5	19.78%	45.02%
6	20.92%	46.75%
7	22.10%	48.50%
8	23.33%	50.25%
9	24.60%	52.00%
10	25.92%	53.74%

## Validation

To validate our findings, we chose to create a confusion matrix to show the accuracy of predicted clusters compared to actual clusters. In a good model, the predicted clusters will be the same as the actual clusters. Summing the diagonal numbers and dividing by the sum of all pitches we found the accuracy at 38.16%. This is a slight improvement over having no information to help predict pitches, with a no information rate of 36.1%. The model's accuracy can be improved, but the overall methodology and findings begin to shed a light on predicting pitch location.

**Table 8. Confusion Matrix**

		Predicted Clusters					Total	Percentage
Actual Clusters		1	2	3	4	5		
	1	166	428	0	7	0	601	29.83%
	2	131	588	0	9	0	728	36.13%
	3	37	140	0	3	0	180	8.93%
	4	80	244	0	15	0	339	16.82%
	5	23	139	0	5	0	167	8.29%
		437	1539	0	39	0	2015	
		Diagonal	769					
		Accuracy	38.16%					



## IV. DISCUSSION

In this study, we set out to create the first method to predict pitch location. Using this data to help players create an approach based on pitch location may be at-bat data's greatest power. If players are receptive and administrators are willing to invest in teaching players how to implement these strategies into their game, teams will begin to outcompete other teams who are not using these tactics. Teams with not much money or status to obtain the best players should find ways to capture this data and use this information to help compete with higher ranked teams. Throughout this process, each part of the analysis begins to shed new light on how to optimize the hitter's approach. In the following section, the importance of each step of our methodology will be explained as well as discussion about rounding out the hitter's approach and future directions of pitch prediction to improve player performance.

### **Benefits of Contour Maps**

The first findings of the contour maps give the hitter a great starting point because the contour maps are user-friendly. The hitter does not have to think too much to interpret the data. Most hitters who never make it to the major leagues go through their entire baseball career and never see this type of analysis. If a hitter has access to a contour map of their hitting-zone they can easily see how pitchers throw them in different situations, what pitches they handle best, and what areas of the strike-zone they can improve their results. The hitter also learns what parts of the strike-zone he can eliminate.

Through the results of the contour map as seen in Figure 5 we see the upper right-hand corner can be eliminated for most at-bats. The contour maps filtered by pitch result in Figure 6 shows the hitter's lack of discipline in certain parts of the strike-zone. Look at

all the pitches not swung at that were strikes in Figure 6d and all the pitches that were swung at that were balls in Figure 6e. This shows the hitter can improve his approach by learning how to hit the pitches that cross through cluster 1 more often, and how difficult the bottom right corner of the strike zone is to distinguish between a ball and a strike. To combat these challenges, we created the clusters to see where pitches occur most often to give the hitter the best place to look for a pitch in different situations in a game.

### **Improving Approach using GMM**

Clusters give us a way to minimize the limitless pitch locations to just five, a manageable number for the hitter to control. We also learn from Figure 6 and Table 3 how each part of the strike-zone has different percentages of pitch results, teaching the hitter what clusters he had the best and worst results.

Having an approach with only the three pitches in the strike-zone, the clusters cover the parts of the plate where the hitter can be disciplined with only one side of the strike-zone. In figure 5, clusters 1,2, and 4 are in the strike zone and only have one border of the strike-zone near their location. If the hitter looks for a pitch in cluster 2, they only must know the inside limit of the strike-zone. If they look for cluster 1, they only need to be disciplined with the outside part of the strike-zone. If the hitter chooses cluster 4, they only need to be disciplined on the bottom half of the strike-zone. These clusters keep the hitter from having to be aware of two limits of the strike-zone at a time, which is not true in clusters 3 and 5.

The hitter did not know until we ran the Gaussian mixture model how to be selectively aggressive. With the hitting-zone broken up into five clusters the hitter learns most hits occur in cluster 2 and learns to avoid swinging at pitches in clusters 3 and 5.

For situations that demand the hitter to swing at a lower percentage pitch, the hitter may try expanding further off of a cluster within the strike-zone as opposed to looking for a pitch located in cluster 3 or 5. If the hitter were to focus on cluster 1 or 4 for a two-strike approach and expand the zone they will still be able to cover the bottom-right corner of the plate, but if they were to set their approach to the bottom-right corner of the plate, they are more likely to swing at pitches outside of the strike-zone. More pitches in the strike-zone will be swung at and fewer pitches will be chased after by the hitter that are located outside of the strike- zone. Next, we needed to understand when to look in certain clusters dependent on different situations in a game, especially when to look for pitches close to clusters 3 or cluster 5, but thrown in the strike-zone. When to look in these areas to create the hitter's approach is important but was not understood until the regression was run.

### **Importance of Regression Analysis**

The Gaussian Mixture Model teaches the hitter which cluster resulted in a hit most frequently, but the multinomial logistic regression teaches the hitter when to come off his favorite pitch and go after other pitches. The results of the regression help the hitter create an approach for all circumstances but especially help the hitter create a more disciplined approach when handling the pitch that is near cluster 3 in the strike-zone, and out of the strike-zone in cluster 3.

The best cluster to look for a pitch given no prediction information is cluster 2 because the hitter finds most success when swinging at pitches in this cluster. Cluster 1 and cluster 2 are difficult to distinguish, but when a right-handed pitcher is on the mound the hitter should look for pitches in cluster 1 as opposed to looking for a pitch in cluster 2

for a left-handed pitcher. If the hitter knew this while they were playing, they might have swung at more pitches in cluster 1, instead of taking so many pitches.

The other cluster in the strike-zone is cluster 4. The best time to look for a pitch in cluster 4 is with baserunners on second and third. This may be because the double play is not in order and the pitcher loses his ability to get a ground ball out at any base. Knowing the easiest way to let a run score is by a passed ball, the pitcher does not want to throw a pitch off the plate, so the pitcher throws more towards the middle of the plate. But, since pitchers, in general, do not throw the ball up in the zone, the hitter can look for the pitcher throw the ball in the middle of the plate and down in the zone. It is smart for the hitter to look for pitches in cluster 4 when runners are on second and third when a lefty and right-handed pitcher are throwing. Only when a lefty is throwing with 0 outs should the hitter look in cluster 2 when runners are on second and third. Every other situation with runners on second and third the hitter will maximize his chances to predict pitch location correctly if he chooses to look for a pitch in cluster 4 as opposed to cluster 1 or 2.

To cover the low and outside corner properly the hitter needs to know when pitchers try to throw to cluster 3. When the hitter has two strikes in the count for a right and left-handed pitcher cluster 3 is predicted more than 50% of the time. Therefore, the hitter should look for pitches in the bottom right-hand corner of the strike-zone with two strikes. This creates the hitter's two-strike approach. Because a pitch that crosses into cluster 3 is a ball, the hitter should choose either cluster 1 and expand down or cluster 4 and expand out to the border of the strike-zone during two-strike at-bats.

Pitchers also look to go to cluster 5 later in the game. The hitter should not swing at pitches going into cluster 5 ever because this pitch is considered a ball. If the hitter were to want to cover the up and inside corner the only time of the game to really look for pitches in this area is in the eighth, ninth, or tenth inning with baserunners on second and third. Much like the two-strike approach, if the hitter looks in cluster 2 and expands his zone up a little, he will be able to cover the top inside corner of the plate without going too far up and inside outside of the strike zone, limiting his swings at balls outside of the strike zone in cluster 5.

### **Overall Model Performance**

The purpose of this study was to determine if the variables of the game could help predict the location of pitches before they enter the hitting-zone. If hitters create an approach to look for a pitch in a certain part of the strike-zone, they will have a lower out-of-zone swing rate because they have more narrowly defined the area with which they are looking for the ball to enter before they swing. Also, fewer pitches outside of the strike-zone will be swung at, and if the hitter gets outside of his three-and-a-half-inch zone, the likelihood that they swing at a strike is much higher than if the hitter was looking at covering the entire strike-zone during an at bat.

When using an approach to wait for a certain pitch located in the strike-zone, it is like the hitter knows where the ball is going and is a step ahead of the ball as opposed to chasing a pitch that looks good enough to hit solid. This is very valuable to a hitter because instead of chasing behind a pitch in a few hundred milliseconds, the hitter is now ahead of the action and can react on time. Setting a location within the strike-zone before the ball is released allows the hitter to focus his attention more on the release point and

focus less attention on covering the entire strike zone, helping him become more of an expert hitter, as Faddie suggests in his study on how expert hitters track the ball from the release point better than non-expert hitters.

Competing with pitch location prediction allows each player to compete against opponents in an entirely new way. Where most organizations in baseball are interested in the player's performance and not in developing them, this approach allows hitters a chance to continue to develop their approach with statistical evidence. Hitters can use their time and energy at training what pitches to look for in all situations of a game. This will allow players to start competing in aspects of the game other than just skill or natural ability. Now players can pre-program counts and situations into their approach and find more success, where this was not an option before. When the hitter implements this approach, they will have more places to turn to discover tendencies in their approach, instead of just tweaking mechanics of the swing to improve results. The best players will now be those who have an ability to study and be prepared instead of competing only on a good swing or natural ability.

### **Improving Model**

To improve the model, we might start using other variables to predict pitch location. Previous pitch characteristics were not used in our analysis, but this may be a good predictor for pitch location in future studies. Many times, pitchers will decide which pitch to throw dependent on the result of the previous pitch. The score of the game, and where the hitter is batting in the line-up may also improve the regression model. Game location and ERA of the pitcher are also variables that could strengthen the model.

Another variable that could be used is the tendency of the pitcher to throw a certain pitch in certain situations. Each of these variables could benefit the strength of this model.

Another variable not mentioned in any of the literature is pitcher style. We did not use this variable in our analysis because it was unstable, but this may be a telling variable if we can get this variable to stabilize in future studies. A pitcher throwing over-the-top tends throw like other pitchers who throw over-the-top. The same is true for all pitcher styles (over-the-top, three-quarters, side-arm, and submarine). Also, the hitter must start their eye-sight in a different area to pick up each pitcher's style unique release point. If the data was filtered dependent on pitcher style the results may come back even more predictable than the original analysis and the hitter may learn how different pitcher styles work the strike-zone.

Another option to improve the model is to use data from major league baseball through Pitchf/x or Statcast. The model may be improved because different predictors could be used to explain more of the model because major league baseball's system is so thorough and advanced. Computer systems are probably more accurate than the hitter recording data after games, therefore the data may be more accurate and predictable.

### **Other's Models**

The regression explains where to predict pitches during different situations in a game. We found the pitcher being a right-handed or left-handed pitcher, outs, inning, strikes and baserunners at second and third to be the only significant variables to describe certain clusters. Balls, pitch number, and inning were all found insignificant in the regression, where balls and innings were found to be significant variables in studies conducted before. In our study balls were not significant, but count is also found to be

significant in other studies. Handedness of the pitcher was not found to be significant in any of the studies conducted in the literature review, but in our study, it was the most consistent variable. This may be because the location of the pitch is much more dependent on the handedness of the pitcher as opposed to which pitch type is thrown.

Our model also improves the hitter's ability to predict pitches in less favorable counts, where Ganashapilai and Gutttag found the pitcher is less predictable in counts favoring the pitcher. This gives the hitter a stronger chance for success when the hitter is at a disadvantage. Using at-bat data may also help the hitter decide to swing or wait. If pitch location were added to the swing or wait concept introduced by Bickel, the hitter will know whether he has a better opportunity swinging at the pitch if the pitch is located in the cluster he is looking. Our model may also help hitters handle the increase in relief pitching. The hitter may be able to create a strategy before the new relief pitcher enters the game, taking away the advantage of relief pitchers.

### **Collection and Implementation**

For teams who do not have the means to purchase millions of dollars of equipment to set up a pitch tracking system like major league baseball, finding ways to record pitches as they cross through the zone will be beneficial. As more teams begin to compete on pitch prediction it will be a necessity for teams to have this type of information. This data is valuable at lower levels, but where major league baseball is using this information to compete, lower levels are not using this data. This is important to use at lower levels, but finding an affordable way to record this data at lower levels is a challenge. Inputting data manually is a low-cost option, but it could be too time



consuming to bring value. Ways to capture the data without having to spend too much money or put too much time entering the data should be explored for lower level teams.

Considering how this information is presented to the player in the game is another strategy that needs to be handled with care, as teams do not want to confuse the hitter with too much information. Most of hitting happens so fast and too much information could cloud the hitter's ability to compete with a clear mind. A way to implement a predictable strategy into a game during at-bats might have players studying before the game, reviewing notes in the dugout, wearing a wristband for plate appearances, or by relaying the approach electronically. Lastly, in theory this model works, but to test this method a player or team would have to see if the results improve from previous at-bats or seasons.

## **Conclusion**

Creating an approach dependent on predicting the pitch location was created by using contour maps, the Gaussian mixture model to create five areas for the hitter to set as approaches, and a multinomial logistic regression to explain when to change from one location to another during different situations in a game. The hitter learns to create approaches to handle at-bats when facing a right-handed or left-handed pitcher, when the hitter has strikes on the count, when the game increases in innings played, when the inning increases the outs, and when baserunners are on second and third. The future is bright for predicting pitch location. Teams who can implement these strategies will have an advantage over teams who do not.








## APPENDIX SECTION

### Cluster Results

Cluster	Count	Proportion
1	601	0.30993
2	728	0.37750
3	180	0.08539
4	339	0.15342
5	167	0.07376

Cluster	PitchHeight	PitchWidth
1	-11.354209	11.1691922
2	-10.505841	4.70481972
3	-18.376871	11.5146122
4	-14.999887	8.15161399
5	-1.9364645	6.92822769

### Multinomial Regression Results

Source	LogWorth	PValue
PitcherR/L	5.140 	0.00001
Outs	2.915 	0.00122
Baserunners	2.565 	0.00272
Strikes	1.091 	0.08118
Balls	0.587 	0.25898
PitchNumber	0.527 	0.29693
Inning	0.501 	0.31531

### Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	67.1064	52	134.2127	<.0001*
Full	2856.0222			
Reduced	2923.1286			

RSquare (U)	0.0230
AICc	5827.3
BIC	6138.11
Observations (or Sum Wgts)	2015

### Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-0.3538833	0.2472683	2.05	0.1524
PitcherR/L[left]	-0.1634931	0.060851	7.22	0.0072*
Inning	0.00938859	0.0219302	0.18	0.6686

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Outs	0.06651665	0.072103	0.85	0.3563
Baserunners[0]	0.07471967	0.1344751	0.31	0.5785
Baserunners[1]	0.30148665	0.1769827	2.90	0.0885
Baserunners[1,2]	0.26263916	0.2209887	1.41	0.2346
Baserunners[1,2,3]	-0.1225372	0.3346525	0.13	0.7142
Baserunners[1,3]	0.1066228	0.2653093	0.16	0.6878
Baserunners[2]	0.06864701	0.204446	0.11	0.7370
Baserunners[2,3]	0.11607273	0.5386678	0.05	0.8294
PitchNumber	-0.1903685	0.1679816	1.28	0.2571
Balls	0.29393437	0.1827623	2.59	0.1078
Strikes	0.23274886	0.2034617	1.31	0.2526
Intercept	-1.0740446	0.3906624	7.56	0.0060*
PitcherR/L[lefty]	-0.3016828	0.0985384	9.37	0.0022*
Inning	0.02883204	0.0335225	0.74	0.3897
Outs	-0.0294653	0.1104827	0.07	0.7897
Baserunners[0]	-0.3578808	0.1946894	3.38	0.0660
Baserunners[1]	0.3408929	0.2412313	2.00	0.1576
Baserunners[1,2]	0.04058252	0.3289315	0.02	0.9018
Baserunners[1,2,3]	-0.1459183	0.4693541	0.10	0.7559
Baserunners[1,3]	0.32012808	0.3507308	0.83	0.3614
Baserunners[2]	0.28045063	0.2761679	1.03	0.3099
Baserunners[2,3]	0.17454547	0.7307896	0.06	0.8112
PitchNumber	-0.464815	0.2846116	2.67	0.1024
Balls	0.06852605	0.3094796	0.05	0.8248
Strikes	0.81021569	0.3318145	5.96	0.0146*
Intercept	-0.9301525	0.3075636	9.15	0.0025*
PitcherR/L[lefty]	-0.3557114	0.0779754	20.81	<.0001*
Inning	0.0273673	0.0264677	1.07	0.3011
Outs	0.33154051	0.0857024	14.97	0.0001*
Baserunners[0]	-0.1705779	0.1413606	1.46	0.2276
Baserunners[1]	0.12470398	0.1945863	0.41	0.5216
Baserunners[1,2]	-0.547724	0.2875378	3.63	0.0568
Baserunners[1,2,3]	-0.4912463	0.404577	1.47	0.2247
Baserunners[1,3]	-0.1485497	0.3021687	0.24	0.6230
Baserunners[2]	-0.2624635	0.2335705	1.26	0.2611
Baserunners[2,3]	1.25659442	0.4542344	7.65	0.0057*
PitchNumber	-0.3698102	0.2278426	2.63	0.1046
Balls	0.37333933	0.2441316	2.34	0.1262
Strikes	0.4867098	0.2671933	3.32	0.0685
Intercept	-1.752391	0.3848329	20.74	<.0001*
PitcherR/L[lefty]	-0.0059926	0.0921368	0.00	0.9481
Inning	0.06815948	0.0337606	4.08	0.0435*
Outs	-0.0344206	0.1115441	0.10	0.7576
Baserunners[0]	0.22523556	0.2002493	1.27	0.2607
Baserunners[1]	-0.4139642	0.3201428	1.67	0.1960
Baserunners[1,2]	0.22203482	0.3421843	0.42	0.5164
Baserunners[1,2,3]	-0.0458445	0.5096737	0.01	0.9283
Baserunners[1,3]	-0.6331438	0.5548428	1.30	0.2538
Baserunners[2]	0.27275364	0.3065498	0.79	0.3736
Baserunners[2,3]	1.19560676	0.5881114	4.13	0.0421*
PitchNumber	-0.068676	0.263578	0.07	0.7944
Balls	-0.177116	0.2904022	0.37	0.5419
Strikes	0.15549681	0.3182039	0.24	0.6251

For log odds of  $1/2$ ,  $3/2$ ,  $4/2$ ,  $5/2$

## LITERATURE CITED

- Albert, J. (2010). Baseball data at season, play-by-play, and pitch-by-pitch levels. *J Stat Educ*, 18(3).
- Bickel, J. E. (2009). On the decision to take a pitch. *Decision Analysis*, 6(3), 186-193.
- Choe, J., & Kim, J. S. (2016). Minimax after Money-Max: Why Major League Baseball Players Do Not Follow Optimal Strategies.
- Cox, D. J., Sosine, J., & Dallery, J. (2017). Application of the matching law to pitch selection in professional baseball. *Journal of Applied Behavior Analysis*, 50(2), 393-406.
- Dellaert, F. (2002). The expectation maximization algorithm. Georgia Institute of Technology.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 1-38.
- Fadde, P. J. (2006). Interactive video training of perceptual decision-making in the sport of baseball. *Technology, Instruction, Cognition and Learning*, 4(3), 265-285.
- Fast, M. (2010). What the heck is PITCHf/x. *The Hardball Times Annual*, 2010, 153-158.
- Flanagan, J. (2015). *Examining MLB's Strike Zone* (Doctoral dissertation, Worcester Polytechnic Institute).
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, pp. 241-249). New York: Springer series in statistics.
- G. Moore, C., & Müller, S. (2014). Transfer of expert visual anticipation to a similar domain. *The Quarterly Journal of Experimental Psychology*, 67(1), 186-196.
- Ganeshapillai, G., & Guttag, J. (2012, March). Predicting the next pitch. In *Sloan Sports Analytics Conference*.
- Hamilton, M., Hoang, P., Layne, L., Murray, J., Padget, D., Stafford, C., & Tran, H. (2014, March). Applying machine learning techniques to baseball pitch prediction. In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (pp. 520-527). SCITEPRESS-Science and Technology Publications, Lda.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Harrison, W. K., & Salmon, J. L. *Bullpen Strategies for Major League Baseball*.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Law, K. (2018). *SMART BASEBALL: the story behind the old stats that are ruining the game, the new ones that are ruining it, and the right way to think about*. NEW YORK: WILLIAM MORROW.

McLachlan, G. J., & Krishnan, T. (1997). Wiley series in probability and statistics. The EM Algorithm and Extensions, Second Edition, 361-369.

Quintana, F. A., Müller, P., Rosner, G. L., & Munsell, M. (2008). Semi-parametric Bayesian inference for multi-season baseball data. *Bayesian analysis (Online)*, 3(2), 317.

Superak, H. M. (2011). *Analyzing Batting Patterns of Major League Baseball Players For Advance Scouting Reports: Using R to Generate High-Level Spatial Plots of PITCH/x Data* (Doctoral dissertation, Emory University).

Shawchik, T. (2017, May 10). Eric Thames and the Transformative Power of Boredom. Retrieved October 31, 2017, from <https://www.fangraphs.com/blogs/eric-thames-and-the-transformative-power-of-boredom/>.

Williams, T., Underwood, J., & Cupp, R. (1986). *The science of hitting*. New York, NY: Simon & Schuster.