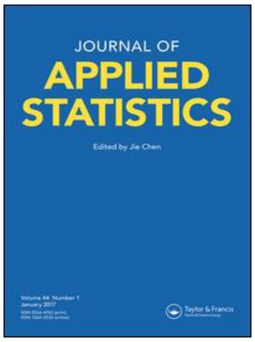

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Measuring sport performances under pressure by classification trees with application to basketball shooting

Rodolfo Metulini & Mael Le Carre

To cite this article: Rodolfo Metulini & Mael Le Carre (2019): Measuring sport performances under pressure by classification trees with application to basketball shooting, Journal of Applied Statistics, DOI: 10.1080/02664763.2019.1704702

To link to this article: https://doi.org/10.1080/02664763.2019.1704702

Published online: 19 Dec 2019.

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2019.1704702

Taylor & Francis
Taylor & Francis Group

Measuring sport performances under pressure by classification trees with application to basketball shooting

Rodolfo Metulini [a] and Mael Le Carre [b]

[a] Department of Economics and Statistics, University of Salerno, Fisciano, Italy; [b] Big & Open Data Innovation Laboratory (BODal-Lab) - Department of Economics and Management, University of Brescia, Brescia, Italy

ABSTRACT

Measuring players' performance in team sports is fundamental since managers need to evaluate players with respect to the ability to score during crucial moments of the game. Using Classification and Regression Trees (CART) and play-by-play basketball data, we estimate the probabilities to score the shot with respect to a selection of game covariates related to game pressure. We use scoring probabilities to develop a player-specific shooting performance index that takes into account for the difficulty associated to score different types of shots. By applying this procedure to a large sample of 2016–2017 Basketball Champions League (BCL) and 2017–2018 National Basketball Association (NBA) games, we compare the factors affecting shooting performance in Europe and in the United States and we evaluate a selection of players in terms of the proposed shooting performance index with the final aim of providing useful guidelines for the team strategy.

ARTICLE HISTORY
Received 10 April 2019
Accepted 7 December 2019

KEYWORDS
Classification trees; basketball; play-by-play; performance evaluation; strategical decisions

1. Introduction

Measuring the performance of a player in team sports has recently gained interest in both the academic and business worlds. Managers face the need to optimize their money by adopting a cost/benefit approach [36]. In their favor, the possibility to retrieve useful information from online platforms is rapidly increasing, thanks to improvements in storage capacity and computing power. In the era of *Money Ball Thinking* [16], scientific literature covered several aspects of research about basketball, ranging from delineating the main characteristics of a game by means of basic descriptive statistics [15] to exploring complex problems, such as forecasting the winning of a game or a tournament [3,11,18,19,23,34,38,39] or identifying the best game strategy [1,24].

The performance of the single player has been evaluated by departing from just using traditional statistics such as points, steals, turnovers. A large range of statistical and machine learning techniques [4,7,9,30–32] has been recently applied. Players must also be evaluated in a more comprehensive way that considers how players cooperate together

CONTACT Rodolfo Metulini ✉ rodol.metulini@gmail.com Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II132, Fisciano SA 84084, Italy

by looking at their position in the court [25–28] rather than at the relation with the performance of their team-mates [22].

Probably the first contribution on offensive performance of single players, proposed by Oliver [29], is called “Offensive Rating” and captures the number of points scored by a player per 100 possessions. Hollinger [13] improves Offensive Rating by returning a per-minute rating of players performance.

More recently, thanks to the increasing availability of play-by-play data, offensive performance of single players turned to be analyzed in terms of shooting percentage. Examples are Csataljay et al. [6], who studied the role of field goal (FG) percentage on the outcome of the game by using 2007 European Basketball Championship data; Piette et al. [33], who applied a new offensive performance measure to 2004–2005 NBA data and Zuccolotto et al. [40], who employed a measure accounting for high-pressure situations to 2016 Olympic Games basketball tournament.

In this paper, employing play-by-play data from 1230 games of the 2017–2018 NBA regular season and from 324 games of the 2016–2017 BCL tournament, we develop a shooting performance index for players that accounts for the different difficulties of the shots and for the fact that players may react differently to game situations that are different according to the level of pressure [5,20,21].

The index is based on estimating, by means of a classification tree, the scoring probability of each type of shot. Made/missed-shot dichotomous variable is the dependent of the classification tree, while a selection of variables related to the type of shot and to the moment in which the shot is attempted are considered as covariates. Scoring probabilities are then used to develop the performance index.

When applied to real case studies, a comparison of the variables affecting BCL shooting performance with those in the NBA league and a visual comparison of a sample of selected players in terms of our performance index are proposed.

Evaluating players’ performance in terms of shooting ability presents some limitations, in fact, a comprehensive assessment should take into account different player’s skills. A player may be bad in shooting but its overall performance may be considered good because, for example, he is good at passing the ball or at taking offensive rebounds. Therefore, this work allows an evaluation of the shooting performance, while it does not provide coaches and managers with information for a global player’s assessment.

The structure of this paper is as follows. Section 2 enters into the details of our methodological strategy for developing the performance index, Section 3 presents the application to case studies and Section 4 is dedicated to the discussion of results. Section 5 concludes.

2. Methodology

2.1. Scoring probabilities using classification trees

Let $shot_{ij}$ be the dichotomous variable assuming value 1 if player i made shot j, 0 otherwise,

$$shot_{ij} = \begin{cases} 1 & \text{if player } i \text{ made shot } j \\ 0 & \text{if player } i \text{ missed shot } j. \end{cases}$$

To model scoring probabilities, we use a classification tree [2] with $shot_{ij}$ as the dependent variable.

Similarly to us, random forests are used to estimate the probability to win football games [17,35] while decision trees are used to model the National Collegiate Athletic Association (NCAA) selection process [8].

Classification trees belong to the more general methodology which goes with the name of Classification and Regression Trees (CART) and it is generally adopted when the dependent variable is categorical. Classification trees have some advantages in comparison with alternative parametric models for categorical data, such as the logistic regression model [14]. Logistic regression model accounts, in a single predictive formula, the entire set of variables and the full sample; when the data have lots of features that interact each others, estimating a single global model can be difficult and confusing. With classification trees, we partition the sample into smaller regions, where the interactions are more manageable. Therefore, we consider classification trees the most suitable method to our case.

We assume that the probability of scoring the shot may depends on several aspects directly related to the type of shot or to particular moments of the match.

For the choice of covariates, we also consider factors related to game pressure, assuming that the reaction to difficult situations is different among players. The definition in Goldman & Rao [10] quantified pressure as a measure of the marginal impact of an additional point as a function of score margin and time remaining. Tango et al. [37] defined high-pressure, those game situations in which the result being on the line.

In light of these definitions, we identify high-pressure situations those being more demanding and troublesome compared to a normal situation. According to the definition that moments can be classified as it follows: (a) the shot clock is going to expire; (b) the score margin between the two teams is close to zero; (c) the team has performed poorly up to that particular moment in the game; (d) the player has missed his previous shot.

Let:

- $shot.clock_{ij}$ be a numerical variable with valid values in the range [0,24] denoting the time (in seconds) on the shot-clock when $shot_{ij}$ has been attempted;
- $sc.diff_{\ ij}$ be a numerical variable denoting the score difference (in points) with respect to the opponents when $shot_{ij}$ has been attempted;
- $miss.t_{ij}$ be a numerical variable with valid values in the range [0, 1], denoting the ratio of missed shots for the whole team in the match when $shot_{ij}$ has been attempted;
- $miss.pl_{ij}$ be a categorical variable assuming value 1 if the previous shot ($j-1$) by the same player in the same match ($miss.pl_{ij-1}$) has been made, 0 otherwise,

$$miss.pl_{ij} = \begin{cases} 1 & \text{if } miss.pl_{ij-1} = 1 \\ 0 & \text{otherwise;} \end{cases}$$

- $shot.type_{ij}$ be a categorical variable with categories $2P$, $3P$ and FT (2-point, 3-point shots and free throws, respectively),

$$shot.type_{ij} = \begin{cases} \text{2P} & \text{if } shot_{ij} \text{ is a 2-point shot} \\ \text{3P} & \text{if } shot_{ij} \text{ is a 3-point shot} \\ \text{FT} & \text{if } shot_{ij} \text{ is a free-throw shot;} \end{cases}$$

- $time_{ij}$ be a numerical variable denoting the time (in seconds) to the end of the quarter when $shot_{ij}$ has been attempted;

- $poss.type_{ij}$ be a categorical variable assuming category *original* if $shot_{ij}$ has been attempted during the original 24 seconds on the shot clock, assuming category *reset* if $shot_{ij}$ has been attempted after the shot clock has been reset to additional 14 seconds,

$$poss.type_{ij} = \begin{cases} \text{original} & \text{if } shot_{ij} \text{ has been attempted during the original 24 seconds} \\ \text{reset} & \text{if } shot_{ij} \text{ has been attempted after the reset of the shot clock;} \end{cases}$$

- $quarter_{ij}$ be a categorical variable indicating the game quarter in which $shot_{ij}$ has been attempted,

$$quarter_{ij} = \begin{cases} 1 & \text{if} shot_{ij} \text{ has been attempted in the first quarter} \\ 2 & \text{if } shot_{ij} \text{ has been attempted in the second quarter} \\ 3 & \text{if } shot_{ij} \text{ has been attempted in the third quarter} \\ 4 & \text{if } shot_{ij} \text{ has been attempted in the fourth quarter} \\ 5 & \text{if } shot_{ij} \text{ has been attempted in overtime.} \end{cases}$$

We use $shot.clock_{ij}$, $sc.diff_{\ ij}$, $miss.t_{ij}$, $miss.pl_{ij}$, $shot.type_{ij}$, $time_{ij}$, $poss.type_{ij}$ and $quarter_{ij}$ as covariates for the classification tree.

To control for the tree instability (e.g. to prevent trees from growing too deeply) and to improve the interpretability of results, a conversion of numerical covariates into categorical is applied.

Converting numerical covariates deals with defining proper thresholds for the categories. To this scope, Thresholds Important Measure (TIM) diagnostic is adopted [40]. TIM aims to choose the thresholds by looking to the total decrease of heterogeneity of the response variable when the feature space is partitioned recursively. The tree splits shots on a particular threshold of the selected split variable: TIM considers all of the thresholds used in the tree to split nodes, then it sums up all of the decreases in the heterogeneity index allowed by each threshold of each covariate.

Let the following model be the classification tree in which all the original covariates are included:

$$\begin{aligned} shot_{ij} \sim\ & shot.clock_{ij}, sc.diff_{ij}, miss.t_{ij}, miss.pl_{ij}, shot.type_{ij}, time_{ij}, \\ & poss.type_{ij}, quarter_{ij}. \end{aligned} \tag{1}$$

To estimate scoring probabilities, we proceed as follows:

(1) TIM diagnostic is applied on the tree defined in Equation (1) to choose the best categorization for the new categorical covariates $time.C_{ij}$, $sc.diff.C_{ij}$, $miss.t.C_{ij}$ and $shot.clock.C_{ij}$;
(2) the following classification tree with all categorical covariates is performed:

$$\begin{aligned} shot_{ij} \sim\ & shot.clock.C_{ij}, sc.diff.C_{ij}, miss.t.C_{ij}, miss.pl_{ij}, shot.type_{ij}, \\ & time.C_{ij}, poss.type_{ij}, quarter_{ij}. \end{aligned} \tag{2}$$

The outcome of this procedure is represented by different estimated scoring probabilities ϕ_j associated to different types of shots.

For example, a 3-point shot attempted in the last minute of a quarter ($shot.type_{ij} = 3P$ and $time_{ij} <= 60$) by a generic player i is expected to has a lower scoring probability compared to a 2-point shot attempted before the last minute begin ($shot.type_{ij} = 2P$ and $time_{ij} > 60$).

Since no player-specific covariates are included in the model, scoring probabilities do not vary along index i.

2.2. The shooting performance index

A shooting performance index for each player is developed here by taking into consideration that different shots present different scoring probabilities (namely, different difficulties).

The (obvious) idea is that it is ‘easy’ to score easy shots while it is ‘difficult’ to score difficult shots. So, we do not want to give to each attempted shot the same importance. The performance index is computed by normalizing the importance of each attempted shot. The index gives more importance to the made shots to which a low scoring probability is associated (difficult shots) and it gives less importance to the made shots to which an high scoring probability is associated (easy shots). Conversely, the index penalised less the missed shots to which a low scoring probability is associated and it penalised more the missed shots to which an high scoring probability is associated.

For each category T of the variable $shot.type_{ij}$ (2P, 3P, FT), let J_T be the set of attempted shots of type T, having different probabilities to be scored. Let denote with x_{ij} the indicator assuming value 1 if shot j of player i scored a basket and 0 otherwise and with ϕ_j its scoring probability according to the classification tree.

For each shot, the difference $x_{ij} - \phi_j$ can be used as a performance measure of the shot. In fact, $x_{ij} - \phi_j$ always assumes a positive value if the shot j of player i scored a basket and it always assumes a negative value if the shot is missed. Moreover, for made shots, the smaller (the larger) is ϕ_j the large (the small) is $x_{ij} - \phi_j$, the large (the small) is the positive contribution of that shot to the performance index. Likewise, for missed shots, the smaller (the larger) is ϕ_j the small (the large) is $x_{ij} - \phi_j$, the small (the large) is the negative contribution of that shot to the performance index.

Taking this logic into consideration, the shooting performance index of player i for shot type T is defined as:

$$P_i(T) = \underset{j \in J_T}{\text{avg}}(x_{ij} - \phi_j). \tag{3}$$

For $T = 2\text{P}$, the value $P_i(T)$ assumes positive values if the player i along all the considered matches performed better at 2-point shooting compared to the average player, negative otherwise. The same holds for T = 3P and for T = FT.

The performance index is computed for a selection of players having a number of shots above a threshold (which is defined “ad-hoc” in each case study) and for 2-point shots ($T = 2\text{P}$), 3-point shots ($T = 3\text{P}$) and free-throws ($T = \text{FT}$).

3. Application

We apply the proposed methodological strategy to BCL and NBA case studies.

BCL is an annual professional basketball competition for European clubs, organized by International Basketball Federation (FIBA), existing since 2015. Among European club competitions, it is a rival to the EuroCup, at a level below the EuroLeague.

The NBA is a professional basketball league in the US composed of 30 teams (29 in the US and 1 in Canada). It is widely considered to be the premier professional basketball league in the world.

FIBA[1] and NBA[2] rules have a lot in common, but there are some differences that might affect shooting performance. One is the dimension of the court —28 by 15 m for BCL and 94 by 50 feet (corresponding to 28.7 by 15.2 m) for NBA – that may have an indirect impact on shooting performance. Another difference is the duration of the game – according to BCL is 40 min (divided into four periods of 10 min each), according to NBA is 48 (four periods of 12 min each) – that is directly connected to our covariate $time_{ij}$.

3.1. Data

Information on whether each shot was made or missed and the additional information on the type of the shot and to the moment in which the shot is attempted have been extracted from the play-by-play of the games.

BCL data for the season 2016–2017 has been web-scraped and processed from the play-by-play available online at http://www.championsleague.basketball/16-17. The total number of shots stands to 47,849. Hence, the sample is large enough to guarantee robust estimation of the scoring probabilities using the classification tree.

NBA data have been made available thanks to a friendly agreement with BigDataBall Company (UK) (www.bigdataball.com). BigDataBall collected the play-by-play of all the NBA games (both regular season and play-off) from the year 2004–2005 to the year 2017–2018, for a total of 14 seasons. In this application data from the latest regular season (2017–2018) are used to avoid some missing games occurring in previous seasons. The total number of shots in season 2017–2018 stands to 265,008.

3.2. ‘Basketball champions league’ case study

Scoring probabilities with classification trees

We perform the classification tree in Formula 1. To control for the dimension of final nodes, we set the parameter related to the minimum number of shots in each node to 0.8% of the total number of shots in the sample[3].

In order to categorize quantitative covariates ($shot.clock_{ij}$,$sc.diff_{\ ij}$, $time_{ij}$, $miss.t_{ij}$), we apply TIM diagnostic. Results are reported in Figure 1. After a careful visual inspection, we define the thresholds for the numerical variables as listed below (s is for seconds, p is for points, % is for percentage):

$$shot.clock.C_{ij} = \begin{cases} \text{time end,} & \text{if } shot.clock_{ij} \leq 5\text{s} \\ \text{middle end} & \text{if } 6\text{s} \leq shot.clock_{ij} \leq 10\text{s} \\ \text{early middle} & \text{if } 11\text{s} \leq shot.clock_{ij} \leq 16\text{s} \\ \text{early} & \text{if } shot.clock_{ij} \geq 17\text{s.} \end{cases}$$

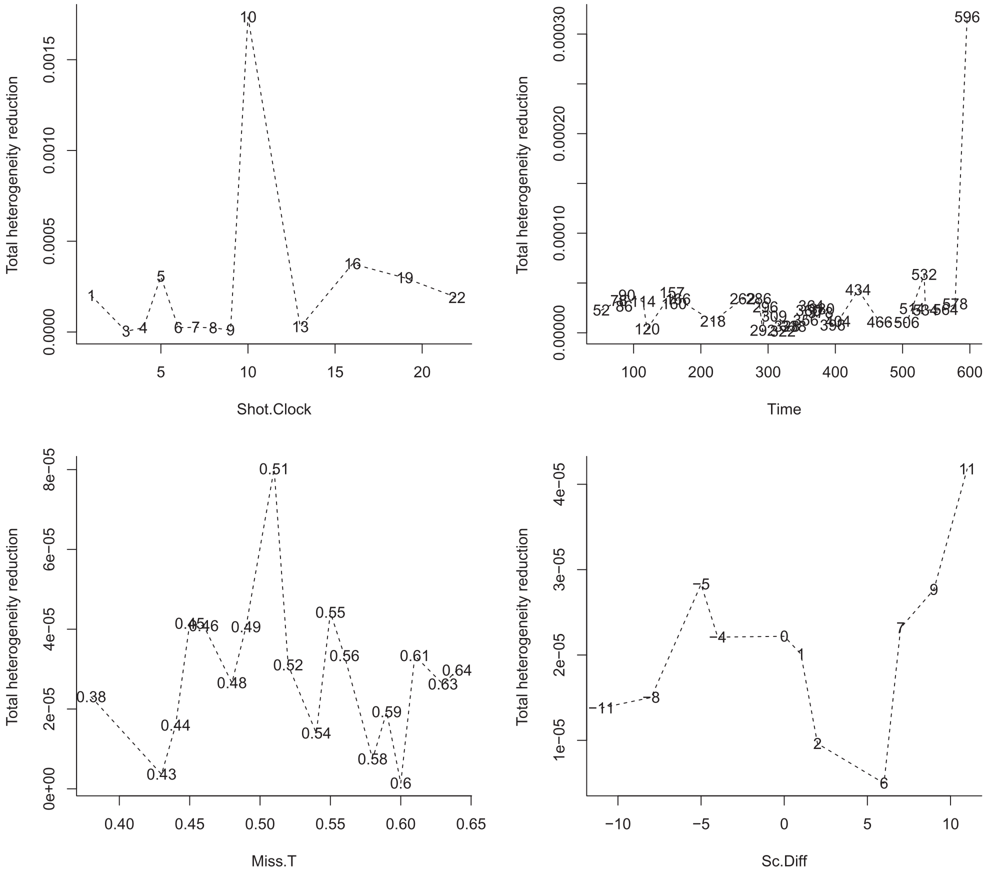

Figure 1. TIM diagnostic for the choice of thresholds for the numerical covariates. BCL data.

$$sc.diff.C_{ij} = \begin{cases} -5 \text{ or less}, & \text{if } sc.diff_{ij} \leq -5\text{p} \\ [-4, 0] & \text{if } -4\text{p} \leq sc.diff_{ij} \leq 0\text{p} \\ [1, 7] & \text{if } +1\text{p} \leq sc.diff_{ij} \leq +7\text{p} \\ [8, 11] & \text{if } +8\text{p} \leq sc.diff_{ij} \leq +11\text{p} \\ +12 \text{ or more} & \text{if } sc.diff_{ij} \geq +12\text{p}; \end{cases}$$

$$time.C_{ij} = \begin{cases} \text{time end} & \text{if } time_{ij} \leq 99\text{s} \\ \text{normal} & \text{if } time_{ij} \geq 100\text{s}; \end{cases}$$

$$miss.t.C_{ij} = \begin{cases} \text{bad} & \text{if } miss.t_{ij} \leq 0.45\% \\ \text{medium} & \text{if } 0.45\% \leq miss.t_{ij} \leq 0.51\% \\ \text{good} & \text{if } miss.t_{ij} \geq 0.51\% \end{cases}$$

Having defined the new categorical covariates, we estimate the classification tree in Formula 2 with all categorical covariates. We ‘prune’ the tree in order to obtain a small number of final nodes and to improve its interpretability. As a robustness check, we also

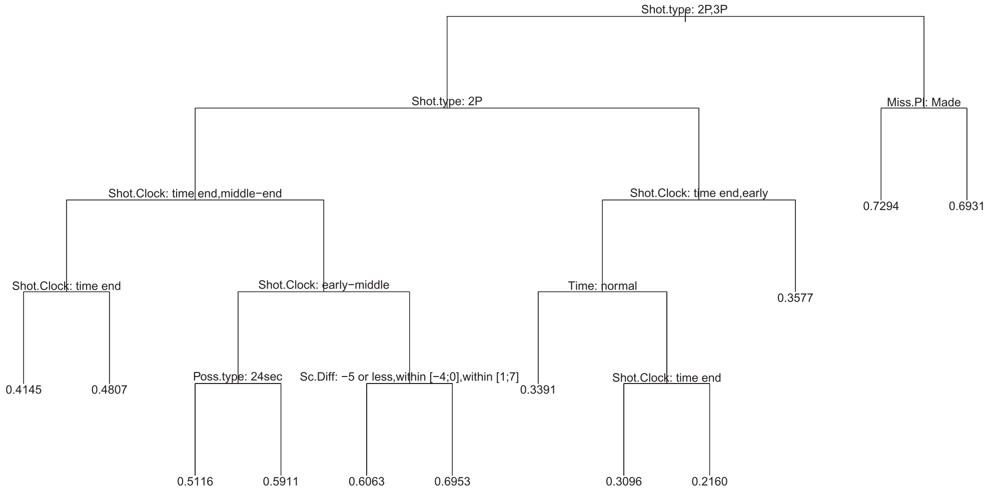

Figure 2. Classification tree with *shot*$_{ij}$ on *shot.clock*$_{ij}$, *sc.diff* $_{ij}$, *miss.t*$_{ij}$, *miss.pl*$_{ij}$, *shot.type*$_{ij}$, *time*$_{ij}$, *poss.type*$_{ij}$, *quarter*$_{ij}$. BCL data.

estimate a tree without pruning the nodes and then we compare the two (the one with the pruning, having 12 final nodes, and the other without pruning, consisting of about 180 final nodes) in terms of Receiving Operation Characteristic (ROC) curves [12]. The Area Under the Curve (AUC) is 0.6617 for the first and 0.7084 for the second. This difference is marginal and proves that, considering 12 nodes, we do not lose relevant information.

Most relevant results from the classification tree reported in Figure 2 are that the first level splits 2P and 3P shots against FT shots. At the second level, 2P and 3P shots are splitted, while another split separates FT shots where the previous FT was made by those where the previous was missed. FTs where the previous FT was made are scored with a probability of 0.7294; when the previous FT was missed, shots are scored with a probability of 0.6931. At the third level, 2P shots are splitted on time end and middle end versus other shot-clock moments. Those moments are further splitted at the fourth level in time end versus middle end. Time end 2P shots are scored with a probability of 0.4145. Middle end 2P shots are scored with a probability of 0.4807. Moreover, at the fourth level, early and early middle 2P shots are splitted. Early middle 2P shots are scored with a probability of 0.5116 during the regular 24 s, with a probability of 0.5911 if they have been attempted during the additional 14 s. Early 2P shots where the scoring difference is less than + 7 are scored with a probability of 0.6063. Early 2P shots where the scoring difference is more than +7 are scored with a probability of 0.6953. 3P shots split, in the third level, into early and time end shots versus middle end and early middle. In the fourth level, early and time end 3P shots split into normal shots and time end shots. Normal shots are scored with a probability of 0.3391. 3P shots during normal time are scored with a probability of 0.3096 if they are attempted at the end of the action (time end); with a probability of 0.2160 when they are attempted early during the action. 3P shot attempted on middle end and early middle has been scored with a probability of 0.3577.

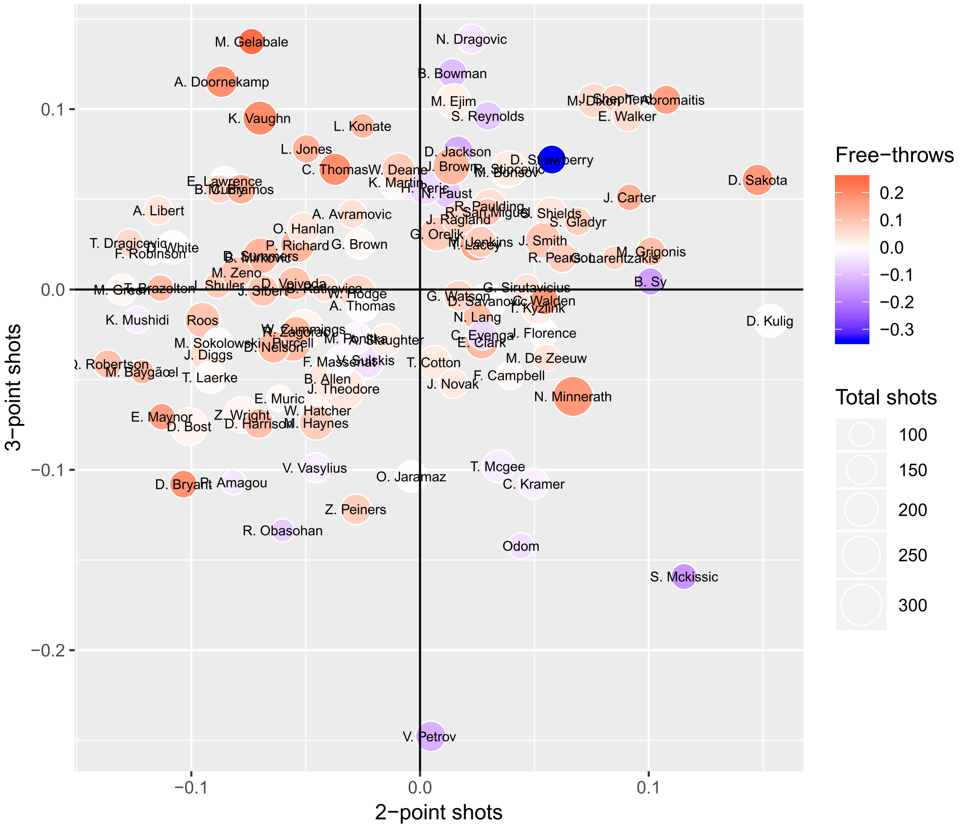

Figure 3. Scatter-plot displaying the shooting performance index for a selection of players. $P_i(T = 2P)$ (x-axis), $P_i(T = 3P)$ (y-axis), $P_i(T = FT)$ (bubbles' colors). BCL data.

The performance index

We compute the player performance index by considering all the players who attempted, summing up all the matches of the season, at least 25 shots in each category of *shot.type*$_{ij}$ variable (i.e. $|J_T| \geq 25$, $\forall\ T = \{2P, 3P, FT\}$, where $|\cdot|$ denotes cardinality). We compute the value of $P_i(T)$ for all these players and we report their performance values in a two-dimensional scatter-plot with $P_i(T = 2P)$ in the x-axis, $P_i(T = 3P)$ in the y-axis and $P_i(T = FT)$ displayed through the bubbles' colors. The dimension of the bubbles represents the number of total shots.

The scatter-plot is displayed in Figure 3. This visual strategy allows to evaluate players in terms of their shooting performance considering 2 points, 3 points and free-throws together. Players are positioned in one of four quadrants of the plot. Players on the top right quadrant are those who perform better than league average in 2 points and 3 points shots. Bottom left quadrant includes players who poorly perform in 2 points and 3 points shots. The other two quadrants represent intermediate situations.

To report some examples, top-right quadrant includes, among others, D. Sakota, J. Carter, E. Walker and T. Abromaitis. These players shot better than the average from both 2 points and 3 points. D. Strawberry also performs above the average on 2 points and 3 points shots, however he poorly performs on free-throws.

Bottom-left quadrant includes, among others, R. Obasohan, P. Amagou, E. Maynor and D. Bryant. These players perform below the average on both 2 points and 3 points shots.

M. Gelabale, A. Doornekamp and K. Vaughn and others are placed on the top-left quadrant, as they better perform on 3 points shots, but they perform below the average on 2 points shots. S. McKissic, N. Minnerath and D. Odom and others are placed on the bottom-right quadrant, as they better perform on 2 points shots and they perform below the average on 3 points shots.

3.3. 'National basketball association' case study

Scoring probabilities with classification trees

As for the BCL case study, we perform the classification tree in Formula 1 on NBA data by controlling for the dimension of final nodes.

Actually, we do not use $miss.t_{ij}$ because, considering the way NBA data were recorded by BigDataBall, it was not possible to correctly count the number of missed shots by the team at the moment of the shot[4].

To categorize numerical covariates, we applied TIM diagnostic. According to the results reported in Figure 4, we define the thresholds for the numerical covariates as listed below:

$$shot.clock.C_{ij} = \begin{cases} \text{time end,} & \text{if } shot.clock_{ij} \leq 5\text{s} \\ \text{middle end} & \text{if } 6\text{s} \leq shot.clock_{ij} \leq 13\text{s} \\ \text{early middle} & \text{if } 14\text{s} \leq shot.clock_{ij} \leq 19\text{s} \\ \text{early} & \text{if } shot.clock_{ij} \geq 20\text{s}. \end{cases}$$

$$sc.diff.C_{ij} = \begin{cases} -6 \text{ or less,} & \text{if } sc.diff_{ij} \leq -6\text{p} \\ [-5,-2] & \text{if } -5\text{p} \leq sc.diff_{ij} \leq -2\text{p} \\ [-1,1] & \text{if } -1\text{p} \leq sc.diff_{ij} \leq +1\text{p} \\ [2,4] & \text{if } +2\text{p} \leq sc.diff_{ij} \leq +4\text{p} \\ [5,19] & \text{if } +5\text{p} \leq sc.diff_{ij} \leq +19\text{p} \\ +20 \text{ or more} & \text{if } sc.diff_{ij} \geq +20\text{p}; \end{cases}$$

$$time.C_{ij} = \begin{cases} \text{time end} & \text{if } time_{ij} \leq 99\text{s} \\ \text{normal} & \text{if } time_{ij} \geq 100\text{s}; \end{cases}$$

Having defined the new covariates, we estimate the classification tree with all categorical covariates. We prune the tree and we obtain 11 final nodes. As a robustness check, we estimate a tree without pruning the nodes (consisting of more than 200 final nodes) and then we compare the two trees. The difference between the AUC is marginal (it is 0.658 with 11 nodes and 0.6638 without pruning) and it proves that by considering 11 nodes, we are not losing relevant information.

The 11 nodes final tree applied to NBA data are represented in Figure 5. The first level splits 2P and 3P shots versus FT shots. At the second level, 2P and 3P shots are splitted, and another split separates FT shots where the previous FT was scored by those where the

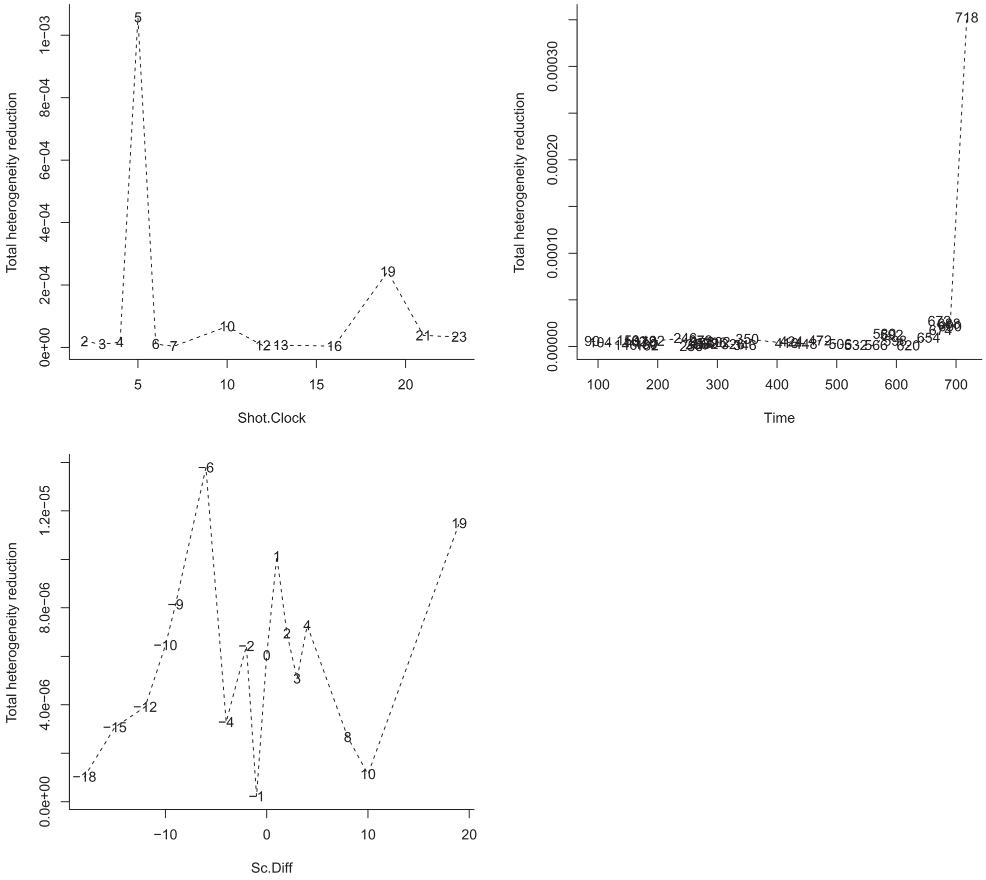

Figure 4. TIM diagnostic for the choice of thresholds for the numerical covariates. NBA data.

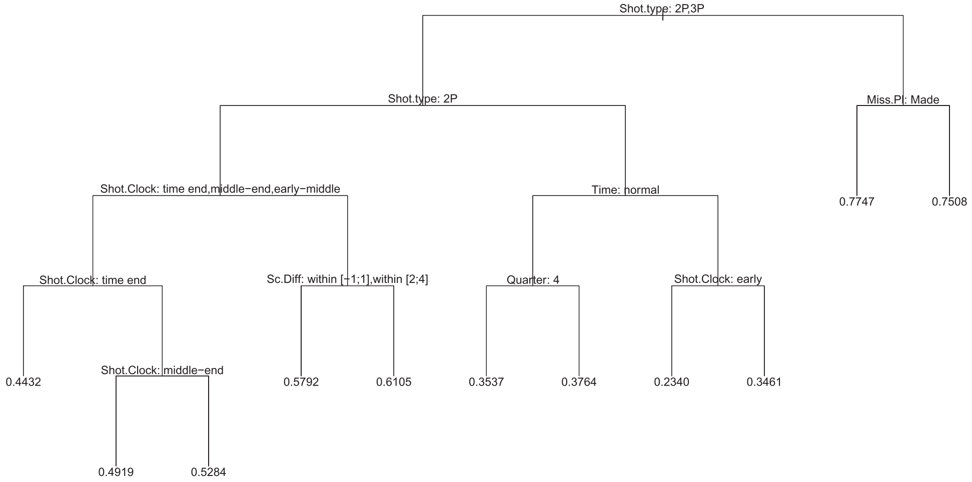

Figure 5. Classification tree with *shot*$_{ij}$ on *shot.clock*$_{ij}$, *sc.diff* $_{ij}$, *miss.t*$_{ij}$, *miss.pl*$_{ij}$, *shot.type*$_{ij}$, *time*$_{ij}$, *poss.type*$_{ij}$, *quarter*$_{ij}$. NBA data.

previous FT was missed. FTs where the previous FT was made are scored with a probability of 0.7747; where the previous FT was missed, FT are scored with a probability of 0.7508. At the third level, 2P shots are splitted on early and other shot-clock moments. Those shots are further splitted at the fourth level in time end shots versus middle end and early middle shots. Time end 2P shots are scored with a probability of 0.4432. Middle end 2P shots are scored with a probability of 0.4919. Early middle 2P shots are scored with a probability of 0.5284. Moreover, at the fourth level, early 2P shots are splitted in [−2,4] shots versus other early 2P shots. Early 2P shots where the scoring difference was in [−2,4] are scored with a probability of 0.5792. Early 2P shots where the scoring difference is not in [−2,4] are scored with a probability of 0.6105. 3P shots split, at the third level, into normal and quarter end shots. At the fourth level, normal 3P shots split into fourth quarter shots and other shots. Normal 3P shots attempted in the fourth quarter are scored with a probability of 0.3537. Normal 3P shots attempted in other quarters are made with a probability of 0.3764. Time end 3P shots split, at the fourth level, into early versus other shots. Time end 3P shots attempted early are scored with a probability of 0.2340. Time end 3P shots attempted at middle or time end are scored with a probability of 0.3461.

The performance index

We compute $P_i(T)$ for all the players who, summing up all the matches of the season, attempted at least 300 shots in each category of *shot.type*$_{ij}$ variable ($|J_T| \geq 300, \forall\ T = \{2P, 3P, FT\}$). We report the performance values in a 2 dimensional scatter-plot with $P_i(T = 2P)$ in the x-axis, $P_i(T = 3P)$ in the y-axis and $P_i(T = FT)$ displayed through the bubbles' colors (Figure 6).

Top-right quadrant includes, among others, S. Curry, K. Durant, K. Irving, K. A. Towns and C. Paul. They perform better than the average in 2 points and 3 points shots. According to the general agreements of basketball experts, they are all considered *all-star* players. L. James, another all-star player, enters the quadrant; however, he is very close to the average in terms of 3 points performance and he poorly performs on free-throws shots.

Bottom-left quadrant includes, among others, D. Fox, D. Smith Jr., D. Wade and J. Jackson. These players perform below the average in 2 points and 3 points shots. Most of these players are on their first year in the league.

J. Tatum, P. George and CJ McCollum belong to top-left quadrant. They better perform in 3 points shots and they perform lower than the average in 2 points shots.

J. Embid, G. Antetokounmpo, D. Green and A. Davis belong to bottom-right quadrant. These players better perform than the average in 2 points shots and they are lower than the average in 3 points shots.

4. Discussion

Measuring players' shooting performance is one of the most relevant issue in team sport since, to win a match, the team needs to score one basket more than its opponent in crucial moments. In this regard, the manager, the coach and the staff need to cooperate in order to define the best strategy to increase the shooting performance of their team. It should go without saying that two point shots have higher probabilities to be scored than 3-point shots. Our procedure based on classification trees statistically confirms this

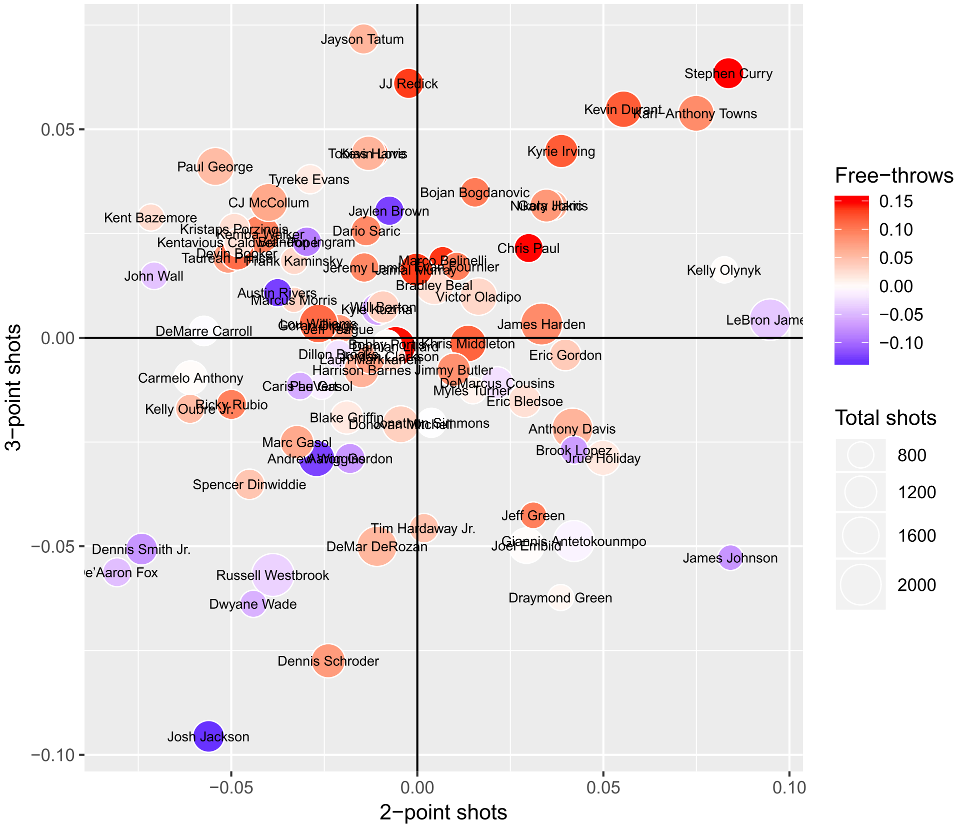

Figure 6. Scatter-plot displaying the shooting performance index for a selection of players. $P_i(T = 2P)$ (x-axis), $P_i(T = 3P)$ (y-axis), $P_i(T = FT)$ (bubbles' colors). NBA data.

evidence and goes more deeply on finding less known evidences about the factors affecting shooting performance, highlighting the differences among European basketball and North American basketball.

In both BCL and NBA, the moment of the action when the shot is attempted (variable *shot.clock.*C_{ij}) has an effect on the probability of scoring the shot. More in detail, 2-point shots at the end of the action (last 5 seconds) present a lower scoring probability compared to 2-point shots in the middle or during the first seconds of the action. Another relevant factor is the scoring difference. A 2-point shot has a lower probability to be scored when the two teams are close each others in the score; this effect appears stronger in NBA compared to BCL. In both NBA and BCL, 3-point shots attempted on the last seconds of the quarter (variable *time.*C_{ij}) present a lower probability to be scored compared to a 3-point shot in normal conditions. Moreover, the last quarter of the game affects the shooting performance in NBA. 3-point shots present a lower probability to be scored in the fourth quarter compared to a 3-point shot attempted during the first three quarters. The type of possession (variable *poss.type*$_{ij}$) has a significant impact on shooting performance only in BCL. More precisely, a 2-point shot attempted after the reset of the time has an higher probability to be scored compared to a 2-point shot attempted during the original 24 s.

Staff should also evaluate single players in the roster. In fact, the shooting performance varies among players. Moreover, the same player may perform differently in different circumstances: one better performs in normal situations while another one better performs in the crucial moments of the game. By developing a shooting performance index for each category of shot type taking into consideration that some shots are more difficult than others, we group players with similar shooting performance.

We define four quadrants. The first quadrant includes the players that perform better than the average in both 2-point and 3-point shots. We recommend managers and coaches to let handle the ball to these players in the crucial moments of the game. A second quadrant includes the players that perform lower than the average in both 2-point and 3-point shots. Players here are often on their first year in the league (the so-called *rookies*). In case those players are not rookies, the manager should think about trade those players or give them less shooting responsibility. The third and the fourth quadrants help the staff to detect those players that should handle the ball, respectively, when the playing strategy opts for a 2-point shot or when the playing strategy opts for a 3-point shot.

5. Conclusions

Data analytics in basketball, likewise in other professional sports, is increasingly used. More and more, managers of professional teams face the need of monitoring the performance of their team and their players. In the era of big data online platforms make available live streams of data to extract useful information, so, from the moment when staff have to choose players to create the roster in summer to the moment when it has to evaluate them playing together, players are constantly monitored. Shooting performance is one of the most important evaluation criteria.

In this work, we proposed a methodological strategy that allows, by selecting proper covariates related to the type of the shot and to the moment in which the shot is attempted, to estimate, with a classification tree, the scoring probabilities of different types of shots and to develop, using that scoring probabilities, a new player performance index that gives merit to players who perform better during difficult game circumstances.

This work can be seen as an agile and easy-to-interpret instrument that (i) by defining the game situations that mostly affect players' shooting performance, (ii) by estimating the scoring probabilities related to those game situations and (iii) by developing a player's performance index that gives merit to whom score difficult shots, it targets to help managers, coaches and the staff that, by adapting the playing strategies accordingly, may improve the performance of the team by increasing their knowledge on the performance of single players.

Notes

[1] http://www.fiba.basketball/documents

[2] https://official.nba.com/rulebook/

[3] According to the value chosen in [40].

[4] As a matter of fact, variable $miss.t_{ij}$ has no relevant impact on BCL case study and we believe that ignoring this variable does not alter the results in this case study.

Acknowledgments

The authors would like to thank Paola Zuccolotto, Marica Manisera and Serhat Ugur for the valuable contribution.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Research carried out in collaboration with the Big&Open Data Innovation Laboratory (BODaI-Lab), University of Brescia (project nr. 03-2016), title Big Data Analytics in Sports (http://bodai.unibs.it/bdsports/index.htm), granted by Fondazione Cariplo and Regione Lombardia.

ORCID

Rodolfo Metulini http://orcid.org/0000-0002-9575-5136

References

[1] D.H. Annis, *Optimal end-game strategy in basketball*, J. Quant. Anal. Sports 2 (2006), pp. 1–9.
[2] L. Breiman, *Classification and Regression Trees*, Routledge, New York, 2017.
[3] M. Brown and J. Sokol, *An improved LRMC method for NCAA basketball prediction*, J. Quant. Anal. Sports 6 (2010), pp. 1–23.
[4] W.W. Cooper, J.L. Ruiz, and I. Sirvent, *Selecting non-zero weights to evaluate effectiveness of basketball players with DEA*, Eur. J. Oper. Res. 195 (2009), pp. 563–574.
[5] P.R. Crocker and T.R. Graham, *Coping by competitive athletes with performance stress: gender differences and relationships with affect*, Sport. Psychol. 9 (1995), pp. 325–338.
[6] G. Csataljay, P.O Donoghue, M. Hughes, and H. Dancs, *Performance indicators that distinguish winning and losing teams in basketball*, Int. J. Perform. Anal. Sport 9 (2009), pp. 60–66.
[7] S.K. Deshpande and S.T. Jensen, *Estimating an NBA players impact on his teams chances of winning*, J. Quant. Anal. Sports 12 (2016), pp. 51–72.
[8] S. Dutta and S.H. Jacobson, *Modeling the NCAA basketball tournament selection process using a decision tree*, J. Sports Anal. 4 (2018), pp. 65–71.
[9] P. Fearnhead and B.M. Taylor, *On estimating the ability of nba players*, J. Quant. Anal. Sports 7 (2011), pp. 1–16.
[10] M. Goldman and J.M. Rao, *Effort vs. concentration: the asymmetric impact of pressure on NBA performance*, Proceedings of the MIT sloan sports analytics conference, Boston, 2012, pp. 1–10.
[11] A.A. Gupta, *A new approach to bracket prediction in the NCAA mens basketball tournament based on a dual-proportion likelihood*, J. Quant. Anal. Sports 11 (2015), pp. 53–67.
[12] J.A. Hanley and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology 143 (1982), pp. 29–36.
[13] J. Hollinger, *Pro Basketball Forecast: 2005–2006*, Potomac, Dulles, VA, 2005.
[14] D.W. Hosmer Jr, S. Lemeshow, and R.X. Sturdivant, *Applied Logistic Regression*, Vol. 398, John Wiley & Sons, New York, 2013.
[15] J. Kubatko, D. Oliver, K. Pelton, and D.T. Rosenbaum, *A starting point for analyzing basketball statistics*, J. Quant. Anal. Sports 3 (2007), pp. 1–22.
[16] M. Lewis, *Moneyball: The Art of Winning An Unfair Game*, WW Norton & Company, New York, 2004.
[17] D. Lock and D. Nettleton, *Using random forests to estimate win probability before each play of an NFL game*, J. Quant. Anal. Sports 10 (2014), pp. 197–205.
[18] B. Loeffelholz, E. Bednar, and K.W. Bauer, *Predicting NBA games using neural networks*, J. Quant. Anal. Sports 5 (2009), pp. 1–15.

[19] M.J. Lopez and G.J. Matthews, *Building an NCAA mens basketball predictive model and quantifying its success*, J. Quant. Anal. Sports 11 (2015), pp. 5–12.

[20] C.C. Madden, R.J. Kirkby, D. McDonald, J.J. Summers, D.F. Brown, and N.J. King, *Stressful situations in competitive basketball*, Aust. Psychol. 30 (1995), pp. 119–124.

[21] C.C. Madden, J.J. Summers, and D.F. Brown, *The influence of perceived stress on coping with competitive basketball*, Int. J. Sport. Psychol. 21 (1990), pp. 21–35.

[22] M. Manisera, R. Metulini, and P. Zuccolotto, *Basketball Analytics Using Spatial Tracking Data*, in *New Statistical Developments in Data Science: SIS*, Florence, June 28–30, 2017, pp. 305–318.

[23] H. Manner, *Modeling and forecasting the outcomes of NBA basketball games*, J. Quant. Anal. Sports12 (2016), pp. 31–41.

[24] P. McFarlane, *Evaluating NBA end-of-game decision-making*, J. Sports Anal. 5 (2019), pp. 17–22.

[25] R. Metulini, *Spatio-Temporal movements in team sports: A visualization approach using motion charts*, Electron. J. Appl. Stat. Anal. 10 (2017), pp. 809–831.

[26] R. Metulini, *Players movements and team shooting performance: a data mining approach for basketball*, Book of short papers SIS2018 *49th Scientific Meeting of the Italian Statistical Society*. Antonino Abbruzzo, Eugenio Brentari, and Marcello Chiodi e Davide Piacentino, eds., Pearson, 2018.

[27] R. Metulini, M. Marisera, and P. Zuccolotto, *Modelling the dynamic pattern of surface area in basketball and its effects on team performance*, J. Quant. Anal. Sports 14 (2018), pp. 117–130.

[28] R. Metulini, M. Marisera, and P. Zuccolotto, *Space-time analysis of movements in basketball using sensor data*, in *Statistics and Data Science: New Challenges, New Generations—Proceedings of the conference of the Italian Statistical Society*, Alessandra Petrucci, and Rossana Verde, eds., Firenze University Press, Firenze, 2019, pp. 701–706.

[29] D. Oliver, *Basketball on Paper: Rules and Tools for Performance Analysis*, Potomac Books, Inc., Washington, DC, 2004.

[30] M.U. Ozmen, *Foreign player quota, experience and efficiency of basketball players*, J. Quant. Anal. Sports 8 (2012), pp. 1–18.

[31] G.L. Page, B.J. Barney, and A.T. McGuire, *Effect of position, usage rate, and per game minutes played on nba player production curves*, J. Quant. Anal. Sports 9 (2013), pp. 337–345.

[32] G.L. Page, G.W. Fellingham, and C.S. Reese, *Using box-scores to determine a position's contribution to winning basketball games*, J. Quant. Anal. Sports 3 (2007), pp. 1–18.

[33] J. Piette, S. Anand, and K. Zhang, *Scoring and shooting abilities of NBA players*, J. Quant. Anal. Sports 6 (2010), pp. 1–23.

[34] F.J. Ruiz and F. Perez-Cruz, *A generative model for predicting outcomes in college basketball*, J. Quant. Anal. Sports 11 (2015), pp. 39–52.

[35] G. Schauberger and A. Groll, *Predicting matches in international football tournaments with random forests*, Stat. Model. 18 (2018), pp. 460–482.

[36] B. Stewart, *Sport Funding and Finance*, Routledge, London, 2017.

[37] T.M. Tango, M.G. Lichtman, and A.E. Dolphin, *The Book: Playing the Percentages in Baseball*, Potomac Books, Inc., Herndon, 2007.

[38] B.T. West, *A simple and flexible rating method for predicting success in the NCAA basketball tournament: updated results from 2007*, J. Quant. Anal. Sports 2 (2008), pp. 3–8.

[39] L.H. Yuan, A. Liu, A. Yeh, A. Kaufman, A. Reece, P. Bull, and L. Bornn, *A mixture-of-modelers approach to forecasting NCAA tournament outcomes*, J. Quant. Anal. Sports 11 (2015), pp. 13–27.

[40] P. Zuccolotto, M. Manisera, and M. Sandri, *Big data analytics for modeling scoring probability in basketball: the effect of shooting under high-pressure conditions*, Int. J. Sports Sci. Coach. 13 (2018), pp. 569–589.