Universitat Politècnica de Catalunya

Facultat d'Informàtica de Barcelona

Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona

Facultat de Matemàtiques i Estadística

Degree in Data Science and Engineering

Bachelor's Degree Thesis

Development of value metrics for specific basketball contexts: evaluating player contribution by means of regression

Armand Alarcón Román

Supervisors:

Sergi Oliva	Portland Trail Blazers, NBA
Jordi Cortadella	Dept. of Computer Science, UPC
Ferran Marqués	Dept. of Signal Theory and Communications, UPC

June, 2022

I would like to thank Sergi Oliva, Jordi Cortadella and Ferran Marqués for their patience, time and dedication throughout this project. Thanks to their help and advice this adventure has been possible. I would also like to express my thanks to Arnau Turch for facilitating the parts where we worked together in order to enrich our work. Finally, I can not forget to mention my family and friends for encouraging and giving me support when I needed it most.

Finally, I would like to express my gratitude to NBA Properties, Inc. for genereously agreeing to provide NBA basketball data that has been essential to carry out the research of this project.

Abstract

NBA clubs invest hundreds of millions of dollars yearly in acquiring players to help them win at the highest level. With such pivotal - and expensive - decisions to be made, their appetite for player evaluation and analysis has grown together with their technical ability to extract data from the game - going from manual to optical tracking over the last decade.

In this project we will introduce some of these analysis as it respects to certain parts of the basketball game, and will do so using a combination of manually-tracked data (play-by-play) and optical data (player-tracking). In particular, we aim to evaluate the real contribution of players in specific actions, in this case, defensive rebounding. We will approach this problem by means of regression, and propose different techniques to obtain more accurate results, including hybrid methods incorporating player-tracking data.

Keywords

basketball, play-by-play data, tracking data, defensive rebounding, data analysis, regression, player contribution

Contents

1. Introduction

Summer has just arrived and five friends, Ashley, Ben, Chloe, Daniel and Ed, have arranged to meet at the town's playground to play basketball. Being an odd number of people, they can not easily split in teams, so they decide to play two versus two by rotating and swapping pairs. After several games they feel exhausted, even more due to the scorching heat. They opt to have some energetic drink and buy a snack. While recovering and recharging their batteries in the shade, they start a lively discussion about who has been the best player of the afternoon. Daniel, who is an extremely proud person, reiterates that he has been the MVP given that he has not lost a single game. Ashley quickly reminds him that all his wins have been very close and at the last moment, with his teammate being the most decisive player. After a thoughtful silence, Ed defends that the most valuable player should be the one that, when on the court, helps his team to enlarge the scoring margin. No matter if the player scores a lot or a few points, perhaps, he prevents the opposing team from scoring thanks to his fierce defense. For this reason, he suggests to check the results of the matchups...

A, B	vs	C, D	resulted in 18 - 21.
A, B	vs	C, E	resulted in 10 - 19.
D, E	vs	A, C	resulted in 25 - 21.
A, E	vs	B, C	resulted in 23 - 15.
B, E	vs	C, D	resulted in 14 - 16.
A, D	vs	B, E	resulted in 22 - 17.

... and use plus-minus (+/-), the traditional measure to determine the contribution of a player. This number is the difference between the points scored by your team and the opponent ones when the player in question is on the court.

Ashley :	$-3-9-4+8+5=-3$
Ben :	$-3-9-8-2-5=-27$
Chloe :	$+3+9-4-8+2=+2$
Daniel :	$+3+4+2+5=+14$
Ed :	$+9+4+8-2-5=+14$

If we trusted these results, we would be quite far from reality. Is Daniel as good as Ed? Reviewing the games, we can observe that Daniel has benefited from the fact that he has avoided playing alongside Ben, who has had a very poor performance, whereas Ed has played twice with him. Without taking into account the matchups where he has been accompanied by Ben, Ed had been reaping good results. Using this method, we are not taking into consideration either with whom a player had to play alongside or against whom he had to compete. A similar case is that of Ashley. She has been defeated every time except when she has shared team with Ed or Daniel since she could benefit from their skill. As we will see later, we are in the need of a measure that acknowledges these relationships and therefore pays attention to the context.

NBA clubs invest hundreds of millions of dollars yearly in acquiring players to help them win at the highest level. With such pivotal - and expensive - decisions to be made, their appetite for player evaluation and analysis has grown together with their technical ability to extract data from the game - going from manual to optical tracking over the last decade. For this reason, mainly in the most popular USA leagues, commonly referred to as the "Big Four", the NBA in basketball, NFL in American football, NHL in ice hockey and the MLB in baseball, clubs have been adding to their coaching staffs personnel in charge of studying and analyzing

this information with the purpose of assisting in decision-making. But this data is not only useful knowledge for teams, it is also of high interest to fans, federations, journalists, etc., that may want to look up for historical statistics about games collected throughout the past seasons. While raw data collection has been carried out for decades, it is over the last two decades that more complex metrics and models have emerged - both derived from historical data, and as the result of new data sources. Professionals like data analysts, etc. are in a position to take this wealth information and analysis to understand every single aspect of the game and translate it into decision-making, on or off the court.

In this project, we will introduce some of these analysis as it respects to certain parts of the basketball game, and will do so using a combination of manually-tracked data (play-by-play) and optical data (player-tracking).

Play-by-Play data, which can be seen like the chronological description of a game, consists in records that contain all type of information concerning traditionally-tracked actions taking place in a game. In this manner, we can understand how a match has developed. As hinted in the initial example, we want to evaluate the real contribution of players in specific actions, in this case, defensive rebounding. We will approach this problem by means of regression, and propose different techniques to obtain more accurate results, including hybrid methods incorporating player-tracking data. Defensive rebounding is a key aspect of a basketball game, consisting on securing possession of the ball after a missed shot by your opponent - thus allowing the rebounding team to start their offensive possession and potentially score.

The idea of this research is to implement and analyze the present metrics that allow to measure the actual impact of each player in the league, not only in terms of scoring but also evaluating their involvement in defensive rebounding. After detecting and describing the main weaknesses of the different models currently used, new alternatives are proposed to overcome these issues. In addition, an innovative model is developed to refine the player contribution by informing and enriching each equation in the system with specific information from the possession that it represents, derived from player-tracking data.

The resulting methods may be interesting for coaches since they will be able to verify which players are more influential given a certain context of the game and, consequently, know which of them to play depending on the aspect to boost. While a good coach should already know his players and how they interact with their teammates, since he sees them every day in training, this type of analysis provides both validation and exploratory hints to that process. On the other hand, sports directors and scouts may be more excited to use these results when seeking for talented or undervalued players that may not be as flashy as others but are really doing a significantly impactful dirty job for the team when on court. There are even cases of players who have made use of data analysis of their performances when negotiating a contract with a club seeking for a higher salary. As an example, in football we have the case of Kevin De Bruyne, who hired a team of data analysts that assessed his influence within the team to broker a better contract at Manchester City [1].

At the end, we will see that the applied techniques and regression models throughout the work can be slightly modified and adapted to figure out the players' contribution in any facet of the game subject to analysis.

2. Goals of the project

As briefly discussed in the previous section, before we got down to business a set of objectives were established to mark the guidelines of the project. The main goals are listed below:

- To analyze and comprehend the current models and metrics used to objectively assess the players' contribution within the team based on the point differential they bring to the squad when on court.
- To realise which are the limitations and flaws in these methods and why they occur.
- To propose solutions and suggest enhancements that could be introduced in order to overcome these drawbacks. Even go further and define alternative models implementing these changes.
- To adapt the models so players' involvement in defensive rebounding can be studied.
- To develop an innovative method that integrates optical tracking data about the positioning of the players on the court to enrich our base dataset with more knowledge and obtain a more precise rating system. This part of the project has been carried out jointly with the results obtained by Arnau Turch during his research work [2].

3. Previous work and state of the art

There are a lot of rating systems to rank the best players from any sport, but, is assessing a player performance as simple in individual sports as in team sports? In individual sports like golf or tennis there are lists that dynamically order the athletes based on their results and achievements, allowing the entrance to tournaments depending on their current position. Nevertheless, designing individual metrics to objectively evaluate the player contribution within the team becomes a more complicated task. How do we know which actions are more important for the team to take the win? Who really has the merit, the one who finally scores or the player that dribbles the opponents to end up assisting him? Players in team sports have different roles depending on their position and physical traits, then, what attributes should we pay attention to when rating them? Not only that but also the level of difficulty to which the player is subjected, we need to consider the context, this is the rival players and even the teammates. All these factors make identifying the real impact of a single player to the whole team even more difficult to isolate.

As mentioned before, these rating systems may have different applications. In the case of individual sports, they are used as seedings in tournaments to separate top players and avoid them to meet in the early rounds of a competition or as a criterion to enter them. But in the sports in general, these metrics are useful for federations and leagues to award their best players, for scouters to identify rising stars, to negotiate contracts and transfer fees, etc. Ratings can serve as help for coaches to prepare the strategy for the upcoming games as they can detect which opposing players to focus on and what are their strengths and weaknesses. They play an important role in the betting industry too, bookmakers may rely on these scores to make more accurate predictions.

One may think that these rankings could be made without any metric and at the hands of a committee, but subjective evaluations tend to be biased by many factors. Moreover, spectators

normally are inclined to appreciate more offensive actions and impressive moves rather than defensive ones. Another thing to bear in mind is that it is impossible to watch every game from a season. As Dean Oliver, one of the basketball's pioneers of advanced analytics, said in a discussion about its place in the game [3]:

> *"Your eyes see the game much better than the numbers. But the numbers see all the games."* - Dean Oliver.

In this research we put our eyes on plus-minus ratings, a specific class of player ratings. Plus-minus ratings estimate the player contribution within the whole team by comparing the performance when he is on and off the court. Next, different variants from the literature that have led us to this point will be presented.

Let's put ourselves in situation. Pick any player and count all the points his team scores while he is playing, after that, subtract the total number of points that his team has conceded during the exact time intervals. The resulting value will be what we call the traditional or basic plus-minus of the player. This kind of rating had been used unofficially since 2003 until NBA adopted them in 2007. Weaknesses were quickly found from the outset, like the fact that they completely ignore the quality of players around you, both teammates and opponents, leading to overrate bad players in good lineups and also to underrate good players in bad teams. To solve this issue, the concept of Adjusted plus-minus ratings (APM) was introduced, proposing a multiple linear regression to acknowledge the context of the game.

The first person to publish research on Adjusted plus-minus was Dan Rosenbaum in 2004. The method he suggested splits every game into segments where no substitutions are made, in other words, the players stay the same during these intervals. This division results in multiple observations that are written down as equations that compound a system of equations that will be solved by running a multiple regression:

$$Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j X_{ij} + \epsilon_i$$

where the response variable Y_i contains the difference of home team points per possession and away team points per possession for the corresponding observation. The predictor variables X_{ij} stand for the participation of the player:

$$X_{ij} = \begin{cases} 1, & \text{if player } j \text{ is playing at home} \\ -1, & \text{if player } j \text{ is playing away} \\ 0, & \text{otherwise} \end{cases}$$

β_0 represents the importance of the home-court advantage and β_j for $j = 1, ..., K$ will be the rating obtained by the player j using this system. The error term ϵ_i refers to the difference between the observed value of the dependent variable and the value obtained from the independent variables.

Rosenbaum found out that the derived ratings were quite noisy, probably due to the parameters of the model having high standard errors associated. Therefore, he introduced weights to the model, by assigning to each observation a weight depending on the importance of it. To do that, he took into account the level of competitiveness of the game phase, for garbage time observations he assigned lower weights whereas for crunch time ones he assigned higher weights. He noticed that using additional seasons and hence more data, he could significantly reduce the noise of the estimators. For this reason, he also gave weights to observations depending on how recent the season to which they belong was.

Another procedure that Rosenbaum came up with to achieve less noisy ratings was to regress the adjusted plus-minus previously computed on a set of game statistics like rebounds, assists, shot attempts, etc. In this way, the model consists of observations where the independent variables are the player's game stats, and the dependent variable is his adjusted plus-minus rating. From this method, the actions of the game that have a major impact to the rating

can be discriminated, thus, which game events correlate the most with the previous rating. Once the influence of these factors are estimated, the known as statistical plus-minus ratings are calculated. Finally, an overall rating is computed for every player based on a combination of the adjusted plus-minus and the statistical plus-minus.

At the same time, Jeff Sagarin and Wayne Winston developed another version of adjusted plus-minus called WINVAL. This variant used a completely different response variable Y_i, they used instead the change in win probability. Such value was calculated from factors like the current score or the time remaining to finish the game.

Other pioneers in the development of adjusted plus-minus ratings were Steve Ilardi and Aaron Barzilai around 2007. Their method was pretty similar to Rosenbaum's method, the difference was that they computed both an offensive and defensive rating for each player, so in the end an assessment for each aspect is obtained.

$$Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j^O X_{ij}^O + \sum_{j=1}^{K} \beta_j^D X_{ij}^D + \beta_{K+1} X^H + \epsilon_i$$

In this model, the dependent variable Y_i refers to the points per possession the team playing offense scores. The independent variables work in a similar way as before:

$$X_{ij}^O = \begin{cases} 1, & \text{if player } j \text{ is playing offense} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ij}^D = \begin{cases} -1, & \text{if player } j \text{ is playing defense} \\ 0, & \text{otherwise} \end{cases}$$

In case it is the home team playing on offense, then X^H takes the value 1, and 0 otherwise.

Ilardi and Barzilai, like Rosenbaum, emphasized the relevance of trying to reduce the noise in the estimated ratings. Given that some pairs or group of players tend to play together most

of the time, the model suffers from high multicollinearity. They partly resolve this by adding multiple seasons, managing to reduce the level of noise.

In this sense, Joseph Sill made an important contribution in 2010 to somewhat overcome the high levels of multicollinearity and overfitting. As opposed to previous models, which added more seasons or removed players with poor minutes, Sill advocated the use of regularization and cross validation to improve prior models. Taking Rosenbaum's model and using ordinary least squares, the ratings would be obtained by minimizing:

$$\sum_i \left(Y_i - \beta_0 - \sum_{j=1}^{K} \beta_j X_{ij} \right)^2$$

Instead, if we apply regularization, a Bayesian technique known as ridge regression, we seek to minimize the following:

$$\sum_i \left(Y_i - \beta_0 - \sum_{j=1}^{K} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=0}^{K} (\beta_j)^2$$

Applying this technique corresponds to using a Gaussian prior distribution that helps to stabilize the estimated ratings, both when we have few samples and when multicollinearity is present. The value of the regularization parameter λ can be fixed by carrying out cross validation beforehand.

4. Methodology

In order to ensure the smooth progress of the project at the same time that quality results are achieved within the term, it is extremely important to attentively organize and plan the required tasks and resources. A good procedure helps to detect upcoming problems and fix them in time before they could affect the results or the accomplishment of any objective set.

Accordingly, we need an agile methodology that enables us to continuously set milestones to be met. And if any issue occurs during the execution of these, we can react rapidly, by adjusting slightly the previous goals and redirecting the research so plans come to fruition before the deadline. In addition, by setting these short objectives that can be fulfilled in a week time, immediately allows us to see results, which motivates us to continue working.

These agile techniques understand the development of a project as an incremental and iterative process. There may not be a final state that is completely closed from the beginning, but it can evolve gradually depending on the outcome of the investigation. Throughout the work, cycles take place. These cycles consist in three phases that are planning, executing and evaluating.

Therefore, for this project we arranged weekly meetings together with the directors and tutors in order to present them the advanced progress over the week. It is in these meetings where the planned and already executed tasks from the previous week are evaluated. The results from the experiments are commented and discussed, later on, the next steps are scheduled for the following week.

5. Development and implementation of models

In the following section, the different models that have been designed and implemented will be presented with their results, justifying and proving the reasons why we need alternatives that overcome the existing issues. Additionally, the whole process involving the acquisition of the data and its preprocessing to prepare it for the models will be explained in detail.

5.1 Data acquisition

The data analyzed during this research comes as a result of an agreement between Universitat Politècnica de Catalunya (UPC) and NBA Properties, Inc (NBAP). The agreement enables the use of NBA basketball data provided by NBAP, giving access during the period of time that the study is carried out. The data set we obtained consisted of both tracking and play-by-play data from the 2020-21 NBA season.

5.2 Data analysis

Before starting to work on models' development, a phase to dive into the data we will be working with is indispensable to understand it. This exploration is also very important to verify which steps are required in order to prepare the data for the models.

5.2.1 Data exploration

Inside the repository we had been given access, tracking and play-by-play data concerning the whole 2020-21 NBA season could be retrieved. Every game, either from regular season or playoffs, has its own file containing the data. Such information is structured in a JSON file that contains multiple fields like *"chances"*, *"shots"*, *"rebounds"*, *"possessions"*, *"fouls"*, *"passes"*, etc. However, we will only work with the first three. Within these branches are listed all the actions of that kind that take place during that game in particular. Every record contains attributes that characterize and describe that event.

One of the tables used is *"chances"*, chances divide possessions into smaller continuous units. To describe it, we have information like the current period and remaining time, the team that is on offense and the one that is on defense, the list of players on the court for each side, the actual score, the points scored in that action, etc.

For the *"shots"* table, apart from attributes similar to those in *"chances"*, we also have the outcome of the shot - whether it was made or missed, if it was a three point attempt, the shot type, etc.

Finally, the *"rebounds"* table has an attribute that allows to link the information from the shot that preceded the rebound opportunity, it also shows which team secured the rebound and whether it was caught by the defensive team, etc.

Regarding the optical tracking data about the positioning of the players, as expected, they consist of the coordinates for each player on the court and the ball too. The action is divided into frames, allowing to reproduce the development of the game from a cenital perspective if properly represented in an animated visualization, showing the players' movements like in Fig. 1.

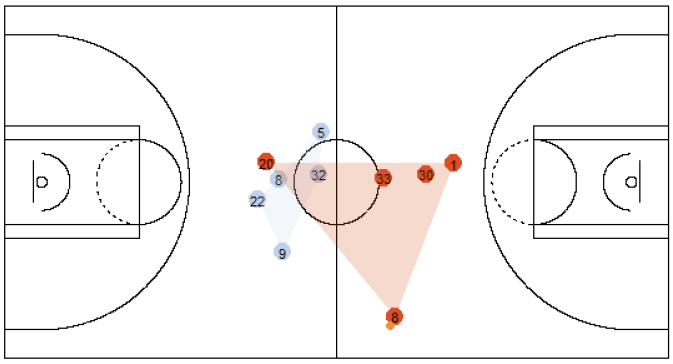

Figure 1: Example of tracking data visualization. Reprinted from *Analyzing NBA basketball data with R* by David Smith [4].

5.2.2 Data preprocessing

Our universe of data will be slightly different from the original one. We are going to apply some filters to the data in order to obtain more accurate results by considering the context of the game. Next, the distinct steps to prepare the data and the decision to apply these will be explained and justified.

First of all, we are going to discard all the actions that take place during garbage time. Garbage time is a term used in timed sports to refer to the period of time that occurs toward the end of a game, in which the winner is already decided since one team's score is so much higher than the other that the probability of a comeback is almost nonexistent. As the outcome is already determined, the game pace drops significantly and coaches start making substitutions to give rest to their best players. They are normally replaced with less experienced or younger players to gain the experience they are lacking and to protect their stars from possible injuries. For this reason, we do not take into account the events taking place in a game from the point where it can already be considered that the outcome will stand. Top players could share court with less experienced young players, making the model to overrate their performances, as now the opposing level should be lower and they can boost their stats. In addition, despite the quality of the players on the court during garbage time, the intensity of the defenses will not

be the most optimal as they tend to relax and avoid great unnecessary efforts.

Garbage time stage will start at the point where the following condition holds true until the end of the game [5]:

$$\textit{period} > 3 \quad \textbf{AND} \quad |\textit{scoring margin}| > \textit{time remaining} \cdot 0.022 + 6$$

Then, it is tagged as garbage time when we are in the last regular period or in overtime period and the scoring margin difference is big enough for the time remaining. In this way, we preserve all the actions taking place during competitive game time while getting rid of low intensity ones.

The second filter we introduce in our preprocessing pipeline is applied on the players of the league. What it exactly does is to filter out any player whose total playing time during the season is below a certain threshold of minutes. Hence, players that are under the cutoff will not be considered in our models and will not obtain a rating for their contribution. The choice of this filter was because the ratings of players with very few minutes and great performances tended to skyrocket prior to regularization. Moreover, they affected the assessment of teammates that played many more minutes, as they were not as consistent since they participated in a lot of games throughout the season. This is connected to collinearity, some players that accumulated fewer minutes had a more beneficial impact on their ratings than starting players because they took advantage from playing alongside them and performing on average at a higher level, whereas players with a lot of minutes may have a bigger variability in their performances.

For our models we considered that using a threshold of 400 minutes was appropriate, looking at the following figure, where the distribution of minutes for all the players in the league is depicted:

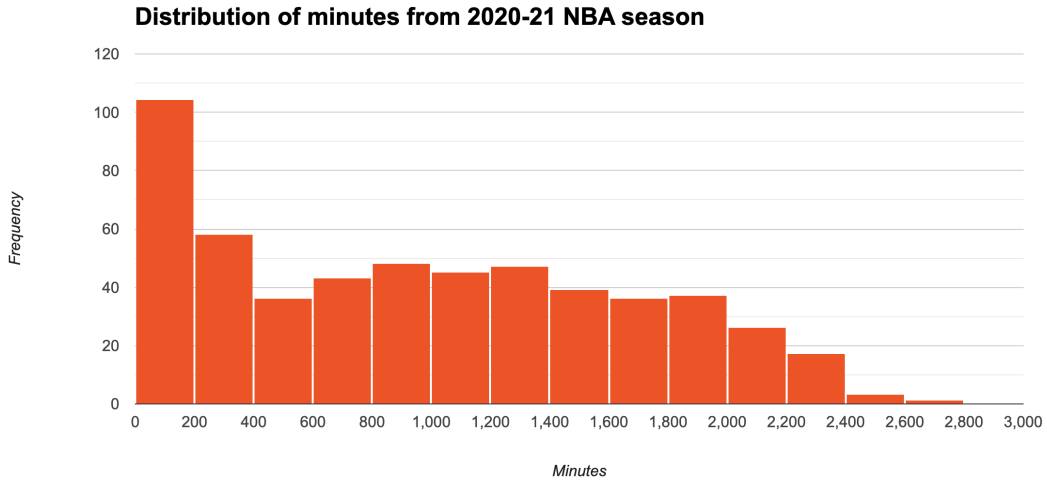

Figure 2: Distribution of minutes played by all players in the 2020-21 NBA season.

The total number of players below the 400 minutes limit is 162, representing the 30% of all the players in the league.

5.3 Models

5.3.1 Introduction to +/- (Plus/Minus) models

As already mentioned in the introduction and previous work sections, the most basic metric to measure the contribution of the players in basketball is the traditional plus-minus, which is nothing other than *net rating*. Recapitulating, it allows gauging how an individual player affects his team performance when on and off the court by taking the difference between the points his team scored and the points they allowed during his participation. The issue was that, despite obtaining an assessment of a specific player impact, this system was not taking into consideration the players around him, both teammates and opponents, with the result that we do not know if it is his merit or that of his environment. Normally, these ratings are on a per 100 possessions basis, so they are not affected by the pace of play - i.e., if a team

plays slower or faster than another. The *net rating* could be written as:

$$\textit{Net Rating} = 100 \cdot \frac{\text{Points}}{\text{Possessions}} - 100 \cdot \frac{\text{Opponent Points}}{\text{Opponent Possessions}}$$

But as you may think, this measure is too simple and is based on too little information given all the data we have at our disposal, so it is not enough to evaluate a player. We could make use of the records that are being generated during a game in order to introduce the players that are on court during a possession. This is what Rosenbaum, Ilardi and Barzilai worked on to develop the *Adjusted Plus-Minus* (APM). The difference with *net rating* was that now we are considering the quality of the players one faces and with whom one shares team. Following this latest model, Sill presented a regularized version of it, the *Regularized Adjusted Plus-Minus* (RAPM). Using regularization, Sill wanted to reduce the level of noise of the estimators obtained previously, which turned out to be quite high.

From what had already been done in the literature, a totally original model is developed. This model is based on the *Regularized Adjusted Plus-Minus*, but this time, instead of having a single regularization parameter that stays the same for all the variables corresponding to the players, we are assigning a different value as regularization parameter for each player. In this case, the penalization effect when regularizing will depend on the sample size of a particular player, which relates very deeply to the noise of his estimators. The reason behind this alternative was to penalize more the players with fewer minutes, hence less data, and not so much the players who already play sufficient minutes that we have data enough about the variability of his performances. The method has been called *Individualized Regularization Adjusted Plus-Minus* and will be explained in further detail in its corresponding section.

5.3.2 Introduction to defensive rebounding models

So far we have been talking about the contribution of the players to their teams in terms of enlarging the scoring margin, by looking at the points scored and allowed during the presence of

a player on the court. Now it is time for evaluating the player impact in defensive rebounding, by adapting the previous approach.

Starting with the most basic and simple metric, which is looking directly at the stat of the defensive rebounds secured by an individual player may seem quite unfair, and it is. Looking at the following table,

Rank	Player	Defensive Rebounds
1	Rudy Gobert	720
2	Nikola Vučević	671
3	Russell Westbrook	641
4	Julius Randle	639
5	Clint Capela	606
6	Domantas Sabonis	592
7	Nikola Jokić	575
8	Giannis Antetokounmpo	574
9	Jonas Valančiūnas	523
10	Enes Freedom	511

Table 1: Players with the most defensive rebounds in the 2020-21 NBA season.

should Russell Westbrook be so high up in the rankings being a guard? Is he inflating his stats by catching uncontested defensive rebounds as his teammates are positioned in the best spot and boxing out the opposing players or does he really hustle efficiently? An individual defensive rebound is credited when a player in the team playing defense secures a live ball immediately following a missed field goal attempt by the team on offense. Perhaps a player has been struggling with the biggest opponent during the action and in the end who gets credit is only the one that grabs it without opposition. Should we also be rewarding these players that ease the task by blocking out other players? Even further, do all players in the league have the same opportunities?

You can imagine the answers to the above questions, we need to considerate more factors. There are teams that due to their strategy and positioning when defending, allow more or less reboundable shots or force the opponent to attempt more difficult shots followed by a rebound opportunity. For this reason, we need to look at the total available defensive rebounds a player

has the opportunity to catch when on the court. This results in the *defensive rebound %* stat, shown in the following table.

Rank	Player	Defensive Rebound %
1	Clint Capela	34.3
2	Rudy Gobert	33.5
3	Enes Freedom	31.9
4	Jonas Valančiūnas	31.9
5	Nikola Vučević	30.9
6	Joel Embiid	29.1
7	Giannis Antetokounmpo	28.9
8	Domantas Sabonis	28.6
9	Russell Westbrook	28.6
10	Mason Plumlee	28.1

Table 2: Players with the higher defensive rebound percentage in the 2020-21 NBA season.

Despite everything, we are not yet taking into account the players around, in order to distribute the merit of a defensive rebound. It is here where models seen when working with plus-minus come into play. In this context, where we want to analyze the contribution in terms of defensive rebound actions, we will modify the equations of the model by adapting the dependent variable. We will be comparing two models after this variation on the response variable, one that is conditioned to the distribution and probability for the defensive team to collect the rebound depending on the shot type and another that isolates this information and only looks at the team that finally collected it. We decided to implement both models in order to compare the results and check if players that were securing more difficult rebounds than easier ones, were receiving enough reward on their ratings. In this manner, we would also be overcoming the problem that some teams facilitate easier rebounds for them because of their fierce defense strategy, so the distribution of the shots they face is acknowledged now.

In the case of rebounding, we will have an equation for each rebound action unlike in plus-minus, where possessions were grouped by lineups. The equations for the model without conditioning to the shot distribution are of the form:

$$Y_i = \sum_{j=0}^{K} \beta_j^O X_{ij}^O + \sum_{j=0}^{K} \beta_j^D X_{ij}^D$$

where:

$$Y_i = \begin{cases} -1, & \text{if defensive team grabs the rebound} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ij}^{O} = \begin{cases} 1, & \text{if player } j \text{ is playing offense} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{ij}^{D} = \begin{cases} -1, & \text{if player } j \text{ is playing defense} \\ 0, & \text{otherwise} \end{cases}$$

Since we only want to assess the contribution in defensive rebounding, we decided to split a player into two ratings: the offensive and the defensive - although we are only interested in the latter. This way, offensive rebound actions for a player do not penalize his defensive rebounding performance.

And, for the model that is informed by the distribution of the shots, the equations are:

$$Y_i - \mathbb{E}[Y_i] = \sum_{j=0}^{K} \beta_j^{O} X_{ij}^{O} + \sum_{j=0}^{K} \beta_j^{D} X_{ij}^{D}$$

where:

$$Y_i - \mathbb{E}[Y_i] = (I_i^{O} - I_i^{D}) - (P^{O}(\text{shot type}_i) - P^{D}(\text{shot type}_i))$$

$$I_i^{O} = \begin{cases} 1, & \text{if offensive team grabs the rebound} \\ 0, & \text{otherwise} \end{cases}$$

$$I_i^{D} = \begin{cases} -1, & \text{if defensive team grabs the rebound} \\ 0, & \text{otherwise} \end{cases}$$

$P^O(\text{shot type}_i)$: probability for the offensive team to grab a defensive rebound given the type of the i^{th} shot (always 0, it is impossible).

$P^D(\text{shot type}_i)$: probability for the defensive team to grab a defensive rebound given the type of the i^{th} shot.

What we are doing here in the outcome is to compare what really took place on the court - who grabbed the rebound - with what was expected to happen given the shot type. Thus, if the team on defense does not secure the rebound but what was expected was them to collect it with a high probability, those players will be highly penalized. In the same manner, if a team fulfills what we expected based on what happens on average in the league, they will not be as rewarded as if they captured a harder rebound.

The probability for a missed shot to be rebounded by the team on defense given its nature, that is the shot type, has been derived from the 2020-21 NBA season's data, by looking at the proportion of shots of a specific kind that were rebounded by the defensive team.

5.3.3 Adjusted Models

5.3.3.1 Adjusted Plus/Minus (APM)

Do you remember the story from the Introduction where 5 friends played some two versus two? Ed and Daniel had the same *traditional plus-minus*, but Daniel benefited from the fact that he avoided to play with the worst player, Ben. With *Adjusted Plus-Minus* we are going to inform the model about these interactions.

A, B	vs	C, D	resulted in 18 - 21.
A, B	vs	C, E	resulted in 10 - 19.
D, E	vs	A, C	resulted in 25 - 21.

A, E	vs	B, C	resulted in 23 - 15.
B, E	vs	C, D	resulted in 14 - 16.
A, D	vs	B, E	resulted in 22 - 17.

As our APM model's equations are of the form:

$$Y_i = \sum_{j=0}^{K} \beta_j X_{ij} \quad ,$$

the system of linear equations that represents the actual players on the court for that problem would be:

$$\begin{aligned}
-3 &= \beta_A + \beta_B - \beta_C - \beta_D \\
-9 &= \beta_A + \beta_B - \beta_C - \beta_E \\
+4 &= \beta_D + \beta_E - \beta_A - \beta_C \\
+8 &= \beta_A + \beta_E - \beta_B - \beta_C \\
-2 &= \beta_B + \beta_E - \beta_C - \beta_D \\
+5 &= \beta_A + \beta_D - \beta_B - \beta_E
\end{aligned}$$

The value of Y_i is the point differential of home team over away team. But in this small example we will keep it simple and assume that the same number of possessions have been played in every matchup. The value of X_{ij} is 1 for the home team players and -1 for the away team players on court, the rest are assigned a 0 and are not considered.

Our goal now is to obtain the values of β that satisfy this set of equations as accurately as possible. To do this, we need to perform a regression that can be done by means of Least Squares. Writing the above equations in matrix form:

$$\begin{bmatrix} -3 \\ -9 \\ 4 \\ 8 \\ -2 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & -1 & 0 \\ 1 & 1 & -1 & 0 & -1 \\ -1 & 0 & -1 & 1 & 1 \\ 1 & -1 & -1 & 0 & 1 \\ 0 & 1 & -1 & -1 & 1 \\ 1 & -1 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \\ \beta_D \\ \beta_E \end{bmatrix}$$

Following the Least Squares formulation, we multiply both sides by the transpose of the player matrix - the one containing the information about who is currently playing.

$$\begin{bmatrix} -3 \\ -27 \\ 2 \\ 14 \\ 14 \end{bmatrix} = \begin{bmatrix} 5 & 0 & -2 & -1 & -2 \\ 0 & 5 & -2 & -3 & 0 \\ -2 & -2 & 5 & 1 & -2 \\ -1 & -3 & 1 & 4 & -1 \\ -2 & 0 & -2 & -1 & 5 \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \\ \beta_D \\ \beta_E \end{bmatrix}$$

Now, the matrix in the left hand side contains the *traditional plus-minus* for each player, that coincide with those computed in the Introduction section. Whereas the result of multiplying the player matrix by its transpose is the player interaction matrix. This is a nice property since we can see in the diagonal the number of games where the player has participated. Daniel is the only one who has played 4 games, the rest 5. The other elements in this matrix show us if two players have coincided more as teammates or opponents, it is the difference between both situations. For example, as stated earlier, Daniel confronts Ben three times, meanwhile he shares team more times than faces with Chloe. The 0s in the matrix reflect that this pair has played the same number of games together as against.

Finally, to obtain the values of the β we multiply by the inverse of the players interaction matrix. Since this matrix cannot be inverted due to a condition number problem, we apply

pseudoinverse.

$$\begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \\ \beta_D \\ \beta_E \end{bmatrix} = \begin{bmatrix} 0.9 \\ -4.9 \\ -0.1 \\ 0.9 \\ 3.3 \end{bmatrix}$$

As we imagined, Daniel had not the same merit as Ed, since Daniel faced Ben three times and never played with him. In this case, the difference in the impact to their team between both players is even more evident and significant. Indeed, Daniel contributes to his team as Ashley does, who seemed to be not that good as Daniel. On the other side, we verify that Ben was the worst player of the day, with a very bad rating. But how do we interpret this rating value? Well, in this example where we considered that every game had the same number of P possessions to simplify it, we would say that Ed is expected to contribute to his team a $+3.3$ point differential per P possessions or $\frac{+3.3}{P}$ point differential per possession.

Now it is time to do this using the 2020-21 NBA season data to obtain an overall rating for each player above the minutes cutoff, explained in the Data Preprocessing segment, for that season. Unlike in the previous example, now the system of equations will have an equation for every lineup combination played during that season. All the possessions involving those specific players will be grouped, and point differential will be accumulated and computed on a per 100 possession basis. Given all the universe of data at our disposal, we will have a lot of relations and interactions between the players in the league, allowing our ratings to be more consistent and precise.

As the equations' dependent variable will be defined and normalized per 100 possessions,

we need to somehow give more prominence to the lineups that played more possessions than another as they are more stable. To do this, we use the method of Weighted Least Squares, where each observation is weighted by the number of possessions that lineup played together. With this modification, lineups that tend to be used very few and consequently have less samples, will not be as relevant. Thus, the weighted least squares estimator is:

$$\hat{\beta} = (X^\intercal W X)^{-1} X^\intercal W y \quad ,$$

and the top 20 players from that season using *Adjusted Plus-Minus* (APM) are:

Player	APM Rating
Stephen Curry	9.98
Karl-Anthony Towns	9.96
LeBron James	9.67
Devonte' Graham	9.21
Kawhi Leonard	9.19
Dorian Finney-Smith	8.91
Cameron Payne	8.33
Mike Muscala	8.22
Giannis Antetokounmpo	8.01
Jrue Holiday	7.45
Jayson Tatum	7.39
Paul George	7.38
Bojan Bogdanovic	7.36
Mike Conley	7.31
Buddy Hield	7.30
Damian Lillard	7.28
Georges Niang	7.20
Clint Capela	7.17
Fred VanVleet	7.15
Seth Curry	7.12

Table 3: Top 20 players with the highest contribution to their team in the 2020-21 NBA season according to APM.

These ratings are interpreted as the point differential per 100 possessions they contribute to their team score when on the court. For example, when Stephen Curry is playing for Golden State Warriors, his team is expected to have a $+9.98$ scoring margin per 100 possessions.

Some ratings might be not real at all since problems of collinearity take place for some players that tend to share many minutes together. Perhaps, Dorian Finney-Smith is masking the performance of Luka Dončić by reducing his rating value as they play most of their minutes together and we do not notice it. In addition, the rating values are quite high for some players that are slightly above the 400 minutes cutoff due to good performances on average. For this reason, we need a technique that helps to control this issue and penalizes players in this situation. In the next section, where regularization will be applied on *Adjusted Plus-Minus* seeking to reduce the noise of the estimators, we will try to enhance the rating system.

5.3.3.2 Adjusted Defensive Rebounding

Leaving plus-minus aside and focusing on defensive rebounding, as previously explained we will experiment with two models that only differ by their response variable. One model is informed by the shot distribution and the other only accounts for who grabbed the rebound. Being said that, this latter and simpler model does not consider if the team where a specific player plays for, forces the opponent to try more difficult and risky shots that can be easily rebounded without opposition. Nevertheless, the model conditioned to the shot type rewards those players that when on court, his team grabs more defensive rebounds that were not expected considering the league average. At the same time, easier rebounds will not have that big impact on the final rating.

The resulting ratings concerning the player's contribution in defensive rebounding for both models can be seen in the following tables, and indicate how a player increments or decrements his team chances to grab a defensive rebound when on the court:

Player	Rating
Anthony Lamb	0.23
Serge Ibaka	0.19
Ivica Zubac	0.18
Mike Muscala	0.16
Frank Kaminsky	0.16
Drew Eubanks	0.16
DeMarcus Cousins	0.16
Damian Jones	0.16
LaMarcus Aldridge	0.15
Tony Bradley	0.15
Mitchell Robinson	0.15
Jakob Poeltl	0.15
Dario Šarić	0.15
Aaron Nesmith	0.15
Killian Hayes	0.15
DeAndre Jordan	0.15
Taj Gibson	0.15
Isaac Bonga	0.14
Royce O'Neale	0.14
Jusuf Nurkić	0.14

Table 4: Top 20 players with the greatest impact on defensive rebounding for the 2020-21 NBA season, using the Adjusted Defensive Rebounding without conditioning model.

Player	Rating
Ivica Zubac	0.17
Serge Ibaka	0.16
Anthony Lamb	0.15
Mike Muscala	0.13
Jusuf Nurkić	0.12
Julius Randle	0.11
Royce O'Neale	0.10
Enes Kanter	0.10
Mitchell Robinson	0.10
DeMarcus Cousins	0.10
Jonas Valančiūnas	0.09
LaMarcus Aldridge	0.08
Al Horford	0.08
Taj Gibson	0.08
DeAndre Jordan	0.07
Derrick Favors	0.07
Cam Reddish	0.07
Anthony Davis	0.07
Clint Capela	0.07
Bismack Biyombo	0.07

Table 5: Top 20 players with the greatest impact on defensive rebounding for the 2020-21 NBA season, using the Adjusted Defensive Rebounding conditioned to shot distribution.

Comparing both rankings as depicted in the next figure, we can realize which players were being benefited from the fact that his team defensive strategy forces the opponent to try shots that are effortlessly rebounded. These are the ones that appear on the left side ranking but disappear when the shot distribution is introduced on the right side approach, as they are now penalized.

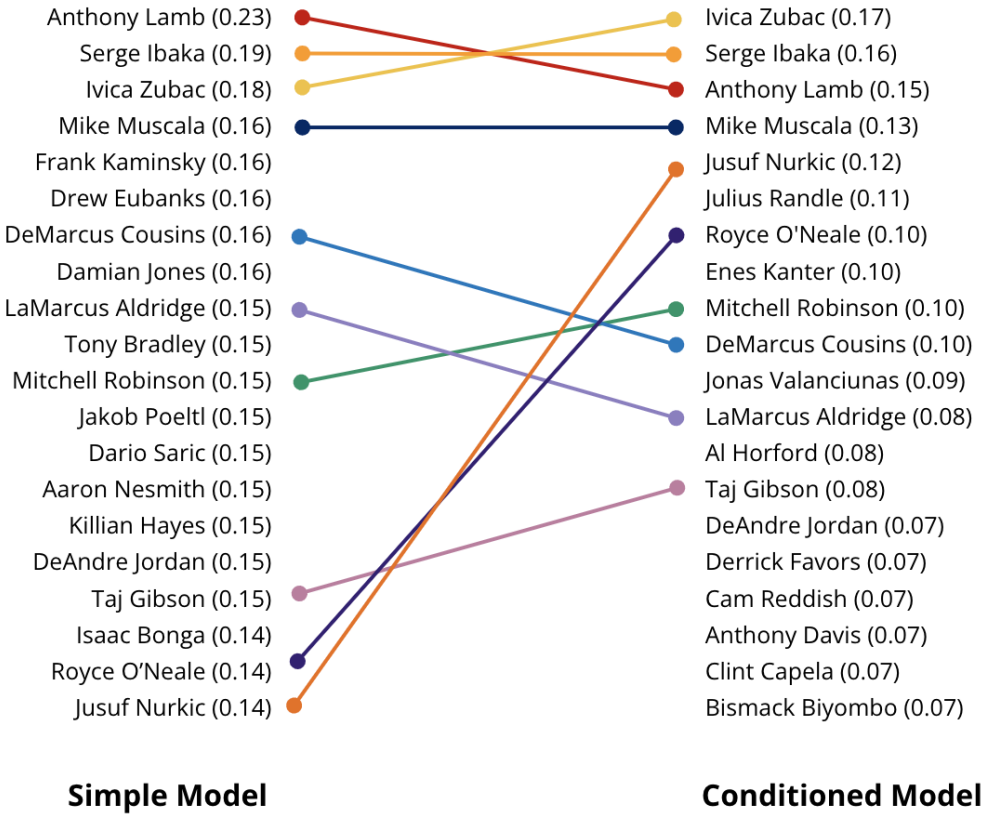

Figure 3: Bump chart between both simple and conditioned models.

From now on we will use the model that is enriched by the shot distribution, as the resulting rating system is fairer than the simple approach, which considered that all rebounds have the same merit.

5.3.4 Regularized Adjusted Models

5.3.4.1 Regularized Adjusted Plus/Minus (RAPM)

When performing the regression in the previous approach we were minimizing:

$$\sum_i \left(Y_i - \sum_{j=0}^{K} \beta_j X_{ij} \right)^2$$

now, when applying the regularization technique we seek to minimize the following expression, where a square term is added in order to minimize the noise of the β estimators:

$$\sum_i \left(Y_i - \sum_{j=0}^{K} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=0}^{K} (\beta_j)^2$$

The optimal value for this regularization parameter λ is obtained via cross-validation prior to the regression step.

Therefore, applying the *Regularized Adjusted Plus-Minus* (RAPM) technique to our data and looking at the following figure, we may consider using a λ value around 2500 since using a bigger value might be too strong and the error does not decrease and stabilizes there.

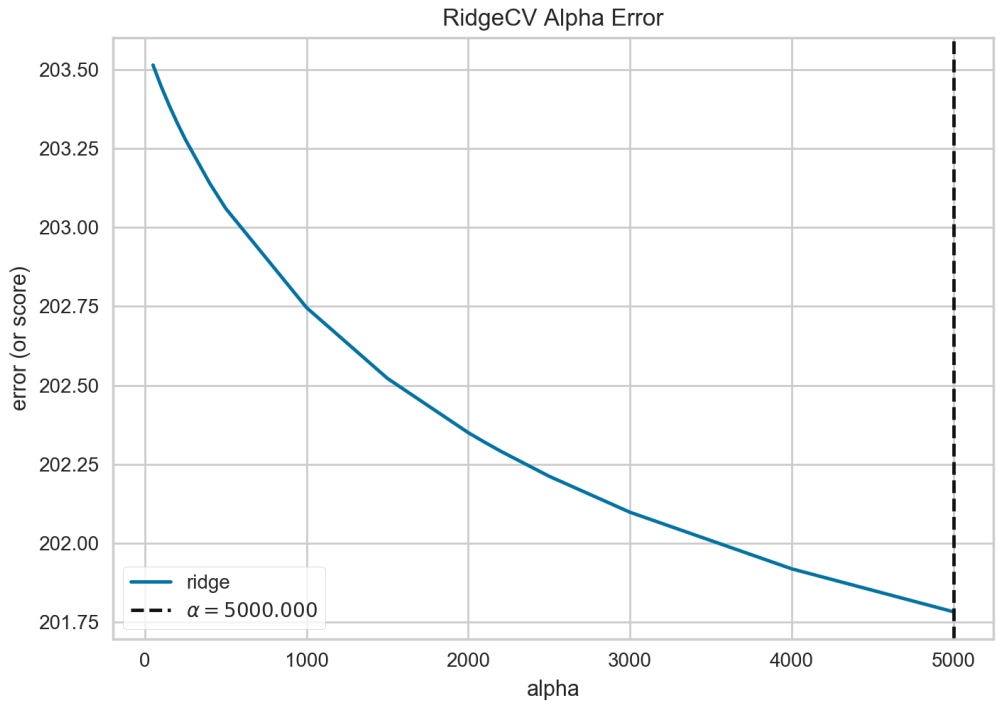

Figure 4: Error plot for the RAPM model depending on the λ value.

The ratings assigned to the top 20 players for that season after applying ridge regression are listed in the next table:

Player	RAPM Rating
LeBron James	4.29
Giannis Antetokounmpo	3.92
Kawhi Leonard	3.76
Dorian Finney-Smith	3.71
Mike Conley	3.69
Jrue Holiday	3.69
Karl-Anthony Towns	3.39
Joel Embiid	3.37
Paul George	3.36
Stephen Curry	3.34
Rudy Gobert	3.24
Joe Harris	3.14
Clint Capela	3.06
Thaddeus Young	2.97
Devonte' Graham	2.93
Jayson Tatum	2.81
Damian Lillard	2.78
Jamal Murray	2.71
Draymond Green	2.67
Dario Šarić	2.63

Table 6: Top 20 players with the highest contribution to their team in the 2020-21 NBA season according to RAPM.

If we compare these ratings with those obtained previously with APM, we will notice that now the values are not that big, they are relative to the regularization parameter defined above that controls that they do not shoot up.

As we commented in the Introduction, rating systems have many applications, but I wanted to mention a derived metric from these that allows to estimate the percentage of wins a team is expected when one specific player is on the court and if the rest of the players are average. We are talking about *Projected Winning Percentage*, a formula that takes point differential and translates it to the expected winning percentage over the course of the season.

$$\text{Projected Win \%} = [\text{Points Differential} * 2.7 + 41]/82$$

Each point differential translates to 2.7 wins along the season and each team plays 82 games during the regular season. Hence, if a team had an approximate $+0$ point differential record, a 50% of wins will be expected from them. Whereas an average team that line up LeBron James, who contributes with a $+4.29$ point differential, should win about the 64.13% of regular season games.

Since some pioneers in *plus-minus* added to their models an estimator for the home-court advantage effect, we decided to carry out some experiments related to that, but using instead a particular β for each team court. Nevertheless, the ratings for the players where very similar to those previously obtained. Despite the fact that it did not help to enhance the rating system at all, it allowed to verify some hypothesis we had.

The first experiment we performed was to verify that the teams that received a larger rating for their home-court advantage, had a significant average points scored at home per 100 possessions, given that this value is in the end derived from scoring. The following figure displays a scatter plot showing the relation between the *"Home-court rating"* and the *"Average of points scored at home per 100 possessions"*:

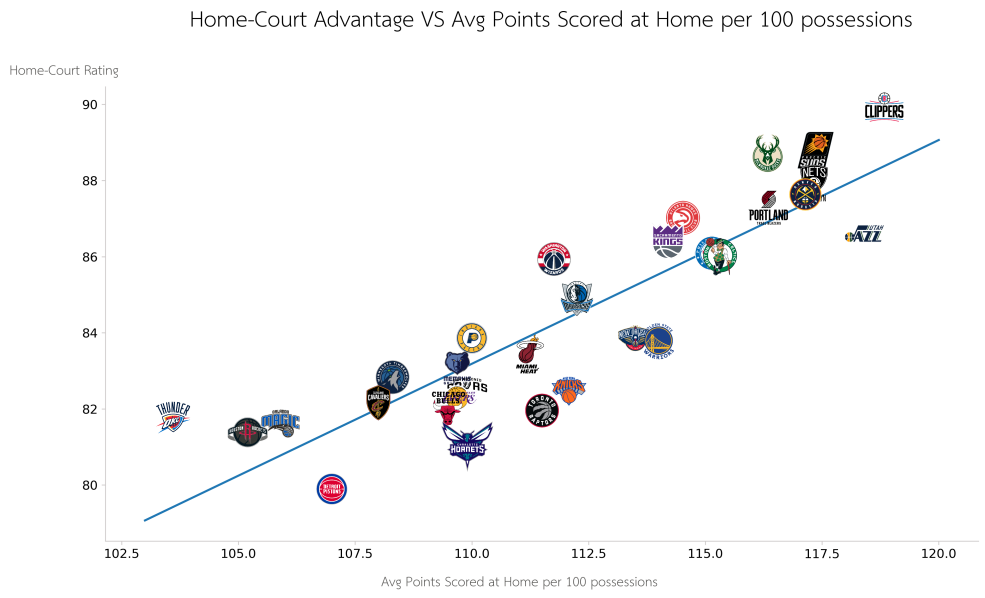

Figure 5: *Home-court rating* against *Average points scored at home per 100 possessions* scatter plot.

The correlation between both variables is ρ=0.8877, as we expected it is very high, teams that score more points at home will have a rating for their court that reflects this reality.

Another similar experiment was to check if the home winning percentage is also related to the rating of the home-court advantage. As depicted in the next figure, it is highly correlated as well, but less than before, $\rho = 0.7424$. It seems logical since the home-court rating is based on points scored and a team that scores a lot of points does not mean that they will win the game, the opponent also plays.

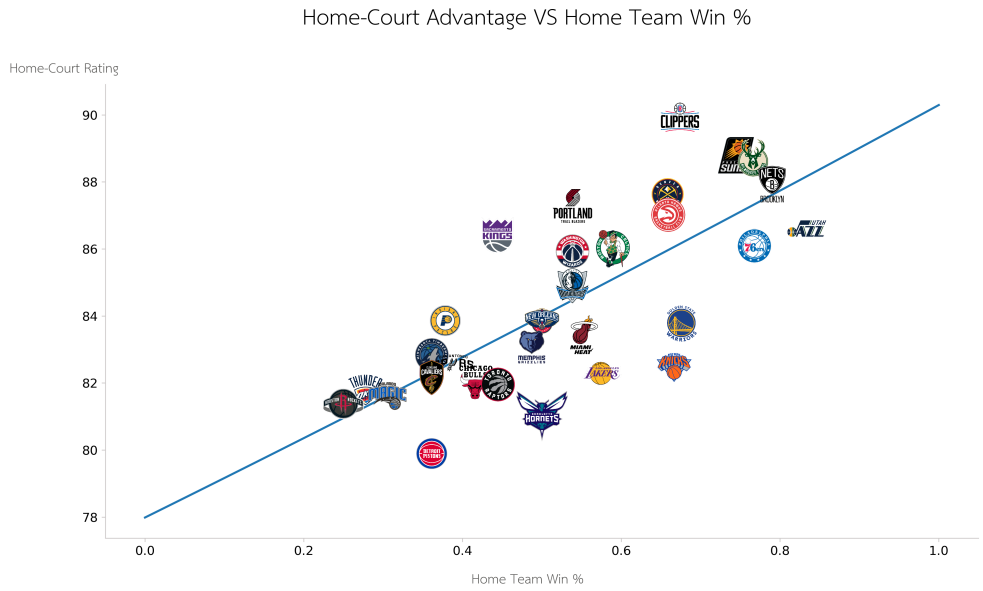

Figure 6: *Home-court rating* against *Home winning percentage* scatter plot.

A third experiment was carried out in order to find out how these ratings help to predict the outcome of a game. The approach we used to predict the winner of a matchup was to consider all the available players for both sides in that particular game and for each team accumulate the result of weighting every player's rating by their average minutes per game stat (MPG). Thus, as the minutes are related to the possessions one plays and the ratings indicate the point differential a player contributes with when on court, for each team we will be getting an expected value for the points scored. By comparing both numbers, a prediction will be made in favour of the team with a higher value. The formula for the team prediction

is:

$$\text{team prediction} = \sum_{i \in T} rating(i) \cdot MPG(i)$$

For the overall season, both regular and playoffs games, using this measure we were able to correctly predict the 68.92% of matchups. Then, we broke down this accuracy to find out which teams were easier to predict. We had the hypothesis that teams that tend to win more games or on the other hand, lose the majority, would have a greater accuracy. We used a scatter plot to test this and we were able to verify it.

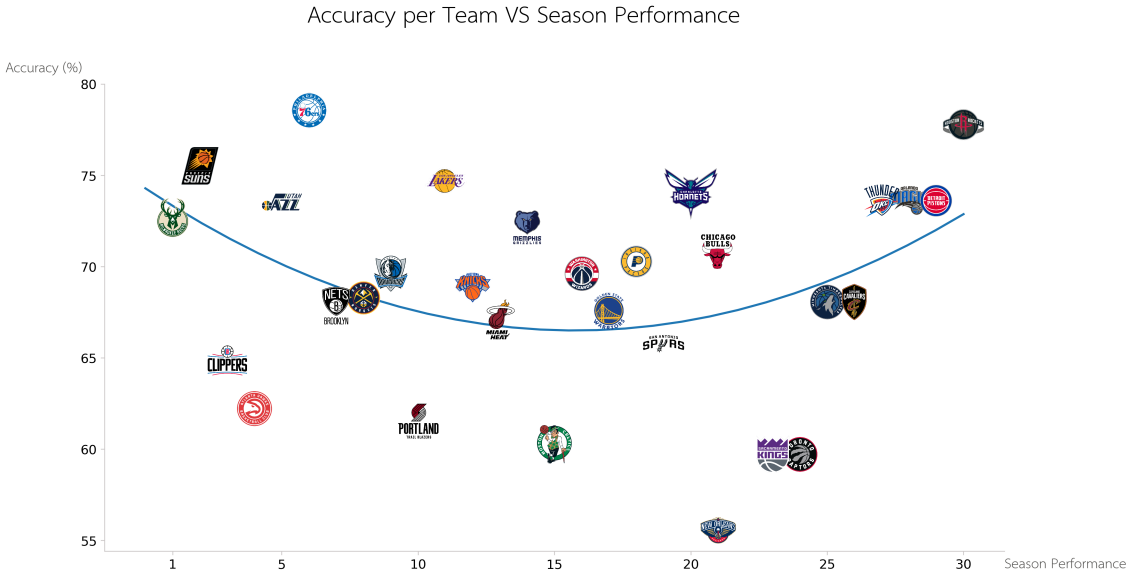

Figure 7: *Accuracy (%)* against *Season performance* scatter plot.

In this scatter plot we represented on the vertical axis the accuracy and on the horizontal axis we displayed the team's season performance. This value indicates the teams that went the furthest in that season - Milwaukee Bucks won that year, so they are assigned a "1". In case multiple teams were eliminated in the same phase, their winning percentage unties it. As we supposed, both the teams that went furthest and the ones that lost the most are easier to predict, some with an accuracy over 75%. Whereas mid-table teams, since they lose roughly as much as they win, their prediction accuracy is lower.

Finally, we wanted to observe the evolution of the teams throughout the season, such as losing streaks or good shape phases. We did that by splitting the data by months and the instant where playoffs started, so a rating for each player and stage of the season was computed. To assess the teams' performance from these individual contributions, we averaged the *"team prediction"* values obtained from the aforementioned formula for the games played during a particular phase. These values can be seen as an average point differential per 100 possessions for the specific window.

Figure 8: Team performance evolution throughout the 2020-21 NBA season.

To make it clearer, we highlighted the two finalists from the 2020-21 season, Milwaukee Bucks and Phoenix Suns. All along the season both teams have a similar performance until the beginning of playoffs from the May 22^{nd} where they have opposite peaks. The Bucks win their first round 4-0, whereas Suns needed 6 games. Then, the Bucks had more disputed rounds than the Suns. Finally, when the finals took place in July, we can observe the Bucks present a better shape as they won.

So far we have realized that applying regularization helps to control the estimators of the players and consequently reduce their noise. However, is it possible to have a more fine-tuned regularization? This is what we will try to implement in the following method explained in *Individualized Regularization Adjusted Models* section, a penalization method that does not restrict players with enough samples as much as players with fewer minutes, thus, fewer samples.

5.3.4.2 Regularized Adjusted Defensive Rebounding

In the same way we applied regularization to the *Adjusted Plus-Minus*, we also did it for the *Adjusted Defensive Rebounding*. On this occasion, the optimal value for the regularization parameter λ is 500, found via cross-validation, and it is the one the minimizes the error of the model.

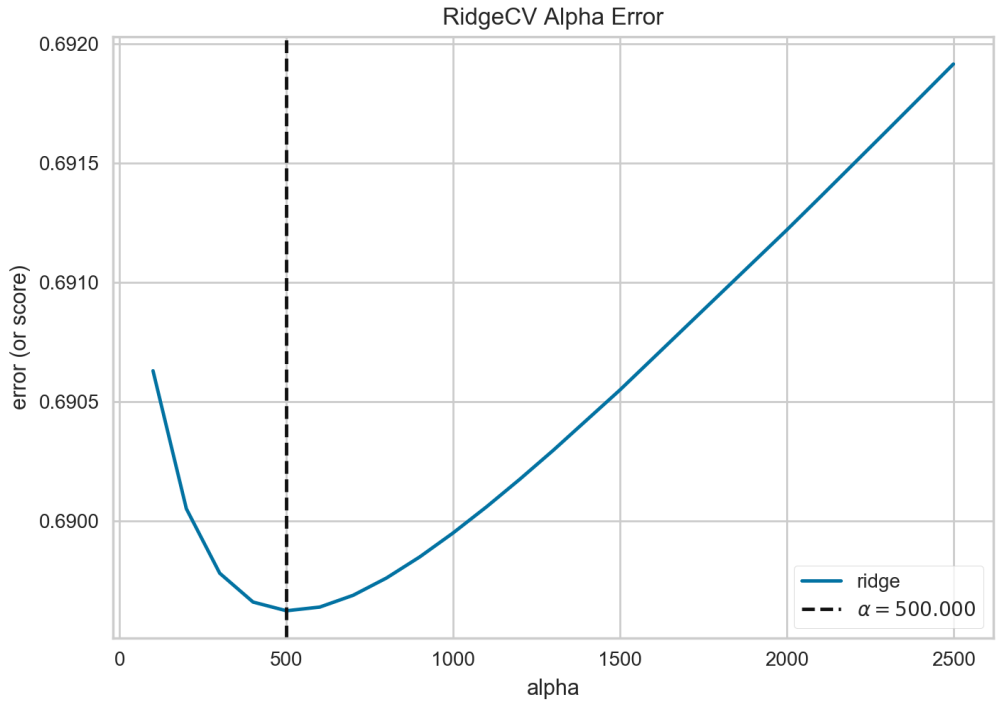

Figure 9: Error plot for the *Regularized Adjusted Defensive Rebounding* model depending on the λ value.

And the top players resulting from this method are listed in the following table:

Player	Rating
Ivica Zubac	0.05
Serge Ibaka	0.04
Nikola Vučević	0.03
Giannis Antetokounmpo	0.03
Jonas Valančiūnas	0.03
PJ Dozier	0.03
Cam Reddish	0.03
Anthony Lamb	0.03
Mike Muscala	0.02
Anthony Davis	0.02
Clint Capela	0.02
Kyle Kuzma	0.02
Coby White	0.02
Royce O'Neale	0.02
DeAndre Jordan	0.02
Jusuf Nurkić	0.02
Justise Winslow	0.02
Alex Len	0.02
Brandon Goodwin	0.02
Kawhi Leonard	0.02

Table 7: Top 20 players with the greatest impact on defensive rebounding for the 2020-21 NBA season, using the Regularized Adjusted Defensive Rebounding model conditioned to the shot distribution.

If we go back and review the players in the ranking for the model without regularization, we will realize that most of the outstanding players remain in this new ranking. Regarding the values for the ratings, these are reduced due to the regularization effect applied.

5.3.5 Individualized Regularization Adjusted Models

In this section we will explain in detail an original model that we implemented in order to perform a more refined regularization, as we have been hinting along previous points. Basically, we want our regularization method to penalize more the players with fewer minutes than the ones that play many more minutes. To understand it, the players' minutes are directly related to the number of samples at our disposal, hence, players with enough data do not need much regularization on their estimators as their value will not shoot up.

We looked for a package that would allow us to easily implement this personalized regularization for each estimator of the model. However, despite the research we were not fortunate enough to find a library that had already implemented this functionality. For this reason, we opt to adapt our equations in order to somehow carry out this desired individualized regularization.

We will try to explain our method with a very simple example involving only two players. If previously the equation was:

$$Y = \beta_A X_A - \beta_B X_B \quad ,$$

now, if player A was on the court twice as many minutes as player B, our idea was to add a weight to each β:

$$Y = 2\beta'_A X_A - 1\beta'_B X_B \quad ,$$

Doing this, a player with more minutes will obtain a lower β' estimator than before, whereas for a player with few minutes playtime, its estimator will not reduce as much. Then, when applying ridge regression we will be minimizing the summatory of the square β' instead of β. So players that used to play more, will not have such a penalized valuation since when we apply the regularization on the β', these are already small enough and they are not affected

as much as the estimators for players with fewer minutes, whose β' will be higher and hence more penalized. This is because regularization forces the estimators to shrink towards zero, with more strength for bigger values. Evidently, when getting the values for the β', we need to fix these values to get back β:

$$\beta_A = 2 \cdot \beta'_A \qquad \beta_B = 1 \cdot \beta'_B$$

In the end, what we are now seeking to minimize is:

$$\sum_i \left(Y_i - \sum_{j=0}^{K} \gamma_j \beta'_j X_{ij} \right)^2 + \lambda \sum_{j=0}^{K} (\beta'_j)^2$$

and despite λ being unique, thanks to the trick we do by playing with the estimators, it will be kind of an individualized regularization for each player. Always bearing in mind that the real assessment for the players is β, the result of fixing the obtained β':

$$\beta_j = \gamma_j \cdot \beta'_j$$

The function we have used to define the γ_j factor for each player is:

$$\gamma_j = \frac{\text{minutes}_j}{400} \quad ,$$

where *minutes*$_j$ is the total minutes played by player *j* during the season 2020-21 and *400* is the minutes cutoff that we defined previously. We divide by this value so every weight factor is greater than 1.

5.3.5.1 Individualized Regularization Adjusted Plus/Minus

Applying this method in order to obtain which players contribute the most within their teams in terms of enlarging the scoring margin, we get the following players for $\lambda = 2500$:

Player	Rating
Karl-Anthony Towns	8.58
LeBron James	8.31
Stephen Curry	8.13
Kawhi Leonard	7.84
Dorian Finney-Smith	7.74
Devonte' Graham	7.29
Giannis Antetokounmpo	7.17
Jrue Holiday	6.45
Paul George	6.44
Cameron Payne	6.27
Clint Capela	6.26
Damian Lillard	6.21
Luguentz Dort	6.17
Jayson Tatum	6.17
Buddy Hield	6.04
Joe Harris	5.91
Mike Conley	5.86
Thaddeus Young	5.83
Fred VanVleet	5.83
Seth Curry	5.68

Table 8: Top 20 players with the highest contribution to their team in the 2020-21 NBA season according to the Individualized RAPM.

We can observe that few new players appear in the ranking compared to the previous ones. Personally, the presence of Luguentz Dort surprised me because he had a -252 *traditional plus-minus*, then I consulted *basketball-reference.com*, where historical statistics can be found and checked his stats. He had a -7.7 $+/-$ per 100 possessions when on court, however, most of his teammates were even worse. But the stat that helps to understand his presence is the $+/-$ difference between when he was playing and when he was on the bench, a $+5.6$ point differential. This means that his team was bad during that season on average, and even though when Luguentz Dort played his team was still poor, he managed to improve the Oklahoma City Thunder's performance.

5.3.5.2 Individualized Regularization Adjusted Defensive Rebounding

Meanwhile, using this personalized regularization trick for the defensive rebounding actions we obtained the next players with a $\lambda = 2500$:

Player	Rating
Ivica Zubac	0.09
Royce O'Neale	0.08
Jonas Valančiūnas	0.06
Julius Randle	0.06
Nikola Vučević	0.05
Serge Ibaka	0.05
Clint Capela	0.05
Giannis Antetokounmpo	0.05
Isaac Okoro	0.05
Kyle Kuzma	0.04
Coby White	0.04
DeAndre Jordan	0.04
Domantas Sabonis	0.04
Anthony Davis	0.04
PJ Dozier	0.03
Ben Simmons	0.03
Kawhi Leonard	0.03
Alex Len	0.03
Jusuf Nurkić	0.03
Rui Hachimura	0.03

Table 9: Top 20 players with the greatest impact on defensive rebounding for the 2020-21 NBA season, using the Individualized Regularization Adjusted Defensive Rebounding model conditioned to the shot distribution.

During all this time we have been assessing the players contribution in defensive rebounding without taking into account where the players were positioned at the moment that a rebound opportunity occurs. In the next section we will develop an original model that enriches the equations with this information, helping to refine the rating system for defensive rebounding.

6. Enriching Adjusted Defensive Rebounding with positioning data

This last model is a totally original and innovative concept, we modify the equations of the *Adjusted Plus-Minus* regression with the pictures of what actually happened during a specific rebound action, that is, the knowledge about the players on the court and their interactions is combined with the usage of positioning data about these players. The idea behind this method is to tackle the problem of collinearity between players' variables. As already said, players that usually play most part of the minutes together, tend to mask each others' real performances as the model is not capable of recognizing who should be credited for a good action. Plugging these data we are introducing an informed bias that is based on the reality itself, unlike the models that applied regularization techniques, where a bias without foundation about what was taking place on the court was being introduced. Furthermore, this innovative method has a lot of potential for future studies as we will comment in the Future Work section, given its expandability and flexibility to analyze other facets of the basketball game.

Perhaps you are wondering why this had not been done before. The main reason may be that historically, when these techniques began to be developed with the objective to assess players' contribution, these optical data concerning the player's position were not available at that moment. Later on, after all the devices and systems that allow to gather this kind of data, tracking data is not popularized nor available for the public, its usage is exclusive.

The way we are going to introduce the knowledge about the positioning of the players is by adding a different weight factor for each player and rebound action. These weights require a previous computation and have been provided by Arnau Turch after his research work [2]. These values reflect the player's prior probability to grab a rebound given the position of all the players in the court for that action and the zone of the court where the shot was attempted.

Thus, the new equations are of the form:

$$Y_i - \mathbb{E}[Y_i] = \sum_{j=0}^{K} \alpha_{ij}\beta_j^O X_{ij}^O + \sum_{j=0}^{K} \alpha_{ij}\beta_j^D X_{ij}^D$$

where α_{ij} is the probability for player j in rebound action i to grab the ball.

When performing the regression for the above system of equations we obtain the following list of top players in terms of defensive rebounding:

Player	Rating
Anthony Lamb	0.29
LaMarcus Aldridge	0.26
Ivica Zubac	0.25
Serge Ibaka	0.25
Mitchell Robinson	0.25
Dario Šarić	0.24
Julius Randle	0.24
Frank Kaminsky	0.23
Jonas Valančiūnas	0.23
Mike Muscala	0.22
Al Horford	0.22
Jakob Poeltl	0.22
Derrick Favors	0.22
Drew Eubanks	0.22
Rudy Gobert	0.22
Jusuf Nurkić	0.21
Bismack Biyombo	0.20
DeAndre Jordan	0.20
Deandre Ayton	0.20
Aaron Nesmith	0.20

Table 10: Top 20 players with the greatest impact on defensive rebounding for the 2020-21 NBA season, using a model that features the positioning of the players.

As you can observe, Rudy Gobert enters the top 20 contributors in defensive rebounding ranking, something that had not happened in previous lists regarding rebound contribution presented along the work. On the other side, Royce O'Neale had been a permanent figure in all the preceding standings. And what do these two players have in common? Exactly, they both played for the Utah Jazz during the 2020-21 season and they were two of the most used

players by the coach. We could say that O'Neale was masking Gobert's performances and the former models were not capable to recognize his real contribution due to collinearity. Digging a little deeper into this comparison, Gobert almost doubled O'Neale in terms of defensive rebounds captured, being the first, the player with most rebounds of this type in the league for that season.

7. Conclusions

In this work, we have presented several models that allow to assess the contribution of the different players in the league, both their impact in enlarging the scoring margin and in defensive rebounding. Various alternatives are expounded in order to overcome the limitations that initial models have. The takeaway of this research is that with the appropriate data and right approach we are able to develop such analytics that help to objectively rate players' performances.

Recapitulating, *Adjusted Plus-Minus* are on the right track but imply some limitations that can be overcome in different manners as done in our methodology:

- **Modifying the dependent variable**. As we have seen in defensive rebounding models, by adapting the response variable, we can analyze players' effect on other aspects of the game. Moreover, it can be adjusted to introduce prior knowledge like the shot distribution to refine the rating system.
- **Introducing weights**. To add information about the positioning of the players on the court, we enriched the equations of the defensive rebound model plugging some weights for each variable. By doing this, we tackled the collinearity issue, so players did not mask each others' performances.
- **Individualized Regularization.** This third approach was intended to penalize more those players that play so few minutes that their ratings tend to shoot up, since we do not have as much data as for other players. On the other hand, the estimators of players with a higher participation are not so affected.

8. Future Work

Other adaptations of the already presented models could be implemented and designed in the future, with the objective of adding even more value to these methods. New factors that until now have not been considered could be introduced to better guide the rating system, so the evaluation of the players is more accurate.

On the other side, for this project we have adapted the *Adjusted Plus-Minus* equations in order to study the impact of the players in terms of defensive rebounding context, but these could be modified to analyze the players in different aspects of the game by defining it properly. This shows how flexible and expandable these kind of models are, changing the response variable is enough.

Despite the proposed alternatives, these techniques still suffer from some variance in their estimators. This drawback could be reduced by adding more data from previous seasons to stabilize the ratings and reduce the noise. We would have much more samples and consequently more relations and interactions between the players in the league. Nevertheless, we could not do this because we only had one season's data at our disposal. Other researchers in the literature have tested this approach by weighting the equations representing events in the past seasons with lower weights, so that more importance is given to recent ones.

Another approach could be customizing possessions in order to add even more knowledge to the models with other available data. More equations with distinct weights could be inserted to the system, concerning data about the position where a shot was attempted and where was grabbed the rebound. Then, the dependent variable would be a state variable.

Finally, during this research we have only been considering the outcome of an action regardless of what moves or attitudes players have towards this. Only the picture concerning the positioning of the players on the court was used to enrich the model of defensive rebound,

but what about everything that happens from the moment there is a throw until the ball is grabbed? We could analyze the video of any play or possession by segmenting it into frames, therefore, information about the players positioning value could be extracted from these in order to plug more equations into the model. By doing this, as there is not yet an outcome result for an intermediate frame, the response variable would represent the current expected number of points given the snapshot of an instant of the possession [6].

Glossary

play-by-play data: chronological description of a game, consists in records that contain all type of information concerning traditionally-tracked actions taking place in a match.

optical tracking: process that consists in determining the position of an object in real-time. This monitoring is done with multiple camera systems that receive data at a rate of 25 frames per second. By means of computer vision algorithms, the positional data for all players and the ball is extracted.

point differential: it is the difference between the number of points one team has scored and the number of points the opponent has scored.

plus-minus (+/-): it is a metric that keeps track of the changes in the scoring margin when a player has been on the court.

defensive rebound: a rebound occurs after a missed shot, when a player retrieves the ball. In case the player grabbing the ball is playing on defense, it is called a defensive rebound.

garbage time: term used in timed sports to refer to the period of time that occurs toward the end of a game, in which the winner is already decided since one team's score is so much higher than the other that the probability of a comeback is almost nonexistent.

box out: term that refers to the protective rebounding position that a player gets to achieve the best spot and prevent his opponent from getting the rebound.

Bibliography

[1] McDonnell, David. (April 7, 2021). *Kevin De Bruyne uses data analysts to broker £83m Man City contract without agent*. The Mirror. https://www.mirror.co.uk/sport/football/news/kevin-de-bruyne-uses-data-23870686.

[2] Turch, Arnau. (2022). *Development of value metrics for specific basketball contexts: a positional approach for the defensive rebound value.* Bachelor's Degree Project. Universitat Politècnica De Catalunya (UPC).

[3] Beck, Howard. (February 26, 2015). *Oliver: Eyes Better Than Numbers But Numbers See All The Games*. RealGM. https://basketball.realgm.com/wiretap/236833/Oliver-Eyes-Better-Than-Numbers-But-Numbers-See-All-The-Games.

[4] Smith, David. (September 2, 2016). *Analyzing NBA basketball data with R*. Revolutions. https://blog.revolutionanalytics.com/2016/09/analyzing-nba-basketball-data-with-r.html.

[5] Beuoy, Michael. (April 19, 2015) *Live Win Probabilities for the NBA*. inpredictable. http://stats.inpredictable.com/nba/wpBox_live.php

[6] Cervone, D., D'Amour, A., Bornn, L. & Goldsberry, K. (August 4, 2014) *A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes.* New York University, Harvard University, Simon Fraser University.

[7] Hvattum, Lars Magnus. (2019). *A comprehensive review of plus-minus ratings for evaluating individual players in team sports.* International Journal of Computer Science in Sport, vol. 18, Issue 1.

[8] Sill, Joseph. (March 6, 2010). *Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing*. MIT Sloan Sports Analytics Conference.

[9] Jacobs, Justin. (September 18, 2017). *Deep Dive on Regularized Adjusted Plus-Minus I: Introductory Example*. squared2020. https://squared2020.com/2017/09/18/deep-dive-on-regularized-adjusted-plus-minus-i-introductory-example/.

[10] Jacobs, Justin. (September 18, 2017). *Deep Dive on Regularized Adjusted Plus Minus II: Basic Application to 2017 NBA Data with R*. squared2020. https://squared2020.com/2017/09/18/deep-dive-on-regularized-adjusted-plus-minus-ii-basic-application-to-2017-nba-data-with-r/.

[11] Jacobs, Justin. (December 24, 2018). *Regularized Adjusted Plus-Minus Part III: What Had Really Happened Was. . . .* squared2020. https://squared2020.com/2018/12/24/regularized-adjusted-plus-minus-part-iii-what-had-really-happened-was/.

[12] Rosenbaum, Dan. (April 30, 2004). *Measuring How NBA Players Help Their Teams Win*. 82games. http://www.82games.com/comm30.htm.

[13] Rosenbaum, Dan. (August, 2004). *Defense is All about Keeping the other Team from Scoring*. 82games. http://82games.com/rosenbaum3.htm.

[14] Ilardi, Steve. (October 28, 2007). *Adjusted Plus-Minus: An Idea Whose Time Has Come*. 82games. http://www.82games.com/ilardi1.htm.

[15] Ilardi, Steve & Barzilai, Aaron. (2008). *Adjusted Plus-Minus Ratings: New and Improved for 2007-2008*. 82games. http://www.82games.com/ilardi2o.htm#table.

[16] Schuckers, M. E., Lock, D. F., Wells, C., Knickerbocker, C.J. & Hock, R.H. (2011). *National Hockey League Skater Ratings Based upon All On-Ice Events: An Adjusted Minus/Plus Probability (AMPP) Approach*. St. Lawrence University, Statistical Sports Consulting, LLC, Iowa State University, Sensis Corporation.

[17] Ghimire, S., Ehrlich, J.A. & Sanders, S.D. (August 25, 2020). *Measuring individual worker output in a complementary team setting: Does regularized adjusted plus minus isolate*

individual NBA player contributions?. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237920.

[18] Cheema, Ahmed. (August 1, 2021). *Calculating Regularized Adjusted Plus-Minus for 25 Years of NBA Basketball*. The Spax. https://www.thespax.com/nba/calculating-regularized-adjusted-plus-minus-for-25-years-of-nba-basketball/.

[19] Gramacy, R.B., Jensen, S.T. & Taddy, M. (January 15, 2013). *Estimating Player Contribution in Hockey with Regularized Logistic Regression*. University of Chicago Booth School of Business, University of Pennsylvania.

[20] Macdonald, Brian. (March 2011). *An Improved Adjusted Plus-Minus Statistic for NHL Players*. MIT Sloan Sports Analytics Conference.