

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

PREDICTING NBA EFFICIENCY FROM NCAA STATISTICS USING PLAY-BY-
PLAY AND ON-OFF STATISTICS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

in Data Science and Analytics

By

ALEXANDER BEENE

Norman, Oklahoma

2018

PREDICTING NBA EFFICIENCY FROM NCAA STATISTICS USING PLAY-BY-
PLAY AND ON-OFF STATISTICS

A THESIS APPROVED FOR THE
GALLOGLY COLLEGE OF ENGINEERING

BY

Dr. Charles Nicholson, Chair

Dr. Sridhar Radhakrishnan

Dr. Randa Shehab

Abstract

New data is available for college basketball since 2009-10. However, when adding interactions, there is a sparse dataset with over 30,000 variables. This causes problems, because there are far more variables than observations. Further, communicating the rigorous mathematical findings with any credibility within an industry that is relatively new to data science approaches is a challenge. With this in mind, I use this new, contextual data to predict future NBA on-court efficiency.

Understanding that analytics are only a piece of the bigger puzzle of drafting players in the NBA, the goal is to use this new data to build a simple model to predict who will become a maximum contract NBA player, with a focus on explainability. I use a novel approach to splitting players into three positions instead of five, by using this contextual information as a proxy. By being able to discuss a simple model with specific context, I believe this is a good process for the NBA Draft when used in tandem with scouting analyses. This allows for clear and transparent takeaways to discuss with the vast basketball knowledge that employees in NBA organizations bring to the table. This should be helpful when given little time to make a decision that has the potential to impact the legacy of an organization. I finish with visualizations of model results from 2007 thru 2018.

Acknowledgements

This Thesis would not exist without the wonderful people I have been blessed to know. Allison—your love made this work possible; you pushed me to finish this work, and you inspired me the entire way. Dad—your love and the countless hours spent learning basketball under you have shaped this work. Mom—your love and the countless hours and help you have provided me, pushing me to pursue mathematics and academia, have shaped this work. Nick, Nathan and Austin—your love and the countless hours spent learning and brainstorming about the game of basketball with you have shaped this work. Grandpa Beene—your love, life-work and hard work have inspired the entirety of this work. Grandma Beene—your love and support throughout life have made this work possible.

Conrad—the time spent chatting with you has very much helped bring this work to fruition. Derek—your advice and directions in many areas have made this work possible. Luke and my friends in Sacramento—the time spent learning under you has very much brought this work to finality. Spencer and my friends in Indiana—this work would not be possible or a reality without your help, guidance and the time spent learning under you. Scott, Karlis and my friends in Oklahoma City—the time spent learning under you very much shaped the way I think, and this work would not be possible without you.

Dr. Nicholson—the time spent learning under you as the chair of my committee, as well as both of your classes, have carved the foundation of my Data Science skillsets. This entire work would not be possible without you. Your guidance, not only in understanding theory, but in learning how to apply problems in the real world, have been invaluable. You very much guided me in bringing this work to finality. Dr. Shehab—from the many hours you have spent going

back-and-forth with me via e-mail, to your advisement throughout the program, to your being on my committee—this Master’s degree would be possible without you. Dr. Sridhar—this work would not be possible without you, as your awesome leadership in the space of computer programming have very much shaped my thought processes and skillset.

This work is done in memory of Grandma Dempsey. As she has passed on to be with the Lord, her love for her grandchildren have shaped and inspired the entirety of this work.

Above all, I want to thank God. Colossians 3:23 reads “Whatever you do, work at it with all your heart, as working for the Lord, not for human masters”. God’s grace is the reason I am here, the reason for this work and the reason for anything good that you see in me.

“Thanks be to God! He gives us the victory through our Lord Jesus Christ.”

—1 Cor. 15:57

Table of Contents

| | |
|---|-------------|
| <u>Abstract</u> | <u>iv</u> |
| <u>Acknowledgements</u> | <u>v-vi</u> |
| <u>List of Tables</u> | <u>vii</u> |
| <u>List of Figures</u> | <u>viii</u> |
| <u>Chapter 1: Introduction and Literature Review</u> | <u>1</u> |
| <u>Chapter 2: Basketball Statistics and Definitions</u> | <u>14</u> |
| <u>Chapter 3: Methodology and Data Collection</u> | <u>28</u> |
| <u>Chapter 4: Results and Discussion</u> | <u>38</u> |
| <u>Chapter 5: Conclusion and Future Work</u> | <u>46</u> |
| <u>References</u> | <u>49</u> |
| <u>Appendix</u> | <u>53</u> |

List of Tables

| | |
|--|--------------|
| <u>Table. 1: Descriptions of Abbreviations</u> | <u>3-9</u> |
| <u>Table. 2: Mappings of Old Positions to New Positions</u> | <u>25</u> |
| <u>Table. 3: Top 20 Metrics for each Position</u> | <u>32-34</u> |
| <u>Table. 4: Top 10 Metrics with Respect to Coefficient of Variation</u> | <u>35</u> |
| <u>Table. 5: Model Coefficients for Guards</u> | <u>39</u> |
| <u>Table. 6: Model Coefficients for Wings</u> | <u>39</u> |
| <u>Table. 7: Model Coefficients for Bigs</u> | <u>40</u> |
| <u>Table. 8: Model Coefficients for Full Model</u> | <u>41</u> |

List of Figures

| | |
|---|-----------|
| <u>Figure 1. Box Plus-Minus vs Win Shares</u> | <u>13</u> |
| <u>Figure 2. Example of Play-by-Play Information</u> | <u>16</u> |
| <u>Figure 3. NBA Box-Minus for NBA players with at least 30,000 MP since 1974</u> | <u>20</u> |
| <u>Figure 4. Traditional Positions and Spacing in Basketball</u> | <u>22</u> |
| <u>Figure 5. Three Positions in Modern Basketball</u> | <u>22</u> |
| <u>Figure 6. Three Positions and “5-Out” Spacing</u> | <u>22</u> |
| <u>Figure 7. Shooting Guards: How they get divided into Guards and Wings</u> | <u>26</u> |
| <u>Figure 8. Power Forwards: How they get divided into Wings and Bigs</u> | <u>27</u> |
| <u>Figure 9. Missing Data by Last NCAA Season</u> | <u>29</u> |
| <u>Figure 10. Model Results for Guards from 2007 to 2018</u> | <u>42</u> |
| <u>Figure 11. Model Results for Wings from 2007 to 2018</u> | <u>43</u> |
| <u>Figure 12. Model Results for Bigs from 2007 to 2018</u> | <u>44</u> |
| <u>Figure 13. Model Results for All Players Combined from 2007 to 2018</u> | <u>45</u> |

Chapter 1: Introduction and Literature Review

Introduction

The goal of this project is to predict a maximum contract National Basketball Association (NBA) player from Men's National Collegiate Athletic Association (NCAA) data. Of all players drafted in the NBA, 80% played at least one year of NCAA basketball (Berri et al., 2010), so predicting NBA performance from NCAA performance is a critical factor in an NBA organization understanding who they should draft with their pick(s). NBA players on a *Maximum Contract* have negotiated the maximum contractual compensation according to the NBA rules. Such players are commonly considered to be the best players. These players often help the organization who drafts them, as the organizations are able to give the players more money and a longer contract than the other 29 teams in the NBA. It became valuable for both players and organizations to risk betting on the NBA at a younger age once the fourth-year option was added to Rookie contracts (Groothuis et al., 2005). These players are important for a thriving franchise in winning championships. The previous 15 NBA Champions have all drafted at least one player who started for them.

Typically, draft modeling is done by predicting either a player's NBA production in his first 2, 3, 4 or 5 years (Berri et al., 2010; Moxley and Towne, 2015; Evans, 2017), his draft position (Berri et al., 2010; Sailofsky, 2018; Evans, 2017) or his career NBA production (Sailofsky, 2018). Early NBA, overall production and career, NBA production are logical to predict. However, if we predict the average of a player's 3rd, 4th and 5th years, this will give us a predictor of whether they will be good enough to warrant a maximum contract in year 4.

Literature Review

Various researchers and industry specialists have investigated mathematical methods for predicting high quality NBA players. In this section, I provide a brief overview of such work. Given the amount of industry-specific terms in defining performance indicators and predictive modeling features in general, for added clarity, Table 1 provides a brief definition of terms. Some of these definitions are taken directly from online resources and “Basketball on Paper” (Basketball-Reference.com, 2005; nba.com, 2014; wagesofwins.com, 2012; Paine, 2013; Myers, 2014; Schreefer, 2018; Basketball-Reference, 2013; KenPom.com, 2018; Oliver, 2004; Goldstein, 2018).

In 2011, David Berri et al. explored the relationships of NCAA statistics and NBA performance. Their measure of performance was Wins Produced per 48, and they included all players drafted from 1995 to 2007 who played in both the NCAA and the NBA (Berri et al., 2010). Berri developed Wins Produced metric as a metric that results from a model to estimate a player's contribution to team wins (wagesofwins.com, 2012). Wins Produced per 48 minutes was the lowest correlated metric to Real Plus-Minus. For this reason, I did not choose to use this as my metric of NBA performance.

They found that the following factors were positively correlated and statistically significant when predicting the first 2, 3, 4 and 5 years of a player's NBA *Wins Produced* per 48 minutes: rebounds (REB), steals (STL), and two-point percentage (2P%). They also found that NCAA points and winning the NCAA championship the year prior to entering the NBA draft were negatively correlated to NBA Wins Produced per 48 for all four models—years 2, 3, 4 and 5 (Berri et al., 2010).

Table. 1. Descriptions of basketball abbreviations

| Basketball Definitions | | |
|------------------------|----------------------------------|---|
| Abbreviation | Description | Definition |
| 2P% | Two-Point Percentage | Two-Point Field Goals Made divided by Two-Point Field Goals Attempted, assuming at least one Two-Point Field Goal has been attempted. |
| 2PA | Two-Points Attempted | Made + Missed shots inside the 3-point line. |
| 2PM | Two-Points Made | Made shots inside the 3-point line. |
| 3P% | Three-Point Percentage | Three-Point Field Goals Made divided by Three-Point Field Goals Attempted, assuming at least one Three-Point Field Goal has been attempted. |
| 3PA | Three-Points Attempted | Made + Missed shots behind the 3-point line. |
| 3PM | Three-Points Made | Made shots behind the 3-point line. |
| AST | Assists | A player passes to another player and leads directly to a basket |
| AST% | Assist Percentage | Assist Efficiency: An estimate of the percentage of teammate field goals a player assisted while he was on the floor. |
| Big | many Power Forwards; all Centers | Centers and Power Forwards who shoot a moderate or low volume of 3-point shots and get a moderate or high number of own miss putbacks. |
| BLK% | Block Percentage | Block Efficiency: An estimate of the percentage of opponent two-point field goal attempts blocked by a player while he was on the floor. |
| Box-Score Stats | Box-Score Statistics | Statistics commonly used in the game of basketball. These are typically aggregates by games played. Examples of these are points scored, rebounds accrued, and turnovers made. These are available in the NBA since 1949, and they have been the traditional standard for quantifying player and team production. They are available for virtually every league in the game of basketball. |

Table. 1 (Continued). Descriptions of basketball abbreviations

| Basketball Definitions | | |
|-------------------------|---------------------------------|---|
| Abbreviation | Description | Definition |
| BPM | Box Plus-Minus | A box-score estimate of the points per 100 possessions that a player contributed above a league-average player, translated to an average team. |
| C | Center | Typical big man, who plays on the block, and has been able to step out and be more agile in modern basketball. Many refer to this position as the “5”. |
| Combine | NBA Combine | Before the NBA draft, teams bring prospective draftees into their facilities and gather their body measurements, as well as having the prospects perform athletic tests and drills. These measurements include height, weight, sprinting, agility, other body measurements and measures of athleticism. |
| DBPM | Defensive Box Plus-Minus | A box-score estimate of the points per 100 possessions that a player contributed on defense above a league-average player, translated to an average team. Per basketball-reference, this is shown to be less reliable for indicating poor or strong defensive players, since most box-score statistics are offensive metrics- besides BLK and STL. |
| Dean Oliver's Net Rtg | Dean Oliver's Net Rating | Net Rating is Offensive Rating - Defensive Rating, both of which Dean Oliver created. Oliver defines Offensive Rating as "the number of points produced by a player per hundred total possessions" and Defensive Rating as "how many points the player allowed per 100 possessions he individually faced while on the court". |
| Defensive Three Seconds | Defensive Three Seconds | A violation that prohibits players from being in the paint on defense for more than three seconds, unless the player is guarding an opponent in legal guarding position. |
| DRB | Defensive Rebound | Retrieving the ball after a missed field goal or free throw during a possession in which your opponent has the ball. |
| DRB% | Defensive Rebounding Percentage | Defensive Rebounding Efficiency: An estimate of the percentage of available defensive rebounds a player grabbed while he was on the floor. |

Table. 1 (Continued). Descriptions of basketball abbreviations

| Basketball Definitions | | |
|--|--|--|
| Abbreviation | Description | Definition |
| Dunk | Dunk | A shot made by slamming the ball down through the hoop from above with one or both hands. It counts as both a 2PM and FGM. It typically requires a high level of height, agility or athleticism to accrue a high number of dunks in games. |
| Efficiency | Efficiency | A general term used to describe how consistent a player is at helping their team while on the court. For individual statistics such as assists and rebounds, assist percentage and rebounding percentage are efficiencies with respect to each statistic. In this paper, my definition of overall basketball efficiency is Box Plus-Minus. |
| eFG% | Effective Field Goal Percentage | Field Goal Percentage, adjusting for the fact that a 3-Point shot made is worth 1.5 times as much as a 2-Point shot made: $(2\text{-Pt FGM} + 1.5*(3\text{-Pt FGM}))/ (2\text{-Pt FGA} + 3\text{-Pt FGA})$ |
| FGA | Field Goals Attempted | Made + Missed Shots (Includes both 2-point field goals and 3-point field goals.) |
| FGM | Field Goals Made | Made shots (Includes both 2-point field goals and 3-point field goals.) |
| Fouls Drawn | Fouls Drawn | A foul assessed to an opposing player while having the basketball. |
| Fouls Ending in Made Basket | Fouls Ending in Made Basket | Typically referred to as "And 1s", these are made baskets after a foul, which result in one free throw attempt for the foul penalty. |
| FTA | Free Throw Attempted | An unhindered attempt worth one point commonly awarded for a foul. |
| FTM | Free Throw Made | A made, unhindered attempt worth one point. |
| Goldstein's Adjusted, Defensive On-Off | Goldstein's Adjusted, Defensive On-Off | Player on-off, per 36 minutes, adjusted for number of possessions played, team statistics and league statistics. |
| Guard | PG, some SG | Point Guards and Shooting Guards/Ball-handlers who are in control enough to not commit transition, charge turnovers (Trans, Offensive Foul TOVs per poss) at a high rate. |

Table. 1 (Continued). Descriptions of basketball abbreviations

| Basketball Definitions | | |
|------------------------|--|---|
| Abbreviation | Description | Definition |
| Hand Checking | Hand Checking | A rule that was made increasingly a violation over the years, which allowed players to be more physical when playing defense. |
| Jumper | Jump Shot | NBA defines this as a "shot taken after a player jumps in the air". Per Schreefer, "any 2-point shot that is not a tip-in, layup or dunk" is registered as a two-point jump shot. |
| KenPom Top 100 Teams | KenPom Top 100 Teams | kenpom.com is a popular basketball analytics website. They post team statistics, which adjust for position and strength of schedule. KenPom Top 100 statistics only include games when a player is playing against a team who is in the Top 100 in KenPom's Adjusted EM, which includes Margin of Victory per possession, adjusted for schedule in the NCAA. |
| Max Contract | Maximum Contract | The total money a player is allowed to make, per the NBA's Collective Bargaining Agreement. The team a player plays for can offer him more money than any other team - between 25 and 30% of the salary cap. |
| NBA | National Basketball Association | The foremost professional basketball league in the United States. It comprises thirty franchised teams, twenty-nine of which are located in the US, and one in Canada. |
| NCAA | National Collegiate Athletic Association | The largest collegiate athletic association in the United States, whose Division 1 basketball includes players who make up a large majority of the NBA draft each season. |
| Net DRtg | Net, On-Off Defensive Rating | Points given up by the team while a player is on the court, subtracted by points given up by a team when the same player is off the court. |
| Non-Garbage | Non-Garbage | Refers to the time in each game when the game is still in reach and highly competitive. In many cases, the best teams in NCAA or the NBA have games that become out of reach, in which case the statistics can be less difficult to accrue. Per Schreefer, Non-Garbage is a "function of score differential and time remaining, meant to remove the...end of blowouts". |

Table. 1 (Continued). Descriptions of basketball abbreviations

| Basketball Definitions | | |
|------------------------|---------------------------------|--|
| Abbreviation | Description | Definition |
| OBPM | Offensive Box Plus-Minus | A box-score estimate of the points per 100 possessions that a player contributed on offense above a league-average player, translated to an average team. |
| On-Off Stats | On-Off Stats | Uses play-by-play information to calculate how well a team performs with each player on the court and with each player off the court. |
| Opp | Opponent | The team a player is playing against. In play-by-play data, many opponent stats, such as Box-Score Stats, are aggregated for the teams a player played against. |
| ORB | Offensive Rebound | Retrieving the ball after a missed field goal or free throw during a possession in which your team has the ball. |
| ORB% | Offensive Rebounding Percentage | Offensive Rebounding Efficiency: An estimate of the percentage of available offensive rebounds a player grabbed while he was on the floor. |
| per 48 | per 48 Minutes | Statistics that are aggregated every 48 minutes of play. Many use this metric, as there are 48 minutes in a game. Per 48 or Per 36 are good metrics to use if you are not able to calculate possession statistics. |
| PF | Power Forward | Typical big man, many of whom have stepped out to shoot 3s in more recent years. Many refer to this position as the “4”. |
| PG | Point Guard | Ball-handler and starts the offense in traditional basketball. Many refer to this position as the “1”. |
| Play-by-Play Stats | Play-by-Play Stats | Uses play-by-play information to add more context and specifications to player statistics. |

Table. 1 (Continued). Descriptions of basketball abbreviations

| Basketball Definitions | | |
|------------------------|-----------------------|---|
| Abbreviation | Description | Definition |
| Poss | Possession | <p>The time a team (or player on a team) gains offensive possession of the ball until it scores, loses the ball or commits a violation or foul (or is on defense during this time, as a defensive possession).</p> <p>This can be calculated, per Dean Oliver, as:</p> $0.5 * ((Tm\ FGA + 0.4 * Tm\ FTA - 1.07 * (Tm\ ORB / (Tm\ ORB + Opp\ DRB)) * (Tm\ FGA - Tm\ FG) + Tm\ TOV) + (Opp\ FGA + 0.4 * Opp\ FTA - 1.07 * (Opp\ ORB / (Opp\ ORB + Tm\ DRB)) * (Opp\ FGA - Opp\ FG) + Opp\ TOV)).$ <p>As offensive rebounds have made this formula farther from accurate in different leagues, one can now use play-by-play data to calculate.</p> |
| Primary Ball-Handler | Primary Ball-Handler | <p>The player on the court who handles the basketball and typically controls the tempo, or number of possessions, a team plays. In modern basketball, some Wings, or even Bigs, may start the offensive.</p> <p>For this reason, primary ball-handler is a better description than Point Guard or Guard.</p> |
| Putback | Putback | A player secures the ball off a missed shot while on offense and quickly scores. |
| SF | Small Forward | Typical guard or swing man in traditional basketball who may shoot well from outside and may be longer to defend. Many refer to this position as the “3”. |
| SG | Shooting Guard | Typical guard in traditional basketball who shoots well from outside. Many refer to this position as the “2”. |
| Short 2 | Short 2-Pt Field Goal | Shots attempted or made which are listed in play-by-play as either "tip-in", "layup" or "dunk". |
| Space | Space | Having players spaced around the basketball court in a way that benefits your team and opens up the area inside the three-point line to be exploited by players that are skilled and gifted at creating opportunities. |
| STL% | Steal Percentage | Steal Efficiency: An estimate of the percentage of opponent possessions that end with a steal by the player while he was on the floor. |

Table. 1 (Continued). Descriptions of basketball abbreviations

| Basketball Definitions | | |
|------------------------|--|--|
| Abbreviation | Description | Definition |
| Team Recovered Blocks | Team Recovered Blocks | Blocked shots resulting in one's own team retrieving the basketball. |
| Technical Fouls | Technical Fouls | A penalty assessed to a player, coach or team which is assessed by a referee for various reasons such as disrespect or profanity towards a referee, physical contact, excessive timeouts or having six players on the court. |
| Tm | Team | Statistics that have been accrued by a team, instead of a player. Most of the stats in this paper are player statistics, so I use “Tm” to indicate team statistics. |
| TOV | Turnover | A player or team loses possession of the ball to the opposing team before a player takes a shot at their basket. |
| TOV% | Turnover Percentage | Turnover Inefficiency: An estimate of turnovers per 100 plays. A higher number is a less efficient player or team. |
| TRB% | Total Rebounding Percentage | An estimate of the percentage of available rebounds a player had while he was on the floor. |
| Trans | Transition | Per Schreefer, this is the first 10 seconds of a team's (or player's) possession. |
| TS% | True Shooting Percent | A measure of shooting efficiency which takes into account field goals, 3-pt shots and free throws. |
| USG% | Usage Percentage | An estimate of the percentage of team plays used by a player while he was on the floor. |
| Wing | some Shooting Guards, Small Forwards, and few PF | Perimeter players who are typically taller, more versatile defender than Guards. A SG who is a Wing gets more Offensive TOVs on Fouls in Transition. A PF who is a Wing shoots an extremely high rate of 3-point shots and gets fewer of his own putbacks. |
| Win Shares | Win Shares | A metric to distribute team success to the appropriate players on each team. |
| Wins Produced | Wins Produced | A metric that results from a model to estimate a player's contribution to team wins. Wins Produced per 48 minutes was the lowest correlated metric to Real Plus-Minus. |

Sailofsky did a similar study, and he included players who played at least 500 total minutes in the NCAA and the NBA between 2006 and 2013 (Sailofsky, 2018). He used Win Shares per game, given that they played in the NBA, as his measure of NBA performance. *Win Shares* is a metric which attempts to distribute team success to the appropriate players on each team. Further, he adjusted all NCAA metrics for the position they play. The positions he used are discussed in Chapter 2. He found that the following factors, adjusted by position, were positively correlated to NBA performance: Rebounding Percentage (REB%), Assist Percentage (AST%), Steal Percentage (STL%), Turnover Percentage (TOV%), and playing in the Pacific 10 conference (which has since become the Pacific 12). Note that TOV% is a negative coefficient, but fewer turnovers are considered “good” in basketball, so this is included in positively correlated variables. The only variable Sailofsky found to be statistically significant and negatively correlated was year of NCAA eligibility. This is an intuitive result, as most of the top players in high school basketball have played one season of NCAA basketball since the NBA implemented a rule that requires all players to be at least 19 years old and one year removed from high school (nba.com, 2005). This would mean that most of the highly ranked NBA prospects have played one season of NCAA basketball, and have heavily weighted the lower years of eligibility—1 for these players who played one season of NCAA basketball. Conversely, players who stay all four years—4—are weighted lower since, on average, they may not have had the same draft expectations early in their NCAA career as players who declare for the NBA draft after their first year.

Groothuis et al. explored the relationship of NCAA statistics to NBA salary and making All-Star teams (Groothuis et al., 2005). He finds that Blocks per game have a positive effect on

salary for both 1997 and 2002. Moxley and Towne used first 3-year NBA Win Share as their metric of success, and they use NBA performance data from the 2001 through 2006. They found that the variables most predictive of high NBA Win Shares were age, quality of college program and college win shares (Moxley and Towne, 2015).

Evans analyzed drafts between 2006 and 2013. He also used first 3-year, NBA Win Shares as his metric of efficiency (Evans, 2017). Evans also adjusted player statistics by position. He found that Turnovers per 40 minutes were positively correlated and statistically significant to first 3-year, NBA Win Shares. He also found that staying all four years in college was negatively correlated to first 3-year, NBA Win Shares, and age is negatively correlated to first 3-year, NBA Win Shares.

Box Plus-Minus and Win Shares per 48 are compared below (Fig. 1). As efficiency in the game of basketball is critical to helping a team win, *Real Plus-Minus* (RPM) is a metric which answers the question of how good a player is for his team when considering confounding factors such as who he plays with and against. As seen in Figure 1, both Box Plus-Minus and Win Shares per 48 have some of the best players in NBA history as the best players according to both metrics.

It should be noted that it is a great feat for a basketball player to play 30,000 minutes in the best basketball league in the world. So, all players, including the players at the bottom of Figure 1, are some of the best basketball players that have played the game. In differentiating between a good NBA player and a great NBA player statistically, a stable sample size of minutes—more than 30,000—gives us a good description of what these metrics look like. Both metrics seem reasonable; however, Box Plus-Minus has over 66% Coefficient of Determination to Real

Plus-Minus. For this reason, I use Box Plus-Minus as the overall measure of performance (Myers, 2014).

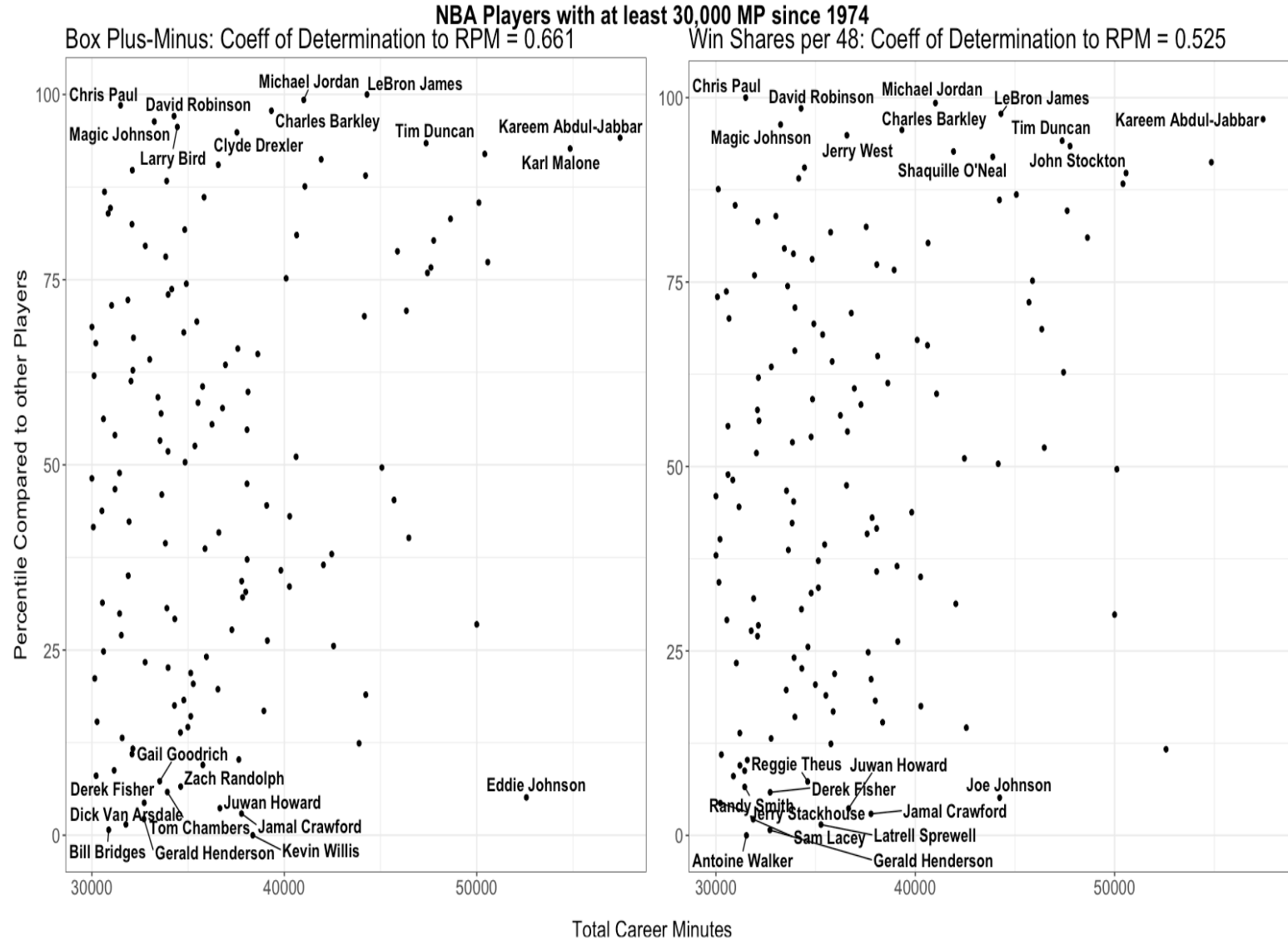


Fig. 1. NBA Box Plus-Minus and Win Shares per 48

Chapter 2: Basketball Statistics and Definitions

Measurements and Box-Score Stats

The *NBA combine* is where NBA teams bring prospective draftees into their facilities and gather their body measurements, as well as having the prospects perform athletic tests and drills. These measurements include height, weight, sprinting, agility, other body measurements and measures of athleticism. Pre-Draft measurements are used to account for the size, length, speed and strength of players. Not all players workout in the NBA Combine, so players' data with missing information are imputed.

Box-Score Stats are commonly used in the game of basketball. These are total aggregates for each game played. Examples of these are points scored, rebounds accrued, and turnovers made. These are available in the NBA since 1949, and they have been the traditional standard for quantifying player and team production. They are available for virtually every league in the game of basketball. However, Box-Score stats are inept at comparing players and teams, because players and teams play a different number of possessions per game. It has been shown and proven that calculating player and team statistics are better for comparison when adjusting for the number of possessions and player or team plays (Oliver, 2004).

My statistics are adjusted for faster or slower tempo by dividing by the number of possessions played. This is estimated by taking the team possessions included in the dataset (Schreefer, 2018). This data is extremely valuable, because Schreefer was able to count the number of possessions from play-by-play information. Possession calculations from box-score stats can often be skewed because the of the coefficients in the estimate (Femrite, 2017).

Advanced Statistics

Advanced statistics are typically defined as any statistic that requires more calculations beyond the box score. Dean Oliver's ground-breaking book in this space, "Basketball on Paper", explains that pace has no correlation to winning. Knowing this, he estimates possession statistics based on the box score. As he shows, there are a finite number of events that can end a play. These events typically show up in a box score in metrics such as made shots—Field Goal Made (FGM), Three-Point Field Goal Made (3PM), Two-Points Made (2PM), Free Throw Made (FTM); missed shots—Field Goal Attempts (FGA) - FGM, Three-Points Attempted (3PA) - 3PM, Two-Points Attempted (2PA) - 2PM, Free Throw Attempted (FTA) - FTM; rebounds—Offensive Rebounds (ORB), Defensive Rebounds (DRB) and turnovers (TOV). Dividing box-score statistics by number of possessions played normalizes teams and players that play faster and get more possessions with players and teams who play slower and get fewer possessions.

However, with play-by-play data, which shows what happened on each possession, possessions can be counted for all players and teams in NCAA basketball since 2009. This data helps in quantifying the metrics of efficiency that Oliver defines (Oliver, 2004), as well as Box Plus-Minus.

Play-by-Play Statistics

Play-by-play data is a powerful addition to box-score statistics. Basketball games typically have a written summary of what happened on each play. Consider the following example from the beginning of the University of Oklahoma vs Wichita State University game played on 12/17/2018:

^ 1st Half

































| TIME | TEAM | PLAY | SCORE |
|-------|---|---|----------------|
| 20:00 |  | Jump Ball won by Wichita State | 0 - 0 |
| 19:51 |  | Shaquille Morris made Dunk. Assisted by Landry Shamet. | 0 - 2 |
| 19:38 |  | Trae Young missed Three Point Jumper. | 0 - 2 |
| 19:38 |  | Rashard Kelly Defensive Rebound. | 0 - 2 |
| 19:32 |  | Rashard Kelly made Dunk. | 0 - 4 |
| 19:16 |  | Khadeem Lattin Turnover. | 0 - 4 |
| 18:58 |  | Shaquille Morris missed Jumper. | 0 - 4 |
| 18:58 |  | Oklahoma Defensive Rebound. | 0 - 4 |
| 18:51 |  | Brady Manek missed Three Point Jumper. | 0 - 4 |
| 18:51 |  | Conner Frankamp Defensive Rebound. | 0 - 4 |
| 18:21 |  | Landry Shamet missed Three Point Jumper. | 0 - 4 |
| 18:21 |  | Shaquille Morris Offensive Rebound. | 0 - 4 |
| 18:16 |  | Shaquille Morris Turnover. | 0 - 4 |
| 18:07 |  | Trae Young made Jumper. | 2 - 4 |
| 17:45 |  | Rashard Kelly missed Jumper. | 2 - 4 |
| 17:45 |  | Khadeem Lattin Defensive Rebound. | 2 - 4 |
| 17:33 |  | Brady Manek made Jumper. Assisted by Trae Young. | 4 - 4 |
| 17:09 |  | Conner Frankamp made Jumper. | 4 - 6 |
| 16:51 |  | Brady Manek made Three Point Jumper. Assisted by Trae Young. | 7 - 6 |
| 16:32 |  | Shaquille Morris missed Three Point Jumper. | 7 - 6 |
| 16:32 |  | Trae Young Defensive Rebound. | 7 - 6 |
| 16:24 |  | Trae Young made Three Point Jumper. | 10 - 6 |
| 16:09 |  | Shaquille Morris made Layup. | 10 - 8 |
| 15:54 |  | Brady Manek missed Three Point Jumper. | 10 - 8 |
| 15:54 |  | Wichita State Offensive Rebound. | 10 - 8 |
| 15:54 |  | Foul on Shaquille Morris. | 10 - 8 |
| 15:54 |  | Official TV Timeout | 10 - 8 |
| 15:50 |  | Official TV Timeout | 10 - 8 |
| 15:45 |  | Trae Young missed Three Point Jumper. | 10 - 8 |
| 15:45 |  | Zach Brown Defensive Rebound. | 10 - 8 |
| 15:21 |  | Zach Brown made Jumper. | 10 - 10 |
| 15:07 |  | Trae Young missed Layup. | 10 - 10 |
| 15:07 |  | Darral Willis Jr. Defensive Rebound. | 10 - 10 |

Fig. 2. Example of basketball, play-by-play information (espn.com, 2018)

This data can be aggregated and provides context to what happened on each play. Consider “Trae Young made Three Point Jumper” with 16:24 (16 minutes and 24 seconds) left in the first half (Fig. 2). A standard box score will register a 3PM. Play-by-play data provides far more context and information. Per Schreefer’s data, transition is defined as the “first 10 seconds of the possession” (Schreefer, 2018). Since Trae grabbed the rebound with 16:32 and scored at 16:24, he scored within 8 seconds of the possession and Schreefer’s data would register this as “transition”. So, with play-by-play data, Trae will register a Three-Point Jump Shot (3PJ), Transition, Three-Points Made (Trans 3PM) and Transition, Three-Point Jump Shot Made (3PJM).

With the progression of modern basketball, this provides important context to quantifying a player’s on-court production in NCAA basketball. This data is used to help predict NBA efficiency, and it is also used as a proxy to classify players by NBA positions (Fig. 7; Fig. 8).

On-Off Statistics

Another use of this play-by-play information is calculating On-Off statistics. One can aggregate how good a team is with a player on the court and off the court. Further, play-by-play information can be aggregated to provide which areas a team is good or bad in with a player on or off the court. For example, on-off statistics include a team’s AST per possession, ORB per possession and STL per possession with a player on or off the court. On-Off data is also aggregate and made publicly available by Schreefer (thesteppen.com, 2018).

Box Plus-Minus as our Measure of Performance

One Box-Score statistic that attempts to quantify a player’s effect on a team is plus-minus. This statistic is the amount of points your team scores compared to your opponent while you are on

the floor. However, this stat is largely influenced by teammates, which makes this stat ineffective for rating a player based on their influence on team success (Shea and Baker, 2013). However, with new NBA data that has over 230,000 possessions per year, Jeremias Engelmann of ESPN created a stat that adjusted plus minus with who they are playing with and against (Wagner, 2014).

Efficiency is used in basketball to describe how consistent a player is at helping their team while on the court. For individual statistics such as assists and defensive rebounds, assist percentage (AST%) and defensive rebounding percentage (DRB%) are efficiencies with respect to their respective statistic. In this paper, my definition of overall efficiency is Box Plus-Minus.

The issue with Real Plus-Minus is that this data only goes back to 2013 and is not currently available for the NCAA and most leagues outside of the NBA. Myers took Oliver's possession-adjusted statistics and regressed them onto 14 years of Real Plus-Minus. He coined this metric as Box Plus-Minus. This stat is highly correlated with Real Plus-Minus and is a good metric for quantifying how efficient a player is while on the court (Fig. 3).

Box Plus-Minus (BPM) is a metric which was built by regressing box-score efficiencies onto Real Plus-Minus (Myers, 2014). This metric includes individual player efficiencies—DRB%, Offensive Rebounding Percentage (ORB%), AST%, Steal Percentage (STL%), Block Percentage (BLK%), Usage Percentage (USG%), Turnover Percentage (TOV%), Three-Point Attempt Rate (3PAr). BPM also adjusts a player's overall effectiveness for his team's overall shooting percentage and the league's 3PAr.

It should be noted that the same BPM with more minutes is better for a team statistically than the same BPM with lower minutes (Myers, 2014). I predict average, minutes-weighted

NBA BPM in years 3 through 5. Career BPM for Players in the last 45 years who have played at least 30,000 are shown below (Fig. 3).

Along with being statistically relevant, this metric passes the eye test, as many players with the highest BPM are also accepted by many basketball experts as some of the best players in NBA history. For example, in 2016 Sports Illustrated (SI) selected their Top 50 players of all time (McCallum, 2018). Considering only players since 1974, the top seven players on SI's list were all in the Top 14 in BPM since 1974. These players are LeBron James (current player; 1st in BPM; 4th in SI as of 2016), Michael Jordan (2nd in BPM; 1st in SI), Magic Johnson (6th in BPM; 3rd in SI), Larry Bird (7th in BPM; 5th in SI), Kareem Abdul-Jabbar (9th in BPM with the most minutes played; 2nd in SI), Jerry West (14th in BPM; 6th in SI) and Tim Duncan (10th in BPM; 7th in SI). The player in SI's Top 5 with the lowest rank according to BPM, Kareem Abdul-Jabbar, accrued the most minutes played at 57,446. The next highest minutes played in SI's Top 5 is LeBron James with 44,298 (as of September 2018). Keeping in mind that more minutes played makes it more difficult to accrue a high BPM, Kareem's ranking by BPM accounting for minutes is higher than 9th.

3 Positions in Basketball

In traditional basketball, players are typically segmented into five positions—one position for each player on the court. These traditional positions are Point Guard (PG), Shooting Guard (SG), Small Forward (SF), Power Forward (PF) and Center (C) (Fig. 4). These positions have shaped the style that teams and players play. For example, the traditional Center grows up playing near the basket their entire careers. Worse, these positions not only have a tendency to force players into a certain style, but they also have a tendency to force teams to play a certain style. Coaches

NBA Historical Box Plus-Minus

Players with at least 30,000 Total Minutes Played Since 1974

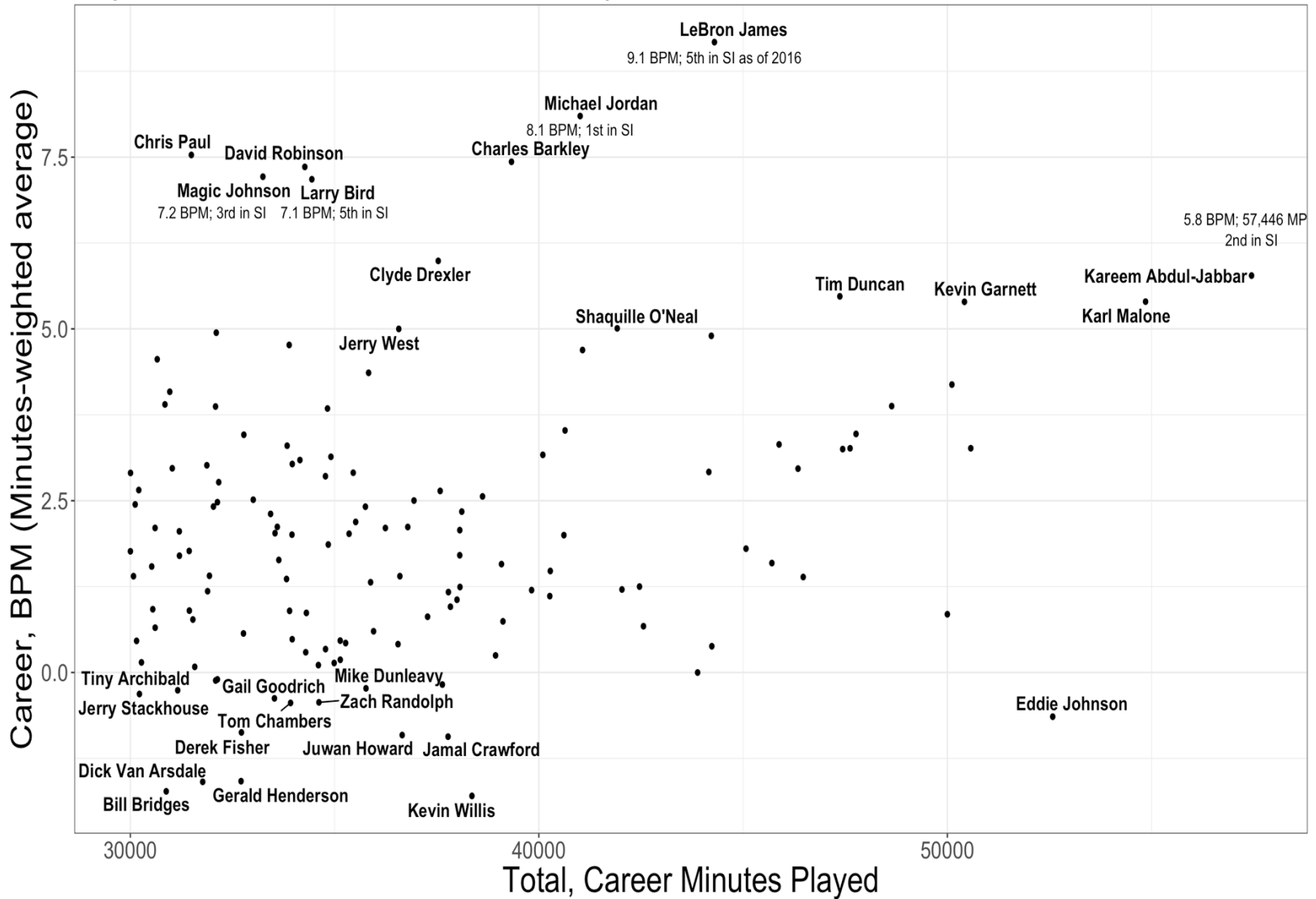


Fig. 3. NBA Box Plus-Minus for NBA players with at least 30,000 MP since 1974

may attempt to conform their teams to fit all five positions, instead of simply placing their best basketball players on the court and figuring how to make that work.

In Figure 4, a simple court diagram of the traditional 5 positions shows how many have spaced the court with two players inside the three-point line. In doing so, there is far less room for players to penetrate to the basket. Restricting one or two players inside the three-point line also makes it far easier to help their teammates who may not be able to stay in front of their opponent. In 2004, the NBA created a defensive three second violation and restricted hand checking (nba.com, 2008). With the defensive 3 second rule, defensive players can remain in the paint as long as they are in guarding position and within three feet of their opponent (NBA.com, 2001). Further, offensive players who are fast and skilled enough to get to the basket were given an advantage by the curtailing of hand checking. These rule changes made it even more imminent for a team to have proper spacing on offense.

Grouping players into only 3 positions has been used more recently in basketball (Sailofsky, 2018). This allows for bigger sample sizes for each position, which will give us a better chance of knowing who the best basketball players are altogether. Fewer positions will also adjust for players who may play many positions. However, there can still be players who play 2 or 3 of the positions, such as playing Guard and Wing. Some examples below show how switching players on defense and offensive with versatility, such as playing both inside the 3-point line and out, can be much simpler when using only three positions. Further, three positions allow you to more easily think about playing with space and having four, or even five, players outside the 3-point line (Fig. 5. and Fig. 6).

Traditional Positions in Basketball

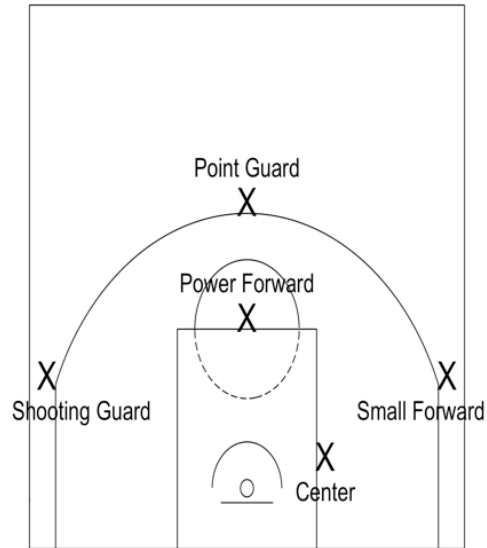


Fig. 4. Traditional Positions and Spacing in Basketball

3 Positions in Basketball

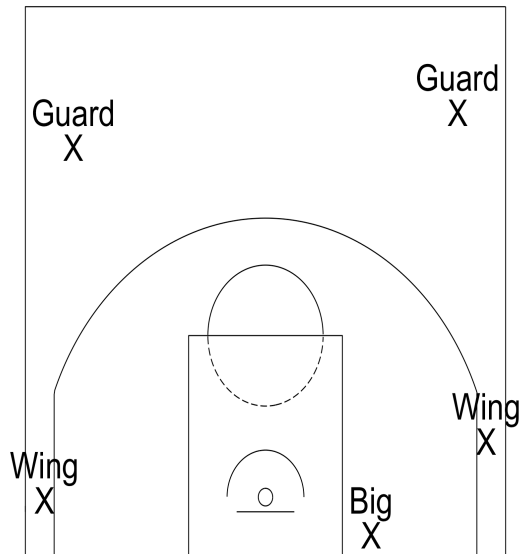


Fig. 5. Three Positions in Modern Basketball

5-Out Spacing

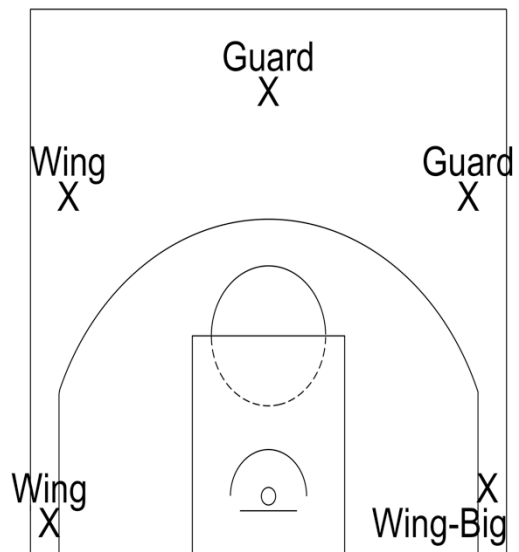


Fig. 6. Three Positions & “5-out” Spacing

Figure 5 and Figure 6 show examples of modern basketball spacing. This type of spacing allows for huge opportunity for players who are able to create opportunities for themselves and their teammates. Thinking in terms of three positions can open up the philosophy of how and where players can play. Further, with more players in modern basketball being able to shoot three-point shots, and shoot shots with range, this spacing is increasingly more effective.

I split players into Guards, Wings and Bigs in a different manner than has been done in the past. Some have split players into Guards (PG and SG), Forwards (SF and PF) and Centers (C) (Sampaio et al., 2006; Abrams et al., 2008; Sampaio et al., 2013; Moxley and Towne, 2015). Others have split players into Ball-Handlers (PG), Wings (SG and SF) and Bigs (PF and C) (Sailorsky, 2018). However, in today's game, some Shooting Guards are commonly the primary ball handlers for their team. Conversely, some Shooting Guards are never the primary ball handler and can defend more positions than typical Guards. With similar logic, some Power Forwards in today's game are able to play and defend beyond the 3-point line. As it seems consensus that a PG is a guard, SF is a wing and C is a big, these positions will remain. However, I use proxies to separate SGs into Guards and Wings and PFs into Wings and Bigs.

Play-by-play information is good for providing context, and this is what I will use as our proxy to separate players into 3 positions. In this paper, *Guards* are PGs and SGs who have lower team turnovers on offensive fouls in transition per possession than average Guards/Wings. *Wings* are SGs who have higher team turnovers on offensive fouls in transition per possession than average Guards/Wings, SFs and PFs who have fewer own miss putbacks than average Wings/Bigs. *Bigs* are PFs who have more own miss putbacks than average Wings/Bigs, and Cs.

A *putback* is when a player secures the ball off a missed shot while on offense and quickly scores.

Transition TOVs on Offensive Fouls per possession are chosen as a proxy for position, because this metric is negatively correlated to NBA BPM for Guards and positively correlated to NBA BPM for Wings (Fig. 7). Own Miss Putbacks are chosen as a proxy for PF because this metric is 14% correlated to Wing, NBA BPM but +52% correlated to Big, NBA BPM (Fig. 8).

Offensive fouls in transition makes intuitive sense to separate guards and wings, since guards are typically more ball dominant and require a high ability of ball control. Figure 7 shows Shooting Guards separated into Guards and Wings. Players, such as Zach Levine and Ben McLemore, who have a high number of Turnovers in transition because of offensive fouls, divided by number of possessions, are classified as Wings. Conversely, players such as Devin Booker and Joe Harris, who have extremely low turnovers in transition by committing offensive fouls are Guards. This makes basketball intuitive sense, because you want guards to handle the ball in many situations. Also, you typically want wings creating and attacking the basket. Therefore, with this proxy, you can take players who have more of a tendency to attack the rim in transition (since they are accruing offensive fouls in transition) and classify of them as wings. Conversely, players who have better ball control in transition will be classified as Guards.

Own miss putbacks makes intuitive sense to separate Power Forwards into Bigs and Wings, as players who play near the basket have far more opportunity to retrieve their misses. In Figure 8, we see that players with a large amount of own miss putbacks, putbacks on their own shots, become Bigs. Players like Larry Sanders and Ed Davis rebound a high number of their own missed shots, and they are classified as Bigs. Conversely, players such as Markieff Morris,

who do not rebound many of their own missed shots and have an extremely high number of 3PA compared to FGA, 3PA_r, are classified as Wings. As with all missing data, the missing data is imputed.

Table 2 is the summary table for the general mapping of the statistical proxies that map a player's traditional position to his new, NBA position. If a player is a PG, he will always be mapped to be an NBA Guard. If a player is a SF, he will always be mapped to be an NBA Wing. If a player is a C, he will always be mapped to be an NBA Big.

More interestingly, if a player is a SG, he will be mapped as either an NBA Guard or Wing. Similarly, if a player is a PF, he will either be mapped to be an NBA Wing or an NBA Big.

Table. 2. Table with Mappings of Old Position to New Position

| Table with Mappings of Old Positions to New Positions | | | | |
|---|---------------------------------|----------------------------------|------------------------------------|--------------|
| Old Position | Transition, Offensive Foul TOVs | Own Miss Putbacks per Possession | Three Point Attempt Rate (3PA/FGA) | New Position |
| PG | - | - | - | Guard |
| SG | Low | - | - | Guard |
| SG | High | - | - | Wing |
| SF | - | - | - | Wing |
| PF | - | Extremely Low | High | Wing |
| PF | | Moderate / High | Low | Big |
| C | - | - | - | Big |

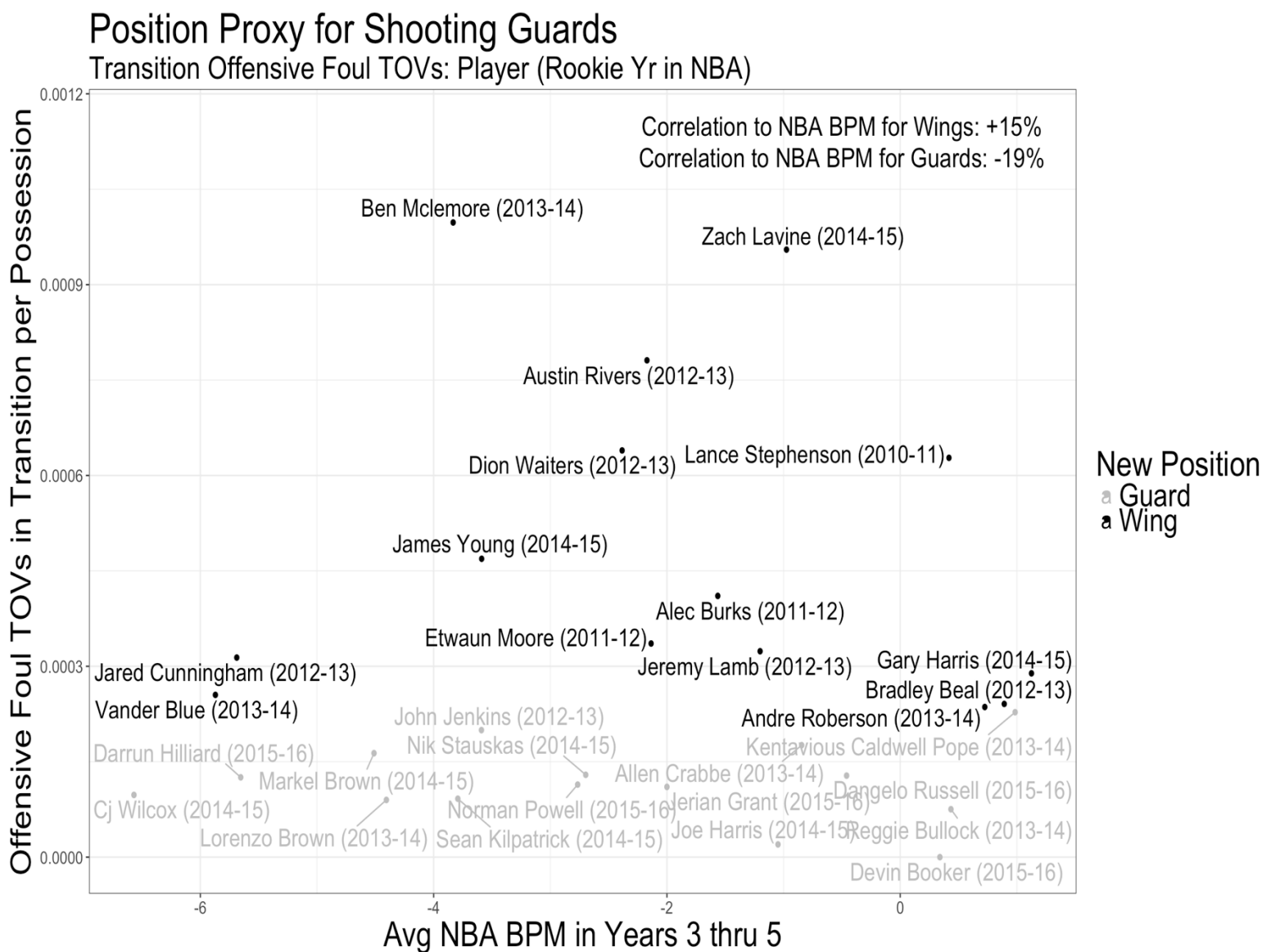


Fig. 7. Shooting Guards: How they get divided into Guards and Wings

Position Proxies for Power Forwards

Own Miss Putbacks & Three-Point Attempt Rate: Player (Rookie Yr in NBA)

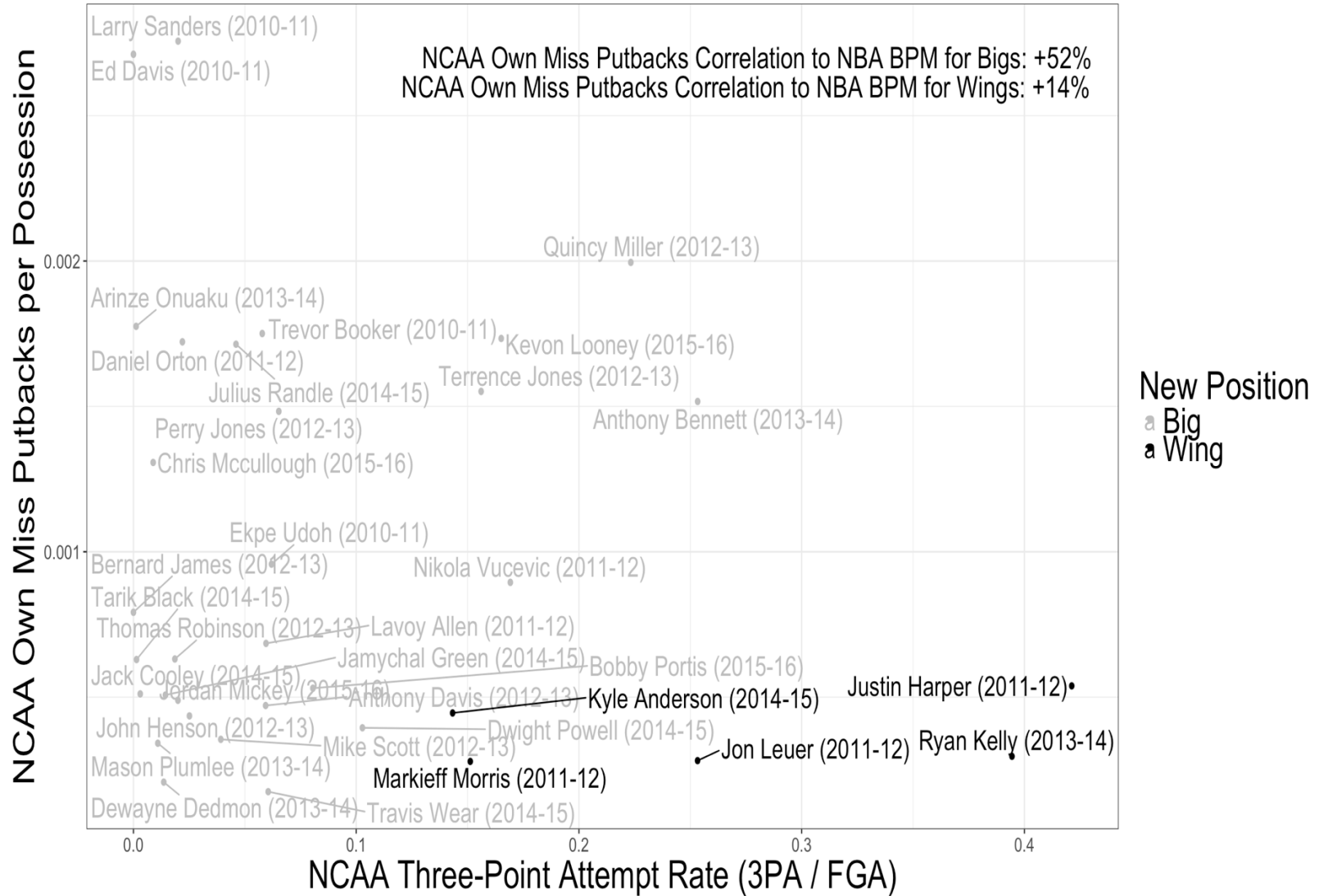


Fig. 8. Power Forwards: How they get divided into Wings and Bigs

Chapter 3: Methodology and Data Collection

CRISP-DM

CRISP-DM, a process for data science, is used as the process of building out our model. CRISP-DM was created in 1996 and has become the most favored methodology in data science, because it is based on practical, real-world data mining projects (Bošnjak et al., 2009; Chapman et al., 2000). While designing the database, business understanding and data understanding are simultaneously the most amount of time spent on the technical work. Play-by-play data, though a huge dataset, is largely understood by the repeatable processes that formed the database. Schreefer's definitions of his dataset is an integral part of understanding the high-covariate dataset. Understanding the context, as well as removing variables with too many 0s, or null values, is critical to constructing the design matrix. The next parts are data preparation, modeling, evaluation and deployment. These will be discussed in detail below.

Variable Selection Method by Correlations

After adding all interaction terms, there are over 30,000 variables in the design matrix. This included all interactions of these variables for each position. The top 130 play-by-play and on-off metrics based on correlation to NBA BPM are kept for each position, as all data before 2010 had to be imputed.

Imputing Missing Data

Though 34% of the original dataset is missing, the reason for the missing data is largely because there is no play-by-play data available before 2010. To see this, Figure 9 shows a chart of missing data by each player's last NCAA season. Since the majority of data is play-by-play or on-off data, NCAA seasons without this data are the main source of missing data.

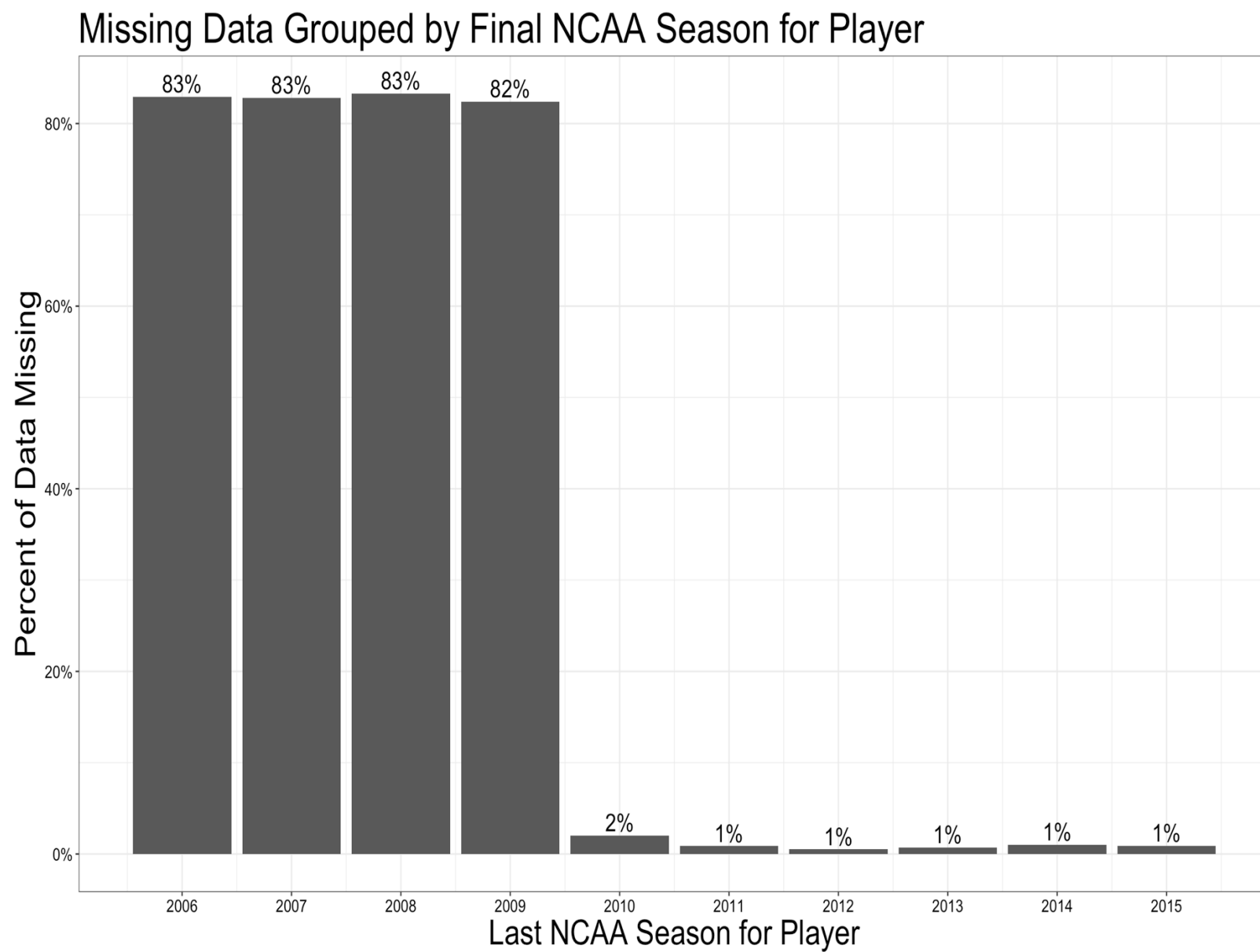


Fig. 9. Missing Data by Last NCAA Season

Multivariate Imputation by Chained Equations (MICE) is used for missing data. MICE is based on Fully Conditional Specifications, where each incomplete variable is imputed by a separate model (van Buuren et al., 2011). Since the data primarily has missing data before 2010 (Fig. 9), predicting the missing data with Multivariate Imputation is a great way to use the relationships within the data to impute. I split the data before imputing, so the relationships with other variables in the data being imputed are related. The data is not contaminated with NBA data, the variable of interest. This includes separately imputing measurements, BPM statistics, NCAA box-score statistics, 3-point statistics, block statistics, assist statistics, rebound statistics, turnover statistics and dunk statistics. Multiple imputation by chained equations allows for three properties that make it ideal for imputing this data: it accounts for the process that created the missing data, preserves the relations in the data and preserves the uncertainty about these relations (van Buuren et al., 2011).

Data Wrangling

One full game, 48 minutes, of -2 BPM (replacement level) is added to every player's Box Plus-Minus to reduce outliers. Examples of these outliers are the numbers for Deandre Liggins and Jarnell Stokes. Liggins only played 1 total minute for the Miami Heat in his years 3 through 5, and Stokes only played 7 minutes for the Denver Nuggets in his years 3 through 5. Without adding this full game of -2 BPM, their data points have the largest residual. However, sample sizes of 1 minute and 7 minutes should not carry heavy weight in predicting a player's efficiency.

After the k conferences the players played in are made into k binary variables, conferences without at least 10 players who played in the NBA are removed from the design

matrix. The list of conferences who remain are the Atlantic 10 (A10), Atlantic Coast (ACC), Big 12, Big East, Conference USA, Mountain West, Big 10, Pacific 12 (Pac 12) and Southeastern Conference (SEC).

After adding interactions, there are 33,933 covariates in the dataset. The way this is handled is by first ordering each position's dataset by Pearson correlation to NBA BPM in years 3 thru 5. Pearson correlation is chosen, because the outliers have already been addressed by including 48 minutes of -2 BPM to all players. Due to this, I care about the outliers and extreme values when looking at correlations. Pearson correlation is a measure of linear relationship between two numerical fields x and y as follows (Nicholson, 2015):

$$r_{xy} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$

For each position—Guards, Wings and Bigs—the 130 covariates with the highest Pearson correlations are kept. Though this statistical method is not well-documented, Pearson correlation has been used for feature selection and shown to perform well (Hall, 1998). Further, with player BPM and each position's covariates, there are still almost 400 variables. The correlation cutoff for guards is 0.121, for wings is 0.164 and for bigs is 0.213. The primary reason for this step is to achieve a model that has a high level of simplicity and interpretability. In doing this, I removed a total of 25,548 variables, with 8,516 variables from each position. Table 3 gives the top 20 metrics for each position, ordered by Pearson correlation to NBA BPM in years three through five.

For measuring variation, I look at *Coefficient of Variation* (CV). If s_x is the standard deviation of a metric, and \bar{x} is the mean of the same metric, then CV is given in Equation 1.

Table. 3. Top 20 Metrics for Guards

| Top 20 Correlations for Guards | |
|---|-------------------------------|
| NCAA Metric | Correlation to NBA BPM |
| Team Blocked when not on the court : BPM | 0.179 |
| Dunk Assists : BPM | 0.178 |
| Dunk Assists : DBPM | 0.171 |
| 3PA : Non-Garbage Percent of Dunks Assisted | -0.167 |
| FTA per FGA : BPM | 0.163 |
| DPM : Transition ORB per Year | 0.154 |
| FGA : Non-Garbage Dunks Assisted | -0.154 |
| Transition Dunks Made : BPM | 0.152 |
| Transition Assists for Short 2s : BPM | 0.151 |
| Team Assists when not on the court : BPM | 0.150 |
| Team Blocks when not on the court : BPM | 0.149 |
| Transition AST for Short 2 : DBPM | 0.148 |
| TOV% : BPM | 0.148 |
| Transition Dunks Made : DBPM | 0.147 |
| DBPM : Non-Garbage ORB per Year | 0.147 |
| Non-Garbage Dunks Made : BPM | 0.147 |
| FTA per FGA : DBPM | 0.147 |
| Non-Garbage Dunks Made : DBPM | 0.145 |
| AST% : BPM | 0.145 |
| Dunks Made vs KenPom Top 100 Teams : BPM | 0.145 |

Table. 3 (Continued). Top 20 Metrics for Wings

| Top 20 Correlations for Wings | |
|---|-------------------------------|
| NCAA Metric | Correlation to NBA BPM |
| BPM : DBPM | 0.213 |
| STL% : BPM | 0.199 |
| $\sqrt{\text{AST\%} * \text{TRB\%}}$: BPM | 0.198 |
| Non-Garbage Pct of TOVs on Offensive Fouls : BPM | 0.197 |
| DRB% : BPM | 0.195 |
| TRB% : BPM | 0.194 |
| Wingspan : BPM | 0.194 |
| Non-Garbage AST from Short 2 : BPM | 0.190 |
| ORB% : BPM | 0.189 |
| Non-Garbage, Short 2s Miss Rebounds : BPM | 0.188 |
| Percent of Blocks from Short 2 : BPM | 0.188 |
| Non-Garbage Percent of AST from Short 2 : BPM | 0.187 |
| BPM : Defensive Win Shares | 0.186 |
| Percent of Team Buckets : DBPM | 0.186 |
| DBPM : Dean Oliver Net Rating | 0.186 |
| Team 3PA while on the floor : BPM | 0.185 |
| Non-Garbage Percent of TOVs on Offensive Fouls : DBPM | 0.184 |
| BPM : Transition DRB per Year | 0.183 |
| Non-garbage AST from Short 2 : DBPM | 0.182 |
| Team 3PM while on the Floor : BPM | 0.182 |

Table. 3 (Continued). Top 20 Metrics for Bigs

| Top 20 Correlations for Bigs | |
|--|-------------------------------|
| NCAA Metric | Correlation to NBA BPM |
| Non-Garbage Percent of AST for Short 2s : BPM | 0.298 |
| $\sqrt{\text{AST}\% * \text{TRB}\%}$: BPM | 0.291 |
| TRB% : BPM | 0.290 |
| BPM : Dean Oliver's Net Rating | 0.289 |
| DRB% : BPM | 0.287 |
| Team Blocks while on the floor : BPM | 0.285 |
| BPM : Dean Oliver's Offensive Rating | 0.282 |
| BPM : Non-Garbage Short 2 FG% | 0.281 |
| BPM : Non-Garbage ORB per Year | 0.280 |
| FG% : BPM | 0.279 |
| BPM : Win Shares | 0.279 |
| AST% : BPM | 0.278 |
| eFG% : BPM | 0.278 |
| TS% : BPM | 0.278 |
| Non-garbage Percent of Dunks Assisted : BPM | 0.278 |
| Vertical Jump : BPM | 0.278 |
| BPM : ORB vs KenPom Top 100 Teams | 0.277 |
| Vertical Reach : BPM | 0.278 |
| Height : BPM | 0.276 |
| Number of Opponent TOVs while on the Floor : BPM | 0.276 |

$$CV = \frac{s_x}{\bar{x}} * 100 \quad (1)$$

Play-by-by metrics, on-off metrics and Offensive BPM have the largest variation of any of the metrics. Table 4 shows the Top 10 metrics with respect to CV.

Table. 4. Top 10 Metrics with respect to Coefficient of Variation

| Top 10 Metrics with Respect to Coefficient of Variation | | | | |
|---|---|--------------------|----------|--------------------------|
| Type | Metric | Standard Deviation | Mean | Coefficient of Variation |
| On-Off | Net DRtg | 0.05 | 0.0008 | 6,256 |
| BPM | OBPM | 2.12 | 0.1309 | 1,620 |
| Play-by-Play | Transition Fouls Ending in a 3PM by Opponent | 0.09 | 0.0079 | 1,120 |
| Play-by-Play | Transition Fouls Ending in a Made 2P Jump Shot by Opponent | 0.06 | 0.0073 | 799 |
| Play-by-Play | Transition Fouls Ending in a Made, Unassisted 3PM for Team | 0.12 | 0.0215 | 543 |
| Play-by-Play | Transition Blocks on 3PA | 0.000032 | 0.000006 | 510 |
| Play-by-Play | Fouls against KenPom Top 100 Teams Ending in a 3PM for the Opponent | 0.14 | 0.0298 | 480 |
| Play-by-Play | Transition, Team Recovered Blocks on 3PA | 0.13 | 0.0272 | 479 |
| On-Off | Goldstein's Adjusted Defensive On-Off | 3.89 | 0.8355 | 465 |
| Play-by-Play | Transition Fouls Ending in a Short 2PM by Opponent | 0.13 | 0.0291 | 462 |

Also, the following play-by-play metrics were removed, because they had a mean and standard deviation of 0: Non-Garbage Technical Fouls, Technical Fouls against KenPom Top 100 Teams, Non-Garbage Fouls Drawn ending in a made basket, Fouls Drawn ending in a made basket against Transition Technical Fouls, KenPom Top 100 Teams, Transition Fouls Drawn ending in a made basket, Transition Fouls ending in a made 3 for the opponent and Number of 20 assist games.

Modeling

After removing and imputing the data, a suite of feature selection techniques are tested. Even after filtering data down by removing data with little variation and data that have the highest correlations to the response variable, there are still over 300 variables to choose from.

The first regression assessment I look at is Akaike's Information Criterion (AIC). For the following, if we let L be the log likelihood and k be the number of estimated parameters, then AIC is the following: $2k - 2\ln(L)$. AIC tends to choose more complex models with higher numbers of variables (Nicholson, 2015). For this reason, feature selection using AIC keeps far too many variables to be explainable.

The next two feature selection methods I look at are Bayesian Information Criterion (BIC) and Least Absolute Shrinkage and Selection Operator (LASSO). Once again, if we let L be the log likelihood and k be the number of estimated parameters, then, BIC is the following: $k \ln(n) - 2 \ln(L)$. Since BIC has a heavy penalty on complexity, the models are far more explainable. However, the model built using BIC for feature selection has a Pearson correlation of 0.49, compared to a Pearson correlation of 0.501 with the model resulting from LASSO for feature selection.

Using LASSO, I am left with explainable models with the highest correlation to the response variable. LASSO minimizes the sum of the squared error, with an upper bound on the sum of the absolute value of the model parameters (Fonti, 2017). That is, if we let N models be given by $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$, then our LASSO parameter, $\hat{\beta}^{lasso}$, is given by minimizing:

$$\frac{1}{2} \sum_{i=1}^N (y_i - (\beta_0 + \sum_{j=1}^p \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^p |\beta_j|$$

where $\lambda > 0$ is a tuning parameter to scale the penalty (Hastie et al, 2017).

In doing so, this takes the dataset with almost 400 variables and shrinks many of the coefficients to 0. The LASSO method for feature selection allows us to take this high-variable dataset and condense it to important variables that are explainable. For cross-validation, I use *Leave-One Out Cross-Validation* (LOOCV)—214-fold for all 214 observations. This means that I fit a model using LASSO 214 times, leaving one observation out each time. The final model performance is based on hold-out predictions and error, by averaging across all 214 models. This allows for use of every observation as a test point to minimize overfitting. The time taken to do this is not insignificant; however, with only 214 observations, LOOCV is worth the time.

After going through this process, we are left with only 15 variables. However, after splitting the variables into Guard Variables, Wing Variables and Big Variables (for each position), we are left with 14 variables that only apply to a player only if he is either a guard, wing or big (but no more than one). Therefore, our model ends up only having one variable that applies to all players, and 14 that apply only when a player that position. Therefore, if we present these models to industry experts, we can think of our model as three simple models: a guard model, a wing model and a big model. This allows for better context for an organizational decision, as well as simplicity in understanding what type of player will be a good fit for each of the three positions.

Chapter 4: Results and Discussion

As mentioned, the results can be thought of as 3 different models for each position. Let y be the predicted NBA, BPM for years 3 thru 5 and NCAA Calculated BPM be x_1 , then the “base” model is given by:

$$y = 1.99 + 0.199 * x_1 \quad (2)$$

More covariates will be added to the base model shown in Equation 2, depending on the position of the player. Given $j-1$ coefficients and covariates for a position, the model for each position is given by:

$$y = 1.99 + 0.199 * x_1 + \sum_{i=2}^j \beta_j x_j$$

NCAA BPM, estimated by using RealGM’s advanced statistics, has a 40.4% correlation to NBA BPM in year’s 3 thru 5 (Table. 5.; Table. 6.; Table. 7.; Table. 8.).

For diagnosing this model, I use *Variance Explained by Predictive Models Based on Cross-Validation* (VEcv). This is shown to be a good indicator of model accuracy when using cross-validation methods (Li et al, 2017). The simple, explainable model that I use has a VEcv of 23.77. Another model diagnostic I look at is the Pearson correlation between the fitted values and the actual values. The Pearson correlation to NBA BPM in years three thru five for the entire model is 0.503. The following are the position models for each of the 3 positions, along with the fitted values from 2007 to 2018 by position.

The predictive model and results for guards who are rookies between 2006-07 and 2017-18 are presented in Table 5 and Figure 10. The VEcv for the Guard portion of the model is 12.39, and the Pearson correlation to NBA BPM in years three thru 5 is 0.36.

Table. 5. The Guard Model (when Guard = 1)

| Guard Model | | | |
|--------------------|---|--------------------|-----------------------|
| Rank | Variable | Coefficient | Cor to NBA BPM |
| 1 | BPM | 0.19901 | 0.404 |
| 2 | Tm BLK by Opp when player is shooting Short 2 vs KenPom Top 100 Teams : DBPM | 0.02536 | 0.171 |
| 3 | AST on 2-Pt Jump Shots vs KenPom Top 100 Teams: BPM | 0.00145 | 0.162 |

The predictive model and results for wings who are rookies between 2006-07 and 2017-18 are presented in Table 6 and Figure 11. The VECv for the wing portion of the model is 34.6 and the Pearson correlation to NBA BPM in years three thru 5 is 0.63.

Table. 6. The Wing Model (when Wing = 1)

| Wing Model | | | |
|-------------------|---|--------------------|-----------------------|
| Rank | Variable | Coefficient | Cor to NBA BPM |
| 1 | BPM | 0.19901 | 0.404 |
| 2 | BPM : DBPM | 0.05622 | 0.213 |
| 3 | OBPM :DBPM | 0.01799 | 0.196 |
| 4 | Non-Garbage AST for Short 2s : BPM | 0.00123 | 0.194 |
| 5 | Non-Garbage BLK by Opp when player is shooting 2-Pt Jump Shots : BPM | 0.00366 | 0.189 |
| 6 | Non-Garbage AST for Short 2s : DBPM | 0.00002 | 0.178 |

The predictive model and results for bigs who were rookies between 2006-07 and 2017-18 are presented in Table 7 and Figure 12. The VECv for the big portion of the model is 27.78, and the Pearson correlation to NBA BPM in years three thru 5 is 0.541.

Table. 7. The Big Model (when Big = 1)

| Big Model | | | |
|------------------|---|--------------------|---------------------------|
| Rank | Variable | Coefficient | Cor to NBA BPM |
| 1 | BPM | 0.19901 | 0.404 |
| 2 | BPM : Putbacks on 2-Pt Jump Shots vs KenPom Top 100 Teams | 0.01424 | 0.323 |
| 3 | BPM : Dean Oliver's Net Rating | 0.00079 | 0.305 |
| 4 | $\sqrt{AST\% * TRB\%}$: Dean Oliver's Net Rating | 0.00102 | 0.245 |
| 5 | DRB% : Dean Oliver's Net Rating | 0.00002 | 0.233 |
| 6 | AST% Putbacks on 2-Pt Jump Shots vs KenPom Top 100 Teams | 0.00200 | 0.229 |
| 7 | DRB% : Putbacks on 2-Pt Jump Shots vs KenPom Top 100 Teams | 0.00022 | 0.219 |

Though the model has 15 covariates total, since 14 of them only affect players with Guard = 1 or Wing = 1 or Big = 1, it is simple to think of it as 3 separate models by position. However, the predictive model and results for all players who were rookies between 2006-07 and 2017-18 are presented in Table 8 and the results in Figure 13.

Table. 8. The Full Model

| Overall Model | | | | |
|----------------------|-------------------------|--|--------------------|-----------------------|
| Rank | Players Included | Variable | Coefficient | Cor to NBA BPM |
| 1 | All Players | BPM | 0.19901 | 0.404 |
| 2 | Big | BPM : Putbacks on 2-Pt Jump Shots vs KenPom Top 100 Teams | 0.01424 | 0.323 |
| 3 | Big | BPM : Dean Oliver's Net Rating | 0.00079 | 0.305 |
| 4 | Big | $\sqrt{\text{AST}\% * \text{TRB}\%}$: Dean Oliver's Net Rating | 0.00102 | 0.245 |
| 5 | Big | DRB% : Dean Oliver's Net Rating | 0.00002 | 0.233 |
| 6 | Big | AST% Putbacks on 2-Pt Jump Shots vs KenPom Top 100 Teams | 0.00200 | 0.229 |
| 7 | Big | DRB% : Putbacks on 2-Pt Jump Shots vs KenPom Top 100 Teams | 0.00022 | 0.219 |
| 8 | Wing | BPM : DBPM | 0.05622 | 0.213 |
| 9 | Wing | OBPM :DBPM | 0.01799 | 0.196 |
| 10 | Wing | Non-Garbage AST for Short 2s : BPM | 0.00123 | 0.194 |
| 11 | Wing | Non-Garbage BLK by Opp when player is shooting 2-Pt Jump Shots : BPM | 0.00366 | 0.189 |
| 12 | Wing | Non-Garbage AST for Short 2s : DBPM | 0.00002 | 0.178 |
| 13 | Guard | Tm BLK by Opp when player is shooting Short 2 vs KenPom Top 100 Teams : DBPM | 0.02536 | 0.171 |
| 14 | Guard | AST on 2-Pt Jump Shots vs KenPom Top 100 Teams: BPM | 0.00145 | 0.162 |

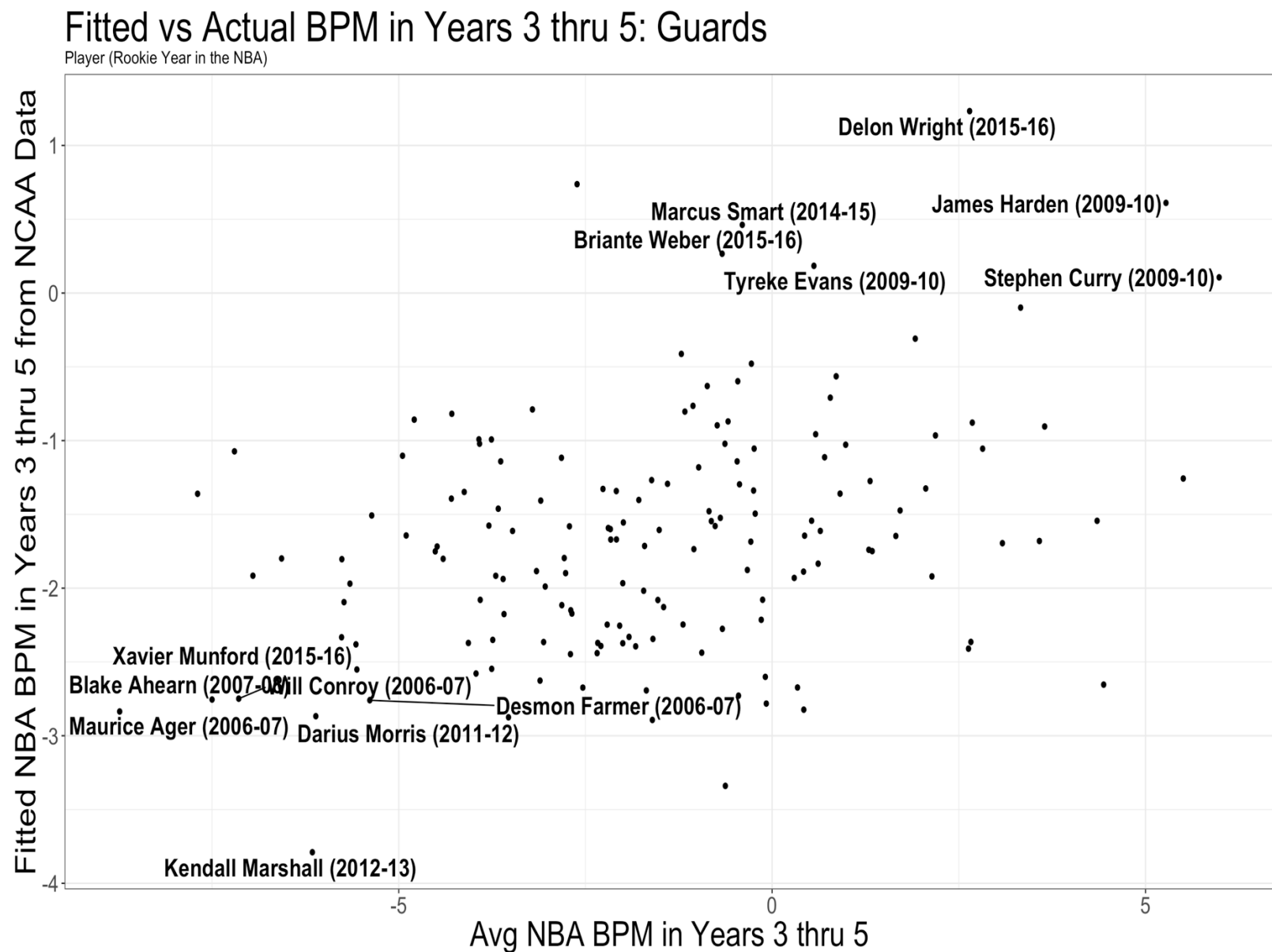


Fig. 10. Results for Guards from 2007 – 2018

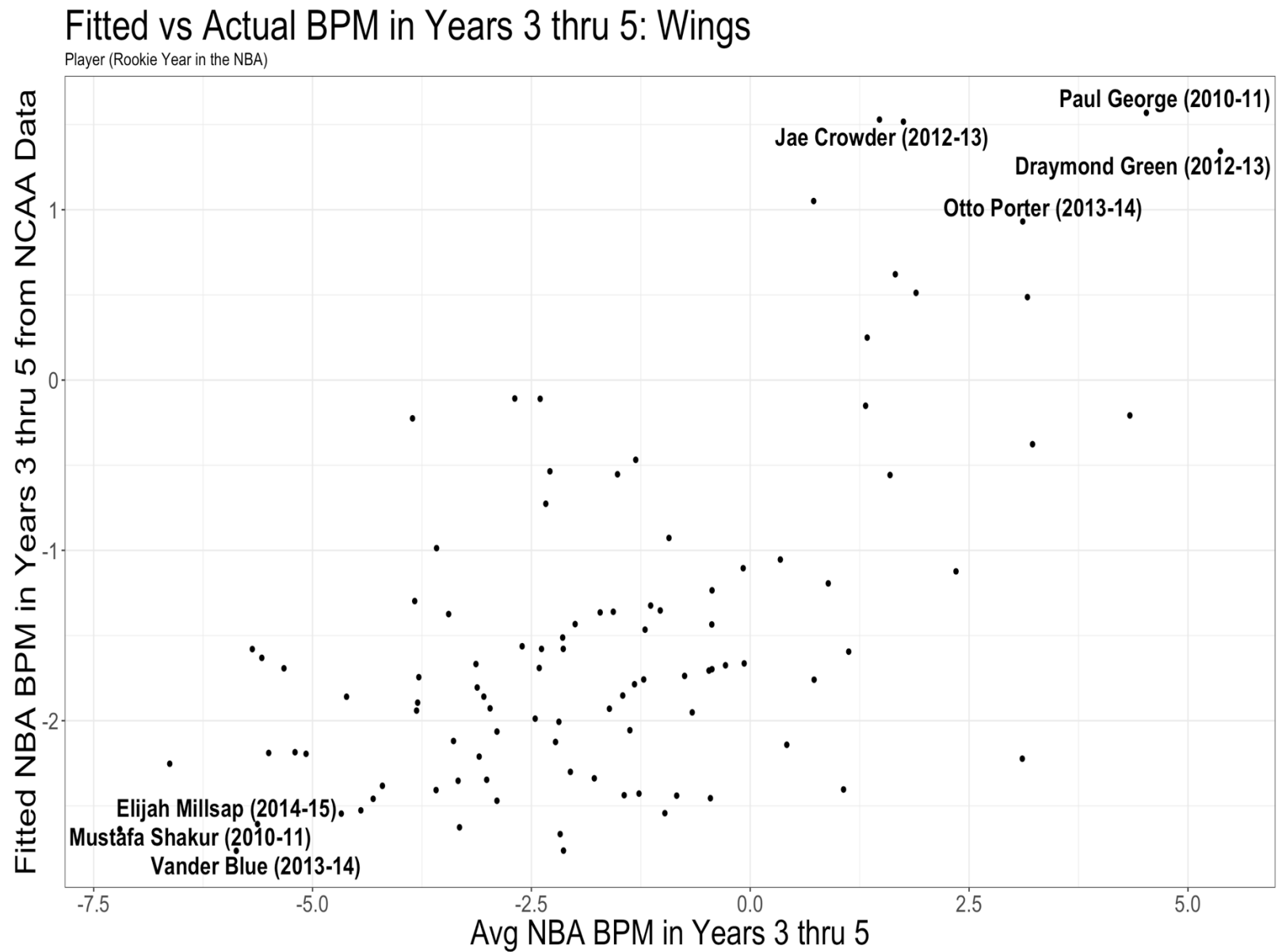


Fig. 11. Results for Wings from 2007 – 2018

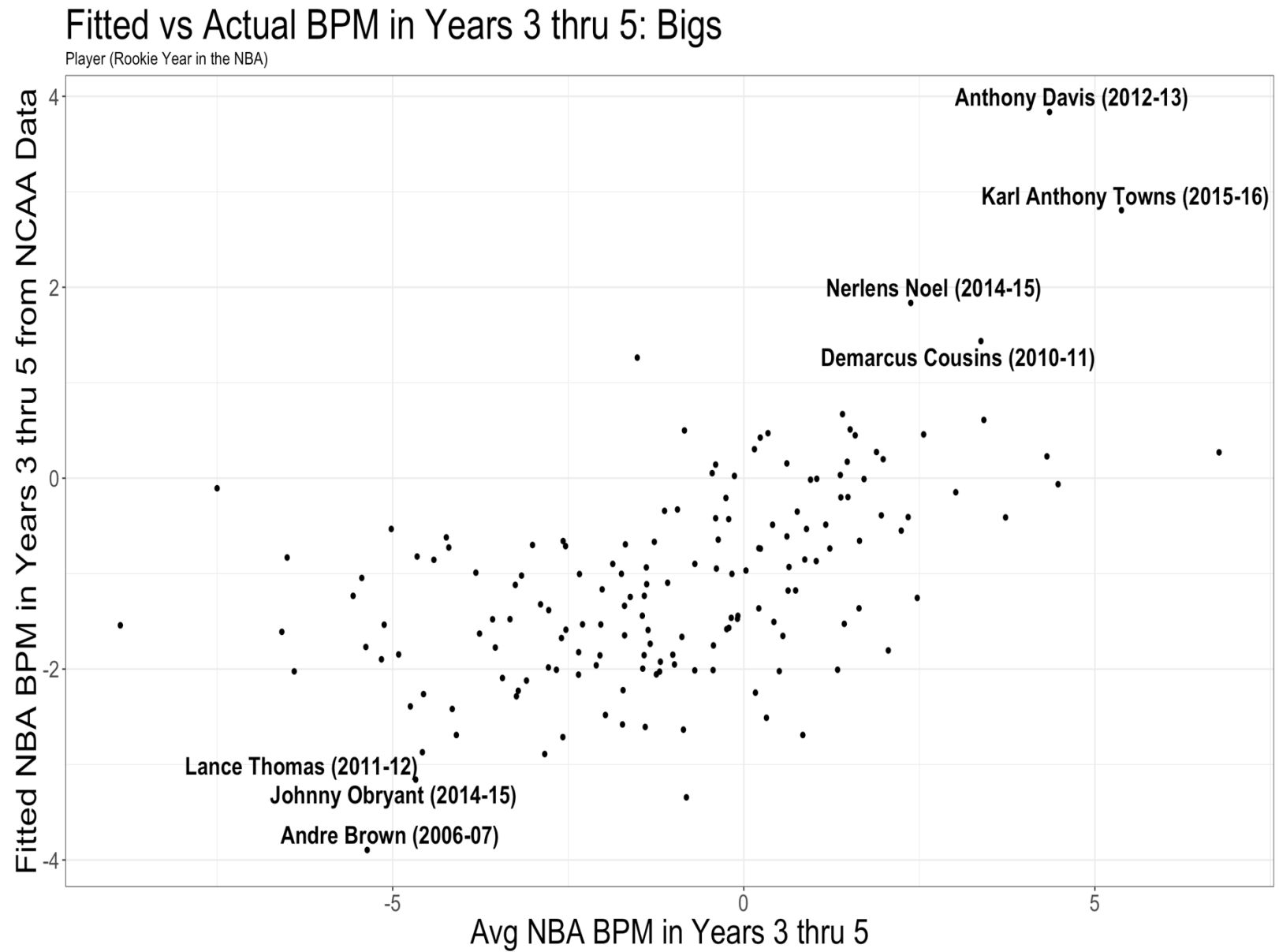


Fig. 12. Results for Bigs from 2007 - 2018

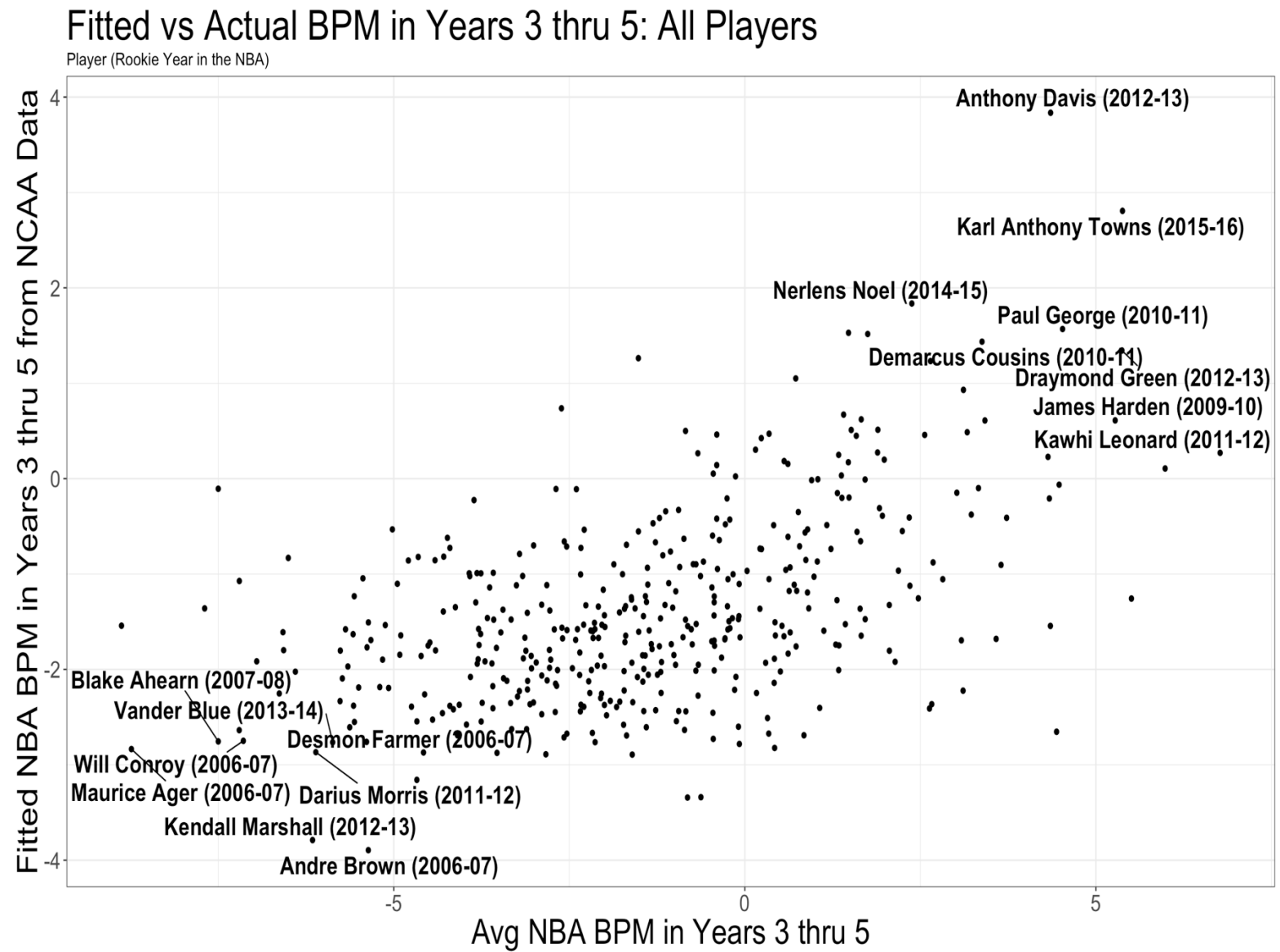


Fig. 13. Results for all players from 2007 - 2018

Chapter 5: Conclusion and Future Work

Conclusion

Given that we started with more than 30,000 variables, this model is simple and explainable. Further, it has good contextual information to be able to discuss with basketball experts and decision makers. It is critical having logic to split players into only three positions. Using three positions instead of five is a logical approach in modern basketball, and it is nice to keep larger sample sizes for finding the best player while still being able to find the best fit for the organization. This specific context of position and a simplistic model should provide a GM or President with good information, as analytics and basketball expertise in collaboration are invaluable in drafting the best players.

For example, in the 2014 draft, Marcus Smart is predicted to be good by this model. The biggest factor is that he is in the 98th percentile in College BPM in the last 11 years. The base statistic will be the answer to the question “is he efficient in college, taking into account his team and the NCAA” (BPM)? In his case, he is highly efficient based on BPM. But where play-by-play, specific context comes in, is that he is in the 97th percentile in assists for 2-pt jump shots against KenPom Top 100 teams. As scouts have watched him play, this will be good context to bring to the discussion. Our model fitted him to be efficient in the NBA, in part, because when playing one of the top 100 teams in NCAA basketball, he assisted 17.5 2-point jump shots per season while accruing a high BPM. This is extremely valuable to know that this specific, contextual information is a big part of our prediction heading into the draft. As there will be lots of detailed, basketball expert opinion brought to the discussion, I believe it is critical to walk into the room with specific context over a highly complex model such as non-parametric modeling.

They will have the basketball knowledge to give context on why we might or might not be overly eager on Smart's assists on 2-pt jump shots against KenPom Top 100 teams for specific basketball reasons after watching Marcus Smart and Oklahoma State play a considerable number of times. Considering that you are likely to have at most one high draft pick, all personnel needs to bring collaborative and specific information to the room. In doing so, the GM or President can more readily make a clear decision with all information readily at their fingertips.

Future Work

The data used goes back to rookies from 2007, as the 2006 NBA Draft was the first draft players could not enter directly from high school. Because of this, data from 2007 to present is similar, whereas, data before 2007 would have to account for players such as LeBron James and Kevin Garnett who went straight from high school to the NBA. An important addition will be normalizing international basketball data to predict NBA success. This would account for players like Dirk Nowitzki and Giannis Antetokounmpo. As EuroLeague has been the International league with the most NBA players, data can be regressed onto EuroLeague statistics before modeling NBA efficiency.

This is a great start to knowing what measures and statistical analyses can be used to build explainable models for the best player to award a maximum contract through the NBA Draft incorporating play-by-play and on-off data. These models are simple, clearly stated and can be used in an efficient manner with scout analyses to help a GM or President of an NBA team limit their risk in predicting a player's likelihood of becoming a Maximum Contract player who can help the organization win championships. As stated in the beginning, all of the recent NBA champions are built, in part, by drafting players that they could award maximum contracts

at the end of their 4-year, rookie contract. Hopefully this work can be used in tandem with scouting analyses to help an NBA team select the player in the draft that is most likely to receive a maximum contract in 4 years.

References

- Abrams, W., Barnes, J. and Clement, A. (2008). *Relationship of Selected Pre-NBA Career Variables to NBA Players' Career Longevity*. [online] The Sport Journal. Available at: <http://thesportjournal.org/article/relationship-of-selected-pre-nba-career-variables-to-nba-players-career-longevity/> [Accessed 15 Sep. 2018].
- Basketball-Reference.com. (2005). *Glossary*. [online] Available at: <https://www.basketball-reference.com/about/glossary.html> [Accessed 16 Sep. 2018].
- Basketball-Reference.com. (2005). *NBA Win Shares*. [online] Available at: <https://www.basketball-reference.com/about/ws.html> [Accessed 15 Sep. 2018].
- Basketball-Reference.com. (2013). Calculating Individual Offensive and Defensive Ratings. [online] Available at: <https://www.basketball-reference.com/about/ratings.html> [Accessed 18 Sep. 2018].
- Berri, D., Brook, S. and Fenn, A. (2010). From college to the pros: predicting the NBA amateur player draft. *Journal of Productivity Analysis*, 35(1), pp.25-35.
- Bošnjak, Z., Grljevic, O. and Bošnjak, S. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. In: *Proceedings of the International Conference on Mathematics Textbook Research and Development*. [online] Timisoara, Romania: IEEE, pp.159 - 160. Available at: https://www.researchgate.net/publication/224544887_CRISP-DM_as_a_framework_for_discovering_knowledge_in_small_and_medium_sized_enterprises_data/citations [Accessed 15 Sep. 2018].
- Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, [online] 45(3), pp.1-2. Available at: <https://www.js-tatsoft.org/article/view/v045i03> [Accessed 15 Sep. 2018].
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). *CRISP-DM 1.0*. [online] The-modeling-agency.com. Available at: <https://www.the-modeling-agency.com/crisp-dm.pdf> [Accessed 15 Sep. 2018].
- ESPN.com. (2018). *Oklahoma vs. Wichita St - Play-By-Play - December 16, 2017 - ESPN*. [online] Available at: <http://www.espn.com/mens-college-basketball/playbyplay?gameId=400988013> [Accessed 15 Sep. 2018].

- Evans, B. (2017). From college to the NBA: what determines a player's success and what characteristics are NBA franchises overlooking?. *Applied Economics Letters*, 25(5), pp.300-304.
- Femrite, M. (2017). *Nylon Calculus: Possessions have been overestimated lately*. [online] FanSided. Available at: <https://fansided.com/2017/05/17/nylon-calculus-posessions-overestimated/> [Accessed 15 Sep. 2018].
- Fonti, V. (2017). *Feature Selection using LASSO*. [ebook] Amsterdam, Netherlands: VU Amsterdam, p.8. Available at: https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf [Accessed 15 Sep. 2018].
- Goldstein, J. (2018). *Introducing Player Impact Plus-Minus*. [online] FanSided. Available at: <https://fansided.com/2018/01/11/nylon-calculus-introducing-player-impact-plus-minus/> [Accessed 22 Sep. 2018].
- Groothuis, P., Hill, J. and Perri, T. (2005). *Early Entry in the NBA Draft: The Influence of Unraveling, Human Capital, and Option Value*. [online] Boone, NC: Appalachian State University, pp.6-30. Available at: <http://www.appstate.edu/~perritj/earlynba.pdf> [Accessed 15 Sep. 2018].
- Hall, M. (1998). *Correlation-based Feature Selection for Machine Learning*. [online] Hamilton, New Zealand: The University of Waikato, p.149. Available at: <https://www.cs.waikato.ac.nz/~mhall/thesis.pdf> [Accessed 15 Sep. 2018].
- Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*. 2nd ed. [ebook] New York: Springer, pp.68-69. Available at: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf [Accessed 29 Sep. 2018].
- Kenpom.com. (2018). 2018 Pomeroy College Basketball Ratings. [online] Available at: <https://kenpom.com/> [Accessed 18 Sep. 2018].
- Li, J., Alvarez, B., Siwabessy, J., Tran, M., Huang, Z., Przeslawski, R., Radke, L., Howard, F. and Nichol, S. (2017). Selecting predictors to form the most accurate predictive model for count data. In: *International Congress on Modelling and Simulation*. [online] Canberra, ACT, pp.13-14. Available at: https://www.researchgate.net/publication/321921431_Selecting_predictors_to_form_the_most_accurate_predictive_model_for_count_data [Accessed 16 Sep. 2018].
- McCallum, J. (2018). <https://www.si.com>. [online] SI.com. Available at: <https://www.si.com/nba/2016/02/09/michael-jordan-lebron-james-stephen-curry-nba-greatest> [Accessed 27 Sep. 2018].

- Moxley, J. and Towne, T. (2015). Predicting success in the National Basketball Association: Stability & potential. *Psychology of Sport and Exercise*, 16, pp.128-136.
- Myers, D. (2014). *About Box Plus/Minus (BPM)*. [online] Basketball-Reference.com. Available at: <https://www.basketball-reference.com/about/bpm.html> [Accessed 15 Sep. 2018].
- Nba.com. (2001). *RULE NO. 10-VIOLATIONS AND PENALTIES*. [online] Available at: http://www.nba.com/analysis/rules_10.html [Accessed 15 Sep. 2018].
- Nba.com. (2005). *NBA Collective Bargaining Agreement Ratified and Signed*. [online] Available at: http://www.nba.com/news/CBA_050730.html [Accessed 15 Sep. 2018].
- Nba.com. (2008). *NBA Rules History*. [online] Available at: http://www.nba.com/analysis/rules_history.html [Accessed 15 Sep. 2018].
- NBA.com. (2014). *Stats Glossary*. [online] Available at: <https://stats.nba.com/help/glossary/> [Accessed 15 Sep. 2018].
- Nicholson, C. (2015). ISE 5103. Class Lecture, Topic: *Data Understanding*.
- Nicholson, C. (2015). ISE 5103. Class Lecture, Topic: *Principles of Modeling*.
- Oliver, D. (2004). *Basketball on paper*. Washington, D.C.: Potomac Books, Inc., pp.1-28.
- Paine, N. (2013). Is WP a legitimate stat?. [Blog] *Is WP a legitimate stat?*. Available at: <http://apbr.org/metrics/viewtopic.php?p=15334#p15334> [Accessed 15 Sep. 2018].
- Paine, N. (2013). *Is WP a legitimate stat? - Page 2 - APBRmetrics*. [online] Apbr.org. Available at: <http://apbr.org/metrics/viewtopic.php?p=15334#p15334> [Accessed 15 Sep. 2018].
- Realgm.com. (2018). *RealGM*. [online] Available at: <https://www.realgm.com> [Accessed 15 Sep. 2018].
- Sailofsky, D. (2018). Drafting Errors and Decision Making Bias in the NBA. In: *MIT Sloan Sports Analytics Conference*. [online] Boston, MA: MIT Sloan Sports Analytics Conference, pp.4-7. Available at: <http://www.sloansportsconference.com/wp-content/uploads/2018/02/2004.pdf> [Accessed 15 Sep. 2018].
- Sampaio, J., Ibáñez, S., Ruano, M., Calvo, A. and Ortega, E. (2013). Game location influences basketball players' performance across playing positions. *International Journal of Sports Psychology*, 39(3), p.3.
- Sampaio, J., Janeira, M., Ibáñez, S. and Lorenzo, A. (2006). Discriminant analysis of game-related statistics between basketball guards, forwards and centres in three professional leagues. *European Journal of Sport Science*, [online] 6(3), pp.173-178. Available at: https://www.researchgate.net/publication/232908563_Discriminant_analysis_of_game-

related_statistics_between_basketball_guards_forwards_and_centres_in_three_professiona
l_leagues [Accessed 15 Sep. 2018].

Schreefer, W. (2018). *College Basketball 'Draft Model Starter Kit' Database*. [online] thestepien.com. Available at: <https://www.thestepien.com/2018/05/15/college-basketball-draft-model-starter-kit-database/> [Accessed 15 Sep. 2018].

Shea, S. and Baker, C. (2013). *Basketball Analytics*. St. Louis, MO: Advanced Metrics, LLC, pp.77-78.

Wagesofwins.com. (2012). *How to calculate Wins Produced*. [online] Available at: <http://wagesofwins.com/how-to-calculate-wins-produced/> [Accessed 15 Sep. 2018].

Wagner, K. (2014). *Just What The Hell is Real Plus-Minus, ESPN's New NBA Stat?*. [online] Deadspin.com. Available at: <https://deadspin.com/just-what-the-hell-is-real-plus-minus-espns-new-nba-s-1560361469> [Accessed 15 Sep. 2018].

Appendix

ERD for Database: Play-by-Play, On-Off and BPM

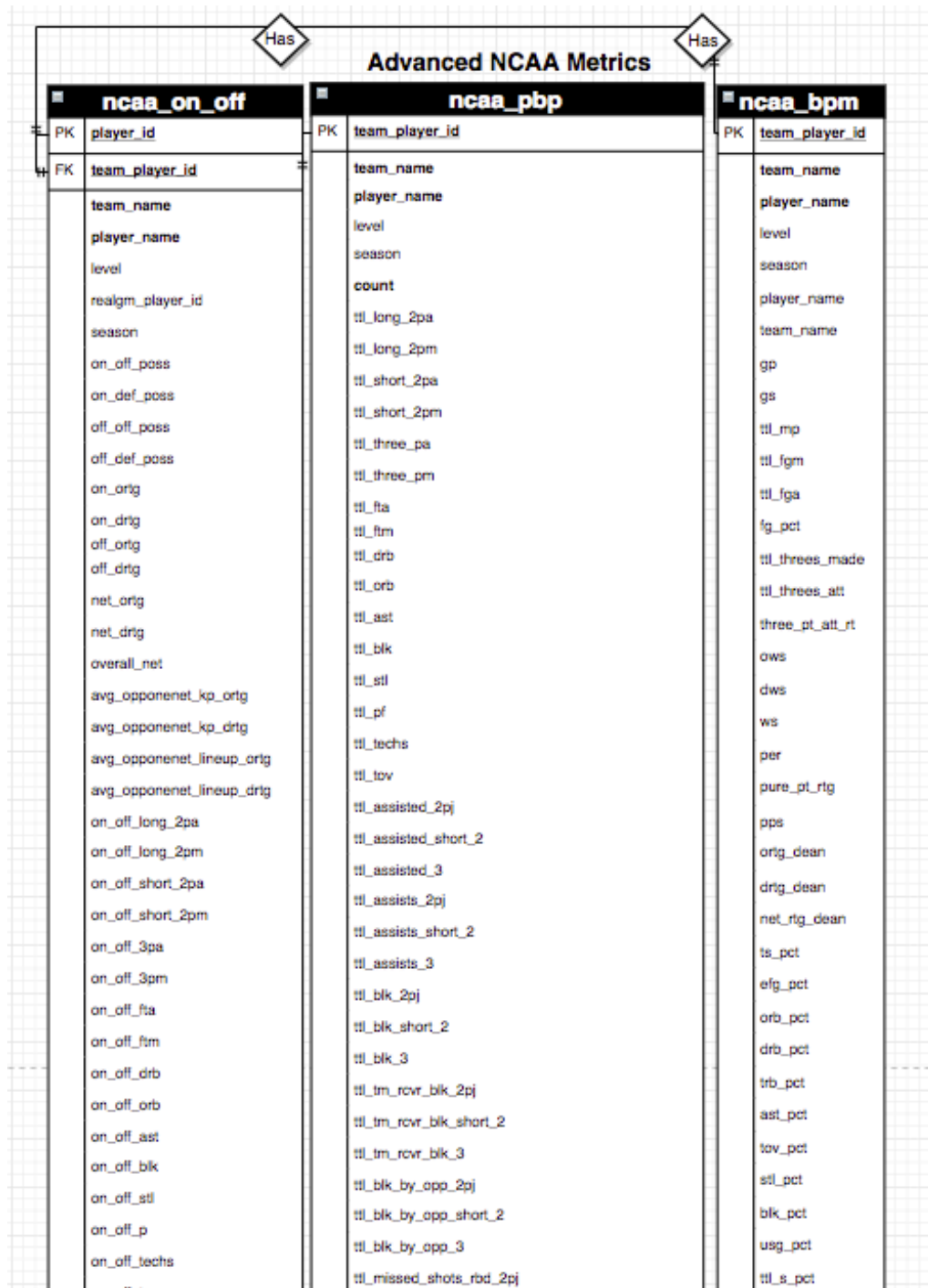


Fig. A1. Excerpt from Entity Relationship Diagram for database

| | | |
|-------------------|-----------------------------------|--------------------|
| on_off_to | ttl_missed_shots_rbd_2pj | ttl_s_pct |
| on_def_long_2pa | ttl_missed_shots_rbd_short_2 | team_ts_pct |
| on_def_long_2pm | ttl_missed_shots_rbd_3 | lg_three_pt_att_rt |
| on_def_short_2pa | ttl_tm_and_1_unast_2pj | player_pct_of_min |
| on_def_short_2pm | ttl_tm_and_1_unast_short_2 | bpm_realgm |
| on_def_3pa | ttl_tm_and_1_unast_3 | obpm_realgm |
| on_def_3pm | ttl_tm_and_1_ast_2pj | dbpm_realgm |
| on_def_fta | ttl_tm_and_1_ast_short_2 | obpm_sref |
| on_def_ftm | ttl_tm_and_1_ast_3 | dbpm_sref |
| on_def_drb | ttl_opp_and_1_surrendered_2pj | bpm_sref |
| on_def_orb | ttl_opp_and_1_surrendered_short_2 | |
| on_def_ast | ttl_opp_and_1_surrendered_3 | |
| on_def_blk | ttl_putbacks_2pj | |
| on_def_stl | ttl_putbacks_short_2 | |
| on_def_pf | ttl_putbacks_3 | |
| on_def_techs | ttl_putbacks_ft | |
| on_def_tov | ttl_putbacks_blk | |
| off_off_long_2pa | ttl_exact_fouls_drawn_2 | |
| off_off_long_2pm | ttl_exact_fouls_drawn_3 | |
| off_off_short_2pa | ttl_exact_fouls_drawn_and_1 | |
| off_off_short_2pm | ttl_dunks_made | |
| off_off_3pa | ttl_dunks_missed | |
| off_off_3pm | ttl_dunks_assisted | |
| off_off_fta | ttl_tov_foul | |
| off_off_ftm | ttl_tov_stl | |
| off_off_drb | ttl_tov_misc | |
| off_off_orb | trans_within_10s_long_2pa | |
| off_off_ast | trans_within_10s_long_2pm | |
| off_off_blk | trans_within_10s_short_2pa | |
| off_off_stl | trans_within_10s_short_2pm | |
| off_off_pf | trans_within_10s_3pa | |
| off_off_techs | trans_within_10s_3pm | |
| off_off_tov | trans_within_10s_fta | |
| off_def_long_2pa | trans_within_10s_ftm | |
| off_def_long_2pm | trans_within_10s_drb | |
| off_def_short_2pa | trans_within_10s_orb | |
| off_def_short_2pm | trans_within_10s_ast | |
| off_def_3pa | trans_within_10s_blk | |
| off_def_3pm | trans_within_10s_stl | |
| off_def_fta | trans_within_10s_pf | |
| off_def_ftm | trans_within_10s_techs | |
| off_def_drb | trans_within_10s_tov | |
| | trans_within_10s_assisted_2pj | |

Fig. A1 (Continued). Excerpt from Entity Relationship Diagram for database

| | |
|---------------------------|--|
| off_def_orb | trans_within_10s_assisted_short_2 |
| off_def_ast | trans_within_10s_assisted_3 |
| off_def_blk | trans_within_10s_assists_2pj |
| off_def_stl | trans_within_10s_assists_short_2 |
| off_def_pt | trans_within_10s_assists_3 |
| off_def_techs | trans_within_10s_blk_2pj |
| off_def_tov | trans_within_10s_blk_short_2 |
| jacobs_adj_on_off_poss | trans_within_10s_blk_3 |
| jacobs_adj_off_off_poss | trans_within_10s_tm_rcvr_blk_2pj |
| jacobs_adj_on_def_poss | trans_within_10s_tm_rcvr_blk_short_2 |
| jacobs_adj_off_def_poss | trans_within_10s_tm_rcvr_blk_3 |
| jacobs_adj_on_orlg | trans_within_10s_blk_by_opp_3 |
| jacobs_adj_off_orlg | trans_within_10s_blk_by_opp_short_2 |
| jacobs_adj_on_drtg | trans_within_10s_blk_by_opp_2 |
| jacobs_adj_off_drtg | trans_within_10s_missed_shots_reb_2pj |
| jacobs_adj_offense_on_off | trans_within_10s_missed_shots_reb_short_2 |
| jacobs_adj_defense_on_off | trans_within_10s_missed_shots_reb_3 |
| jacobs_adj_on_off | trans_within_10s_tm_and_1_unast_2pj |
| wills_check | trans_within_10s_tm_and_1_unast_short_2 |
| | trans_within_10s_tm_and_1_unast_3 |
| | trans_within_10s_tm_and_1_ast_2pj |
| | trans_within_10s_tm_and_1_ast_short_2 |
| | trans_within_10s_tm_and_1_ast_3 |
| | trans_within_10s_opp_and_1_surrendered_2pj |
| | trans_within_10s_opp_and_1_surrendered_short_2 |
| | trans_within_10s_opp_and_1_surrendered_3 |
| | trans_within_10s_putbacks_2pj |
| | trans_within_10s_putbacks_short_2 |
| | trans_within_10s_putbacks_3 |
| | trans_within_10s_putbacks_ft |
| | trans_within_10s_putbacks_blk |
| | trans_within_10s_exact_fouls_drawn_2 |
| | trans_within_10s_exact_fouls_drawn_3 |
| | trans_within_10s_exact_fouls_drawn_and_1 |
| | trans_within_10s_dunks_made |
| | trans_within_10s_dunks_missed |
| | trans_within_10s_dunks_assisted |
| | trans_within_10s_tov_foul |
| | trans_within_10s_tov_stl |
| | trans_within_10s_tov_misc |
| | v_kp_top100_long_2ps |
| | v_kp_top100_long_2pm |

Fig. A1 (Continued). Excerpt from Entity Relationship Diagram for database

| | | |
|--|---|--|
| | v_kp_top100_short_2pa | |
| | v_kp_top100_short_2pm | |
| | v_kp_top100_short_3pa | |
| | v_kp_top100_3pm | |
| | v_kp_top100_fta | |
| | v_kp_top100_ftm | |
| | v_kp_top100_drb | |
| | v_kp_top100_orb | |
| | v_kp_top100_ast | |
| | v_kp_top100_blk | |
| | v_kp_top100_stl | |
| | v_kp_top100_pf | |
| | v_kp_top100_techs | |
| | v_kp_top100_tov | |
| | v_kp_top100_assisted_2pj | |
| | v_kp_top100_assisted_short_2 | |
| | v_kp_top100_assisted_3 | |
| | v_kp_top100_assists_2pj | |
| | v_kp_top100_assists_short_2 | |
| | v_kp_top100_assists_3 | |
| | v_kp_top100_blk_2pj | |
| | v_kp_top100_blk_short_2 | |
| | v_kp_top100_blk_3 | |
| | v_kp_top100_tm_rcvr_blk_2pj | |
| | v_kp_top100_tm_rcvr_blk_short_2 | |
| | v_kp_top100_tm_rcvr_blk_3 | |
| | v_kp_top100_tm_blk_by_app_2pj | |
| | v_kp_top100_tm_blk_by_app_short_2 | |
| | v_kp_top100_tm_blk_by_app_3 | |
| | v_kp_top100_missed_shots_reb_2pj | |
| | v_kp_top100_missed_shots_reb_short_2 | |
| | v_kp_top100_missed_shots_reb_3 | |
| | v_kp_top100_tm_and_1_unast_2pj | |
| | v_kp_top100_tm_and_1_unast_short_2 | |
| | v_kp_top100_tm_and_1_unast_3 | |
| | v_kp_top100_tm_and_1_ast_2pj | |
| | v_kp_top100_tm_and_1_ast_short_2 | |
| | v_kp_top100_tm_and_1_ast_3 | |
| | v_kp_top100_opp_and_1_surrendered_2pj | |
| | v_kp_top100_opp_and_1_surrendered_short_2 | |
| | v_kp_top100_opp_and_1_surrendered_3 | |
| | v_kp_top100_putbacks_2pj | |
| | v_kp_top100_putbacks_short_2 | |

Fig. A1 (Continued). Play-by-pay table continued

| | | |
|--|-------------------------------------|--|
| | v_kp_top100_putbacks_3 | |
| | v_kp_top100_putbacks_ft | |
| | v_kp_top100_putbacks_blk | |
| | v_kp_top100_exact_fouls_drawn_2 | |
| | v_kp_top100_exact_fouls_drawn_3 | |
| | v_kp_top100_exact_fouls_drawn_and_1 | |
| | v_kp_top100_dunks_made | |
| | v_kp_top100_dunks_missed | |
| | v_kp_top100_dunks_assisted | |
| | v_kp_top100_tov_foul | |
| | v_kp_top100_tov_stl | |
| | v_kp_top100_tov_misc | |
| | non_garbage_long_2pa | |
| | non_garbage_long_2pm | |
| | non_garbage_short_2pa | |
| | non_garbage_short_2pm | |
| | non_garbage_3pa | |
| | non_garbage_3pm | |
| | non_garbage_fta | |
| | non_garbage_ftm | |
| | non_garbage_drb | |
| | non_garbage_orb | |
| | non_garbage_ast | |
| | non_garbage_blk | |
| | non_garbage_stl | |
| | non_garbage_pf | |
| | non_garbage_techs | |
| | non_garbage_tov | |
| | non_garbage_assisted_2pj | |
| | non_garbage_assisted_short_2 | |
| | non_garbage_assisted_3 | |
| | non_garbage_assists_2pj | |
| | non_garbage_assists_short_2 | |
| | non_garbage_assists_3 | |
| | non_garbage_blk_2pj | |
| | non_garbage_blk_short_2 | |
| | non_garbage_blk_3 | |
| | non_garbage_tm_rcvr_blk_2pj | |
| | non_garbage_tm_rcvr_blk_short_2 | |
| | non_garbage_tm_rcvr_blk_3 | |
| | non_garbage_blk_by_opp_2pj | |
| | non_garbage_blk_by_opp_short_2 | |

Fig. A1 (Continued). Play-by-pay table continued

| | | |
|--|---|--|
| | non_garbage_blk_by_opp_short_2 | |
| | non_garbage_blk_by_opp_3 | |
| | non_garbage_missed_shots_reb_2pj | |
| | non_garbage_missed_shots_reb_short_2 | |
| | non_garbage_missed_shots_reb_3 | |
| | non_garbage_tm_and_1_unast_2pj | |
| | non_garbage_tm_and_1_unast_short_2 | |
| | non_garbage_tm_and_1_unast_3 | |
| | non_garbage_tm_and_1_ast_2pj | |
| | non_garbage_tm_and_1_ast_short_2 | |
| | non_garbage_tm_and_1_ast_3 | |
| | non_garbage_opp_and_1_surrendered_2pj | |
| | non_garbage_opp_and_1_surrendered_short_2 | |
| | non_garbage_opp_and_1_surrendered_3 | |
| | non_garbage_putbacks_2pj | |
| | non_garbage_putbacks_short_2 | |
| | non_garbage_putbacks_3 | |
| | non_garbage_putbacks_ft | |
| | non_garbage_putbacks_blk | |
| | non_garbage_exact_fouls_drawn_2 | |
| | non_garbage_exact_fouls_drawn_3 | |
| | non_garbage_exact_fouls_drawn_and_1 | |
| | non_garbage_dunks_made | |
| | non_garbage_dunks_missed | |
| | non_garbage_dunks_assisted | |
| | non_garbage_tov_foul | |
| | non_garbage_tov_stl | |
| | non_garbage_tov_misc | |
| | transition_blk | |
| | own_miss_putbacks | |
| | dunk_putbacks | |
| | pick_6s | |
| | putback_3_reb | |
| | dunk_assists | |
| | late_shotclock_long_2pa | |
| | late_shotclock_long_2pm | |
| | late_shotclock_short_2pa | |
| | late_shotclock_short_2pm | |
| | late_shotclock_3pa | |
| | late_shotclock_3pm | |
| | transition_blk_2pj | |
| | transition_blk_short_2 | |

Fig. A1 (Continued). Play-by-pay table continued

| |
|--|
| transition_blk_3 |
| transition_tm_rcvr_blk_2pj |
| transition_tm_rcvr_blk_short_2 |
| transition_tm_rcvr_blk_3 |
| transition_opp_and_1_surrendered_2pj |
| transition_opp_and_1_surrendered_short_2 |
| transition_opp_and_1_surrendered_3 |

Fig. A1 (Continued). Play-by-pay table continued

This database is built and used to query data, tie it together and build out a design matrix. The analysis is all built from this data. The database is largely populated by Schreefer's data (Schreefer, 2018) and scraping basketball-reference.com.