Classification trees to define spatial performance indicators in basketball

Marica Manisera [1], Paola Zuccolotto [2] & Marco Sandri [3]

[1] *University of Brescia, C.da S. Chiara, 50 - 25122 Brescia, Italy, email: marica.manisera@unibs.it*
[2] *University of Brescia, C.da S. Chiara, 50 - 25122 Brescia, Italy, email: paola.zuccolotto@unibs.it*
[2] *DMS StatLab, C.da S. Chiara, 50 - 25122 Brescia, Italy, email: sandrimarco@gmail.com*

Résumé. Dans cette contribution, nous proposons la définition d'indicateurs de performance spatiale pour les joueurs et les équipes de basket-ball. L'analyse de la probabilité de marquer des joueurs et des équipes dans différents secteurs du terrain est très pertinente pour l'analyse du basket-ball, car elle permet aux entraîneurs et aux experts de définir de meilleures stratégies de jeu et programmes d'entraînement. Dans cette contribution, nous proposons une méthode basée sur les arbres de classification, qui définit une partition du terrain de jeu en rectangles avec des probabilités de marquer différentes au maximum. Des autres mesures d'efficacité de jeu calculées dans les rectangles peuvent tre utilisées pour définir des indicateurs de performance de scoring spatial. De plus, chaque équipe/joueur analysé a sa propre partition, de sorte que des comparaisons peuvent facilement être effectuées entre différentes équipes/joueurs.

Mots-clés. Analyse de sport, de performance, arbres de régression et classification.

Abstract. In this contribution we propose the definition of spatial performance indicators for basketball players and teams. The analysis of players' and teams' scoring probability in different areas on the court is very relevant in basketball analytics, because it enables coaches and experts to define better game strategies and training programmes. In this contribution we propose a method based on classification trees, which defines a partition of the court in rectangles with maximally different scoring probabilities. Shooting efficiency measures computed within the rectangles can be used to define spatial scoring performance indicators. Also, each analyzed team/player has its/his own partition, so comparisons can be easily made among different teams/players.

Keywords. Sport analytics, performance analysis, classification and regression trees.

1 Introduction

The use of statistical analysis and modelling techniques in sports has gained an increasing attention over the last years, as documented by the extensive scientific production on this topic and the publication of several insightful collections of statistical analyses applied to data from a wide range of sports, including soccer, tennis, American football, baseball, basketball amongst many others (see, for example, Albert et al., 2017). Developments, extensions, and applications on sports analytics have been proposed for several aims, including predicting the results of a match or a tournament, determining the drivers that affect the probability of winning a game, analyze players' performance or a team's performance, investigating the relationships between players' behavior and team success, identifying playing patterns, also with reference to players' movements, trajectories and the network of passing actions.

Performance analysis (PA) plays a crucial role in sports analytics, although there are several other research areas in sports (among others: sports psychology, medicine and management, ...). It has become essential not only for teams and coaches and athletes, but also for sports organizations and academic researchers. PA can be obtained by collecting performance data and interpreting them appropriately. Results enable coaches to identify the best training programmes, teams and athletes to improve their tactical decisions and game strategies, and sports organizations to manage teams and players more effectively.

This contribution focuses on PA in basketball. Several books and papers have been published on this topic, that can be exploited following several research lines and using different statistical methodologies. PA can be considered in a broad sense (accounting for both offensive and defensive abilities, for example) or with a restricted meaning, paying attention only on shooting performance. In addition, performance can be examined with reference to "time", focusing on the time dynamics during a match or during a season or on specific situations occurring during a game (Zuccolotto et al., 2017).

Another promising research line is PA in "space". The study of the players' position on the court during the game is gaining relevance due to the availability of spatial tracking data, used to investigate game strategies, players' roles and performance. Information Technology Systems allows to collect a large amount of different types of spatio-temporal data from a game: play-by-play data, which report a sequence of relevant events occurring during a game, and tracking data, collected using optical- or device-tracking and processing systems, capturing the movements and trajectories of players on the court (or the ball). With tracking data available, it is possible to consider several aspects, for example the interdependency of one player's trajectory with the other players' movements (Metulini et al., 2018) The final aim is to explain and observe cooperative movement patterns in reaction to a variety of factors, such as coach strategies in order to give suggestions to sports experts on the best strategy to win a game.

This contribution fits in this research line, concerned with PA in "space". In particular, we propose a statistical methodology to define spatial performance indicators with respect to the spots on the court. The idea is to split the court into a number of small areas in order to identify the favorite spots from which a player (or the players of one specific team) like to shoot and the awkward areas. The obtained spatial performance indicators can be used, for example, to compare players and teams.

2 Spatial shooting performance indicators

In this contribution we propose to construct spatial shooting performance indicators, by identifying how a selected measure of shooting performance varies on the court. We split the court into a number of rectangles. Each rectangle is then colored according to (i) scoring percentages; (ii) difference in scoring percentages with respect to other players, teams, etc.; (iii) some aggregate statistics (e.g., average play length when the shot is attempted, average time in the quarter, etc.).

We propose to define the rectangles by means of a classification tree, belonging to the well-known framework of Classification and Regression Trees (CART; Breiman et al., 1984). CART has recently be judged to be one of the top 10 algorithms in data mining thanks to its easy interpretability and accuracy in results (provided that the sample size is large enough to avoid instability issues). In detail, we use a classification tree in a naive way in order to find the best partition of the court according the criterion of maximizing the difference in the scoring percentages between rectangles. Considering, for example, all the shots attempted by a specific player in a sufficiently high number of matches, we build a classification tree where the outcome variable Y is a binary variable indicating whether the attempted shot scored a basket ($Y = 1$) or not ($Y = 0$) and the predictors are the two variables indicating the two space coordinates of each attempted shot. The estimated tree induces a partition of the space of the predictors into rectangles (the tree leaves) that can be colored by the scoring probability (probability of $Y = 1$) characterizing the shots attempted in that rectangle. Then, several shooting efficiency measures can be computed within the rectangles and used as spatial scoring performance indicators, which can be effectively represented by different colors on the partition of the court map.

As an example, Figure 1 shows the spatial performance indicators obtained for Stephen Curry, a well-known NBA player (data are from the regular season 2017-2018). In the left-hand plot, the partition of the court induced by the classification tree is colored according to Curry's scoring percentages, which evidently vary on the different spots on the court. In the middle and right-hand plots, rectangles are colored according to the difference in scoring percentages with respect to Curry's teammates (all the other players of the Golden State Warriors team in that season) and according to the average time in the quarter when the shot is attempted, respectively.

The easy interpretation of the court partition gives interesting suggestions on the scoring performance of the single player and can help the definition of game strategies but also the identification of the best training programmes. The same analysis can be done for all the shots attempted by the players of one team, in order to obtain spatial performance indicators of the whole team, also compared with the performance of all its opponents. Each analyzed team/player has its/his own partition, so comparisons can easily be made among different teams/players. A graphical inspection of the partitions of different players allows, for example, the identification of the court zones most favorable for each player and this can help the coach to define the best training and game strategy

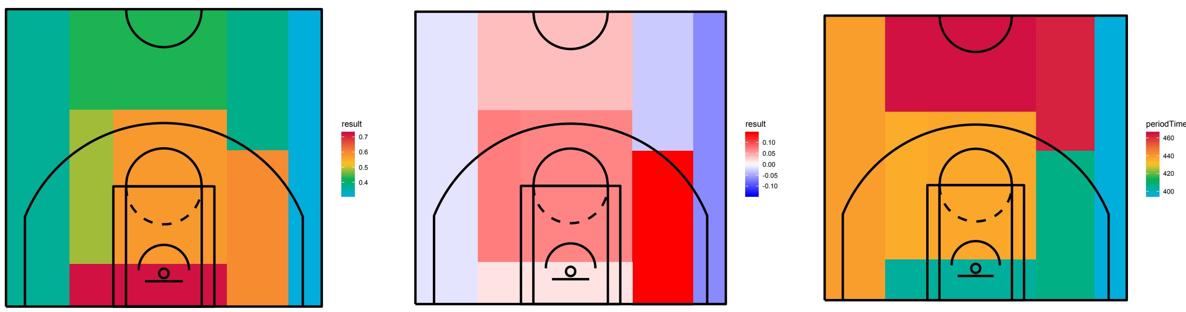

Figure 1: Spatial performance indicators of Stephen Curry on the partition of the court induced by the classification tree: scoring percentages (left), difference in scoring percentages with respect to teammates (middle) and average time in the quarter when the shot is attempted (right) - Data from the NBA regular season 2017-2018

to finally win the match. The R code useful to perform the proposed spatial analysis of scoring performance of players and teams will be soon available in the upcoming R package BasketballAnalyzeR.

Acknowledgments
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").

References

Albert, J., Glickman, M. E., Swartz, T. B. and Koning, R. H. (2017), *Handbook of Statistical Methods and Analyses in Sports*, CRC Press, Boca Raton, FL.

Breiman, L. and Friedman, J. H. and Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA.

Metulini, R., Manisera, M., Zuccolotto, P. (2018). Modelling the dynamic pattern of surface area in basketball and its effects on team performance, *Journal of Quantitative Analysis in Sports*, 14, pp. 117-130.

Zuccolotto, P., Manisera, M. and Sandri, M. (2017). Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions, *International Journal of Sports Science & Coaching* ,13, pp. 569-589.