Minor League Baseball Projection Models

Eric Droegemeier

University of North Texas

Toulouse Graduate School of Business

ADTA 5910 – Capstone

Dr. Denise Philpot

Spring 2024

**TABLE OF CONTENTS**

**INTRODUCTION ............................................................................................ 3**Background ......................................................................................................... 3  
Research Questions ............................................................................................. 6

**LITERATURE REVIEW ................................................................................. 6**  
A Grassroots Movement....................................................................................... 6  
Sabermetrics goes mainstream ............................................................................. 8  
Public Projection Models ..................................................................................... 9

**METHOD........................................................................................................... 14**  
Data Collection and Source................................................................................. 14  
Data Manipulation and Cleaning ........................................................................ 16

**EXPLORATORY DATA ANALYSIS ............................................................ 16**  
Data Description ................................................................................................. 16  
Description of Variables...................................................................................... 17  
Outliers................................................................................................................. 18

**MODELING ...................................................................................................... 21**Linear Modeling................................................................................................... 21  
Ridge Modeling ................................................................................................... 22  
Lasso modeling .................................................................................................... 26  
Best Model Selection............................................................................................ 28

**MODEL EVALUATION.................................................................................... 29**War Formulae........................................................................................................ 29  
Player Projections.................................................................................................. 29  
  
**CONCLUSION ................................................................................................... 33**Discussion ............................................................................................................ 33  
Applications ......................................................................................................... 34  
Limitations and Future Research ......................................................................... 35

**DATA DICTIONARY ....................................................................................... 37  
REFERENCES ................................................................................................... 39**

**Introduction**

Background

Baseball teams have many motivations to create a roster that can win. From financial gain through increased attendance, local and national television revenue, playoff revenue, and fan engagement, to the simple thrill of victory for players, personnel, owners, and executives. Because of the millions of dollars potentially gained or lost due to a successful season, the industry is extremely cutthroat as teams fight for an edge in even the smallest of margins. This evolved slowly over the first 150 years or so of the sport, but since the Moneyball revolution of the early 2000’s, teams have become more analytical in their approach. Teams have become more data driven and generally make better decisions in all aspects of the games, whether it be macro decisions such as player trades, free agent signings, drafting and signing amateurs, and organizational philosophies, or more micro decisions such as platoons, lineup, and bullpen construction, and in game strategical decisions.

My project is to build a model to accurately predict the performance of Major League Baseball players based on their performance in the minor leagues, and have a model that accounts for ballpark factors, run environments, age, strength of league, level, and various other variables that can affect minor league performance. This model is inspired by similar models such as the Fangraphs KATOH projection system. KATOH was developed by Fangraphs Chris Mitchell in 2014 to predict the long-term WAR of minor league players over the course of their careers should they reach the major leagues. Mitchell ran predictions and saw which statistics were significant at predicting future major league performance throughout all minor league levels.

There are other projection models out there, such as ZiPS, designed by Dan Szymborski, which was developed in 2002-2004 and went live to the public in 2005. ZiPS performs two primary tasks: estimating a baseline expectation for a player and estimating future performance using large samples of similar players. It does this by using multiyear statistics and giving a heavier weight to recent seasons. Szymborski has expanded the scope of ZiPS through the years as more data has become publicly available, such as incorporating Statcast data into the projection modeling in 2013. ZiPS uses the baseline player projection and compares it to the baselines of all the other players in the database to project future production. Not only does it calculate future individual production, but also how much better a player is than their compatriots. ZiPS also uses Mahalanobis Distance to find real world historical comparisons for players. It essentially says, for example, “Corey Seager in the next three years is the most similar of a player to Anthony Rendon from 2017-2020, the next similar player is Cal Ripken Jr from 1988-1991, next is Buddy Bell 79-82, and so on.”

The benefits of this study are obvious: scouts watch players and try to predict if that player will ever be a major league player someday, and if so, will they be a good player, a great player, or just an average player. But scouts aren’t perfect and are making guess work. If we can use statistics, actual measurements of player performance, and use models to predict how that player will perform upon reaching the major leagues. This is not meant to be a replacement for traditional scouting, but rather an objective looks on minor league performance when accounting for all the elements of performance. This can identify players who may be overlooked by scouts and evaluators, or may identify players who scouts like, but the projection suggests being cautious about.

What was once a niche corner of baseball fandom with individuals such as Bill James and Tom Tango writing about a more analytical approach to the game has now evolved into teams spending millions of dollars on analytics departments. This was made famous by Michael Lewis Moneyball and its subsequent film adaptation that told the story of the Billy Beane and the Oakland A’s and how they built a team that competed with the likes of the Yankees and Red Sox while keeping a payroll less than half the size. The A’s did this by buying undervalued players for cheap.

Today, every team has an analytics department of some form. Some teams, such as the Tampa Bay Rays, will intentionally not spend on big name free agents or retaining their current stars, and have decided it is a better investment to spend on a top tier analytics department to keep costs down and maximize profits. The way Major League Baseball’s financial system is structured, teams are financially motivated to target young players as they provide the best performance for a low salary before the player is eligible to reach free agency. The Rays have maintained a playoff caliber roster for the last 15-20 years despite being near the bottom of the league in payroll. Others, such as the Dodgers or Yankees, invest heavily into both analytics and player development but also the on-field product, maintaining payrolls well above 200 and sometimes 300 million. Mets Owner Steve Cohen said he views spending money to build a winning team as a public service (Cox, 2024). The Atlanta Braves have spent money on retaining their young stars but have kept the payroll manageable by signing them to long-term contracts for a low price while the players are still young and have not fully developed.

No matter if the goal is to maximize profit or to maximize winning, teams have financial motivation to invest into an analytics department. Having a statistical projection model can help a team analyze players and assist with decision making.

Link to dataset: https://www.fangraphs.com/leaders/minor-league?pos=all&level=0&lg=2,4,5,6,7,8,9,10,11,14,12,13,15,16,17,18,30,32&stats=bat&qual=y&type=1&team=&season=2006&seasonEnd=2023&org=&ind=0&splitTeam=false&players=&sort=23,1

**Research question one**: Which MiLB stats correlate with MLB stats? During the EDA process, I will explore what minor league stats correlate to their major league counterparts. For example, is there a strong positive correlation between a high batting average in the minors and a high batting average in the minors?

**Research question two**: Can an accurate projection model be made to predict major  
league stats based on performance in the minor leagues? My goal is to build a model that can accurately predict future performance for minor league baseball players.

Potential concerns and issues do arise, however. Stats such as fWAR are proprietary and cannot easily be replicated statistically. WAR is positional dependent, as certain positions on a baseball field are more valuable than others. Fielding and baserunning data are unavailable for the minor leagues, so the model can only look at offensive output only.

**LITERATURE REVIEW**

A grassroots movement

Since the invention of computers, baseball analysts and front office personnel have attempted to build accurate models to project the performance of players. Major league teams are highly motivated to gain a competitive edge on the competition by being able to accurately predict the future. Professional baseball teams use models in all aspects of decision making, when it comes to the amateur draft and international amateur free agents, player trades, free agent contracts and extensions, arbitration hearings, and player development. Additionally, a slew of public projection models has emerged that have given fans exposure to baseball statistics modeling.

Statistics have always been a part of baseball, and people working independently from teams have led the charge of educating and informing the public of a more analytical approach to America’s pastime for over a hundred years. Baseball Hall of Fame sportswriter Henry Chadwick invented the first box score, first denoted statistical abbreviations such as strikeouts as K, and tracked statistics as runs, errors, assists, and hits (Pesca, 2011). Chadwick’s work laid the foundation for fans to follow teams and players.

Bill James was a security guard at a pork and beans factory when he began to publish his annual Baseball Abstract in 1977. Bill James invented many modern advanced statistics that are meant to give a better and more accurate picture of a player’s performance than the traditional stats and scouting methods. James invented statistics like Runs Created, which proved to be foundational for statistician Tom Tango, who invented statistics such as wOBA, FIP, and wRC+ among others. Detailed explanations of these stats and James’s work can be found on the Fangraphs Glossary, including the Fangraphs version of WAR, or Wins Above Replacement. WAR quantifies every aspect of a player’s contribution to how many games their team won over a replacement player. WAR “…attempts to estimate a player’s total value relative to a free available player, such as a minor league free agent.” (Slowinski, 2012).

Tom Tango, also known for his online alias Tango Tiger, invented a public projection model known as Marcel. It is a simple projection model and performs about on par with other projection models. According to Tango, the system simply, "...uses three years of MLB data, with the most recent data weighted heavier. It regresses toward the mean. And it has an age factor." (MLB.com, 2024.)

Sabermetrics Goes Mainstream.

Oakland A’s GM Billy Beane was a follower of Bill James’s work and built a competitive team on a much lower payroll than other teams such as the Yankees and Red Sox. Rather than rely on traditional scouting, Beane assembled a team based on statistics such as on base percentage and computer models to find players who were not superstar names but were very good players. Beane thought a player’s physical appearance was not important, whereas it was an important part of scouting, and that college players had a much higher success rate compared to high schoolers in the Amateur Draft. The Michael Lewis book Moneyball told this story to the public, which was later adapted into a film starring Brad Pitt.

In 2005, Wasserman, et al. tested this theory by comparing the statistics of a select group of players from college and high school over four years. The researchers found no significant difference in performance. However, the researchers claimed that Beane’s reasoning for drafting college players was that they would perform better than high school players, which contradicts Michael Lewis’s Moneyball, in that the reasoning Beane gave for drafting college players is that they were safer bets to succeed in the majors, not necessarily that they would perform better upon reaching the league. I believe this research was flawed in its approach. “Because of the minimal significant differences between college and high school players’ “Moneyball” statistics, many MLB teams might want to disregard the notion that cheaper “Moneyball” college drafted players are better investments because they do not do as well as their high school drafted counterparts. However, even though the comparison is not significant statistically, the statistics may be significant to an organization/coach, which is playing the Moneyball way of baseball.” (Wasserman, et al, 2005).

As Bill James (and Billy Beanes) ideas and strategies spread around the baseball industry, it was no longer enough of a competitive edge to identify players by statistics overlooked by other teams. The best and most forward-thinking teams poured resources into their player development. Travis Sawchik and Ben Lindberg wrote on the modern trends and strategies that teams use in their 2019 book The MVP Machine. High speed cameras and radar ball tracking technology such as Trackman and Edgertronic opened a whole new world of player analysis, and Major League Baseball’s roll out of Statcast made this data available to the public in 2015. Statcast provided the public with metrics such as spin rate, exit velocity, sprint speed, and introduced expected statistics via high-speed cameras at all 30 parks. These metrics unfortunately are only available in major league parks, so I cannot include them in my model. However, Sawchik and Lindberg explain that teams have highly secretive proprietary methods of player development that involved internal projection systems that factor in such data from the minor leagues. Many players also seek out development on their own grounds if they feel the organization they play for is not helping them enough, which has led to the growth of companies such as Driveline, which provides analytical and biomechanical focused training.

Public Projection Models

In 2014, Chris Mitchell of Fangraphs launched KATOH, named after Yankees minor league Gosuke Katoh, first attempted to predict simply whether a minor leaguer would make the playoffs. In the next iteration of KATOH, Mitchell sought to identify significant statistics per minor league level that most strongly contributed to WAR in the major leagues.

“The table below gives a summary of which stats proved to be significant at each minor league level. This analysis includes minor league data going back to 1990, the first year in which full-season A-ball was broken up into Class-A (A) and Class-A Advanced (A+). R+ refers to the advanced rookie leagues–the Appalachian and Pioneer Leagues, while R- includes the Arizona and Gulf Coast Leagues.

A table with text on it

Description automatically generated

Figure 1 - Significant Statistics by Level (Mitchell, 2014)

This table shows that age, ISO, BABIP, and K% are significant at all or nearly all minor league levels. However, it is not great at predicting the performance of players in the low minors. Most players at that level are teenagers, and the pitching is not very good compared to the upper minors, so hitting stats don’t correlate strongly with WAR.

An interesting part of this article was the few paragraphs on Aaron Judge. Mitchell says that KATOH was not a big fan of Judge and that his high walk rate was not an indicator that he would be a good major leaguer. “Take away Judge’s walks, and his stat line suddenly looks like that of a nondescript minor leaguer with something of a strikeout problem.” (Mitchell, 2014). The scouts were a fan of Judge however, and now that Judge has won an MVP award and set the American league single season home run record, it shows the limitations of relying on projection models!

Good sports analysts, however, will tweak their models to improve them over time, and Chris Mitchell improved KATOH later with changes like accounting for sample size, adding positional adjustments based on defensive ability, and player height. He also added projections for pitchers, whereas KATOH had previously only looked at hitters.

A chart of a number of green circles

Description automatically generated

Figure 2 - How Select Stats Influence WAR Thru Age-28 for a Typical Pitcher (Mitchell, 2015)

This image shows the stats with the strongest influence on WAR for pitchers at each level of the minor leagues. As the chart shows, age is a huge factor in determining WAR through the age-28 season. This is likely because younger players have more opportunities for pro seasons and thus accrue more WAR. Another reason is that younger players, especially teenagers, have more room to develop and may not have yet reached physical maturity compared to older players.

Public projection models can only be so accurate. Josh Imbriani (n.d.), in his research of neural network-based baseball projection system, concluded that proprietary projection models used by major league teams are far more accurate than public models. Obviously, if a team had an extremely accurate projection model, they would not want it to become public as they would lose the competitive edge that such a model provides. Imbriani compared his models to other public projection models, ZiPS, Marcel, and STEAMER. He found his simple one and two hidden layer neural network was pretty like those models, with it performing slightly better in some areas and slightly worse in others, when looking at individual player comparisons for Evan Longoria, Mike Trout, and Bryce Harper. NeuRayl, his name for his model, was only looking at simple counting stats like home runs, batting average, RBI, etc, while other models calculated more complex statistics such as WAR.

One of the more well-known projection systems is ZiPS, from Dan Szymborski of Fangraphs. What began as a simple projection model in 2004 evolved from a simple model to one more complex. “When estimating a player’s future production, ZiPS compares their baseline performance, both in quality and shape, to the baseline of every player in its database at every point in their career. This database consists of every major leaguer since the Deadball era — the game was so different prior to then that I’ve found pre-Deadball comps make projections less accurate — and every minor league translation since the late 1960s.” (Szymborski, 2021.) ZiPS uses cluster analysis, namely Mahalanobis Distance, to make historical player comps. Brandon Nguyen made an informative blog post on how ZiPS does this, and explains why Mahalanobis distance is used over, for example, Euclidian Distance. “In the case of baseball statistics, some statistics correlate with one another such as OBP and OPS. Then the Mahalanobis distance is a measure that we would want to use, as it considers the correlations between data unlike the Euclidean distance.” (Nguyen, 2021.)

Many MLB teams have relied more and more on sabermetrics and advanced computer models to help inform decision making. Jacob Berish researched if teams may rely too heavily on projections so much so that it causes bias in their decision making. He defines projection bias as “Projection bias refers to an individual’s assumption that their tastes and preferences will remain constant over time.” (Berish, 2021). His study looked at the statistical performance of players before and after they signed a long-term free agent contract. His results showed a slight bias among teams but lacked a strong conclusion to determine if it existed among all executives in the study.

Daniel Acevedo (2017) created a model based on a simple spinner board method. Each spinner contained potential outcomes and were weighted based on a player’s previous career stats, and weighted for age regression. Multiple seasons were projected due to the wide variance of potential outcomes, and rates are obtained so that number of plate appearances can be predicted. For example, Acevedo predicted Wil Myers 2018 season. The first step was to establish if the result would be a ball in play or not a ball in play. From there, the spinner board breaks down further with multiple possible outcomes that result from a ball in play. Acevedo then compared his model to other public projection systems like Marcel, Steamer, ZiPS, and Cairo. Acevedo concluded that his model performed similarly to other public models, but it was significantly weaker when there were fewer than 3 years of data available for the player.

Cairo is the same projection system as Marcel the Monkey, but “includes minor league statistics, adjusts for park and league effects, adjusts the aging curve depending upon the statistic, takes age and position into account when regressing a player’s performance, and uses four years of data instead of three. These projections are then put into the Diamond Mind simulator, and team projections are estimated using the results of 50,000 simulations.” (Slowinski, 2011).

PECOTA is developed by Nate Silver at Baseball Prospectus, and it stands for “Player Empirical Comparison and Optimization Test.” It is named after former utility infielder Bill Pecota and is an extremely complicated model. It is like other projection models in that it calculates a baseline performance based on previous statistics, but it differs from other systems in that it factors in body type, position, and age, to identify various comparison factors. It also has a Breakout Rate, which is “the estimated likelihood that the player will beat his baseline weighted average from the previous three years by at least 20%, calculated by looking at how often the comparison players broke out in that fashion.” (Druschel, 2016).

Based on my research, I believe that a model can be designed that can predict if a player will reach the major leagues, and predict said performance in the major leagues, based on their minor league stats. Ideally, I would like for this model to be extremely accurate, but if it is only moderately accurate, I would be satisfied with that result.

**METHOD**

Data collection and source

The data I will be using is from Fangraphs, which is powered by Sports Info Solutions, a company that tracks pitch type, velocity, batted ball location, and play-by-play data for major league and minor league games. Fangraphs minor league data goes as far back as the 2006 season but can be customized to include only a certain number of seasons. Fangraphs provides advanced statistics as well, such as their own version of WAR (Wins Above Replacement), which is an attempt by statisticians to quantify each aspect of the game (hitting, fielding, and baserunning) into a number which calculates how many wins they gave their team over a typical “replacement” player, or an average minor leaguer. Fangraphs WAR is proprietary, and Chris Mitchell’s original KATOH calculated projected WAR over a player’s career through the age of 28. However, his research found the variables that were statistically significant across all levels were Age, ISO (Isolated Power), BABIP (Batting Average on Balls in Play), and that BB% was significant in the higher levels of the minor leagues. Age is telling, as it would make sense than a 25-year-old in rookie ball would have a worse chance of even making the major leagues than a 20-year-old in double or triple A.

The dataset from Fangraphs is very customizable and easy to manage to include any length of time, as well as split the seasons or combine into an entire minor league career. As stated earlier, their data dates to 2006. It also allows me to filter by playing time. I can choose to include only players who played close to a full season or expand the data set to include every player who has ever appeared in a game during that time. If I were to look at qualified minor league players dating back to 2006 with no split seasons, my dataset would have 13,172 observations. The variables are also flexible and can be customized to my needs, provided it is something Fangraphs tracks. I also can use Fangraphs to get MLB level stats which will be needed to make comparisons to test data. The data contains a player and team name, which are identifiers, and level of the minors played. Minor league levels vary by skill and talent but generally progress from Rookie and DSL, A, AA, and AAA levels. Another variable is the player age, which has been shown to be important when projecting future performance. PA is plate appearances, which is a sample size indicator. It counts the number of times a player came up to bat in the game. The other variables are hitting statistics. The default advanced statistics page contains 20 variables. For this study, I will only be looking at hitting statistics.

Data manipulation and cleaning

Data manipulation was done using Microsoft Excel and analyzed with RStudio. The variables required of the research question were multiple offensive statistics for both minor league and major league players. Fangraphs minor league statistical archive dates back to 2006, so an excel file was exported from Fangraphs of all minor league players who had a qualified number of plate appearances from the 2006 to 2023 seasons. This gave me the total career minor league stats for 13,173 players across all levels and seasons in that range of years.

To clean the data, I exported from Fangraphs a file of every single player who had at least one plate appearance in MLB from 2006-2023 and used an Excel function to match the minor league players with their major league stats using the unique ID number that Fangraphs gives each player. Any player who never had a major league plate appearance was then deleted from the dataset, as well as Major Leaguers who did not have any minor league stats (likely debuted before the 2006 cutoff date or did not spend long enough in the minor leagues to qualify). This left me with 2,057 unique players in the dataset with their major and minor league stats combined into one.

**EXPLORATORY DATA ANALYSIS**

Data description

The dataset consists of 29 variables and 2,057 players. As mentioned previously, the 2,057 players were combed from over 10,000 minor league players, so players who never reached the majors were removed from the dataset. This is because all their MLB stat variables would be missing. As a result, there are no missing values in the dataset. However, there are some players who had as low as one major league plate appearance. If the player did not reach base in their one plate appearance, many of their MLB stats, such as batting average, would be zero and their wRC+ would be –100.

Description of Variables

The variables offered a range of information, from useful identifiers such as player name, the team they play for and their age, as well as the level they played for. In the research section, I saw that correlation between performance can vary by level and this study will test that theory as it will ignore level and simply combine all minor league levels. Another complexity in accounting for the level of the minors is Major League Baseball’s closure of over a hundred of its minor league clubs in 2019 and more research may be needed on the effect having fewer levels of the minors has on player performance and development.

Other variables are simple counting variables, such as PA and MLBPA, which are just a count of the number of times to the plate a player had in the minor leagues and major leagues respectively, or MLBHR and MLBG which counts the number of home runs the player hit in MLB and the number of games appeared in. Some are percentages such as the percent of plate appearances resulting in a walk, there are simple stats like batting average, and more complex ones like weighted runs above average, a formula that uses linear weights to assign a number of runs above (or below) the average a player was worth. Two key variables here are MLBwOBA and MLBWAR as these were two variables I identified as response variables during my modeling. I chose these two variables as wOBA is considered to be a better indicator of a player’s true offensive value compared to other rate statistics. WOBA accounts for the fact that not every way of reaching base is equal, which is a drawback of stats like OBP and Batting Average. Singles are less valuable than doubles, which in turn are less valuable than triples or home runs. Whereas stats like Slugging Percentage simply assign a double to be twice as valuable as a single, wOBA looks at run values and understands it is more complex than that. wOBA also accounts for other ways of reaching base, such as walks, hit by pitches, or beating out a double play ball. It takes the linear weights of every possible batted ball event in a baseball game and adjusts them year by year as league wide offensive trends change.

WAR is considered by sabermetricians to be the best (but not perfect!) overall measure of a player’s contribution. It stands for Wins Above Replacement and is an assessment of hitting, running, and defense to boil everything down to one number: how many wins would X player give over a hypothetical replacement player?

Outliers

There are some interesting outliers that are worth mentioning here. Sergio Santos was a former relief pitcher who appeared in 194 games in the major leagues. Before the Designated Hitter was applied universally in 2020, pitchers would have to bat in the National League. This was very common for starters, but very rare that relief pitchers would need to bat, as they would usually throw only one or two innings and be replaced by a pinch hitter. However, Sergio Santos had one plate appearance in his major league career. However, he had 1459 plate appearances in the minors. He was a shortstop and third baseman before he converted to pitching full time. There are a few other relief pitchers who are in the dataset, such as Jose Ruiz, Mychal Givens, Sean Doolittle, and Pedro Baez.

I did not remove these outliers from the dataset for a few reasons. First, they likely were not very good at hitting. If they were, they would not have switched to pitching. Also, I did not want to assign a cutoff for number of plate appearances as more talented players will obviously receive more plate appearances so doing so would skew the results. And given the data's timing, having a minimum number of plate appearances would introduce bias to players who have been playing longer and penalize younger players, including top prospects such as Gunnar Henderson. For example, Dustin Pedroia has the 22nd most MLB plate appearances in the data set, but only had 566 plate appearances in the minors in this dataset. Pedroia debuted as a rookie in 2006, the first-year data was available. Meanwhile, top Diamondbacks prospect Jordan Lawlar has just 34 big league plate appearances, but at 21 years old he has his whole career ahead of him still.

When we visualize the outliers, it becomes clear how much outliers can effect the dataset. In the figure, most points of data are clustered around 100 wRC+ and the 0 career WAR. We can see a player of even 20 career WAR is an extreme outlier. The dot on the farthest left is Mike Trout, and we can see what an extreme outlier he is!

A graph with many colored dots

Description automatically generated

Figure 3 - MLBWAR and Minors wRC+ colored by team

Another way to visualize the outliers is by arranging our chart by team. This allows us to see which teams have developed more players that stand out as star players. The chart below shows that teams such as the Braves, Red Sox, and Dodgers seem to have more outliers compared to teams like the Marlins. Those teams all have been very successful and won a World Series during the span of the data set. The chart that labels two teams means a player was traded at some point during their minor league career.

A table of black dots

Description automatically generated with medium confidence

Figure 4 - MLBWAR and Minors wRC+ with split Teams

Upon conducting a simple correlation matrix, we can see which statistics correlate strongly with each other. In this correlation matrix, we are especially focused on the variables MLBWAR and MLBwOBA as we are trying to find a simple indicator of production at the major league level. None of the variables show a very strong positive correlation with these two variables. The strongest correlation with WAR was wRC+ at 0.35. Surprisingly, minor league wOBA only correlated with MLBwOBA at 0.3. The correlation matrix does reveal which stats have the strongest correlations- 0.82, 0.85, and 0.94 respectively- to each other, which is understandable as stats like wRC+ is simply an improved version of rate statistics like OBP, SLG, and OPS and measure the same thing. OPS itself stands for On Base Plus Slugging and is the two stats added together. OPS and SLG correlate at 0.95, so to avoid multicollinearity I do not include redundant statistics in my calculations.

**MODELING**

For all modeling a set seed function was used. The dataset was split into training data and test data with 80% in the training data and 20% in the test data.

Linear modeling

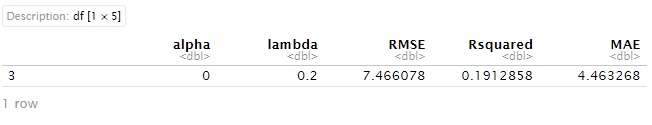
A simple linear regression model was tested to predict two variables, MLBWAR, and MLBwOBA, one for each variable. I tested various linear regression models using different variables. The first model I tested was using variables based off my knowledge of baseball stats and the ones most commonly used to assess player performance. These variables were BB%, K%, BB/K, AVg, OPS, ISO, BABIP, wOBA, and wRC+. PA was also included to control sample size. This model produced a very strong p-value of < 2.2e-16, but the Adj R squared was only 0.2104. The significant coefficients (resulting in a P-value of less than 0.01) showed the statistically significant variables were PA, BB%, K%, OPS, ISO, and wRC+.

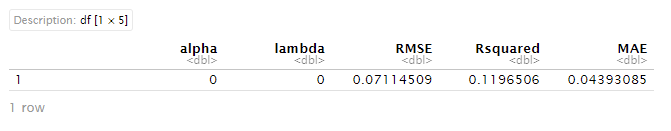
Interpreting these variables is important and tricky. The problem with PA is it can vary wildly and for different reasons. Injury can substantially cut a player’s number of PA and a small sample size of PAs can lead to every other variables in the data being skewed. Baseball stats normalize over time, for example a player with one single plate appearance may walk, which leads to a 100% BB%, 0% K%, and 1.000 OBP. This is one reason I selected just the qualified players in the minors.

Finding the AIC, and BIC of this model resulted in very high numbers. This simple linear model showed that the variables used are statistically significant, but the model does not fit the data very well.

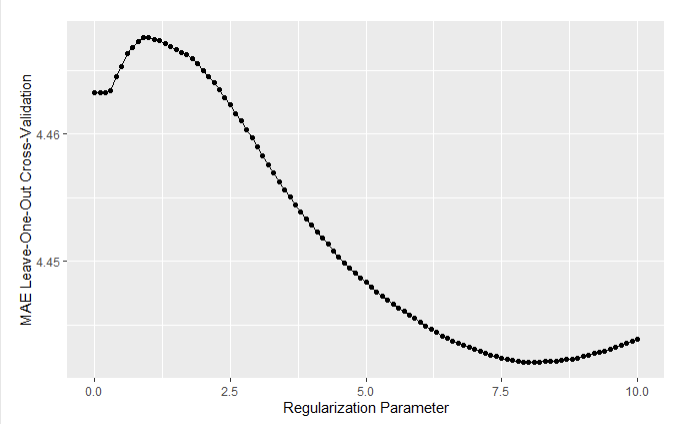
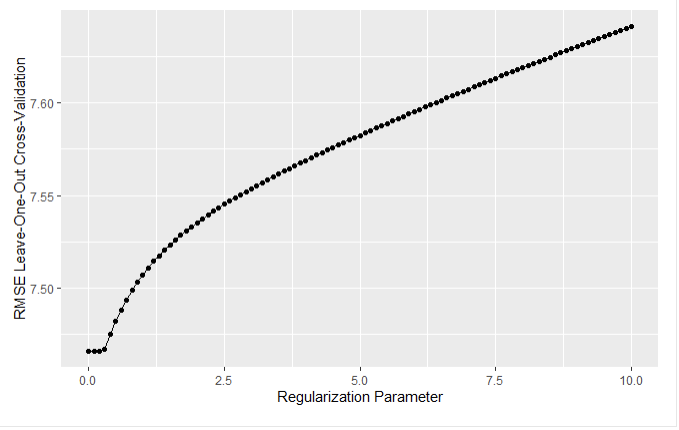
Ridge modeling

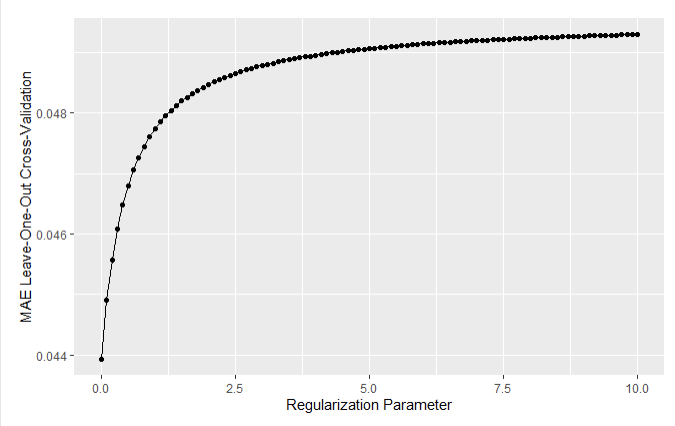
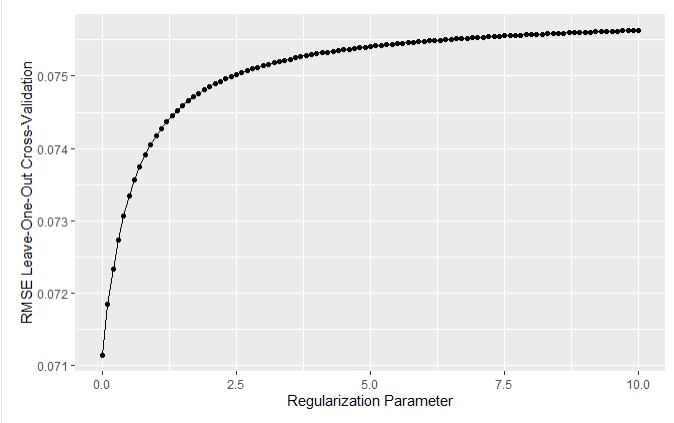
Since my dataset has many variables that mean nearly the same thing, there is a risk for multicollinearity. Ridge regression is a technique to reduce overfitting and corrects multicollinearity. Performing a ridge model to test both WAR and wOBA models meant including every variable in minor league stats. Ridge regression on the MLBWAR predictor resulted in the following values:

And the results of the ridge regression on the MLBwOBA predictor:

As we can see by these results, our lambda values are very low. This means we run the risk of overfitting the data and our model will be very complex. The RMSE and MAE values however show that the ridge model for MLBwOBA is very accurate, especially compared to the ridge model for WAR. However, if we look further into the data, we may not want to rule out the ridge model for WAR completely. MLB career WAR totals can range from the low negative numbers to over 100 for some of the game’s all-time great players. A Mean Absolute Error of around 4 may not be that bad.

Ridge model plots for WAR:

Ridge model plots for MLBwOBA

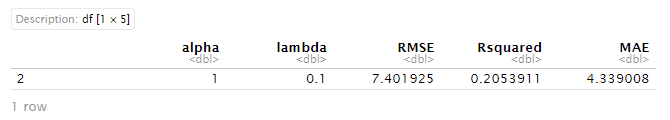


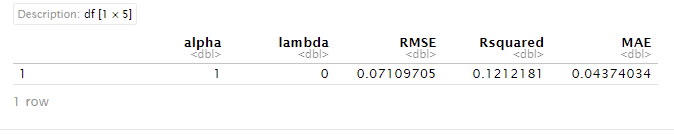
For WAR, the ridge model identified the best coefficients as PA, BB%, K%, Avg, OPS, ISO, BABIP, WOBA, and wRC+. For wOBA, the best coefficients were BB%, K%, AVG, OBP, ISO, BABIP, wRC, wRAA, and wRC+.

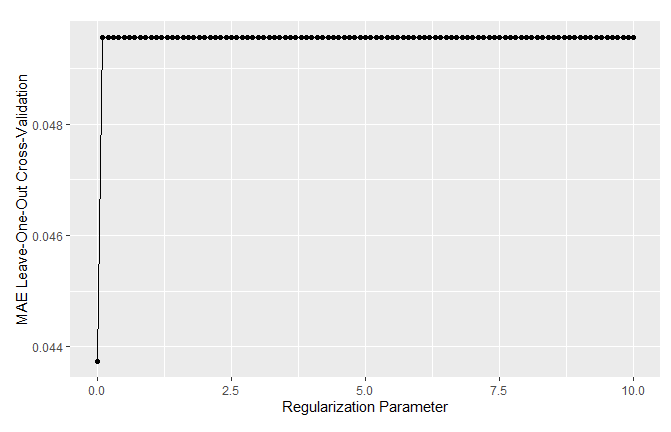
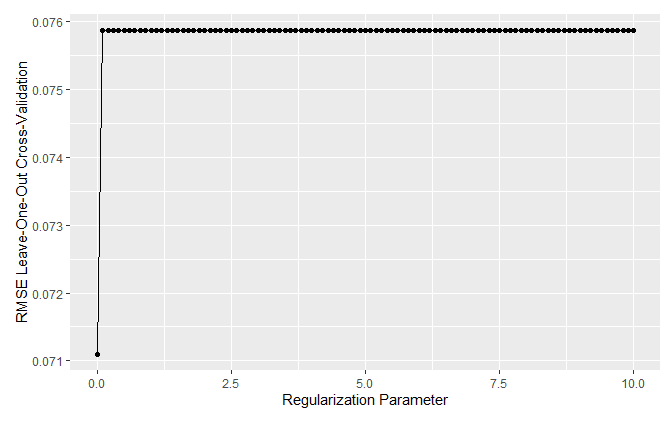
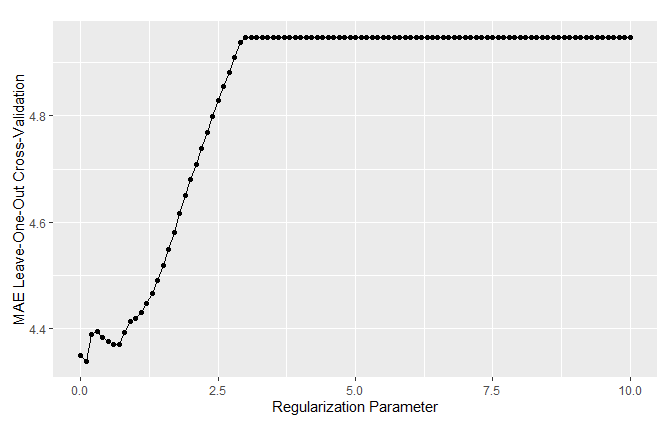
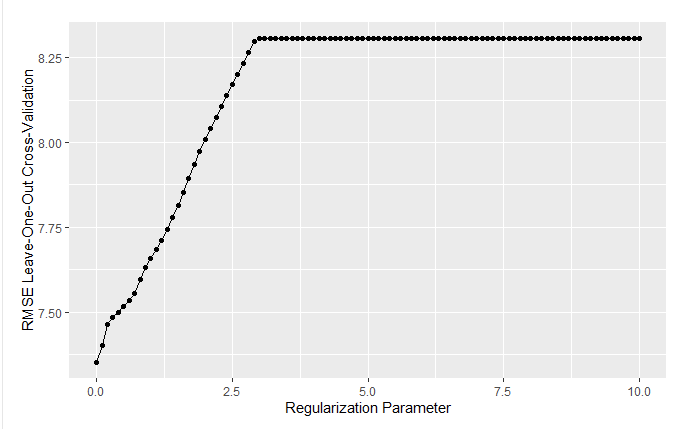
Lasso Modeling

To narrow down the number of variables used, Lasso regression was performed. Lasso regression takes all the variables in a dataset and selects the best ones automatically.

For WAR

For wOBA

These results are very similar to the Ridge regression. Here are graphs of the regularization parameter for RMSE and MAE for WAR and wOBA respectively.

for WAR, the variables removed from the model were PA, BB/K, SLG, OPS, and wOBA. The results of this shows that for wOBA, every variable was included, with OPS being the only variable removed in the lasso regression. Knowing what we know about wOBA, and its attempt to be the ultimate, all-in-one offensive metric, it makes sense that every single offensive statistic would have significant coefficients. However, it makes using a Lasso regression to predict wOBA a waste of time. Every single variable included was extremely strong. For WAR, the strongest coefficient was BABIP, and the strongest negative was OBP.

Best Model Selection

I performed more simple linear regression models using the variables suggested by the ridge and lasso regression. However, this did nothing to improve the adjusted r squared of the model. The resulting adjusted r squared for each linear model was as follows:

WAR using Lasso variables: 0.213

WAR using Ridge variables: 0.2108

WOBA with Ridge variables: 0.1273

WOBA with Lasso variables: 0.1292

Using R, I took the Mean Squared Prediction Error for each model to evaluate the best model overall. The extremely low scores for the wOBA target variable models indicated multicollinearity in the model. I ended up selecting the Multi-linear regression using the Lasso variables for WAR as it provided the most accurate model and did not select a wOBA model due to the multicollinearity. See figure 5 for MSPE numbers.

A screen shot of a computer code

Description automatically generated

Figure 5 MSPE scores for each model

**MODEL EVALUATION**

WAR formulae

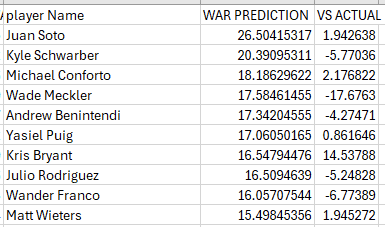
Let’s look at some actual projections of players! When looking at the best model, it’s clear that the models that were predicting WAR were better overall than the models predicting wOBA. Using R, we can determine the coefficients used in predicting WAR to calculate what an individual player’s war would be based on their minor league stats. Let’s look at some of MLBs superstars and top prospects. Perhaps we can gain some insight on how they might perform in the future.

The formula being used is below.

"MLBWAR = 18.99529 - 0.001822582 \* PA + 90.64339 \* `BB%` - 68.64342 \* `K%` - 1.590395 \* `BB/K` + 205.0558 \* AVG - 256.3319 \* OPS + 213.7581 \* ISO + 112.8694 \* BABIP + 99.07734 \* wOBA + 0.2942214 \* `wRC+`"

Player projections

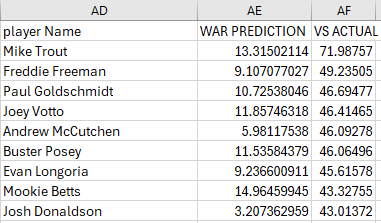
Using an excel spreadsheet, the players stats were entered into the model for all 2,058 players in the dataset. The top 10 highest predicted player WARs are below:



The Vs Actual tab is how much their actual WAR differs from what our formula predicts. Juan Soto has the highest total, at 26.5. He has already exceeded that by almost 2 WAR, and he is just 26 years old. Going down the list, there are a lot of misses here. Michael Conforto, Kris Bryant, and Julio Rodriguez are young enough and good enough to blow their predictions out of the water. Wander Franco was on pace to do so before he was arrested in the Dominican Republic and put his entire career in jeopardy.

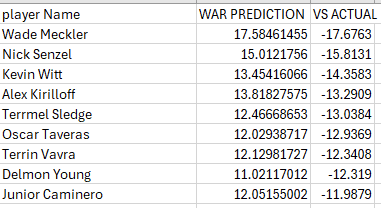
Kyle Schwarber was an extremely heralded prospect when he was younger. He is now a 10-year MLB veteran, and has not exactly lived up to expectations. It is interesting that my WAR projection for him is actually fairly close to his real production, if he has a few more decent seasons, he should come pretty close by the time he retires. Another player the model hit on was Yasiel Puig, which my model came within 1 WAR of. Puig has since retired, and unless he comes back and puts up great numbers, the model pretty much called it. Wade Meckler is also a name that stands out to me. He made his major league debut in 2023, which is why his actual WAR is so much lower than predicted. However, the model is predicting great things for him! He is not a highly ranked prospect, which could mean my model is able to identify undervalued players based only on their minor league production.

Let’s look at the biggest misses. Here are the top 10 players that FAR exceeded what my model predicted.



This is honestly not surprising. Every player that the model was way off on are huge superstars. These players are all MVP candidates and multiple time all-stars. So why does the model suck at identifying the best players? I believe this is a sample size issue. Because we are looking at players entire careers, and not year by year production, the sample size, included in the model as number of plate appearances, is a big contributor. Any player who is in this tier will have far more plate appearances in the majors than they will in the minors. WAR is a counting stat, so the more games they play in, the more (or sometimes less) WAR they will accrue. All these players have had lengthy careers. This is why I hope to get a model that works for wOBA as well as WAR, as wOBA is a rate stat and as sample size increases it will normalize. A possibility to normalize WAR with sample size is to instead measure WAR/162, or the number of WAR a player will put up per 162 game seasons. Fortunately, my model did not predict any of these players to have negative career WAR.

Next, let’s look at the biggest over predictors and see what information can be gleaned.



Wade Meckler finds himself on this list too, at the number one slot. Junior Caminero is the Rays best prospect and #4 on MLB.com’s top prospects list. He is 20 years old, and the model seems to think he can be a solid everyday player for the Rays. Oscar Taveras was a very highly rated prospect in the early 2010’s, but tragically was killed in a car accident before his talent was fully realized. This list also highlights some players that both the model and evaluators missed on. Nick Senzel, Alex Kirilloff, and Delmon Young were all highly regarded prospects, but their careers have either stagnated or fizzled off. Senzel and Kirilloff are bench players who show no signs of improving, and Delmon Young is considered by many to be one of the biggest prospect busts ever. I had no idea who Kevin Witt and Terrmel Sledge are, but they were bench players in the 2000’s who retired after finishing their careers in the minors.

We can use this WAR formula to look at any active or historical player in the minor leagues. The chart below shows the predicted top 10 prospects for A-ball and higher players for current prospects in the minor league system. Here we see Wade Meckler again, as well as Jackson Holliday. Holliday is widely regarded by industry evaluators as being the top prospect in the minor leagues, so it’s good my model is able to identify his potential!

A table with names and numbers

Description automatically generated

Overall, averaging the total predicted war vs the players actual war, the Vs Actual column has a mean of –0.03. So, when looking at the big picture of the 2057 players in the dataset, the average player has a pretty accurate prediction. Using this formula, I can predict the WAR of any minor league player if stats are available for them.

**CONCLUSION**

Discussion

As mentioned earlier, I could not find a model with particularly high accuracy, despite the strong p-values that showed the importance of my variables to the predicted statistic. To predict WAR, a more complex model is needed to account for the complexity of the WAR statistics. As shown in the literature review, projection models usually account for the age of the player, defensive position, and level of competition when making projection models. I think the WAR I calculated with my model may be an ok median career war projection. I would not recommend my model to anyone to seriously predict the performance of minor leaguers.

Reviewing my research questions, question one was: Which MiLB stats correlate with MLB stats? The strongest model was the ridge regression with lasso. The stats that the lasso model selected were plate appearances, BB%, K%, AVG, OPS, ISO, BABIP, wOBA, and wRC+. Going by the correlation matrix, the strongest correlation with MLBWAR was wRC+ with 0.36.

The second research question was Can an accurate projection model be made to predict major league stats based on performance in the minor leagues? The problem with answering this question is defining accuracy. According to traditional data analysis, the adjusted r squared of the models is not the ideal range of 0.75 or higher. However, given the subject matter of baseball, it is hard to say what the industry standard is for accuracy in modeling. Teams have proprietary projection models that they consult in their decision-making process, however the accuracy of those models are kept very secret of course. Even the teams that are the best at drafting and developing players are not hitting on 75% of their prospects. That would be absolutely crazy!

Applications

The truth is that MLB prospects fail more often than they succeed. According to Kevin Goldstein (Goldstein, 2022), around 70% of prospects fail to make an impact at the major league level. Using the traditional scouting 20-80 scale, where a 50 FV is considered an MLB regular and 60 FV is considered a star player, the vast majority of prospects have at least a 40% chance of being considered a bust, the highest percentage of possible outcomes, with the next highest being in the 40-45 range, which is considered to be a bench/role player for a big-league team.

The figure below represents players in the 50-70 range of the 20-80 scouting scale. Note that the standard deviation for a 50 FV position player is nearly 5 standard deviations. A 50 FV player is considered a big league regular who will start every day. Out of thousands of minor leaguers, there are limited spots for regular every day guys. The last two columns show the percentage of guys who were graded in the top 100 bust percentage and star percentage based on career WAR.

A table with numbers and a green border

Description automatically generated

Figure 6 - Edwards (2018)

When taking his research into the success rates of prospects into account, it makes the 0.2 accuracy rate of my models look better. Given the simplicity of my model, this is likely the most accurate it can be without increasing the complexity. I believe that my model is a good baseline for player expectations of the median of a player’s possible career outcomes.

Limitations and future research

I want to improve this model and add to the complexity. My goal is to get an accuracy score of around 0.3, which would match the success rate of big-league prospects. Another thing to keep in mind is that the 30% is just the success rate of the leagues TOP PROSPECTS, usually 100-200 or so players, not accounting for every single one of the thousands of players in the minor leagues every single year.

Future research could be the number of players who had successful careers who were not considered “top prospects” by evaluators. There are also chances to analyze this with college and high school statistics, but the models would need to consider the vast differences in competition level compared to the professionals in the minor and major leagues. A model can lead teams to change their draft strategy in the types of players that they target in trades and drafting.

More future research could be done on pitchers. With the increase in pitcher injuries, teams need more and more available arms to bring up from the minors to fill innings throughout a long and grueling season. In conclusion, my model is not accurate, and I hope to add more complexity to it and account for the overfitting of the variables better in my final presentation.

**DATA DICTIONARY**

Name: Player’s name

Team: Name of the team played for

Level: level of minors played in

Age: age range of player during seasons in dataset

PA: number of plate appearances

BB%: percentage of plate appearances resulting in a walk

K%: percentage of plate appearances resulting in a strikeout

BB/K: Ratio between walks and strikeouts

AVG: Batting average, in what percentage of at bats did the player reach safely via a hit?

OBP: On base percentage, in what percentage of all plate appearances did the batter reach base?

SLG: Slugging percentage, OBP where extra base hits are value higher

OPS: OBP plus SLG

Spd: Fangraphs proprietary baserunning value stat

BABIP: Batting Average on Balls In Play, what is the players batting average only on balls hit fairly into play

wSB: Fangraphs proprietary stolen base weighted value rate

wRC: weighted runs created, invented by bill james and measure the number of runs a player creates

wRAA: weighted runs above average, the runs above (or below) the average of the rest of the league

wOBA: weighted on base average, takes SLG and uses the linear weights of extra base hits to properly value them alongside reaching via walk or hit by pitch

wRC+: wRC normalized to 100 being average for the league, higher is better

PlayerID: a unique number assigned to each player from fangraphs as that players identification

MLBPA: PA, but in the majors

MLBAVG: major league batting average

MLBwOBA: MLB wOBA

MLBwRC+: wRC+ while in the majors

MLBG: Number of games played in the majors

MLBWAR: Wins Above Replacement- the ultimate stat of value that tells how many wins a player aided to his team over a random “replacement level” player

**REFERENCES**

Cox, J. (2024, April 3). *Steve Cohen says it’s early to judge Mets season, sees making team a winner as a “civic responsibility.”* CNBC. https://www.cnbc.com/2024/04/03/steve-cohen-says-its-too-early-to-judge-the-mets.html

Pesca, M. (2009, July 30). *The Man Who Made Baseball’s box score a hit*. NPR. <https://www.npr.org/templates/story/story.php?storyId=106891539>

Slowinski, P. (2012, April 2). *War for position players*. Sabermetrics Library. <https://library.fangraphs.com/war/war-position-players/>

MLB. (n.d.). *Marcel the monkey forecasting system (MARCEL): Glossary*. MLB.com. <https://www.mlb.com/glossary/projection-systems/marcel-the-monkey-forecasting-system>

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. W.W. Norton.

Wassermann, E., Czech, D. R., Wilson, M. J., & Joyner, A. B. (2005). An Examination of the Moneyball Theory: A Baseball Statistical Analysis. *The Sport Journal*, *8*(1). <https://thesportjournal.org/article/an-examination-of-the-moneyball-theory-a-baseball-statistical-analysis/>

Lindbergh, B., & Sawchik, T. (2019). *MVP machine: How baseball’s new nonconformists are using data to build better players*. Basic Books.

Mitchell, C. (2014, December 30). *Katoh: Forecasting Major League hitting with minor league stats*. The Hardball Times. <https://tht.fangraphs.com/katoh-forecasting-a-hitters-major-league-performance-with-minor-league-stats/>

Mitchell, C. (2015, November 25). *A Primer on a new and improved Katoh*. FanGraphs Baseball. <https://blogs.fangraphs.com/a-primer-on-a-new-and-improved-katoh/>

Imbriani, J. (n.d.). *Josh Imbriani*. Personal Blog. <https://www.joshimbriani.com/blog/2016/12/24/neurayl-a-neural-network-baseball-projection-system/>

Szymborski, D. (2020, November 10). *The 2021 zips projections: An introduction*. FanGraphs Baseball. <https://blogs.fangraphs.com/the-2021-zips-projections-an-introduction/>

Nguyen, B. (2021, June 19). *Finding player comps with mahalnobis distance*. Brandon G. Nguyen. <https://brandon.nguyen.vc/2021/06/finding-player-comps-with-mahalnobis-distance/>

Berish, J. (2019). *Is Projection Bias Present within a General Manager’s Decision-Making Process within Major League Baseball (MLB)?* Economics Student Theses and Capstone Projects <https://creativematter.skidmore.edu/econ_studt_schol/119>

Acevedo, D. (2018). Simulation-Based Projections for Baseball Statistics.

Slowinski, P. (2011, February 16). *The projection rundown: The basics on Marcels, Zips, Cairo, Oliver, and the rest*. Sabermetrics Library. <https://library.fangraphs.com/the-projection-rundown-the-basics-on-marcels-zips-cairo-oliver-and-the-rest/>

Druschel, H. (2016, February 22). *A guide to the projection systems*. Beyond the Box Score. <https://www.beyondtheboxscore.com/2016/2/22/11079186/projections-marcel-pecota-zips-steamer-explained-guide-math-is-fun>

Goldstein, K. (2022, March 1). *Managing prospect expectations*. FanGraphs Baseball. <https://blogs.fangraphs.com/managing-prospect-expectations/>

Edwards, C. (2018, December 28). *An update to Prospect Valuation*. FanGraphs Baseball. <https://blogs.fangraphs.com/an-update-to-prospect-valuation/>