View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Chapman University Digital Commons

Chapman University

Chapman University Digital Commons

Computational and Data Sciences (PhD) Dissertations

Dissertations and Theses

Spring 5-2020

Novel Statistical and Machine Learning Methods for the Forecasting and Analysis of Major League Baseball Player Performance

Christopher Watkins
Chapman University, watki115@mail.chapman.edu

Follow this and additional works at: https://digitalcommons.chapman.edu/cads_dissertations

Recommended Citation

C. Watkins, "Novel statistical and machine learning methods for the forecasting and analysis of Major League Baseball player performance," Ph.D. dissertation, Chapman University, Orange, CA, 2020.
https://doi.org/10.36837/chapman.000139

This Dissertation is brought to you for free and open access by the Dissertations and Theses at Chapman University Digital Commons. It has been accepted for inclusion in Computational and Data Sciences (PhD) Dissertations by an authorized administrator of Chapman University Digital Commons. For more information, please contact laughtin@chapman.edu.

Novel Statistical and Machine Learning Methods for the Forecasting and Analysis of Major League Baseball Player Performance

A Dissertation by

Christopher Watkins

Chapman University

Orange, CA

Schmid College of Science and Technology

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational and Data Sciences

May 2020

Committee in charge:

Cyril Rakovski, Ph.D., Chair

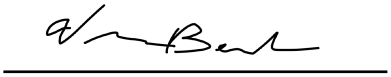
Vincent Berardi, Ph.D.

Adrian Vajiac, Ph.D.

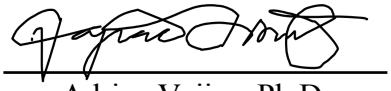
CHAPMAN UNIVERSITY
SCHMID COLLEGE OF SCIENCE AND TECHNOLOGY
Computational and Data Sciences

The dissertation of Christopher Watkins is approved.

Cyril Rakovski, Ph.D., Chair



Vincent Berardi, Ph.D.



Adrian Vajiac, Ph.D.

May 2020

Novel Statistical and Machine Learning Methods for the Forecasting and Analysis of Major League Baseball Player Performance

Copyright © 2020

by Christopher Watkins

ACKNOWLEDGEMENTS

First, I would like to thank my entire committee, Dr. Cyril Rakovski, Dr. Vincent Berardi, and Dr. Adrian Vajiac, for their help and support not only for this dissertation, but my entire time at Chapman. First, thank you Cyril for supporting what I wanted to do, never pressuring me to go a different route, and always helping me with any problem I had. Baseball is not a common topic for a dissertation, and even with little knowledge of the game you supported my endeavors, gave me valuable advice for methods to employ, and how to improve my ideas. I could not have completed this without you advising me. Next, thank you Vinny for your excitement and support of this project. I'm lucky to have taken your interactive data analysis course and learn about your passion for baseball which was so valuable for this dissertation. Finally, thank you Adrian so much for your support for the last about 9 years. I've known you since I was an 18-year old freshman way back in 2011 and you've seen me grow up and go through a lot of ups and downs. Your door was always open for me to ask any questions or bring up any problems I was going through. You consistently helped me even when you were swamped with work. I really appreciate everything you have done for me during my time at Chapman.

Next, I'd like to thank everyone in CADS who supported me throughout my graduate journey. It was not easy, but without you all I would not be where I am today. First, the staff and faculty in the program were fantastic and were always willing to help. Dr. Hesham El-Askary, thank you for allowing me to be in this program and funding my time in CADS. I want to express my thanks and gratitude to Robin Pendergraft, who was the graduate coordinator for the majority of my time in the program. You went above and beyond for us, the students. We all knew you cared

about our well-being and success in the program. Many times, I would email you about a problem I had or a form to be signed and you got it solved quickly for me. You made my graduate school life much easier and I am very appreciative of that. I also want to thank my friends Chelsea-Parlett-Pelleriti and Viseth Sean. Chelsea, thank you for being my first and best friend in the program. We started as partners in CS 510 and never looked back. Text messages, Slack messages, and discussing homework or studying for exams were almost daily. I remember studying all summer for the qualification exams and countless other exams. I am so thankful for our friendship and all the help over the last few years. Viseth, thank you for always being a positive voice. We've gone through a lot, but I always knew you would be supportive and positive even when times were tough. There are many other in the CADS program that have made a positive impact on me, even if we did not know each other that long and I want to thank them too. We went through many battles together, and I would not trade it for anything. The reason why I love school so much is because of the people and my cohort was composed of some of the best people I've ever met.

I would also like to thank everyone that supported me throughout my entire academic journey. It has been a long 9 year run since the beginning of undergrad and I would not have been able to do it without the help of many people. To all the faculty and staff at Chapman University, thank you for everything. The impact you have made on me is immeasurable and I will be forever grateful. To my great friends, Andrew Ferrell, Anibal Hernandez, Chad Walker, and Karynna Okabe-Miyamoto who I have known a long time: Thank you for the positivity and support during this long journey. Escaping the academic world by going to dinner, Disneyland, or an Angels game was always wonderful. It means so much to me to have such great friends I could lean on and be myself with.

Finally, and most importantly, I'd like to thank my family. There isn't a day that goes by that I don't think about how lucky I am to have such a supportive family during this academic journey. In particular I want to thank my parents, Jerry and Hiroko Watkins. They have been my biggest supporters since day one when I decided to go to Chapman, left for Oregon State, and came back to Chapman. It was a difficult journey and we went through a lot over the years. I know it was difficult on you both, seeing the stress, anxiety, and sadness throughout my academic career but through all that I know you are proud of my perseverance and happy that I achieved my dreams. I could never have gone through this without you and no words would be enough to show my gratitude.

ABSTRACT

Novel Statistical and Machine Learning Methods for the Forecasting and Analysis of Major League Baseball Player Performance

by Christopher Watkins

Baseball has quickly become one of the most analyzed sports with significant growth in the last 20 years [1] with an enormous amount of data collected every game that requires professional teams to have a state-of the-art analytics team in order to compete in today's game. Statcast, introduced in 2015, "allows for the collection and analysis of a massive amount of baseball data, in ways that were never possible in the past" [2]. Using this new Statcast data that is updated every pitch, a novel metric was developed, Pitcher Effectiveness, that is updated dynamically throughout a game. It was shown to be predictive of runs in combination with rate of change of the metric as well as effective in evaluating a starting pitcher on the game level and overall. Baseball can be broken down into a Markov Chain with 24 different states based on the combination of outs and baserunners where throughout the game teams will transition from one base/out state to another when events such as hits, outs, walks, and others occur [3]. Using this idea, pitch sequencing was explored on the micro level of each state individually. Looking at the last three pitches in a sequence, certain sequences in particular states were shown to have some predictive power in predicting outs, hits, and strikeouts. In addition, proportion tests showed significant differences in the proportion of outs and strikeouts of sequences depending on the baseball state. From fantasy baseball to Major League Baseball (MLB) front offices, projections of players' future performance are important and are explored quite often. Several machine learning methods were explored for

projecting future weighted on base average (wOBA) [3]. These methods were evaluated and the best were compared to 2020 projections from the reputable Steamer [4].

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

Abbreviation	Meaning
BB/9	Walks per 9 innings
ERA	Earned Run Average
ERA-	Earned Run Average Minus
EV	Effective Velocity
FB%	Fastball Percentage
FIP	Fielding Independent Pitching
GB%	Groundball Percentage
HP	High Performance
HR/9	Homeruns per 9 innings
K/9	Strikeouts per 9 innings
LP	Low Performance
LRPP	Linear Run Pitcher's Performance
MAE	Mean Absolute Error
NN	Neural Network
OBP	On Base Percentage
PTB	Pitcher's Total Bases
PV	Perceived Velocity
RBI	Runs Batted In

RMSE	Residual Mean Squared Error
SCAD	Smoothly Clipped Absolute Deviation
SLG	Slugging Percentage
STB	Strike-To-Ball Ratio
SVM	Support Vector Machine
SwStr%	Swinging Strike Percentage
WHIP	Walks and Hits per Inning Pitches
wOBA	Weighted On Base Average
xBA	Expected Batting Average
xwOBA	Expected Weighted On Base Average

1 Pitcher Effectiveness: A Step Towards in Game Analytics and Pitcher Evaluation

1.1 Introduction

1.1.1 Evolution of Statistics in Baseball

Baseball has quickly become one of the most analyzed sports with significant growth in the last 20 years [1] with an enormous amount of data collected every game that requires professional teams to have a state-of the-art analytics team in order to compete in today's game. As an example, the Houston Astros lost over 100 games each season from 2011-2013. In 2014, sports writer Ben Reiter predicted the Astros to win a World Series sooner rather than later, in 2017, because of the advanced analytics team they built [5].The Houston Astros ended up winning the 2017 World Series, as predicted by Mr. Ritter. In 2019 the Tampa Bay Rays brought Jonathan Erlichman, who has not played baseball past T-ball, into the dugout as "the first full-time analytics coach ever to join a major-league staff. In his new role, he will use his knowledge of data to assist manager Kevin Cash with in-game decisions and provide real-time information to players" [6]. These examples highlight the role of data analytics within this sport.

The analysis of baseball has evolved over the years, with three major categories being sequentially developed. First, there are the traditional statistics such as homeruns, batting average, and earned run average. Next, baseball statisticians created new statistics called Sabermetrics which further improved the analyses of player performance. This concept was

pioneered by Bill James in the 1980's and defined sabermetrics as "the search for objective knowledge about baseball" [7]. Some examples of Sabermetric statistics include on-base percentage (OBP), slugging percentage (SLG), and wins above replacement (WAR). This started the Moneyball movement, where teams analyzed players differently than in the past with general manager Billy Beane of the Oakland Athletics leading the charge [8].

The final category of statistics was made possible by the introduction of Statcast in 2015, which revolutionized Major League Baseball [2]. Statcast tracks every single play on the MLB field and "allows for the collection and analysis of a massive amount of baseball data, in ways that were never possible in the past" [2]. The technology that makes this possible "is a combination of two different tracking systems -- Trackman Doppler radar and high definition Chyron Hego cameras. The radar, installed in each ballpark in an elevated position behind home plate, is responsible for tracking everything related to the baseball at 20,000 frames per second. This radar captures pitch speed, spin rate, pitch movement, exit velocity, launch angle, batted ball distance, arm strength, and more. Separately, each ballpark also has a Chyron Hego camera system, where six stereoscopic cameras are installed in two banks of three cameras apiece down the foul line. The camera system tracks the movement of the people on the field, which allows for the measurement of player speed, distance, direction, and more on every play" [2].This new data from Statcast is easily accessible through BaseballSavant [9]. Proper analyses of these high precision, multidimensional data should be able to provide in game analytics to coaches that would enable them to better evaluate player performance in real-time.

1.1.2 Decision to Remove a Starting Pitcher

One of the most difficult decisions a MLB manager has to make is when to remove a starting pitcher. Remove a starting pitcher too early, you do not maximize his use and risk overworking relief pitchers. Remove a starting pitcher too late when he is fatigued, and he will likely give up many runs and/or place your relief pitchers in difficult situations. A real-time predictive model could help the manager make the optimal decision during the game. One method that was proposed in 2017 [10] considered pitches as time series data and used dynamic time warping and 1-nearest neighbor to classify the outing on an ongoing basis. Using the linear weights for all possible plays and count, the metric Linear Run Pitcher's Performance (LRPP) was built as a rolling sum to classify the performance as High Performance (HP) or Low performance (LP). When the result of 10 pitches were unknown, precision, recall and F1 values had means 0.9, 0.8, and 0.89 and with 30 pitches unknown the F1 value was 0.78 [10]. The models did better as more pitches were thrown and the authors "believe that the model should perform well when starting pitchers exceed [50 pitches], given that the average number of throws per game of the 20 studied pitchers is equal to 101" [10].

Another approach that has been taken was building a regression model that uses past inning at bats, game situation, and historical data to predict Pitcher's Total Bases (PTB) for the following inning; a cut-off value was then used to determine if the pitcher should be taken out [11]. After the PTB model makes a prediction, it was compared to a manager model, which was built from actual manager decisions [11], which predicted the manager's decision correctly for 95% of the innings. The manager model correctly predicted a run being scored for 75% of evaluated innings and the PTB model correctly predicted a run being scored 81% of the innings [11]. Considering the 5th inning on, which was 21,538 innings, the PTB model disagreed with the manager's

decision 48% of the time [11]. Results suggested that the PTB model performed well; when the manager decided to leave the pitcher in and the PTB model agreed, 17.7% of the innings the pitcher gave up at least one run. In contrast, when the manager left a starting pitcher in and the PTB model disagreed, 31.5% of the innings the pitcher gave up at least one run [11].

In 2017 Harrison and Salmon also addressed the question of when to remove a pitcher from the game by using a system that uses pitch counts and strike-to-ball ratio (STB) [12]. Using 700,000 pitches from the 2015 season, linear regressions were built regressing balls and strikes with respect to pitch count for a particular pitcher for the season [12]. This regression line "represents an expected strike-to-ball ratio" for a pitcher, which means for a game that a pitcher was doing better than the expected STB the pitcher was pitching above average and pitching below the trend line indicates a pitcher was faltering [12]. Looking at the number of games a pitcher reaches a certain pitch count and the STB at different pitch counts, the authors found that at high pitch counts there was a drop in the mean STB line with a unique number of pitches where this occurs for each pitcher [12]. The authors suggest "these changes in performance can be used as trigger points to evaluate if a pitcher is tired or reaching his limit in other ways and perhaps needs to be pulled" [12]. They propose that a pitcher is removed at this trigger point and managers using knowledge of the mental and physiological state of the pitcher could improve the decision making [12].

In 2010 Piette, Braunstein, McShane and Jensen developed a point mass Bayesian random effects model to evaluate the effectiveness of a pitcher [13]. They found that the metrics with the highest signal were ground ball percentage (GB%), fly ball percentage (FB%), and strikeouts per nine innings (K/9) for relief pitchers and fielding independent pitching (FIP), homeruns per nine

innings (HR/9), pitchers, earned run average (ERA), and walks per nine innings (BB/9) for starting pitchers [13]. The authors argue that "high signal metrics have a large fraction of players which are different from league average and give high confidence about which players are not league average" [13]. The paper noted that the metrics used were not park, team, or league adjusted [13].

The above-discussed models to aid with the decision to remove a starting pitcher do not use the rich Statcast data that is now available. In this work we design a novel metric, Pitcher Effectiveness, that can be used to evaluate a starting pitcher on both an in-game and overall outing basis. Although it would be exciting to apply this in real time, MLB has restricted the use of technology in the dugout. This newly constructed metric, Pitcher Effectiveness, is unique in comparison to other metrics because it does not take runs into account but is designed as a predictor of runs. The goal of Pitcher Effectiveness is to measure how effective a pitcher is by only taking into account the variables that the pitcher can control. For example, a pitcher cannot control the defense so they should not be evaluated on runs caused by errors, but they do control working ahead of the count. Also, a pitcher who made a great pitch, with soft contact, but resulting in a hit should not be penalized because of a defensive shift or the hit falling between two fielders. Of course, baseball is a game where events are dependent on more than just the starting pitcher, but the starting pitcher has the biggest influence on the game. Using Statcast data, Pitcher Effectiveness is continuously calculated after each pitch to generate a rolling sum throughout the game. This chapter discusses the data used, methods, results, and future work involved with Pitcher Effectiveness.

1.2 Data

The analyzed data was obtained from BaseballSavant and included pitchers that threw 2000 or more pitches in 2018 MLB season. This data set included mostly traditional starting pitchers, but there were also swing pitchers who make multiple spot starts and the opener was used by several teams in 2018. We considered every pitch that these pitchers threw and examined variables such as pitch speed, post-pitch score, fielding alignment, launch angle and exit velocity among others. This resulted in 115 pitchers, 305,633 pitches, and 89 variables. We removed 1,150 pitches because of missing values produced by an error with Statcast, which left 304,483 total pitches. The analysis was restricted to 7 relevant variables, developed through domain knowledge, plus 5 new variables were created using transformations of the original ones. These new variables that were created included Pitcher Effectiveness, pitches against, runs, on base, pitches, hit, slope, and run prediction. A detailed description of these variables is shown in Table 1-1.

Table 1-1: Description of Variables

Variable	Description
Player Name	Pitchers name
Events	Ball in play event (Single, double, strikeout, etc. Null if not in play)
Description	Result of pitch (Hit into play, ball, strike, etc.)
Balls	Number of balls in at bat
Strikes	Number of strikes in at bat
Launch Speed	Exit velocity of batted ball
Post Bat Score	Opposing team score
Picher Effectiveness	New metric created
Runs	Resulting number of runs from pitch
Pitches	Cumulative number of pitches thrown by pitcher
Slope	Fitted slope of Pitcher Effectiveness over the previous 5 or 10 pitches
Run Prediction	Binary variable for x or more runs given up 5 or 10 after the pitches used in the slope (x = 1,2, or 3)

1.3 Methods

A model was designed to predict the number of runs given up by a pitcher using the Pitcher Effectiveness score and the change in this score over a certain number of pitches as covariates. Using 2018 Statcast data, the metric Pitcher Effectiveness was calculated as a time series comprised of each pitch per game. A pitcher starts with a Pitcher Effectiveness of zero and the value was continuously updated with each pitch throughout the game, until the pitcher was taken out of the game. The three variables used to calculate the Pitcher Effectiveness were event, ball and strike count and exit velocity. Each outcome for these three variables has a weight, and the sum of these weights for each pitch determine that pitch's contribution to Pitcher Effectiveness. The weights used for a single, double, triple, homerun, walk, and hit by pitch are linear weights for calculating weighted on base average (wOBA) for the 2018 season [14]. The linear weights are

found by calculating the run expectancy for each event using the data from the 2018 season [14]. All other weights (count, out, swinging strike, and exit velocity) were a carefully chosen and can be found in Table 1-2. An extensive grid search of values between 0.1 to 1.5 by 0.2 (8 values, 4096 combinations) was done to find the optimal weights. The computation took over 72 hours to complete on a 32-core computer. These values were compared to weights built with domain expertise, which ultimately performed better. For example, if the pitcher was ahead of the count, and gave up a single with an exit velocity of 95 mph then the Pitcher Effectiveness score for that pitch would be 0.5-0.88-0.5 = -0.88. The higher the Pitcher Effectiveness, the better the pitcher was doing overall. Negative values for Pitcher Effectiveness indicate that a pitcher was ineffective. Summary statistics of final Pitcher Effectiveness scores for pitchers' entire games can be found in Table 1-3. \

Table 1-2: Pitcher Effectiveness Weights

Event	Weight
Single	-0.88
Double	-1.25
Triple	-1.58
Homerun	-2.03
Walk	-0.69
Hit by pitch	-0.72
Out (except sacrifice fly)	+0.5
Other plays	+0
Swinging Strike	+0.5
Count	Weight
Ahead of count	+0.5
Behind count	-0.5
Even count	+0
Exit Velocity	Weight
95+ mph exit velocity	-0.5
80 mph or less exit velocity	+0.5

Table 1-3: Summary Statistics for Pitcher Effectiveness

Statistic	Value
Minimum	-16.38
Quantile 2	6.93
Mean	15.01
Quantile 3	22.77
Maximum	55.18

We hypothesized that trends in Pitcher Effectiveness would be a more useful predictive metric than the value associated with a single pitch. Therefore, for each pitch a linear regression model was used to estimate the trend (i.e. slope) over the previous 5 pitches; this was repeated for the previous 10 pitches as well. As a result of this procedure, the first 4 or 9 pitches from each game were not considered. The data set used was sufficiently large so that the predictive power of Pitcher Effectiveness was not compromised.

For the run prediction variable, we examined several run-based outcomes including any number of runs, more than 1 run, more than 2 runs, and more than 3 runs given up in the next 5 or 10 pitches. This procedure means that the last 5 or 10 pitches were ignored for each game respectively because there is nothing to predict when the pitcher is taken out of the game. Again, due to the large data, we don't expect these omissions to materially affect predictive ability.

We used the presence and absence of runs as the outcome variable and Pitcher Effectiveness and slope of the recent performance trend as predictor variables in a logistic regression model combined with 5-fold cross validation. We identified the best predictive model by comparing a range of potential models built using different combinations of the number of pitches used to calculate slope, the number of pitches used for run prediction, and the number of runs to predict.

In addition, Pitcher Effectiveness by itself was compared to other metrics that are used to evaluate a pitcher's performance. The grid search tested for the highest cross validated area under the ROC curve for the particular pitch and run prediction combination of 4+ runs, 10 pitches for slope calculation, and 5 pitches for run prediction. The logistic regression model used can be found in Equation 1-1.

$$log - odds(Run\ Prediction) \sim Pitcher\ Effectiveness\ +\ Slope$$

Equation 1-1: Run Prediction Logistic Regression Model

1.4 Results

The models did very well to predict a big inning, 3+ or 4+ runs scored, within the next 5 or 10 pitches. The best combinations can be found in with the 5-fold cross validated area under the curve (CV AUC). Both the Pitcher Effectiveness and slope variables were statistically significant (p-values of) in the models as seen in Table 1-4.

Table 1-4: Logistic Regression Model Results

CV AUC	Slope Coefficient	Pitcher Effectiveness Coefficient	Number of pitches for slope	Number of pitches for prediction	Number of runs
0.716	-1.522	-0.0369	5	5	3+
0.751	-1.901	-0.0304	5	5	4+
0.723	-1.578	-0.0306	5	10	4+
0.778	-2.968	-0.0185	10	5	4+
0.743	-2.407	-0.0254	10	5	3+
0.711	-2.074	-0.0214	10	10	3+
0.744	-2.551	-0.0183	10	10	4+

The models did not do well in predicting 1+ or 2+ runs scored in the next 5 or 10 pitches. All CV AUC values were less than 0.7 for these combinations. There are a few likely reasons why this

was the case. First, it takes much less to score 1-2 runs even if a pitcher was effective. For example, a pitcher could be doing well all game but miss location once and give up a homerun. Another example could be a pitcher giving off a leadoff double and a run scoring without another hit (i.e. Combination of moving the running via groundout or flyout and scoring on a groundout or sacrifice fly, etc.). Next, errors can cause runs to be giving up by a pitcher, although unearned. Both Pitcher Effectiveness and the slope do not take errors into account, which means a pitcher can still be effective, but the defense causes a run to be scored. It is also rare that there are more than 1-2 runs scored from an error. After the initial error, even though runs may be unearned, that pitcher must continue to pitch effectively to prevent runs and has a large effect on more runs being scored. Thus, predicting a larger number of runs was more successful.

To test if Pitcher Effectiveness was a good metric to evaluate average starting pitching performance over a season, it was compared to already-established metrics. Each pitcher's average Pitcher Effectiveness for the 2018 season was calculated and compared to earned run average (ERA) and earned run average minus (ERA-). ERA is the average number of earned runs (not a result of an error) per 9 innings for a pitcher. The lower a pitcher's ERA, the better they have performed overall. When plotting ERA and Pitcher Effectiveness there was a significant, strong negative correlation, which is evidence that Pitcher Effectiveness is a good metric. ERA- is a Sabermetric statistic that is park and league adjusted [15]. This allows pitchers to be compared more accurately regardless of their home ballpark and whether they pitch in the American or National league. For example, a pitcher whose home ballpark is Angels Stadium is at an advantage because it is a pitcher friendly park where less homeruns are hit but at Coors Field a pitcher is at a disadvantage because it is a hitter friendly park where more homeruns are hit. Just like ERA, lower ERA- indicates better performance of the pitcher. The adjustment

makes 100 average, where the amount less than 100 is the percentage they performed better than average while the amount above 100 is the percentage the performed worse than average. For example, a pitcher with an ERA- of 80 performed 20% better than average and an ERA- of 110 indicates the pitcher performed 10% worse than average. Again, there was a statistically significant, strong negative correlation between ERA- and Pitcher Effectiveness which would seem to indicate that Pitcher Effectiveness is a good metric for evaluating pitching performance. Both plots with their associated correlation can be found in Figure 1-1 and Figure 1-2.

To evaluate a pitcher's performance for a game, many look at the traditional pitching line to see the innings pitched, strikeouts, number of hits, runs, walks, and homeruns given up by a pitcher. In a search for one number to describe a pitcher's performance, Bill James developed the metric Game Score in the 1980's [16]. Each pitcher began with a score of 50, then their score would change depending on the play and associated weight from Table 1-5.

Table 1-5: Game Score Weights (Bill James)

Event	**Weight**
Start of Game	+50
Out	+1
Inning completed after 4th	+2
Strikeout	+1
Hit	-2
Earned Run	-4
Unearned Run	-2
Walk	-1

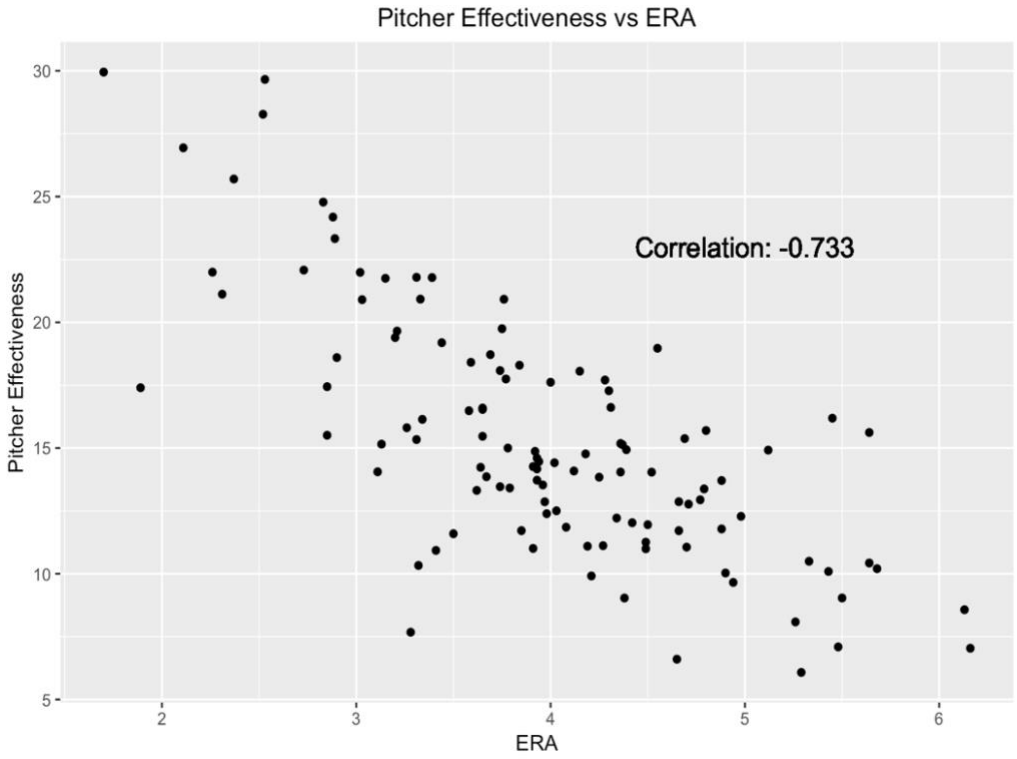


Figure 1-1: Pitcher Effectiveness versus ERA

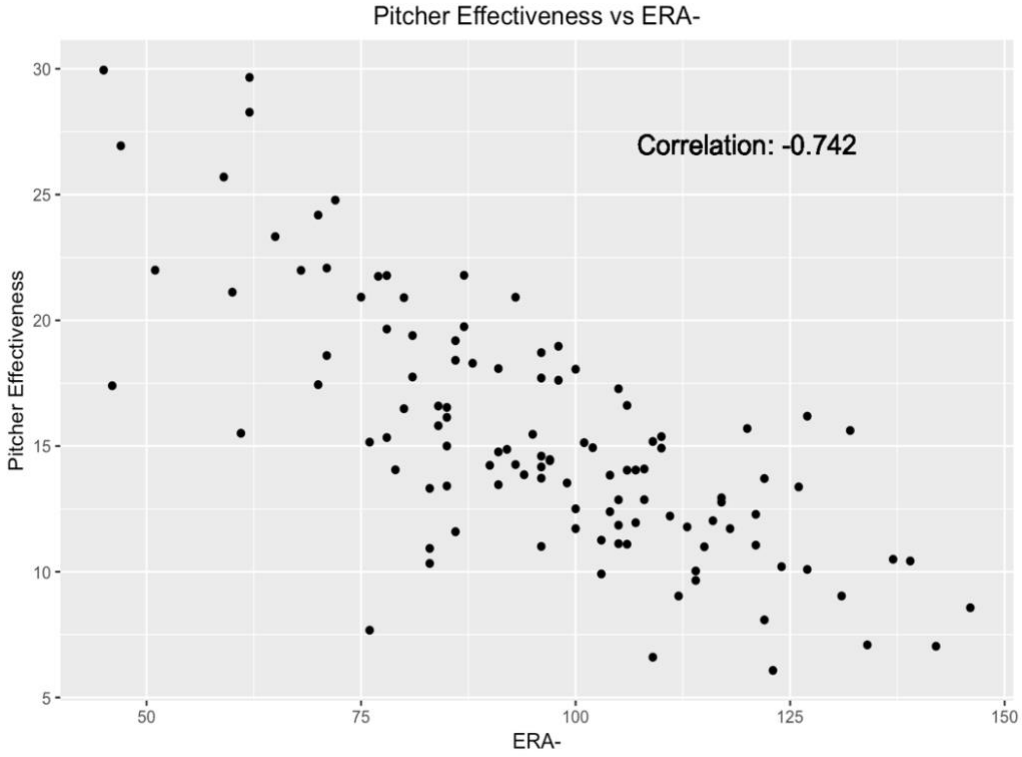


Figure 1-2: Pitcher Effectiveness versus ERA-

In 2014 Tom Tango updated these weights to correlate more with a pitcher's talent level [16]. The major difference from Bill James' formulation is starting from a score of 40, instead of 50, and taking homeruns into account. Associated weights for Tom Tango's formula are in Table 1-6. A Game Score of 50 is an average performance for a pitcher and a Game Score of 40 indicates a replacement level outing [16].

Table 1-6: Game Score Weights (Tom Tango)

Event	**Weight**
Start of game	+40
Out	+2
Strikeout	+1
Walk	-2
Hit	-2
Any Run	-3
Homerun	-6

To evaluate Pitcher Effectiveness as a metric to evaluate a starting pitcher's game performance, it was compared to both the traditional pitching line and Game Score. For both the traditional line score and Game Score, data was used from Baseball Reference where they use Bill James' formula for Game Score [17]. The worst and best pitched games according to Pitcher Effectiveness were more closely examined. For example, Dylan Bundy, on May 8th, 2018, had the lowest Pitcher Effectiveness for a game at -16.38 and the trend during the game can be seen in Figure 1-3. In this game Bundy did not record an out, while giving up 7 runs and a low Game Score of 10 with the pitching line in Table 1-7. Bundy's Pitcher Effectiveness in this outing agreed with Game Score, being the lowest of the season for him.

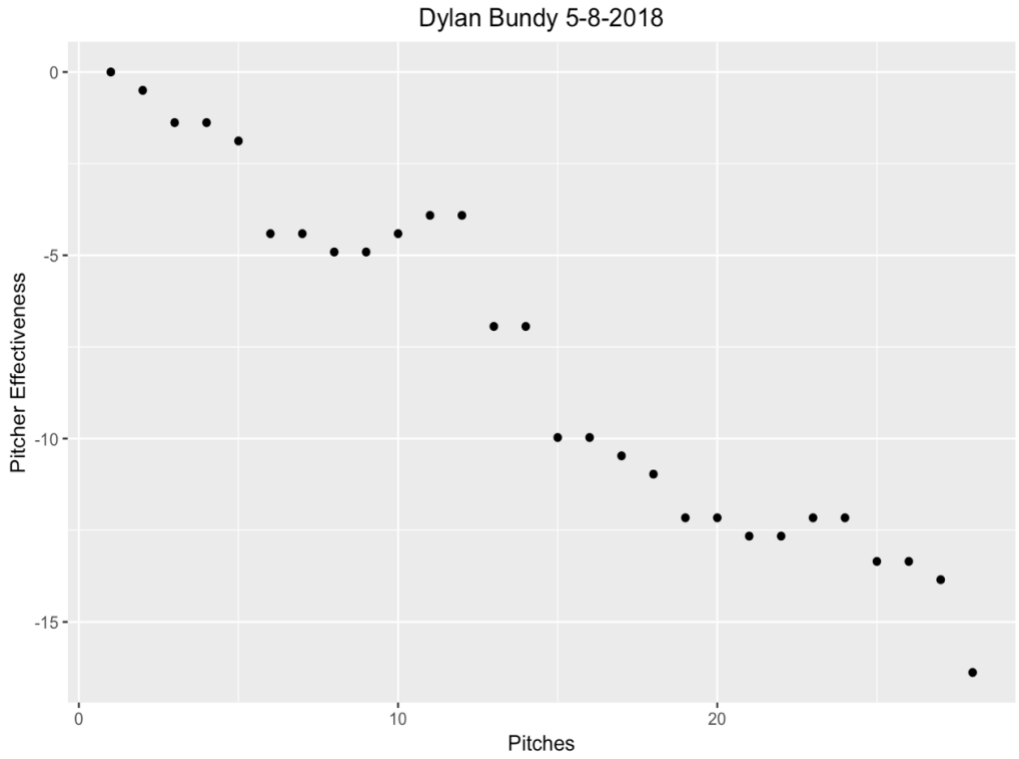


Figure 1-3: Dylan Bundy Pitcher Effectiveness (5-8-18)

Table 1-7: Dylan Bundy 5-8-18 [18]

Innings Pitched	Hits	Earned Runs	Walks	Strikeouts	Homeruns	Game Score
0+	5	7	2	0	4	10

The best Pitcher Effectiveness for the season was an outing by multiple Cy Young award winner Max Scherzer on May 30th. He pitched 8 innings giving up no runs, striking out 12, and a Game Score of 89. A small dip in Pitcher Effectiveness, in Figure 1-4: Max Scherzer Pitcher Effectiveness (5-30-18), for Scherzer at around 94 pitches was due to a double by Manny Machado and walk to Mark Trumbo in the 7th inning of the game. According to Game Score, this was not quite the best pitched game by Scherzer but rather the April 9th game with a slightly better Game Score of 93. The pitching lines were similar, as seen in Table 1-8, but Scherzer pitched a shutout and did not walk anyone. However, the difference comes from Game Score valuing innings pitched after the 4th and Pitcher Effectiveness valuing pitchers missing bats with swing a miss strikes. In

the May 30th game Scherzer had 11 of 12 strikeouts swinging while in the April 9th game he had 8 out of 10 strikeouts swinging. In addition, Game Score uses at-bat based data while Pitchers Effectiveness uses pitch level data.

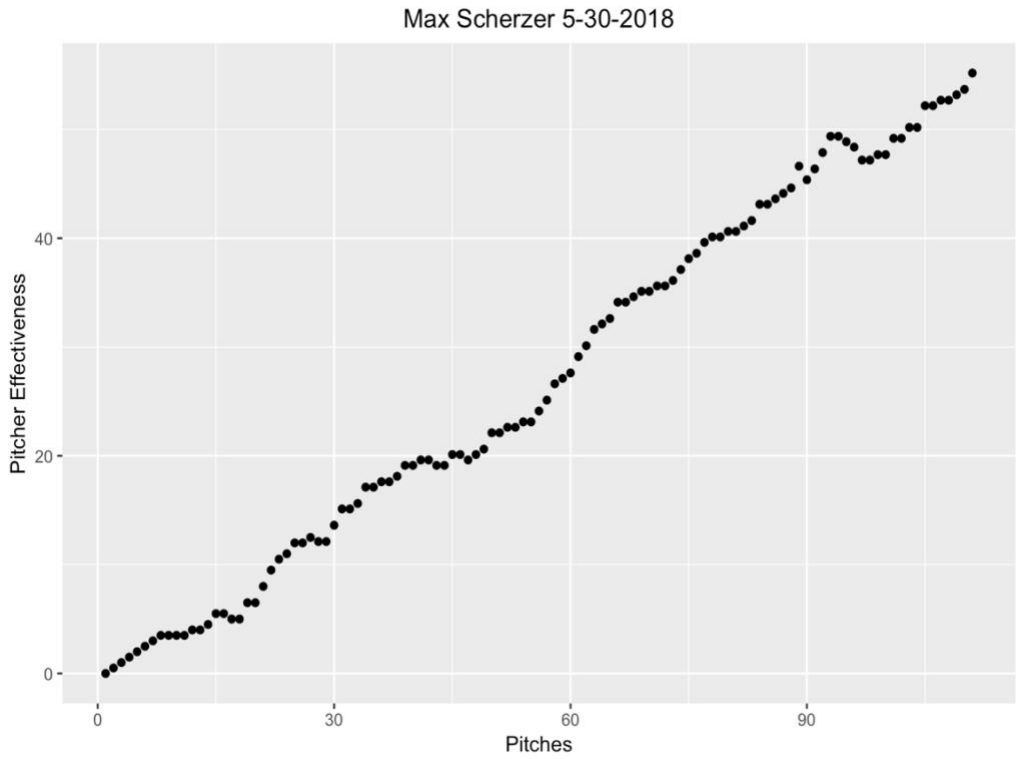


Figure 1-4: Max Scherzer Pitcher Effectiveness (5-30-18)

Table 1-8: Max Scherzer [19]

Date	Innings Pitched	Hits	Earned Runs	Walks	Strikeouts	Homeruns	Game Score
5-30-18	8	2	0	1	12	0	89
4-9-18	9	2	0	0	10	0	93

Overall for the 2018 season the top 5 in Pitcher Effectiveness were Jacob DeGrom (the 2018 NL Cy Young Award Winner), Max Scherzer, Justin Verlander, Chris Sale, and Aaron Nola in Table 1-9. All five of these pitchers are seen as elite, and their spot in the top 5 is consistent with that. It is not surprising the 2018 NL Cy Young award winner was the best, however, the 2018 AL Cy

Young award winner Blake Snell was ranked 34th with an Average Pitcher Effectiveness of 17.4. This could be due to the fact that Snell, although incredibly effective, did not pitch deep into games. Since you need pitches to accumulate Pitcher Effectiveness, the metric favored pitchers that pitched deep into the game. Most metrics indicated that Snell performed at a high level, but Pitcher Effectiveness did not put him at an elite level for the season.

Table 1-9: Average Pitcher Effectiveness per Game Top 5

Player Name	Average Pitcher Effectiveness
Jacob deGrom	29.95
Max Scherzer	29.66
Justin Verlander	28.27
Chris Sale	26.94
Aaron Nola	25.7

1.5 Conclusion

In this paper Pitcher Effectiveness has been shown to be able to predict a pitcher giving up many runs, and a viable metric for evaluating a starting pitcher's performance for a game or across a whole season. Utilizing Statcast data, the metric gives a better evaluation of pitchers using data we never had before. Coaches could look back at the game pitch by pitch to find times the Pitcher Effectiveness dropped. This is a step forward for in game analytics, that already involves scouting reports available in the dugout, and the evaluation of how effective a pitcher is.

1.6 Future Work

There are many areas of future work involving Pitcher Effectiveness. First, the weights and variables can be better optimized for better predictive power and evaluation of pitchers. A wider grid search could be done to find more optimal values. Blake Snell being ranked 34th was odd and could mean some weights are too high. Or it could be the case that some are actually too low.

Instead of looking at Pitcher Effectiveness per game, it could be adjusted per 9 innings like ERA. In addition, choosing the weights as whole numbers rather than some as fractional may be better for simplifying the metric. Statcast data is very new and the weights need further exploration. Also, both slope and number of pitchers for run prediction could be better optimized. There may be a better number of pitches to look at for both variables that would make the prediction better. From a front office perspective Pitcher Effectiveness could also be used to find pitchers that may be undervalued. This would help teams with a limited budget to build a more competitive team.

The most exciting possibility for future work is improving the utilization of Pitcher Effectiveness for use by managers. The model could be compared to a manager's decision by looking at what the model is predicting versus what the manager did, similar to the Gartheeban and Guttag analysis [11]. For example, the model could predict many runs given up in the next 10 pitches, the manager leaves the pitcher in, and the pitcher gives up runs. Pitcher Effectiveness could extend the work of Harrison and Salmon [12] by building a trend line of the average Pitcher Effectiveness per pitch for particular pitchers. Pitcher Effectiveness above the trend line could indicate good performance while under the trend line would indicate bad performance.

Next, exploring Pitcher Effectiveness for use with relief pitchers is an area of future research. The data set that was taken for this paper involved mainly traditional starting pitchers, with a few exceptions. However, getting ahead of the count, missing bats, and getting outs are important for a relief pitcher as well. One may argue that it is more important, especially in tight ballgames. A relief pitcher will pitch one to two innings maximum, in general, in a game. For in game analytics for relief pitchers the goal is to avoid giving up any runs, or few runs depending on the score of the game. With this in mind and less inning pitched by relief pitchers, a coach would want to know very quickly if they are at risk of giving up runs. This means the slope variable would have to use

less pitches since a relief pitcher may go 20-25 pitches in a game. Also, evaluating relief pitchers on their outing using Pitcher Effectiveness would be different than starting pitchers. Their ceiling for Pitcher Effectiveness is much lower than a starting pitcher who are in the game longer. Either an adjustment of giving each relief pitcher an initial Pitcher Effectiveness or separating the evaluation of a starting and relief pitcher could be solutions to this problem. Looking at how a relief pitcher's Pitcher Effectiveness is affected by the number of days off would also be interesting to explore.

Finally, Pitcher Effectiveness can be adapted to incorporate additional efficiency metrics such as quality of contact. For example, less pitches and more outs indicate the pitchers is efficient (i.e. quick outs). This could be captured using Statcasts' pitch number for at bat variable. In the last couple years pitchers have been told to elevate the fastball as hitters are adapting their launch angle for the best contact. Pitchers that have been traditionally effective lower in the strikezone must adapt to this new trend. Rewarding pitchers who induce non-quality contact could be added to Pitcher Effectiveness.

2 A Comprehensive Analysis of Pitch Sequence Effectiveness and Predictability for the 24 Baseball States

2.1 Introduction

Pitchers are involved in every single moment of a baseball game. They initiate action with every pitch that they throw. Batters are tasked with trying to hit one of thirteen types of pitches that are now thrown in Major League Baseball (MLB) [20]. With all the data now available it is easy to find what is the most used pitch by a pitcher, which can be broken down by each ball-strike count, location of pitches, and motion of each pitch [20]. Batters are trying to predict what pitch will be thrown while pitchers want to throw something a batter would not expect. This chess match between batters and pitchers involves pitch sequencing, which is the theory that pitches earlier in the at bat influence a batter's behavior later in the at bat [21].

Two important components of pitch sequencing are perceived velocity (PV) and effective velocity (EV). PV is defined as "an attempt to quantify how fast a pitch appears to a hitter, by factoring the velocity of the pitch and the release point of the pitcher" [22]. For example, if two pitchers throw a 93-mph fastball but one pitcher releases the ball closer to home plate, the perceived velocity would be higher for the pitcher who released the ball closer to home plate [22]. EV was built on the idea that "hitters swing naturally late against the fastball up-and-in, and naturally early against the off-speed pitch down-and-away, due to the bat path required to make

quality barrel contact against pitches in these locations" [21]. Knowing this, the hitters will try to adjust their swing based on the previous pitch location and speed [21], as well as the current ball-strike count. Pitchers utilize these ideas to their advantage and must become less predictable to be successful.

A paper by Joel Bock in 2015 explored the predictability of MLB pitchers and the impact of predictability on predicting performance metrics earned run average (ERA) and fielding independent pitching (FIP) [23]. Using data of MLB pitchers from 2011-2013, excluding pitchers who did not accumulate 1000 pitches, the top four most used pitches for each pitcher were identified [23]. Then, multinomial logistic regression and support vector machine (SVM) were used to predict which of those four pitches would be thrown next [23]. They obtained an overall predictability of the next pitch of 74.5% [23], which was an improvement on another paper predicting a fastball or not a fastball at 70% [24]. Further analysis was done showing the top ten most and least predictable pitchers overall, based on if the hitter was ahead, behind, or even in the count, and handedness matchups [24]. Interesting findings were Joel Hanrahan being highly predictable (95.1%) when the batter is ahead of the count and less predictable (56.1%) when the batter is behind in the count while Luke Gregorson was the opposite, being predictable when the batter is behind in the count (86.2%) and less predictable (66.6%) when the batter is ahead of the count [24]. The author did not expand on other occurrences similar to those two. In predicting ERA and FIP using the predictability of the pitcher using their model, they obtained a significant p-value at the 0.05 level but the R-Squared were very low at 0.0175 and 0.021 respectively [24]. They concluded that high predictability of the pitch sequence did not imply a

higher ERA or FIP, with many examples in the data where highly predictive pitchers has a low ERA and/or FIP or highly unpredictive pitchers that had a high ERA and/or FIP [11].

In 2013 Jon Roegele explored pitch sequences that led to strikeouts [25]. First, using data up to the All-Star Break in 2013, Roegele looked at the most common final two pitches to strike out hitters [25]. The most common final two pitches were two four-seam fastballs (2859) with two sliders as the second most common (1792) but two changeups (693) and two curveballs (609) were significantly less than sliders [25]. When Roegele explored the most common final two pitches for a strikeout on the pitcher level only four of the top fifteen did not use the same pitch back-to-back [25]. Extended to 3-pitch sequences, three four seam fastballs (1656) were the most common and three sliders (625) were the third most common. Interestingly, on the individual pitcher level, only three of the top seventeen sequences used different pitches [25]. In the article it was noted that a pitcher using their best three pitches makes sense, but pitch locations were not considered, which is a vital part of pitch sequencing [25].

In 2014 Roegele considered the effect of "back-to-back pitches from a pitcher to a hitter where each pitch is in a similar location at the swing decision point, but where the two pitches end up crossing the plane of home plate in quite different spots" [26]. By decision point Roegele explains "A 90 mph fastball takes roughly 400 milliseconds to travel between the pitcher's release point and the front of home plate. Estimates of the point where a batter must commit to start swinging range from 150-225 milliseconds before the time the pitch crosses the plate. Once a swing has started, only elite hitters are able to make further swing adjustments to the path initially started at the go/no-go decision point" [26]. Using data from 2013 and 2014 heat map matrices of swinging strike percentage (SwStr%) were built for each season where the rows were

the distance between the two consecutive pitches at the decision point and the columns were the distance between the two consecutive pitches when they crossed home plate [26]. A band of pitches from the matrix were found where SwStr% was higher based on the decision point distance and distance when crossing home plate, which led to the conclusion that "the closer consecutive pitches are to one another at the swing decision point, the less distance apart they need to arrive at home plate to generate higher than normal swing and miss rates" [26]. This would mean that the two pitches would have to be different types for the trajectory to be so different [26]. For example, the first pitch being a four-seam fastball and the second pitch being a slider, which was the most common pitch sequence in that band of pitches [26]. For all pitch types, when the pitch was in the band on the second pitch there was a higher SwgStr% [26]. It was also shown that elite starters such as Cole Hamels and Johnny Cueto pitched a lot in this band and pitchers overall saw an increase in SwStr% [26]. Thus, it was concluded that using these two pitch sequences would lead to a higher SwStr%.

In baseball there are a total of 24 combinations of base/out states that are possible [3]. For example, bases loaded with two outs. Throughout the game teams will transition from one base/out state to another when events such as hits, outs, walks, and others occur [3]. Using this idea, a run expectancy matrix can be built where for each of the 24 states there is a run expectancy attached to it [3]. The 2019 run expectancy matrix with all states is below in Table 2-1. Going further, we are able to assign run values per event, such as a homerun or strikeout [3]. The player that has to navigate all of these states is the pitcher, who must make the correct pitches to get three outs to reset to the starting state of no runners and no outs for the next inning. Based on what state the game is in, a pitcher may approach the batter in a different way. Unlike other papers that do not consider these states when looking at pitch sequencing, the goal of this

chapter is to explore differences of sequences per state and predictability of hits, runs, and strikeouts based on the pitch sequence and state the game is in.

Table 2-1: Run Expectancy Matrix (2019) [27]

Bases	0 Outs	1 Out	2 Outs
_/_/_	0.544	0.298	0.115
1B/_/_	0.935	0.564	0.242
_/2B/_	1.147	0.713	0.339
_/_/3B	1.369	0.953	0.391
1B/2B/_	1.537	0.979	0.467
1B/_/3B	1.759	1.219	0.518
_/2B/3B	1.971	1.368	0.615
1B/2B/3B	2.362	1.634	0.742

2.2 Data

All data for this analysis was pulled from BaseballSavant [20]. Twenty-four data sets were produced representing each baseballs state, which included every pitch thrown in a particular baseball state in 2019. Position players that pitched in 2019 were removed from the data sets. This resulted in a total of 731,083 pitches across the 24 data sets with 89 variables each. Variables that were used from the original 89 variables were pitch number, which was the number of pitches in the at bat, and zone, which is the zone the pitch was in. Zone comes from the definition on BaseballSavant which can be found in Figure 2-1. There were several variables that were created for the analysis. First, a pitch type variable was created where FB represented a fastball type pitch (Four-Seam Fastball, Two-Seam Fastball, Sinker, Cut-Fastball), OS represented an off-speed type pitch (Changeup, Splitfinger, Forkball, Screwball), BR represented a breaking ball type pitch (Slider, Curveball, Knuckle-Curve, Knuckleball, Eeephus), and Other represented all other types of pitches. From this, a variable was built that contained the entire

sequence of pitches where the most recent pitches were listed last. For example, FB, BR, OS is the three-pitch sequence fastball, breaking ball, off-speed (pitch 1, pitch 2, pitch 3). Then the variable Last3 was manufactured from the sequence variable where Last3 represented the last three pitches thrown in the sequence for a sequence that was three or more pitches and the whole sequences if it was only a one or two pitch sequence. For example, for the sequence FB, FB, BR, OS, BR, FB the Last3 variable would be OS, BR, FB. Then logical variables hit, out, and runs were built. These variables were used as outcome variables. All explanations of variables are in Table 2-2. In addition, the data sets were cleaned to only contain pitches that were the final pitch of an at bat, where an event occurred (i.e. Hit, walk, strikeout, etc.) resulting in 189,078 pitches.

Table 2-2: Variable Explanations

Variable	**Explanation**
Pitch Number	Number of pitches in the at bat
Zone	Zone pitch was in
Pitch Type	Type of pitch (FB, BR, OS, Other)
Sequence	Sequence of pitches in at bat
Last3	Last three pitches in sequence
Hit	Logical variable for hit
Out	Logical variable for out
Run	Logical variable for run

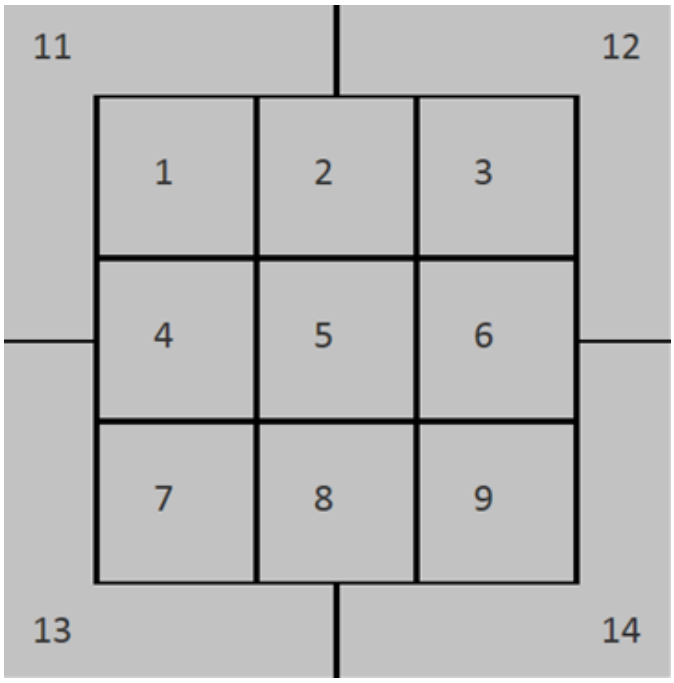


Figure 2-1: Zones [9]

An additional data set was also created for each baseball state. These data sets were frequencies of the variable Last3 for each baseball state in addition to the number of hits, outs, strikeouts, and runs for the different combinations of last three pitches. An example of these data sets is in Table 2-3, which were the top five last three pitch combinations for the state of no runners on base with no outs.

Table 2-3: Frequency Table Example

No runners on with no outs					
Last3	**Total**	**Hits**	**Outs**	**Strikeouts**	**Runs**
FB, FB, FB	8190	1641	5096	2018	291
FB	3222	1125	2003	0	249
FB, FB, BR	3091	556	2316	1154	74
FB, BR, FB	2902	600	1970	759	97
BR, FB, FB	2814	524	1849	757	93

2.3 Methods

To test for differences between sequences per state proportion tests were performed between the top 10 Last3 variables, relative to the state of no runners with no outs, that intersected between data sets. The list of the top 10 can be found in Table 2-4 below. Proportion tests for outs and strikeouts between sequences per state were performed. However, it did not make sense to compare proportions of strikeouts for the sequences FB, and FB, FB since a strikeout cannot occur in these at bats. Therefore, they were omitted in the results.

Table 2-4: Top 10 Last3 Sequences

Top 10 Last3									
FB, FB, FB	FB	FB, FB, BR	FB, BR, FB	BR, FB, FB	FB, FB	FB, BR, BR	BR, BR, FB	BR, FB, BR	FB, OS, FB

For each of the 24 baseball states three logistic regression models were built to predict a hit, out, strikeout, and run. Each model had the same predictor variables, Last3, Pitch Number, and Zone. For each model, only Last3 variables that occurred more than 20 times were considered in the analysis. For hit, out, and run prediction at bats of any number of pitches, three pitch at bats, and at bats with three or more pitches were considered. Strikeouts do not occur for at bats with one or two pitches, thus only three pitch at bats and at bats with three or more pitches were considered. To evaluate these models a 5-fold cross validated area under the ROC curve (AUC) was calculated for each. The models are summarized in Equation 1-1.

$$log - odds(Hit/Out/Run/Strikeout) \sim Last3 + Pitch\ Number + Zone$$

Equation 2-1: Pitch Sequence Outcomes Logistic Regression Model

2.4 Results

2.4.1 Proportion Tests

The results of the proportion test for outs are found below in Table 2-5.

Table 2-5: Proportions of Outs

	State	FB, FB, FB	FB	FB, FB, BR	FB, BR, FB	BR, FB, FB	FB, FB	FB, BR, BR	BR, BR, FB	BR, FB, BR	FB, OS, FB
1	_/_/_ 0 outs	0.622	0.622	0.749	0.679	0.657	0.636	0.786	0.689	0.748	0.648
2	_/_/_ 1 out	0.634	0.624	0.757	0.692	0.654	0.636	0.793	0.663	0.746	0.654
3	_/_/_ 2 outs	0.627	0.609	0.753	0.669	0.645	0.635	0.782	0.684	0.719	0.653
4	1B/_/_ 0 outs	0.61	0.594	0.69	0.615	0.6	0.573	0.709	0.621	0.69	0.613
5	1B/_/_ 1 out	0.616	0.531	0.683	0.635	0.603	0.561	0.724	0.623	0.679	0.608
6	1B/_/_ 2 outs	0.618	0.543	0.682	0.629	0.599	0.534	0.708	0.592	0.645	0.583
7	_/2B/_ 0 outs	0.622	0.577	0.688	0.665	0.645	0.588	0.647	0.689	0.675	0.692
8	_/2B/_ 1 out	0.595	0.55	0.645	0.634	0.579	0.594	0.676	0.578	0.678	0.577
9	_/2B/_ 2 outs	0.591	0.63	0.701	0.629	0.593	0.571	0.709	0.674	0.655	0.554
10	_/_/3B 0 outs	0.573	0.556	0.724	0.593	0.636	0.684	0.619	0.562	0.765	0.889
11	_/_/3B 1 out	0.586	0.652	0.683	0.535	0.716	0.6	0.792	0.551	0.691	0.471

12	_/_/3B 2 outs	0.632	0.651	0.767	0.696	0.577	0.596	0.726	0.565	0.718	0.5
13	1B/2B/_ 0 outs	0.644	0.646	0.777	0.659	0.592	0.625	0.702	0.694	0.758	0.606
14	1B/2B/_ 1 out	0.585	0.575	0.726	0.7	0.569	0.726	0.755	0.63	0.69	0.631
15	1B/2B/_ 2 outs	0.634	0.614	0.714	0.667	0.575	0.617	0.737	0.686	0.769	0.689
16	1B/_/3B 0 outs	0.575	0.561	0.706	0.623	0.6	0.621	0.722	0.567	0.71	0.478
17	1B/_/3B 1 out	0.662	0.556	0.58	0.563	0.53	0.578	0.75	0.614	0.688	0.615
18	1B/_/3B 2 outs	0.609	0.44	0.636	0.539	0.532	0.487	0.667	0.613	0.61	0.617
19	_/2B/3B 0 outs	0.5	0.8	0.656	0.628	0.556	0.739	0.667	0.517	0.737	0.562
20	_/2B/3B 1 out	0.617	0.527	0.565	0.565	0.563	0.576	0.645	0.612	0.567	0.548
21	_/2B/3B 2 outs	0.617	0.566	0.68	0.673	0.634	0.621	0.66	0.678	0.6	0.565
22	1B/2B/3B 0 outs	0.573	0.654	0.784	0.639	0.553	0.585	0.75	0.731	0.714	0.571
23	1B/2B/3B 1 out	0.67	0.57	0.763	0.679	0.567	0.612	0.776	0.627	0.797	0.867
24	1B/2B/3B 2 outs	0.629	0.544	0.778	0.643	0.61	0.644	0.893	0.722	0.745	0.653
	P-Value	0.08	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05

All of the p-values were significant at the 0.05 level, with the exception of the Last3 of FB, FB, FB, meaning we rejected the null hypothesis that the proportion of outs for these Last3 pitch sequences are the same across states. This would suggest that for these Last3 there was a difference in the proportion of outs for at least one state compared to another. Taking a closer look at these values, line plots were built in Figure 2-2 to show the profiles of the different Last3 variables across the 24 different states.

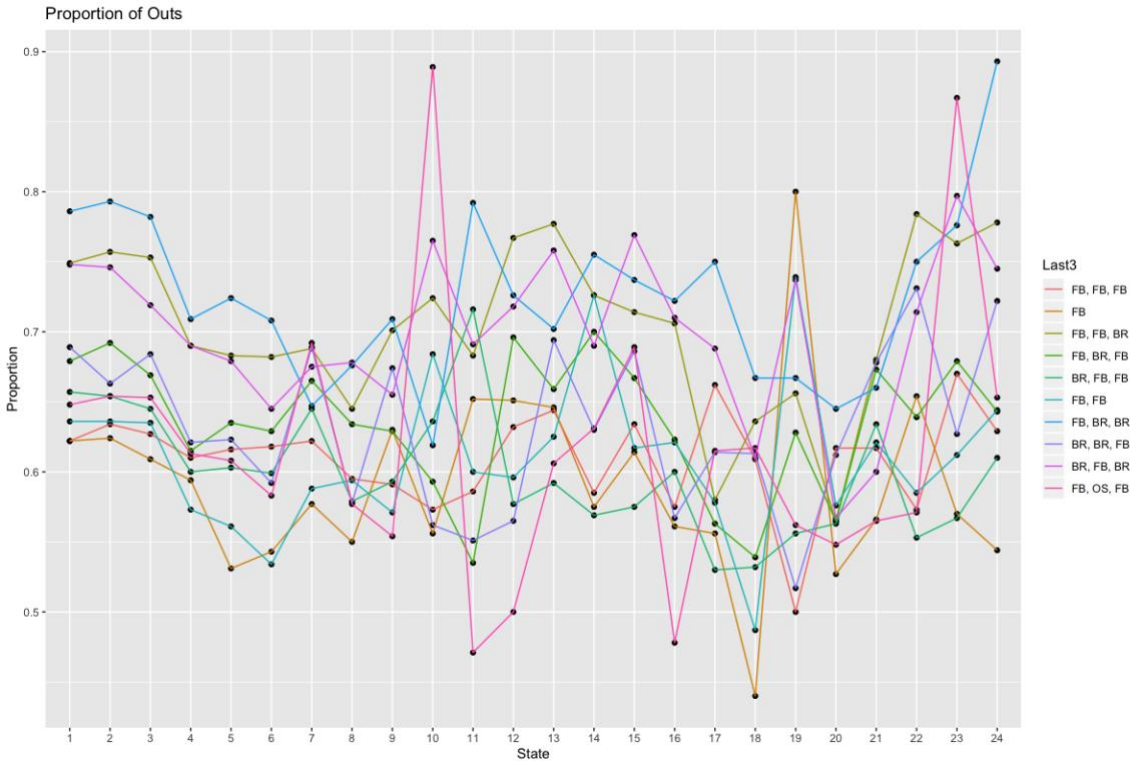


Figure 2-2: Profiles for Proportion of Outs

A few observations that were interesting were the differences in the proportion of outs for certain states. First, looking at FB there was a large difference between state 18 (1B/_3B 2 outs) and state 19 (_/2B/3B 0 outs). The proportion of outs were 0.44 (n = 257) and (n=35) 0.8, respectively. The proportion test between the two resulted in a p-value of 0.00013, which rejected the null hypothesis that the proportions were the same. The other Last3, FB, OS, FB, showed a large difference between state 10 (_/_/3B 0 outs) and state 11 (_/_/_3B 1 out). The proportion of outs were 0.889 (n =9) and 0.471 (n =34) which suggest high variance because the sample sizes were small. The proportion test between the two resulted in a p-value of 0.06, which failed to reject the null hypothesis that the proportions were the same. State 10 is the least populated since it occurs the least in games. Lastly, a Last3 of FB, FB, BR and FB, BR, BR had a proportion of outs of 0.726 and 0.757 respectively across states compared to a proportion of 0.652 of all ten Last3 sequences across states. These two sequences' proportion of outs achieved p-values well below 0.05 (< 2.2e-

16), when compared to the overall proportion of outs which means there is a significant difference between proportions. This means that setting up with a fastball then finishing with a breaking ball was more effective.

Next, the results of the proportion test for strikeouts are found below in Table 2-6.

Table 2-6: Proportions of Strikeouts

	State	**FB, FB, FB**	**FB, FB, BR**	**FB, BR, FB**	**BR, FB, FB**	**FB, BR, BR**	**BR, BR, FB**	**BR, FB, BR**	**FB, OS, FB**
1	_/_/_ 0 outs	0.246	0.373	0.262	0.269	0.427	0.296	0.362	0.229
2	_/_/_ 1 out	0.271	0.408	0.289	0.294	0.451	0.282	0.384	0.235
3	_/_/_ 2 outs	0.262	0.42	0.29	0.293	0.455	0.291	0.378	0.264
4	1B/_/_ 0 outs	0.224	0.337	0.237	0.247	0.366	0.297	0.338	0.199
5	1B/_/_ 1 out	0.229	0.365	0.258	0.245	0.389	0.252	0.312	0.199
6	1B/_/_ 2 outs	0.249	0.386	0.248	0.257	0.401	0.282	0.354	0.188
7	_/2B/_ 0 outs	0.256	0.335	0.225	0.244	0.36	0.295	0.292	0.231
8	_/2B/_ 1 out	0.266	0.34	0.283	0.209	0.426	0.291	0.36	0.201
9	_/2B/_ 2 outs	0.28	0.363	0.281	0.244	0.425	0.295	0.333	0.211
10	_/_/3B 0 outs	0.253	0.345	0.259	0.227	0.286	0.125	0.412	0.333
11	_/_/3B 1 out	0.234	0.337	0.221	0.275	0.403	0.217	0.272	0.118
12	_/_/3B 2 outs	0.278	0.455	0.338	0.289	0.41	0.226	0.382	0.172

13	1B/2B/_ 0 outs	0.245	0.44	0.222	0.2	0.375	0.241	0.363	0.197
14	1B/2B/_ 1 out	0.227	0.389	0.233	0.252	0.408	0.261	0.329	0.246
15	1B/2B/_ 2 outs	0.283	0.402	0.299	0.256	0.441	0.302	0.366	0.262
16	1B/_/3B 0 outs	0.244	0.471	0.151	0.2	0.25	0.3	0.355	0.174
17	1B/_/3B 1 out	0.262	0.277	0.23	0.252	0.382	0.205	0.312	0.192
18	1B/_/3B 2 outs	0.278	0.359	0.227	0.248	0.301	0.301	0.288	0.3
19	_/2B/3B 0 outs	0.227	0.344	0.163	0.194	0.333	0.241	0.395	0.25
20	_/2B/3B 1 out	0.184	0.341	0.271	0.241	0.329	0.239	0.328	0.258
21	_/2B/3B 2 outs	0.31	0.424	0.347	0.28	0.43	0.311	0.291	0.217
22	1B/2B/3B 0 outs	0.226	0.486	0.278	0.237	0.25	0.308	0.4	0.214
23	1B/2B/3B 1 out	0.275	0.392	0.202	0.24	0.403	0.237	0.459	0.2
24	1B/2B/3B 2 outs	0.263	0.484	0.252	0.247	0.507	0.344	0.351	0.204
	P-Value	<0.05	<0.05	<0.05	0.117	<0.05	0.734	0.214	0.480

Contrary to the proportion of outs, only four Last3 were significant at the 0.05 level. This would suggest that for these Last3 there was a difference in the proportion of strikeouts for at least one state compared to another. Similar to outs, plots were built, in Figure 2-3, to show the profiles of the proportion of strikeouts of the different Last3 across the 24 different states.

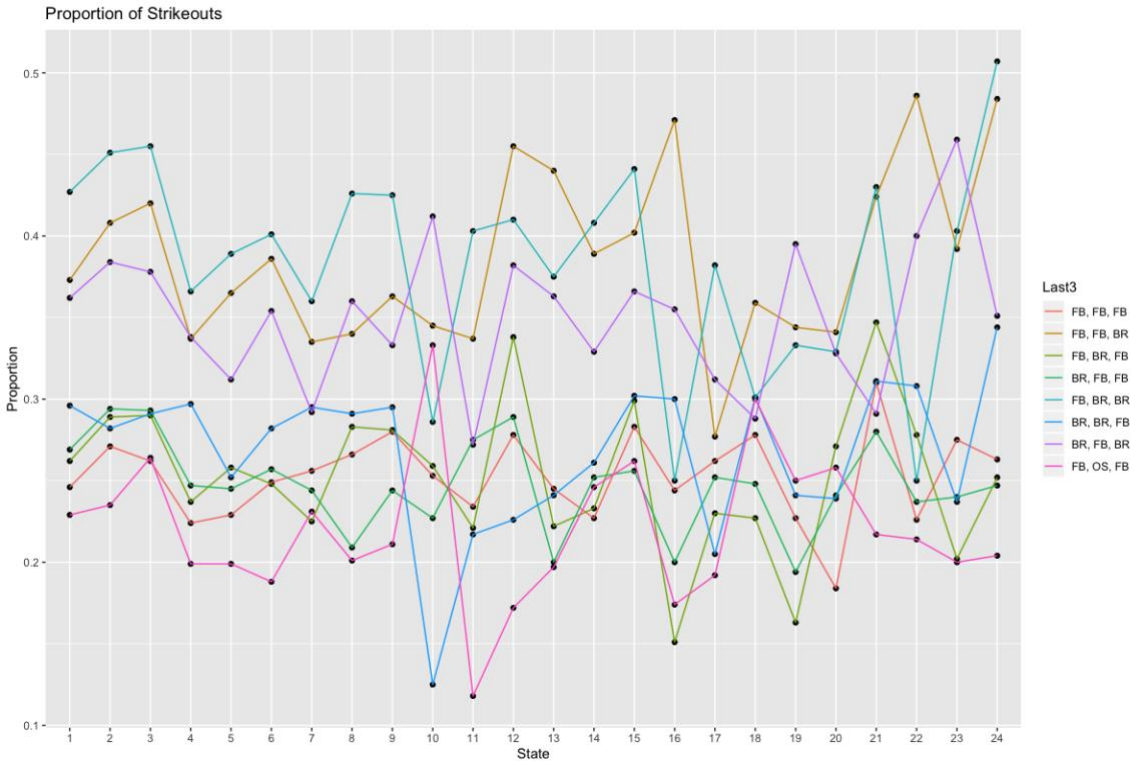


Figure 2-3: Profiles for Proportion of Strikeouts

Three Last3 sequences stood out, FB, FB, BR; FB, BR, BR; and BR, FB, BR as their profiles seemed higher than the others. They had proportions of strikeouts of 0.387, 0.422, and 0.358 respectively across states for the proportion of strikeouts compared to a proportion of strikeouts 0.3 of all eight Last3 sequences across states. These three sequences' proportion of strikeouts achieved p-values well below 0.05 (< 2.2e-16), when compared to the overall proportion of outs which means there is a significant difference between proportions. This suggested that finishing with a breaking ball yielded more strikeouts after being set up by a fastball.

2.4.2 Logistic Regression Models

The AUC for each model was calculated with at bats of any number of pitches (Table 2-7), at bats with three pitches (Table 2-8), and at bats with three or more pitches (Table 2-9) are shown below.

Table 2-7: Logistic Regression Results for Any Number of Pitches for At Bat

State	Hit AUC	Out AUC	Run AUC
_/_/_ 0 outs	0.707	0.605	0.656
_/_/_ 1 out	0.696	0.605	0.664
_/_/_ 2 outs	0.683	0.603	0.664
1B/_/_ 0 outs	0.684	0.595	0.645
1B/_/_ 1 out	0.682	0.610	0.664
1B/_/_ 2 outs	0.684	0.621	0.655
_/2B/_ 0 outs	0.646	0.603	0.651
_/2B/_ 1 out	0.671	0.618	0.649
_/2B/_ 2 outs	0.666	0.615	0.667
_/_/3B 0 outs	0.666	0.470	0.513
_/_/3B 1 out	0.709	0.545	0.653
_/_/3B 2 outs	0.683	0.622	0.674
1B/2B/_ 0 outs	0.643	0.607	0.644
1B/2B/_ 1 out	0.656	0.615	0.661
1B/2B/_ 2 outs	0.664	0.607	0.666
1B/_/3B 0 outs	0.731	0.490	0.588
1B/_/3B 1 out	0.675	0.599	0.643
1B/_/3B 2 outs	0.682	0.658	0.685
_/2B/3B 0 outs	0.676	0.525	0.566
_/2B/3B 1 out	0.733	0.595	0.671
_/2B/3B 2 outs	0.675	0.592	0.681
1B/2B/3B 0 outs	0.628	0.564	0.607
1B/2B/3B 1 out	0.614	0.617	0.657
1B/2B/3B 2 outs	0.596	0.607	0.685

Table 2-8: Logistic Regression Results for Three Pitch At Bats

State	Hit AUC	Out AUC	Run AUC	Strikeout AUC
_/_/_ 0 outs	0.635	0.571	0.593	0.724
_/_/_ 1 out	0.649	0.572	0.598	0.726
_/_/_ 2 outs	0.599	0.571	0.598	0.732
1B/_/_ 0 outs	0.626	0.520	0.606	0.668

1B/_/_ 1 out	0.628	0.525	0.618	0.674
1B/_/_ 2 outs	0.644	0.523	0.644	0.668
_/2B/_ 0 outs	0.487	0.522	0.563	0.603
_/2B/_ 1 out	0.589	0.521	0.576	0.697
_/2B/_ 2 outs	0.631	0.525	0.632	0.665
_/_/3B 0 outs	0.637	0.700	0.693	0.614
_/_/3B 1 out	0.648	0.468	0.500	0.522
_/_/3B 2 outs	0.585	0.486	0.529	0.639
1B/2B/_ 0 outs	0.593	0.641	0.648	0.727
1B/2B/_ 1 out	0.522	0.514	0.599	0.656
1B/2B/_ 2 outs	0.581	0.522	0.568	0.674
1B/_/3B 0 outs	0.631	0.493	0.468	0.576
1B/_/3B 1 out	0.603	0.558	0.624	0.473
1B/_/3B 2 outs	0.667	0.576	0.663	0.677
_/2B/3B 0 outs	0.736	0.503	0.607	0.667
_/2B/3B 1 out	0.651	0.419	0.612	0.600
_/2B/3B 2 outs	0.498	0.549	0.527	0.608
1B/2B/3B 0 outs	0.494	0.398	0.457	0.606
1B/2B/3B 1 out	0.639	0.613	0.564	0.725
1B/2B/3B 2 outs	0.623	0.570	0.5904	0.731

Table 2-9: Logistic Regression Results for Three or More Pitch At Bats

State	**Hit AUC**	**Out AUC**	**Run AUC**	**Strikeout AUC**
_/_/_ 0 outs	0.712	0.615	0.654	0.648
_/_/_ 1 out	0.705	0.613	0.659	0.634
_/_/_ 2 outs	0.684	0.614	0.665	0.636
1B/_/_ 0 outs	0.694	0.583	0.640	0.617
1B/_/_ 1 out	0.699	0.596	0.671	0.634
1B/_/_ 2 outs	0.697	0.605	0.662	0.626
_/2B/_ 0 outs	0.625	0.611	0.625	0.612
_/2B/_ 1 out	0.657	0.624	0.652	0.623
_/2B/_ 2 outs	0.684	0.624	0.676	0.607
_/_/3B 0 outs	0.600	0.452	0.482	0.537

_/_/3B 1 out	0.712	0.590	0.654	0.591
_/_/3B 2 outs	0.678	0.633	0.679	0.626
1B/2B/_ 0 outs	0.662	0.605	0.649	0.639
1B/2B/_ 1 out	0.663	0.602	0.665	0.633
1B/2B/_ 2 outs	0.664	0.607	0.665	0.612
1B/_/3B 0 outs	0.735	0.510	0.580	0.673
1B/_/3B 1 out	0.670	0.593	0.649	0.556
1B/_/3B 2 outs	0.665	0.644	0.677	0.600
_/2B/3B 0 outs	0.688	0.575	0.672	0.579
_/2B/3B 1 out	0.705	0.571	0.649	0.593
_/2B/3B 2 outs	0.634	0.615	0.630	0.580
1B/2B/3B 0 outs	0.554	0.544	0.559	0.655
1B/2B/3B 1 out	0.599	0.623	0.679	0.644
1B/2B/3B 2 outs	0.624	0.631	0.701	0.655

Overall, logistic regression models for out and run did not perform well with an AUC less than 0.7 except for when looking at a runner on 3rd base no outs for three pitch at bats and bases loaded two outs for at bats with three or more pitches for out and run models respectively. There were better results with the logistic models for hit and strikeout. The best performing models for strikeout occurred when looking at just three pitch sequences with several AUC values above 0.7. For hit, there were several states in all three tables where the AUC was over 0.7. Interestingly, the AUC for at bats with three or more pitches were above 0.7 for the same states as at bats with any number of pitches and outperforming most of them.

2.5 Conclusion

Predictability of a hit, run, out, and strikeout did not perform well but showed some promise for certain number of pitches in an at bat. Proportion tests showed that there are differences for certain Last3 sequences in the proportion of outs and strikeouts for certain states. In addition, there could

be more differences in effectiveness between Last3 sequences of certain states that resulted in more outs and strikeouts. Further profile analyses are needed to explore this idea.

2.6 Future Work

The biggest area of future work would be a profile analysis of these sequences in two different ways. First, looking closer at these profiles as in this paper but adding more sequences to see where there is an actual difference. Determining the exact sequences with significant differences could impact which sequence is preferable for getting an out. Second, looking at profiles of sequences in each state individually. In key states, such as those that involve the bases loaded, it is important to get an out, with a strikeout being preferable. Thus, determining the best performing sequence for particular states may be the difference of giving up a run or getting out of an inning with a strikeout. These profile analyses could yield information on which sequences are really making a difference in getting outs and strikeouts.

Another area of future work is performing a similar predictive analysis of these sequences as a whole, rather than separated into particular states. This would fix the problem of sparse data for certain states. Models could instead use the particular state as a predictor variable for a hit, out, run, or strikeout and may lead to better predictability. This would be a more holistic analysis, to get an overall sense of which sequences are working for pitches and which sequences are ineffective but a broader analysis than the one done in this paper.

Separating relief pitchers and starting pitchers is another area that could be explored. Many relief pitchers are summoned into the game in a non-favorable state, one with runners on base, that they must navigate. This makes the strikeout much more important for a relief pitcher. Looking at the profiles for relief pitchers alone in key states would help to identify what is more effective

for them. A starting pitcher throws many more pitches than a relief pitcher, which means that the hitters are able to observe the common sequences being used. It would make sense to use number of pitches thrown for a model with starting pitchers. This difference between the relief pitchers and starting pitchers may paint a different picture than this analysis.

Finally, looking at profiles of just at bats that were two or three pitches could be helpful. Instead of just looking at Last3, restricting to just two or three pitch sequences are easier to analyze and compare but this may cause problems for the amount of data per state. In particular, analyzing three pitch sequences for strikeouts, which are rare, would lead to sparseness in the data sets. However, when looking at outs it would be helpful to see which sequences lead to quicker outs. The lower the number of pitchers per at bat for a pitcher the longer the pitcher can stay in a game, which has a significant repercussion on a team. The health and effectiveness of pitchers is determined by the amount they are used over the course of the season. For a starting pitcher, pitching more innings will help the bullpen avoid being overused. For a relief pitcher, quick outs lead them to be available to pitch again earlier and the team could avoid using another relief pitcher right away.

3 Forecasting wOBA Using High Accuracy Statistical and Machine Learning Algorithms

3.1 Introduction

From fantasy baseball to Major League Baseball (MLB) front offices, projections of players' future performance are important and are explored quite often. Projections of future performance have gotten better over the years as more data has become easily accessible through websites such as Fangraphs [4], Baseball-Reference [17], Retrosheet [28], and Baseball Savant [20]. Fangraphs [4] and Baseball-Reference [17] both offer traditional and Sabermetric [7] statistics for all players and teams while Retrosheet [28] offers play-by-play information for every game. In 2015 Statcast was introduced and described as "a state-of-the-art tracking technology that allows for the collection and analysis of a massive amount of baseball data, in ways that were never possible in the past. Statcast can be considered the next step in the evolution of how we consume and think about the sport of baseball that began over a decade ago, when Major League Baseball Advanced Media installed pitch tracking hardware in each Major League stadium. That was a step that unlocked a new age of baseball fandom, and Statcast builds upon that innovation by measuring everything the previous system did, along with a great deal more" [29]. This new data became easily accessible via Baseball Savant [20], which allows for more accurate projections. There has been an abundance of research on the projection of future player performance using a variety of different methodologies.

Player Empirical Comparison and Test Algorithm, known as PECOTA, has the reputation of being the most accurate at predicting player performance [30]. PECOTA uniquely uses a player's past performance and fits it to a comparable MLB player using nearest neighbor analysis [31]. From this, PECOTA develops a probability distribution for several metrics such as homeruns, strike outs, walks, batting average, among others, for the player's performance in the next several years [31]. This gives a level of confidence PECOTA has in their projection rather than just a point estimate [31]. Another well-known and highly regarded projection system is Steamer, which is updated on Fangraphs [4]. Steamer uses both past performance and age curves to make their projections [4]. For pitchers, Steamer uses pitch-tracking data to make their projections for future seasons [4].

In 2007 Arlo Lyle combined multiple machine learning techniques to improve the accuracy of future performance of MLB players [32]. Lyle used the machine learning techniques of model trees, artificially neural networks (NN), and support vector machines (SVM) with three ensemble learning techniques that included bagging, boosting, and stacking [32]. At bats, runs, hits, doubles, triples, homeruns, on base percentage (OBP), age, and season were used as inputs to predict runs, hits, doubles, triples, homeruns, and runs batted in (RBI) using the various techniques [32]. The data that was used for training included 2,151 player-seasons from 1973-2005 and the testing data was 330 observation from the 2006 season [32]. The best results from methods employed by Lyle only outperformed PECOTA in predicting triples, but outperformed other projection systems Marcels and ZiPS (as cited in Lyle) for MAE, RMSE, and R-Squared for nearly every metric predicted [32]. Looking closer at triples, where Lyle's method beat PECOTA, Lyle's MAE was 1.27 compared to 1.39 for PECOTA, RMSE was 1.86 compared to 1.97 for PECOTA, and finally R-Squared was 0.706 compared to 0.695 for PECOTA. Although

PECOTA was superior overall, results showed that Lyle's methods were close to PECOTA's [32].

In 2009 Jensen, McShance and Wyner, used Hierarchical Bayesian Models to project a MLB player's hitting performance [33]. Their outcome variable was homerun total while their predictor variables were at bats, age, home ballpark, and position (not including pitchers) and the data used for analysis were 10,280 player-season totals from 1990 to 2005 [33]. A hierarchical model was built to predict homeruns with the homeruns outcome variable being binomially distributed with the number of at bats as opportunities and year specific homerun rate as the probability [33]. Then, the log odds of a player's unobserved homerun rate for the year was modeled as a function of home ballpark, position, age and elite status as defined by the authors [33]. The authors full model was tested on 559 players from the 2006 season which resulted in a root mean square error (RMSE) of 5.3 of the predicted means and their 80% confident intervals contained 85.5% of their predicted data [33]. In comparing their model to PECOTA [31] for the top 118 homerun hitters in 2006, the authors' model resulted in a RMSE of 7.33, and mean absolute error (MAE) of 4.4 for all of those players compared to a RMSE of 7.11 and MAE of 4.68 for PECOTA [33]. For young players, players 26 years or younger, in the data set the author's model outperformed PECOTA with a 2.62 RMSE and 1.93 MAE compared to a 4.62 RMSE and 3.44 MAE from PECOTA [33]. However, for older players, those older than 36, PECOTA projections resulted in a RMSE of 7.26 and MAE of 4.79 compared to the author's model that resulted in a RMSE of 7.56 and MAE of 4.48 [33]. The authors noted that the larger RMSE "suggest that our model might be making large errors on a small number of players" [33].

In 2014 Mushimie Lona Panda used penalized linear regression models, which included Lasso, Elastic Net, and Smoothly Clipped Absolute Deviation (SCAD) to predict which metrics are best for measuring a player's talent [34]. In this paper, Panda explores five defensive metrics with 5,585 player-seasons spanning the 1973 to 2012 seasons and forty-five offensive metrics with 7,429 player-seasons for the 2002 to 2012 seasons, since fourteen of the forty-five offensive metrics were unavailable until 2002 [34]. The goal was to determine which metrics are high or low signal to determine which metrics were more informative and had a higher predictive power [34]. The results showed that seven offensive metrics stood out, which were hits, runs, walks, runs batted in, singles, strikeouts, and weighted runs created with a high mean and high signal [34]. Panda notes that this set of metrics "provides a substantial reduction in the dimensionality for hitting metrics" [34]. For defensive metrics two out of the five were identified, which were assists and putouts that had a high signal and mean as well [34].

In 2015 Daniel Herrlin used a Bayesian approach to forecast MLB performance for optimizing a fantasy baseball draft [35]. Herrlin did this for both hitters and pitchers, where outcomes of at bats from the respective point of views [35]. Outcomes of at bats that were modeled by Herrlin were outs, walks, singles, doubles, triples, homeruns, and stolen bases [35]. The Bayesian model for hitters utilized a Dirichlet prior with multinomial data with a quadratic age curve used for accounting for the change over time [35]. After a posterior distribution for the next season was built based on a player's skill, seasons were simulated [35]. Distributions were developed for runs scores, homeruns, RBI, slugging percentage (SLG), batting average, and stolen bases [35]. A z-score was built for each category and the average determined the rankings for players [35]. With this, the model's ranking was compared to the actual ranking, which was the rank based on the actual average of the players performance in the categories, Roto World rankings, and Athlon

Sports rankings [35]. Then, based on all of these rankings and actual performance Herrlin was able to value each player [35]. For pitchers, it was a similar process but the categories used to rank pitchers was wins, strikeouts, earned run average (ERA), and walks and hits per inning pitcher (WHIP) [35]. Herrlin then went on to develop algorithm based on the results of the Bayesian models that were developed for rankings [35].

The purpose of this chapter was to explore and contrast different methods for projecting future weighted on base average (wOBA) [3] with the new Statcast [29] data. wOBA is a better way of evaluating a hitter because it weights outcome based on run value [3]. For example, homeruns are weighted more than a single because the expected runs are higher after a homerun is hit rather than a single [3]. Although it is possible to have a wOBA above 1, across a whole season with sufficient at bats the range of wOBA is between 0 and 1. Several methods were explored for predicting a hitter's next season's numerical wOBA and manufactured factor wOBA.

3.2 Data

The data that was used was hitter data from 2013-2019 for player-seasons where a player accumulated 200 or more plate appearances. This resulted in 2474 observations of player-seasons. Data sets were combined from Fangraphs [4] and Baseball Savant [20] where 45 variables were taken from Fangraphs and 4 variables were taken from Baseball Savant. The four variables from Baseball Savant [20] were expected weighted on base average (xwOBA), expected batting average (xBA), average exit velocity (launch_speed), and average launch angle (launch_angle). Expected values are based on the quality of contact of a baseball that is hit rather than the actual outcome that is influenced by the quality of the opponent's defense and defensive alignment [20]. Since Statcast [29] was unavailable until 2015, "NA" values were placeholders

for data from the 2013 and 2014 seasons. The average wOBA varies from season to season but has stayed around 0.320. The wOBA.factor variable was manufactured as a factor variable where a wOBA of 0.320 and above was grouped as "Above Average" and below 0.320 was grouped as "Below Average". Since the group of exactly average players was scarce, it was decided to combine these players into a group with above average players. In addition, splitting players into more groups made data scarcer, leading to less training data per group. The variables wOBA.next and wOBA.next.factor represented the wOBA and wOBA.factor for the next season. For example, if an observation was for the 2013 season, these variables would be 2014 season values. If a player did not get over 200 plate appearances the following year, "NA" values represented these variables. The description of these variables, as well as the others, can be found in Table 3-1.

Table 3-1: Forecasting wOBA Variable Descriptions

Variable	**Description**
Name	Player Name
Season	Season
HR	Number of Homeruns
R	Number of Runs Scored
RBI	Run Batted In
BB%	Walk Percentage
K%	Strikeout Percentage
ISO	Isolated Power
BABIP	Batting Average on Balls In Play
AVG	Batting Average
OBP	On Base Percentage
SLG	Slugging Percentage
wOBA	Weighted On Base Average
wRC+	Weighted Runs Created Plus
Off	Offensive Runs Above Average
WAR	Wins Above Replacement

Age	Player's Age
GB%	Ground Ball Percentage
FB%	Flyball Percentage
HR/FB	Homeruns per Flyball
WPA	Win Probability Added
RE24	Run Expectancy based on 24 base-out state
O.Swing%	Swing Percentage Outside of Strikezone
Z.Swing%	Swing Percentage in Strikezone
Swing%	Swing Percentage
O.Contact%	Contact Percentage Outside of Strikezone
Z.Contact%	Contact Percentage Inside Strikezone
Contact%	Contact Percentage
Zone%	Percentage of Pitches seen in Strikezone
SwStr%	Swinging Strike
Pull%	Pull Percentage
Cent%	Center Percentage
Oppo%	Opposite Field Percentage
Soft%	Soft Contact Percentage
Med%	Medium Contact Percentage
Hard%	Hard Contact Percentage
Bat	Batting Runs Above Average
RAR	Runs Above Average
AVG+	Batting Average Adjusted for park factors with 100 as average
BB%+	Walk Percentage Adjusted for park factors with 100 as average
K%+	Strikeout Percentage Adjusted for park factors with 100 as average
OBP%+	On Base Percentage Adjusted for park factors with 100 as average
SLG+	Slugging Percentage Adjusted for park factors with 100 as average
ISO+	Isolated Power Adjusted for park factors with 100 as average
BABIP+	Batting Average on Balls In Play Adjusted for park factors with 100 as average

xwOBA	Expected Weighted On Base Average
xBA	Expected Batting Average
launch_speed	Exit Velocity
launch_angle	Launch Angle
wOBA.factor	wOBA as factor (Split at 0.320)
wOBA.next	wOBA for next season
wOBA.next.factor	wOBA factor for next season

After the data was combined the numeric variables were weighted averages per season, with the observation of a particular season weighted with the previous two seasons. For example, for an observation of a player from 2016, the numeric variables were weighted with the player's 2015 and 2014 seasons. The weighted averages were explored though a grid search optimizing predictability, where the weighting for each season was explored for integer values between 1 and 5. This resulted in the current season being weighted by 5 and the previous seasons not being weighted at all (i.e. weighted by 1). If one, or both of the previous two seasons did not exist in the data set, which could be due to injuries or a rookie season for a player, the weighted average would be calculated from the existing seasons.

Once the weighting was complete, the data was subset to included just the 2015-2018 seasons without observations with "NA" values. This was done to include just seasons that included Statcast [29] data. The 2019 season was not used to build predictive models as the 2020 season had not happened to have truth data for outcome variables but was used to project 2020 wOBA for players. This resulted in 1071 observations of player-seasons for 424 different players for the 52 variables. Also, the 2019 data was separated from the other seasons. In the 2019 data set there were 359 players.

3.3 Methods

The data was first split into an 80/20 train/test split which led to 857 training observations and 214 observations in the testing set. Methods that were explored for numerically predicting wOBA were linear regression, ridge regression, elastic net regression, extreme gradient boosting, and a neural network. For predicting wOBA as a manufactured factor variable the methods used were logistic regression, stochastic gradient boosting, extreme gradient boosting, and a neural network. Several models and neural network structures were explored to obtain the best accuracy. For the models, all variables, except Name and Season and outcome variables, were in the model with the addition of a quadratic age variable and four interaction terms as shown in Equation 3-1. Performance trends tend to follow a quadratic path [4] with age and the interaction terms were explored with domain knowledge to increase accuracy. All models were built using the training data with repeated (five times) 5-fold cross-validation and scored on the testing set.

$$\mathrm{w}OBA.next/wOBA.next.factor \sim .+Age^2 + launch\_speed * launch\_angle * K\% * O.Swing\% + wOBA * (wRC+) * (ISO+) + GB\% * FB\% * HR/FB + Pull\% * Hard\%$$

Equation 3-1: Projecting wOBA and wOBA factor

For the neural networks, a grid search was performed for the best one-layer neural network from 1 to 48 nodes, the best two-layer neural network with 1 to 48 nodes in the first layer and 1 to 30 in the second layer, and the best three-layer neural network with 1 to 20 nodes in the first layer, 1 to 10 in the second layer, and 1 to 5 in the third layer. The best neural network for predicting wOBA was a one-layer neural network with 42 nodes with a rectified linear (relu) activation with dropout (0.5 rate) and a 1 node output layer with linear activation. The best neural network for predicting the wOBA factor variable was a two-layer neural network with 6 and 10 nodes respectively with

relu activation and dropout (0.5) and a 2 nodes output layer with softmax. Neural networks were fit on the training data with 100 epochs and a batch size of 25 using the Adam Optimizer. Without the amount of training observations needed, more epochs and changing batch size did not have a positive effect on the accuracy. Both neural networks were scored on the testing set. Due to computation time, deeper architectures and more nodes for the two-layer and three-layer network were not explored. Drastic improvements were not expected with a deeper architecture due to the amount of data.

After the best methods were identified, they were retrained on all of the 2015-2018 data. Then, projections for the 2020 season using 2019 data were compiled and compared to projections from Steamer, a system used by Fangraphs [4].

3.4 Result

The results of projecting wOBA numerically for the next season can be found below in Table 3-2.

Table 3-2: wOBA Results

Method	MAE
Linear Regression	0.0240
Ridge Regression	0.0241
Elastic Net Regression	0.0242
Extreme Gradient Boosting	0.0265
Neural Network	0.0243

The metric used to evaluate the results was MAE rather than RMSE, since wOBA ranges between 0 and 1 as previously mentioned. Linear regression was the best performing metric, edging out ridge and elastic net regression. Both extreme gradient boosting and the neural network did not perform as well as the others with the neural network close behind. Although

linear regression performed the best, we would want a MAE better than 0.0240 in projecting wOBA. This error would mean a player projected to be exactly average (0.320), could actually be above average or poor [36]. This shows the fickle nature of sports, in particular baseball, which leads to difficulty in projecting future performance. The results of projecting wOBA as a factor for the next season can be found below in Table 3-3.

Table 3-3: wOBA Factor Results

Method	**Accuracy**
Logistic Regression	70.09%
Stochastic Gradient Boosting	69.16%
Extreme Gradient Boosting	65.89%
Neural Network	75.23%

Accuracy was the metric used to evaluate the results of the wOBA factor projections. The neural network outperformed the other three methods the next closest being logistic regression, both achieving an accuracy of 70%. An accuracy of over 70% in the context of sports bodes well for predictability with the variability in performance from year to year. This type of projection was superior to the numeric projection since a player could be average, above average, or poor with the MAE that was found as noted before. The MAE between the linear regression model and Steamer projections for wOBA and the similarity between the neural network and Steamer projections for wOBA as a factor for the 2020 season is below in

Table 3-4: MAE Difference and Similarity with Steamer

Metric	**Value**
MAE	0.0128
Similarity	82.78%

The MAE of 0.0128 and only about 17% of results differing with Steamer projections showed there was some agreement between them but still quite apart. One reason for this was the

methods described in this paper having problems with top players that are outliers. In addition, in 2019 several elite young players have come into the league. **Error! Reference source not f ound.** below illustrates this, by comparing projections of the top 5 wOBA projections for the 2020 season to Steamer's projections for 2020.

Table 3-5: Top 5 wOBA Projections Compared to Steamer

Name	2020 wOBA	2020 Steamer wOBA	2020 wOBA factor	2020 Steamer wOBA factor
Christian Yelich	0.383	0.398	Above Average	Above Average
Juan Soto	0.383	0.400	Above Average	Above Average
Mike Trout	0.380	0.427	Above Average	Above Average
Anthony Rendon	0.380	0.367	Above Average	Above Average
Freddie Freeman	0.379	0.382	Above Average	Above Average

Surprisingly, Mike Trout, regarded as the best player in MLB, was ranked 3rd in projected wOBA for 2020 at 0.380 with Steamer projecting 0.427. Trout never had a 0.380 wOBA or less in his eight full seasons [4][17]. Juan Soto, in his second season at age 20, was ranked higher than Trout. He was very young, and not many players at their age succeed right away at the major league level. The average age of players from 2015-2019 was 28, minimum 19 (Juan Soto), maximum 43, and quantile one 25. With the data used, wOBA projections were pulled closer to the mean. In addition, age was significant at the 0.05 level for the linear regression and logistic regression models. With few young players in the data set, and those that were performing at a high level, age may have been overvalued in the model. Age was not a differentiating factor between Christian Yelich and Trout as they were the same age. Trout's WPA has been lower than Yelich's the last two seasons (5.17 and 4.14 compared to 7.34 and 6.02). WPA was a significant variable in the linear model at the 0.05 level. In almost every other metric Trout outperformed Yelich the last three season, which made this result surprising. With a

MAE of 0.0240, that could push Trout above the rest but at 0.404 on the higher end it still falls short of Steamer's. Also, a difference of 0.003 in wOBA is miniscule. Both the logistic regression model and Steamer projections agreed on the top 5 wOBA as a factor variable.

3.5 Conclusion

Results showed that linear regression and a neural network were the best methods for projecting wOBA numerically and as a manufactured factor variable respectively. Although in terms of MAE and accuracy the results were not earth shattering, in the context of projecting future performance in baseball, these were not terrible results. Year to year variability in baseball lead to less accurate results but manufacturing variables into factors led to more useful and accurate information. However, the models may have overvalued age and did not perform well for players that are outliers. A big problem was the amount of data used for analysis. Although Statcast data was helpful in projecting future performance, using only data from 2015-2018 for training limits the amount of training observations for the models, leading to higher variability in projections. As more seasons are played, the incorporation of Statcast data in projecting future performance should lead to higher accuracy in models. However, using more data rather than Statcast data may lead to better results.

3.6 Future Work

There are several areas for future work in this area. First, the weighting of the data could be changed. More seasons could be used in the weighted average and a wider grid search could be explored for better predictability. In addition, a grid search for the best combination of seasons and weighting schemes could be explored as well. Also, regressing the data of players that have played less than three seasons towards the mean as weighting could be explored. In The Book [3] it's been

shown that players' wOBA will regress towards the mean on average [3]. Using this fact, average performance data could be added to weight the performance of players that have not played many seasons.

Neural networks could also be explored more. Different architectures could be explored with more layers and a different number of nodes per layer. Although a wider grid search could be performed, it is limited by computational power. As we add more layers and number of nodes to search, the computations will take exponentially longer to complete. Deeper architectures could me more effective but with the limited amount of data and 48 inputs, the improvements may be modest.

Next, fitting more Fangraphs player-season data rather than limiting to 2015-2018 could improve the models. There is still the issue of baseball changing year to year and a limited amount of player data for players younger than 25, but more data may assuage the issue. Also, imputing the Statcast data for seasons prior to 2015 could be possible but with limited data it may prove unhelpful. Another approach to the limited data issues is to limit the number of plate appearances to 100 and also have the number of plate appearances as a predictor variable in the models. Lowering the number of plate appearances too much could lead to problems of unusually high wOBA. Simulating data is another way to add more data that can be used in the models.

Another approach for better accuracy could be using the projections from other sources, such as PECOTA, Steamer, and others. One way of doing this is using those projections for the next season as predictor variables in the models. With the accuracy of those other projection systems, they could add predictability by being added to the models. A different way of incorporating other projections is to take a weighted average after the models have predicted wOBA for the next season. A grid search could be explored for the optimal weighting of these projections as well.

Finally, approaching age in a different way may lead to better results. Instead of using the raw player's age, clustering players based on age and the other variables and then using that cluster assignment as a predictor variable. This would combat against young players who are outliers. Using an aging curve to build a variable for the predicted drop or increase in wOBA based on age could also be used. In addition, instead of just looking at the predicted drop or increase in wOBA, one could look at the predicted drop or increase in wOBA by cluster assignment.

References

[1] K. Koseler and M. Stephan, “Machine Learning Applications in Baseball: A Systematic Literature Review,” *Appl. Artif. Intell.*, vol. 31, no. 9–10, pp. 745–763, 2017.

[2] MLB Advanced Media, “Statcast.” 2019.

[3] T. Tango, M. Lichtman, and A. Dolphin, *The Book: Playing The Percentages In Baseball*. TMA Press, 2007.

[4] D. Appleman, “Fangraphs.” [Online]. Available: https://www.fangraphs.com. [Accessed: 03-May-2020].

[5] B. Reiter, “Your 2017 World Series Champs,” *Sport. Illus.*, p. 30, 2014.

[6] J. Diamond, “The MLB Coach Who Played Only T-Ball,” *The Wall Street Journal*, 11-Mar-2019.

[7] SABR, “Society For American Baseball Research.” [Online]. Available: https://www.sabr.org. [Accessed: 03-May-2020].

[8] M. Lewis, *Moneyball: The art of winning an unfair game*. W.W. Norton, 2003.

[9] D. Willman, “Baseball Savant.” 2019.

[10] C. Soto-Valero, M. González-Castellanos, and I. Pérez-Morales, “A predictive model for analysing the starting pitchers’ performance using time series classification methods,” *Int. J. Perform. Anal. Sport*, vol. 17, no. 4, pp. 492–509, 2017.

[11] G. Gartheeban and J. Guttag, “A data-driven method for in-game decision making in MLB,” p. 973, 2013.

[12] W. K. Harrison and J. L. Salmon, “Bullpen Strategies for Major League Baseball,” pp. 1–16, 2017.

[13] J. Piette, A. Braunstein, B. B. McShane, and S. T. Jensen, “A Point-Mass Mixture Random Effects Model for Pitching Metrics,” *J. Quant. Anal. Sport.*, vol. 6, no. 3, 2010.

[14] Fangraphs, “Guts!” .

[15] S. Slowinski, “ERA-/FIP-/xFIP-.” 2019.

[16] MLB Advanced Media, “Game Score.” 2019.

[17] S. Forman, “Baseball-Reference.” [Online]. Available: https://www.baseball-reference.com. [Accessed: 03-May-2020].

[18] Baseball Reference, “Dylan Bundy.” 2019.

[19] Baseball Reference, “Max Scherzer.” 2019.

[20] D. Willman, “Baseball Savant,” 2020. [Online]. Available: https://baseballsavant.mlb.com. [Accessed: 03-May-2020].

[21] QuantumSports, “Pitch Sequencing 101 : Fastballs Up-And-In,” 2019.

[22] MLBAM, “Perceived Velocity (PV),” 2019. .

[23] J. R. Bock, “Pitch Sequence Complexity and Long-Term Pitcher Performance,” pp. 40–55, 2015.

[24] G. Ganeshapillai and J. Guttag, “Predicting the Next Pitch,” 2012.

[25] J. Roegele, “Strikeout Pitch Sequences,” 2013.

[26] J. Roegele, “The Effects of Pitch Sequencing,” 2014.

[27] Baseball Prospectus, “Run Expectancy Matrix,” 2019. [Online]. Available:

https://legacy.baseballprospectus.com/sortable/index.php?cid=975409.

[28] D. Smith, “Retrosheet.” [Online]. Available: https://www.retrosheet.org. [Accessed: 03-May-2020].

[29] MLBAM, “Statcast,” 2020. .

[30] MLBAM, “Player Empirical Comparison and Optimization Test Algorithm (PECOTA).” [Online]. Available: http://m.mlb.com/glossary/projection-systems/player-empirical-comparison-and-optimization-test-algorithm. [Accessed: 03-May-2020].

[31] Baseball Prospectus, “PECOTA Projections.” [Online]. Available: https://www.baseballprospectus.com/pecota-projections/. [Accessed: 03-May-2020].

[32] A. Lyle, “Baseball Prediction Using Ensemble Learning,” 2007.

[33] S. T. Jensen, B. B. Mcshane, A. J. Wyner, and A. Lyle, “Hierarchical Bayesian Modeling of Hitting Performance in Baseball,” no. 4, pp. 631–652, 2008.

[34] M. Panda, “Penalized Regression Models For Major League Baseball Metrics,” 2014.

[35] D. Herrlin, “Forecasting MLB Performance Utilizing A Bayesian Approach In Order To Optimize A Fantasy Baseball Draft,” 2015.

[36] S. Slowinski, “wOBA,” 2010. [Online]. Available: https://library.fangraphs.com/offense/woba/. [Accessed: 03-May-2020].