Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze statistiche

Relazione Finale

Ranking methods for data analytics on players' performance in basketball games

Relatore Ch.ma Prof.ssa Laura Ventura
Dipartimento di Scienze Statistiche
Correlatore Ch.mo Prof. Luigi Salmaso
Correlatore Ch.mo Prof. Livio Corain
Dipartimento di Tecnica e Gestione dei sistemi industriali

Laureando Silvio Vadruccio
Matricola N 1106878

Anno Accademico 2017/2018

Contents

List of Figures

List of Tables

Introduction

The aim of this work is to build a ranking method for basketball players with the employment of $play-by-play$ data, in order to answer the question "Who is the best?".

After an introduction the basketball and data analysis in this field, the main statistics we are dealing with are presented. First, a main division between basketball statistics, in $Bottom-up$ and $Top-down$ measures. Then, the most employed performance indicator will be explained, namely $Player\ efficiency\ rating$, $Wins\ produced$ and $Approximate\ Value$.

Once introduced the problem of efficiency and how to assess contribution among players in basketball, there will be introduced the data used in this work. $Box\ scores$ and $play-by-play$ data will be explained and new statistics available trough ETL (Extract, Transform and Load) process will be defined.

Subsequently, ranking theory in dyadic data and pairwise comparison is introduced, explaining the set of hypothesis necessary to build a ranking among players and performing a simulation study based on the model suggested.

After this, application of this method will be shown about $Reyer\ Venezia$ players, for season 2016/2017 in Italian men Basketball Serie A1. Eventually, difference between performance measure and the ranking method presented

will be argued.

Chapter 1

The mathematics of hoops

Basketball is a team sport invented in 1891. Statistics is a bit older than that. In the last five decades, however, they discovered a common ground. Nowadays quantitative analysis proved to be a powerful tool for basketball organisations and teams, in order to keep track, analyse and take advantage of the insights produced to let players develop their game, teams improve their roster and strategies and basketball organisations to increase their visibility. These tools help to understand and measures player's performances as well as to build a tactical strategy on the court. A large variety of data are collected, from box scores, which are tables containing information about every player regarding the most known metrics (such as points, rebounds and assists), to almost continuous spatial data referring to the position of every player on the court, including the ball. In addition, the most recent years have seen the introduction in the field of basketball analytics of really advanced statistical and machine learning methods, as well as the employment of artificial intelligence.

At least in the NBA, the most known basketball league in the world, every

team has a statistical department which offers insights to coaches and players about the game and how the team can improve it. During the years, a lot of players has begun to hire private consultants to analyse their production and help them to highlight strengths and weaknesses in order to become better and better (the most famous case of such kind is former NBA MVP Kevin Durant who hired Justin Zormelo as his personal data science assistant[1]). Teams can use the help of statisticians, economists and data scientists in different areas: with data about the position of the player who makes a shot, we are able to build graphs which show where, on the floor, the player has more advantage; with data about team production in chronological order, we can quantify the efficiency during every phase of the game, helping the coaching staff to improve the game strategy; with data about the position, for every fraction of second, of every player on the court, we are able to detect and recognized which offensive strategy the opponent is performing, adjusting the defensive strategy in real time.

1.1 Basketball meets data

Research in empirical sports analysis began in the 50s, particularly in the field of baseball. In 1858, Henry Chadwick, a sports writer, developed the box score. In this way, analysts were given a summary of the individual and team performance. Unfortunately, even though The Society for American Baseball Research (SABR), founded in 1971, fostered the idea of developing research in baseball, advanced baseball statistics did not get any recognition from baseball teams until 1977, when Bill James published his "Baseball Abstracts".

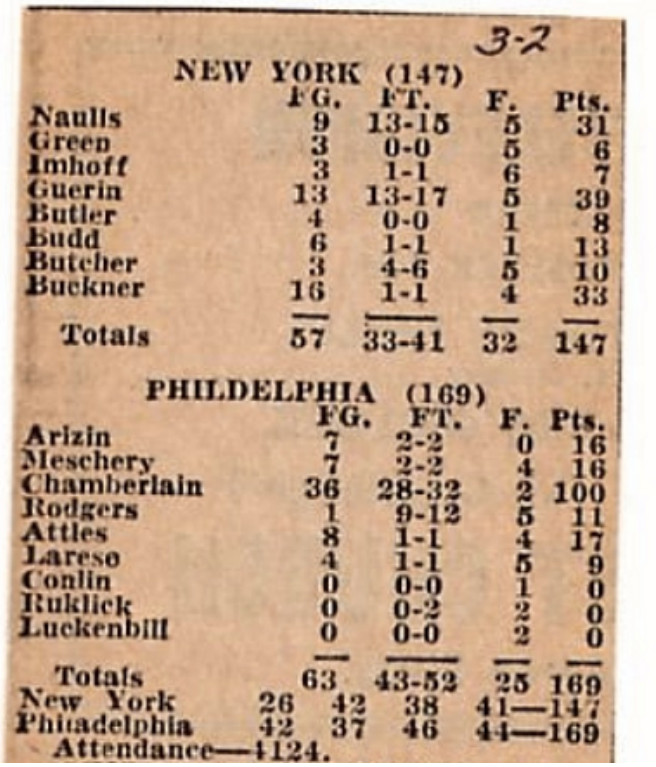
[1] `http://fivethirtyeight.com/features/meet-the-personal-stats-analyst-who-helped-kevin-durant-win-the-mvp/`

In the 90s, however, when Oakland Athletics began to apply quantitative analysis to baseball, and even more in 2002, when through statistical analysis they went on to win 20 games in a row, that this approach earned its public celebrity.

Mathematical approach in basketball came right after and during the first years of the new millennium, these kind of analysis were performed mainly by economists and statisticians for scientific purposes or as a hobby. A large fan based movement of data enthusiasts started to grow daily and even the NBA started to watch more closely. A number of websites, like 82games, fivethirtyeight, Basketball-Reference and the one of NBA now offer a very large and complete variety of data, statistics and deep analysis about the NBA.

Data themselves have become much more complete and insightful than ever. Even if the sport community relies on information provided by classic box scores, new forms of data, like play-by-play, shot charts and more advanced statistics have been introduced in common sports talks. Since the 40s, box scores have improved a lot. An example of box score from 1961/1962 season game between New York and Philadelphia can be seen in Figure 1.1. This game is really famous for the fact that the centre of Philadelphia, Wilt Chamberlain, who was the most dominant player at that time, scored 100 points. As we can see, this box score contains very few information: there are the names of the teams competing and the total score, for every quarter and at the end of the game and then information about players made shots (there was no three point shot in 1962), made and missed free throws, personal fouls and points scored.

The encounter between basketball and data collection dates back in the



3-2

NEW YORK (147)

	FG.	FT.	F.	Pts.
Naulls	9	13-15	5	31
Green	3	0-0	5	6
Imhoff	3	1-1	6	7
Guerin	13	13-17	5	39
Butler	4	0-0	1	8
Budd	6	1-1	1	13
Butcher	3	4-6	5	10
Buckner	16	1-1	4	33
Totals	57	33-41	32	147

PHILDELPHIA (169)

	FG.	FT.	F.	Pts.
Arizin	7	2-2	0	16
Meschery	7	2-2	4	16
Chamberlain	36	28-32	2	100
Rodgers	1	9-12	5	11
Attles	8	1-1	4	17
Larese	4	1-1	5	9
Conlin	0	0-0	1	0
Ruklick	0	0-2	2	0
Luckenbill	0	0-0	2	0
Totals	63	43-52	25	169

New York	26	42	38	41—147
Philadelphia	42	37	46	44—169

Attendance—4124.

Figure 1.1: New York - Philadelphia box score from NBA season 1961/1962

late 40s, largely due to the work of one man, Dick Pfander[2], who started collecting every box score of professional game played in USA as a hobby. This led the Basketball-Reference website in 2012 to have the box scores for every game in NBA history. United States has always been fertile soil for statistics in sports.

1.2 The rise of basketball data analysis

While quantitative records, in the form of more primitive box scores than the present ones, and summary statistics have always been available as additional information in sports papers and players' cards, the aim of using statistical methods to gain insights about the game is relatively modern: on February 10, 2001 Dean Oliver wrote the first post that marked the beginning of basketball analytics as it is[3]. That is the first post from the Association of Professional Basketball Research (APBR). Until the early 2000s, the only way to "learn" about the game of basketball was through intuition: smart players and coaches understood plays and movements on the court (watching hours and hours of games) and were able to adjust and overcome the opponent. With the scientific community entering this field, professionals have been able to define their questions in a mathematical rigorous way and to answer them through statistics.

1. "Does Hack-a-Shaq work? Can we define parameters under which it makes sense, such as game situations, Shaq's FT%, etc.?"

[2]"How Basketball-Reference Got Every Box Score"; http://grantland.com/the-triangle/how-basketball-reference-got-every-box-score/.

[3]`https://groups.yahoo.com/neo/groups/apbr_analysis/conversations/topics/1`

2. "What rule changes can the NBA make to improve the quality of the game? Is increased scoring an improvement? If so, what rule changes will increase scoring? Will the zone do it? Why?"

3. "What additional statistics could be taken to improve individual defensive evaluation? (I have numerous suggestions for later.) How do we evaluate individual defence?"

These are just few examples from Oliver's original post. As we can see, these questions ask really specific questions which covers not only in-game aspects but regarding the NBA league as a whole and require very specific statistical tools to be answered.

Finding a way to express a value for the player's contribution during a game has always been one of the main goal in basketball analysis. Since it is hard to compare a 30 points performance against a 20 points, 7 assist e 4 rebounds one. This led a lot of professionals to experiment methods which relied on statistical principles that could discriminate these kind of performances in order to find the best one.

Later in those years, APBR members, such as Dean Oliver and John Hollinger, published their works. John Hollinger (2005) authored a series a *Pro Basketball prospectus* and *Pro Basketball Forecast*, from 2002 to 2005, and Oliver (2004) *Basketball on paper: Rules and Tools for Performance Analysis*. During these years, NBA league and teams noticed and then hired these individuals, like Daryl Morey, an MIT Sloan School of Management graduate who, among others, created the MIT Sloan Sport Analytics Conference, perhaps the most famous basketball-analytics related annual event.

In the last years of the 90s, Hollinger worked for OregonLive as a sports editor, developing a deep work about the NBA. In these years he came up with an attempt to combine every way a player can contribute to his team in

one number: the Player Efficiency Rating (P.E.R.). This formula, very long and complicated, includes all the statistics in the box score, using adjustments for the pace of the league, in order to give a value, based on the fact that the contribution of the average player has value equal to 15. This index, along with the *Wins Produced*, is the most known summary statistics in player's performance evaluation.

In Berri, Schmidt, and Brook (2006), Berri, Schmidt and Brook published the book *The Wages of Wins: Taking Measure of the Many Myths in Modern Sport*, in which they built a model, *Wins Produced*, to estimate individual player contribution to winning. This method is grounded on strong principles, using linear model to elicit the value of every action made by a player, in order to have a single value which expresses how good the player is[4]. The output of this model is an index which express how many wins, per 48 minutes, an entire game played in the NBA, were produced by the player. By summing the *Wins produced* by every player in a team, they were able to calculate the team's expected Wins. The result shows that this method is very accurate in describing players production.

Although these methods have wide employment in sports analytics, they rely on poor data. The box score itself does not provide a lot of information and metrics included are exclusively *Bottom-up measures*, which means that they credit too much the final player of the play. We will present the main division between metrics later in this work.

The advent of new data led researchers to understand that "on-the-ball" statistics were incomplete. Highlighting the final player of a play does not give enough credit to his teammates. Moreover, players whom usefulness consists in off-the-ball movements, such as screen, cuts or patrolling the basket

[4] `http://wagesofwins.com/how-to-calculate-wins-produced/`

in order to prevent shots is totally neglected by box scores statistics. More evolved data, such as play-by-play, which is a chronological sequence of actions made by players, allow us to know every line up on the court every moment, in order to find teammates who perform particularly better when they play together, or to calculate plus-minus measures, like points made by the team and points made by the opponent when the player is on the court. In this case, analysis based on this kind of data, includes the effect of a lot more of aspects regarding the game, even tough some problems remain.

In the most recent years, Shea and Baker (2013), from basketballanalyticsbook.com, have focused their effort on assessing a value to defence in order to have an index which combine offensive and defensive production. In 2013, they published *Basketball Analytics*, a precious book which advances the researches of previous authors, using measures extractable by play-by-play data and including information about movements "off-the-ball", information that classic box scores do not provide. They were able to calculate an *Approximate Value* measure, which includes information about the player's defensive skills, by looking at the net opponent's production, when the player is on the floor and when he is not. Plus minus measures are a kind of *Top-down metrics*, which consider the team total production, when the player is present or not in the line-up. This kind of metrics can not discriminate credit between player who contributed more and player who did not play good. By combining the two type of metrics, Shea and Baker were able to mitigate the down-sides of every category.

1.3 Basketball data analysis worldwide

Spatio-temporal data are the newest form employed in analysis and they contain more information than ever. The problem of assessing the value of the contribution of a player who helps his team with screens, such as big men, can be addressed with this new type of data. Of course, methods to analyse these data still need to develop in meaningful ways and evolution in data proceeds faster. On the other hand, advanced data are employed in United States. The picture in the rest of the world is different. In Europe, few major leagues collect these information and there is not the cultural background of data enthusiasts present overseas. Few teams have statistical departments and the technology available is still out of date. Major realities, such as *Euroleague Basketball* introduced *SportVU* camera, which can collect data at a rate of 25 times per second about the position of players on the court, but it is not common across Europe.

While United States drive basketball statistical analysis, Europe follows with some years of delay. A lot of European countries, especially Eastern European ones, have a long and noble basketball tradition. Not having a strong unifying continental organisation, such as the NBA, however, led domestic leagues to develop in separate ways and in different times. In addition, a lot of actors have always been reluctant to the idea of approaching analytically to the game of basketball. This left private operators, like Opta above all the others, to enter this market gaining a competitive advantage by acquiring an incredible amount of data about all major sports played. Finally, in Italy, few teams rely on statistical analysis to improve, like Armani Milano and Reyer Venezia, even tough there are not teams statistical departments yet.

1.4 A game of spaces

What I find marvellous about the game of basketball is that it contains two challenges in it: the first one is spacing, the early phase of a play when the players have to find a clear shot by performing tactical choices altogether; the second one, the shot itself, which has a certain probability to happen (due to the spacing ability of the team) and a certain probability of going in the basket, which depends on player's shooting ability and chance.

The first phase of a play is clearly the longest one and unfortunately the one there are less data about. If a shot is made from under the basket, the probability the it goes into the basket is higher. So we can say, without loss of generality, that one main point in basketball is to reach the nearest point available to make a shot. This involves moving.

Using a metaphor, consider the defensive effort a basketball team has to endure like trying to stop a water flow with wooden boards. On the offensive end there is water, every player has a different grade of *viscosity*, meaning that the ability of passing through the boards is different. The goal, however, is to pass trough those boards. On the defensive end, there are the wooden boards, which also has different heights and widths, depending on every player's characteristics. From this point of view, we see that both efforts are team efforts, especially defence. If water flow specifically trough a certain path, the boards can stop it; if water floods everywhere, the basket is in danger. On the other hand, to stop the water, boards need to act in concert, to obstruct water on the path where there is more.

1.4.1 Talking with numbers

It is therefore non trivial trying to value every player: a player who drives undisturbed to the hoop because he is faster than his defender is one thing, a second or third offensive option who drives undisturbed to the hoop because his teammates draw a double or triple team clearing him the way is another one. On the defensive end, this operation is even more ambiguous: let us consider the case of Kawhi Leonard[5], a small forward playing for the San Antonio Spurs. He is a very good player on both the ends of the floor, especially on the defensive end. He was nominated for All-NBA defensive first team for three straight years, 2015, 2016 and 2017 and won the NBA Defensive Player of the Year in two seasons. It is so good at what he does defensively that competitors needed to adjust their offensive strategy to him. In this case, Chicago Bulls opted for a risky choice, isolating their best shooter from the play, which turned out to be one of the reasons of their victory over San Antonio Spurs. They thought that the offensive ability of the team, minus their best shooter, Jimmy Butler, was higher than the defensive ability of the opponent without their best defender, Kawhi Leonard. So they decided to remove from the play Butler, drawing Leonard out of the 2 points area and let the other four players on the court against a worse defence. The result was win. Statistics (and sports writers), however, told a story about Kawhi Leonard making his teammates worse on the defensive end.

It is clear that all this information needs to be taken into account when valuing player's performance. Modern data and mathematical methods do not offer this degree of precision yet, but research is developing more and more. In addition, statistics need context. They cannot be taken as it is, it

[5] https://www.cbssports.com/nba/news/kawhi-leonard-is-so-great-at-defense-hes-actually-hurting-the-spurs/

is necessary interpreting them and use them to help explain what happens on the court with a more scientific approach. That's why, more than talking trough number, it is necessary to talk *with* numbers, keep in mind that the one question we can ask is "Do you think we are going in the right direction?".

1.5 Data, data, data

It is simple. Basketball analysts need data.

At this time, statistical analysis, machine learning and big data are employed in almost all aspect of life and work. If we think what a smart-watch can track about our health condition or smartphones' GPS can track about our movements, we reach the conclusion that all of this information can be very useful in sports analysis. This very up to date type of data can lead to how we think the old ones. Box score can integrate measures about space coming from SportsVU cameras or measures about time coming from play-by-play data. It is also possible to develop the existing statistics considered taking advantage of the basketball literature, for example the NBA considers different categories of shots (*jump shot*, *layup*, *and1*) in order to differentiate more and more the information contained to better interpret how that player perform. Recently, a lot of additional statistics have been suggested. 82games introduced a new measure called *Potential assists*, defined as "A pass that leads directly to a possession event (shot, foul, turnover)."[6]. Since the most employed tracking method, the box score, is relying on actions, the more new actions are defined, the more data it can contain. Moreover, building box score for different phases of the game, like different quarters, or for different areas of the court, as the post or the wings, allow researchers to

[6] `http://www.82games.com/assisted.htm`

enrich the information about players' performance. However, basketball data analysis is in its first years yet and the growth of the scientific community will lead to new data and methods to help team improve.

1.6 Roles in basketball

Basketball is a sports which involves 10 big men in a very little space. So everyone needs to do his job. Different players' characteristics led to the definition of particular spaces and skills that every player should had. Players who are small, fast, with a good ball handling and passing ability tend to play far from the basket, in the *backcourt*, while players who are big and strong play near the basket, in the *frontcourt*.

There are 5 basic positions in basketball:

- Point guards (*Playmaker* in Italian plays): usually the shortest player in the line-up, is the one who runs the play. Ideally, a point guard has good passing skills and the ability to find clear teammates. He is also one of the best shooters in the team;

- Shooting guards (*Guardia* in Italian plays): is the other player which starts the play outside the two points area. A good shooting guards has great shooting and dribbling ability, can cut trough the area to receive passes to finish at the rim or to pass to a clear teammate;

- Small forward (*Ala piccola* in Italian plays): is the middle term between pure *Backcourt* and pure *Frontcourt* roles. A good small forward needs height and ability to play inside the area, where is useful to be more stronger than faster, also reaching to the wings, or shooting from the distance and cutting the area like a shooting guard;

- Power forward (*Ala grande* in Italian plays): is the one, along with the centre, who play near the hoop. A good power forward have great balance and can use his strength and it plays at a higher speed than the centre. He can also be expected to operate on the wings and corner areas;

- Centre (*Centro* in Italian plays): is the tallest of the team, positioned near the basket. It is usually the one demanded for rebounds. A good centre has the ability of obstructing the sight to the opponent's shooter.

In order to win, a line-up needs that all these individual abilities to merge togeter, and better than the opponent. Understanding how and if the line-up can merge and overcome the opponent has been investigated in the most recent years. In the continuation of this work, two types of measure will be introduced, *bottom-up* and *top-down* metrics, and the most employed method of individual's performance analysis.

Chapter 2

Performance measures

In order to summarise all possible contributions by a player, several indexes which take into account data and basketball knowledge have been invented in the last 20 years. Here a main division between measures employed in basketball analysis will be presented, followed by a review on the three most used and scientifically valid summary statistics in professional basketball, namely *Player Efficiency Rating* (P.E.R.), *Wins produced* and *Approximate Value*.

2.1 The player or the team?

In order to quantify player's production, the first statistics considered historically were *Points scored*. As we have seen in Figure 1.1 in Chapter 1, very few data were collected in the past. Nowadays we can rely on a much larger variety of data. Among the statistics available, it is possible to divide them into two main categories: *Bottom-Up* metrics and *Top-Down* metrics.

2.1.1 Bottom-up measures

A *Bottom-up* metric is a measure of an individual's production built upon records of individual accomplishment. All of the statistics in the box score, except for *Plus/Minus* are bottom-up measures. These metrics give the credit for the action to the player holding the ball. If we consider a defensive rebound grabbed after an opponent's shooter bad shot, it seems logical to consider that the player who holds the ball after the shot is the one who should be given credit for that rebound. It is not that simple.

What are other player's doing meanwhile that particular player is grabbing the rebound? Again, basketball is a team effort. So another player's might help his teammate by boxing out an opponent, preventing him to reach the ball. However, rebound statistics gives credit only to the player who grabbed the ball. Another example could be a shot made: a player that helps a shooter with a screen is not credited for his help; if players in the line-up move the ball well and they clear space for the shooter with 4 or 5 passes, bottom-up metrics will credit the shooter for the shot and maybe the player who scored an assist (because there are several definitions of "assist"). All of the other player, who actually contribute to the play, are neglected by this type of metrics.

However, measuring player's production by bottom-up metrics is not the only way to measure as individual's contribution to his team. Since 1967, *NHL* (National Hockey League) started to track a statistic called *Plus/Minus*. This is the total goals scored by the player's team while he was on the ice minus the total goals scored by the team's opponent while he was on the ice. So a player accumulates more positive plus/minus when their team is outscoring the opponent.

2.1.2 Top-down measures

Plus/minus is a typical *Top-down* measure. This kind of metrics take into account the total team's production to represent the player's individual contribution. If a player has a plus/minus of +13, it is clear that his team played well while he was in the line-up. With this approach, in the case of rebounding, a player who boxes out his opponent, letting a teammate grab the rebound is credited for his effort. If the team plays with fast ball movement and scores, all the players are credited for the points scored. So where the bottom-up metrics fall short, the top-down ones excel.

On the other hand, if we consider a team where the best player is highly better than all the others, top-down measures will credit the whole line-up evenly also when that player plays with bad teammates. If much of the points are scored by the best player in isolation, all of the players, who do not contribute to the scoring accumulate a positive plus/minus. To illustrate this situation, Shea and Baker[1] presented the case of Lebron James and Mario Chalmers in their book "*Basketball analysis: Objective and Efficient strategies for understanding how teams win*".

During the 2012-2013 NBA season, the Miami Heat were one of the top 2 teams in the NBA and finished the season by winning the Finals against the San Antonio Spurs. Looking at top-down metrics, we know that while Lebron James and Mario Chalmers were on the court, the Miami Heat were productive. If we look at plus/minus, we see that the two were first and fifth for plus/minus in the league, with +720 for James and +569 for Chalmers. If we look at the pair of them together on the court, we notice that they had +571 plus/minus on the season. Undoubtedly, if the pair was on the court,

[1] Shea, Baker (2013)

the team was outscoring its opponent. Still, it is necessary to find the portion of plus/minus attributable to each one. We can look at plus/minus when the two played singularly. Of his 2068 played in the season, Mario Chalmers played 92.9% of his time on the court with Lebron. In the remaining 146 minutes, he was -2. Unfortunately, the sample size is too small to determine a statistically significant effect. Instead James played 2877 minutes and he was + 149 in the 955 minutes played without Chalmers.

It is clear that the main problem with this type of measures is that it does not weight effort and contribution accordingly to what happens on the court. If a player shares the court with very good teammates, this will boost his plus/minus. Moreover, if a team looses regularly, its players will record, on average, negative plus/minus.

Combining the two approaches to mitigate the weaknesses and enhance the strengths could lead to more accurate performance indexes. We will now review highly employed indexes based on both these type of measures.

2.2 Player value metrics

In this section we present three important player value measures which take into account both bottom-up and top-down metrics. All of these metrics try to aggregate the many ways a player can contribute to his team into one single number. All of the three methods discussed have differences regarding the hypotheses which they use to elicit contribution's value.

When extracting individual's contribution in a game, it needs to be taken into account that players usually plays a different amount of time. So scoring 12 points in 30 minutes or in 15 should resolve in different contribution values. An easy way to standardise production among players who play a different

amount of time is dividing statistics by minutes played. In this way we can better highlight differences in production between two players; in the case above the first player has 0.4 *Points per minutes* while the second one has 0.8 *Points per minute*.

Consider now two players who plays on different teams. By rules, a basketball play can not last more than 24 seconds, in which case the team is called a *24 seconds violation*. However, average seconds per play are different among teams in a league: there will be teams which offensive strategy involves getting to the basket in few seconds (such as Mike D'Antoni "*7 seconds or less Offense*", employed in 2005-2006 with the Phoenix Suns in the NBA) and teams which prefer to hold the ball in order to find a more clear opportunity to reach the basket. In basketball literature, the average speed of a play, thus the different average possessions employed in a game, is called *pace*. Therefore, different teams play at different paces.

2.2.1 Possessions

Considering this, minutes are not the most accurate way to standardise production, because two players who play the same amount of time in teams which plays at different paces, do not have the same opportunity to record statistics. Basketball is designed in a way such that the ball possession alternates between the two facing teams. So, if a team plays at a higher pace, it will employ a high number of possessions than a team which plays at a slower pace. Of course, considering a single game, the number of possessions of the facing teams will roughly be the same, but regarding the entire season, the numbers of possessions employed by different teams can vary a lot. Let us now consider how a possession starts and ends, in order to estimate the total amount.

A possession starts when a team's member has the ball. There are four events which terminates a possession: a made shot, a turnover, a rebound from the other team and a trip to the line:

- When a player shoots the ball, there are two possible outcomes: the ball goes in or the ball goes out the basket. If the ball goes in, the opponent' team get the possession, if the ball goes out, there are other two outcomes: the team grabs the offensive rebound, keeping the possession of the ball, or the opponent' team grab the defensive rebound and get the possession;

- When a player looses the ball, the opponent's team get the possession;

- When an opponent's player fouls a player, letting him shoot two free throws (*and-1* situations do not apply because the possession already finished, since there was a made basket).

John Hollinger, in his book *Pro Basketball forecast 2005/2006*, explain that the best estimate he calculated about possessions that turns into a trip to the line are the 44% in the NBA. So, to define an estimate for the number of possessions employed by a team during a game we can write:

$$\text{Poss.} = \text{FGA} - \text{ORB} + \text{TO} + (0.44 \times \text{FTA}).$$

Possessions are the base currency of Basketball. Since it is impossible to have significantly more possessions than the opponent, the main goal for teams is spending their possessions in a more efficient way, in order to record more positive actions and less negative ones, than the opponent, *per possession.*

2.2.2 Player Efficiency Rating

P.E.R. is a summary statistics exclusively built upon bottom-up measures. This index includes all the statistics recorded in the box score: Minutes played, Points, Rebounds, Assists, Steals, Block, Field Goals and Free Throws.

It is now presented the formula and then we will explain all the parts that the P.E.R. is built on.

$$
\begin{aligned}
\text{P.E.R.} = & \left(\frac{\text{League Pace}}{\text{Team Pace}}\right) \times \left(\frac{15}{\text{League Average}}\right) \times \left(\frac{1}{\text{Mins}}\right) \times \\
& \Bigg[\text{FGM.3P} + (\text{AST} \times 0.67) + \left(\text{FGM}\left\{2 - \left[\left(\frac{\text{team AST}}{\text{team FGM}}\right) \times 0.586\right]\right\}\right) \\
& + \left(\text{FTM} \times 0.5 \times \left\{1 + \left[1 - \left(\frac{\text{team AST}}{\text{team FGM}}\right)\right] + \left[\left(\frac{\text{team AST}}{\text{team FGM}}\right) \times 0.67\right]\right\}\right) \\
& - (\text{VOP} \times \text{TO}) - (\text{FG.Miss} \times \text{VOP} \times \text{League DRB \%}) \\
& - \{\text{FT.Miss} \times \text{VOP} \times 0.44 \times [0.44 + (0.56 \times \text{League DRB \%})]\} \\
& + [\text{DRB} \times \text{VOP} \times (1 - \text{League DRB \%})] \\
& + [\text{ORB} \times \text{VOP} \times \text{League DRB \%}] + (\text{STL} \times \text{VOP}) \\
& + (\text{BLK} \times \text{VOP} \times \text{League DRB \%}) \\
& - \{\text{PF} \times [\text{League FT makes per PF} - (\text{League FTA per PF} \times 0.44 \times \text{VOP})]\} \Bigg]
\end{aligned}
$$

Let us now explain all the parts involved in this equation.

- *Pace Factor*(PF).

  It is a measure of how many possessions a team uses each game. It is defined as:

  $$\text{Pace Factor} = \frac{\text{Team.Poss} + \text{Opp.Poss}}{2 * (\text{Team Min Played}/5)} \times 48$$

  The second adjustment is made in order to represent a measure which has average equal to 15 (Hollinger thought that 15 Points per game

was expected by a decent player if he played about thirty-five minutes per game).

The third adjustment is made in order to compare players who played a different amount of time. Then Hollingers assigns a value to each of the main statistics in the box score.

- *Assist*(AST).

  This statistics is maybe the most ambiguous measure in sports. As Hollinger said, "...but an assist, at its root level, is an opinion" [2]. He considered that the as assist is worth a third of a made basket, since there are three single actions in it: spacing, shooting and passing; so an assist accounts for one of them. Since a made basket is worth 2 points, the value, on a points scale, of an assist is 0.67.

- *Field goals made*(FGM).

  They are worth 2 points, but the value of the possible assist needs to be subtracted. Hollinger thought that also Free Throws value needed to take into account the contribution of the passer to the player who got fouled. So he took from the value of the assist, 0.67, the ratio of "assisted Free Throws" and assisted Field Goals.

- *Three-points field goals made*(3PFGM).

  They just add 1 point to the value of a field goal made, since its due only to the shooter's ability.

- *Free throws*(FTM).

  They are worth one points, but the value of the phantom assist need to be subtracted. In this way, the shooter gets one point if he was unassisted and 2/3 of a point if he was assisted.

[2] "Pro Basketball forecast 2005/2006", p.6, **Hollinger**

- *Turnover*(TO).

  A turnover simply is worth a possession. The value of a possession is the ratio between points made and the number of possession the team employed. In this case is 1.04.

- *Steals*(STL).

  The same logic of turnovers is applied here.

- *Missed Field Goals*(FG.Miss).

  Shooting a bad shot results in the opponent getting the opportunity to grab a rebound and getting the possession. So the value of a missed field goal needs to take into account the value of a possession and the league defensive rebound percentage.

- *Missed free throws*(FT.Miss).

  The same logic is applied, but the impact of the miss is scaled by its amount.

- *Offensive rebounds*(ORB).

  Offensive rebounds works at the opposite of missed field goals, so grabbing a rebound while having the possession ensure the possession continues, thus its value needs to be scaled by the league defensive rebound percentage.

- *Defensive rebounds*(DRB).

  They complete defensive stops. So a missed shot and a defensive rebound should add to the value of a possession.

- *Blocked shots*(BLK).

  Hollinger considers that a blocked shots results in a field goal miss.

So it has the opposite value of a missed field goal. The team will gain possession if they get a rebound, otherwise the opponent will start over.

- *Fouls*(PF).
  The value of a fouls committed involves the value of a free throw. First we calculate the value of the Free Throws made of the amount of fouls committed, subtracting then the value of a Free Throw attempt, multiplied by the value of a possession times 0.44, which is the portion of possessions which result into a trip to the line.

Player Efficiency rating is a great effort to assign a single value to every player, on a points scale, in order to represent the amount of the player's contribution, even tough not necessarily by shooting. Nonetheless, P.E.R. measure presents a number of critical issues. First, it does not account for off-the-ball actions and defence; since it is built exclusively on bottom-up metrics, it suffers the same problems of this type of measures. Moreover, the value of each statistics varies by year and league, so to calculate P.E.R. for not NBA player, it needs to calculate each value again.

In conclusion, Player Efficiency rating is not the perfect rating statistics, but is gives a fair indication about player's production on both ends, more fair on the offensive side than on the defensive one.

Years later, another group of researchers, David Berri, Martin Schmidt and Stacey Brooks, came up with a new way of measuring how many wins a player contributed to during the season. This very elegant attempt to summarise all the statistics recorded in the box score is called *Wins Produced*[1].

[1]Berri, Schmidt, and Brook (2006).

2.2.3 Wins produced

The ultimate goal in basketball, as in all the other competitive sports, is winning. The authors' question, answered in their book, was "How does a team wins?". So, the value of the individual's performance needs to be linked in some way to the event of the player's team winning. They argued Dean Oliver's idea, explained in *Basketball on paper*, that he called "my Personal Difficulty theory for distributing credits in basketball: the more difficult the contribution, the more credit it gets". "Although difficulty may be important", they argued, "ultimately the game is about winning".

As mentioned also by Oliver and Hollinger, wins are determined by how many points per possession a team scored and the opponent score. Points scored per possession and points surrendered per opponent's possession can be seen as a productivity measures regarding team's production. We also know that the number of possessions is estimated using some of the statistics in the box score.

Let us now define to measures of efficiency in basketball, which are called *Offensive Efficiency* and *Defensive Efficiency*.

$$\text{Off. Efficiency} = \frac{\text{Points Scored}}{\text{Possessions employed}}$$

$$\text{Def. Efficiency} = \frac{\text{Points Allowed}}{\text{Possessions acquired}}$$

2.2.3.1 Percentage and efficiency

Given these two measures, the authors built a model which described the effect of points scored and points allowed per possessions on the winning percentage of a team. They estimated the model with data from season 1987–

1988 to 2010–2011 from the NBA. The model explained about the 95% of the variability of the winning percentage. In this way, they expressed winning as solely a function of offensive and defensive efficiency. The step further compared to the work of Hollinger and Oliver was that they explained wins through regression analysis, which gives empirical solidity to the the thought and insight offered from their predecessors.

They went further on modelling winning percentage on points scored, possessions employed, points allowed and possessions acquired. In this way they found a value for those statistics in order to calculate of many wins the player contributed to.

The statistics considered in the box score which value can be extracted from this model are:

- Three points field goal (3PFGM);
- Two points field goal (2PFGM);
- Free throw made (FTM);
- Missed field goal (FG.Miss);
- Missed free throw (FT.Miss);
- Offensive rebound (ORB);
- Turnovers (TO);
- Defensive rebounds (DRB);
- Steals (STL);
- Team rebounds (Team.TRB).

Blocked shot(BKL), personal fouls(PF) and assists(AST) here are not considered in the formula for possessions. So their value needed to be estimated in other ways.

To determine blocked shots' value, a model was estimated connecting opponent's made field goals to opponent's field goal attempts, blocked shots, and dummy variables for teams, years (1973-74 to 2010-11), and leagues (data from both the ABA (American Basketball Association, active from 1967 to 1976) and NBA were employed. The R^2 of the model estimated is equal to 93%.

The value for personal fouls is determined by the value of Opponent's Free Throw made. They considered the percentage of fouls the player committed on his team and multiplied it for the amount of free throws made by the opponent.

Now the production for a single player can be computed as a weighted sum of each statistics in the box score multiplied by the value given by the model. The authors computed also a measure for production per minute, by dividing the total production by the amount of minutes played, then calculating a per 48 (*P48*) minutes measure.

2.2.3.2 Including Assists

Assists are not part of offensive or defensive efficiency, but definitely have impact on the outcomes. Specifically, a player's shooting efficiency is related to the number of assists his teammates accumulate. To see this, the following model was estimated: player's adjusted field goal percentage as a function of player's adjusted field goal percentage last season, age, age squared, percentage of games played last two seasons, dummy variable for position played, dummy variable for new coach, dummy variable for new team, dummy vari-

able for year, stability of roster, the teammates' per-minute production of assists, the teammates' adjusted field goal percentage.

To incorporate the value of assists into Wins Produced, they calculated for each player his Teammates' Assists per Minute (TAPM), defined as:

$$\text{TAPM} = \frac{\text{Team Assists - Player Assists}}{\text{Team Minutes - Player Minutes}}$$

Multiplying TAPM for each player by the coefficient on TAPM from the above model and then by 2, allows us to see how TAPM impact a player's points-per-field goal attempts. Further multiplying the result by field goal attempts shows us how many points a player scored should be credited to his teammates. Multiply by the impact points have on wins represents how much of a player's production of wins should be credited to his teammates. They then allocated the outcome across all players on a team by the percentage of assists on the team that are credited to each player.

Through a series of adjustment, for position played, team defence, total rebounding, they came up with a relative adjusted per 48 minutes measure relative to the position played. To calculate a measure without regard to the role it needs to be added the average number of wins a player contributes per 48 minutes, in order to have a per 48 minutes wins produced measure (*WP48*). Eventually, dividing by 48 and multiplying by the amount of minutes played, gives the final measure of Wins produced by the player in that season.

2.2.3.3 Empirical results

The accuracy of this method is confirmed by the empirical results: for the 2003-2004 NBA season the average error between wins produced and actual wins is 1.67 wins. The framework built to calculate team wins is very

articulate, it takes into account, as said earlier in the chapter, that a team to succeed needs to win, that different roles contribute in different ways and that player's production is connected to his teammates' production. Combining all of these ideas is fundamental to understand how two groups of people interact during a match.

Although in this method is considered an adjustment for team's defence, Wins produced remains a summary statistics based on bottom-up metrics. Until the most recent years there have been few attempts to aggregate different kinds of measure, which could explain how the player behaves on the defensive side of the court. In 2013 appears the last and latest attempt to "aggregate the many ways a player can contribute to his team", by Stephen Shea and Christopher Baker, who also focused their research on efficiency as the engine of wins.

2.2.4 Approximate Value

2.2.4.1 Offensive efficiency

Shea and Baker initially addressed the problem that they needed both bottom up and top down metrics. Then investigated efficiency regarding individual players. They define an *Offensive Efficiency* (OE) measure based on individual performances. This measures accounts all the successful offensive possessions the player is involved in and the player's total number of potential offensive ends. The measure is defined as:

$$\text{OE} = \frac{\text{FGM} + \text{AST}}{\text{FGA - ORB} + \text{AST} + \text{TO}}$$

The main difference from this definition to the Hollinger and Oliver's one is that this one does not account for points but rather for field goals

made. The other one is including assists, since at a team level is a redundant information, but at an individual level absolutely not. The aim of this index is to offer a measure that does not give too much credits to players who make a lot of points but are not so efficient in terms of field goal percentage. The objective is to have a measure which give more credit to players who are more efficient. Moreover, being a relative measure, so that the effect on offensive efficiency of a player who does a lot of efficient actions and adds one bad decision is less detrimental than the effect on the OE of a bad player.

Modelling team win percentage on the mean of Offensive Efficiency for all the players in the team displayed that about the 56% of the variability was explained by the model.

In order to account for the different values of shots, they defined the *Efficient Points Scored* (EPS), which considers the value of shots, adjusted for a quantity depending on the total amount of Points scored in the League. It is defined as:

$$\text{EPS} = \text{F} \times \text{OE} \times \text{PTS},$$

with $F = \frac{\text{League Points}}{\sum(\text{OE*PTS})}$.

In this way, EPS expresses, on a points scale, the amount of points adjusted for efficiency. In this way we know how many points a player scored in a efficient way.

To express the whole *Efficient Offensive Production* (EOP), which considers also for the value added by assists made, they thought that assists have give contribute on different degrees, since an assist on the perimeter that allows the shooter a three pointer can be seen as almost all merit of the shooter, while an assist served at the rim should be credited mainly to the passer. They recorded that about the 38% of assists in the 2012–2013 NBA

season led to a basket at the rim. So assumed that the value that should be credited to the passer is $2 * 0.38 = 0.76$PTS.

They then define the *Efficient Offensive Production* as:

$$\text{EOP} = \frac{\text{League PTS}}{\sum(\text{OE} \times (\text{PTS} + 0.76 \times \text{AST}))} (\text{PTS} + 0.76 \times \text{AST}) \times \text{OE}.$$

Since EOP is higher for guards, they further adjust this quantity for position played. They carries on the already existing idea in the sports literature of *replacement player*. This idea, introduced in Baseball, allows to compare EOP to the team or league average or to a real player.

The results is a summary of the whole individual's offensive production. Introducing a measure specifically designed to describe player's efficiency is a great step forward compared to the previous works. Moreover, EOP is on a points scale, which offers good interpretability. Involving exclusively bottom up measures, however, leads this measure to suffer from the same issues. At this regard, let us now look at another measure introduced by Shea and Baker that involves primarily top-down metrics: *Defensive stops gained*.

2.2.4.2 Defensive stops gained

To measure how good a defender is, we need to take into account the opponent's offensive production. A player can affect the opponent's offensive production by causing more bad shots, by avoiding the opponent's to grab offensive rebounds after a bad shot and causing the opponent's to turn over the ball more. This are the only ways a player has to stop defensively his opponent. To measure team's shooting ability, it is considered the *Effective Field Goals ratio* (eFG%), definded as:

$$\text{eFG\%} = \frac{\text{FGM} + (0.5 \times \text{FGM.3P})}{\text{FGA}}.$$

This measure accounts for the different points values between two and three pointers. When looking at the opponent's offensive rebounds, we can consider the portion of Offensive rebounds grabbed by the opponent divided by the amount of potential offensive rebounds available. This ratio statistics is called *Offensive Rebound percentage* (ORB%), defined as:

$$\text{ORB\%} = \frac{\text{ORB.opp}}{\text{ORB.opp} + \text{teamDRB}}.$$

If a player defends well, the ORB% of the opponent should be lower when he patrols the area. It is possible to express the propensity of the opponent to turn the ball over by taking the portion of possessions employed by the opponent that resulted in a turnover. This measure is called *Turnover percentage* (TO%), defined as:

$$\text{TO\%} = \frac{\text{TO}}{\text{Poss}}.$$

Looking at the opponent's production, it is possible to extract the player's contribution in defensive stops by considering the opponent's offensive production when the player is on the court compared to the opponent's offensive production when he sits on the bench. Authors considered that for NBA season 2012–2013, a team averaged 82 field goal attempts per game and 20 Three points field goal attempts per game. This lead to about 92 effective field goal attempts, so a percentage drop in eFG% result in the opponent missing an additional 0.82 field goals. Regarding offensive rebounds, a team averaged 42 rebounds per game, so a point drop in percentage mean 0.42 less offensive rebounds given up. Moreover, about the 73.5% of the missed shot were rebounded by the defence. So, 0.82 field goals saved leads to $0.735 * 0.82$ defensive stops gained. Regarding turnovers, the average turnover ratio for the league was 13.7% and the average numbers of possessions was 106, so a

points drop in TO% is worth 1.06 defensive stops.

Considering the difference for the measures between when the player is on the court compared to when he is not, it is possible to calculate the *Defensive Stops Gained per full game*, defined as:

$$\begin{aligned} textDGS/G = &- (0.82 \times 0.735 \times \text{eFG\%[Net]}) - (0.42 \times \text{ORB\%[Net]}) + \\ &+ (1.06 \times \text{TO\%[Net]}) \,. \end{aligned}$$

This quantity shows how many defensive stops the player gains if he plays the whole game. Dividing by 48 minutes and multiplying by the number of minutes played gives as a result the number of defensive stops gained (DSG). From this we are able to calculate the number of *Defensive Points Saved* (DPS), multiplying DSG by 2 points.

2.2.4.3 Combining offence and defence

Being expressed in a points scale, *Efficient Offensive Production* and *Defensive Points Saved* can be added up to represent approximately a player's contribution value. This is why this quantity is called *Approximate Value*, defined simply as:

$$\text{AV} = \text{EOP} + \text{DPS}.$$

One of the great merits of this summary statistics is to combine offensive and defensive effort, specifically tailored on single players rather than teams. Extracting the value of the defensive performance looking at the opponent offensive production allow us to use more and more net top-down metrics, that were not taken into account until now.

From this point, we can start thinking to a way to measure and/or compare players, keeping in mind that it is more suited considering the individual's production, his opponent's production but also the production of his teammates when he is on the floor. To add complexity to this, we could add characteristics of both team and the tournament, in order to use as much data as possible.

Let us now introduce a slightly different framework in order to measure how good or bad the player's production is, during a season, compared to his teammates or others.

Chapter 3

Basketball Data

There are multiple ways in which to track down how players are performing during a basketball game. Like baseball, a basketball play can be broke down to basic actions performed by players, such as shot, pass, rebound or lost ball.

Basketball literature includes a high number of actions. There are actions, which are called *individual on ball fundamentals*, which are the basic skills required to every player, namely shot, pass, dribble and layup. Of course, a play involves more actions than these ones: a player can grab a rebound or box the opponent out on a rebound, a shooting guard can cut to the basket or can turn around on a screen, made by a frontcourt player, to receive the pass from the ball-handler and make a shot. It is clear that a lot of these basic actions consider the case that a player can move (and should do it) even without the ball. Setting a screen for a teammate can ensure him a clear shot. Boxing out a defender after a shot can provide that the ball, therefore the possession, remains to the team. Let us now review two ways of tracking and presenting data during a basketball game: the box score and the play by play.

Figure 3.1: 11° first-round game between Reyer Venezia and Emporio Armani Milano.

3.1 The box score

The box score is the main support used to record basketball games. It is employed in all the highest level leagues in the world and even minor leagues teams, such as the Unipd Men Basketball team adopted the box score to record every team game. Legabasket offers the box score for every game since 1987/1988 season. Let us examine the information contained in a box score from Italian men Basketball Serie A1.

First, there are general information about the game: the name of the league, the number of the game and the round, date and time and the name of the venue. Then there are the teams' names and the final score; in addition the score and the progressive score for every quarter is recorded. At the end there are information about the referees. An example can be seen in Figure 3.1.

After the general information, it is presented the real box score: for every team there is a table which contains information about a series of single actions regarding every player. The statistics recorded in Italian men Basketball

Serie A1 are:

- *Points scored* (PTS).
  It is the total amount of points scored by the player. In basketball there are three different types of shots, two points shots, three points shots, both of them are included in the field goals category, and free throws, which are worth one point.

- *Minutes played* (Min).
  It is the total amount of minutes that the player was on the court.

- *Fouls committed* (PF).
  It is the amount of foul plays the player committed during the game. In this case, players can be called for a foul even if they sit on the bench.

- *Fouls drawn* (FD).
  It is the amount of foul plays the player was committed to.

- *2 Points field goal attempts* (2PFGA).
  It is the sum of made and missed shots taken in the two-points area.

- *2 Points field goals* (2PFGM).
  It is the amount of made shots taken in the two-points area.

- *2 Points field goals percentage* (2PFG.Per).
  It is the ratio between the 2 points field goals and the 2 points field goals attempts.

- *Dunks made* (Sc).
  It is the amount of two points field goals in which the player also touched the rim with one or two hands.

- *3 Points field goal attempts* (3PFGA).

  It is the sum of made and missed shots taken out the two-points area.

- *3 Points field goals* (3PFGM).

  It is the amount of made shots taken out the two-points area.

- *3 Points field goals percentage* (3PFG.Per).

  It is the ratio between the 3 points field goals and the 3 points field goals attempts.

- *Free throw attempts* (FTA).

  It is the sum of made and missed free throws.

- *Free throw made* (FTM).

  It is the amount of made free throws.

- *Free throw percentage* (FT.Per).

  It is the ratio between the free throw made and free throw attempts.

- *Offensive rebounds* (ORB).

  It is the amount of rebounds grabbed in the opponent's two points area.

- *Defensive rebounds* (DRB).

  It is the amount of rebounds grabbed in the team's two points area.

- *Total rebounds* (TRB).

  It is the amount of rebounds grabbed.

- *Blocks made* (BLK).

  It is the amount of opponent's shots that the player stops with one or two hands.

- *Shots blocked* (BLKS).

  It is the amount of the player's shots that the opponent stops with one or two hands.

- *Turnovers* (TO).

  It is the number of times that the player surrenders the possession while having the ball.

- *Steals* (STL).

  It is the number of times that the player acquired the possession without the opponent shooting.

- *Assists* (AST).

  It is the amount of passes completed by the player that lead a teammate to a made shot.

- *Valutazione Lega* (VL).

  It is an index expressing the performance value; it is computed as a weighted sum in which it is assigned weight equal to 1 to positive actions, namely points scored, assists, steals, total rebounds, blocks made and fouls drawn, and value -1 is assigned to the negative actions, such as turnovers, shots blocked, fouls committed and missed shots.

- *Offensive Efficiency rating* (OER).

  It expresses the ratio between the points scored and the possessions played, where the estimate of the possession played by the player is equal to the sum of 2 and 3 points field goal attempts, turnovers and the number of free throw attempts divided by 2.

- *Plus/minus.*

  It indicates the difference of the margin, the points made by the team

Umana Reyer Venezia

Umana Reyer Venezia			Falli		Tiri da 2			Sc	Tiri da 3			Tiri Liberi			Rimbalzi			Stoppate		Palle		Ass	Valutaz.		+/-
AllDe Raffaele Walter	Pt	Min	C	S	R	T	%		R	T	%	R	T	%	Off	Dif	Tot	Dat	Sub	Per	Rec		Lega	OER	
* 0 Haynes MarQuez	15	35	0	3	0	2	0.0	0	4	8	50.0	3	4	75.0	1	2	3	0	1	3	0	3	13	1.00	13
* 2 Hagins Jamelle	16	23	4	4	7	9	77.8	1	0	0	0.0	2	2	100.0	2	5	7	0	0	2	0	0	19	1.33	24
3 Ejim Melvin	7	20	5	2	2	5	40.0	0	1	3	33.3	0	0	0.0	2	2	4	2	1	2	2	1	5	0.70	-7
* 4 Peric Hrvoje	11	26	3	4	4	13	30.8	1	0	2	0.0	3	4	75.0	3	3	6	1	1	2	0	2	6	0.58	2
* 6 Bramos Michael	5	31	2	3	0	1	0.0	0	0	3	0.0	5	6	83.3	0	2	2	1	0	1	0	2	5	0.63	5
7 Tonut Stefano	7	14	3	1	1	2	50.0	0	1	1	100.0	2	2	100.0	0	0	0	0	0	2	2	1	5	1.17	-9
9 Visconti Riccardo	0	0	0	0	0	0	0.0	0	0	0	0.0	0	0	0.0	0	0	0	0	0	0	0	0	0	0.00	0
12 Filloy Ariel	6	11	2	1	1	1	100.0	0	1	4	25.0	1	2	50.0	0	1	1	0	0	0	1	0	3	1.00	-17
14 Ress Tomas	1	6	1	2	0	0	0.0	0	0	0	0.0	1	2	50.0	1	0	1	0	0	0	0	0	2	1.00	-4
16 Ortner Benjamin	0	4	3	0	0	1	0.0	0	0	0	0.0	0	0	0.0	0	0	0	0	0	1	0	1	-4	0.00	-5
22 Viggiano Jeff	0	4	0	0	0	1	0.0	0	0	1	0.0	0	0	0.0	0	0	0	0	0	0	0	0	-2	0.00	4
* 25 McGee Tyrus	20	26	5	3	0	0	0.0	0	6	9	66.7	2	2	100.0	1	2	3	0	0	0	2	3	23	2.00	14
Squadra	0	0	1	0	0	0	0.0	0	0	0	0.0	0	0	0.0	2	6	8	0	0	0	0	0	7	0.00	
Totali	88	200	29	23	15	35	42.9	2	13	31	41.9	19	24	79.2	12	23	35	4	3	13	7	13	82	0.97	

Figure 3.2: Box score for Reyer Venezia team.

minus the points made by the opponent, when the player is on the court.

At the bottom of the table there are team statistics and totals. An action scored is assigned to the team when there is no clear indication about who performed that action. For example, if the opponent's shooter misses his shot and the ball bounces from hand to hand, eventually finishing in the hand of a team's player, that rebound is assigned to the team and not the final player. See Figure 3.2 for an example.

In this case, information about who the starters are is represented by the stars near the names. Is it also presented the name of the team's head coach.

Valutazione Lega, *Offensive efficiency rating* and *Plus/minus* can give a glance of the player's performance. We will explore strengths and weaknesses of these indexes (mainly weaknesses) later in this work. See Figure 3.3 for an example.

It is clear that this representation gives a static view of the performances. Box score can contain information about every quarter, about the whole game, about more games and can contain sums of all the statistics considered,

EA7 Emporio Armani Milano

EA7 Emporio Armani Milano				Falli		Tiri da 2			Sc	Tiri da 3			Tiri Liberi			Rimbalzi			Stoppate		Palle		Ass	Valutaz.		+/-
	All. Repesa Jasmin	Pt	Min	C	S	R	T	%		R	T	%	R	T	%	Off	Dif	Tot	Dat	Sub	Per	Rec		Lega	OER	
1	McLean Jamel	11	25	2	10	3	9	33.3	1	0	0	0.0	5	9	55.6	3	2	5	1	2	0	0	1	14	0.81	7
2	Fontecchio Simone	0	12	3	0	0	0	0.0	0	0	1	0.0	0	0	0.0	0	3	3	0	0	3	1	1	-2	0.00	1
9	Kalnietis Mantas	14	29	1	3	3	4	75.0	0	2	5	40.0	2	4	50.0	0	4	4	0	0	5	1	10	20	0.88	16
• 11	Raduljica Miroslav	9	15	4	5	2	6	33.3	1	0	0	0.0	5	6	83.3	3	0	3	0	0	1	0	0	7	0.90	-11
12	Dragic Zoran	14	26	3	4	4	6	66.7	0	2	2	100.0	0	1	0.0	1	4	5	0	1	2	2	1	17	1.33	18
• 14	Pascolo Davide	0	5	0	0	0	1	0.0	0	0	0	0.0	0	0	0.0	0	0	0	1	1	0	0	0	-1	0.00	-18
• 20	Cinciarini Andrea	0	14	2	0	0	0	0.0	0	0	0	0.0	0	0	0.0	0	1	1	0	0	1	0	1	-1	0.00	-17
21	Sanders Rakim	15	23	5	3	3	4	75.0	1	2	3	66.7	3	6	50.0	0	3	3	0	0	1	1	2	13	1.36	3
• 23	Abass Abass Awudu	10	23	2	1	3	4	75.0	0	1	4	25.0	1	1	100.0	2	2	4	1	0	0	0	2	12	1.18	-19
30	Cerella Bruno	0	0	0	0	0	0	0.0	0	0	0	0.0	0	0	0.0	0	0	0	0	0	0	0	0	0	0.00	0
• 43	Simon Krunoslav	11	28	5	1	1	3	33.3	0	3	7	42.9	0	0	0.0	0	2	2	0	0	3	1	1	2	0.85	0
	Squadra	0	0	0	0	0	0	0.0	0	0	0	0.0	0	0	0.0	1	4	5	0	0	0	0	0	5	0.00	
	Totali	84	200	27	27	19	37	51.4	3	10	22	45.5	16	27	59.3	10	25	35	3	4	16	6	19	86	0.95	

Figure 3.3: Box score for Emporio Armani Milano team.

means and so on.

The measures available, except for the Plus/minus, are exclusively bottom-up measures because they give credit to the last player who touches the ball. As we will analyse these measures in the next chapter, we will discuss positive and negative sides of assessing the most part of the contribution to the final player of the play.

One merit of the box score is manageability. This representation offers a dataset very easy to read and to extract information from. Shooting percentages are very useful during the game to let players know how are they shooting. Total team's production also help to highlight what are the keys of the good or bad performance against the opponent. When a team clearly plays better than the opponent, it shows an advantage in more statistics considered in the box score, such as rebounds and field goal percentage. When two teams are quite similar and the score is close, box score information does not show any particular difference. In addition, extracting insights about players from the box score is far more ambiguous from a static point of view of the game.

Let us consider that a player usually does not play the whole game, so it is useful to know which players he shares the court with at any time. In addition, basketball is a team sport in which a team needs to be a team in order to success. In this case, assessing contribution shares to every player becomes a non trivial question.

In order to overcome this problem, it is now shown another method a data tracking heavily employed in basketball; the play-by-play data. At the end of this chapter, we will present a way of combining information provided by both methods.

3.2 Play-by-play

Play-by-play data is a chronological history of the game. While the focus in the box score is on the player, in this form of data the main focus is on the action made. First, there is information about the teams' names and the minute of the game (in Italian men Basketball Serie A2, Euroleague and NBA there are also seconds recorded). Then, for every single action made on the court, it is recorded which action is and who performed it. When points are scored, the score is updated in the time column. See Figure 3.4 for an example.

The actions recorded include:

- *2 Points field goal missed in the paint* (2P.FG.Miss.1).
  A shot from the paint area that the player misses.

- *2 Points field goal made in the paint* (2P.FGM.1).
  A shot from the paint area that the player makes.

- *2 Points field goal missed out of the paint* (2P.FG.Miss.2).

Umana Reyer Venezia	Min.	EA7 Emporio Armani Milano
	1. min	
Peric Hrvoje - Tiro sbagliato da sotto		
Peric Hrvoje - Stoppata Subita (1)		
		Abass Abass Awudu - Stoppata (1)
McGee Tyrus - Canestro da 3 punti	3-0	
Hagins Jamelle - Fallo commesso (1)		
		Raduljica Miroslav - Fallo subito (1)
Hagins Jamelle - Canestro da sotto	5-0	
		Cinciarini Andrea - Assist (1)
	5-3	**Simon Krunoslav - Canestro da 3 punti**
Bramos Michael - Tiro sbagliato da 3 punti		
McGee Tyrus - Rimbalzo offensivo (1)		
Haynes MarQuez - Canestro da 3 punti	8-3	
		Simon Krunoslav - Palla persa (1)
		Cinciarini Andrea - Palla persa (1)
	2. min	
McGee Tyrus - Fallo subito (1)		
		Abass Abass Awudu - Fallo commesso (1)
McGee Tyrus - Tiro libero segnato	9-3	
McGee Tyrus - Tiro libero segnato	10-3	
	3. min	
McGee Tyrus - Palla recuperata (1)		

Figure 3.4: Play by play from Reyer Venezia - Emporio Armani Milano game.

A shot from outside the paint area that the player misses.

- *2 Points field goal made out of the paint* (2P.FGM.2).

  A shot from outside the paint area that the player makes.

- *3 Points field goal missed* (3P.FG.Miss).

  A shot from outside the two points area that the player misses.

- *3 Points field goal made* (3P.FGM).

  A shot from outside the two points area that the player makes.

- *Dunk made* (Sc).

  A made shots touching the rim with one or both hands.

- *Free throw missed* (FT.Miss).

  A free throw the player misses.

- *Free throw made* (FTM).

  A free throw the player makes.

- *Assist* (AST).

  A pass completed by the player which lead to a made shot by a teammate.

- *Offensive rebound* (ORB).

  A rebound grabbed in the opponent's two points area.

- *Defensive rebound* (DRB).

  A rebound grabbed in the team's two points area.

- *Turnover* (TO).

  The event that the player surrenders the possession while having the ball.

- *Steal* (STL).

  The event that the player acquired the possession without the opponent shooting.

- *Foul committed* (PF).

  A foul play the player committed during the game. In this case, players can be called for a foul even if they sit on the bench.

- *Foul drawn* (FD).

  A foul play the player was committed to.

- *Blocks made* (BLK).

  An opponent's shots that the player stops with one or two hands.

- *Shots blocked* (BLKS).

  A player's shots that the opponent stops with one or two hands.

With this type of data, it is possible to link an assist to a made basket. In this way, it is possible to further classify made shots as assisted and unassisted. Analysing assisted and unassisted shots separately is based on the fact if a shot is made after a pass, it has higher probability of finishing in the basket, since passing ideally creates space for the shooter.

Play-by-play data main advantage is that more information can be retrieved than from the box score. From a motion picture like play-by-play, compared to a static picture such as the box score, it is possible to know when the player is on the court, which are the line-ups competing and how the game is going in every of its phases. It is possible to create box scores from it, simply by counting how many single actions the player performs, referred to the whole game or to arbitrarily chosen time intervals, such as half times, single quarters or approximately the last 6 minutes of a close game, the so called *clutch time*.

On the other hand, the main problem with play-by-play is manageability. This data consist in text strings which include information. It is necessary to process them in order to create a box score, which can now contain more information than simple counts regarding actions. It is useful tough perform an ETL (*Extract, Transform, Load*) process on the strings by which it is possible to retrieve, transform and create a manageable dataset.

3.3 Combining the data

There are three main approaches by which we can retrieve information about a player's performance: looking at what he does on the court, looking at what the team does when he is on the court or outside, looking at what the opponent does when he is on the court or outside.

When looking at what a player does on the court, relying exclusively on the box score information, a biased picture is presented, the more the player is not involved in on-ball actions. Trying to identify and measures a player's contribution to his team victory needs to take into considerations this bias. Players who contribute to their team with off-the-ball movements, such as screens or cuts, create space which, as said earlier, is one of the two challenges in a game of basketball. This contribution, however, can be elicited, partially, from top-down measures, like Plus/minus. The next questions to answer are therefore:

- "How bad the opponent is when the player is on the court versus when is not?"
- "How good the team is when the player is on the court versus when is not?"

3.3.1 Extracting player's performance from opponent's production

Consider the first question. How it can happen that a line-up make the other one doing mistakes? They can prevent the opponent from shooting or make him shoot worse than usual, they can contain the opponent from getting available rebounds, defensively and offensively, they can make the opponent loose the ball more or they can prevent the opponent from stealing the ball, all things that has effect on possession and shooting performance.

All this considered, the most natural way to retrieve information about the player's contribution is to take into account the difference between opponent's percentages or sums statistics when the player is on the court and when he sits on the bench.

In this work, we consider 10 different net measures, namely:

- *Effective Field Goals net percentage* (eFG% net).
  The difference between opponent's Effective field goals percentage when the player is on or off the court. Effective field goal percentage is a summary statistics which take into account the higher points value of a three points shot to calculate an adjusted shooting percentage. Effective field goal net percentage is defined as

$$\text{eFG.Per.Net} = \frac{\text{FGM}_{on} + (0.5 \times \text{3PFGM}_{on})}{\text{FGA}_{on}} - \frac{\text{FGM}_{off} + (0.5 \times \text{3PFGM}_{off})}{\text{FGA}_{off}}.$$

- *Offensive rebounds net percentage.*
  It is the net value of Offensive rebound percentage. This statistics consider the ratio of the offensive rebouns grabbed by the opponent on all the available offensive rebounds; Offensive rebound net percentage is defined as

$$\text{ORB.Per.Net} = \frac{\text{Opp.ORB}_{on}}{\text{Opp.ORB}_{on} + \text{Team.DRB}_{on}} - \frac{\text{Opp.ORB}_{off}}{\text{Opp.ORB}_{off} + \text{Team.DRB}_{off}}.$$

- *Defensive rebounds net percentage.*
  Conversely, it is possible to built a net measure representing the difference in defensive rebounds performance by the opponent, i.e.,

$$\text{DRB.Per.Net} = \frac{\text{Opp.DRB}_{on}}{\text{Opp.DRB}_{on} + \text{Team.ORB}_{on}} - \frac{\text{Opp.DRB}_{off}}{\text{Opp.DRB}_{off} + \text{Team.ORB}_{off}}.$$

- *Turnover net percentage.*
  It is the net value between turnover percentages when the player is on or off the court. Turnover percentage indicates the fraction of the opponent's possessions that results into a lost ball. Turnover net percentage is defined as

$$\text{TO.Per.Net} = \frac{\text{Opp.TO}_{on}}{\text{Opp.Poss}_{on}} - \frac{\text{Opp.TO}_{off}}{\text{Opp.Poss}_{off}}.$$

- *Steal net percentage.*
  It is the net value between steal percentages when the player is on or off the court. Steal percentage indicates the fraction of the team's

possessions that results into a stolen ball by the opponent. Steal net percentage is defined as

$$\text{STL.Per.Net} = \frac{\text{Opp.STL}_{on}}{\text{Team.Poss}_{on}} - \frac{\text{Opp.STL}_{off}}{\text{Team.Poss}_{off}}.$$

- *2 Points field goal net percentage, from the post.*

  It indicates the difference between 2 Points field goal percentages from the post; it is defined as

$$\text{2P.FGM.1.Per.Net} = \frac{\text{Opp.2P.FGM.1}_{on}}{\text{Opp.2P.FGA.1}_{on}} - \frac{\text{Opp.2P.FGM.1}_{off}}{\text{Opp.2P.FGA.1}_{off}}.$$

- *2 Points field goal net percentage, from outside.*

  It indicates the difference between 2 Points field goal percentages from outside; it is defined as

$$\text{2P.FGM.2.Per.Net} = \frac{\text{Opp.2P.FGM.2}_{on}}{\text{Opp.2P.FGA.2}_{on}} - \frac{\text{Opp.2P.FGM.2}_{off}}{\text{Opp.2P.FGA.2}_{off}}.$$

- *3 Points field goal net percentage.*

  It indicates the difference between 3 Points field goal percentages; it is defined as

$$\text{3P.FGM.Per.Net} = \frac{\text{Opp.3P.FGM}_{on}}{\text{Opp.3P.FGA}_{on}} - \frac{\text{Opp.3P.FGM}_{off}}{\text{Opp.3P.FGA}_{off}}.$$

- *True shooting net percentage.*

  The difference between True Shooting percentage (TS%) when the player is on or off the court. True Shooting percentage is a summary statistics that consider all types of shot, 2 points, 3 points field goals and free throws and tries to combine them into a single adjusted percentage. It is defined as

$$\text{TS.Per.Net} = \frac{\text{Opp.PTS}_{on}}{2(\text{Opp.FGA}_{on} + (0.44 \times \text{Opp.FTA}_{on}))} - \frac{\text{Opp.PTS}_{off}}{2(\text{Opp.FGA}_{off} + (0.44 \times \text{Opp.FTA}_{off}))}.$$

- *Assisted field goals net percentage.*

  This statistic indicates the difference between the ratios of the opponent's assisted field goals on total field goals when player is on or off the court. It is defined as:

$$\text{Ass.FGM.Per.Net} = \frac{\text{Opp.Ass.FGM}_{on}}{\text{Opp.FGM}_{on}} - \frac{\text{Opp.Ass.FGM}_{off}}{\text{Opp.FGM}_{off}}.$$

These measures are based on top-down measures. This means that they look at the production of the whole line-up. As mentioned earlier, these measures are biased towards players who play with good teammates, because they do not discriminate between individual and team effort. By combining both bottom-up and top-down measures, is it possible to mitigate the downsides of these two kind of measures.

Since we are dealing with team statistics, we could include in this new "Opponent's production box score" also data regarding differences of count statistics, as points, rebounds, assists, steals, turnovers and fouls. Since the data are based on individual's production, we can build two different opponent's box score, one for the opponent's performance when the player is on the court and the other one regarding opponent's production when the player is on the bench.

The new variables available with this reasoning are:

- Opponent's 2 Points field goals, from the post (Opp.2P.FGM.1);
- Opponent's 2 Points field goal attempts, from the post (Opp.2P.FGA.1);
- Opponent's 2 Points field goals percentage, from the post (Opp.2P.FGM.1.Per);
- Opponent's 2 Points field goals, from the outside (Opp.2P.FGM.2);
- Opponent's 2 Points field goal attempts, from outside (Opp.2P.FGA.2);

- Opponent's 2 Points field goals percentage, from outside (Opp.2P.FGM.2.Per);
- Opponent's 3 Points field goals (Opp.3P.FGM);
- Opponent's 3 Points field goal attempts (Opp.3P.FGA);
- Opponent's 3 Points field goals percentage (Opp.3P.FGM.Per);
- Opponent's free throws made (Opp.FTM);
- Opponent's free throw attempts (Opp.FTA);
- Opponent's free throws percentage (Opp.FT.Per);
- Opponent's offensive rebounds (Opp.ORB);
- Opponent's defensive rebounds (Opp.DRB);
- Opponent's steals (Opp.STL);
- Opponent's turnovers (Opp.TO);
- Opponent's fouls committed (Opp.PF);
- Opponent's assists (Opp.AST).

These data can be manipulated into net measures, representing the difference in opponent's production when the player in on the court versus the case when he is not playing or, along with information abut the player's team production, in plus/minus measures, which consider both team's production when the player is on the court.

3.3.2 Extracting player's performance from team's production

Conversely, it is possible to retrieve information about player's performance by looking at what the line-up does when the player is in of out. It is possible to calculate the same indexes as for the opponent's case with regard to the team. Every definition is equal as the above case, but team statistics are switched with opponent's ones. To this, the differences of the main box score variables are computed.

In this way, other two "Team's production box score" are available, one with *on measures* and the other one with *off measures*.

3.3.3 Levelling players performances

As explained in the previous chapter, counts data about player who play a significantly different amount of time are difficult to compare. In order to standardise counts statistics, from play-by-play data, it is possible to estimate the number of possessions one the court for the player and the opponent, as the number of possessions employed by the teams facing when the player is not playing.

Following the definition of the possessions estimate, we can now rely of four different estimates for possessions regarding the player, namely *Possession on*, *Possession off*, *Opponent's Possession on*, *Opponent's Possession off*.

Of course, the number of possessions employed by the teams in the two cases, are roughly the same. On the other hand, by using standardised measures, problems relative to the different scales are avoided.

Chapter 4

Ranking method for sport data analysis

Keeping in mind the aim of measuring an individuals' contribution to the game, let us now introduce some new features to the theoretical framework. As already seen previously, the main tool employed is the linear model. When estimating player's performance measures, his performance needs to be included into the game context. Players have teammates and opponent's; moreover, in most of the major sports, including basketball, teams face each other multiple times during the season, the same number of games played at home and away. Role played also gives information about the performance, since every role has its specific characteristics. It need to be taken into account that a player can only play for one team, at the time, and can only have one role. This means that his performance is strongly linked to his team, the opponent, where he is playing and what role he is playing for his team. Every team has at least one player per role (usually two or more), so every player can be thought as nested within his role, which is nested within the team. With this framework in mind, it is possible to assign different value to contributions made by good players who play on different levels team. The

contribution of a good player to a bad team can be considered more valuable than the one of a good player to a good team. Of course the same consideration can be made for player who play different role, since the things they are required to do on the court are different.

By modelling performance measures as fixed effect linear models, it is possible to do inference about the player's net strength and weakness in that particular performance. Linear models output allows also to test whether one player is significantly better than another one, regarding that particular skills. In order to have a complete picture which consider the most possible information, a viable option is to merge all the result from single hypothesis tested into a general one which answer to the question "Who is the best?".

4.1 Round robin design

The round robin design is a setting in which there are C agents, such as teams, players or other groups, who are n times pairwise connected by generating two outcomes that are dependent by the feature of that given pair of subjects. This type of network are called dyads. The most known tournament design for team sport has always been round robin design. In this framework there are competitors which challenge each other, at least one time. Usually, as the case of Serie A1 Italian men Basketball league, tournaments are organized as a double round robin design, in which every team faces each other two times, one time is its own home court and the other one in the opponent's home court. In this case, round robin is said to be balanced. However, round robin does not only refers to sports. Psychologists who investigates social network and interactions models widely employ round robin design based models.

This type of models are specifically designed to analyse comparison data. In addition, there is a considerable existing literature on modelling the scores of two opposing teams for data sport analytics purposes. Regarding teams' performance and comparison, there have been suggested to model differences in score between two teams and the result of teams effects and home court advantage. With a modification of the least squares estimation method, it is possible to build a ranking of the teams within the league. An alternative approach, which consists in modelling teams' outcome separately has also been employed, applying a bivariate Poisson distribution, introducing a dependence parameter for the goals scored by the opposing teams. An extension of this model, consisting in modelling scores as the joint distribution of Possion ones, representing the total number of goals and a binomial distribution, representing the goals of one team given the total number of goals. Focusing on the probability of winning, as in the case of *Wins produced*, a whole lot of literature has developed, employing the Bradley-Terry model. In order to make pair comparison, data sport analytics has seen a wide employment of linear models for paired comparisons, the Bradley-Terry model and the Thurstone-Mosteller model.

Modelling performances as fixed effect linear models allow us to draw pair comparison between each pair of players in the league. Then combining every pair comparison, it is possible to build a ranking among all the players. Different subgroups of players can be considered, such as all the players in a team who play the in the same role, or all the players in a team, or all the players who plays a certain role or it is possible to consider all the players in the league. This perspective is slightly different from the logic of *Wins produced*, *P.E.R.* and *Approximate Value*, since all these measures gives a value to each statistics, without regard to players. The goal, however, is the

same, because both methods answer to the same question: "How much a player contributes to his team?".

4.2 Testing hypothesis and ranking on round robin design for Data Sports analytics

Let us now consider a round robin designed tournament, in which there are $\Pi_1, ..., \Pi_C$ competitors, the teams, which challenge each with one another, so competitors j and h meet n_{jh} times, $j \neq h$. In the case of Serie A1 Italian men Basketball league, $n_{jh} = 2, \forall j, h = 1, ..., C, j \neq h$, therefore the design is balanced.

The output is in a form of scalar or, more often, a vector for each pairwise comparison. For each player in the league, we can see the output, during the whole season, as longitudinal multivariate observations of performance indicators. Borrowing terminology from social relations models employed for round robin design, when teams j and h meet on the $i-th$ occasion, for every pair of players we obtain a pair of observations $\boldsymbol{y}_{il(j)}$ and $\boldsymbol{y}_{il'(j)}$ as a realization of a p-variate random variable. In this case, $\boldsymbol{y}_{il(j)}$ represent the response of player l, within team j as an *actor* towards teammate l', on the $i-th$ occasion against team j. Of course, for $\boldsymbol{y}_{il'(j)}$ roles are reversed.

Without loss of generality, let us assume that for each ${}_k y_{il(j)}$ univariate component, $k = 1, ..., p$, a higher value of y means that the performance is better. In this study, we examined three different performances domains; the first one considered involves all the four types of shot:

- 2 Points field goals from the post (2P.FGM.1.p);
- 2 Points field goals from outside (2P.FGM.2.p);

- 3 Points field goals (3P.FGM.p);
- Free throw made (FTM.p).

The second domain considered consist in the two different types of rebounds:

- Offensive Rebounds (ORB.p);
- Defensive rebounds (DRB.p).

The last one of which this analysis is focused tries to take into account the three main statistics which describe better the individual's offensive production:

- Points (PTS.p);
- Offensive rebounds (ORB.p);
- Assists (AST.p).

In this way it is possible to have a strong suggestion of which player in the team contributes most on the offensive end of the court.

For example, ${}_k y_{il(j)}$, in the first case, might represent the 2 Points field goals made by player l, of team j, against team h on the $i-th$ occasion.

The interaction between each pair of players l and l', nested within one mutually exclusive team and one mutually exclusive role, playing on team j against opponent h on the $i-th$ occasion produces a pair of outcome Y, we can model as:

$$ {}_k Y_{il(j)} =_k \mu +_k \tau_j + \mathbf{x}_{ij} \cdot \boldsymbol{\beta} +_k \mu_{l(j)} + \varepsilon_{ilj}, \tag{4.1} $$

where ${}_k \mu_{l(j)} =_k \tau_{l(j)} +_k \mathbf{z}_{ilj} \cdot_k \boldsymbol{\gamma}$, and

$$ {}_kY_{il'(j)} =_k \mu +_k \tau_j + \mathbf{x}_{ij} \cdot \boldsymbol{\beta} +_k \mu_{l'(j)} + \varepsilon_{il'j}, \tag{4.2} $$

where ${}_k\mu_{l'(j)} =_k \tau_{l'(j)} +_k \mathbf{z}_{il'j} \cdot_k \boldsymbol{\gamma}$, with $j, h = 1, \ldots, C, j \neq h$, $i = 1, \ldots, n_{jh}$, $l, l' = 1, \ldots, s_j, s_h$, $k = 1, \ldots, p$, and $\varepsilon_{ilj} \sim IID(0; \sigma^2(\tau_{l(j)}))$.

In particular, μ is the league-related global mean across all the competitors, τ_j is the team-related mean, β includes home-away effect when team j and h meet, $\tau_{l(j)}$ refers to the effect of the player and ε are p-variate $IID(0, \sigma^2(\tau_{l(j)}))$ random errors; in this way we allow parameters for the variance of every player to be different from each other. In addition, γ represent coefficients for every possible covariate effect that should not be referred to any related-performance information but to different performance indicators, such as those referred to the whole team or opponent's production or metrics referring to action that do not belong to the domain considered. In order to highlight the net effect between the two players l and l', of team j, facing team h on the $i-th$ occasion, let us consider the net performance:

$$ \begin{aligned} \Delta_k Y_{il(j)l'(j)} &= ({}_k\tau_{l(j)} -_k \tau_{l'(j)}) + (({}_k\boldsymbol{z}_{ilj} \cdot_k \boldsymbol{\gamma} + \varepsilon_{il(j)}) - ({}_k\boldsymbol{z}_{il'(j)} \cdot_k \boldsymbol{\gamma} + \varepsilon_{il'(j)})) \\ &= \tau_{il(j)} - \tau_{il'(j)} + \boldsymbol{\gamma}\Delta\boldsymbol{z}_{il(j)l'(j)} + \boldsymbol{u}_{il(j)l'(j)} \end{aligned} \tag{4.3} $$

In this way, design effect are excluded and $\tau_{il(j)}$ represent the total ability of player l, of team j, facing opponent h at the $i-th$ occasion; $\boldsymbol{u}$ are p-variate $IID(0, \sigma^2(\tau_{l(j)}))$ random errors. Note that the expressions are a fixed effects multivariate multi-way ANOVA model, in which each univariate component can be expressed as a linear combination of design effect and other covariates effects.

We are now concerned with inference about net ability between players l and l' within team j. Considering $\tau_{l(j)}$, we can test parameter significance

and test $\tau_{l(j)}$ and $\tau_{l'(j)}$ against each other. The set of hypothesis of interest for significance testing can be expressed as:

$$
\begin{aligned}
H_{0l(j)} : \bigcap_{i=1}^{2} \boldsymbol{\tau}_{il(j)} = \mathbf{0} &\equiv \bigcap_{i=1}^{2} \bigcap_{k=1}^{p} {}_k\tau_{il(j)} = 0 \equiv \bigcap_{i=1}^{2} \bigcap_{k=1}^{p} {}_kH_{0i(l(j))} \\
H_{1l(j)} : \bigcup_{i=1}^{2} \boldsymbol{\tau}_{il(j)} \neq \mathbf{0} &\equiv \bigcup_{i=1}^{2} \bigcup_{k=1}^{p} {}_k\tau_{il(j)} \neq 0 \equiv \bigcup_{i=1}^{2} \left[{}_kH^{-}_{1il(j)} \cup {}_kH^{+}_{1il(j)} \right] \\
&\equiv \bigcup_{i=1}^{2} \left[(\boldsymbol{\tau}_{il(j)} < 0) \cup (\boldsymbol{\tau}_{il(j)} > 0) \right] \\
&\equiv \bigcup_{i=1}^{2} \bigcup_{k=1}^{p} \left[({}_k\tau_{il(j)} < 0) \cup ({}_k\tau_{il(j)} > 0) \right] \\
&\equiv \bigcup_{i=1}^{2} \bigcup_{k=1}^{p} \left[{}_kH^{-}_{1il(j)} \cup {}_kH^{+}_{1il(j)} \right],
\end{aligned}
$$

where $l(j) = 1, \ldots, n_j$. On the other hand, the set of hypothesis for pairwise comparison can be expressed as:

$$
\begin{aligned}
H_{0(l(j)l'(j))} : \bigcap_{i=1}^{2} \boldsymbol{\tau}_{il(j)} = \boldsymbol{\tau}_{il'(j)} &\equiv \bigcap_{i=1}^{2} \bigcap_{k=1}^{p} {}_k\tau_{il(j)} = {}_k\tau_{il'(j)} \equiv \bigcap_{i=1}^{2} \bigcap_{k=1}^{p} {}_kH_{0i(l(j)l'(j))} \\
H_{1(l(j)l'(j))} : \bigcup_{i=1}^{2} \boldsymbol{\tau}_{il(j)} \neq \boldsymbol{\tau}_{il'(j)} &\equiv \bigcup_{i=1}^{2} \bigcup_{k=1}^{p} {}_k\tau_{il(j)} \neq {}_k\tau_{il'(j)} \\
&\equiv \bigcup_{i=1}^{2} \left[{}_kH^{-}_{1i(l(j)l'(j))} \cup {}_kH^{+}_{1i(l(j)l'(j))} \right] \\
&\equiv \bigcup_{i=1}^{2} \left[(\boldsymbol{\tau}_{il(j)} < \boldsymbol{\tau}_{il'(j)}) \cup (\boldsymbol{\tau}_{il(j)} > \boldsymbol{\tau}_{il'(j)}) \right] \\
&\equiv \bigcup_{i=1}^{2} \bigcup_{k=1}^{p} \left[({}_k\tau_{il(j)} < {}_k\tau_{il'(j)}) \cup ({}_k\tau_{il(j)} > {}_k\tau_{il'(j)}) \right] \\
&\equiv \bigcup_{i=1}^{2} \bigcup_{k=1}^{p} \left[{}_kH^{-}_{1i(l(j)l'(j))} \cup {}_kH^{+}_{1i(l(j)l'(j))} \right],
\end{aligned}
$$

where $l(j) = 1, \ldots, n_j$.

The Union-Intersection Roy's principle allow us to express null and alternative hypothesis as described. In this way, we are able to take into account

possible interaction between teams and home/away effect, teams and roles and players and teams. Decomposing the multivariate hypothesis into k univariate ones let us model every performance measure separately and then combining the univariate p-values, which are not independent, into multivariate tests.

Regarding the alternative hypothesis, it is worth noting that the two separated hypothesis highlight in which direction the possible difference actually takes place. Under this framework, under the alternative hypothesis al least one of the possible one-sided directions must be true in order to reject the null hypothesis. On the other hand, it is possible, regarding players comparison, that for some univariate performance player $l(j)$ is above the level of player $l'(j)$ and vice versa, which would be the case of $H^{-}_{1l(j)}$ and $H^{+}_{1l(j)}$ be jointly true (as well as $H^{-}_{1(l(j)l'(j))}$ and $H^{+}_{1(l(j)l'(j))}$).

By exploiting the multivariate one-sided alternatives in the latter expression, a ranking can be constructed among all the players in the league, or in a team, or within the same role within the same team. By suitable combining information from directional multivariate p-values, the underline possible latent ordering among τs parameters can be properly estimated. The rationale behind this ranking method within a multivariate setting is the following: if not all $H_{0l(j)l'(j)}$ are true, it must exist an ordering $[1], [2], ..., [n_j]$ among τs, such that:

$$\boldsymbol{\tau}_{[1]} \leq \boldsymbol{\tau}_{[2]} \leq ... \leq \boldsymbol{\tau}_{[n_j]}$$

We are able to say that when $\tau_{l(j)} < \tau_{l(j)}$ if there exists at least one univariate ${}_k\tau_{l(j)} < {}_k\tau_{l(j)}$ and at the same time there is not any univariate p-value for which the opposite inequality holds. If the last condition is not met, the two players are ranked at the same level. It is also stated that, in a multivariate setting, the parameters are not tied not only when all univariate

p-values are equal, but also when $H^{-}_{1(l(j)l'(j))}$ and $H^{-}_{1(l(j)l'(j))}$ are jointly true.

Let ${}_kp^{-}_{il(j)}$ and ${}_kp^{-}_{il(j)}$, ${}_kp^{-}_{il(j)l'(j)}$ and ${}_kp^{-}_{il(j)l'(j)}$, the univariate p-values employed to infer on the possible equality versus directional alternatives presented in the expressions above; assuming normality on the p-variate random errors (i.e. $\varepsilon \sim IIN(0, \sigma^2(\tau_{l(j)}))$) and in order to test the two different hypothesis

- $H_{0l(j)}$ versus $H^{-}_{1l(j)}$ or $H^{+}_{1l(j)}$, and
- $H_{0l(j)l'(j)}$ versus $H^{-}_{1l(j)l'(j)}$ or $H^{+}_{1l(j)l'(j)}$,

a set of suitable univariate test statistics can be expressed as

$$ {}_kt_{il(j)} = \frac{{}_k\hat{\tau}_{il(j)}}{se({}_k\hat{\tau}_{il(j)})} \sim t_{gdl(error)} \tag{4.4} $$

and

$$ \begin{aligned} {}_kt_{il(j)l'(j)} &= \frac{{}_k\hat{\tau}_{il(j)} - {}_k\hat{\tau}_{il'(j)}}{se({}_k\hat{\tau}_{il(j)}) + se({}_k\hat{\tau}_{il'(j)}) - 2cov({}_k\hat{\tau}_{il(j)}, {}_k\hat{\tau}_{il'(j)})} \\ &\dot{\sim} t_{gdl(error)-2}, \end{aligned} \tag{4.5} $$

where $gdl(error)$ are the error's degrees of freedom in the linear model.

Regarding multivariate p-values $p^{-}_{il(j)}$ and $p^{+}_{il(j)}$ for testing $H_{0l(j)}$ *vs* $H^{-}_{1l(j)}$ or $H^{+}_{1l(j)}$, equivalently $p^{-}_{il(j)l'(j)}$ and $p^{+}_{il(j)l'(j)}$ for testing $H_{0l(j)l'(j)}$ *vs* $H^{-}_{1l(j)l'(j)}$ or $H^{+}_{1l(j)l'(j)}$, it has been suggested in the work of Arboretti et al. (2014) a combination procedure, assuming normality regarding the p-values random errors, which employs an empirical adaptation of Brown's Method for dependent p-values which is appropriate for high correlated data, such as sports data.

This solution needs to be viewed as approximated and its behaviour and the robustness against normality assumption under finite samples evaluated via simulation study.

Let ${}_kp^-_{(l(j),l'(j))}$ and ${}_kp^+_{(l(j),l'(j))}$ be the multivariate directional p-values related respectively to alternative hypothesis ${}_kH^-_{1l(j)l'(j)} : {}_k\tau_{l(j)} < {}_k\tau_{l'(j)}$ and ${}_kH^+_{1l(j)l'(j)} : {}_k\tau_{l(j)} > {}_k\tau_{l'(j)}$. Since, by definition,

$$ {}_kp^+_{l(j)l'(j)} = 1 - {}_kp^-_{(l(j),l'(j))} = {}_kp^-_{(l'(j),l(j))}, $$

all one sided inferential result can be expressed as:

$$ P^+ = \left[\begin{bmatrix} - & {}_1p^+_{(1,2)} & {}_1p^+_{(1,3)} & \cdots & {}_1p^+_{(1,n_j)} \\ {}_1p^+_{(2,1)} & - & {}_1p^+_{(2,3)} & \cdots & {}_1p^+_{(2,n_j)} \\ \cdots & \cdots & - & \cdots & \cdots \\ {}_1p^+_{((n_j-1),1)} & {}_1p^+_{((n_j-1),2)} & \cdots & - & {}_1p^+_{((n_j-1),n_j)} \\ {}_1p^+_{(n_j,1)} & {}_1p^+_{(n_j,2)} & \cdots & {}_1p^+_{(n_j,(n_j-1))} & - \end{bmatrix} \right. $$
$$ \left. , \cdots , \right. $$
$$ \left[\begin{bmatrix} - & {}_pp^+_{(1,2)} & {}_pp^+_{(1,3)} & \cdots & {}_pp^+_{(1,n_j)} \\ {}_pp^+_{(2,1)} & - & {}_pp^+_{(2,3)} & \cdots & {}_pp^+_{(2,n_j)} \\ \cdots & \cdots & - & \cdots & \cdots \\ {}_pp^+_{((n_j-1),1)} & {}_pp^+_{((n_j-1),2)} & \cdots & - & {}_pp^+_{((n_j-1),n_j)} \\ {}_pp^+_{(n_j,1)} & {}_pp^+_{(n_j,2)} & \cdots & {}_pp^+_{(n_j,(n_j-1))} & - \end{bmatrix} \right] $$

Let $p^+_{(l(j),l'(j))}$ the directional p-value statistics related to the alternative hypothesis $H^-_{1l(j)l'(j)} : \tau_{l(j)} < \tau_{l'(j)}$ and $H^+_{1l(j)l'(j)} : \tau_{l(j)} > \tau_{l'(j)}$ respectively. All of the $(n_j \times (n_j - 1))$ $p^+_{l(j)l'(j)}$ can be expressed as:

$$ P^+_\bullet = \begin{bmatrix} - & p^+_{(1,2)} & p^+_{(1,3)} & \cdots & p^+_{(1,n_j)} \\ p^+_{(2,1)} & - & p^+_{(2,3)} & \cdots & p^+_{(2,n_j)} \\ \cdots & \cdots & - & \cdots & \cdots \\ p^+_{((n_j-1),1)} & p^+_{((n_j-1),2)} & \cdots & - & p^+_{((n_j-1),n_j)} \\ p^+_{(n_j,1)} & p^+_{(n_j,2)} & \cdots & p^+_{(n_j,(n_j-1))} & - \end{bmatrix} $$

Note that p-value statistics in this expression indicates if and which of the two alternative hypothesis is verified; we need also to recall that, when dealing with multivariate hypothesis, can happen that both of the alternative hypothesis is verified, meaning that the ranking level of the two players is the same. It is also worth noting that what happen for univariate directional p-values does not happen when dealing with multivariate hypothesis, i.e. $p^{+}_{(l(j),l'(j))} \neq 1 - p^{+}_{(l(j),l'(j))}$.

Now let α be the chosen significance α-level and let S be the $n_j \times n_j$ matrix which transforms the adjusted (by multiplicity) p-values $p^{+}_{(l(j),l'(j))adj}$ into 0-1 scores where each element $s_{(l(j),l'(j))}$ take values 0 is $p^{+}_{(l(j),l'(j))adj} > \tau/2$ and 1 if $p^{+}_{(l(j),l'(j))adj} \leq \tau/2$, i.e.:

$$S = \begin{bmatrix} - & s_{(1,2)} & s_{(1,3)} & \cdots & s_{(1,n_j)} \\ s_{(2,1)} & - & s_{(2,3)} & \cdots & s_{(2,n_j)} \\ \cdots & \cdots & - & \cdots & \cdots \\ s_{((n_j-1),1)} & s_{((n_j-1),2)} & \cdots & - & s_{((n_j-1),n_j)} \\ s_{(n_j,1)} & s_{(n_j,2)} & \cdots & s_{(n_j,(n_j-1))} & - \end{bmatrix}$$

S can be viewed as a more synthetic representation of results from all multivariate directional pairwise comparisons suitable for testing all the possible pairwise inequalities. If we consider the sum of the $s_{(l(j),l'(j))}$ scores along the $l(j) - th$ row or the $l'(j) - th$ column, then we are respectively counting the players who, at that particular level of significance, are considered to be worse or better. Hence, we are able to define the estimate $\hat{r}(l(j))$ and $\hat{r}(l'(j))$ of the rank $r(l(j))$ and $r(l'(j))$, i.e. the ordering of each player compared with all the others players considered by referring to the ranking definitions:

$$\hat{r}(l'(j))^{D} = 1 + \sum_{l(j)=1}^{n_j} s_{(l(j),l'(j))}, \quad l(j) \neq l'(j), l'(j) = 1, \ldots, n_j \tag{4.6}$$

$$\hat{r}(l(j))^U = 1 + \left\{ \# \left[\left(n_j - \sum_{l'(j)=1}^{n_j} s_{(l(j),l'(j))} \right) > \left(n_j - \sum_{l'(j)=1}^{n_j} s_{(l''(j),l'(j))} \right) \right], \right.$$
$$\left. l''(j) = 1, \ldots, n_j, l(j) \neq l''(j) \right\}, \tag{4.7}$$

with $l(j) = 1, \ldots, n_j$. Here, D and U stand for downward and upward rank estimates respectively. The ranking estimators defined above are deriving by counting, on the basis of empirical evidence, of how many players are significantly better/worse than $l(j)-th/l'(j)-th$ players at the chosen significance α-level. The two estimates are intentionally denoted with a different notation in order to highlight that sometimes they could provide different rank estimates for the same player because of the intransitivity issue.

4.2.1 Properties of the ranking estimator

Since we are dealing with inference about the players, the ranking estimator $\bar{r} = \left\{ \bar{r}_1, \bar{r}_2, ..\bar{r}_{n_j} \right\}$ of the true ranking $r = \left\{ r_1, r_2, ..r_{n_j} \right\}$ related to the n_j players is affected both by type I and type II errors. The estimator is afflicted also by type III error, which occurs when one accepts a specific directional alternative hypothesis when in fact the other alternative hypothesis is true. In this case, when a given players takes a false-better/false-worse ranking than the ranking of a worse/better player this means that type III error occurs.

Let the *Correct Global Ranking* (CGR) and the *Correct Individual Ranking* ($CIR_{l(j)}$) be the event which occur when $\bar{r} \equiv r$ and $\bar{r}_{l(j)} \equiv r_{l(j)}$ that is when the ranking estimator

$$\bar{r}_{l(j)} = 1 + \left\{ \sharp(\hat{r}^U_{l(j)} + \hat{r}^D_{l(j)})/2 > (\hat{r}^U_{l'(j)} + \hat{r}^D_{l'(j)})/2, l'(j) = 1, ..., n_j, l'(j) \neq l(j) \right\}, \tag{4.8}$$

with $l(j) = 1, \ldots, n_j$, jointly and singularly correctly estimates the ranks for all and for the $l(j) - th$ player respectively.

It can be proved that the ranking estimator presented above satisfies the following properties:

$$
\begin{aligned}
&(1) \quad \Pr\{CGR|Homogeneity\} = \Pr\left\{\bar{r}_{l(j)} = r_{l(j)} = 1, \forall l(j)|Homogeneity\right\} \\
&\qquad\qquad = 1 - \alpha. \\
&(2) \quad \Pr\left\{CIR_{l(j)}|Homogeneity\right\} = \Pr\left\{\bar{r}_{l(j)} = r_{l(j)} = 1|Homogeneity\right\} \\
&\qquad\qquad \geq 1 - \alpha^*_{l(j)} \\
&\qquad l(j) = 1, \ldots, n_j. \\
&(3) \quad \text{if } \alpha_{l(j)} > \alpha_{l'(j)}, \text{ then} \\
&\qquad \Pr\left\{\bar{r}_{l(j)} < \bar{r}_{l'(j)}|non - homogeneity\right\} > \alpha^*,\ l(j), l'(j) = 1, \ldots, n_j, \\
&\qquad l(j) \neq l'(j). \\
&(4) \quad \lim_{n\to\infty} \Pr\{CGR|homogeneity\} = \lim_{n\to\infty} \Pr\{CIR|non - homogeneity\} \\
&\qquad\qquad = 1.
\end{aligned}
$$

Here α and α^*_j are respectively the chosen significance α-level and the resulting adjusted individual α-level in all the pairwise comparison involving the $l(j) - th$ player, $n = \min_j(n_j)$ and *homogeneity* and *non-homogeneity* refer to the situation in which we assume that the null hypothesis of equality of all players is true or false, respectively.

The second property means that the probability of rejecting a false pairwise null hypothesis is greater than the adjusted individual α-level, which means in turn that the combined test is an unbiased test; the third property means that the combined hypothesis are consistent so that all the sample sizes increase the probability that the estimated ranking matches the true one, that is to reject either one and all false null hypothesis converges to one.

4.3 A dynamic approach to testing and ranking on round robin design for Data Sports analytics

Once we proceed to build a ranking based on all the data in the season, we are concerned on a more dynamic and update available way to present the results. In this way, game by game, we are able to characterize team and individuals' performances in order to highlight how a player contributes to his team in each period of the season. In order express the new approach, we need to rethink the hypothesis considered to estimate univariate a multivariate p-values, regarding significance testing and pairwise comparison.

The set of hypothesis of interest for significance testing can be now expressed as:

$$
\begin{aligned}
H_{0tl(j)} : \bigcap_{t=1}^{T}\bigcap_{i=1}^{2} \boldsymbol{\tau}_{til(j)} = \mathbf{0} &\equiv \bigcap_{t=1}^{T}\bigcap_{i=1}^{2}\bigcap_{k=1}^{p} {}_k\tau_{til(j)} = 0 \\
&\equiv \bigcap_{t=1}^{T}\bigcap_{i=1}^{2}\bigcap_{k=1}^{p} {}_kH_{0ti(l(j))} \\
H_{1tl(j)} : \bigcup_{t=1}^{T}\bigcup_{i=1}^{2} \boldsymbol{\tau}_{til(j)} \neq \mathbf{0} &\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2}\bigcup_{k=1}^{p} {}_k\tau_{til(j)} \neq 0 \\
&\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2} \left[{}_kH^{-}_{1til(j)} \cup {}_kH^{+}_{1til(j)}\right] \\
&\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2} \left[(\boldsymbol{\tau}_{til(j)} < 0) \cup (\boldsymbol{\tau}_{til(j)} > 0)\right] \\
&\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2}\bigcup_{k=1}^{p} \left[({}_k\tau_{til(j)} < 0) \cup ({}_k\tau_{til(j)} > 0)\right] \\
&\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2}\bigcup_{k=1}^{p} \left[{}_kH^{-}_{1til(j)} \cup {}_kH^{+}_{1til(j)}\right],
\end{aligned}
$$

with $l(j) = 1, \ldots, n_j$, whereas the set of hypothesis for pairwise comparison

can be expressed as

$$
\begin{aligned}
H_{0t(l(j)l'(j))} : \bigcap_{t=1}^{T}\bigcap_{i=1}^{2} \boldsymbol{\tau}_{til(j)} = \boldsymbol{\tau}_{til'(j)} &\equiv \bigcap_{t=1}^{T}\bigcap_{i=1}^{2}\bigcap_{k=1}^{p} {}_k\tau_{til(j)} = {}_k\tau_{til'(j)} \\
&\equiv \bigcap_{t=1}^{T}\bigcap_{i=1}^{2}\bigcap_{k=1}^{p} {}_kH_{0ti(l(j)l'(j))} \\
H_{1t(l(j)l'(j))} : \bigcup_{t=1}^{T}\bigcup_{i=1}^{2} \boldsymbol{\tau}_{til(j)} \neq \boldsymbol{\tau}_{til'(j)} &\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2}\bigcup_{k=1}^{p} {}_k\tau_{til(j)} \neq {}_k\tau_{til'(j)} \\
&\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2} \left[{}_kH^{-}_{1ti(l(j)l'(j))} \cup {}_kH^{+}_{1ti(l(j)l'(j))} \right] \\
&\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2} \left[(\boldsymbol{\tau}_{til(j)} < \boldsymbol{\tau}_{til'(j)}) \cup (\boldsymbol{\tau}_{til(j)} > \boldsymbol{\tau}_{til'(j)}) \right] \\
&\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2}\bigcup_{k=1}^{p} \left[({}_k\tau_{til(j)} < {}_k\tau_{til'(j)}) \cup ({}_k\tau_{til(j)} > {}_k\tau_{til'(j)}) \right] \\
&\equiv \bigcup_{t=1}^{T}\bigcup_{i=1}^{2}\bigcup_{k=1}^{p} \left[{}_kH^{-}_{1ti(l(j)l'(j))} \cup {}_kH^{+}_{1ti(l(j)l'(j))} \right],
\end{aligned}
$$

with $l(j) = 1, \ldots, n_j$.

In practice, we select all the data available until game $t, t = 1, ..., 30$ and proceed to build a ranking for all the players considered, then we include data from game $t + 1$ and so on. Since it is not possible to estimates parameters coefficient with too few data, i.e. building a ranking on 2 or 3 games only (since it can happen that a player of those considered does not play and there are no data available), is has been set the minimum number of games necessary to build our model in this case to 4 games.

4.4 Simulation Study

In order to evaluate the accuracy of this ranking methodology, a simulation study has been performed. The setting of the simulation consists in a league with 16 teams (such as the *Italian men Basketball serie A1 league*), with 10 players per team. The number of response variables it has been set up to 2 and the errors has been designed as heteroscedastic and with a correlation equal to $\rho = 0.2$. In addition, several players has been chosen to have same rank level, in order to simultaneously work either under H_0 and H_1. We also considered two scenarios, in which the true means, on both responses by a quantity equal to two times the standard deviation. In this case, response variables employed are those of the second domain $ORB.p$ and $DRB.p$.

Regarding errors multivariate distributions, three types were chosen:

- *Normal distribution*,
- *Student-t distribution with 3 degrees of freedom*,
- *Right-skewed g and h distribution.*

Figure 4.2 represents the performances of the five simulated players (we considered the front court players from Reyer Venezia) under different type of bivariate distributions. Note that players 1 and 2, as well as 3 and 4, do perform equally in mean but have different scatter parameters.

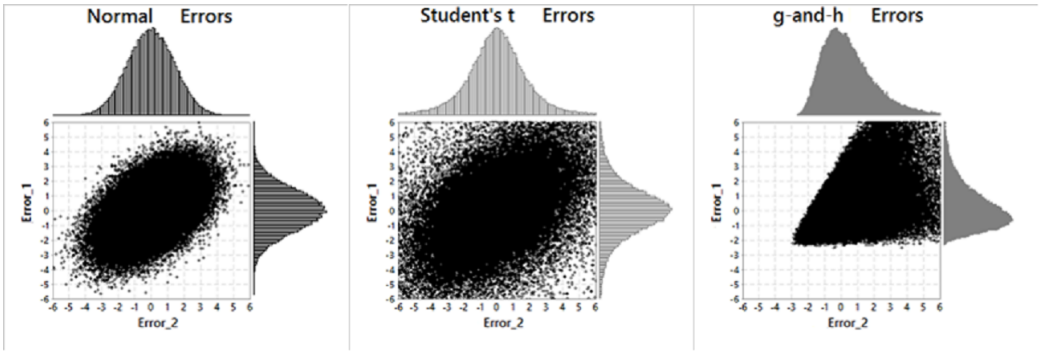


Figure 4.1: Bi-variate errors distribution three types: 1) Normal, 2) t-Student, and 3) g and h.

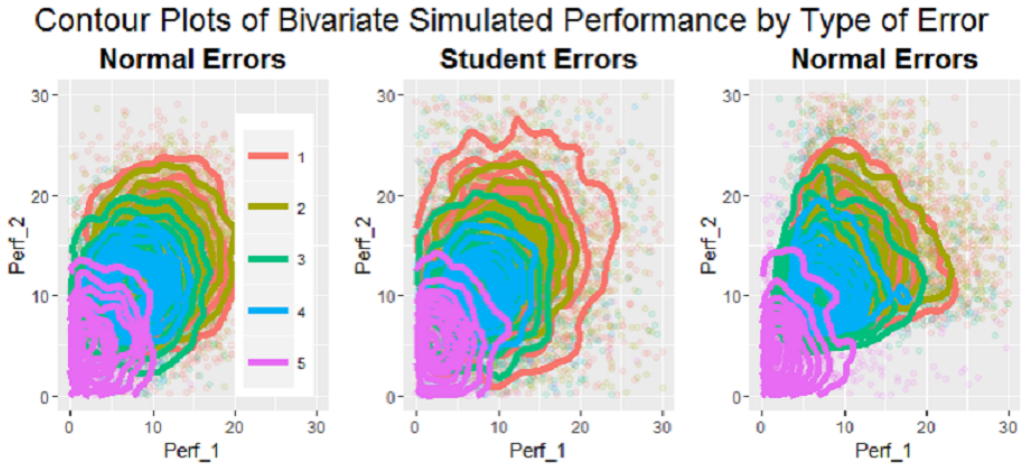


Figure 4.2: Contour plots of bivariate simulated performances by type of error.

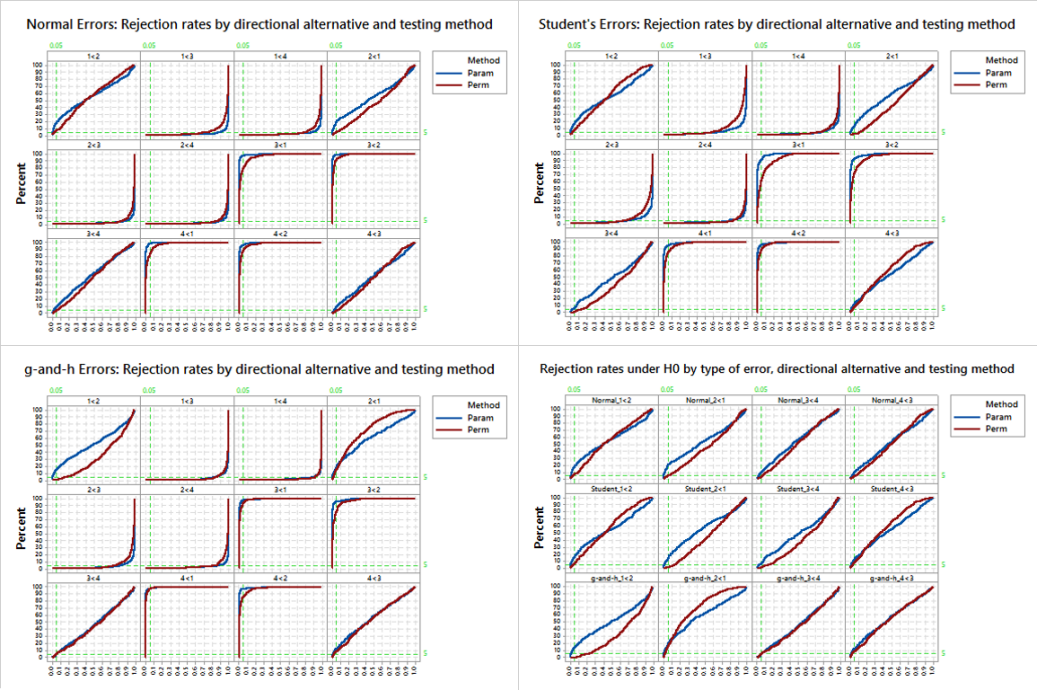


Figure 4.3: Rejection rates by type of error.

By looking at the too large rejection rates, in Figure 4.3, under the alternative hypothesis, we note that the simulation setting was not very well calibrated in what it is concerned with to the shifts between the means of simulated players. Anyway this is not up to now a big issue because the main present goal is on investigating the behaviour of the two testing procedures under the null hypothesis. In this connection let us notice that the permutation tests are more suitable to respect the nominal rejection rates while the parametric tests appear as a biased testing procedure for the specific problem at hand. The reasons behind this unexpected result can be manifold, from the heteroscedastic errors, up to the not negative responses forced to zero when they was simulated as negative or finally to the inaccuracy of the Kost's combination to properly take into account for the joint dependency between univariate p-values. We are currently facing this issue, either theoretically

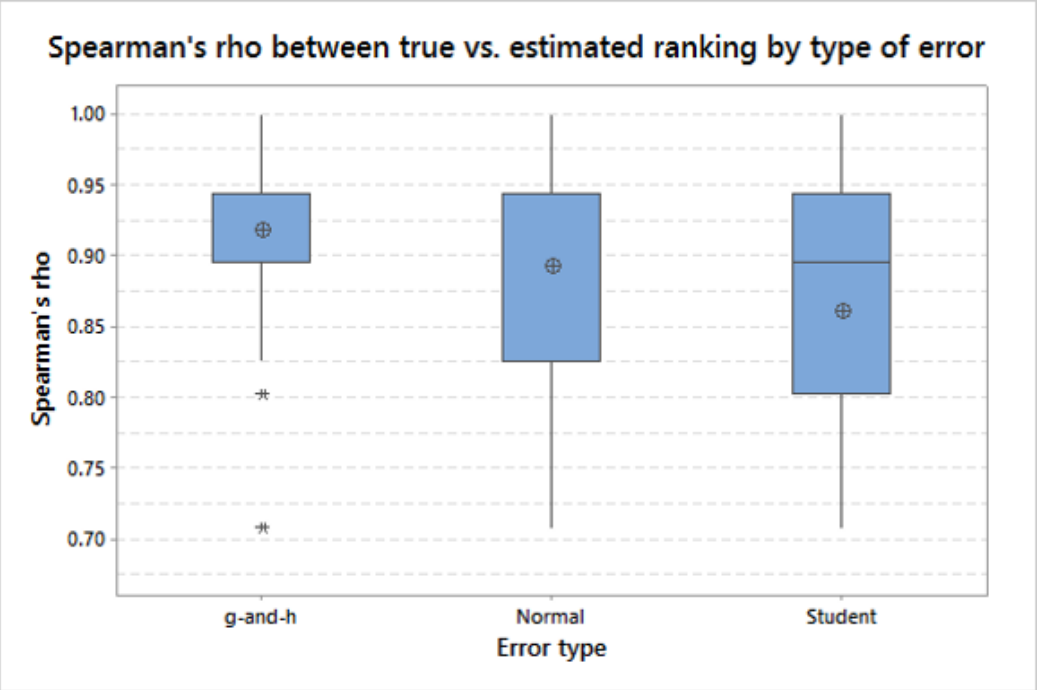


Figure 4.4: Spearman's correlation by type of error.

and computationally, to better try to explain the simulation results.

As about the ability of the permutation testing procedure to estimate the true underlying ranking across players, Figure Figure 4.4 represent the Spearman's correlation rho between the estimated and the true ranking by type of error. Note that the ranking procedure is quite robust with respect to heavy tailed errors when compared to normal errors and it is surprisingly good in case of skewed errors. The last point is unexpected and currently under scrutiny.

Finally, we applied the testing and ranking permutation method to the real case study by using Reyer Venezia players, from season 2016/2017, divided into two main role categories: 'Backcourt' and 'Frontcourt' players, in Figure 4.5 and Figure 4.6.

VE-Backcourt testing and ranking

	1	2	3	4	5
1		.090	**.010**	1.00	1.00
2	1.00		.876	1.00	1.00
3	1.00	1.00		1.00	1.00
4	1.00	.162	**.010**		.906
5	1.00	.552	**.018**	1.00	
ranking=	1	4	5	1	1

VE-Frontcourt testing and ranking

	1	2	3	4	5
1		.372	.360	.640	**.006**
2	**.010**		.444	**.010**	**.006**
3	.894	1.00		.525	**.006**
4	1.00	1.00	1.00		**.006**
5	1.00	1.00	1.00	1.00	
ranking=	3	1	2	3	5

Figure 4.5: Pairwise multivariate p-values and ranking for Backcourt and Frontcourt players. Backcourt players: 1) Bramos Michael, 2) Filloy Ariel, 3) Haynes MarQuez, 4) Mcgee Tyrus, and 5) Tonut Stefano. Frontcourt players: 1) Ejim Melvin, 2) Hagins Jamelle, 3) Ortner Benjamin, 4) Peric Hrovje, and 5) Viggiano Jeff.

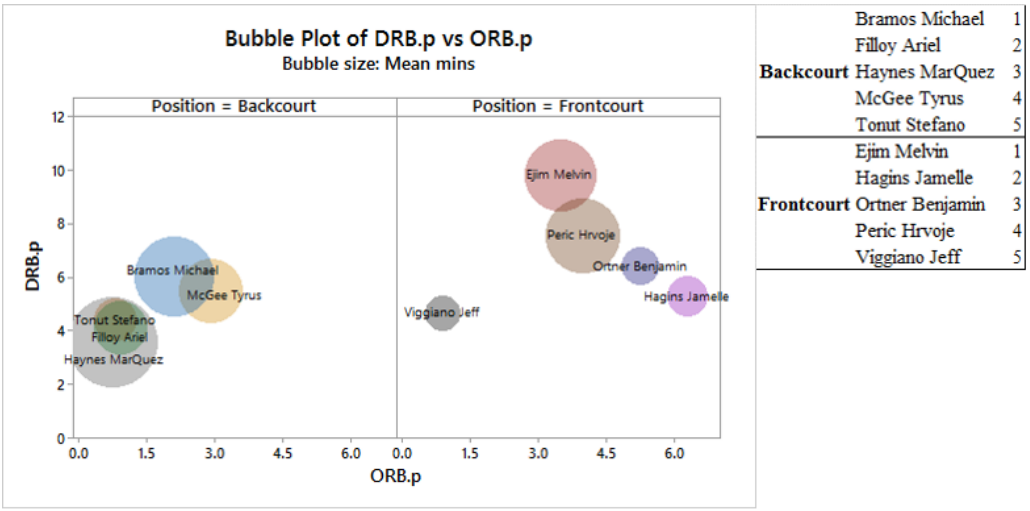


Figure 4.6: Bubbleplot for ORB.p and DRB.p.

Chapter 5

Application of ranking estimation of Italian men Basketball Serie A1

After presenting the ranking methodology, let us now present a real case: the data come from the Italian men Basketball Serie A1 season 2016/2017, in particular our focus is on Reyer Venezia team. The players have been divided into two main subgroups, depending on their role: "Frontcourt" players, which are the ones who play next to the basket, and "Backcourt" players, which are the ones who handle the ball more. In Table 5.1 is shown a summary of the covariates and response variables employed within the model estimated. All the measures considered have been standardized by possessions.

Table 5.1: Summary statistics for variables considered.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Minuti	10	18.903	6.520	10.830	12.853	23.628	29.230
FGA.2P.1.p	10	0.074	0.047	0.020	0.032	0.100	0.150
FGM.2P.1.p	10	0.037	0.025	0.010	0.012	0.057	0.070
FG.Per.2P.1	10	0.407	0.088	0.240	0.338	0.475	0.500

Continued on next page

Table 5.1 – *Continued from previous page*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
FGM.2P.1.AST.Per	10	0.245	0.138	0.060	0.157	0.332	0.460
FGA.2P.2.p	10	0.021	0.011	0.000	0.020	0.028	0.040
FGM.2P.2.p	10	0.006	0.005	0.000	0.000	0.010	0.010
FG.Per.2P.2	10	0.171	0.098	0.030	0.098	0.252	0.310
FGM.2P.2.AST.Per	10	0.118	0.083	0.000	0.075	0.195	0.220
FGA.2P.p	10	0.100	0.049	0.040	0.053	0.130	0.170
FGM.2P.p	10	0.052	0.027	0.020	0.030	0.068	0.100
FG.Per.2P	10	0.448	0.098	0.260	0.402	0.505	0.630
FGM.2P.AST.Per.p	10	0.307	0.134	0.090	0.210	0.410	0.490
FGA.3P.p	10	0.067	0.042	0.000	0.045	0.100	0.110
FGM.3P.p	10	0.026	0.017	0.000	0.015	0.038	0.050
FG.Per.3P	10	0.265	0.153	0.000	0.215	0.373	0.420
FGM.3P.AST.Per	10	0.170	0.160	0.000	0.058	0.292	0.430
Sc.p	10	0.007	0.009	0	0	0.01	0
Sc.AST.Per	10	0.087	0.122	0.000	0.022	0.105	0.400
FTA.p	10	0.040	0.016	0.020	0.030	0.050	0.070
FTM.p	10	0.031	0.011	0.020	0.020	0.040	0.050
FT.Per	10	0.412	0.151	0.150	0.332	0.485	0.640
FGA	10	0.169	0.028	0.130	0.152	0.195	0.210
FGM.p	10	0.078	0.018	0.050	0.070	0.095	0.100
eFG.Per	10	0.519	0.061	0.440	0.480	0.558	0.630
PTS.p	10	0.209	0.041	0.140	0.192	0.242	0.260
TS.Per	10	54.623	6.637	44.820	51.793	58.155	66.980
ORB.p	10	0.027	0.018	0.010	0.010	0.038	0.060
DRB.p	10	0.057	0.020	0.030	0.042	0.068	0.100
AST.p	10	0.041	0.019	0.010	0.030	0.057	0.070
TO.p	10	0.036	0.014	0.010	0.030	0.040	0.060
STL.p	10	0.020	0.007	0.010	0.020	0.020	0.030
BLK.p	10	0.013	0.007	0.000	0.010	0.020	0.020
PF.p	10	0.062	0.021	0.040	0.042	0.070	0.100
FD	10	1.926	0.733	1	1.6	2.4	3
Poss.on	10	35.280	12.413	19.920	23.265	44.787	54.070
Poss.off	10	40.904	12.416	22.110	31.392	52.917	56.270
eFG.net.opp	10	−0.062	0.061	−0.200	−0.068	−0.018	−0.010
ORB.per.net.opp	10	0.006	0.043	−0.070	−0.022	0.035	0.070
DRB.per.net.opp	10	−0.008	0.024	−0.040	−0.028	0.010	0.030
TO.per.net.opp	10	0.002	0.017	−0	0	0.01	0
STL.per.net.opp	10	0.002	0.017	−0.020	−0.010	0.010	0.030
FG.per.2P.1.net.opp	10	−0.052	0.076	−0.190	−0.078	0.002	0.050

Continued on next page

Table 5.1 – *Continued from previous page*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
FG.per.2P.2.net.opp	10	−0.036	0.090	−0.140	−0.110	0.035	0.090
FG.per.3P.net.opp	10	−0.045	0.064	−0.150	−0.098	0.010	0.020
TS.per.net.opp	10	−6.671	7.028	−22.010	−7.620	−2.095	0.300
AST.FGM.per.net.opp	10	−0.048	0.061	−0	−0.1	−0.02	0

In Table 5.2 the same information about covariates is provided, but the means for each player in the team are showned. In this way, we have an indication of the performance of that player at the end of the season.

In Table 5.3 are presented the respective outputs from three different linear models employed to estimate the effect of the net value of each covariate on each univariate response variable. In particular, the first model is a linear model for the first domain, the second model is a linear model for the second domain, and the last models is a linear model on the third domain.

The result does not show a particular good fit on most of the variables singularly. This is because, by modelling positive quantities via linear model, we are not taking into account one of the critical issues of the linear model: it is not well suited for strictly positive data as response variables. On the other hand, linear model is a easy and widely employed method of estimation, which is easy to interpret.

Table 5.2: Reyer Venezia players' mean summary statistics

Nome	Bramos Michael	Filloy Ariel	Haynes MarQuez	McGee Tyrus	Tonut Stefano	Ejim Melvin	Hagins Jamelle	Ortner Benjamin	Peric Hrvoje	Viggiano Jeff
Ruolo	Backcourt	Backcourt	Backcourt	Backcourt	Backcourt	Frontcourt	Frontcourt	Frontcourt	Frontcourt	Frontcourt
minuti	25.77	16.83	29.23	20.93	14.00	23.20	12.47	12.00	23.77	10.83
FGA.2P.1.p	0.03	0.03	0.04	0.07	0.06	0.10	0.14	0.10	0.15	0.02
FGM.2P.1.p	0.01	0.01	0.02	0.04	0.03	0.06	0.07	0.05	0.07	0.01
FG.Per.2P.1	0.33	0.33	0.36	0.43	0.46	0.50	0.48	0.44	0.50	0.24
FGM.2P.1.AST.Per	0.22	0.06	0.06	0.15	0.25	0.42	0.34	0.31	0.46	0.18
FGA.2P.2.p	0.02	0.03	0.02	0.03	0.04	0.02	0.00	0.02	0.01	0.02
FGM.2P.2.p	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00
FG.Per.2P.2	0.26	0.22	0.26	0.12	0.31	0.09	0.05	0.14	0.23	0.03
FGM.2P.2.AST.Per	0.22	0.11	0.09	0.11	0.21	0.07	0.00	0.15	0.22	0.00
FGA.2P.p	0.05	0.05	0.06	0.10	0.10	0.13	0.17	0.13	0.17	0.04
FGM.2P.p	0.03	0.03	0.03	0.04	0.05	0.07	0.10	0.06	0.09	0.02
FG.Per.2P	0.45	0.38	0.40	0.41	0.49	0.51	0.63	0.44	0.51	0.26
FGM.2P.AST.Per.p	0.35	0.16	0.09	0.24	0.32	0.47	0.43	0.32	0.49	0.20
FGA.3P.p	0.10	0.11	0.11	0.10	0.06	0.06	0.00	0.00	0.04	0.09
FGM.3P.p	0.04	0.05	0.04	0.03	0.03	0.03	0.00	0.00	0.01	0.03
FG.Per.3P	0.35	0.38	0.34	0.30	0.38	0.42	0.00	0.00	0.19	0.29
FGM.3P.AST.Per	0.15	0.39	0.43	0.34	0.15	0.08	0.00	0.00	0.11	0.05
Sc.p	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.01	0.01	0.01
Sc.AST.Per	0.03	0.00	0.03	0.04	0.00	0.17	0.40	0.06	0.12	0.02
FTA.p	0.02	0.02	0.03	0.04	0.05	0.05	0.05	0.04	0.07	0.03
FTM.p	0.02	0.02	0.03	0.04	0.04	0.03	0.04	0.02	0.05	0.02
FT.Per	0.34	0.33	0.46	0.47	0.58	0.43	0.49	0.23	0.64	0.15
FGA	0.15	0.16	0.17	0.20	0.16	0.20	0.18	0.13	0.21	0.13
FGM.p	0.07	0.07	0.07	0.08	0.08	0.10	0.10	0.06	0.10	0.05
eFG.Per	0.55	0.56	0.50	0.48	0.54	0.57	0.63	0.44	0.48	0.44
PTS.p	0.19	0.21	0.20	0.22	0.22	0.25	0.25	0.14	0.26	0.15
TS.Per	57.41	57.60	53.12	51.63	58.34	58.62	66.98	44.82	52.28	45.43
ORB.p	0.02	0.01	0.01	0.03	0.01	0.03	0.06	0.05	0.04	0.01
DRB.p	0.06	0.04	0.03	0.05	0.04	0.10	0.05	0.07	0.07	0.06
AST.p	0.03	0.06	0.07	0.06	0.05	0.04	0.01	0.03	0.04	0.02
TO.p	0.01	0.04	0.04	0.04	0.03	0.06	0.05	0.03	0.04	0.02
STL.p	0.01	0.03	0.02	0.02	0.03	0.02	0.01	0.02	0.02	0.02
BLK.p	0.01	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.01
PF.p	0.04	0.05	0.04	0.07	0.04	0.07	0.10	0.09	0.07	0.05
FD	2.00	1.59	2.37	2.19	2.37	2.37	1.62	1.11	3.10	0.54
Poss.on	48.37	32.57	54.07	38.80	26.31	43.79	22.25	21.60	45.12	19.92
Poss.off	27.81	43.61	22.11	37.39	49.88	32.39	53.93	54.59	31.06	56.27
eFG.net.opp	-0.01	-0.07	-0.04	-0.04	-0.20	-0.01	-0.13	-0.06	-0.01	-0.05

Continued on next page

Table 5.2 – *Continued from previous page*

Nome	Bramos Michael	Filloy Ariel	Haynes MarQuez	McGee Tyrus	Tonut Stefano	Ejim Melvin	Hagins Jamelle	Ortner Benjamin	Peric Hrvoje	Viggiano Jeff
ORB.per.net.opp	-0.07	-0.03	0.07	0.00	0.04	0.01	0.02	0.01	-0.04	0.05
DRB.per.net.opp	0.01	0.03	-0.03	-0.04	0.01	-0.01	-0.04	-0.02	0.00	0.01
TO.per.net.opp	0.00	0.00	-0.04	0.01	0.01	0.02	0.00	0.02	0.00	0.00
STL.per.net.opp	-0.01	0.00	0.02	0.01	-0.01	0.01	0.03	-0.02	0. 01	-0.02
FG.per.2P.1.net.opp	-0.06	0.01	0.05	-0.08	-0.19	0.04	-0.14	-0.07	-0.02	-0.06
FG.per.2P.2.net.opp	0.02	-0.11	0.08	-0.01	-0.13	0.04	-0.11	-0.09	0.09	-0.14
FG.per.3P.net.opp	0.01	-0.13	-0.06	0.02	-0.15	-0.03	-0.11	0.01	-0.02	0.01
TS.per.net.opp	-1.67	-6.15	-3.37	-4.95	-22.01	0.30	-15.73	-8.11	-1.35	-3.67
AST.FGM.per.net.opp	0.00	-0.04	-0.02	-0.03	-0.16	0.06	-0.12	-0.04	-0.06	-0.07

Table 5.3: Output of the linear models for the I domain, II domain and III domain, respectively.

	I domain				*II domain*		*III domain*		
	FGM.2P.1.p	FGM.2P.2.p	FGM.3P.p	FTM.p	ORB.p	DRB.p	PTS.p	ORB.p	AST.p
	(1)	(2)	(3)	(4)	(1)	(2)	(1)	(2)	(3)
home1	-0.001	-0.001*	-0.0002	-0.0005	−0.002***	−0.002**	−0.004**	−0.002***	−0.003***
	(0.001)	(0.0003)	(0.0004)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
m.squadra	0.003	-0.002*	0.006***	-0.0002	0.002	-0.003	0.020***	0.002	0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.003)
avversario1	0.0003	0.001	-0.002	0.002	-0.0004	0.0003	-0.003	-0.0004	-0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario2	-0.001	-0.002*	-0.001	-0.001	0.001	-0.001	-0.003	0.002	-0.001
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario3	-0.002	-0.001	0.005***	0.001	0.001	0.001	0.010	0.0003	0.004*
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario4	-0.002	-0.001	0.002	-0.0004	-0.0002	0.001	0.0002	-0.001	-0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario5	-0.006***	-0.001	0.0003	0.007***	0.002	0.005	-0.009	0.001	−0.008***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario6	0.005**	0.003***	-0.003*	-0.002	-0.001	-0.001	0.008	-0.0002	0.004*
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario7	0.003	-0.001	0.003*	-0.002	0.002	0.003	0.008	0.001	0.00004
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario8	-0.002	-0.001	0.001	-0.005**	-0.001	−0.012***	-0.007	0.001	0.0001
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)

Continued on next page

Table 5.3 – *Continued from previous page*

	I domain				*II domain*		*III domain*		
	FGM.2P.1.p	FGM.2P.2.p	FGM.3P.p	FTM.p	ORB.p	DRB.p	PTS.p	ORB.p	AST.p
	(1)	(2)	(3)	(4)	(1)	(2)	(1)	(2)	(3)
avversario9	0.005**	-0.002	0.001	0.001	0.001	0.004	0.013*	0.001	0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario10	-0.0002	0.001	-0.004***	-0.001	-0.003	0.005*	−0.015**	-0.003	-0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario11	0.002	0.001	-0.001	0.002	-0.001	-0.004	0.003	-0.001	0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario12	0.001	0.002*	-0.001	-0.001	0.002	0.001	0.002	0.001	0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.003)
avversario13	0.0001	-0.003***	-0.001	-0.002	−0.004**	-0.004	−0.012*	−0.003*	-0.0005
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario14	0.003	0.0002	0.003**	0.001	0.002	0.003	0.015**	0.002	0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
avversario15	-0.006***	0.004***	0.003*	-0.002	0.003	-0.00005	-0.001	0.002	-0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
minuti	0.001*	0.0001	0.002***	0.0002	0.002***	0.003***	0.007***	0.001***	0.002***
	(0.0004)	(0.0002)	(0.0003)	(0.0004)	(0.0004)	(0.001)	(0.001)	(0.0004)	(0.0005)
I(minuti^2)	-0.00002***	0.00001	-0.00002***	-0.00003***	−0.00002***	−0.00003***	−0.0001***	−0.00002***	−0.00002**
	(0.00001)	(0.00000)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00002)	(0.00001)	(0.00001)
ORB.p	0.261***	-0.024***	-0.105***	-0.007	—	—	—	—	—
	(0.018)	(0.009)	(0.013)	(0.018)	—	—	—	—	—
DRB.p	0.045***	-0.003	-0.043***	0.013	—	—	0.069*	0.109***	−0.077***

Continued on next page

Table 5.3 – *Continued from previous page*

	I domain				*II domain*		*III domain*		
	FGM.2P.1.p	FGM.2P.2.p	FGM.3P.p	FTM.p	ORB.p	DRB.p	PTS.p	ORB.p	AST.p
	(1)	(2)	(3)	(4)	(1)	(2)	(1)	(2)	(3)
	(0.012)	(0.006)	(0.009)	(0.012)	—	—	(0.038)	(0.011)	(0.014)
FG.Per.2P.1	—	—	—	—	0.006^{***}	0.006^{**}	—	—	—
	—	—	—	—	(0.002)	(0.002)	—	—	—
FGM.2P.1.AST.Per	—	—	—	—	0.003^{*}	0.004^{*}	—	—	—
	—	—	—	—	(0.001)	(0.002)	—	—	—
FG.Per.2P.2	—	—	—	—	-0.005^{**}	-0.003	—	—	—
	—	—	—	—	(0.002)	(0.003)	—	—	—
FGM.2P.2.AST.Per	—	—	—	—	0.002	0.004	—	—	—
	—	—	—	—	(0.002)	(0.003)	—	—	—
FG.Per.3P	—	—	—	—	-0.007^{***}	-0.007^{**}	—	—	—
	—	—	—	—	(0.002)	(0.003)	—	—	—
FGM.3P.AST.Per	—	—	—	—	-0.007^{***}	-0.004	—	—	—
	—	—	—	—	(0.002)	(0.003)	—	—	—
Sc.AST.Per	—	—	—	—	0.010^{***}	0.011^{***}	—	—	—
	—	—	—	—	(0.002)	(0.003)	—	—	—
FT.Per	—	—	—	—	-0.001	0.001	—	—	—
	—	—	—	—	(0.001)	(0.002)	—	—	—
AST.p	-0.029^{**}	-0.009	0.020^{*}	-0.014	-0.062^{***}	-0.085^{***}	—	—	—
	(0.014)	(0.007)	(0.010)	(0.014)	(0.013)	(0.019)	—	—	—
TO.p	0.013	0.003	-0.016	-0.009	0.049^{***}	0.068^{***}	-0.017	0.042^{***}	0.064^{***}
	(0.016)	(0.008)	(0.012)	(0.016)	(0.015)	(0.022)	(0.053)	(0.015)	(0.019)

Continued on next page

Table 5.3 – *Continued from previous page*

	I domain				*II domain*		*III domain*		
	FGM.2P.1.p	FGM.2P.2.p	FGM.3P.p	FTM.p	ORB.p	DRB.p	PTS.p	ORB.p	AST.p
	(1)	(2)	(3)	(4)	(1)	(2)	(1)	(2)	(3)
STL.p	0.035	-0.0001	-0.018	-0.002	-0.015	-0.080^{***}	-0.001	-0.021	0.146^{***}
	(0.023)	(0.011)	(0.017)	(0.023)	(0.021)	(0.031)	(0.073)	(0.021)	(0.026)
BLK.p	0.074^{***}	-0.009	-0.107^{***}	-0.018	0.235^{***}	0.253^{***}	0.154^{**}	0.258^{***}	-0.168^{***}
	(0.023)	(0.012)	(0.018)	(0.023)	(0.021)	(0.032)	(0.074)	(0.021)	(0.027)
PF.p	-0.001	0.008	0.017^{*}	-0.007	0.053^{***}	0.023	0.099^{**}	0.053^{***}	-0.003
	(0.012)	(0.006)	(0.009)	(0.012)	(0.011)	(0.017)	(0.039)	(0.011)	(0.014)
FD	0.002^{***}	0.0004^{**}	-0.0005^{**}	0.016^{***}	0.003^{***}	0.001^{***}	0.022^{***}	0.002^{***}	-0.001
	(0.0003)	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0005)	(0.001)	(0.0003)	(0.0004)
Poss.on	0.0003^{*}	-0.0001	-0.0001	-0.0001	-0.001^{***}	-0.001^{***}	-0.00002	-0.001^{***}	-0.0001
	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.001)	(0.0001)	(0.0002)
eFG.net.opp	0.013	-0.003	0.004	-0.053^{***}	0.017	0.044^{***}	-0.009	0.013	0.018
	(0.011)	(0.006)	(0.009)	(0.011)	(0.010)	(0.015)	(0.037)	(0.010)	(0.013)
ORB.per.net.opp	-0.0002	0.001	0.001	0.001	0.002	-0.028^{***}	0.009	0.005^{*}	-0.004
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.005)	(0.011)	(0.003)	(0.004)
DRB.per.net.opp	0.003	-0.001	-0.005^{**}	-0.00004	-0.043^{***}	-0.001	-0.021^{*}	-0.042^{***}	-0.006
	(0.004)	(0.002)	(0.003)	(0.004)	(0.003)	(0.005)	(0.011)	(0.003)	(0.004)
TO.per.net.opp	-0.006	0.001	-0.002	-0.004	0.005	-0.010	-0.012	0.007	-0.016^{***}
	(0.005)	(0.002)	(0.004)	(0.005)	(0.004)	(0.006)	(0.015)	(0.004)	(0.006)
STL.per.net.opp	-0.002	-0.007^{**}	-0.011^{**}	-0.010	-0.021^{***}	-0.014	-0.067^{***}	-0.017^{***}	-0.008
	(0.007)	(0.003)	(0.005)	(0.007)	(0.006)	(0.009)	(0.022)	(0.006)	(0.008)

Continued on next page

Table 5.3 – *Continued from previous page*

	I domain				*II domain*		*III domain*		
	FGM.2P.1.p	FGM.2P.2.p	FGM.3P.p	FTM.p	ORB.p	DRB.p	PTS.p	ORB.p	AST.p
	(1)	(2)	(3)	(4)	(1)	(2)	(1)	(2)	(3)
FG.per.2P.1.net.opp	0.009***	0.0003	-0.003	-0.004	-0.002	-0.006	0.005	-0.0005	-0.001
	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.004)	(0.010)	(0.003)	(0.004)
FG.per.2P.2.net.opp	-0.004**	0.004***	-0.001	-0.001	0.001	-0.0003	-0.004	0.001	-0.001
	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.006)	(0.002)	(0.002)
FG.per.3P.net.opp	-0.003	-0.002	0.006**	0.004	0.002	0.007	0.012	-0.0002	0.004
	(0.004)	(0.002)	(0.003)	(0.004)	(0.004)	(0.005)	(0.013)	(0.004)	(0.005)
TS.per.net.opp	-0.0001	0.0001	-0.00000	0.001***	−0.0002*	−0.001***	0.001*	-0.0001	-0.0001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0001)	(0.0001)
AST.FGM.per.net.opp	0.005**	-0.0004	-0.004**	0.002	-0.002	−0.007**	0.002	-0.001	0.004
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.003)
m.ruoli	-0.011***	0.002	0.007***	-0.002	−0.006***	−0.008***	-0.008	−0.008***	0.010***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
VE.Backcourt.Bramos.Michael	-0.012**	-0.0001	0.002	-0.008	0.006	0.010	-0.024	0.007	−0.021***
	(0.006)	(0.003)	(0.004)	(0.006)	(0.005)	(0.008)	(0.018)	(0.005)	(0.007)
VE.Backcourt.Filloy.Ariel	-0.007	0.005*	0.012***	-0.007	0.0004	0.002	0.024	-0.002	0.009
	(0.006)	(0.003)	(0.004)	(0.006)	(0.005)	(0.008)	(0.019)	(0.005)	(0.007)
VE.Backcourt.Haynes.MarQuez	-0.005	-0.003	-0.001	-0.0003	-0.004	-0.011	-0.017	-0.004	0.006
	(0.006)	(0.003)	(0.004)	(0.006)	(0.005)	(0.008)	(0.018)	(0.005)	(0.007)
VE.Backcourt.McGee.Tyrus	0.013**	-0.005*	-0.003	0.005	0.009*	0.002	0.007	0.009*	0.007
	(0.006)	(0.003)	(0.004)	(0.006)	(0.005)	(0.008)	(0.019)	(0.005)	(0.007)
VE.Frontcourt.Ejim.Melvin	0.0001	-0.001	0.012***	-0.004	-0.005	0.028***	0.013	−0.010*	0.007

Continued on next page

Table 5.3 – *Continued from previous page*

	I domain				*II domain*		*III domain*		
	FGM.2P.1.p	FGM.2P.2.p	FGM.3P.p	FTM.p	ORB.p	DRB.p	PTS.p	ORB.p	AST.p
	(1)	(2)	(3)	(4)	(1)	(2)	(1)	(2)	(3)
	(0.006)	(0.003)	(0.004)	(0.006)	(0.005)	(0.008)	(0.018)	(0.005)	(0.007)
VE.Frontcourt.Hagins.Jamelle	0.013**	-0.002	-0.013***	0.009	0.011*	−0.025***	0.034*	0.018***	−0.015**
	(0.006)	(0.003)	(0.005)	(0.006)	(0.006)	(0.009)	(0.021)	(0.006)	(0.007)
VE.Frontcourt.Ortner.Benjamin	-0.005	0.004	-0.010**	-0.007	0.013**	-0.002	−0.040**	0.013**	0.007
	(0.006)	(0.003)	(0.004)	(0.006)	(0.005)	(0.008)	(0.019)	(0.005)	(0.007)
VE.Frontcourt.Peric.Hrvoje	0.016***	0.001	-0.006	0.001	0.002	-0.001	0.002	0.002	0.003
	(0.006)	(0.003)	(0.004)	(0.006)	(0.005)	(0.008)	(0.018)	(0.005)	(0.007)
home1:m.squadra	-0.002	0.0002	-0.002	0.001	0.002	0.001	-0.007	0.003	−0.005**
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.003)
home1:avversario1	0.001	0.0004	0.001	0.001	-0.001	-0.002	0.009	-0.001	0.004*
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario2	-0.001	-0.002**	-0.0002	-0.001	-0.001	0.0001	-0.007	-0.001	-0.001
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario3	-0.0001	-0.0004	0.0003	-0.00004	0.0004	-0.002	-0.001	0.001	-0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario4	-0.004**	-0.0004	-0.0001	0.002	0.002	0.002	-0.009	0.002	-0.004
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario5	-0.001	-0.001	0.001	0.001	0.0004	0.003	0.001	-0.0001	−0.006***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario6	0.0002	0.001	-0.002	-0.0004	-0.003	0.002	-0.001	-0.003	0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)

Continued on next page

Table 5.3 – *Continued from previous page*

	I domain				*II domain*		*III domain*		
	FGM.2P.1.p	FGM.2P.2.p	FGM.3P.p	FTM.p	ORB.p	DRB.p	PTS.p	ORB.p	AST.p
	(1)	(2)	(3)	(4)	(1)	(2)	(1)	(2)	(3)
home1:avversario7	0.003	-0.0003	-0.0001	-0.002	-0.0004	-0.001	0.005	-0.0002	-0.001
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario8	0.004*	0.001	0.002	-0.003	0.001	−0.005*	0.013*	0.002	0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario9	-0.005**	0.002*	-0.0004	-0.002	0.001	0.006**	−0.012*	0.0005	−0.006**
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario10	-0.003	0.0002	-0.0002	-0.002	0.0005	0.002	-0.011	0.0001	0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario11	0.003	0.0002	-0.002	0.003	-0.002	-0.004	0.004	-0.002	-0.003
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario12	0.003	-0.001	0.001	-0.0001	-0.0002	0.0005	0.001	-0.001	0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario13	0.001	0.001	0.002	-0.00003	0.00002	0.001	0.007	-0.0005	0.006**
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario14	0.003	-0.0002	-0.002	0.002	0.002	-0.004	-0.001	0.003	-0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
home1:avversario15	-0.002	-0.002**	0.001	0.001	-0.0002	0.003	-0.002	-0.0003	-0.0002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
Constant	0.005	0.006***	0.005*	0.013***	0.014***	0.034***	0.046***	0.009***	0.018***
	(0.004)	(0.002)	(0.003)	(0.004)	(0.003)	(0.005)	(0.012)	(0.003)	(0.004)

Continued on next page

Table 5.3 – *Continued from previous page*

	I domain				*II domain*		*III domain*		
	FGM.2P.1.p	FGM.2P.2.p	FGM.3P.p	FTM.p	ORB.p	DRB.p	PTS.p	ORB.p	AST.p
	(1)	(2)	(3)	(4)	(1)	(2)	(1)	(2)	(3)
Observations	3,921	3,921	3,921	3,921	3,921	3,921	3,921	3,921	3,921
R^2	0.155	0.054	0.132	0.401	0.211	0.108	0.203	0.194	0.095
Adjusted R^2	0.142	0.038	0.118	0.391	0.197	0.092	0.190	0.181	0.081
Residual Std. Error I (df = 3857)	0.033	0.017	0.025	0.033	—	—	—	—	—
Residual Std. Error II (df = 3851)	—	—	—	—	0.030	0.046	—	—	—
Residual Std. Error III (df = 3859)	—	—	—	—	—	—	0.108	0.031	0.039
F Statistic I (df = 63; 3857)	11.259^{***}	3.491^{***}	9.289^{***}	40.937^{***}	—	—	—	—	—
F Statistic II (df = 69; 3851)	—	—	—	—	14.902^{***}	6.727^{***}	—	—	—
F Statistic III (df = 61; 3859)	—	—	—	—	—	—	16.104^{***}	15.227^{***}	6.644^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

5.1 Significance testing on players' parameters

In this part, univariate and multivariate p-values matrices will be presented. If we look at the univariate p-values matrices, we notice information about individual's performances which are under, in line or above the average performance of the same role players belonging to Reyer Venezia, for every single response variable within the same domain. If we look at the multivariate p-values matrices, we notice information about whether individual's performance, regarding the whole domain considered, is under, in line or above the average performance of the same role players belonging to Reyer Venezia. It can occurs that both directional p-values regarding player $l(j)$ is better than player $l'(j)$ and vice-versa are statistically significant. This means that his perform is average overall and that focusing on some univariate performance, player $l(j)$ is better than the average performancefor some other ones, he performs worse than the average performance.

Table 5.4: Directional univariate p-values for significance testing first domain Backcourt.

	2P.FGM.1.p	2P.FGM.2.p	3P.FGM.p	FTM.p	2P.FGM.1.p	2P.FGM.2.p	3P.FGM.p	FTM.p
Bramos Michael	0.981	0.515	0.339	0.931	0.019	0.485	0.661	0.069
Filloy Ariel	0.895	0.046	0.003	0.875	0.105	0.954	0.997	0.125
Haynes MarQuez	0.792	0.870	0.558	0.518	0.208	0.130	0.442	0.482
McGee Tyrus	0.012	0.966	0.779	0.204	0.988	0.034	0.221	0.796
Tonut Stefano	0.179	0.261	0.127	0.179	0.821	0.739	0.873	0.821

Table 5.5: Directional univariate p-values for significance testing first domain Frontcourt.

	2P.FGM.1.p	2P.FGM.2.p	3P.FGM.p	FTM.p	2P.FGM.1.p	2P.FGM.2.p	3P.FGM.p	FTM.p
Ejim Melvin	0.495	0.653	0.002	0.737	0.505	0.347	0.998	0.263
Hagins Jamelle	0.018	0.764	0.996	0.072	0.982	0.236	0.004	0.928
Ortner Benjamin	0.804	0.112	0.990	0.901	0.196	0.888	0.010	0.099
Peric Hrvoje	0.003	0.428	0.908	0.411	0.997	0.572	0.092	0.589
Viggiano Jeff	0.020	0.457	0.031	0.487	0.980	0.543	0.969	0.513

Table 5.6: Directional univariate p-values for significance testing second domain Backcourt.

	ORB.p	DRB.p	ORB.p	DRB.p
Bramos Michael	0.140	0.098	0.860	0.902
Filloy Ariel	0.469	0.388	0.531	0.612
Haynes MarQuez	0.758	0.918	0.242	0.082
McGee Tyrus	0.048	0.386	0.952	0.614
Tonut Stefano	0.140	0.401	0.860	0.599

Table 5.7: Directional univariate p-values for significance testing second domain Frontcourt.

	ORB.p	DRB.p	ORB.p	DRB.p
Ejim Melvin	0.845	0.0001	0.155	1.000
Hagins Jamelle	0.032	0.998	0.968	0.002
Ortner Benjamin	0.008	0.608	0.992	0.392
Peric Hrvoje	0.340	0.554	0.660	0.446
Viggiano Jeff	0.028	0.480	0.972	0.520

Table 5.8: Directional univariate p-values for significance testing third domain Backcourt.

	PTS.p	ORB.p	AST.p	PTS.p	ORB.p	AST.p
Bramos Michael	0.902	0.095	0.999	0.098	0.905	0.001
Filloy Ariel	0.099	0.638	0.091	0.901	0.362	0.909
Haynes MarQuez	0.825	0.761	0.176	0.175	0.239	0.824
McGee Tyrus	0.348	0.048	0.145	0.652	0.952	0.855
Tonut Stefano	0.402	0.167	0.447	0.598	0.833	0.553

Table 5.9: Directional univariate p-values for significance testing third domain Frontcourt.

	PTS.p	ORB.p	AST.p	PTS.p	ORB.p	AST.p
Ejim Melvin	0.235	0.969	0.160	0.765	0.031	0.840
Hagins Jamelle	0.049	0.001	0.977	0.951	0.999	0.023
Ortner Benjamin	0.983	0.007	0.140	0.017	0.993	0.860
Peric Hrvoje	0.448	0.381	0.313	0.552	0.619	0.687
Viggiano Jeff	0.400	0.016	0.437	0.600	0.984	0.563

Table 5.10: Directional multivariate p-values for significance testing first domain Backcourt.

	$Pr(\tau_{l(j)} < 0)$	$Pr(\tau_{l(j)} > 0)$
Bramos Michael	0.885	0.050
Filloy Ariel	0.019	0.363
Haynes MarQuez	0.919	0.244
McGee Tyrus	0.129	0.246
Tonut Stefano	0.090	0.990

Table 5.11: Directional multivariate p-values for significance testing first domain Frontcourt.

	$Pr(\tau_{l(j)} < 0)$	$Pr(\tau_{l(j)} > 0)$
Ejim Melvin	0.056	0.629
Hagins Jamelle	0.087	0.081
Ortner Benjamin	0.753	0.026
Peric Hrvoje	0.053	0.543
Viggiano Jeff	0.023	0.954

Table 5.12: Directional multivariate p-values for significance testing second domain Backcourt.

	$Pr(\tau_{l(j)} < 0)$	$Pr(\tau_{l(j)} > 0)$
Bramos Michael	0.075	0.970
Filloy Ariel	0.489	0.684
Haynes MarQuez	0.944	0.100
McGee Tyrus	0.094	0.893
Tonut Stefano	0.220	0.851

Table 5.13: Directional multivariate p-values for significance testing second domain Frontcourt.

	$Pr(\tau_{l(j)} < 0)$	$Pr(\tau_{l(j)} > 0)$
Ejim Melvin	0.001	0.442
Hagins Jamelle	0.144	0.014
Ortner Benjamin	0.033	0.750
Peric Hrvoje	0.499	0.649
Viggiano Jeff	0.073	0.845

Table 5.14: Directional multivariate p-values for significance testing third domain Backcourt.

	$Pr(\tau_{l(j)} < 0)$	$Pr(\tau_{l(j)} > 0)$
Bramos Michael	0.554	0.005
Filloy Ariel	0.113	0.874
Haynes MarQuez	0.620	0.346
McGee Tyrus	0.062	0.972
Tonut Stefano	0.321	0.857

Table 5.15: Directional multivariate p-values for significance testing third domain Frontcourt.

	$Pr(\tau_{l(j)} < 0)$	$Pr(\tau_{l(j)} > 0)$
Ejim Melvin	0.357	0.249
Hagins Jamelle	0.003	0.267
Ortner Benjamin	0.031	0.204
Peric Hrvoje	0.438	0.819
Viggiano Jeff	0.069	0.898

5.2 Significance testing on pairwise player comparison parameters

In this part, univariate and multivariate p-values matrices for testing pairwise comparison among all Backcourt and Frontcourt Reyer Venezia players will be presented. If we look at the univariate p-values matrices, we gain information on every player compared to all the others playing the same role. Focusing on the values on the same line, we can verify if player $l(j) - th$ performance is better than all the other players, while focusing on the values on the same column, we can verify if player $l(j) - th$ performance is worse than all the other players.

If we look at the multivariate p-values matrices, we can verify which player is the best, on the whole domain, considering each pair of players in the same role in Reyer Venezia. Regarding every pair of players, it can occurs that both directional p-values verifying if player $l(j)$ is better than player $l'(j)$ and vice-versa are statistically significant. This means that they have the same ranking overall and that each of them and that, for some univariate performance, player $l(j)$ is better than player $l'(j)$ and for some other ones, player $l'(j)$ is better than player $l(j)$.

Table 5.16: Pairwise multivariate p-values first domain Backcourt.

	Bramos Michael	Filloy Ariel	Haynes MarQuez	McGee Tyrus	Tonut Stefano
Bramos Michael	1	0.352	0.583	0.366	0.595
Filloy Ariel	0.396	1	0.250	0.072	0.594
Haynes MarQuez	0.384	0.121	1	0.560	0.594
McGee Tyrus	0.084	0.006	0.305	1	0.591
Tonut Stefano	0.589	0.589	0.590	0.592	1

Table 5.17: Pairwise multivariate p-values first domain Frontcourt.

	Ejim Melvin	Hagins Jamelle	Ortner Benjamin	Peric Hrvoje	Viggiano Jeff
Ejim Melvin	1	0.290	0.050	0.919	0.588
Hagins Jamelle	0.845	1	0.204	0.974	0.586
Ortner Benjamin	0.579	0.351	1	0.735	0.589
Peric Hrvoje	0.056	0.019	0.0003	1	0.580
Viggiano Jeff	0.596	0.597	0.595	0.604	1

Table 5.18: Pairwise multivariate p-values second domain Backcourt.

	Bramos Michael	Filloy Ariel	Haynes MarQuez	McGee Tyrus	Tonut Stefano
Bramos Michael	1	0.001	0.034	0.225	0.588
Filloy Ariel	1.000	1	0.931	0.993	0.589
Haynes MarQuez	0.983	0.152	1	0.897	0.589
McGee Tyrus	0.639	0.006	0.117	1	0.583
Tonut Stefano	0.596	0.595	0.595	0.601	1

Table 5.19: Pairwise multivariate p-values second domain Frontcourt.

	Ejim Melvin	Hagins Jamelle	Ortner Benjamin	Peric Hrvoje	Viggiano Jeff
Ejim Melvin	1	0.00004	0.001	0.104	0.589
Hagins Jamelle	0.277	1	0.756	0.664	0.587
Ortner Benjamin	0.291	0.234	1	0.693	0.587
Peric Hrvoje	0.419	0.013	0.118	1	0.584
Viggiano Jeff	0.595	0.597	0.597	0.600	1

Table 5.20: Pairwise multivariate p-values third domain Backcourt.

	Bramos Michael	Filloy Ariel	Haynes MarQuez	McGee Tyrus	Tonut Stefano
Bramos Michael	1	0.001	0.034	0.225	0.588
Filloy Ariel	1.000	1	0.931	0.993	0.589
Haynes MarQuez	0.983	0.152	1	0.897	0.589
McGee Tyrus	0.639	0.006	0.117	1	0.583
Tonut Stefano	0.596	0.595	0.595	0.601	1

Table 5.21: Pairwise multivariate p-values third domain Frontcourt.

	Ejim Melvin	Hagins Jamelle	Ortner Benjamin	Peric Hrvoje	Viggiano Jeff
Ejim Melvin	1	0.00004	0.001	0.104	0.589
Hagins Jamelle	0.277	1	0.756	0.664	0.587
Ortner Benjamin	0.291	0.234	1	0.693	0.587
Peric Hrvoje	0.419	0.013	0.118	1	0.584
Viggiano Jeff	0.595	0.597	0.597	0.600	1

Table 5.22: Reyer Venezia Backcourt first domain results.

	Ranking	Score	Mean TS%
Bramos Michael	3	2.360	57.410
Filloy Ariel	5	1.350	57.600
Haynes MarQuez	1	1.250	53.120
McGee Tyrus	1	2.160	51.630
Tonut Stefano	3	0.950	58.340

Table 5.23: Reyer Venezia Frontcourt first domain results.

	Ranking	Score	Mean TS%
Ejim Melvin	1	3.970	58.620
Hagins Jamelle	4	2.410	66.980
Ortner Benjamin	4	3.300	44.820
Peric Hrvoje	1	3.350	52.280
Viggiano Jeff	3	1.830	45.430

Table 5.24: Reyer Venezia Backcourt second domain results.

	Ranking	Somma punteggio	Mean TRB%
Bramos Michael	1	0	0.080
Filloy Ariel	5	0	0.050
Haynes MarQuez	4	0	0.040
McGee Tyrus	2	0	0.080
Tonut Stefano	3	0	0.050

Table 5.25: Reyer Venezia Frontcourt second domain results.

	Ranking	Somma punteggio	Mean TRB%
Ejim Melvin	1	0	0.130
Hagins Jamelle	5	0	0.110
Ortner Benjamin	4	0	0.120
Peric Hrvoje	2	0	0.120
Viggiano Jeff	3	0	0.070

Table 5.26: Reyer Venezia Backcourt third domain results.

	Ranking	Somma punteggio	Mean OE
Bramos Michael	4	7.020	0.590
Filloy Ariel	5	7.650	0.550
Haynes MarQuez	1	8.240	0.510
McGee Tyrus	2	8.520	0.550
Tonut Stefano	3	5.340	0.560

Table 5.27: Reyer Venezia Frontcourt third domain results.

	Ranking	Somma punteggio	Mean OE
Ejim Melvin	1	9.710	0.560
Hagins Jamelle	4	6.710	0.710
Ortner Benjamin	4	6.200	0.750
Peric Hrvoje	1	9.700	0.570
Viggiano Jeff	3	5.150	0.610

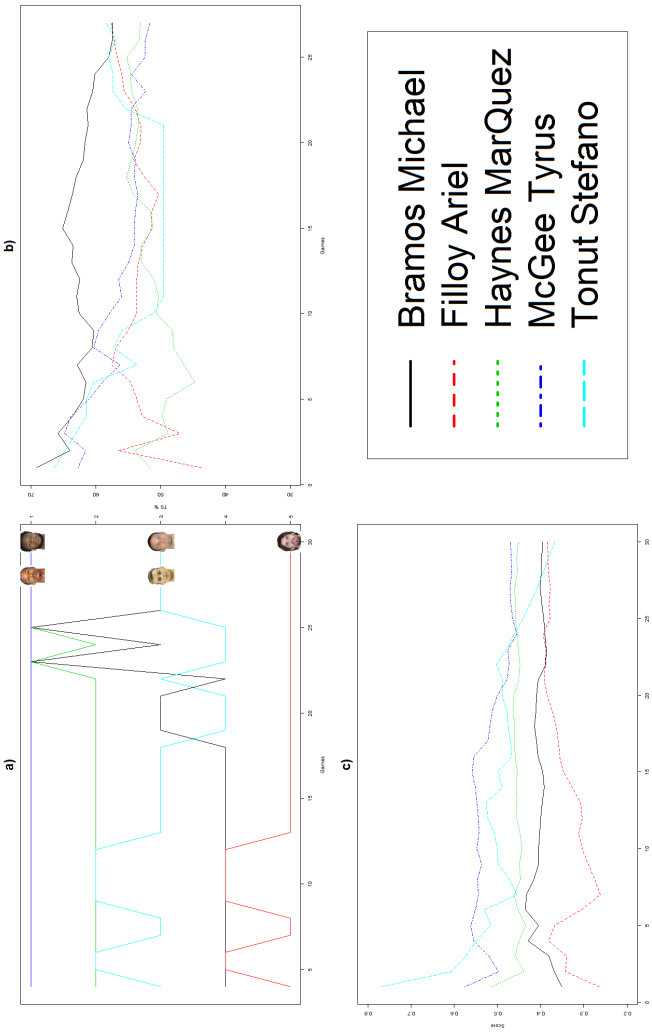


Figure 5.1: a) Reyer Venezia Backcourt players dynamic ranking first domain. b) Reyer Venezia Backcourt players TS % over the season. c) Reyer Venezia Backcourt players Dynamic score first domain.

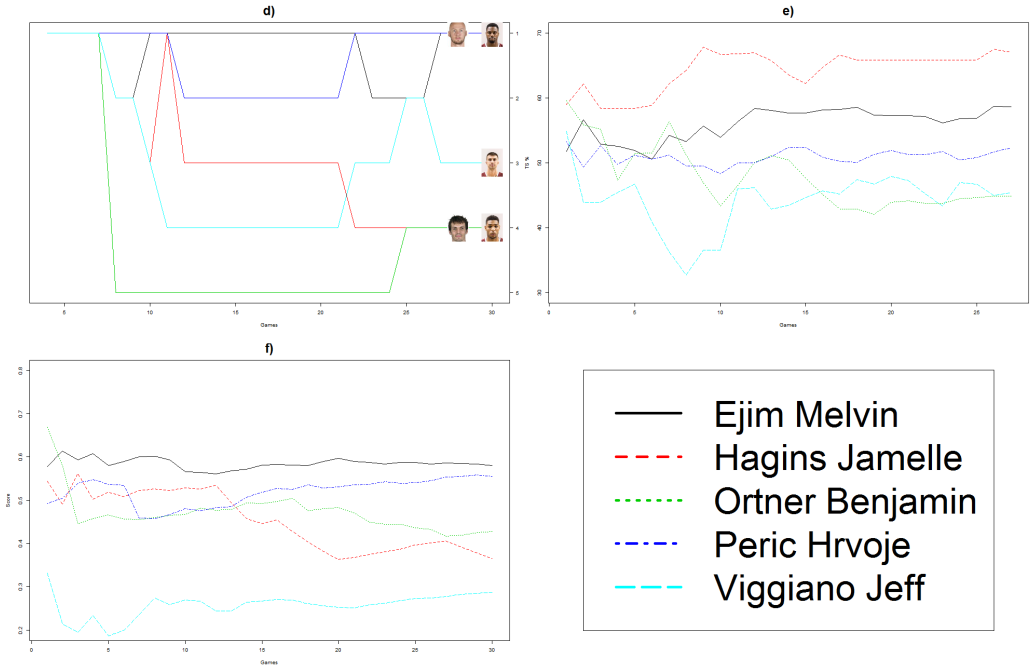


Figure 5.2: d) Reyer Venezia Frontcourt players dynamic ranking first domain. e) Reyer Venezia Frontcourt players TS % over the season. f) Reyer Venezia Frontcourt players Dynamic score first domain.

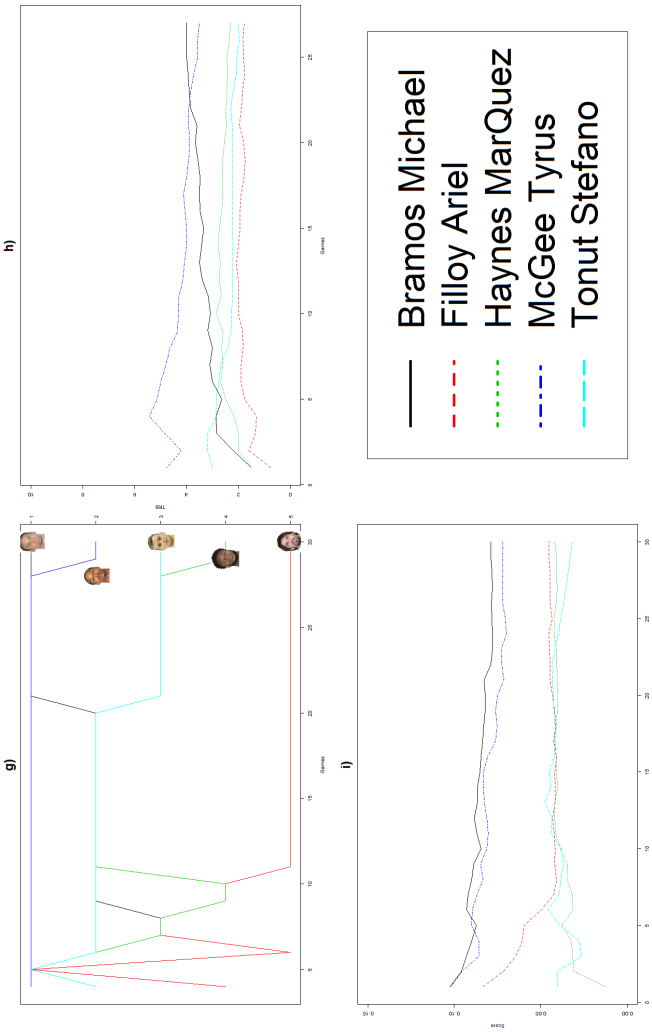


Figure 5.3: g) Reyer Venezia Backcourt players dynamic ranking second domain. h) Reyer Venezia Backcourt players TRB over the season. i) Reyer Venezia Backcourt players Dynamic score second domain.

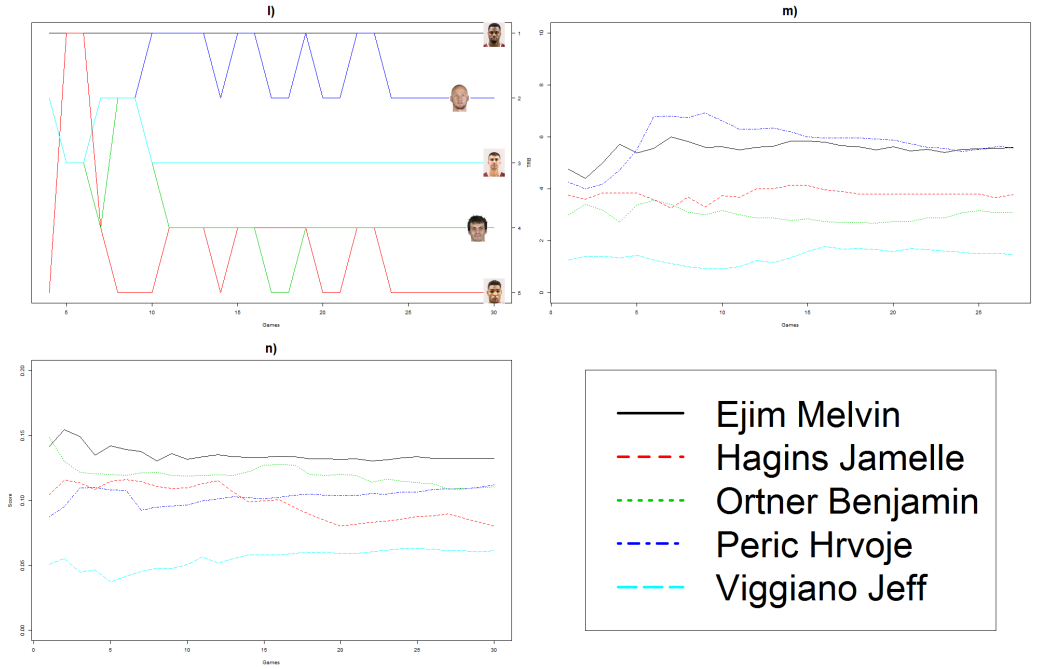


Figure 5.4: l) Reyer Venezia Frontcourt players dynamic ranking second domain. m) Reyer Venezia Frontcourt players TRB over the season. n) Reyer Venezia Frontcourt players Dynamic score second domain.

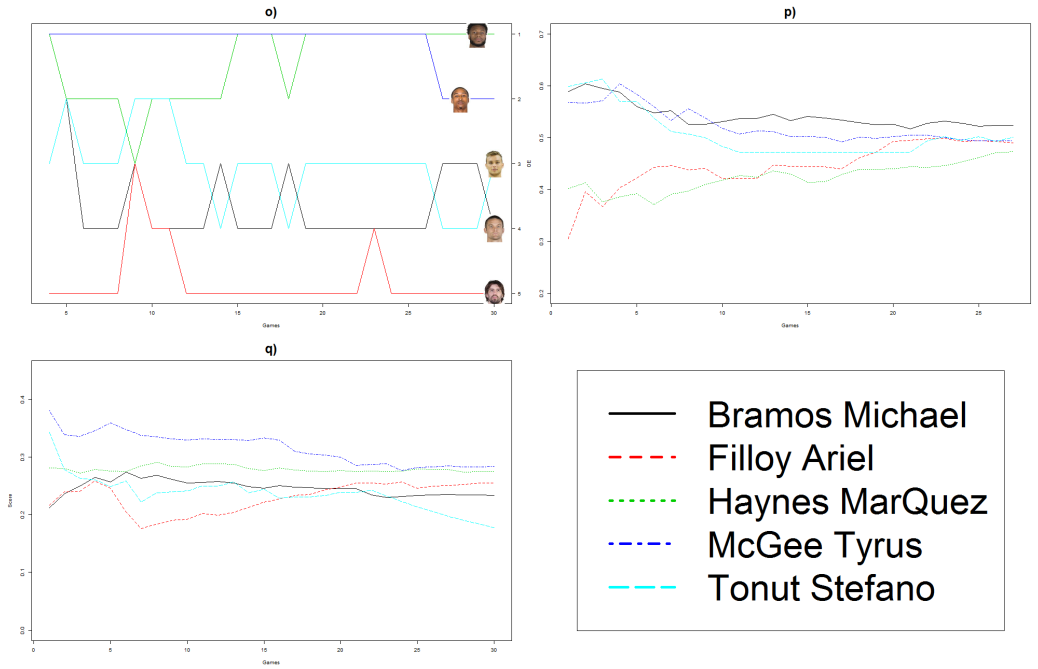


Figure 5.5: o) Reyer Venezia Backcourt players dynamic ranking third domain. p) Reyer Venezia Backcourt players OE over the season. q) Reyer Venezia Backcourt players Dynamic score third domain.

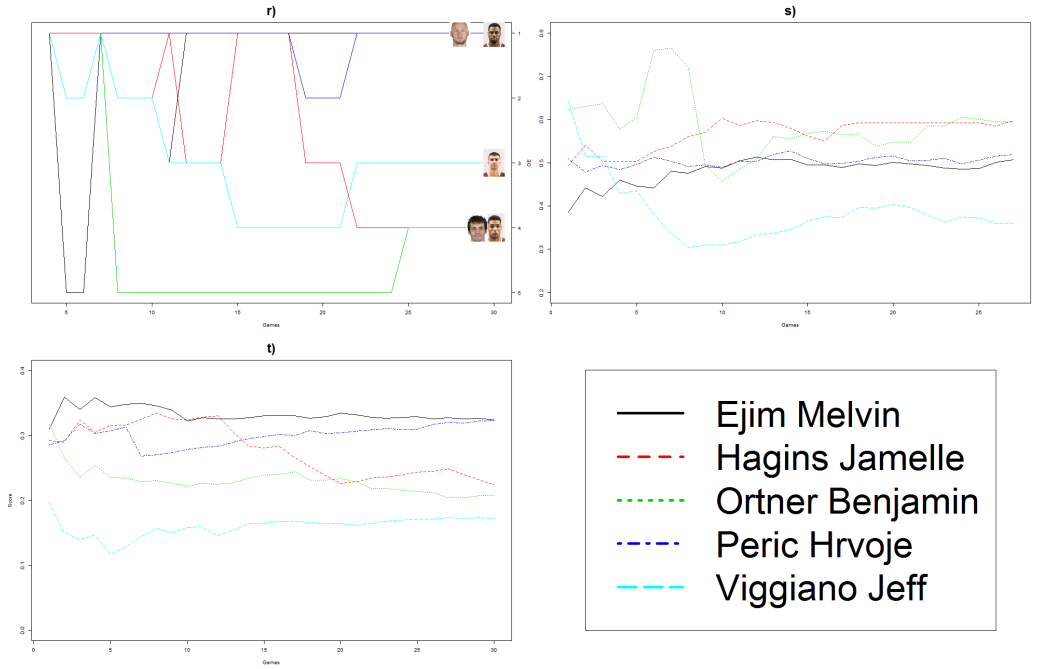


Figure 5.6: r) Reyer Venezia Frontcourt players dynamic ranking third domain. s) Reyer Venezia Frontcourt players OE over the season. t) Reyer Venezia Frontcourt players Dynamic score third domain.

5.3 Conclusions

Once the methodology has been presented, let us analyse strenghts and weaknesses of it. As we have seen in Chapter 2, the most employed performance indicators, *Player efficiency rating*, *Wins produced* and *Approximate value* extrapolate contributions values from all the data in the league, considering more seasons for the estimate of the effect of each action a player can make on the floor. In this way, within the same actions, players contributions are weighted the same. Ranking players with this methodology allow us to asses a specific value to a specific player, for that specific move, in that particular game, considering a setting richer in information.

Among these performance measures, only *Approximate Value* is specifically tailored on individuals, as in the case of our method. Both of them, in addition, employ both bottom-up and top-down measures, in order to mitigate the bad effects carried by using only one type of measures, as explained previously. The abundance of data allow us to build several indexes which can be useful in the composition of the covariates for every model to estimate.

On the other hand, this method represent a new way of considering players comparison. Borrowing for social science ranking methodology on dyadic design data, we are able to look at each player compared to all the others considered. Therefore it is possible to build a ranking designed on that particular season (or those specific games), without the need of a lot of data, which could be difficult to have. By building a ranking with a dynamic approach it is also possible monitor the evolution of the individuals and the entire team during the season, helping the front office to monitor and adjust strategies in real time.

Regarding model estimation, we have seen that not all the models give a

good description of the single performance. Considering that we are dealing with positive quantities, linear models are not the best tool with which analyse such data, although it is definitely a fast and widely employed method of modelling sport performances. Regarding this problem, a great variety of models can be considered, both parametric and non parametric ones. A good feature of this ranking method is that the models used for single performances does not need to be the same and considering the same number of covariates, so we are able also to perform variable selection through all-subset regression, lasso regularization or stepwise regression.

Regarding player roles, it is possible to exploit further classifications in order to better divide player which contributes in the most similar way, since it does not seem that a two level classification between backcourt and frontcourt players seems best suited. This problem can be analysed through cluster analysis or directly working with coaches and professionals.

Regarding the ranking itself, we need to take into account the easy communicability of such information. On the other hand, when dealing with the question "Who is the best?" with same level ranking players, it is difficult to provide an answer with such method. Harville (2003) suggested to calculate a score instead of performing a ranking, in order to have a single number to represent individual's contribution. We have showed in the previous chapter that this can be done with this method, by considering the model matrix, and results are in line with those of the ranking.

All things considered, we can say that this ranking methods for basketball players present numerous improvements to its predecessors.

Bibliography

Adam, A. (2016). "Generalised linear model for football matches prediction". In: *CEUR Workshop Proceedings Vol. 1842, Proceedings of the Workshop on Machine Learning and Data Mining for Sports Analytics, Riva del Garda, Italy, September 19, 2016*.

Arboretti, R., S. Bonnini, L. Corain, and L. Salmaso (2014). "A permutation approach for ranking of multivariate populations". In: *Journal of Multivariate Analysis* 132, pp. 39–57.

Basketball reference, Glossary. URL: `https://www.basketball-reference.com/about/glossary.html`.

Berri, D., M. Schmidt, and S. Brook (2006). *The Wages of Win: Taking measures of the many Myths in modern sports*. Standford, California: Standford University Press.

Bradley, R. A. and M. E. Terry (1952). "Rank analysis of incomplete block designs: I, The method of paired comparisons". In: *Biometrika* 39, 324–345.

Bretz, F., T. Hothorn, and P. Westfall. *Multiple Comparisons Using R*. CRC Press. Boca Raton.

Cattelan, M., C. Varin, and D. Firth (2013). "Dynamic Bradley–Terry modelling of sports tournaments". In: *Journal of the Royal Statistical Society Series C Applied Statistics* 62, pp. 135–150.

Chan, V. (2011). "Prediction Accuracy of Linear Models for Paired Comparisons in Sports". In: *Journal of quantitative analysis in sports* 7, pp. 3–18.

Chen, T. and Q. Fan (2016). "A functional data approach to model score difference process in professional basketball games". In: *Journal of Applied Statistics*.

Franks, A., A. Miller, L. Bornn, and K. Goldsberry (2014). "Characterizing the Spatial Structure of Defensive Skill in Professional Basketball". In: *The Annals of Applied Statistics* 9.

Goddard, J. and I. Asimakopoulos (2004). "Forecasting football results and the efficiency of fixed-odds betting". In: *Journal of Forecasting* 23, 51–66.

H., Koning R. (2000). "Balance in competitions in Dutch soccer". In: *Statistician* 49, 419–431.

Harville, D. A. (2003). "The Selection or Seeding of College Basketball or Football Teams for Postseason Competition". In: *Journal of the American Statistical Association* 98, pp. 17–27.

Harville, D. A. and M. H. Smith (1994). "The Home-Court Advantage: How Large Is It and Does It Vary From Team to Team". In: *The American Statistician* 48, pp. 22–28.

Hollinger, J. (2005). *Pro Basketball Forecast 2005-2006 Edition*. Dulles, Virginia: Potomac Books.

How to calculate Wins Produced. URL: `http://wagesofwins.com/how-to-calculate-wins-produced/`.

Karlis, D. and I Ntzoufras (2003). "Analysis of sports data by using bivariate Poisson models". In: *Statistician* 52, 381–393.

Kenny, D. A., D. A. Kashy, and W. L. Cook (2006). *The Analysis of Dyadic Data*. New York: Guilford Press.

Kost, J. and M. McDermott (2002). "Combining dependent P-values". In: *Statistics & Probability Letters* 60, pp. 183–190.

Lega Basket Serie A, glossario. URL: `http://web.legabasket.it/statistics/`.

M., Cattelan (2012). "Paired Comparison Data: A Review with Emphasis on Dependent Data". In: *Statistical Science* 27, pp. 412–433.

Oliver, D. (2004). *Basketball on paper: Rules and Tools for Performance Analysis*. Dulles, Virginia: Potomac Books.

Poole, W., D. L. Gibbs, I. Shmulevich, B. Bernard, and A. T. Knijnenburg (2016). "Combining dependent P-values with an empirical adaptation of Brown's method". In: *Bioinformatics* 32, i430–i436.

Potential assist definition. URL: `http://www.82games.com/assisted.htm`.

Rasmussen, R. V. and M. A. Trick (2008). "Round robin scheduling – a survey". In: *European Journal of Operational Research* 188(3), pp. 617–636.

Roy, S. N. (1953). "On a heuristic method of test construction and its use in multivariate analysis". In: *Annals of Mathematical Statistics* 24, pp. 220–238.

S., Stern H. (1992). "Who's Number One? Rating Football Teams". In: *Proceedings of the Section on Statistics in Sports, American Statistical Association*, pp. 1–6.

— (1995). "Who's Number One in College Football? And How Might We Decide?" In: *Chance* 8, pp. 7–14.

Shea, S. (2014). *Basketball Analytics: Spatial Tracking*. Stephen Shea.

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

Stefani, R. T. (1980). “Improved Least Squares Football, Basketball, and Soccer Predictions”. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-10, pp. 116–123.

Vracar, P., E. Štrumbelj, and I. Kononenko (2016). “Modeling basketball play-by-play data”. In: *Expert Systems With Applications* 44, 58–66.

Ringraziamenti

Vorrei ringraziare in primis la professoressa Laura Ventura che ha reso possibile questo lavoro di tesi e per la disponibilità offerta.

Ringrazio i professori Luigi Salmaso e Livio Corain per il supporto, la pazienza, i consigli che mi hanno dato durante questi mesi.

Alla dottoressa Ilaria Bussoli va la mia ammirazione, i miei ringraziamenti più sentiti e tutte le coccole del mondo perchè si. Perchè è la più bellissimissima, intelligentissima e bravissimissimissima orsacchiotta che c'è. Inoltre perchè, senza di lei, non dico sarei in alto mare con la tesi, ma proprio starei ancora studiando statistica progredito.

A mamma, papone e Dani che mi vogliono bene nonostante io sia io e mi hanno supportato in questi anni di magistrale (e pure prima) in modo amorevole (fortunatamente per me anche economico), dedico questa tesi perchè, anche se solo una fotografia di quello che ho imparato fino ad ora, è merito loro se ho potuto imparare tutte queste cose da principio.

Ora la lista inizia a diventare veramente lunga, di questo mi sento molto fortunato, di persone che, tra Padova e Bari, dal Perù all'Australia, vorrei ringraziare. Spero possa bastare in questa sede CasaPorto, CasaNave-CasaFaggin, VillaFowst, CasaBari, il Locale e tutte le generazioni di amici, abitanti, affini, abusivi, concubini, giocolieri, musicisti e fricchettoni che si sono susseguite e i miei compari della sala studio senza i quali il mio sog-

giorno a Padova sarebbe stato un'esperienza vuota e poco istruttiva, al pari di dare esami (*coff coff*).

Sono sicuro che nessuna delle persone incluse in questi gruppi preferisca una citazione in un documento che non leggerà nessuno piuttosto che un abbraccione orsacchiottone e una bella birra o un goto de vin. Sono anche positivo sul fatto che a questo punto in molti staranno già bevendo o saranno già brilli, quindi non leggeranno neanche queste parole. Prosit.

Un grazie finale a Riccardo Zampinetti, che ha fatto un faticosissimo e bel lavoro, che mi ha dato una grossa mano durante questa tesi.