ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

School of Engineering
Master Degree in Artificial Intelligence

ADVANCED BASKETBALL ANALYTICS

Thesis in
Machine Learning

Supervisor
Chiar.mo Prof.Ing. Claudio Sartori

Co-supervisor
Dr. Sergio Scariolo

Candidate
Francesco Olivo

Fifth Session
Academic Year 2022-23

Keywords

Basketball Analytics
Plus-Minus
Euroleague
RAPM
Elastic Net Regression

Index

Introduction

The realm of sports has witnessed a remarkable transformation over the past twenty-five years, a period characterized by groundbreaking developments. This evolution in sports is multifaceted, attributed to advancements in several key areas: the enhancement of athletic performance, a deeper understanding of medical knowledge to bolster athletes' conditions, significant improvements in sports equipment, and the pivotal role of sports analytics.

The concept of sports analytics is deeply rooted in the history of modern sports, dating back to its early days[1]. This discipline emerged from the fundamental need to quantify team and individual performances in sports, offering a precise and objective lens. While sports analytics have always been integral to sports, their modern application is a more recent phenomenon, originating in the early 21st century.

A quintessential example of this modern incarnation of sports analytics is story of the Oakland Athletics[2], a baseball team that in 2002 dramatically enhanced its performance compared to previous seasons by adopting a sabermetric approach to player evaluation. This strategy, which marked a significant departure from traditional methods, set a precedent in the sports world.

The Oakland Athletics' success story has served as a catalyst, inspiring sports directors and managers across various disciplines to embrace statistical and analytical methods. This trend has extended beyond baseball, influencing sports like ice hockey, basketball, and eventually, football and soccer. It's noteworthy how sports analytics initially flourished in baseball, a sport that lends itself well to discrete mathematical modeling. Baseball actions, primarily involving the pitcher and the hitter, can be easily quantified and analyzed.

Conversely, sports such as basketball and ice hockey pose greater challenges for analytics due to the dynamic interactions of multiple players, each contributing to the team's success in less quantifiable ways. These sports can be seen as *continuous* from a mathematical perspective, a complexity further amplified in sports like soccer and football, where the number of players and the fluidity of

play increase significantly.

The transformation of the sports landscape through analytics has been predominantly an American phenomenon, heavily influenced by the unique synergy between sports and academia in U.S. colleges. This relationship has flourished in an environment where sports are not just a form of entertainment but an integral part of cultural and educational systems. The U.S. boasts one of the richest sports environments globally, notably distinct from its European and Asian counterparts.

In the United States, the *Big Four* sports, American football, baseball, basketball, and ice hockey, have been the primary benefactors and drivers of sports analytics. However, these sports do not enjoy the same level of popularity worldwide. In contrast, soccer reigns supreme in Europe and South America, while cricket is the sport of choice in India and Oceania.

There are exceptions, such as China and the Philippines, where basketball has garnered immense popularity, and Japan, where baseball enjoys a similar status. Yet, these regional preferences have not significantly influenced the global sports analytics landscape. This disparity is particularly pronounced in Europe, where the implementation of advanced analytics in sports remains relatively nascent, especially in sports other than soccer.

The landscape of basketball analytics in Europe, when compared to the advancements in the United States, particularly in the NBA and NCAA, shows a notable gap. This disparity can be attributed to a mix of financial and cultural factors, which often limit the ability of European clubs outside the NBA to implement long-term strategies fully leveraging the benefits of analytics.

Despite these challenges, there are commendable exceptions in the European basketball scene that have embraced analytics to enhance their strategies and performances. In Italy, teams like Virtus Bologna, Pallacanestro Varese, and Pallacanestro Trieste have shown a keen interest in incorporating data-driven approaches. France's Paris Basketball is another notable example, as are Baskets Bonn in Germany and the Spanish National team. These teams have been at the forefront of integrating advanced analytics into their operations, setting a precedent within their respective leagues and countries.

The adoption of analytics by these teams represents a significant step forward in the evolution of basketball strategy and management in Europe. It is a positive sign that paves the way for a broader acceptance and integration of analytics in European basketball. As more teams witness the benefits reaped by these early adopters, it is hoped that the coming years will see a growing number of European teams joining this trend. Such a development would not

only enhance the competitive landscape but also contribute to the enrichment and expansion of the culture of analytics in European basketball, bringing it closer to the sophistication seen in American basketball analytics.

Chapter 1

Background

1.1 Review of Basketball Analytics

For many years, basketball analytics primarily revolved around the statistical interpretation of the traditional box score. The box score, as exemplified in Table 1.1, provides a basic framework to understand game events, albeit in a limited scope.

Player	S5	MIN	PTS	2FGM	2FGA	2FG%	3FGM	3FGA	3FG%	FTM	FTA	FT%	DREB	OREB	REB	AST	STL	TOV	PF	BLK	+/-
M. Jordan	1	43:41	45	12	28	42.9%	3	7	42.9%	12	15	80%	1	0	1	1	4	1	2	0	2
T. Kukoc	1	42:06	15	6	12	50%	1	2	50%	0	0		3	0	3	4	0	0	3	0	8
D. Rodman	0	38:59	7	3	3	100%	0	0		1	2	50%	4	4	8	1	2	2	5	1	5
R. Harper	1	28:34	8	3	3	100%	0	1	0%	2	2	100%	3	0	3	3	1	1	2	2	5
S. Pippen	1	25:43	8	4	7	57.1%	0	0		0	0		3	0	3	4	2	2	2	1	16
S. Kerr	0	24:05	0	0	0		0	0		0	0		0	0	0	3	1	1	3	0	-3
L. Longley	1	14:34	0	0	1	0%	0	0		0	0		2	0	2	0	1	0	4	0	-4
S. Burrell	0	10:18	0	0	1	0%	0	0		0	0		0	0	0	0	0	0	0	0	-17
J. Buechler	0	8:00	2	1	1	100%	0	0		0	0		1	1	2	1	0	0	1	0	3
B. Wennington	0	4:00	2	1	1	100%	0	0		0	0		0	0	0	0	0	2	1	0	-10
Total	0	240:00	87	30	57	52.6%	4	10	40%	15	19	78.9%	17	5	22	17	11	9	23	4	2

Table 1.1: Boxscore of the Chicago Bulls in the 6th game of 1998 NBA Finals

A significant shortcoming of the box score format lies in its bias towards offensive statistics. Defensive contributions are often underrepresented, with metrics like steals, blocks, defensive rebounds, and to a certain extent, fouls. However, these stats do not comprehensively capture defensive prowess. Fouls, in particular, can be ambiguous, not clearly delineating between offensive and defensive actions.

This bias is rooted in historical practices as well as the inherent challenges in objectively quantifying defensive play. For instance, while an offensive shot is a discrete action directly attributed to a player, its defensive counterpart, such as a contested shot, lacks such straightforward attribution. This ambiguity extends to steals, where credit is often given to the player who ultimately secures the

ball, ignoring the efforts of players who may have facilitated the turnover. This is also true, up to a certain degree, for some offensive contributions such as screens, or for rebounds.

Another complexity in measuring defensive performance is the collaborative nature of professional basketball defense. Unlike certain offensive statistics, which can be attributed to individual efforts, defense typically requires coordinated team effort, making it challenging to isolate and quantify individual contributions

Recent advancements, particularly in the NBA, have seen the integration of Computer Vision-based tagging systems that offer more nuanced insights into defensive play. These systems can track player movements and interactions more precisely, providing a richer dataset to analyze defensive effectiveness. However, such advanced analytical tools are predominantly limited to the NBA, leaving a gap in the analytical capabilities in other leagues, including European basketball.

1.1.1 Measuring efficiency using possessions

One of the most significant contributions to basketball analytics was made by Dean Oliver in 2004. In his seminal book Basketball on Paper[3], Oliver proposed a novel method for measuring efficiency, shifting the focus from a game-based normalization to a possession-based approach. A possession in basketball is defined as a sequence of consecutive events where a team controls the ball, concluding with a made shot, a missed shot rebounded by the opponents, or a turnover. This method of analyzing data *per possession* offers a more insightful analysis by inherently accounting for varying styles of play, especially differences in playing speeds.

The concept of measuring a team's pace of play is particularly insightful. Pace is quantified by normalizing the number of possessions per 48 minutes, the duration of an NBA game. This normalization uses minutes instead of games to accommodate for variations in game lengths, such as those extended by overtime periods. When a team does not play in overtime, the pace and number of possessions effectively coincide. The formula for calculating possessions (POSS) and pace (PACE) are as follows:

$$POSS = FGA + 0.44 * FTA + TOV - OREB \tag{1.1}$$

$$PACE = 48 * \frac{POSS}{MIN} \tag{1.2}$$

The 0.44 factor is applied to the number of free throw attempts (FTA) to estimate how many of those attempts actually end a possession. The rationale behind this factor is that not every free throw attempt results in the conclusion of a possession. In basketball, various scenarios, such as and-one plays, technical foul shots, or the first shot of a set of free throws, may lead to the continuation of the possession after the free throw.

When dealing with traditional box score data, which doesn't provide granular details about each play, applying the 0.44 multiplier to FTAs becomes a necessary approximation. It serves as a practical means to estimate the number of possessions used, given the limitations in the level of detail available in box scores.

However, with the availability of play-by-play (PBP) data, the dynamics of possession counting change: PBP data offers a detailed record of every event in a game, including each free throw attempt. By using PBP data, analysts can accurately determine whether a specific free throw concluded a possession or not. This precise information allows for a more exact count of possessions, moving beyond the approximations required when using box scores.

Similar to the definition of possessions is the definition of *plays*, which nonetheless do not account for the number of captured offensive rebounds. Differently from a possession, a play counts all of the single scoring opportunities that a team has:

$$PLAYS = FGA + 0.44 * FTA + TOV \tag{1.3}$$

Oliver's approach led to the development of three crucial metrics that qualitatively evaluate the effectiveness of offenses and defenses, rather than merely quantifying scores or points allowed per game. These metrics are the Offensive Rating (OFFRTG), Defensive Rating (DEFRTG), and Net Rating (NETRTG). They provide a measure of a team's ability to score or defend within a single possession. The formulas for these metrics normalize team efficiency per 100 possessions, aligning with the average number of possessions in an NBA game:

$$OFFRTG = \frac{100 * PTS_{team}}{POSS_{team}} \tag{1.4}$$

$$DEFRTG = \frac{100 * PTS_{opp}}{POSS_{opp}} \tag{1.5}$$

$$NETRTG = OFFRTG - DEFRTG \tag{1.6}$$

1.1.2 Four Factors

In his work, Oliver highlighted the four advanced metrics, known as *Four Factors*, which mostly impact on a team capability of performing on both sides of the game. These metrics are the *Effective Field Goal Percentage* (EFG%), the *Turnover Percentage* (TOV%), the *Offensive Rebound Rate* (OREB%), and the *Free Throw Rate* (FTr).

These statistical metrics, traditionally used for team evaluations, can also be effectively adapted to assess individual player performance. By modifying the formulas to focus on a specific player's contribution rather than the team's collective output, these metrics offer valuable insights into individual skill and efficiency. For example, the formula for Offensive Rebound Percentage (OREB%) can be adjusted to reflect an individual player's ability to secure offensive rebounds in comparison to the total available while they are on the court.

Additionally, these metrics can be extended to analyze defensive performance by considering the statistics of the opponents. By doing so, they provide a measure of how effectively a player or team is performing on the defensive end.

Effective Field Goal Percentage

Effective Field Goal Percentage adjusts the traditional field goal percentage to account for the added value of three-point shots. By giving extra weight to three-pointers (0.5 times the number of made three-point shots), EFG% provides a more accurate reflection of a player's or team's overall shooting efficiency, considering that three-pointers are worth more than two-point shots.

$$EFG\% = \frac{FGM + 0.5 * 3FGM}{FGA} \tag{1.7}$$

Turnover Percentage

Turnover Percentage is a statistical metric that measures the rate at which a player or team commits turnovers relative to their total plays. A turnover occurs when a player loses possession of the ball to the opposing team, typically due to mistakes like bad passes, traveling violations, losing the ball out of bounds, or having the ball stolen by an opponent. It accounts for the pace

of the game and the overall number of possessions, offering a more accurate representation of turnover propensity than simply counting turnovers

$$TOV\% = \frac{TOV}{FGA + 0.44 * FTA + TOV} = \frac{TOV}{PLAYS} \tag{1.8}$$

Offensive Rebound Rate

Offensive Rebound Rate is a statistical measure used to assess a team's or player's efficiency in securing offensive rebounds. It is calculated as the percentage of available offensive rebounds a team or player successfully retrieves during a game. The formula for Offensive Rebound Rate is:

$$OREB\% = \frac{OREB_{tm}}{OREB_{tm} + DREB_{opp}} \tag{1.9}$$

Free Throw Rate

Free Throw Rate is used to evaluate how often a team or player gets to the free-throw line relative to how often they attempt field goals. It is a measure of a team's or player's ability to draw fouls and earn free throw opportunities. The formula for Free Throw Rate is:

$$FTr = \frac{FTA}{FGA} \tag{1.10}$$

1.1.3 Other Advanced Metrics

Subsequently, many advanced metrics were developed in order to better estimate the impact of teams and players in various aspects of the game, without being limited to the amount of played minutes:

Team offensive and defensive ratings are macro-level metrics that evaluate the overall efficiency of a team's offense and defense, respectively. The offensive rating for a team reflects how effectively the team scores points, considering the number of possessions they have in a game. It encapsulates the collective output of the team's offensive efforts, including shooting efficiency, turnover rates, and ability to secure offensive rebounds. Conversely, the defensive rating measures a team's effectiveness in preventing the opposing team from scoring, encompassing aspects like opponent shooting efficiency, forced turnovers, and defensive rebounding capabilities.

On the other hand, individual offensive and defensive ratings delve into the contribution of each player to their team's performance. These ratings are

more complex as they attempt to isolate a player's impact from the team context. The individual offensive rating assesses how efficiently a player uses possessions when they are on the court, factoring in their scoring ability, assist rates, and turnover tendencies. The individual defensive rating, meanwhile, gauges a player's effectiveness in limiting the opponent's scoring opportunities, considering their contributions to steals, blocks, defensive rebounds, and overall defensive presence.

The distinction between team and individual ratings is significant because it allows analysts to understand not just the collective strength of a team but also the specific contributions of each player. This differentiation helps in identifying the value players bring to their team's offensive and defensive systems, and can be instrumental in player evaluation, game strategy, and team building in basketball.

These advanced metrics represent a significant evolution in basketball analytics, providing deeper insights into the efficiency and effectiveness of teams in both offensive and defensive aspects of the game. By focusing on per-possession analysis, these metrics offer a more accurate and nuanced understanding of a team's performance, taking into account the varying paces and styles of play in modern basketball. These metrics can focus on aspects such as shooting, but also on other side of the game such as rebounding or passing.

True Shooting Percentage

True Shooting Percentage (TS%) is a measure of shooting efficiency that takes into account field goals, three-point field goals, and free throws. This metric is more comprehensive than EFG% as it includes free throws, providing a holistic view of a player's shooting efficiency.

$$TS\% = \frac{PTS}{2 * (FGA + 0.44 * FTA)} \tag{1.11}$$

Points per Possession

Points per Possession (PPP) is an analytical metric closely related to offensive rating, yet it operates on a different scale and is typically applied in more specific contexts. While offensive rating is often used to gauge overall team efficiency, PPP is employed to assess the efficiency of a player or team within particular playtypes, such as Pick&Roll, Post-ups, Isolations, and others. Contrary to what the name might imply, PPP is traditionally calculated as a measure of points per play, not per possession.

$$PPP = \frac{PTS}{FGA + 0.44 * FTA + TOV} = \frac{PTS}{PLAYS} \tag{1.12}$$

Points per Shot

Points Per Shot (PPS) is an intuitive metric that measures the average number of points scored per field goal attempt. It offers a direct way to assess a player's or team's scoring efficiency by linking the total points scored to the number of shots taken. This metric is particularly useful for evaluating a player's or team's ability to convert shooting opportunities into points.

The formula for PPS can be seen as a practical application of the Effective Field Goal Percentage (EFG%), scaled to reflect actual points rather than a percentage. EFG% is designed to account for the extra value of three-point shots. When you multiply EFG% by 2, you essentially convert this percentage into a points per shot value.

$$PPS = \frac{2 * 2FGM + 3 * 3FGM}{FGA} = 2 * EFG\% \tag{1.13}$$

1.1.4 Play-by-Play Data

The landscape of basketball analytics underwent a transformative shift with the meticulous tracking of play-by-play (PBP) data starting in the 1996-97 NBA season. PBP data represents a detailed event log, recording player actions on the court in a chronological sequence. This data collection method not only captures the outcome of each play but also provides critical context regarding the timing and location of events, particularly for shots. This expansion of data collection enhanced the analytical capabilities, allowing for a richer, more nuanced understanding of the game.

With PBP data, analysts could now incorporate spatial and temporal dimensions into their evaluations. This meant that the analysis of plays included not just the *what* but also the *where* and *when*, offering a deeper insight into the strategies and dynamics of the game. It enabled a more comprehensive view of player movements and team formations, revealing patterns and tendencies that were previously indiscernible.

A significant advantage of PBP data is its ability to evaluate all ten players on the court simultaneously. This comprehensive view extends beyond individual performance, shedding light on team dynamics, interactions, and strategies in real-time. This holistic approach has proven invaluable in understanding the complexities of team sports like basketball.

One of the earliest and most pivotal advancements from PBP data was the development of the Plus-Minus (+/−) statistic. This metric measures the point differential for a team with a specific player on the court. It's calculated by subtracting the points allowed by the team from the points scored while the player is in the game. The Plus-Minus statistic serves as a quantifier of a player's net impact on the team's performance during their time on the floor, providing a simple yet powerful insight into player effectiveness.

PBP data also spurred innovation in evaluating player combinations and team lineups. This aspect, known as *Lineups Analysis*, examines the performance of various player groupings, whether it's a five-player lineup or pairings of players. This analysis has been instrumental in understanding player synergy, optimizing team composition, and devising strategic matchups. It allows teams to assess not just individual contributions but also how players' styles and skills complement each other, leading to more informed decisions in team management.

1.1.5 Spatial Description

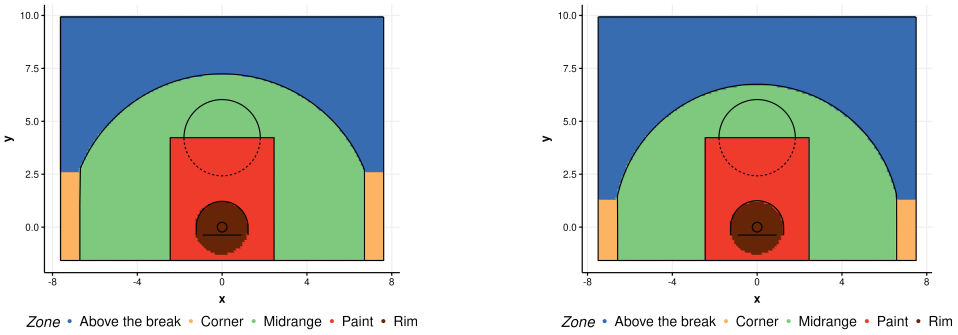
The spatial tracking of shots across the entire competition has been particularly revolutionary. It enables analysts to examine not only team tendencies in specific zones but also, and more crucially, the efficiency of shots in these areas. The rules of basketball classify shots into two main types: shots within the three-point line, worth two points, and shots from outside the three-point line, which yield three points. This distinction has significant strategic implications.

Breaking down the basketball court into five major shot zones enhances the granularity of this analysis:

- Rim: the area within the basket and the no-charge zone, which is the circle of 1.5 metres radius around the basket.
- Paint: often painted in a contrasting color, this rectangular area extends from the baseline under the basket to the free-throw line. Offensive players are restricted from staying in the paint for more than three seconds in a row to prevent camping near the basket.
- Midrange: this area is between the paint and the three-point line. It includes the elbows (the corners of the free-throw line) and the wings (the sides of the court).
- Corner: Located at the intersection of the baseline and the three-point line on each side of the court. The distance to the basket from the corner

is slightly shorter than from other parts of the three-point line, making it a favored spot for shooters.

- Above the Break: This refers to the area of the three-point line that extends from the corners to the point where the arc approaches the top of the key.



(a) NBA court zones

(b) FIBA court zones

Figure 1.1: Difference between the NBA court and the FIBA court, used in European and international leagues

While there are some differences between the FIBA court, used in international competitions and European basketball, and the NBA court, which is slightly larger with longer corner zones, the fundamental zones of play can be considered analogous across both court types. This uniformity in court zoning allows for a consistent approach in analyzing player and team performances in different leagues.

1.1.6 The Impact of Analytics

The impact of analytics in basketball extends far beyond mere game and player analysis, significantly influencing strategic approaches and playstyles. One of the most transformative figures in this regard has been Daryl Morey, former General Manager of the Houston Rockets. Morey's analytical approach brought to light the disparity in efficiency and value between 2-point and 3-point shots. Specifically, analysis of Points Per Shot (PPS) by shot zone, as detailed by Goldsberry[4], revealed a stark difference in efficiency between shots at the rim, 2-point shots outside the rim, and 3-point shots, particularly from the corner.

Morey also recognized the strategic advantage of increasing the game's pace.

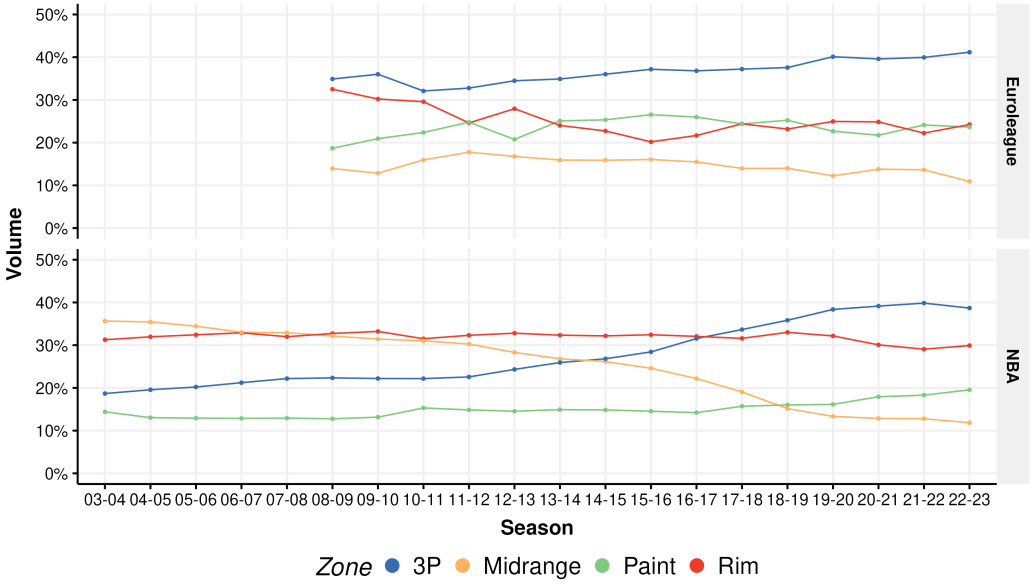


Figure 1.2: Variations of shot volumes by zone in the last 20 years in NBA and Euroleague

A faster pace not only heightened the entertainment value but also correlated with an increase in scoring opportunities and, consequently, the probability of winning. This insight led to a gradual yet profound shift in team strategies across the league, emphasizing greater reliance on 3-point shots, especially over mid-range attempts, as noted by Partnow[5]. This strategic evolution has contributed to what can be described as an *offensive inflation* in the NBA, a trend that continues to shape the game.

As illustrated in Figure 1.2, this uptick in 3-point shooting volume is more pronounced in the NBA than in the Euroleague, though a similar trend is observable in Europe's premier basketball league. It is crucial to acknowledge how differences in court size, rules, and other contextual factors between the NBA and Euroleague influence these playstyle adaptations. Additionally, the defensive 3-second rule in the NBA markedly affects team spacing and shot selection, further differentiating the strategic landscape between the two leagues.

The evolution of basketball in the NBA, characterized by strategic shifts influenced by analytics, has been more pronounced compared to changes observed in the Euroleague. However, considering the increasing exchange of players and ideas between these two prominent leagues, it is reasonable to anticipate

that trends prominent in the NBA will begin to manifest more distinctly in European basketball as well.

1.2 Tools

1.2.1 Data Collection

The approach to data analysis in basketball varies significantly between the NBA and European leagues. The NBA has been a forerunner in embracing data analytics, encouraging stakeholders to engage in data-driven projects. It facilitates this by offering an API (Application Programming Interface) that allows for easy access and downloading of a wide range of basketball data. This progressive stance has not only enhanced game strategies and player evaluations but also fostered a culture of data-driven decision-making within the league.

Contrastingly, the European basketball scene faces unique challenges that hinder the widespread adoption and implementation of advanced analytics. One primary issue is the relatively lower emphasis placed on data analysis in these leagues. However, a more pressing problem is the lack of standardization in statistical systems across different European leagues. Each league often operates its own statistical system, characterized by variability in the types of data collected, the level of detail (granularity), and accessibility. This fragmentation makes it difficult to conduct comprehensive and comparative analyses across leagues, posing a significant barrier to the development of a unified analytical approach in European basketball.

Sdeng

To address these challenges, the `Sdeng` library was developed. This Python-based web scraper[6] is designed to navigate the disparate statistical systems of various European leagues, extracting data and converting it into a common, unified format. The library's primary goal is to standardize the data collection process, thereby enabling analysts to compare and contrast data across different leagues more effectively. By harmonizing the data, the `Sdeng` library opens up new possibilities for in-depth analysis, benchmarking, and cross-league studies that were previously impractical due to the lack of standardized data.

The introduction of tools like the `Sdeng` library is a pivotal step in elevating the role of data analytics in European basketball. By overcoming the barriers of data fragmentation and inaccessibility, it paves the way for more sophisticated

analytical approaches, bringing European basketball closer to the analytical rigor seen in the NBA.

The `Sdeng` library, initially developed to accommodate a few specific leagues, has remarkably expanded its reach, encompassing a wide array of basketball competitions across the globe. Its coverage now includes not only European leagues like the Italian Legabasket, Euroleague, Eurocup, Basketball Champions League, and Eurobasket but also major domestic leagues throughout Europe. The addition of the Australian League, the NBA, the G League, and all FIBA international competitions signifies a major milestone.

Moreover, the `Sdeng` library incorporates sophisticated error-handling mechanisms in its data downloading phase. It automatically addresses common issues that often plague basketball data sets, such as errors in the sequencing of actions, incorrect substitutions, inaccuracies in shot location data, and challenges in correctly matching (or *fuzzy joining*) player information across different data sources. These features are particularly crucial given the complexity and variety of data involved in basketball analysis.

The scraping component of the library was developed using `Python`[7], renowned for its robustness in handling web data extraction. `Python`'s object-oriented nature and a rich ecosystem of libraries make it an ideal choice for scraping tasks. In particular, the library `BeautifulSoup`[8], a well-known `Python` tool for web scraping, plays a pivotal role in this process. `BeautifulSoup`'s ability to navigate, search, and modify the parse tree of HTML and XML files makes it exceptionally suited for extracting data from various basketball league websites with varying structures and formats.

1.2.2 Statistical Analysis

On the other hand, the statistical computation component of Sdeng is written in `R`[9], a language widely recognized for its capabilities in statistical analysis and data visualization. `R`'s extensive package ecosystem, particularly the `tidyverse`[10] collection of packages for data manipulation and the `ggplot2`[11] library for data visualization, makes it exceptionally powerful for processing and analyzing complex datasets. The `tidyverse`, with its emphasis on readable and expressive syntax, streamlines data manipulation tasks, allowing for more efficient and intuitive handling of the scraped basketball data. Meanwhile, `ggplot2` offers a versatile and powerful tool for creating sophisticated visualizations, enabling users to uncover patterns and insights from the data.

This bifurcated approach, utilizing Python for data acquisition and R for data analysis and visualization, leverages the respective strengths of each

language. Python's effectiveness in scraping and handling web data combined with R's advanced capabilities in data manipulation, statistical analysis, and visualization results in a potent synergy.

Other significant libraries that were crucial to this work were `Gt Table`[12], `magrittr`[13], `Unidecode`[14], `glmnet`[15] and `BasketballAnalyzeR`[16].

Chapter 2

Individual Advanced Metrics

2.1 Overview

Over the past two decades, the field of basketball analytics has seen the development of numerous metrics aimed at evaluating a player's performance. These metrics, often categorized as *all-in-one*[17], strive to encapsulate a player's overall impact on the game into a single numerical value. This approach represents an effort to distill the multifaceted nature of basketball performance into a more digestible and comparative format.

Historically, these all-in-one metrics have their roots in weighted averages of traditional basketball statistics. The methodology typically involves assigning positive values to statistics that reflect beneficial contributions to the game, such as scoring points, grabbing rebounds, dishing out assists, and making steals. Conversely, actions deemed detrimental to team performance, like missing shots, committing fouls[1], or turnovers, are assigned negative weights. This approach creates a balance sheet of sorts, crediting players for positive actions and debiting for negatives.

One notable example of an all-in-one metric is the Efficiency (EFF) metric, developed by Martin Manley[19]. Another is the Performance Index Rating (PIR), which is widely used in Euroleague basketball. Both metrics follow the principle of assigning weighted values to various statistical components to arrive at an overall performance score for a player. These metrics aim to provide a quick, yet comprehensive, assessment of a player's contribution to the game, taking into account a wide range of statistical inputs.

[1] Fouling is not intrinsically negative, as it can be adopted as a defensive strategy[18]

$$EFF = \frac{(PTS+REB+AST+STL+BLK-FG_{missed}-FT_{missed}-TOV)}{GAMES} \tag{2.1}$$

$$PIR = \frac{(PTS+REB+AST+STL+BLK+PFD)-(FG_{missed}+FT_{missed}+TOV+BLKA+PF)}{GAMES} \tag{2.2}$$

While early all-in-one metrics like EFF and PIR provided a foundational approach to player evaluation, they were limited in their ability to fully capture the diverse aspects of basketball performance. These metrics, though useful, did not adequately differentiate between the varying impacts of different statistical contributions. Recognizing this limitation, more sophisticated metrics were developed to offer a more nuanced view of player performance, one of which is the Player Efficiency Rating (PER) developed by John Hollinger[20].

PER marked a significant advancement in basketball analytics by incorporating the context of team performance into the evaluation of individual players. This metric goes beyond the simple aggregation of positive and negative statistics. Instead, it adjusts these statistics to account for factors such as the pace of the game and the overall efficiency of the team. By doing so, PER provides a more balanced and context-aware assessment of a player's performance.

One of the key innovations of PER is its recognition that not all statistical contributions are equal in terms of their impact on a game. For instance, it factors in the efficiency of scoring by considering field goal attempts, free throw attempts, and three-point shots. It also adjusts for the team's pace of play, allowing for a fair comparison between players in different systems or playing styles. This adjustment is crucial in modern basketball, where the pace and style of play can vary significantly from team to team.

2.2 Plus-Minus

The Plus-Minus metric, originally conceived by the Montreal Canadiens of the National Hockey League (NHL) in the 1950s, was designed to quantify the point differential a team experiences while a specific player is on the ice. This concept, rooted in ice hockey, has since been adopted in basketball, becoming a widely used metric to assess player impact. Its transition to basketball was facilitated by the availability of PBP data, first in the NBA and subsequently in leagues around the world. It has seen subsequently been developed also for other sports[21], such as football[22] and volleyball[23]

In basketball, the Plus-Minus metric reflects the net score difference (points scored by the player's team minus points allowed) during the time a player is on the court. It can yield both positive and negative values: a positive Plus-Minus indicates that the team outscored its opponents while the player was in the game, suggesting a positive impact, whereas a negative value implies the team was outscored, indicating a potentially less effective performance.

While Plus-Minus represents a significant advancement in quantifying a player's impact, it is not without its flaws. One key limitation is its failure to account for the quality of teammates and opponents on the floor. A player's Plus-Minus can be significantly influenced by the performance of others, meaning that it may not always accurately reflect an individual player's contribution. For instance, a player might have a high Plus-Minus simply by being on the court with a strong lineup, or conversely, a good player might have a lower Plus-Minus due to playing with weaker teammates or against superior opposition.

Another issue with Plus-Minus is its susceptibility to noise, especially when used to evaluate performance in a single game. In such cases, the metric can be influenced by a small sample of events, many of which may be beyond the control of a single player. This variability means that Plus-Minus can sometimes provide a misleading representation of a player's impact in individual games.

Despite these limitations, Plus-Minus remains a valuable metric in basketball analytics. It provides a straightforward, if broad, indicator of a player's overall impact on team performance. However, for a more accurate and comprehensive assessment, it is often used in conjunction with other metrics that can provide additional context and account for the nuances that Plus-Minus alone may overlook.

2.3 On-court advanced stats

The introduction of play-by-play data, coupled with advanced statistical methods, has significantly deepened the analysis of a player's impact on a team's performance. This data allows for a detailed examination of how team dynamics shift when a player is on the court versus when they are off it.

By analyzing PBP data, it is possible to calculate a team's performance metrics when a specific player is playing. Additionally, by comparing these on-court statistics to the team's overall stats, analysts can extract valuable insights about the team's performance during the player's absence. This comparative analysis offers a clearer understanding of the player's influence on various aspects of the game.

The real value of this analysis emerges when these statistics are normalized. As highlighted in Section 1.1.1, normalization by possessions is an effective approach, allowing for the calculation of on-court stats per 100 possessions for any given player. This normalization accounts for the pace of the game, thereby enabling a fair comparison across different contexts.

$$OFFRTG_{tm\_on} = \frac{100 * PTS_{tm\_on}}{POSS_{tm\_on}} \tag{2.3}$$

$$DEFRTG_{tm\_on} = \frac{100 * PTS_{opp\_on}}{POSS_{opp\_on}} \tag{2.4}$$

$$NETRTG_{tm\_on} = OFFRTG_{tm\_on} - DEFRTG_{tm\_on} \tag{2.5}$$

Furthermore, the on-off differential, which is the difference in team performance when a player is on the court versus off, is particularly revealing. This metric, also known as *impact*, is instrumental for coaching staffs and front offices in pinpointing where a player most significantly affects the team's performance.

$$Impact = NETRTG_{tm\_on} - NETRTG_{tm\_off} \tag{2.6}$$

Where the ratings *off-the-court* can be simply computed by subtracting from the total non-normalized team values (points and possessions) the values *on-the-court*, which can be computed from PBP data.

While efficiency ratings are common metrics used to assess a team's overall performance in these respective areas, the on-off differential allows for an examination of more specific elements of play. For instance, it can shed light on how the presence of a player affects the frequency of shots from the paint or influences the opponents' three-point shooting percentage.

This approach represents a considerable advancement in player evaluation, but it is not without limitations. One of the persistent challenges is the difficulty in isolating the individual impact of a player from the collective contribution of their teammates and opponents. The interdependent nature of basketball means that a player's on-court value is invariably intertwined with the performance of others around them.

Chapter 3

Advanced Plus-Minus

As highlighted in Section 2.2, the Plus-Minus metric, while informative, is subject to a significant limitation due to high multicollinearity. This issue arises when players from the same team, especially those belonging to exceptionally strong or weak teams, exhibit similar Plus-Minus values. Such a scenario makes it challenging to discern the individual contribution of each player from the team's collective performance. A striking example of this can be seen in the 2022-23 NBA season, where the Plus-Minus leaders, as shown in Table 3.1, predominantly belonged to the Denver Nuggets – the team that eventually clinched the league title.

Player	Team	+/-
N. Jokic	DEN	+9.3
A. Gordon	DEN	+7.6
J. Holiday	MIL	+7.2
M. Porter Jr.	DEN	+6.7
K. Caldwell-Pope	DEN	+6.5

Table 3.1: NBA Plus-Minus leaders in 2022-23 Regular Season with at least 50 played games

While advanced on-court statistics have mitigated multicollinearity to some extent, the quest for more refined analytical tools has led researchers to develop new advanced metrics. These metrics employ statistical models, moving beyond merely descriptive analytics. An exemplary outcome of this research is the development of Adjusted Plus-Minus, which represents a significant advancement in isolating individual player impact from the overall team performance.

Traditionally, advanced metrics like Adjusted Plus-Minus have predominantly been computed for the NBA, largely due to the wider availability of detailed game data and the global interest the league garners. However, this particular study shifts its focus to the Euroleague, a domain that has garnered significantly less attention in the realm of advanced basketball analytics. By concentrating on Euroleague basketball, the study aims to uncover and explore aspects of the game that have remained relatively unexamined in European contexts.

The primary objective of this Euroleague-focused analysis is twofold. First, it seeks to illuminate the nuances and unique characteristics of European basketball, which might differ from those observed in the NBA due to variations in playing styles, rules, and competitive dynamics. Second, the study endeavors to evaluate the correlation between the insights derived from these advanced metrics and the general public perception or conventional wisdom about the game in Europe.

3.1 Adjusted Plus-Minus

Adjusted Plus-Minus (APM), as conceptualized by Rosenbaum in 2004[24], represents a significant advancement in basketball analytics. This statistical method is specifically designed to isolate the impact of an individual player on the game, distinguishing it from the contributions of the other nine players on the court. The foundation of APM analysis lies in examining segments of the game referred to as stints. These are specific periods during which there are no substitutions, meaning the same ten players remain on the court throughout the duration of the stint.

APM is particularly focused on evaluating team performances per 100 possessions during these stints. It assesses how the team fares with a specific combination of ten players on the floor. The core of APM's utility is in its ability to calculate an individual coefficient for each player. This coefficient represents the estimated impact of the player per 100 possessions, essentially quantifying their contribution to the team's performance during the time they are on the court.

The point differential per 100 possessions for each stint is represented as Y_i, and the player presence for the home team (A) and the road team (B) during each stint j is indicated by $A_{i,j}$ and $B_{i,j}$. The coefficients $\beta_1, \beta_2, \ldots, \beta_n$ quantify the individual impact of each player.

$$\begin{aligned} Y_1 &= \beta_1 A_{1,1} + \beta_2 A_{2,1} + \ldots + \beta_{n-1} B_{n-1,1} + \beta_n B_{n,1} \\ Y_2 &= \beta_1 A_{1,2} + \beta_2 A_{2,2} + \ldots + \beta_{n-1} B_{n-1,2} + \beta_n B_{n,2} \\ &\vdots \\ Y_s &= \beta_1 A_{1,s} + \beta_2 A_{2,s} + \ldots + \beta_{n-1} B_{n-1,s} + \beta_n B_{n,s} \end{aligned} \tag{3.1}$$

The primary goal of APM is to determine the values of β that best correlate with the observed point differentials. This is achieved by expressing the problem in matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{2,1} & \ldots & B_{n,1} \\ A_{1,2} & A_{2,2} & \ldots & B_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,s} & A_{2,s} & \ldots & B_{n,s} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \tag{3.2}$$

Which can be reformulated as

$$Y = P\beta \tag{3.3}$$

Since the matrix P (representing player presence) is neither square nor invertible, a least-squares approach is used, involving the multiplication of both sides by P's transpose:

$$P^T Y = P^T P\beta \tag{3.4}$$

The matrix $P^T P$ is a Gram matrix, the elements of which indicate the number of stints shared between players. It is invertible if and only if it has full rank. Due to potential dependencies in the rows of P, the pseudo-inverse of $P^T P$ is utilized:

$$\beta = (P^T P)^+ P^T Y \tag{3.5}$$

The values in β represent the Adjusted Plus-Minus, indicating the impact per 100 possessions of the players, adjusted for the effect of players they share the court with. The leaders in APM between 2018 and 2023 with at least 50 games are displayed in Table 3.2.

The examination of the results, while showcasing some of the league's most distinguished players in recent seasons, has brought to light an issue of notably high

Player	Games	Minutes	APM	sd
W. Tavares	173	4045	+220.3	145.1
M. Fall	108	2414	+218.3	145.1
I. Canaan	63	1202	+217.1	145.1
T. Walkup	174	3970	+213.8	145.0
P. Lacombe	53	695	+213.4	145.1
N. Weiler Babb	93	2330	+212.7	145.1
J. Dibartolomeo	133	1756	+212.0	145.0
B. Baron	118	2594	+212.0	145.1
L. Olinde	76	1428	+211.8	145.1
P. Henry	100	2644	+211.5	145.0

Table 3.2: Best APM for Euroleague players with at least 50 games between 2018-19 and 2022-23

standard deviation values in the results. This variability, as Rosenbaum has previously articulated, arises from several factors intrinsic to APM's methodology and basketball data characteristics.

A significant source of this variability is the inherent noise within the data, primarily because most game segments analyzed in APM, known as stints, are short, usually encompassing just about three possessions. This brevity often leads to fluctuations in data, challenging the accuracy of APM values. Additionally, the traditional APM model's lack of a regularization term contributes to this issue. For instance, the exceptionally high APM value of 220.3 for a player like Walter Tavares seems implausible when compared to the typical net rating of around +20 for elite Euroleague teams.

In response to these challenges, two primary solutions are suggested. Firstly, extending the analysis to a broader timespan, spanning 5 or 10 seasons, can help dampen the short-term fluctuations and offer a more comprehensive view of a player's performance, which is why traditionally APM is not analyzed for a single season. Secondly, the introduction of a regularization term in the APM calculation is proposed to address the issue of extreme values. This term would help moderate the influence of outliers and provide a more balanced representation of a player's impact.

It is also crucial to acknowledge the distinct differences in game dynamics between the NBA and Euroleague when interpreting APM data. The NBA's regular season comprises 82 games per team, with each game lasting 48 minutes, leading to an average of about 100 possessions per game. Conversely, the

Euroleague features a shorter regular season with only 34 games of 40 minutes each, resulting in an average of approximately 70 possessions per game. This disparity highlights a significant difference in pace between the two leagues. Even when normalized to a 40-minute game, an NBA match features substantially more possessions than its Euroleague counterpart, underscoring the need for a contextual understanding of APM values across different leagues.

3.2 Regularized Adjusted Plus-Minus

As table 3.2 shows, the variance of APM is extremely high, enough to make it completely unreliable in assessing the contribute provided by each player. To solve this problem, a regularization method is introduced. The outlined metric is called Regularized Adjusted Plus-Minus (RAPM).

The challenge in calculating APM arises from the potential non-invertibility of the player matrix, which is a common issue in situations with multicollinearity or when the matrix is not square. *Ridge Regression*[25] offers a solution by introducing a regularization term, λI, to Equation 3.4:

$$P^T Y = (P^T P + \lambda I)\beta \tag{3.6}$$

Where $(P^T P + \lambda I)$ becomes an invertible matrix. Thus, 3.5 can be reframed as:

$$\arg\min_{\beta}(Y - X\beta)^T(Y - X\beta) \tag{3.7}$$

Which becomes, by applying the regularization term:

$$\arg\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta \tag{3.8}$$

The solution for this equation can be found as:

$$\beta = (X^T X + \lambda I)^{-1} X^T Y \tag{3.9}$$

This regularization term serves two main purposes. Firstly, it helps to control multicollinearity among players. Multicollinearity occurs when players' performances are highly correlated, making it difficult to isolate the individual

impact of each player. By adding the regularization term, Ridge Regression effectively reduces the Residual Sum of Squares (RSS), thus mitigating the effects of multicollinearity.

Secondly, the term $\lambda\beta^T\beta$ acts as a *shrinkage penalty*. This penalty is crucial in controlling the magnitude of the coefficients, β, thus preventing overfitting. Overfitting occurs when a model becomes too complex, capturing the noise in the data rather than the underlying pattern. The regularization term ensures that the coefficients do not reach extreme values that overemphasize the impact of any single player.

The tuning parameter λ plays a critical role in balancing these two objectives. When $\lambda = 0$, the Ridge Regression model reduces to the standard APM calculation, with no regularization applied. As λ increases, the shrinkage penalty becomes more dominant, reducing the magnitude of the coefficients. However, if λ approaches infinity, it excessively penalizes the coefficients, driving them towards zero and diminishing the model's effectiveness in capturing player impacts.

Finding the optimal value for λ is crucial for the effectiveness of the Ridge Regression model in RAPM calculation, particularly in the context of minimizing model variance. Since RAPM typically does not involve labeled data in the traditional sense, the focus shifts to minimizing the variance of the model to enhance its predictive accuracy and reliability. This is achieved through *k-Fold Cross-Validation*[26], a technique that assesses the model's performance across different subsets of the data. By partitioning the data into k subsets ($k = 5$ in this case), the model is trained on $k - 1$ subsets while the remaining subset is used for testing. This process is repeated k times, each time with a different subset used for validation, ensuring a comprehensive evaluation of the model's performance.

The role of k-Fold Cross-Validation in this context is intricately linked to the *bias-variance trade-off* in statistical modeling. The bias-variance trade-off highlights the challenge in balancing the complexity of the model (bias) with its ability to perform consistently across different data sets (variance). A high λ value in Ridge Regression introduces more bias, simplifying the model but potentially making it less sensitive to the true relationships in the data. Conversely, a lower λ value reduces bias but can increase variance, leading to a model that fits the training data well but may not generalize effectively. The k-Fold Cross-Validation technique aids in finding the λ value that achieves the optimal balance between bias and variance, thereby enhancing the model's overall effectiveness and reliability in predicting a player's impact.

Player	Games	Minutes	RAPM
F. Campazzo	81	1982	+2.51
W. Tavares	173	4045	+2.45
D. Balbay	101	807	+2.29
S. Sanli	132	1861	+2.14
K. Simon	125	3020	+2.09
I. Canaan	63	1202	+2.00
N. Mirotic	128	3284	+1.98
R. Fernandez	138	2543	+1.92
M. Fall	108	2414	+1.75
J. Dibartolomeo	133	1756	+1.71

Table 3.3: Best RAPM for Euroleague players with at least 50 games between 2018-19 and 2022-23, for $\lambda = 354.5$

The application of RAPM in Euroleague basketball has yielded intriguing results, displayed in Table 3.3, spotlighting not only some of the league's widely celebrated players like Facundo Campazzo, Walter Tavares, and Nikola Mirotic but also unveiling some unexpected names, such as Dogus Balbay. Balbay's inclusion is particularly notable given his average of only 8 minutes per game over 101 games, illustrating the depth and nuance that these metrics can offer beyond traditional playing time or scoring averages.

One key observation from the results is the alignment of the scale of values with realistic expectations of player impact. For instance, an average team net rating of around 2.5 throughout a season, as suggested by the data, is typically indicative of a playoff-bound team in Euroleague competition. This coherence in scale marks a significant improvement over traditional APM, and suggests a more accurate reflection of a player's contribution to their team's performance.

However, it is important to note that, similar to APM, RAPM is most effectively analyzed over multiple seasons. This approach is taken to mitigate the inherent variability and potential anomalies that can arise from single-season data. By examining player performance across a broader timespan, RAPM offers a more stable and reliable assessment, smoothing out short-term fluctuations and providing a clearer picture of a player's consistent impact over time.

Chapter 4

Improvements of the model

4.1 Offensive and Defensive RAPM

While the RAPM model offers an effective approach to evaluate a player's impact, it does not differentiate between the contributions made on offense and defense. To address this, RAPM can be split into Offensive RAPM (ORAPM) and Defensive RAPM (DRAPM), providing a more granular analysis of a player's impact.

To achieve this, the base model for RAPM 3.2 is modified to separately account for offensive and defensive contributions:

$$\begin{bmatrix} Y_{1,A} \\ Y_{1,B} \\ Y_{2,A} \\ Y_{2,B} \\ \vdots \\ Y_{s,A} \\ Y_{s,B} \end{bmatrix} = \begin{bmatrix} A_{1,1,A,off} & A_{2,1,A,off} & \dots & B_{n,1,A,def} \\ A_{1,1,B,off} & A_{2,1,B,off} & \dots & B_{n,1,B,def} \\ A_{1,2,A,off} & A_{2,2,A,off} & \dots & B_{n,2,A,def} \\ A_{1,2,B,off} & A_{2,2,B,off} & \dots & B_{n,2,B,def} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,s,A,off} & A_{2,s,A,off} & \dots & B_{n,s,A,def} \\ A_{1,s,B,off} & A_{2,s,B,off} & \dots & B_{n,s,B,def} \end{bmatrix} \begin{bmatrix} \beta_{1,off} \\ \beta_{1,def} \\ \beta_{2,off} \\ \beta_{2,def} \\ \vdots \\ \beta_{n,off} \\ \beta_{n,def} \end{bmatrix} \tag{4.1}$$

Here, $Y_{s,T}$ represents the offensive rating for team T (A for the home team, B for the one on the road) during stint s, and $P_{i,s,T,side}$ indicates the presence of player i during the s-th stint on the *side* of team T. It takes a value of 1 if the side of the player's team is the same of the stint, -1 if the side is different, 0 if he is not on the court.

This model results in two β values for each player, one for each side of the court. The system can be solved using ridge regression, as shown in the RAPM

calculation. Once β_{off} and β_{def} are obtained for each player, they represent ORAPM and DRAPM, respectively.

Unlike the Defensive Rating, where lower values indicate better defense, a higher DRAPM represents a better defensive performance, as it can be interpreted as the number of points per 100 possessions that the player prevented.

The overall RAPM is usually computed as the sum of the two values:

$$RAPM = ORAPM + DRAPM \tag{4.2}$$

Nonetheless, Jacobs[27] highlighted that this is not inherently correct, since players can play a different number of possession on the two ends of the court. This may appear counter-intuitive, but is explained with the tendency of coaches, especially in the last minutes of tied games, to change lineup at every change of possession, whether the situation allows, in order to use more offensive or defensive lineups. Table 4.1 shows the player having the having the largest difference between offensive and defensive possessions per game, among players with at least 25 played games. Notably, they are known as specialists on either offense or defense, and are usually offensive players.

Player	GAMES	POSS off	POSS def	Tendency	Difference
D. Balbay	101	13.3	15.1	Defense	1.87
S. Wilbekin	149	47.7	46.3	Offense	1.41
A. Muhammed	73	28	26.8	Offense	1.25
J. Carroll	84	27.5	26.3	Offense	1.23
N. De Colo	140	45.3	44.1	Offense	1.21
N. Nedovic	110	40.2	39	Offense	1.17
J. Anderson	133	25.1	26.2	Defense	1.17
S. Larkin	150	49.8	48.7	Offense	1.08
D. Bacon	56	49.8	48.7	Offense	1.05
M. James	161	56	55	Offense	1.04

Table 4.1: Players with at least 25 played games between 2018-19 and 2022-23, having the largest difference in offensive and defensive possessions per game

To account for this difference, a new weighted formulation for the overall RAPM is proposed:

$$RAPM = 2 * \frac{ORAPM * POSS_{off} + DRAPM * POSS_{def}}{POSS_{off} + POSS_{def}} \tag{4.3}$$

As Table 4.2 shows, some of the players with the largest difference appear in Table 4.1, stating the relevance of accounting for the different number of possessions. It is also relevant to note that the differences are in fact very small, which makes 4.2 still suited in the cases where the number of possessions on each side is not available.

player	GAMES	POSS off	POSS def	ORAPM	DRAPM	RAPM sum	RAPM weight
D. Balbay	101	1340	1529	0.24	1.21	1.44	1.51
M. Birsen	51	635	658	-2.38	1.03	-1.34	-1.28
N. Nedovic	110	4419	4290	0.72	-1.22	-0.5	-0.47
G. Ricci	65	1249	1280	-1.72	0.61	-1.11	-1.08
I. Ukhov	66	971	1006	-0.38	1.2	0.82	0.85
S. Antonov	92	1289	1331	-0.45	1.05	0.6	0.63
A. Muhammed	73	2047	1956	0.51	-0.53	-0.02	0.01
U. Garuba	53	1403	1430	-2.15	0.34	-1.81	-1.79
D. Bertans	63	1936	1914	1.72	-1.86	-0.14	-0.12
S. Karasev	72	1887	1916	-2.27	0.23	-2.04	-2.02

Table 4.2: Largest differences in RAPM using *sum* or *weighted sum* for Euroleague players with at least 50 games between 2018-19 and 2022-23, with $\lambda = 292.4$

4.2 Weighting

In the realm of basketball analytics, RAPM and its derivatives such as Offensive RAPM and Defensive RAPM, are built on foundational assumptions that merit further scrutiny for enhanced accuracy. A pivotal assumption in these models is the equal weighting of all possessions during a game. However, the reality of basketball dynamics suggests a more nuanced approach, as not all possessions hold the same level of importance. For instance, segments of a game colloquially known as *garbage time*, where the outcome is largely decided and competitive intensity may decrease, can skew the perceived impact of a player. These periods contrast sharply with *clutch* moments, the final, often decisive minutes of a close game or overtime, where player performance under high-pressure conditions is particularly telling. Additionally, playoff games, with their elevated stakes and intensity, arguably carry more weight than regular-season encounters.

The original RAPM model also does not take into account the varied lengths of player stints. This oversight can lead to an imbalance, as making a significant impact over a longer stint is generally more indicative of a player's contribution than brief appearances, which are more susceptible to statistical noise.

To address these limitations, the model can be refined with the introduction of a weight vector W in the following form:

$$W = L \circ G_t \circ G_p \tag{4.4}$$

Here, L_i denotes the length of the i-th stint, recognizing that sustained performance is a more reliable indicator of a player's impact. G_{ti} assigns weights based on the type of game, giving playoff games a higher value (2) compared to regular-season games (1). G_{p_i} adjusts for the game phase, with clutch moments rated at 2, garbage time at 0.5, and standard periods at 1.

A similar approach has been used by Grassetti[28], when adopting the concept of RAPM for lineups. However, the weighting system developed in his paper is quite arbitrary with respect of what happens on the court, rather then on the context in which events happen. The mentioned weighting system, assigning a specific value to game events, such as made or missed shot, biases the model towards what the author believes to be more relevant in a basketball game, instead of focusing on objective efficiency per 100 possessions.

This solution also has the advantage of embedding weighting directly within the computation of ORAPM and DRAPM, thus making unnecessary to compute the overall RAPM as proposed in 4.3, and allowing to use the traditional formulation as described in 4.2.

Player	Games	Minutes	ORAPM	DRAPM	RAPM
F. Campazzo	81	1982	1.34	1.88	3.23
W. Tavares	173	4045	0.76	2.17	2.93
R. Fernandez	138	2543	1	1.67	2.66
D. Balbay	101	807	0.66	1.87	2.53
I. Ukhov	66	557	0.23	2.26	2.49
I. Canaan	63	1202	0.15	2.25	2.4
K. Simon	125	3020	1.82	0.57	2.39
S. Sanli	132	1861	1.77	0.6	2.37
J. Dibartolomeo	133	1756	1.31	0.99	2.3
R. Sorkin	58	770	0.76	1.53	2.29

Table 4.3: Best ORAPM and DRAPM for Euroleague players with at least 50 games between 2018-19 and 2022-23 using Ridge regression, with $\lambda = 185.2$

The results displayed in Table 4.3 provide some insights on the best overall contributes, and shed light on the aspect of the game where players are more impactful, either offense, defense or both. Notably, many players are the same provided by both APM and the traditional RAPM, displaying the ability

of the model to predict consistent values while allowing a deeper level of inspection.

4.3 Shrinkage Methods

Ridge regression is a valuable technique in linear regression that aims to address multicollinearity issues by introducing a regularization term. It endeavors to shrink the coefficients of all p predictors towards zero, without forcing any of them to reach exactly zero. This regularization technique ensures that all predictors retain some level of influence on the model's output. While Ridge Regression effectively handles multicollinearity, it does not inherently differentiate between predictors that have minimal impact and those that are truly irrelevant.

To address these challenges, two alternative regression techniques, Lasso Regression[29] and Elastic Net Regression[30], are proposed as solutions. These methods offer distinct approaches to handling predictor selection and improving model interpretability, which can be particularly beneficial when dealing with high-dimensional datasets or when the distinction between minimal impact and irrelevant predictors is crucial.

4.3.1 Lasso Regression

Lasso regression shares similarities with ridge regression in terms of pushing coefficient values towards zero. However, it differs crucially in its ability to set some coefficients exactly to zero when the regularization parameter λ is sufficiently large. This characteristic of lasso regression makes it a more suitable choice for models where sparsity is desired, effectively identifying and excluding less impactful variables (or players, in this context). Lasso regression thus provides a means to refine the model further, allowing for a more accurate representation of each player's distinct impact on the court.

The mathematical formulation of lasso regression in this context is expressed as:

$$\beta = \arg\min_{\beta} \left(\frac{1}{2n} \|W^{1/2}(Y - X\beta)\|_2^2 + \lambda \|\beta\|_1 \right) \tag{4.5}$$

This equation represents the optimization problem at the core of lasso regression, balancing the fit of the model (measured by the residual sum of squares) with the complexity (measured by the L1 norm of the coefficients). The inclusion of the weight matrix W allows for differential weighting of observations, further refining the model's accuracy and relevance to real-game scenarios.

Player	Games	Minutes	ORAPM	DRAPM	RAPM
W. Tavares	173	4045	0.69	6.15	6.84
W. Clyburn	124	3524	6.64	0	6.64
K. Simon	125	3020	5.34	0	5.34
D. Exum	63	1300	4.66	0	4.66
T. Black	85	1330	0	4.46	4.46
I. Canaan	63	1202	0	4.42	4.42
N. Mirotic	128	3284	2.95	0.91	3.87
F. Campazzo	81	1982	2.25	1.5	3.75
K. Punter	96	2393	3.77	-0.19	3.58
R. Fernandez	138	2543	1.27	2.22	3.49

Table 4.4: Best ORAPM and DRAPM for Euroleague players with at least 50 games between 2018-19 and 2022-23 using Lasso regression, for $\lambda = 0.43$

Player	Games	Minutes	ORAPM	DRAPM	RAPM
A. Shved	86	2570	0	-4.72	-4.72
M. Birsen	51	362	-4.66	0	-4.66
I. Diop	87	848	-4.45	0	-4.45
T. Schneider	89	1002	0	-4.43	-4.43
S. Karasev	72	1048	-4.31	0	-4.31
L. Radosevic	104	1400	-3.98	0	-3.98
M. Delow	84	1001	-3.96	0	-3.96
J. Lauvergne	89	1583	0	-3.69	-3.69
L. Nnoko	55	1114	0	-2.89	-2.89
M. Eric	71	1180	-2.84	0	-2.84

Table 4.5: Worst ORAPM and DRAPM for Euroleague players with at least 50 games between 2018-19 and 2022-23 using Lasso regression, for $\lambda = 0.43$

The data presented in Tables 4.4 and 4.5 reveal a distinctive trend in Lasso regression: it tends to favor players who excel significantly on one end of the court. This tendency is evident as only two players among the top 10, Facundo Campazzo and Rudy Fernandez, exhibit both a DRAPM and an ORAPM greater than 1. In contrast, five of the top six players demonstrate a zero impact on either offense or defense. Notably, the highest values observed in Lasso regression are more pronounced than those seen in Ridge regression. For example, Walter Tavares's overall RAPM in Lasso regression is more than double

his value in the Ridge regression (referenced in 4.3). However, while Facundo Campazzo maintains a comparable value to his Ridge regression performance, he is no longer ranked as the top player in the Lasso regression framework. This shift underscores the unique impact of Lasso regression in emphasizing distinct, one-sided contributions and its influence on player rankings.

4.3.2 Elastic Net Regression

Elastic Net regression shares similarities with Ridge and Lasso regression in its goal of mitigating multicollinearity and reducing the impact of less relevant predictors. However, it offers a unique blend of both L1 (Lasso) and L2 (Ridge) regularization, striking a balance between feature selection and coefficient shrinkage.

One of the crucial distinctions of Elastic Net regression is its ability to set some coefficients exactly to zero when the regularization parameters, denoted as λ_1 and λ_2, are sufficiently large. This characteristic makes it particularly suitable for situations where sparsity is desired, enabling the automatic identification and exclusion of less impactful variables or players in the context of your analysis.

The mathematical formulation of Elastic Net regression can be expressed as follows:

$$\beta = \arg\min_{\beta} \left(\frac{1}{2n} |W^{1/2}(Y - X\beta)|_2^2 + \lambda_1 |\beta|_1 + \lambda_2 |\beta|_2^2 \right) \tag{4.6}$$

In this equation, β represents the coefficient vector, Y is the target variable, X is the predictor matrix, and W is a weight matrix that can be used to assign varying importance to observations. The regularization parameters λ_1 and λ_2 control the L1 (Lasso) and L2 (Ridge) regularization terms, respectively.

In Elastic Net regression, the parameter α plays a crucial role in determining the balance between L1 (Lasso) and L2 (Ridge) regularization. It essentially dictates the mix of regularization techniques applied in the model. An α value of 1 equates to pure Lasso regression, emphasizing feature selection by setting some coefficients to zero, while an α of 0 aligns with Ridge regression, focusing more on shrinking coefficients to address multicollinearity. Intermediate values of α indicate a blend of both approaches, enabling Elastic Net to leverage the feature selection capability of Lasso and the coefficient shrinkage of Ridge.

$$\beta = \arg\min_{\beta} \left(\frac{1}{2n} \|W^{1/2}(Y - X\beta)\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2) \right) \tag{4.7}$$

Player	Games	Minutes	ORAPM	DRAPM	RAPM
W. Tavares	173	4045	0.79	6.17	6.96
W. Clyburn	124	3524	6.77	0	6.77
K. Simon	125	3020	5.34	0.1	5.44
D. Exum	63	1300	4.76	0	4.76
T. Black	85	1330	0	4.56	4.56
I. Canaan	63	1202	0	4.53	4.53
N. Mirotic	128	3284	3.14	0.94	4.09
F. Campazzo	81	1982	2.29	1.73	4.02
R. Fernandez	138	2543	1.43	2.41	3.84
K. Punter	96	2393	3.84	-0.41	3.44

Table 4.6: Best ORAPM and DRAPM for Euroleague players with at least 50 games between 2018-19 and 2022-23 using Elastic Net regression, with $\lambda = 0.78$ and $\alpha = 0.5$

Player	Games	Minutes	ORAPM	DRAPM	RAPM
M. Birsen	51	362	-5.47	0	-5.47
T. Schneider	89	1002	-0.15	-4.87	-5.02
A. Shved	86	2570	0	-4.86	-4.86
S. Karasev	72	1048	-4.69	0	-4.69
I. Diop	87	848	-4.88	0.33	-4.54
L. Radosevic	104	1400	-4.27	0	-4.27
M. Delow	84	1001	-4.17	0	-4.17
J. Lauvergne	89	1583	0	-4.1	-4.1
L. Nnoko	55	1114	0	-3.27	-3.27
M. Eric	71	1180	-3.02	-0.05	-3.07

Table 4.7: Worst ORAPM and DRAPM for Euroleague players with at least 50 games between 2018-19 and 2022-23 using Elastic Net regression, with $\lambda = 0.78$ and $\alpha = 0.5$

The results provided by this variation of the model are quite similar to the ones provided by Lasso regression.

4.3.3 Comparison

RAPM models, using all of the described regressions, encounter significant challenges in accurately predicting the value of players with few possessions. This limitation is vividly illustrated in image 4.1, where a phenomenon known as the *funnel effect* is evident. The funnel effect, more pronounced in Ridge regression, also occurs in Lasso and Elastic Net regression, albeit to a lesser extent.

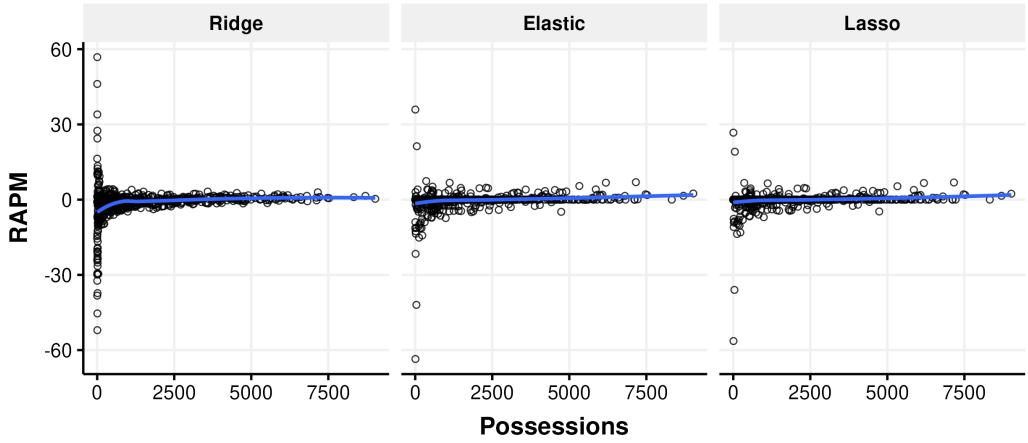


Figure 4.1: Funnel effect using ridge regression (left), elastic net regression (center) lasso regression (right)

The funnel effect refers to a pattern where the accuracy of predictions or evaluations disproportionately decreases for players with fewer possessions or playing time. In RAPM models, this effect manifests as a widening variance or uncertainty in player evaluation as the number of possessions decreases. This effect is akin to a statistical funnel where the right (representing players with many possessions) has narrow variability, indicating more reliable and consistent evaluations. As one moves left the funnel (towards players with fewer possessions), the variability increases, leading to less reliable predictions.

The presence of the funnel effect is particularly problematic when analyzing statistics per 100 possessions for players who had minimal court time. These players often participate during garbage time, therefore their contributions may not accurately reflect their true abilities or potential impact under normal game conditions.

To address this issue and enhance the model's reliability, a filtering criterion has been applied, setting a minimum threshold of 50 games for player inclusion

in the analysis. This approach ensures that the players evaluated have had a sufficient amount of playtime, providing a more stable and representative dataset for RAPM calculations.

Method	Stat	Mean	Sd	Min	Median	Max	Zeros%	Shapiro Test
Ridge	DRAPM	-0.12	1.06	-3.72	-0.06	4.11	0	0.98
	ORAPM	-0.25	1.15	-3.79	-0.21	2.67	0	0.99
	RAPM	-0.37	1.52	-5.98	-0.29	3.95	0	0.99
Lasso	DRAPM	-0	1.06	-5.99	0	7.34	82.5%	0.48
	ORAPM	0.01	1.17	-7.02	0	6.64	80%	0.52
	RAPM	0.01	1.55	-7.02	0	6.84	66.9%	0.7
Elastic	DRAPM	-0.01	1.17	-6.39	0	7.81	79%	0.51
	ORAPM	-0.01	1.29	-7.36	0	6.77	75.9%	0.56
	RAPM	-0.02	1.7	-7.36	0	6.96	61.6%	0.74

Table 4.8: Statistical comparison across different regression models for players with at least 50 games between 2018-19 and 2022-23

The results presented in Table 4.8 highlight a significant disparity between the original model utilizing Ridge Regression and the newer models. The former yields more constrained values, falling within the range of $[-5.98, +3.95]$ for the overall RAPM, despite exhibiting a similar standard deviation. Notably, when applying a filter for players with at least 50 games played in the 2018-2023 timeframe, the results demonstrate a clear adherence to a normal distribution pattern, as visualized in Figure 4.2, and confirmed by the high Shapiro-Wilk Test statistic[31], which closely approaches 1.

In contrast, both the Lasso and Elastic Net regression models display notably lower Shapiro-Wilk values, approximately 0.5 for offensive/defensive estimates and 0.7 for the overall RAPM. This phenomenon is largely influenced by the high prevalence of zero values, constituting around 80% for ORAPM and DRAPM, and around 60-65% for the overall RAPM.

The results obtained from Lasso and Elastic Net regression models reveal a significant limitation: a high prevalence of players being assigned a zero impact. This outcome is somewhat counterintuitive, as it suggests that as many as 80% of players have a negligible impact on their teams, which is unlikely to be the case. This phenomenon is particularly apparent in Figure 4.3, where it's observed that many players with RAPM values within the $[-2, 2]$ range, as determined by Ridge Regression, are reduced to a value of zero in models using Lasso or Elastic Net. The figure also indicates that the distinction between the

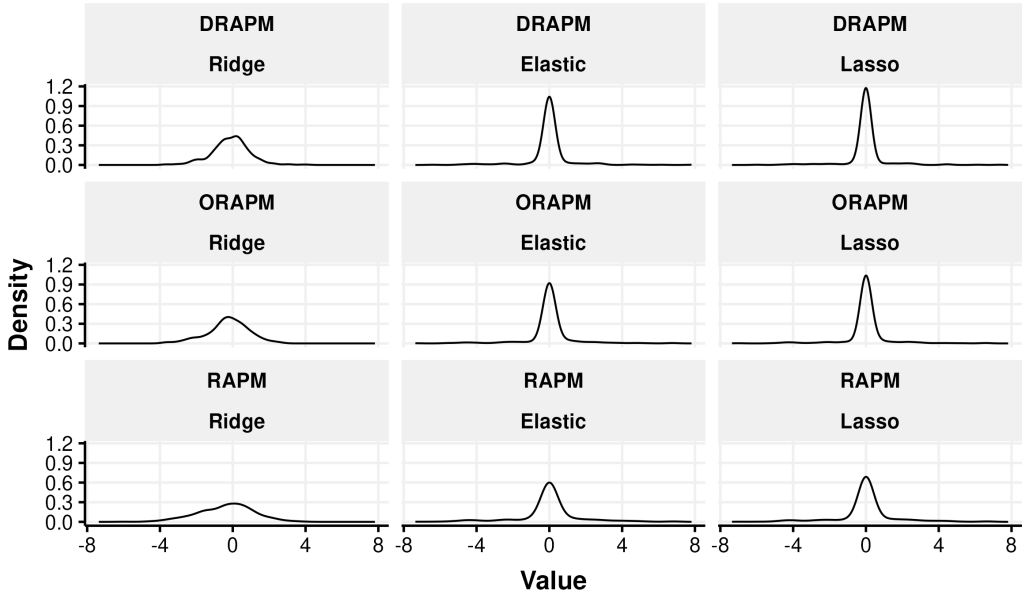


Figure 4.2: Distribution of values across different regression models for players with at least 50 games between 2018-19 and 2022-23

Lasso and Elastic Net models is relatively subtle in terms of their treatment of these players.

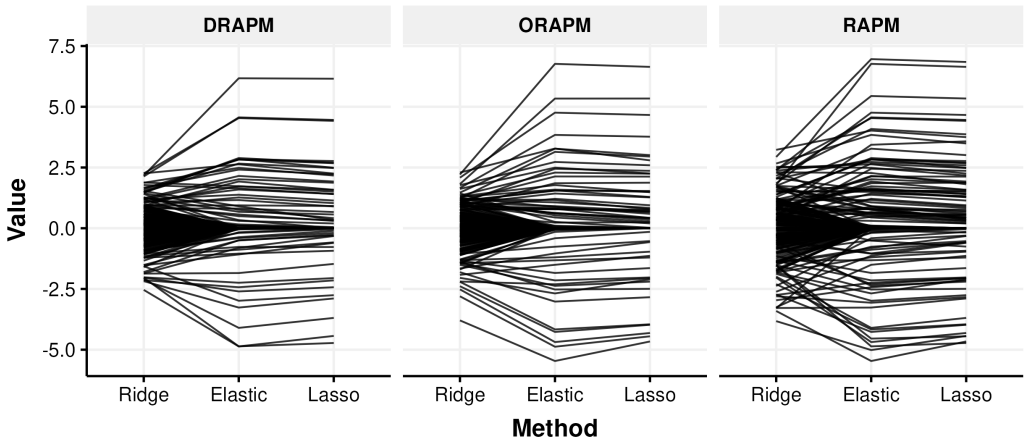


Figure 4.3: RAPM using different regression models for players with at least 50 games between 2018-19 and 2022-23

To mitigate these issues was addressed the flexibility of Elastic Net and its

parameter α. Following thorough analysis, a finely tuned value of $\alpha = 1.25 \times 10^{-3}$ was selected. This specific value predominantly leans towards the Ridge Regression approach while subtly integrating aspects of Lasso Regression, offering an improvement over the standard Elastic Net approach with $\alpha = 0.5$. This nuanced application of Elastic Net Regression helps to overcome the challenges of zero impact assignment, providing a more realistic and balanced representation of player contributions.

The results of this approach are summarized in Table 4.8:

Method	Stat	Mean	Sd	Min	Median	Max	Zeros%	Shapiro Test
	DRAPM	0.04	0.86	-3	0	2.78	36.7%	0.89
$\alpha = 0.00125$	ORAPM	0.02	0.9	-4.08	0	2.85	33.8%	0.9
	RAPM	0.06	1.25	-4.16	0	3.59	12.9%	0.97

Table 4.9: Statistical report of Elastic Net having $\alpha = 0.00125$ and $\lambda = 112.1$ for players with at least 50 games between 2018-19 and 2022-23

This approach enhances the traditional RAPM computation by driving a small subset of coefficients to zero, aligning with empirical experience, but without disrupting the original RAPM values significantly. By doing so, it is possible strike a balance that preserves the integrity of the RAPM metric while harnessing the advantages of Lasso Regression in a controlled manner.

4.4 Results

The findings of this study, leveraging the RAPM model, intriguingly align with the qualitative assessments typically derived from the *eye-test* in basketball. This alignment lends credence to the model's accuracy, particularly noteworthy given the unsupervised nature of the problem where direct validation through a supervised dataset is not feasible. The fact that many of the top 10 most impactful players identified over the 2018-2023 period are also highly acclaimed and awarded players further reinforces the model's credibility.

In the realm of Euroleague basketball, RAPM analysis over the past five years has spotlighted Facundo Campazzo as the competition's most impactful player. Campazzo, celebrated for his prowess, stands out for his versatile skills that extend across both ends of the court. His high ratings in both Offensive and Defensive RAPM are a testament to his well-rounded game and justify his reputation as one of Euroleague's premier talents.

Player	Games	Minutes	ORAPM	DRAPM	RAPM
F. Campazzo	81	1982	1.51	2.08	3.59
W. Tavares	173	4045	0.8	2.78	3.58
R. Fernandez	138	2543	1.07	1.92	2.99
K. Simon	125	3020	2.24	0.52	2.75
I. Canaan	63	1202	0	2.68	2.68
W. Clyburn	124	3524	2.85	-0.21	2.65
N. Mirotic	128	3284	1.46	1.04	2.5
T. Black	85	1330	0	2.49	2.49
D. Exum	63	1300	2.49	0	2.49
S. Sanli	132	1861	2.03	0.42	2.45
I. Ukhov	66	557	0	2.41	2.41
J. Dibartolomeo	133	1756	1.44	0.94	2.38
D. Balbay	101	807	0.19	1.99	2.18
M. Fall	108	2414	1.2	0.89	2.09
A. Abrines	112	1902	0.2	1.8	2.01

Table 4.10: Best ORAPM and DRAPM for Euroleague players with at least 50 games between 2018-19 and 2022-23 using Elastic Net regression, with $\alpha = 0.00125$ and $\lambda = 112.1$

Walter Tavares closely trails Campazzo in terms of overall impact, but with a distinct skew towards defensive excellence. Tavares's defensive capabilities, as reflected in his RAPM scores, resonate with his accolades, including three Defensive Player of the Year (DPOY) awards. His dominance on the defensive end is well-acknowledged and aligns seamlessly with public opinion and his award history, underscoring the model's ability to capture real-world performance accurately.

The analysis also sheds light on players who excel predominantly on one side of the court, sometimes in unexpected ways. Isaiah Canaan and Tarik Black, despite their differing roles as guard and center respectively, both emerge as defensive standouts. Their high defensive ratings challenge conventional role-based expectations and highlight the diverse skill sets present in the league.

Conversely, Will Clyburn and Dante Exum are recognized for their offensive contributions. Clyburn's acclaim for his scoring prowess is mirrored in his ORAPM scores. More intriguing, however, is the case of Dante Exum, typically lauded for his defensive skills. The RAPM model reveals a substantial impact on the offensive end, diverging from the general perception of his defensive orien-

tation. This insight into Exum's offensive contribution underscores the nuanced understanding of player capabilities that RAPM analysis can offer.

Player	Games	Minutes	ORAPM	DRAPM	RAPM
T. Schneider	89	1002	-1.17	-3	-4.16
M. Birsen	51	362	-4.08	0	-4.08
L. Nnoko	55	1114	-0.83	-2.39	-3.22
M. Eric	71	1180	-2.12	-0.92	-3.04
S. Enoch	56	963	-1.83	-1.19	-3.02
S. Karasev	72	1048	-2.91	0.08	-2.84
L. Radosevic	104	1400	-2.74	0	-2.74
S. Monia	68	1099	-0.43	-2.21	-2.64
A. Tyus	75	1234	-0.95	-1.57	-2.51
J. Lauvergne	89	1583	0	-2.51	-2.51
M. Delow	84	1001	-2.49	0	-2.49
A. Kurucs	70	472	-1.31	-1.14	-2.46
J. Puerto	50	614	-0.21	-2.21	-2.42
A. Shved	86	2570	0.32	-2.56	-2.24
U. Garuba	53	801	-2.2	0	-2.2

Table 4.11: Worst ORAPM and DRAPM for Euroleague players with at least 50 games between 2018-19 and 2022-23 using Elastic Net regression, with $\alpha = 0.00125$ and $\lambda = 112.1$

The study also sheds light on players who, according to the model, negatively impact their teams. The case of Metecan Birsen is particularly intriguing; despite having a positive win-loss record, he is identified as one of the worst impacting players. This might suggest that while Birsen's presence does not hinder his team's ability to win games, his individual contributions might be less critical to their success. The model also brings into focus players like Alexej Shved, whose defensive shortcomings are highlighted alongside a limited offensive value.

One of the significant findings in the application of RAPM is its relatively lower correlation with team record (a correlation coefficient of 0.599) compared to the traditional Plus-Minus statistic (which shows a higher correlation of 0.783 with team record), as shown in image 4.4. This difference in correlation is particularly revealing and underscores the strengths of the RAPM model. Of course, a certain correlation is still expected and desirable, since players with a good contribution help the team to win.

The traditional Plus-Minus metric, while straightforward and easy to under-

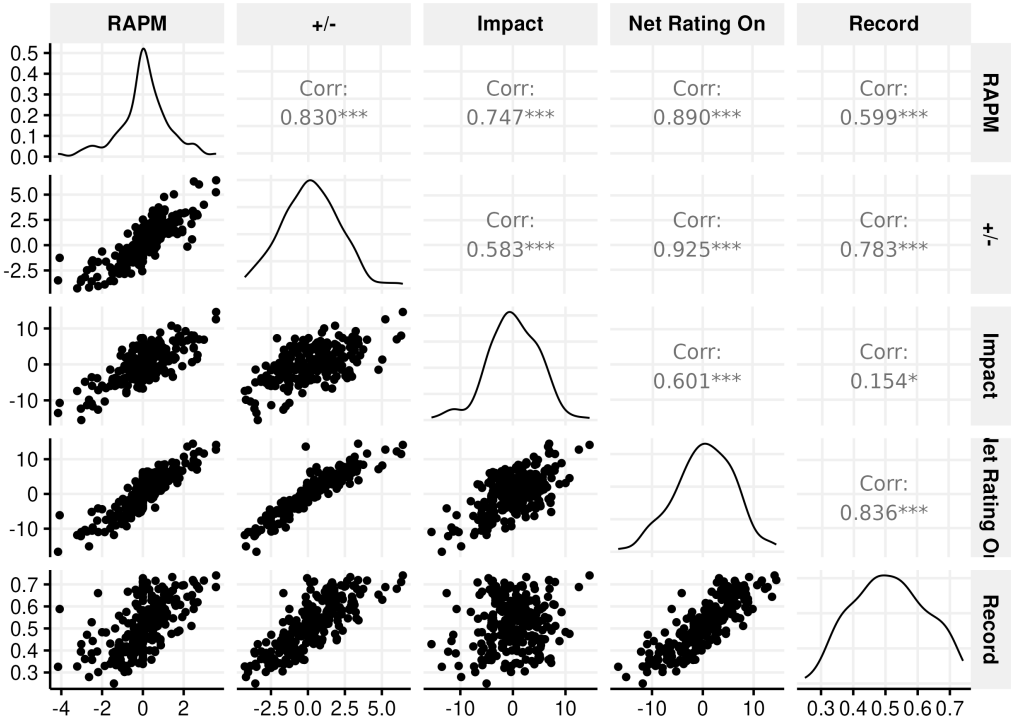


Figure 4.4: Correlations between the most adopted metrics in European basketball to evaluate a player's impact on the team

stand, often oversimplifies player impact by not distinguishing between the individual contributions of a player and the overall performance of the team. As a result, players in strong teams can have inflated Plus-Minus figures simply by virtue of playing alongside excellent teammates, even if their individual contribution is not as significant. This can lead to misleading evaluations where the metric reflects more about the team's overall strength rather than the player's specific impact.

In contrast, RAPM offers a more nuanced approach. By reducing the influence of playing within a particularly strong or weak team, RAPM focuses more on the individual player's contribution. The lower correlation with team record suggests that RAPM is less prone to the confounding effect of a team's overall performance, offering a clearer insight into the actual impact a player has on the court. This makes RAPM a more reliable metric for evaluating individual player performance, especially in diverse team contexts.

This characteristic of RAPM is particularly valuable for teams and analysts

in assessing player performance independently of the team context. It allows for more objective evaluations, especially in cases where players are part of exceptionally strong or weak teams. By focusing on the individual rather than the collective, RAPM provides a truer reflection of a player's contribution and effectiveness.

In conclusion, the lower correlation of RAPM with team record compared to traditional Plus-Minus not only demonstrates the validity of RAPM in assessing individual player impact but also highlights its advantage in mitigating the effect of a player's context.

Chapter 5

Applications

5.1 Fine-Tuning BPM

RAPM serves as starting point for several *all-in-one* metrics in basketball analytics, such as Box Plus-Minus (BPM)[32], LEBRON[33], Expected Plus-Minus (EPM)[34], and Real Plus-Minus (RPM)[35]. These metrics use RAPM as a baseline, correlating it with traditional and advanced statistics over extended periods (like 5 or 10 years) to derive coefficients for each stat. The final all-in-one metrics are then computed as a weighted sum of individual stats, using these coefficients as weights. This method melds the insights from RAPM with other statistical perspectives, offering a comprehensive view of player performance.

However, it's important to note that most of these advanced metrics have been developed for NBA basketball. European basketball presents different dynamics and styles, which could potentially necessitate a distinct set of coefficients for an accurate analysis. Despite this, even leading resources on European basketball stats, such as *Hackastat*[36], often rely on the traditional coefficients developed for the NBA. This adoption might overlook subtle but critical nuances of European basketball.

This study on RAPM paves the way for fine-tuning the coefficients of metrics like Box Plus-Minus to align more closely with the nuances of European basketball, especially the Euroleague. BPM, in contrast to RAPM, relies solely on traditional box score data, making it a more accessible metric. It assesses a player's contribution per 100 possessions without factoring in playing time, offering a rate-based evaluation of performance.

Created by Myers, BPM was intentionally designed to utilize only box score

stats, enhancing its accessibility. However, this approach comes with a notable limitation: the box score's inherent bias towards offensive stats, as described in Section 1.1.

As a result, BPM inherently skews towards evaluating offensive contributions more effectively than defensive ones. Nevertheless, it serves as a solid foundation for developing more comprehensive metrics that encompass defensive prowess.

In the European-focused implementation of Box Plus-Minus (BPM), additional adjustments such as player roles, team context, or Bayesian Regression[37] to incorporate playing time are not included. This results in a *vanilla* version of BPM. This simplified form serves as a baseline for the development of a Euroleague-adapted BPM, providing a foundational model that can be further refined to suit the specific nuances and requirements of European basketball leagues.

To compute Box Plus-Minus, a Lasso Regression is performed, utilizing offensive and defensive RAPM as target variables and box score statistics per 100 possessions as predictors. This analysis covers a span of 12 seasons, from 2011 to 2023, and is divided into four segments, each encompassing a three-year period. Following the calculation of RAPM for each segment, as detailed in Section 4.3.2, the data is then merged with the box score statistics per 100 possessions. To ensure the robustness of the analysis and eliminate outliers, players with fewer than 70 total possessions are excluded from the study, this threshold reflecting the average length of a Euroleague game. The coefficients derived from this model are presented in Table 5.1.

Stat	Coefficient off	Coefficient off
PTS	0.139	-0.008
3FGM	0.133	0.028
AST	0.163	0.029
TOV	-0.317	-0.03
OREB	0.062	-0.06
DREB	0.05	0.078
STL	0.065	0.272
BLK	0.001	0.102
PF	0.023	-0.016
FGA	-0.121	-0.008
FTA= $0.44 * FTA$	0.011	-0.011

Table 5.1: BPM Coefficients for Euroleague

Interestingly, the coefficients obtained are lower than those proposed by Myers, reflecting the stylistic differences between European leagues and the NBA. On offense, turnovers (TOV) emerge as a significant negative factor, while a penalty on field goal attempts (FGA) rewards efficient scoring. Defensively, steals (STL) are the most valuable stat, followed by blocks (BLK) and defensive rebounds (DREB).

These coefficients enable player performance evaluation over the defined three-year segments. By multiplying each stat by its respective coefficient and summing these values, it is possible to obtain Offensive BPM (OBPM) and Defensive BPM (DBPM). The overall BPM is simply the sum of OBPM and DBPM:

$$BPM = OBPM + DBPM \tag{5.1}$$

Since the stats are normalized per 100 possessions, there's no need to adjust for the number of possessions played on each end. Table 5.2 showcases the players with the highest BPM over these segments. Notably, most top performers, such as Shane Larkin, Luka Doncic, and Sergio Rodriguez, are renowned for their offensive skills. This metric, while offensively biased, serves as a valuable tool for talent identification across European leagues without the need for Play-By-Play data or extensive computational resources required by other models.

Segment	Player	GAMES	OBPM	DBPM	BPM
2017-2020	S. Larkin	59	3.83	0.53	4.35
2017-2020	L. Doncic	33	3.37	0.82	4.19
2014-2017	S. Rodriguez	55	3.35	0.57	3.91
2020-2023	D. Thompson	34	2.96	0.94	3.9
2014-2017	L. Doncic	47	2.76	1.07	3.83
2017-2020	L. Sikma	28	2.88	0.9	3.78
2020-2023	S. Vezenkov	109	3.14	0.63	3.76
2020-2023	N. Mirotic	100	3.2	0.55	3.75
2014-2017	N. Bjelica	29	2.53	1.22	3.75
2017-2020	N. Mirotic	28	3.1	0.6	3.7

Table 5.2: Best BPM segments in Euroleague from 2011 to 2023

5.2 Multi-League Analysis

In the domain of NBA analytics, a significant focus has been on predicting the potential and performance of new players, especially those transitioning from the NCAA college league. This is due to the fact that a substantial number of players entering the NBA predominantly come from NCAA basketball. This one-directional flow from NCAA to NBA simplifies predictive analytics for NBA organizations. They typically use NCAA player performance data as a known variable and NBA statistics as target variables. This approach is central to evaluating potential draft picks, helping teams gauge the future impact of these players in the NBA. While attention is also given to players from other leagues, such as the GLeague, NBL, or European leagues, the predominant and steady stream of talent from NCAA to NBA makes this predictive task more manageable and reliable.

In contrast, the landscape of European basketball presents a far more complex scenario for similar predictive endeavors. European basketball is characterized by a multitude of leagues, each with its unique style, level of competition, and player dynamics. This diversity creates a challenge in predicting the impact of a player moving to a new league within Europe, especially if they have no prior experience in that specific league. For instance, predicting how a player from Spain's ACB league would perform in Italy's LBA involves navigating a web of variables and uncertainties that are not as prevalent in the NCAA-to-NBA transition.

The primary challenge in European basketball analytics lies in the scarcity of training data and the *noisy* context. Unlike the NCAA-to-NBA pathway, where there is a wealth of data on players making this specific transition, the movements between European leagues, such as from ACB to LBA, involve far fewer players. This limited data pool hinders the development of robust predictive models. Additionally, the varied contexts of different European leagues – in terms of playing styles, tactical approaches, and overall league competitiveness – add layers of complexity, making it challenging to accurately forecast a player's performance in a new league.

As previously highlighted, RAPM is a metric that benefits significantly from extensive training data, typically necessitating multiple seasons' worth of information to enhance its accuracy and reduce susceptibility to noise. This requirement presents a particular challenge in the context of European basketball leagues, which, compared to the NBA, generally exhibit lower player retention rates. The implication of these lower retention rates is that RAPM, when applied to individual European leagues, tends to be noisier and potentially less reliable due to the frequent player turnover. This is even more relevant

when considering that European leagues usually play less games, which are shorter and played at a significant lower pace, which shrinks the number of stints each player plays throughout a season.

To mitigate this issue and address the complexities inherent in evaluating players across different European leagues, a novel approach can be employed: Multi-League Analysis using RAPM. Instead of restricting the RAPM model to a single league over various seasons, this method involves training the model across multiple leagues concurrently, still spanning multiple seasons. Such an approach capitalizes on the substantial number of players who switch between domestic leagues or participate in international competitions. Consequently, there are numerous instances where players' performances intersect across different league contexts, providing a richer, more diverse dataset for the RAPM analysis.

Implementing RAPM for Multi-League Analysis offers a significant advancement in European basketball analytics, particularly addressing the challenges faced by scouts and front-office personnel. This approach marks a significant shift from traditional methods, allowing for a comprehensive and unified evaluation of player performances across different leagues on the same scale. Such a unified evaluation is immensely beneficial in European basketball, where the diverse range of leagues often presents a hurdle in accurately assessing player quality and potential.

Additionally, the broader data base encompassing multiple leagues significantly enhances the predictive power of the RAPM model. This expansion of data not only increases the model's accuracy but also its reliability in forecasting player impact and effectiveness. By encompassing a wider array of player performances and contexts, the model becomes more robust, capable of making more precise predictions about a player's future performance. This is particularly valuable in a sport like basketball, where player performance can be influenced by a myriad of factors, including team dynamics, coaching styles, and league characteristics.

Table 5.3 provides an overview of the top 20 players in Multi-League RAPM for the seasons spanning from 2020-21 to 2022-23. These players have all participated in at least 50 games across five of the most prominent European basketball competitions. It's important to note that incorporating additional leagues into the model significantly increases computational demands, which is why this example includes only five leagues rather than all the major ones. The challenges associated with expanding the model will be further addressed in the upcoming chapter.

Player	Games	Minutes					DRAPM	ORAPM	ORAPM
		EL	EC	LBA	ACB	PROA			
W. Tavares	227	2740			2615		3.03	2.4	5.43
T. Satoransky	76	880			887		1.92	2.91	4.83
A. Hanga	204	1598			1784		3.13	1.19	4.32
C. Moneke	64	202			865	225	0.92	3.23	4.14
J. Dibartolomeo	92	1295					1.6	2.54	4.14
I. Cordinier	109	342	825	1276			2.52	1.59	4.11
N. Mirotic	210	2506			2473		1.52	2.57	4.08
S. Sanli	175	1584			1112		1.75	1.98	3.72
Y. Fall	207	1407			363	1548	1.32	2.38	3.7
M. Bilan	51			1351			0	3.44	3.44
W. Clyburn	85	2503					0	3.41	3.41
N. Melli	140	1610		1709			3.31	0	3.31
E. Muric	52		1340				1.01	2.08	3.08
G. Ricci	209	745	410	1985			2.66	0.41	3.07
R. Sorkin	58	770					2.08	0.98	3.06
M. Teodosic	151	507	884	2075			1.2	1.85	3.05
D. Thompson	111	916	358	891	747		0.47	2.44	2.91
A. Abrines	184	1533			1711		1.24	1.64	2.88
S. Shields	157	1886		2231			1.57	1.3	2.87
J. Carroll	68	460			576		0	2.8	2.8

Table 5.3: Best ORAPM and DRAPM for players with at least 50 games between 2020-21 and 2022-23 in multiple leagues

The findings in Table 5.3 exhibit a high degree of consistency with the Euroleague RAPM results presented in Table 4.7. One notable exception is the absence of Facundo Campazzo due to his limited number of games during this period. An interesting trend that emerges from these results is the overall increase in RAPM values for most of the top players, indicating an even more significant impact in their respective domestic leagues. This underscores the varying degrees of influence that players can exert across different competitions, with many players demonstrating their prowess on the home front.

Additionally, these results bring attention to players who were not prominently featured in the previous analysis due to their limited game time. Players like Chima Moneke and Isaia Cordinier stand out as impactful contributors, highlighting the depth of talent across multiple leagues and emphasizing the importance of considering a broader spectrum of competitions when evaluating player impact.

Table 5.4 offers valuable insights into the potential influence of players who have not participated in the Euroleague during the seasons under consideration. This list includes highly regarded players such as Miro Bilan and Giorgi Shermadini, both of whom have Euroleague experience in previous seasons.

Player	Games	Minutes					DRAPM	ORAPM	ORAPM
		EL	EC	LBA	ACB	PROA			
M. Bilan	51			1351			0	3.44	3.44
E. Muric	52		1340				1.01	2.08	3.08
G. Shermadini	106				2428		-0.06	2.81	2.75
Z. Nikolic	64		422			648	2.39	0.25	2.64
A. Feliz	114		767		1491		1.82	0.69	2.51

Table 5.4: Best ORAPM and DRAPM for players outside of Euroleague with at least 50 games between 2020-21 and 2022-23

Conclusions and future work

The findings from this study on Regularized Adjusted Plus-Minus in European basketball are highly encouraging and align well with the prevailing perceptions of the sport. The models successfully identified many players as impactful, corroborating their status as highly regarded figures in European basketball. This alignment between the RAPM results and public opinion validates the effectiveness of RAPM as a tool for assessing player impact and performance.

However, this research should be viewed as an initial foray into the application and significance of RAPM within the European basketball context. The field of Basketball Analytics is dynamic, continually evolving, and ripe for the adaptation and interpretation of new metrics tailored to European leagues. Although the current data availability in European basketball is somewhat limited, there is optimism that the coming years will see a more extensive and refined collection of data, enhancing the analytical capabilities in the region.

Beyond serving as a standalone metric, RAPM can act as a foundational element for a range of other advanced metrics. One such application, as discussed in Section 5.2, is the multi-league approach. This methodology could pave the way for the adaptation of metrics like Value Over Replacement Player (VORP)[38], providing team managers not only with insights into a player's overall impact but also with valuable information for salary negotiations and player valuations in the diverse and intricate landscape of European basketball.

Moreover, the combination of RAPM with powerful tools like `Sdeng` opens the possibility of developing predictive models akin to those used in the NBA, such as the DARKO model[39]. While the unique characteristics of European basketball add an extra layer of complexity to model development, the multi-league approach offers a viable solution to simplify this challenge.

In conclusion, this study serves as a stepping stone towards a more comprehensive and nuanced understanding of player performance in European basketball.

The promising results of RAPM, coupled with the potential for future analytical developments, underscore the growing significance of advanced metrics in enhancing strategic decisions, player evaluations, and the overall appreciation of the game in Europe. As data availability improves and analytical techniques evolve, the future of basketball analytics in Europe looks bright, with a myriad of opportunities for deeper insights and innovative applications.

Acknowledgments

First and foremost, I wish to thank Sofia, for her overwhelming support and help during these five years, for the laughter, the adventures, and for always having my back during tough times. It is not an understatement to say that part of this achievement is hers, too.

I wish to thank my family for their love and support, even when they did not see clearly the path I was following. My mom, for her patience and her love, my dad, for his suggestions and his trust in me, and my brother, for his friendship and the good-times.

A special acknowledgment goes to Sergio, for believing in me and for sharing his incredible expertise and knowledge with me, and to Iacopo, to whom I will always be grateful, for his unbelievable kindness and for making all of this possible. Also, a special thanks to all of the amazing coaches I had the privilege of calling colleagues, Andrea, Alberto, Matteo, Cristian, Luca, and Lassi. Thank you for helping me grow, professionally and personally.

I wish to thank my course mates, Yuri, Flavio, Enrico, and Marco, for their friendship and their inspiration to be a better student and engineer.

I wish to thank Nicolò, for being a friend and a mentor in the most sincere and genuine possible way.

Special thanks are also due to the BDSports team, professors Zuccolotto, Manisera, and Sandri, whose work has been an incredible inspiration, and whose help has been profoundly necessary in many circumstances.

Finally, the most sincere thanks to the entire Virtus organization and to all of its amazing people, for helping turn a dream into reality.

Bibliography

[1] Henry Chadwick. Beadle's dime base-ball player : a compendium of the game. *Beadle's Dime*, 1(1), 1860.

[2] M. Lewis. *Moneyball: The Art Of Winning An Unfair Game*. Business book summary. WW Norton, 2003.

[3] Dean Oliver. *Basketball on Paper: Rules and Tools for Performance Analysis*. Brassey's, Incorporated, 2004.

[4] K. Goldsberry. *Sprawlball: A Visual Tour of the New Era of the NBA*. Houghton Mifflin, 2019.

[5] S. Partnow. *The Midrange Theory*. Triumph Books, 2021.

[6] Ryan Mitchell. *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc., 1st edition, 2015.

[7] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[8] Leonard Richardson. Beautiful soup documentation. *April*, 2007.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[10] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[11] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[12] Richard Iannone, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, and JooYoung Seo. *gt: Easily Create Presentation-Ready Display Tables*, 2024. R package version 0.10.0.9000, https://github.com/rstudio/gt.

[13] Stefan Milton Bache and Hadley Wickham. *magrittr: A Forward-Pipe Operator for R*, 2022. https://magrittr.tidyverse.org, https://github.com/tidyverse/magrittr.

[14] Solc Tomaz. *Unidecode: ASCII transliterations of Unicode text*, 2023. https://pypi.org/project/Unidecode/.

[15] Jerome Friedman, Robert Tibshirani, and Trevor Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[16] Marica Manisera, Marco Sandri, and Paola Zuccolotto. `BasketballAnalyzeR`: the R package for basketball analytics. In *Conference Smart Statistics for Smart Applications*, SIS 2019, pages 395–402. Pearson, 2019. 19st-21st June 2019.

[17] Michele Conti. *L'impatto delle Analytics sulla trasformazione del Basket Moderno*. Master thesis, Università IULM, 2023.

[18] Franklin Kenter. An analysis of the basketball endgame : When to foul when trailing and leading ! 2015.

[19] Martin Manley. *Basketball Heaven*. Doubleday, 1989.

[20] J. Hollinger. *Pro Basketball Forecast*. Pro Basketball Forecast Series. Potomac Books, Incorporated, 2005.

[21] Lars Magnus Hvattum. A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport*, 18:1–23, 07 2019.

[22] Tarak Kharrat, Javier López Peña, and Ian McHale. Plus-minus player ratings for soccer, 2017.

[23] Zachary Hass and Bruce Craig. Exploring the potential of the plus/minus in ncaa women's volleyball via the recovery of court presence information. *Journal of Sports Analytics*, 4:1–11, 08 2018.

[24] Dan T. Rosenbaum. Measuring how nba players help their teams win. *82games*, 2004.

[25] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[26] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2014.

[27] Justin Jacobs. Regularized adjusted plus-minus part iii: What had really happened was. . . , 2018.

[28] Luca Grassetti, Ruggero Bellio, Giovanni Fonseca, and Paolo Vidoni. Estimation of lineup efficiency effects in basketball using play-by-play data, 2019.

[29] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[30] Hui Zou and Trevor Hastie. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 03 2005.

[31] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, dec 1965.

[32] Daniel Myers. About box plus/minus (bpm), 2020.

[33] Narsu Krishna. Lebron: The man, the myth, the metric?, 2021.

[34] Snarr Taylor. What is estimated plus-minus (epm)?, 2020.

[35] Engelmann Jeremias Ilardi Steve. The next big thing: real plus-minus, 2014.

[36] Cappelletti Luca. Hack a stat: where advanced stats happen, 2016.

[37] G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley Classics Library. Wiley, 2011.

[38] Woolner Keith. Introduction to vorp: Value over replacement player, 2001.

[39] Patton Andrew Medvedovsky Kostya. Daily adjusted and regressed kalman optimized projections, 2020.