POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Basketball shooting performance evaluation through spatial probability maps: an application to the 2022/2023 Italian first league

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Mirko Luigi Carlesso, 991872

Advisor:
Prof. Andrea Cappozzo

Co-advisors:
Paola Zuccolotto

Academic year:
2022-2023

Abstract: This work focuses on Basketball analytics, specifically on evaluating shooting performance using innovative maps that vividly illustrate the probability of scoring from every position of the court. These maps can uncover subtle areas with higher or lower effectiveness in shooting performance, providing valuable insights within the Basketball framework. To construct the map, an inferential procedure is employed: a model is trained using actual shot data, which includes coordinates and binary outcome (shot made or missed), and subsequently used to generate predictions for the court grid. We propose two main approaches, the former based on Adaboost method, which is a boosting technique that uses decision-trees as base learners, and the latter based on Indicator Kriging, a geostatistical prediction technique. We also discuss the possibility to add categorical variables to the model, leading to the production of visual impactful maps. We then tackle the issue of the non-homogeneous density of the basketball shots data. All the employed models are then compared using an index able to assess the graphical goodness of the maps. The data application involves a precise and rich dataset containing all the shots taken in the 2022/2023 Italian first league (LBA).

Key-words: Basketball Analytics, Performance Analysis, Adaboost, Geostatistic, Indicator Kriging

1. Introduction

Data Science serves as a powerful tool for extracting knowledge from data and its application in the world of sports is experiencing rapid growth [3]. In recent years, more and more people engaged in the sports industry have recognized the importance of data in conducting their operations, ranging from company management to player performance evaluation. The availability of data has expanded considerably, presenting ample opportunities to measure and assess athletes' performance in greater detail than ever before.
Books have been written and even films have been made on the topic of Sports Analytics [2, 17, 20], highlighting how this area is rapidly developing. For instance, in some sports the scouting process is increasingly managed through data and statistics, and even athletes sometimes look after their own economic interests by showcasing their impact on the team through data-driven insights. This is certainly possible because of the increasing presence of data capture techniques, spanning from traditional match tagging to the use of sensors and artificial intelligence. As such, Sport analytics encompasses several key areas that can be summarized within the two following macroblocks:

- Business enterprise analytics

- Analysis of players' technical and physical performance

The former area concerns classic industry-wide statistical analysis, while the latter, which will be the primary focus of this work, concerns player and team performance.

Focusing specifically on Basketball, over the years there has been a tangible evolution in the concept of statistics itself. The pioneer in this regard was Dean Oliver with his famous book Basketball On Paper [20]. He introduced the concept of possession to scan the pace of games leading to a more accurate system of players and teams comparisons. Essentially, a possession can be defined as an opportunity for a team to make a basket. Since the introduction of this concept, there has been a significant rise in the number of possessions during games leading to an increase in the overall pace of the game. The introduction of the so-called Four Factors of Basketball Success, still a concept introduced in Oliver's book, is an important turning point in the word of Basketball Analytics. Oliver started the era of advanced statistics in basketball, an area that is still developing rapidly.
When it comes to implementing these concepts on the basketball court, Daryl Morey was undoubtedly the pioneer, making his entrance onto the stage during his period as the General Manager of the Houston Rockets, from 2006 to 2020. He developed a Sport Analytics team within the franchise and brought revolutionary concepts to the field that are still changing the way Basketball is viewed today. For instance, he developed a precise shot selection strategy according to which the most effective shots are behind the three-point line and close to the rim.
These ideas about shot selection were recently brought to Italy by Pallacanestro Varese, specifically by its CEO Luis Scola, a former player of Houston Rockets of the GM Morey. In the summer of 2022, Scola brought Matt Brase to Varese as head coach, who in previous years worked as an assistant for Houston Rockets. The idea of the game they brought in is a strong analytical one, preferring shots from 3 or close to the rim, while also aiming to significantly increase the pace of games. The result was surprising as Varese, despite one of the lowest budgets in the league to operate in the market, finished the championship in sixth place with 17 wins and 13 losses. During the months of April and May 2023, while still manuscript was being prepared, I had the opportunity to intern at Pallacanestro Varese, allowing me to experience their innovative basketball vision from the inside.

For this work, we chose to analyze the shooting performance of teams and players of the Italian first league (LBA) 2022/2023 by producing maps that could highlight areas of higher or lower shooting effectiveness. Maps of this kind could assist coaches and scouts in gaining a deeper understanding of shooting efficiency. Unlike classic descriptive techniques, these maps generate zones with different scoring probability based on a statistical model trained on the data. For example, they can reveal small areas of high efficiency within larger areas of lower efficiency. The analysis of shooting performance was inspired by the book *Basketball Data Science, with applications in R* [28]. In the book a lot of different research topics about Basketball Analytics are covered, analysing data of the National Basketball Association (NBA). Talking specifically on shooting maps, two important paper was published recently. The first [29] introduced methods for visualization of the shots in terms of scatter-plot or density plot and then proposed a method based on CART (Classification And Regression Trees) [5] to produce shooting probability maps. While the second one [30] tackled the same problem changing the coordinates system of the shots to a polar one and employed ensemble methods to obtain more meaningful and robust maps. An index to evaluate the visual effectiveness of the produced maps was also introduced in the latter. In these works play-by-play data of NBA season 2017/2018 were been analysed, which contains coordinates of each shots and some other information about them. On the other hand play-by-play data of the Italian league (LBA) 2022/2023 will be used in this work which, as explained later in Section 2, contains a lot of problems regarding the shot coordinates that have been fixed thanks to the analytics team of Pallacanestro Varese.

From a methodological point of view, maps are produced predicting the probability of scoring a basket in each point of a grid built over the Basketball half-court. The prediction is generated by a model that has been trained using all the shots from the specific team being analysed. In this model, the shot coordinates are used as predictors, while the binary outcome of whether the shot is successful or not serves as the response variable. Building on what has been done in the aforementioned works [29, 30], here we will first employ methods based on decision trees and then a geostatistical approach in which the basketball court is interpreted as a two-dimensional spatial domain. In the first case, after trying CART [5] and Random Forest [4], the Adaboost [23] algorithm will be used, which, being a boosting technique, is capable of adapting to the data on which it is trained, resulting in meaningful maps that effectively assess shooting performance. In the spatial statistic context, Indicator Kriging [7] will be used, which is a geostatistical technique suitable for predicting a binary response variable. In addition, two extensions will be developed and discussed later: the first incorporates categorical variables into the model for each shot leading to the creation of as many maps as the levels of the categorical variable being considered, while the second concerns an adjustment of the prediction in areas where there are

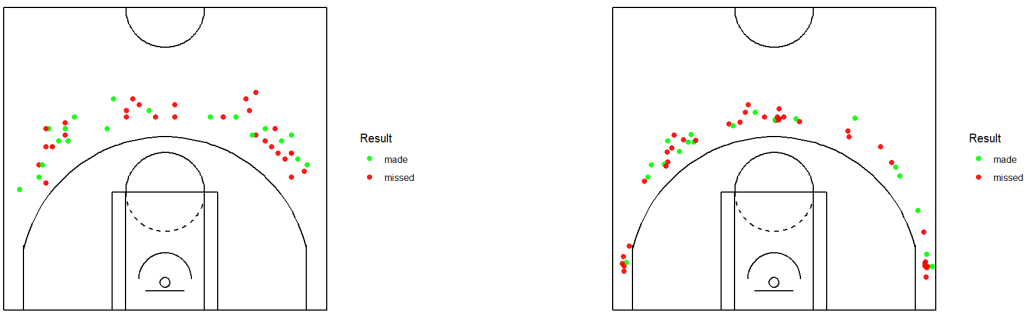
"few" shots and is meant to produce meaningful maps for any kind of customization that a final user may desire.

The paper is organized as follows: in Section 2 the data used for the analysis will be presented, highlighting their great worth. In Section 3 the statistical problem concerning the creation of shooting performance maps will be described in a more technical way, and Sections 4 and 5 will present the two different approaches with which this problem was solved: the former based on decision trees and the latter based on geostatistics. Section 6 will present 2 extensions regarding categorical variables and adjustment of model predictions in low-density shot areas. In Section 7 the proposed models will be compared using an index proposed by Zuccolotto et al. [30] while some concluding remarks and a discussion on possible future research directions are reported in Section 8.
To ensure good usability of the maps, it was developed a Shiny application that allows a user to view maps with a very high level of customization. In the app one could visualize maps both of the teams and of some players. A detailed description of the application can be found in Appendix A.

2. LBA 2022/2023 Tournament Shots Dataset: Collection and Accuracy Challenges

To do statistical analysis about basketball in recent years, we are relying more and more on play-by-play data, that is, a dataset that presents, line by line, every event in the game. The various columns represent the basic information about the game (teams, day and time), the ten players on the court and various details about the event just recorded. Obviously, the more detailed the dataset, the more useful information that can be derived from it. European basketball in this respect is still lagging behind the amount of information that is collected for NBA games, but some organizations, such as Pallacanestro Varese, are helping to make our basketball more aware of the usefulness that certain data can have.

Data considered in this work are all the shots of the LBA 2022/2023 tournament, which are taken from the play-by-play made available by the league. In these data, for each shot we have information on its coordinates in the court, the player who attempts it, its result, either made or missed, i.e. our response variable and some game's information of the moment in which the shot is taken. Although these data are freely accessible, their accuracy is, unfortunately, compromised. This is due to their real-time acquisition during matches without subsequent updates, leading to a considerable number of errors despite the critical relevance of the analysis derived from these data. At the beginning of 2022/2023 championship, the analytics team of Pallacanestro Varese noticed these tagging errors. As depicted in Figure 1 the difference between actual coordinates and coordinates tagged by the league can be remarkable. By looking at this example one can notice very serious errors, but, within each match, numerous minor errors exist that can render the data supplied by the Italian League capable of generating entirely inaccurate analyses. Therefore, in order to have reliable coordinates, Pallacanestro Varese created a system that corrects the position of each shot and also adds other information about it. This was done on every game in the championship, thus leading Varese have accurate and incredibly important data to evaluate the teams' shooting performance. This provides Varese with a competitive edge over other teams, as they possess a dataset containing accurate information. In contrast, many other teams often base their strategies on incorrect or unreliable data.



(a) Shots tagged by the League

(b) Shots corrected by Pallacanestro Varese

Figure 1: Example of shot coordinate correction: three-point-shots of the match Pallacanestro Varese-Umana Reyer Venezia, November 2022. Clearly one can see that all the corner-threes were wrongly tagged by the league play-by-play (Figure 1a). While in Figure 1b are displayed the real positions of the shots.

The work done for by Pallacanestro Varese, called tagging, involves reviewing each shot of every game by correcting the position and adding other features to it. Going into detail, the following operations are manually carried out for each shot:

- The position is corrected
- Seconds left on 24-shot-clock are added
- Whether or not the shot was contested by the defense is indicated
- Information is added on each missed shot as to whether it was assisted or not
- If the shot was scored after an assist, the position of the assist is also corrected.

The benefits of having a complete, accurate, and enriched dataset of all shots along with other useful details was one of the reasons that helped Varese have a high-profile season. Analysing the shooting spectrum of a game, which entails examining the distribution of shots across the court, holds immense significance for them. Indeed, it allows the coaching staff to assess whether the observed shot distribution aligns with their strategic game plan and overall objectives. It can be done also on data of the other teams, proving to be an important tool in pre-game scouting.
In the context of our work, accurate shooting position data holds paramount significance, especially when constructing a model that relies on shots' coordinates. Therefore, we will use the data furnished by Pallacanestro Varese for our analysis.

In the dataset, each shot is accompanied by a column indicating the corresponding court area from which it was taken. These zones, visualized with different colors in Figure 2, represent a simple partition of the court based on its lines.
For the analysis, it was decided to exclude shots that are not considered significant in basketball, namely those taken from very far away, falling within the "3W" zone. Such shots are not typical of regular game play as they are attempted only in 'desperate' offensive situations and are significantly fewer in number compared to all other shots (153 out of 30,197 total shots in LBA 2022/2023). By excluding these shots, the resulting maps also neglect the "3W" zone of the court.

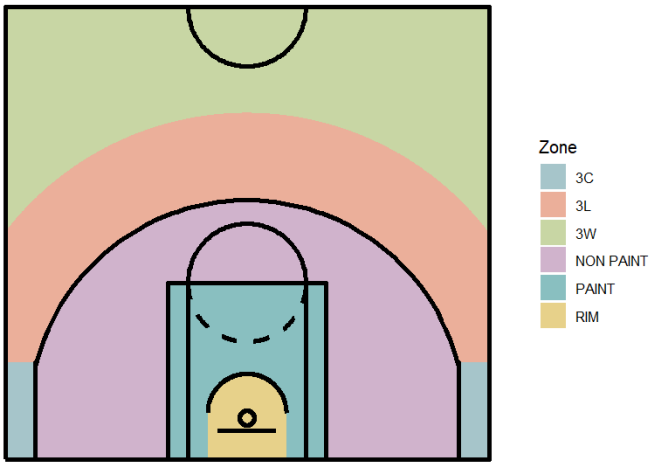
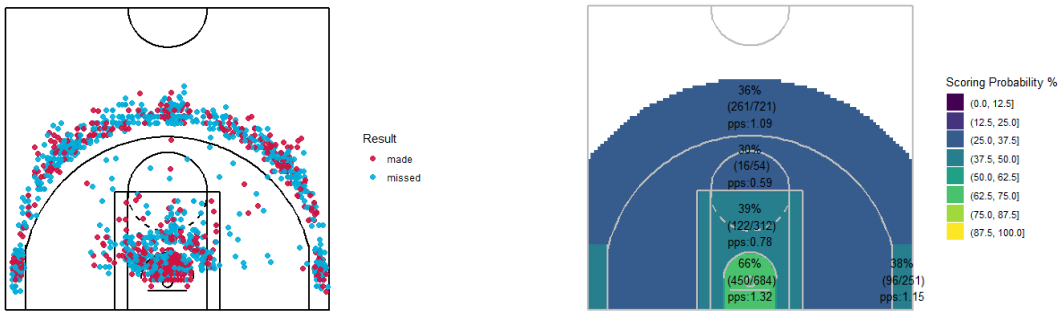


Figure 2: Zones in which i partitioned the court. These are simply induced by the court-lines.

3. Problem Description

The first step to visualize shots in Basketball is to display a shot chart where one can read the performance for each predetermined zone of the court. These maps are the classic ones that can be found in every game report along with the scatter plot of all the taken shots, like the example reported in Figure 3. In addition to the aforementioned examples, a density plot can be used to visualize the shots distribution of a team or player, providing valuable insights into their preferred shooting locations on the field. In this work, we will discuss an innovative and valuable tool introduced by Zuccolotto et al. [29], which involves creating maps to evaluate shooting performance using a model built on the data. Moving beyond the descriptive methods mentioned earlier, we transition into the realm of inferential statistics. Here, our conclusions, namely the maps, originate from a statistical model. These maps provide insights into a team's shooting patterns across diverse court regions, with the delineation of these areas being shaped by statistical modeling rather than predetermined.



(a) Scatter plot of the shots

(b) Probability of scoring by zones

Figure 3: Examples of classic approach to analyze shooting performance of a Team on Pallacanestro Varese shots in LBA 2022/2023. In Figure 3a shots are displayed in the court colored by their result, while in Figure 3b for each one of court-zones (Figure 2) the following statistics are displayed from the top to the bottom: shooting percentage; (shots made/ shots missed); point per shot, i.e. (shots made/ shots missed) multiplied by the point of the shot (2 or 3). The zones are also colored according to their shooting percentage. The codes for generating these figures are adapted from the functions provided in the `BasketballAnalyzeR` package [22].

The underlying statistical question guiding the creation of these maps is as follow: given a sequence of team's shots, represented by their 2D coordinates on the court and their outcome, the objective is to predict the probability of scoring for every point of the basketball half-court. Thus the dataset S consists of M available couples $(\mathbf{x}_i, y_i)$ representing shots, where $\mathbf{x}_i$ consists of two components, $x_{h,i}$ and $x_{w,i}$, representing the height

and width of the shot in the court, and $y_i \in \{0, 1\}$ representing the binary outcome of shot $i, i = 1, ..., M$. It is noteworthy that in some cases the components $x_{h,i}$ and $x_{w,i}$ will be replaced by polar coordinates $x_{r,i}$ and $x_{\theta,i}$, representing respectively radius and angle of shot with respect to the basket. In addition, the analysis can be extended by adding a categorical component $x_{c,i}$ for each shot. Operationally, for the prediction part we have a 100×100 grid consisting of points in the court with their $x_{h,j}$ and $x_{w,j}$ coordinates, $j = 1, ..., 10000$. We want to predict $P(y_j = 1|x_{h,j}, x_{w,j})$ for the grid based on the M shots we have available. It is crucial to highlight that the primary objective is not to maximize model performance on new data, but rather to generate a map that accurately reflects the team's actual performance during the period under analysis. Therefore, the evaluation of the models will not be based on accuracy calculated on new data. Instead, other parameters will be employed to assess the graphical quality and goodness of the maps. This approach ensures that the resulting maps capture the true shooting performance of the team under consideration, offering valuable visual insights into their performance distribution across the basketball court.

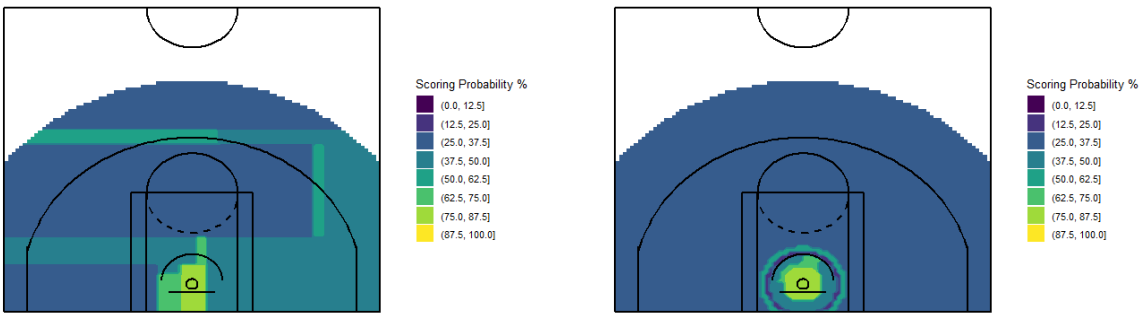
Thus, a classification model seems to be the correct approach to deal with this problem. The model needs to be trained on the M shots having the $(x_{h,i}, x_{w,j})_{i=1}^{M}$ coordinates as predictors (or the polar ones) and the outcomes y_i as response. Once the model is trained, it is used to obtain $P(y_j = 1|x_{h,j}, x_{w,j})$ for the grid. For the goal of creating areas characterized by different scoring probability, as pointed out by Zuccolotto et al. [29], tree-based classification models seem to be the most suitable ones due to their interpretability in the case of a two-dimensional predictor space. We will discuss them in Section 4.

An alternative approach involves treating the basketball court as a spatial domain by employing geostatistical techniques to create the maps. In geostatistics the procedure involves an initial modeling of spatial variability, followed by the application of prediction techniques. In Section 5, we will explore in-depth the implementation and results of using Indicator Kriging [7], which allowed us to create highly significant and visually striking maps.

4. Methods based on Decision Trees

4.1. CART

Given the objective of creating maps that partition the basketball court into zones with different probabilities of making a basket, a straightforward and effective solution involves the usage of methods based on decision trees [29]. Specifically, CART (Classification and Regression Trees) [5] offer an optimal partition in the predictor space by maximizing the variability between the induced division based on classification or regression criteria. In the context of predicting the value of a dependent variable Y from a set of predictors $X_1, X_2, ..., X_p$, CART proceed through a binary partitioning process. At each step, the algorithm divides the predictor space and fits a simple model to each resulting partition, the so-called nodes. For classification tasks, the simple model fitted in each node often represents the mode of the partitioned data. Classification trees employ primarily two criteria for partitioning: information gain or Gini impurity. These criteria aim to create homogeneous groups within each node, optimizing the classification process. The splitting process may continue until each node contains only one unit, and to control excessive tree growth, pruning criteria are introduced to prevent overfitting. For a through description of the method and the quantities involved, the interested reader is referred to [15].
By leveraging CART in this manner, we can efficiently generate maps that facilitate the identification of distinct shooting zones and provide valuable insights into players' scoring probabilities across the basketball court. In our case the dependent variable is represented by the binary outcome of each shot and the predictors can be either the regular coordinates $(x_{h,i}, x_{w,j})_{i=1}^{M}$, i.e. height and width in the half court, or the polar coordinates $(x_{r,i}, x_{\theta,j})_{i=1}^{M}$, i.e. the radius and the angle with respect to the basket. As previously mentioned, using CART the obtained partition is very easy to interpret from a visual point of view in the Basketball court. In the case of classic coordinates, rectangles are created in the court while if polar coordinates are used, circular sectors or annuli are drawn. Following what Zuccolotto et al. did [29], we report in Figure 4 two partitions of the half-court generated with CART on the shots of Openjobmetis Varese in the 2022/2023 season of the Italian league. Models are built using `rpart` function from the R package `rpart` [25] using a complexity parameter `cp = 0.003` and the minimum number of observations that must exist in a node in order for a split to be attempted equal to `minsplit = 85`. The analysis reveals that both models indicate a higher scoring probability when shooting near the basket, which is a result not particularly useful for those involved in basketball. The model employing regular coordinates also identifies other certain areas with good performance, but these regions have peculiar shapes that might be challenging to interpret on a basketball court.



(a) CART grown on classic coordinates

(b) CART grown on polar coordinates

Figure 4: Examples of shooting performance maps produced via CART on classic coordinates (Figure 4a) and on polar coordinates (Figure 4b). Data are all the shots taken by Pallacanestro Varese in the 2022/2023 LBA tournament.

4.2. Random Forest

Using CART in this context is limiting because, despite the interpretability of the maps, they lack flexibility and suffer from instability [15]. For this reason, ensemble methods have gained popularity and are also more effective for our specific objective. They combine a certain number of weak learners, often tree-based method like CART, to create a more robust and stable classifier. The prediction of an ensemble model is derived from the predictions of its constituent elements, often through a weighted majority vote in classification tasks. Among ensemble techniques we find a division between bagging and boosting. The former aims to reduce variance and improve stability by averaging the predictions of independently trained basic models. Examples of bagging methods are Random Forest [4] and Extra Randomized Trees [14]. Boosting techniques, instead, train basic models sequentially by focusing on hard-to-classify data with the goal of reducing bias without overfitting. Examples are AdaBoost algorithm [23], Gradient Boosting Machines [10] and Extreme Gradient Boosting [6].

Moving to ensemble methods, we firstly implement a Random Forest (RF) [4] method on polar coordinates. RF constructs, in parallel, multiple decision trees using random subsets of the data and of the features. Each individual tree contributes its vote to determine the prediction, which is done by majority voting in the classification framework. This method mitigates overfitting and helps in getting more accurate and stable predictions. Since our features are only two, the radius and the angle of each shot with respect to the basket, we decided not to exploit the random choice of the feature in order to split a node, setting the number of features used equal to 2 for each decision tree. This bagging strategy is built in R using the `randomForest` package [18] where we grown 5000 trees with a minimum size of the terminal nodes equal to 150 (about 10% of total shots), which seems to be a reasonable choice for each team. We report in Figure 5 a shooting performance map produced via Random Forest model. The model is trained on the available shot of the selected team and then predictions of the scoring probability are made on a 100×100 grid built on the court. It can be notice that RF is more effective in recognizing different areas in terms of shooting percentage on the court. According to the map Germani Brescia, during 2022/2023 season, has been effective in shooting from the baseline and the wings, while they performed worse in the central part of the court.

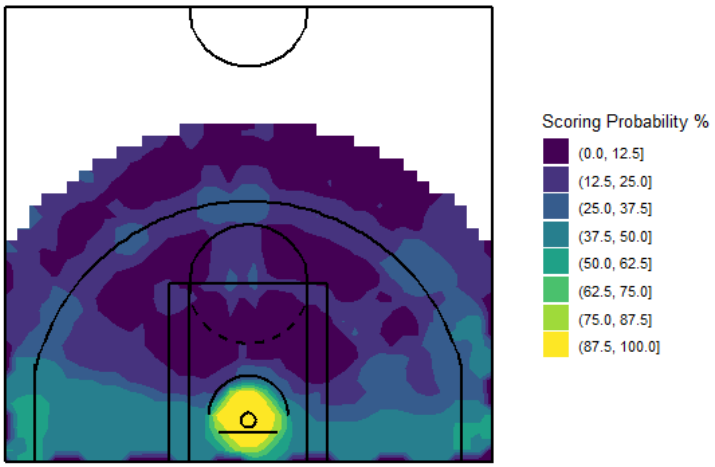
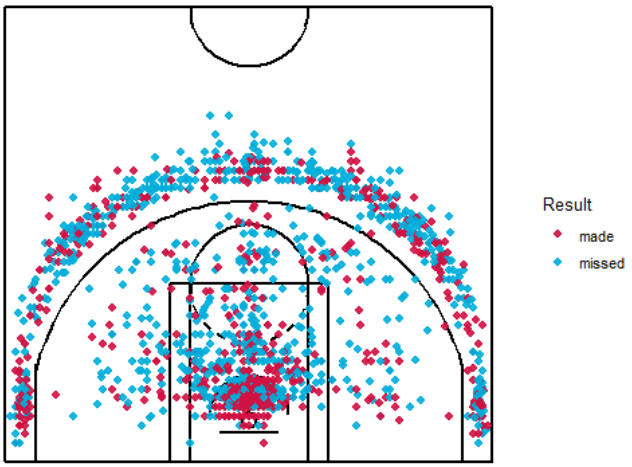
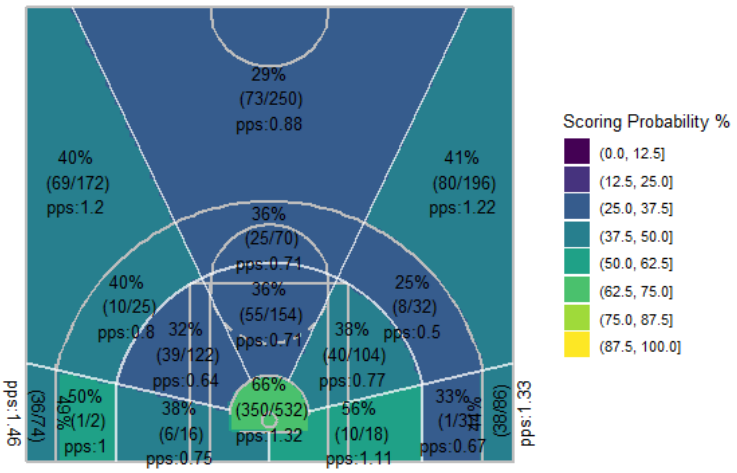


Figure 5: Shooting performance map obtained with Random Forest on shots taken by Germani Brescia in the 2022/2023 LBA tournament.

Looking at the maps made with Random Forest more in details, it came to light that they suffer from a major problem. In fact, as it can be seen, they produce large areas of the court where the probability of making a basket is between 0% and 12.5% and others (almost always under the basket) where we are in the range (87.5, 100]%, which seem to be a little extreme. This, in addition to being a pattern shared by all models on Italian teams, is also noticeable in the maps produced with Random Forest and Extra Randomized Trees by Zuccolotto et al. [30] on NBA player data. These areas, despite being zones where there is indeed a low/high probability, are exaggerated negatively and positively by the model. Figure 6 shows the scatter plot and statistics by sector for Germani Brescia, based on the same shots with which the model in Figure 5 was constructed. It can be seen that both the dark purple and yellow areas do not match the actual data, which is something we would like maps of this type to do. They should indeed be a visually impactful tool that gives a correct assessment of what is the shooting performance of teams or players. In conclusion, Random Forest is capable in generating maps that delineate areas of varying scoring effectiveness. However, it's important to note that while these maps identify general trends of good and poor shooting ability, the predicted probabilities may not precisely align with the actual probabilities exhibited by the team. Consequently, the next logical step involves exploring alternative approaches that can replicate the prowess of Random Forest in capturing nuanced fluctuations in shooting efficacy across localized zones, while concurrently generating predictive probabilities that harmonize with a team's authentic performance.



(a) Scatterplot of the shots



(b) Statistics by sector

Figure 6: Scatterplot of all the shots colored by their binary result(Figure 6a) and statistics (shots made/missed, percentage and point-per-shot) for each sector of the half-court (Figure 6b). Data are all the shots taken by Germani Brescia in the 2022/2023 LBA tournament. The codes for generating these figures are adapted from the functions provided in the `BasketballAnalyzeR` package [22].

4.3. Adaboost

Given the limitations encountered with bagging methods like Random Forest, transitioning to boosting techniques proves to be a promising choice. Boosting methods present a compelling advantage in creating models that possess a deeper understanding of the training data, which addresses the main issue that arose with RF. Throughout the training process, boosting methods construct a robust model by iteratively emphasizing misclassified instances. Sequentially growing a series of simple models known as weak learners, they adjust the focus on data points to better capture the underlying process generating the data. In our context, where the aim is creating shooting performance maps for accurate team evaluation, we share the same objective as boosting techniques, that is to be more data-focused and derive meaningful insights from the information provided by the data. Therefore, adopting boosting methods aligns perfectly with our goal to produce comprehensive and informative shooting performance maps. Here we present the results obtained using Adaboost algorithm [23].

Adaboost, that stands for adaptive boosting, is a method developed by Robert Schapire and Yoav Freund in 1995 [9]. It was the first practical boosting algorithm, and remains one of the most widely used and studied, with applications in numerous fields [16, 23, 27]. The method takes as input a training set S of M samples with each sample composed by a couple $(\mathbf{x}_i, y_i)$, with $\mathbf{x}_i$ instance drawn from some space X and $y_i \in Y$ class

label associated with $\mathbf{x}_i$. The algorithm (Algorithm 1) calls a Weak Learner (usually decision stumps - shallow decision trees with only one split) at every round t. At each round t the booster provides Weak Learner with a distribution D_t over the training set S. In response, the Weak Learner computes a classifier or hypothesis $h_t : X \to Y$ which minimizes the training error ϵ_t over D_t. This process continues for T rounds, and, at last, the booster combines the weak hypothesis $h_1, ..., h_T$ into a single final hypothesis h_{fin}.
The initial distribution D_1 is uniform over S so $D_1(i) = 1/M$ for all i. To compute distribution D_{t+1} from D_t and the last weak hypothesis h_t, we multiply the weight of example i by some number $\beta_t \in [0, 1)$ if h_t classifies x_i correctly, and otherwise the weight is left unchanged. The weights are then renormalized dividing by the normalization constant Z_t. Thus, AdaBoost focuses the most weight on the examples which seem to be hardest to correctly classify for the weak learner. The number β_t is computed as a function of the training error ϵ_t. The final hypothesis h_{fin} is a weighted vote of the weak hypotheses. That is, for a given instance x, h_{fin} outputs the label y that maximizes the sum of the weights of the weak hypothesis predicting that label. The weight of hypothesis h_t is defined to be $\log(1/\beta_t)$ so that greater weight is given to hypothesis with lower error.

Algorithm 1 AdaBoost.M1 for binary data

Input: sequence of M examples $\langle (x_1, y_1), ..., (x_M, y_M) \rangle$ with labels $y_i \in Y = \{0, 1\}$; weak learning algorithm **WeakLearn**; integer T specifying number of iterations
Initialize $D_1(i) = 1/M$ for all i
for $t = 1, 2, ..., T$ **do**
    1. Call **WeakLearn**, providing it with the distribution D_t.
    2. Get back a hypothesis $h_t : X \to Y$.
    3. Calculate the error of h_t: $\epsilon_t = \sum_{i:h_t(x_i) \neq y_i} D_t(i)$.
    If $\epsilon_t > 1/2$, then set $T = t - 1$ and abort loop
    4. Set $\beta_t = \frac{\epsilon_t}{2(1-\epsilon_t)}$.
    5. Update distribution D_t:
    $$D_{t+1} = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t, & \text{if } h_t(x_i) = y_i \\ 1, & \text{otherwise} \end{cases}$$
    where Z_t is a normalization constant (chosen so that D_{t+1} will be a distribution).
end for
Output the final hypothesis:
$h_{fin} = \arg\max_{y \in Y} \sum_{t:h_t(x)=y} \log(1/\beta_t)$.

In our case, $\mathbf{x}_i$ is composed by the polar coordinates $x_{r,i}$ and $x_{\theta,i}$ and the target $y_i \in 0, 1$ represents shot's outcome. In `R` Adaboost.M1 is implemented in the package `adabag` [1] through the function `boosting`. The total number of rounds T was set to 100, the default value. Once the training procedure is done on the available data, a prediction grid of scoring probabilities is made over the court in order to produce the desired shooting performance map. We report in Figure 7 an example of these maps built on data of Germani Brescia, so we can make a comparison between this map and the one produced via Random Forest (Figure 5). Both the maps highlights the fact that the team has been better in shooting from the side and the baseline, while having some small good area also in the central part of the court. This tendency to shoot worse from the top, shared among almost all teams, is maybe due to the fact that most the isolation pull-up shots [1] are taken there and they are the less effective type of shot in basketball. The good point of the map built with Adaboost is that the percentage that can be read in the map is reliable, since here there is a matching between the map and the real performance (see Figure 6b). So with Adaboost algorithm we can conclude that we achieve the goal of producing maps able to assess in a precise way the shooting performance of the teams.

[1] Isolation refers to a scenario in which the o□ensive player engages in a one-on-one matchup with a defender within an open area of the court. In such situations, the o□ensive player frequently attempts challenging shots, often taking them directly o□ the dribble (pull-up). These dribble-initiated shots are among the most di□cult in basketball.

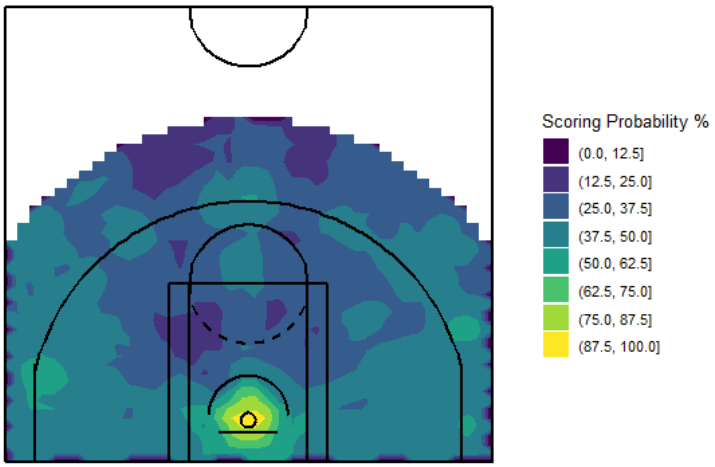


Figure 7: Shooting performance map obtained with Adaboost algorithm on shots taken by Germani Brescia in the 2022/2023 LBA tournament.

Adaboost algorithm turns out to be a great tool in producing meaningful shooting performance evaluation maps. Highlighting also small area with peculiar shooting percentage, they are very useful for inspecting the behaviour of the teams. Referring again on Figure 7, despite the central zone of lower probability, we notice there is a small light area across the three point line. Watching Germani Brescia's games, one could see that Amedeo Della Valle [2] is used to shoot from there with a great efficiency and so it is not surprising to have this little area with higher scoring probability in that position. Typically, within a predefined sector, there may exist variations in performance that traditional descriptive methods fail to capture. From our perspective, this nuanced understanding represents the true benefit of using our shooting performance evaluation maps.

Unfortunately, the Adaboost algorithm, being a boosting technique with a sequential training procedure, can be relatively slow. Consequently, when aiming to produce shooting performance maps for users, this approach is not ideal, as running the entire procedure might take a considerable amount of time. Therefore, we recognize the necessity of exploring alternative solutions to enhance user-friendliness within the context of our Shiny application. The primary objective remains achieving the same visual effectiveness in generating the maps while significantly improving computational efficiency to ensure timely map production. In the subsequent section, through geostatistic, we successfully achieved a very promising outcome in terms of computational speed and map generation.

5. Method based on Geostatistic

5.1. Introduction and notation

Spatial statistic deals with data observed at different locations within a spatial domain [7]. In these cases, the observed data frequently display a spatial dependence structure, necessitating careful attention to ensure precise statistical evaluations. Spatial statistics aims to develop inferential methods that appropriately account for this spatial dependence when dealing with georeferenced observations. In this framework, the analyzed phenomenon is modelled through a random field

$$\{Z_{\mathbf{s}}, \quad \mathbf{s} \in D\}. \tag{1}$$

In details, $D \subseteq \mathbb{R}^d$ is the spatial domain, usually being $d = 2, 3$ and $Z_{\mathbf{s}}$ observed at locations $\mathbf{s}_1, ..., \mathbf{s}_M$. $Z_{\mathbf{s}}$ could be either a continuous or a discrete variable. Typical geostatistical analyses consist of (a) identify appropriate models to describe the spatial variability of the phenomenon, (b) estimate the corresponding parameters and (c) perform predictions at unobserved locations in the system.

In the context of basketball shots' data analysis, employing geostatistics is a sensible approach due to the inherent spatial nature of the problem. Indeed Geostatistic is well-suited for situations where data exhibits spatial autocorrelation, meaning that nearby locations tend to have similar value of $Z_{\mathbf{s}}$. In our case, basketball shots

[2] Italian player of Germani Brescia. He averages 16.5 points per game during the 2022/2023 season.

are taken from specific locations on the court ($d = 2$), and it is reasonable to assume that shots' outcomes, that will be our $Z_{\mathbf{s}} \in \{0, 1\}$, in the same area are more likely to be similar due to factors like shooting angles, distance from the basket and player roles. By treating the basketball court as a two-dimensional spatial domain, where each location $\mathbf{s}_i$ is composed by coordinates $x_{h,i}$ and $x_{w,i}$, we can leverage geostatistical techniques such as Variogram modeling to analyze the spatial autocorrelation of successful and unsuccessful shots. Then, employing this model, we will make use of Indicator Kriging (IK) [7], a prediction technique in the geostatistical context, to produce shooting performance evaluation maps. We will estimate through IK the probabilities of scoring at unsampled locations, which for us are the grid points on the half-court, enabling the creation of our spatial probability maps highlighting areas with a higher or lower probability of successful shots.

To make inference in the context of Spatial Statistic one has to keep in mind that $\mathbf{Z} = (Z_{\mathbf{s}_1}, ..., Z_{\mathbf{s}_M})$ is a random vector. Two key assumptions that we will use in the following are second order stationarity and isotropy.

Definition 5.1. Process $\{Z_{\mathbf{s}}, \mathbf{s} \in D\}$ is said *second-order stationary* if the following conditions hold
i) $\mathbb{E}[Z_{\mathbf{s}}] = a$, for all $\mathbf{s}$ in D;
ii) $Cov(Z_{\mathbf{s}_i}, Z_{\mathbf{s}_j}) = \mathbb{E}[(Z_{\mathbf{s}_i} - a)(Z_{\mathbf{s}_j} - a)] = C(\mathbf{h})$, for all $\mathbf{s}_i$,$\mathbf{s}_j \in D$, $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$.

C is called *covariogram* in the context of Spatial Statistic.

Definition 5.2. A second-order stationary process $\{Z_{\mathbf{s}}, \mathbf{s} \in D\}$ is said *isotropic* if the following condition hold

$$Var(Z_{\mathbf{s}_i} - Z_{\mathbf{s}_j}) = 2\gamma(h), \quad h = ||\mathbf{h}|| = ||\mathbf{s}_i - \mathbf{s}_j||, \quad \mathbf{s}_i, \mathbf{s}_j \in D.$$

Otherwise, it is said to be *anisotropic*.

Isotropy is a property that includes second-order stationarity and directional homogeneity. Indeed it is verified when the covariance structure is homogeneous over all the directions of $\mathbb{R}^d$.
2γ is said to be the *variogram*, with γ being *semivariogram*. Covariogram and variogram are related by the following identity if the process is second-order stationary

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}), \quad \mathbf{h} \in \mathbb{R}^d$$

Real applications rely often on second-order stationarity, while for Ordinary Kriging the process should also be isotropic.
The main tool in Spatial Statistic is the variogram, since its knowledge is required both for an exploratory analysis of the phenomenon and for Kriging. The variogram estimation procedure can be divided into two phases: (1) Compute a sample variogram directly from the available data and (2) fit a theoretical model to this sample variogram. Later we will mention some structural properties of the variogram that we now introduce. A valid semivariogram $\gamma(\cdot)$ is symmetric and it may present a discontinuity at the origin, associated to a non-zero limit as h approaches 0

$$\lim_{h \to 0} \gamma(h) = \tau^2 \neq 0 = \gamma(0).$$

τ^2 is called *nugget*. Other structural features of a valid semivariogram are the *sill* and the *partial sill*. These are defined taking the limit of the semivariogram as h approaches infinity.

$$\tau^2 + \sigma^2 = \lim_{h \to +\infty} \gamma(h),$$

where σ^2 being the partial sill and the sum $\tau^2 + \sigma^2$ defined to be the sill. The presence of a finite sill indicates that the process is second-order stationary. The last quantity to be introduced is the *range*, defined as the value of h for which the sill has been reached

$$\gamma(R) = \tau^2 + \sigma^2.$$

The range R has to be interpret as the range of influence of the process: for distances greater than the range, two elements of the process are uncorrelated.
As said, after a sample variogram is computed from the data, a theoretical model is fit to this sample variogram. A number of parametric models for this purpose has been introduced. The simpler one is the *Pure Nugget*, which models the absence of spatial correlation. Then we cite also the *Exponential*, the *Spherical* and the *Matern*, which are more complex models often used in practice.
The main goal for our purpose remains doing predictions and in the geostatistical setting this is done employing kriging techniques. Giving a set of locations $\mathbf{s}_1, ..., \mathbf{s}_M$ in D along with the observations of the process $Z_{\mathbf{s}_1}, ..., Z_{\mathbf{s}_M}$, the interest is to predict an unobserved element $Z_{\mathbf{s}_0}$ at $\mathbf{s}_0$, or to perform prediction over a spatial grid in D. Kriging is a probabilistic approach to solve this problem that produces an interpolation function

based on covariance or variogram model derived from data. The Kriging predictor $Z^*_{\mathbf{s}_0}$ of an element $Z_{\mathbf{s}_0}$ is the Best Linear Unbiased Predictor (BLUP) $Z^*_{\mathbf{s}_0} = \sum_{i=1}^{M} \lambda_i Z_{\mathbf{s}_i} + \lambda_0$, whose weights $\lambda_0, ..., \lambda_m$ solve

$$\begin{aligned} \min \quad & \mathbb{E}[(Z_{\mathbf{s}_0} - Z^*_{\mathbf{s}_0})] \\ \text{subject to} \quad & \mathbb{E}[Z^*_{\mathbf{s}_0}] = \mathbb{E}[Z_{\mathbf{s}_0}] \end{aligned} \tag{2}$$

In our specific scenario, we find ourselves within a second-order stationary setting with an unknown process mean. Consequently, the appropriate method to employ is Ordinary Kriging (OK). Given that our response variable is binary, we will use Indicator Kriging (IK) – an approach that incorporates OK in situations where $Z_{\mathbf{s}}$ is an indicator variable. IK is frequently employed in practical applications where the aim is to predict the probability of a certain variable surpassing a threshold or the presence or absence of an element at a particular location [19, 24]. In these instances, IK tackles the challenge by utilizing the principles of Ordinary Kriging. The optimal weights $\boldsymbol{\lambda}$ in OK are computed solving the following linear system:

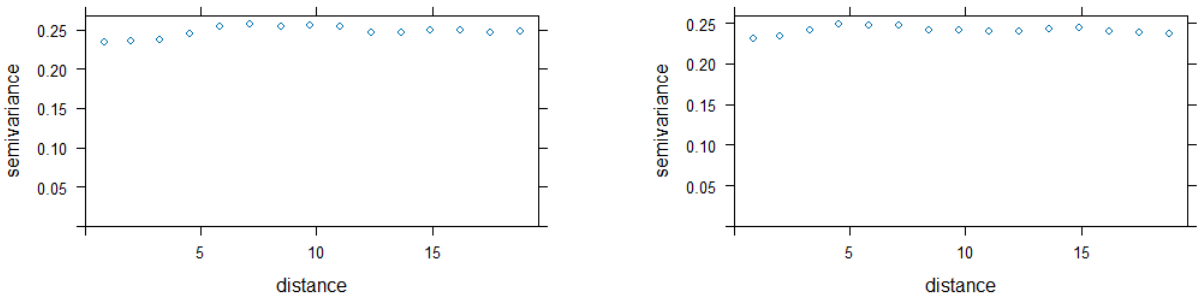
$$\begin{pmatrix} \Sigma & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \xi \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma_0} \\ 1 \end{pmatrix} \tag{3}$$

ξ being a Lagrange multiplier, $\Sigma = [Cov(Z_{\mathbf{s}_i}, Z_{\mathbf{s}_j})]$ and $\boldsymbol{\sigma_0} = [Cov(Z_{\mathbf{s}_i}, Z_{\mathbf{s}_0})]$.

For a complete discussion of these topics we direct the reader's attention to the book *Geostatistics: Modeling Spatial Uncertainty* [7].

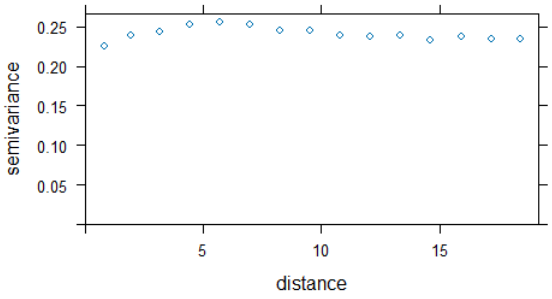
5.2. Data Application

To perform variogram analysis and kriging in `R` we used the package `gstat` [21]. First we focus on variogram estimation on the available shots of each team. The sample variogram computed on data are approximately equal for each team, as we can see in Figure 8, which is to be expected as we consider the teams as a whole. By looking at the graphs, we notice that since there is a well-defined sill, the process turns out to be second-order stationary. On the other hand, correlation seems not to be dependent from the distance, suggesting that the best theoretical model to fit can be a pure-nugget or an exponential/spherical that reaches the sill immediately. It is noteworthy that randomizing the choice of data, i.e. selecting random numbers of matches for a team, the resulting sample variogram does not change. In addition, as introduced before, we have to verify isotropy in order to perform ordinary kriging. In practice, it is done by inspecting the so-called *directional variograms* for a number of fixed directions in $\mathbb{R}^d$. As reported in Figure 9, changing direction does not affect the resulting sample variogram, so also isotropy is verified for teams' shots. To establish a theoretical model, we experiment with several models and subsequently choose the best-fitting one. Our options include the Pure Nugget, Exponential, Spherical, and Matern models. In Figure 10 we report an example of fitted variogram.

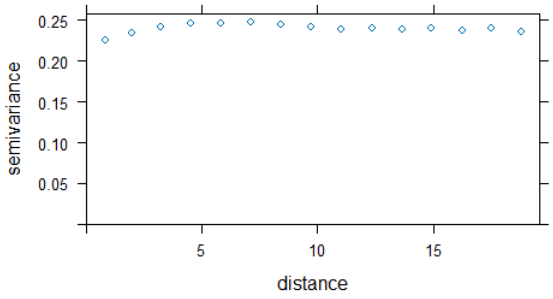


(a) Banco Di Sardegna Sassari

(b) Tezenis Verona



(c) Openjobmetis Varese



(d) Germani Brescia

Figure 8: Four examples of sample variogram computed on shot taken by the teams in 2022/2023 LBA tournament. Below each variogram, the name of the team is indicated.

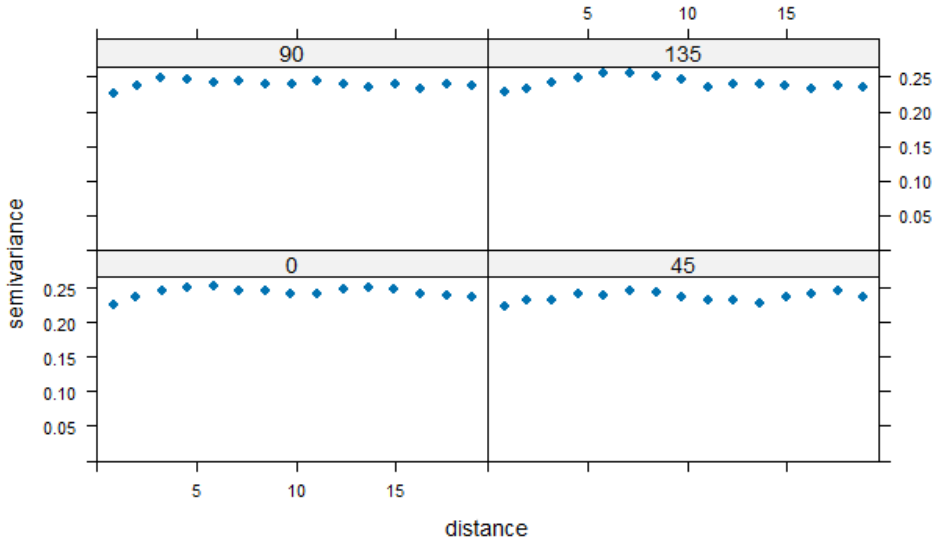


Figure 9: Directional Variograms for shots of Germani Brescia in LBA 2022/2023 tournament. 0,45,90,135 are the directions in the spatial domain. 0 represents the north (y axis).

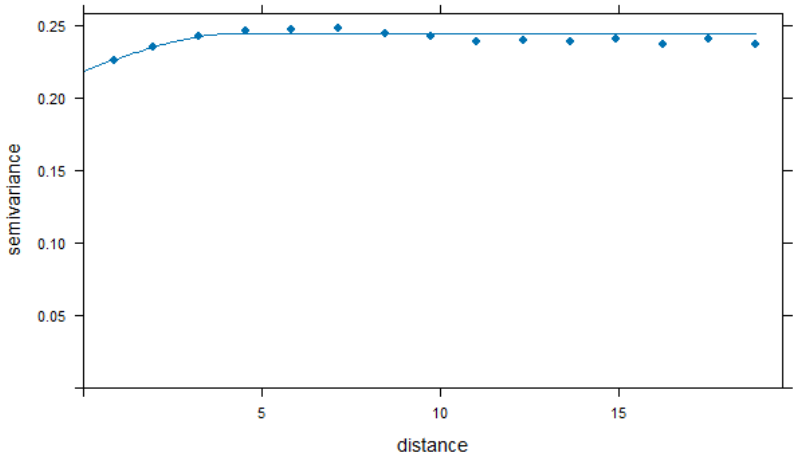


Figure 10: Fitted variogram for shots of Germani Brescia in LBA 2022/2023 tournament. The resulting model is a composition of a nugget and a spherical model.

Once the variogram fitting process is complete, we can leverage it for kriging purposes. Our choice is to apply Indicator Kriging for estimating the probability of making a basket at every point within our 100×100 court grid. As previously discussed in the theoretical introduction, Indicator Kriging predicts the value for a new location $Z_{\mathbf{s}_0}$ by calculating a linear combination of existing observations, with the weights determined by the underlying covariance structure of the problem. In our implementation, we chose to use a subset of 75 observations for calculating predictions at each grid point. This decision was taken to achieve a balance between maintaining graphical quality and ensuring computational efficiency. In Figure 11 we report an example of shooting performance evaluation map produced via Indicator Kriging. The team under analysis is again Germani Brescia, so one could compare properly all the graphs proposed. The pattern highlighted by this map is the same of RF and Adaboost (see Figure 7 and Figure 5 respectively), with a good performance from the baseline and from the wing. In terms of the accuracy of the percentages displayed on the map, we are on par with Adaboost, ensuring that IK can effectively assess the team's shooting capabilities with precision. The significant improvement with respect to Adaboost lies in the computational aspect. Indeed, the process of employing Indicator Kriging to produce the desired map is approximately 3.5 times faster than that of Adaboost.

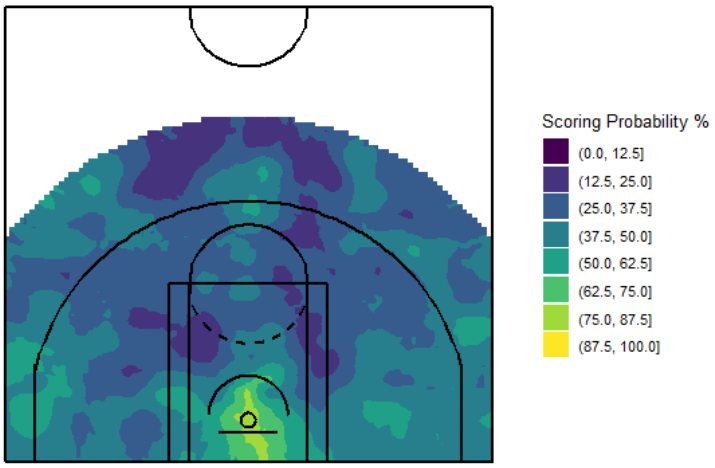


Figure 11: Shooting performance evaluation map built using Indicator Kriging, for shots of Germani Brescia in LBA 2022/2023 tournament.

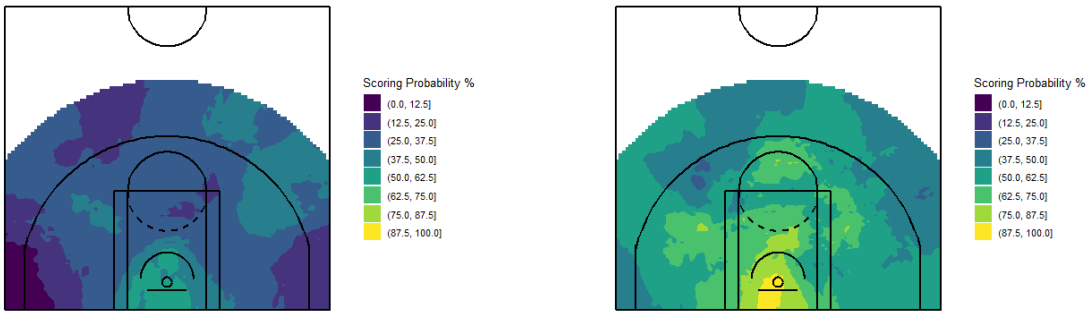
6. Model Extensions

6.1. Categorical Variables

The maps generated through Adaboost and Indicator Kriging, as seen so far, represent a highly valuable tool for identifying even small regions of high or low effectiveness in a team's shooting performance. In Basketball, especially in recent times, an increasing amount of detailed information is being captured about in-game events. Concerning shots, as outlined in Section 2, Pallacanestro Varese has gone beyond the conventional data found in league-provided play-by-plays by collecting other categorical information on each shot. Therefore, it becomes very interesting to investigate the importance of additional variables through the shooting performance evaluation maps. From a statistical perspective, this exploration is done by incorporating the categorical variable of interest into the model alongside the spatial coordinates. By training on the complete dataset, the model discerns the variable importance in predicting shot percentages. In the prediction phase, as many maps as the level of the categorical variable are produced, always using a grid on the court. We report in Figures 12,13,14 and 15 the shooting performance evaluation maps obtained using Indicator Kriging for each additional variable, since most of them are really impressive from a visual point of view. The considered variables are the following:

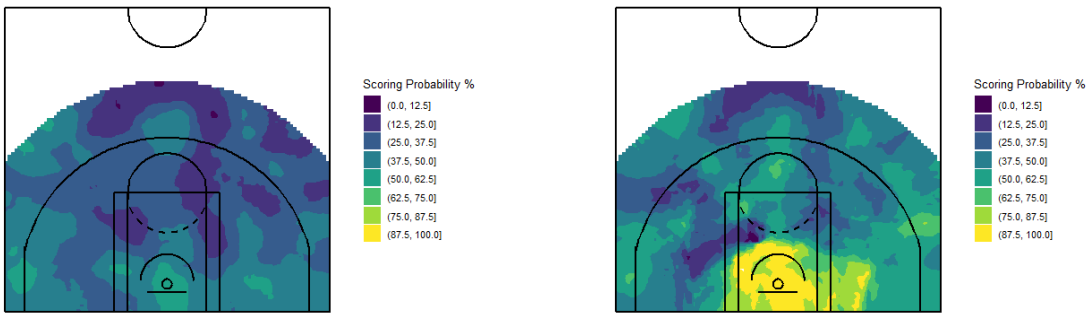
- *contested*: a binary variable indicating whether the shot has been contested by the defense or not;
- *assisted*: a binary variable indicating whether the shot has been assisted by a teammate or not;
- *24clock*: a three-level categorical variable indicating the time on the shot clock when the shot has been attempted. The ranges are: [17,24], [9,16] and [1,8];
- *home*: a binary variable indicating whether the shot has been attempted in a home or away-game;

Let us denote with X_c a generic categorical variable added to the model as predictor in order to fit the variogram. Then, during the prediction phase, L different 100×100 grid are used, being L the number of levels of X_c. Finally, L maps are produced, one for each level $l = 1, ..., L$.



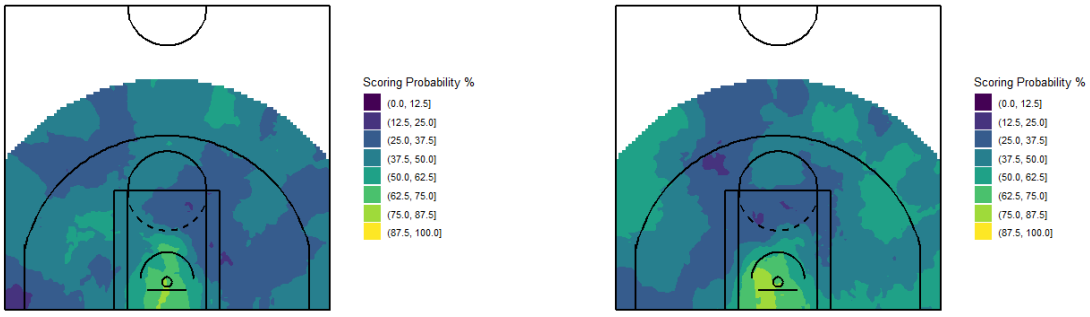
(a) Unassisted Shots (b) Assisted Shots

Figure 12: Shooting Performance Evaluation maps with additional categorical variable. The variable is a binary one indicating whether the shot is assisted or not. Two maps are produced, one predicting the outcome for unassisted shots(Figure 12a) and the other for assisted shots (Figure 12b). Data are all the shots taken by Banco di Sardegna Sassari in the 2022/2023 LBA tournament.



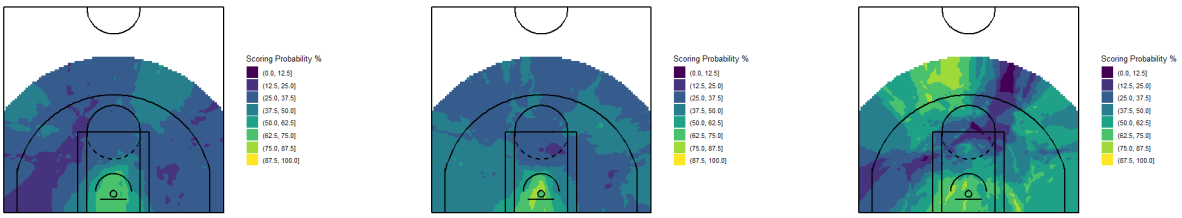
(a) Contested Shots (b) Uncontested Shots

Figure 13: Shooting Performance Evaluation maps with additional categorical variable. The variable is a binary one indicating whether the shot is contested or not by the defense. Two maps are produced, one predicting the outcome for contested shots(Figure 13a) and the other for uncontested shots (Figure 13b). Data are all the shots taken by Germani Brescia in the 2022/2023 LBA tournament.



(a) Away-games Shots (b) Home-games Shots

Figure 14: Shooting Performance Evaluation maps with additional categorical variable. The variable is a binary one indicating whether the shot is taken in an away or home game. Two maps are produced, one predicting the outcome for away-game shots(Figure 14a) and the other for home-game shots (Figure 14b). Data are all the shots taken by Banco di Sardegna Sassari in the 2022/2023 LBA tournament.



(a) Late-Clock Shots (b) Regular-Clock Shots (c) Early-Clock Shots

Figure 15: Shooting Performance Evaluation maps with additional categorical variable. The variable is a 3-level factor indicating whether the shot is taken late in the play (1-8 seconds on the shotclock), in regular time (9-16 seconds) or early (17-24 seconds). Three maps are produced, one predicting the outcome for late-clock shots(Figure 15a), the second for regular-clock shots (Figure 15b) and the other for early-clock shots (Figure 15c). Data are all the shots taken by Givova Scafati in the 2022/2023 LBA tournament.

In Figure 12, a substantial disparity in performance is evident between the two levels of the categorical vari-

able, which indicates whether a shot was assisted by a teammate or not. Throughout various court regions, assisted shots display a notably higher probability of success. This phenomenon can be attributed to the fact that players tend to shoot more rhythmically after receiving a pass. This analytical insight is made possible by our possession of data on assists for both successful and unsuccessful shots, information not available in the league's play-by-play data. Some teams exhibit less pronounced differences between the maps. For these teams, the distinction between assisted and unassisted shots might be less significant, suggesting either a considerable number of missed assisted shots or a proficiency in unassisted shooting.

The second categorical variable considered involves whether shots were contested by the defense or not. To clarify, a shot is classified as contested if a defensive player attempts to interfere with the offensive player's shooting motion. Figure 13 vividly illustrates a significant discrepancy between the two levels. The improvement in performance for uncontested shots holds considerable importance for every team, as capitalizing on open shots is a highly desirable benefit.

In basketball, as well as in the broader realm of sports, playing a home game is often considered advantageous. This is because the home team is more familiar with the court, and the presence of fans can significantly influence the dynamics of a match. Naturally, one might expect the home team to exhibit higher scoring probabilities. This anticipated trend is slightly noticeable in Figure 14. However, when compared to the first two variables examined in this context, the differences observed here are not substantial. Indeed based on our dataset, the shooting probability during home games (46.1%) only slightly exceeds that of away games (45.7%). This suggests that the advantage of playing at home may be found in aspects of the game other than shooting efficiency. In contrast, the disparity for the other two variables is more pronounced: 53.2% versus 39.9% for assisted shots and 58.9% versus 39.7% for uncontested shots. The focus on searching for uncontested and assisted shots represents indeed the primary objectives of a team's offensive plays.

Finally, the last categorical variable considered is a three-level factor concerning the time remaining on the 24-seconds shot-clock when the shot is taken. The division made was the following: (a) late-shots, when remaining time is in range [1,8], (b) regular-shots, range [9,16] and (c) early-shots, range [17,24]. As displayed in Figure 15 early-shots, since taken in fast-break situations [3], are the better ones in terms of efficiency. On the other hand, late-shots are the worst ones, since when the shot-clock is running out bad shots are often taken. For the whole Italian league we notice the following shooting percentage for these three categories: 53.7% for early-shots, 47.8% for regular-shots and 40.8% for late-shots. The information regarding the time remaining on the 24-seconds shot-clock is also present in the play-by-play data provided by the league, but it is in most cases incorrect. Again, having correct data available thanks to Pallacanestro Varese results in an advantage in carrying out consistent analysis.

Incorporating categorical information into an analysis demands a more extensive dataset. This is due to the requirement of training the model to comprehend the impact of each level of the categorical variable on the response one. As a consequence, this form of analysis can be limited when a small number of games are analyzed. It becomes particularly insightful and meaningful when conducted after the championship concludes, while in the middle of it it is better to analyse these variables through tables and numbers.

6.2. Weighted Predictions

The main issue of Basketball shots is their density in the court space, which is not homogeneous. In fact, as it can be seen in Figure 16, the most frequent shot in basketball is the one attempted from a zone close to the basket that represents a very small portion of the court. The other area with a significant density is the three-point-area from which in the last years teams are increasing their number of attempts.

[3] A fast-break situation encompasses all instances where the offense attacks before the defense is fully prepared, so typically in the first seconds of the offensive play.

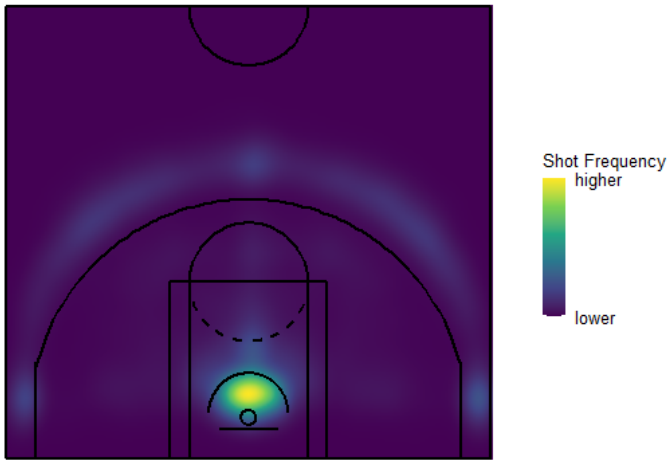


Figure 16: Density of all the shots taken in LBA-2022/2023 season.

Given the distribution of shots across the court, it becomes evident that forecasting the probability of successfully scoring a basket based on its location poses a significant challenge. Moreover, during the analysis of a specific team, it's possible to encounter low-density shot areas in comparison with the league average. In such cases, the model may encounter difficulties in generating meaningful predictions. For us the quantity of interest, that has to be computed for each team $t = 1, ..., 16$, is $P(y_j = 1|x_{h,j}, x_{w,j})$, with j index of the point in the prediction grid. In the following this quantity will be called $pshot_t(\mathbf{s}_j)$, being $\mathbf{s}_j$ the location of the shot, which components are $x_{h,j}$ and $x_{w,j}$, and t the team under analysis. In regions where teams have taken relatively few shots, both Adaboost and Indicator Kriging encounter difficulties in generating meaningful predictions $pshot_t(\mathbf{s}_j)$. Hence, the idea is to formulate a rule that, after the training process, rectifies the model's predictions within the aforementioned low-shot-density areas. We call $pmodel_t(\mathbf{s}_j)$ the probability predicted by the model in location $\mathbf{s}_j$. Furthermore, we introduce an additional probability denoted as $p_{zone}(\mathbf{s}_j)$, which encapsulates a priori information regarding the league's scoring probability within the specific $zone$ to which the location $\mathbf{s}_j$ belongs. This $p_{zone}(\mathbf{s}_j)$ is simply the shooting probability of the whole league within the $zone$, where the shooting probability is the ratio between shots made and shot attempted, while the considered zones are the ones depicted in Figure 2, except the '3W'. $pmodel_t(\mathbf{s}_j)$ and $p_{zone}(\mathbf{s}_j)$ are the two candidates for predicting the scoring probability across the prediction grid with a key distinction: $pmodel_t(\mathbf{s}_j)$ originates from a statistical model trained on the existing shots, whereas $p_{zone}(\mathbf{s}_j)$ is computed directly from the data itself. The method that follows computes a weighted average between these two alternatives, with the weighting factor determined by the frequency of shots attempted in the analysed zone by the team. We say that a team t has pulled a sufficient number of shot from a certain $zone$ if

$$\frac{f_{zone,t}}{f_{zone}} > 1,$$

where $f_{zone,t}$ and f_{zone} are respectively the relative shooting frequency of the team and the league from $zone$. So, for the team t we can define a weight $w_{\text{zone},t}$ computed in the following way:

$$w_{\text{zone},t} = \begin{cases} \frac{f_{\text{zone},t}}{f_{\text{zone}}}, & \text{if } \frac{f_{\text{zone},t}}{f_{\text{zone}}} < 1 \\ 1, & \text{otherwise} \end{cases}, \quad \text{for each} \quad \text{zone} = 1, ..., 5. \tag{4}$$

Then, employing this weight, we can compute our $pshot_t(\mathbf{s}_j)$ as a weighted average of $pmodel_t(\mathbf{s}_j)$ and $p_{\text{zone}}(\mathbf{s}_j)$

$$pshot_t(\mathbf{s}_j) = w_{zone,t} \cdot pmodel_t(\mathbf{s}_j) + (1 - w_{zone,t}) \cdot p_{\text{zone}}(\mathbf{s}_j) \quad \text{for} \quad j = 1, ..., 10000, \tag{5}$$

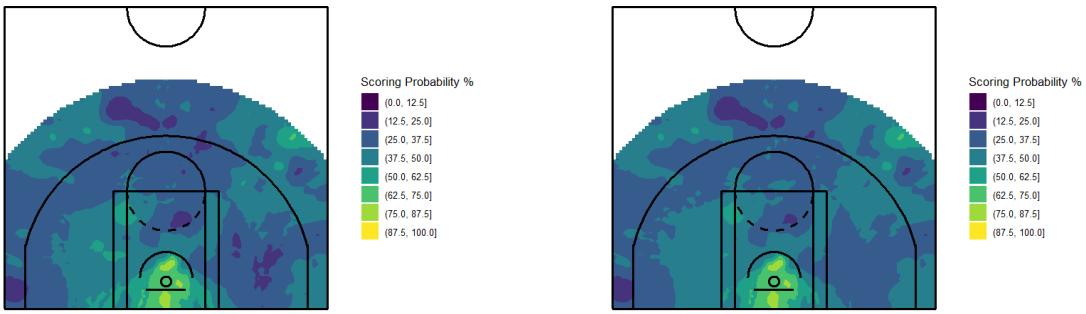
where $zone$ s.t. $\mathbf{s}_j \in zone$ and the weight $w_{zone,t}$ computed as in equation (4).
Through this approach, we can capitalize on the concept of employing the model-predicted probability solely when the analysed team has attempted a substantial number of shots within a specific zone. Alternatively, if the shot count is inadequate, a predefined league-value is employed to compute the prediction. $pshot_t(\mathbf{s}_j)$ is then used to produce the shooting performance evaluation map. This procedure is established based on the premise that predictions generated by the statistical models ($p_{\text{model},t}(\mathbf{s}_j)$) are constructed using the information contained within the existing data. Consequently, when dealing with court regions where the analysed team has failed to generate a significant number of shots, relying solely on the model for predictions in those areas

could potentially be insecure.

To provide a clearer comprehension of the outlined process, we present the output of the model for Openjobmetis Varese in Table 1. This display includes only 5 rows out of a total of 10,000, serving as illustrative examples for each of the 5 selected zones. As evident from the table, the team's relative frequency is below the league value in the 'NON PAINT' and 'PAINT' zones. As a result, for computing the final prediction $pshot_t(\mathbf{s}_j)$, both the model-predicted probability ($pmodel_t(\mathbf{s}_j)$) and the a-priori value ($p_{zone}(\mathbf{s}_j)$) are used there. In the remaining zone, where the weight $w_{zone,t}$ is 1, only the model-predicted probability is employed. Throughout this explanation, the model has been discussed in a generalized context, as this procedure is applicable to all the previously used models (Adaboost and Indicator Kriging). Figure 17 illustrates a comparison between the Indicator Kriging map and the weighted Indicator Kriging map. The difference, as shown in Table 1, can be found only in 'NON PAINT' and 'PAINT' zones.

loc	*zone*	$f_{zone,t}$	f_{zone}	$w_{zone,t}$	$p_{zone}(\mathbf{s}_j)$	$pmodel_t(\mathbf{s}_j)$	$pshot_t(\mathbf{s}_j)$
s1	3C	0.12	0.08	1	0.39	0.44	0.44
s2	3L	0.36	0.33	1	0.36	0.38	0.38
s3	NON PAINT	0.03	0.14	0.19	0.38	0.28	0.36
s4	PAINT	0.15	0.18	0.84	0.41	0.35	0.36
s5	RIM	0.34	0.27	1	0.68	0.70	0.70

Table 1: Example of output of the weighted model for shots taken by Pallacanestro Varese in 2022/2023 LBA tournament. Only one location (loc) example for each zone is reported. Indicator Kriging model prediction are used as $pmodel_t(\mathbf{s}_j)$.



(a) Map produced via Indicator Kriging (b) Map produced via Weighted Indicator Kriging

Figure 17: Comparison between shooting performance evaluation maps produced via Indicator Kriging and Weighted Indicator Kriging. Data are all the shots taken by Openjobmetis Varese in 2022/2023 LBA tournament.

This approach is particularly tailored for the Shiny application, where the user's high level of customization necessitates the consistent generation of meaningful maps on every occasion. For us, this adjustment in predictions serves as a rapid solution for addressing situations where predictions might lose their meaningfulness. It is important to note that this approach is not a rigorously designed statistical model, but rather a pragmatic solution customized for this particular issue. From our perspective, this type of reasoning could lay the groundwork for approaching the problem in a novel manner, taking into account the spatial distribution of attempted shots across both densely and sparsely populated court areas.

7. Model Comparison

In this section, we delve into a comparison of the various models designed to create shooting performance evaluation maps. These maps offer valuable insights into a team's or player's efficiency in scoring across different

areas of the basketball court. As the demand for accurate and actionable insights from these maps increases, it becomes essential to evaluate the performance of different modeling techniques in capturing the intricate facets of shooting behavior. It is important to understand that our aim is not primarily focused on creating a model for optimal performance on new data or striving for high accuracy on training data. Instead, the aim is to create a metric that matches our goal: producing high-quality maps that accurately represent the team's shooting behavior. In the upcoming analysis, we will use the approach introduced by Zuccolotto et al. [30], as their methodology aligns with the objective of the maps. Essentially, they devised an index that rewards maps which exhibit two characteristics: (1) low variance in the neighbourhood of each grid point, signifying visual homogeneity, and (2) a cumulative probability distribution significantly divergent from that of a uniform distribution in the interval [0,1]. The second aspect aims to ensure that maps do not have an excessive number of predictions clustered around 0 and 1, as these extreme probabilities might not accurately reflect a team's actual shooting behavior. The index, denoted as Φ, is calculated as the ratio of these two attributes (see Formula 6). The numerator represents the first characteristic, indicating that a lower value results in low variance and visually homogeneous maps. On the other hand, the denominator pertains to the second attribute, where higher values signify a departure from the uniform distribution. Therefore, a lower Φ index corresponds to a more favorable map based on this metric.

$$\Phi = \frac{\sigma_N}{H}, \tag{6}$$

where σ_N is defined as:

$$\sigma_N = \sqrt{\frac{1}{g}\sum_{i=1}^{g}\sigma_{N_i}^2}, \tag{7}$$

being g is the number of grid points and σ_{N_i} is the standard deviation of the scoring probabilities estimates of the points adjacent in space to the ith grid point.
While H is defined as:

$$H = \sup_y |\hat{F}(y) - F_U(y)|, \tag{8}$$

where $F_U(y)$ is the cumulative distribution function of a Uniform random variable and $\hat{F}(y)$ is the empirical distribution function of the estimated scoring probabilities of a given map.
For a comprehensive explanation of the index, we direct the reader to the detailed description provided in [30].

The models investigated in this study are Random Forest, Adaboost, and Indicator Kriging. The chosen approach for comparing these models about this index involves assessing their performance across different teams. To achieve this, we calculated the Φ index for each model for every team. The resulting Φ values for each team are illustrated in Figure 18. The preferred model, following this evaluation, is the one with lower Φ values, which, in this case, is Indicator Kriging the most of the times. Also Adaboost performs well for our task, having lower value of Φ 4 times out of the total 16. On average, Indicator Kriging emerges as the top performer with a mean index value of 0.09, followed by Adaboost with 0.11 and Random Forest with 0.20. As seen the maps generated by these two models are quite visually engaging and effectively depict the subtle differences in shooting capabilities among the teams. As a result, they prove to be valuable tools in addressing the specific challenges we are focused on.

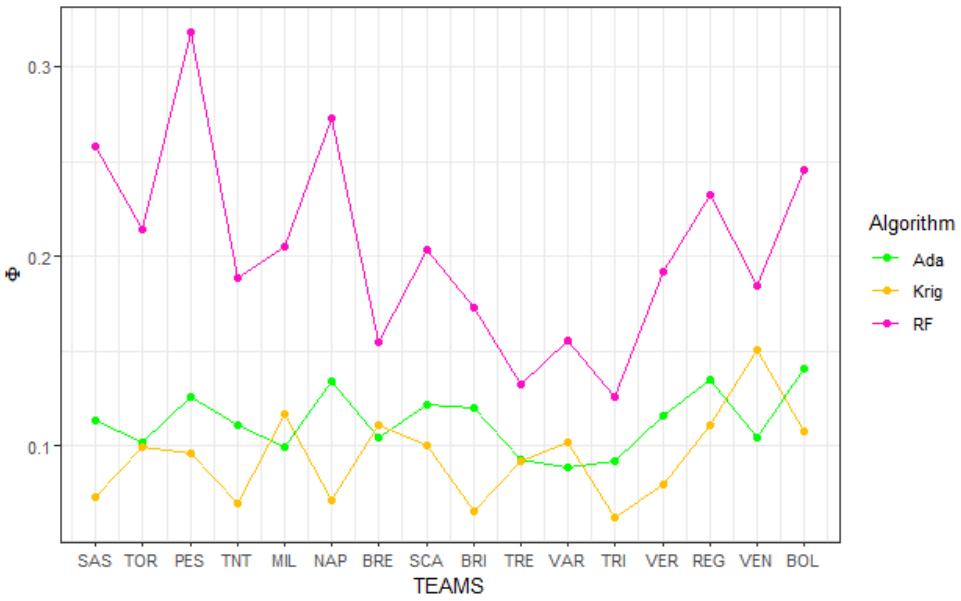


Figure 18: Comparison between Random Forest, Adaboost and Indicator Kriging using the Φ index for each one of the 16 LBA teams.

Based on this index analysis and considering the significantly faster speed of Indicator Kriging compared to Adaboost, it seems that IK is the most suitable method for creating visually appealing and accurate shooting performance evaluation maps. However, it is important to note that the assumptions underlying Indicator Kriging are not easy to satisfy. Furthermore, in a spatial context like a basketball half-court, constructing a robust geostatistical model relies heavily on having sufficient and well-distributed data, which is not always the case. Indeed it is not surprising that for a team like Openjobmetis Varese with an extreme shot distribution pattern (illustrated in Figure 3a), the Adaboost model yields lower Φ index value (Figure 18) than IK.
Therefore, it is challenging to determine a clear winner among the models. Ultimately, we can conclude that both Indicator Kriging and Adaboost, despite their individual shortcomings, are well-suited tools for effectively tackling our problem.
As we introduced the model with weighted prediction in Section 6.2, we need to compare it to the original procedure using the Φ index to gain a comprehensive understanding of the methods. We performed this comparison for Indicator Kriging in Figure 19 and for Adaboost in Figure 20. The results show that employing weighted prediction does not significantly impact the visual performance of the model. Thus, the weighted approach, which refines predictions in zones with limited data, proves to be a good choice since depending solely on the model in such zones could potentially lead to misleading conclusions about shooting performance.

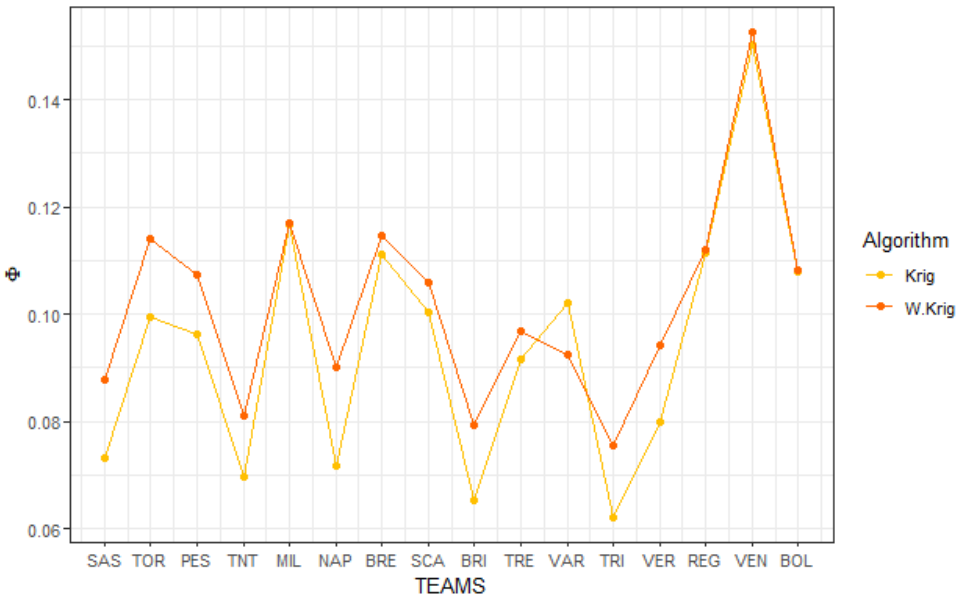


Figure 19: Comparison between Indicator Kriging (Krig in the graph's legend) and Indicator Kriging with weighted predictions (W.Krig in the graph's legend), using Φ index for each one of the 16 LBA teams.

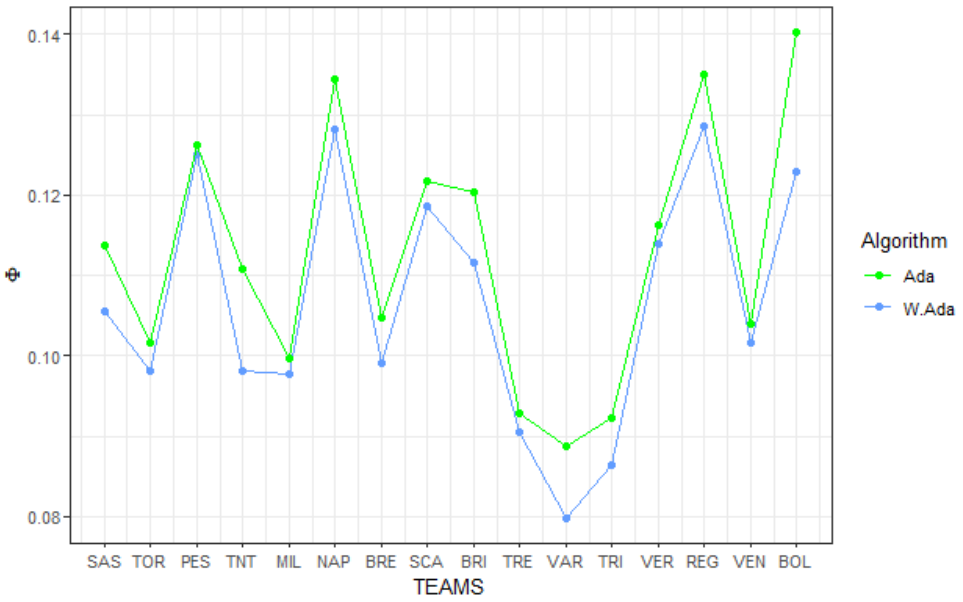


Figure 20: Comparison between Adaboost (Ada in the graph's legend) and Adaboost with weighted predictions (W.Ada in the graph's legend), using Φ index for each one of the 16 LBA teams.

8. Concluding remarks

In this study, we build on the insights of Zuccolotto et al. [29, 30] to delve into the shooting performance analysis of both teams and players. Our aim is to create spatial maps that not only offer precise insights into shooting performance but also possess visual appeal. Notably, the data used for this effort is meticulously collected by Pallacanestro Varese's analytical team, resulting in a comprehensive dataset for evaluating the shooting performance of the 2022/2023 Italian League. Our first innovative modelling approach involves the usage of Adaboost, a boosting technique that outperforms bagging methods in closely fitting data. This method generates a prediction grid that closely aligns with the actual team or player performance and gives rise to eye-appealing maps. Additionally, we employ geostatistical techniques, treating the mid-court as a spatial domain, and apply Indicator Kriging to craft visually striking and accurate maps. We also incorporate categorical

variables to yield diversified shot maps based on the chosen variable, revealing significant insights particularly in the context of assisted/unassisted or contested/uncontested shots. To address prediction challenges in areas with low shooting frequencies, we introduce a technique to rectify model predictions in these regions. Lastly, using an index proposed by Zuccolotto et al. [30], we compare the models, highlighting their strengths and weaknesses. Adaboost, despite the computational complexity of its sequential training procedure, consistently delivers favorable outcomes across nearly all cases without being reliant on specific assumptions. On the other hand, Indicator Kriging produces visually striking maps and operates faster than Adaboost. However, it rests upon substantial assumptions and necessitates a more homogeneous shot distribution for generating robust results. Additionally, a Shiny application has been developed, as outlined in Appendix A, to facilitate the exploration of all the generated maps. This interactive tool offers panels for evaluating and comparing the shooting performance of both teams and players using our maps, scatter plots, and tables.

These maps represent an innovative tool in the field of Basketball analytics, providing a fresh perspective on shooting performance through spatial visualization. Unlike conventional approaches that depend solely on tables and figures, these maps have the capability for example to uncover subtle areas of notable skill within zones of lower effectiveness. Moreover, their integration with categorical variables could prove invaluable for coaches and scouts, enabling them to design tailored training routines and strategic game plans.

The primary constraint of this study lies in the necessity for a substantial and well-distributed dataset across the basketball court to generate accurate maps. Achieving this homogeneity in data distribution is challenging, particularly given the relatively small sample sizes in basketball matches. While the weighted approach partly mitigates this limitation, in scenarios with limited data availability, utilizing traditional methods such as tables and scatter plots might mitigate the risk of performing unreliable inference.

Discussing this work, we have exclusively focused on team-level results due to the greater complexity associated with analysing individual player performances. Players' datasets are typically limited, if we consider a single tournament of 30 games like LBA, and each player has a distinctive playing style translating into a unique shooting density across the court. While we attempted to generate shooting performance evaluation maps for players, they frequently lack meaningfulness due to the sparse input of shots for each player. As of now, we believe that tables and numerical statistics provide a clearer understanding of players' shooting performance.

Several potential areas for future research have been left unexplored in this study. These include: (1) Spatial Point Patterns analysis, a technique commonly used in the Spatial Statistics framework [12, 13]. This method focuses on characterizing the distribution of points in space and making inferences about the underlying process that generated the observed pattern. To adapt this technique to basketball shot analysis, each spatial point (i.e., shot) would need to be marked with its binary outcome (successful or not), leading to have a marked spatial point process. (2) Gaussian Process models for spatial data, whose aim is to model the spatial dependencies through a Gaussian Process [8]. This approach is typically integrated into a Bayesian framework to address our problem, establishing a spatial prior for the covariance structure of made/missed shots across the court. The weighted predictions approach developed in Section 6.2 is a specific and problem-dependent solution, which could potentially serve as a foundation for more comprehensive Bayesian analysis in the future. (3) Mixed-effects models [11] offer a natural extension of our work. These models use the entire dataset of shots, encompassing all teams, and introduce a 'mixed effect' component by treating the team identity hierarchically. In this approach, when predicting the scoring probabilities for a specific team t within an area of low shot density, the model could automatically incorporate a league-wide value. This not only enhances the statistical robustness but also extends the concept described in Section 6.2.

References

[1] Esteban Alfaro, Matias Gamez, and Noelia Garcia. adabag: An r package for classification with boosting and bagging. *Journal of Statistical Software*, 54:1–35, 2013.

[2] Chris Anderson and David Sally. *The numbers game: Why everything you know about soccer is wrong*. Penguin, 2013.

[3] Zhongbo Bai and Xiaomei Bai. Sports big data: management, analysis, applications, and challenges. *Complexity*, 2021:1–11, 2021.

[4] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[5] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[7] Jean-Paul Chiles and Pierre Delfiner. *Geostatistics: modeling spatial uncertainty*, volume 713. John Wiley & Sons, 2012.

[8] Andrew O Finley, Abhirup Datta, and Sudipto Banerjee. spnngp r package for nearest neighbor gaussian process models. *arXiv preprint arXiv:2001.09111*, 2020.

[9] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[10] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[11] Andrzej Gałecki, Tomasz Burzykowski, Andrzej Gałecki, and Tomasz Burzykowski. *Linear mixed-e ects model*. Springer, 2013.

[12] Anthony C Gatrell, Trevor C Bailey, Peter J Diggle, and Barry S Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, pages 256–274, 1996.

[13] Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

[14] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63:3–42, 2006.

[15] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[16] Julian Hatwell, Mohamed Medhat Gaber, and R Muhammad Atif Azad. Ada-whips: explaining adaboost classification with applications in the health sciences. *BMC Medical Informatics and Decision Making*, 20(1):1–25, 2020.

[17] Michael Lewis. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.

[18] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[19] Wonho Oh and Brent Lindquist. Image thresholding by indicator kriging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):590–602, 1999.

[20] Dean Oliver. *Basketball on paper: rules and tools for performance analysis*. U of Nebraska Press, 2011.

[21] Edzer J Pebesma. Multivariable geostatistics in s: the gstat package. *Computers & geosciences*, 30(7):683–691, 2004.

[22] Marco Sandri, Paola Zuccolotto, Marica Manisera, Maintainer Marco Sandri, P Zuccolotto, and M Manisera. The r package basketballanalyzer. *Basketball Data Science: with Applications in R*, 2020.

[23] Robert E Schapire. Explaining adaboost. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 37–52. Springer, 2013.

[24] Jeffrey L Smith, Jonathan J Halvorson, and Robert I Papendick. Using multiple-variable indicator kriging for evaluating soil quality. *Soil Science Society of America Journal*, 57(3):743–749, 1993.

[25] Terry Therneau, Beth Atkinson, and Brian Ripley. rpart: Recursive partitioning and regression trees. *R package version*, 4:1–9, 2015.

[26] Hadley Wickham. *Mastering shiny*. " O'Reilly Media, Inc.", 2021.

[27] Yanli Wu, Yutian Ke, Zhuo Chen, Shouyun Liang, Hongliang Zhao, and Haoyuan Hong. Application of alternating decision tree with adaboost and bagging ensembles for landslide susceptibility mapping. *Catena*, 187:104396, 2020.

[28] Paola Zuccolotto and Marica Manisera. *Basketball data science: With applications in R*. CRC Press, 2020.

[29] Paola Zuccolotto, Marco Sandri, and Marica Manisera. Spatial performance indicators and graphs in basketball. *Social Indicators Research*, 156:725–738, 2021.

[30] Paola Zuccolotto, Marco Sandri, and Marica Manisera. Spatial performance analysis in basketball with cart, random forest and extremely randomized trees. *Annals of Operations Research*, 325(1):495–519, 2023.

A. Shiny Application: maps visualization

Given our goal to create shooting performance evaluation maps and the uniqueness of our precise dataset, we have developed a Shiny application (https://b9fagl-mirkoluigi99.shinyapps.io/census-app/). This holds particular significance in our context, given our comprehensive analysis of the complete LBA 2022/2023 season using a detailed and extensive dataset. Since our textual explanation could not include maps for all teams, an interactive dashboard to display team and player maps was essential. Using the `R` package Shiny [26], we have crafted an application with multiple panels. Each panel empowers users to view our maps with a high degree of customization. In the following we provide an overview of each panel.

- **Documentation:** Here the user can find a brief introduction of the Shiny Application and the other panels.
- **Shooting Performance Analysis:** This panel offers a complete overview on the shooting performance of the selected team. Our approach for offering a comprehensive insight into team shooting performance entails two main elements: (1) a table and graph that present basic statistics for each court sector, with graph sectors color-coded according to shooting percentages, and (2) a shooting performance evaluation map created using Indicator Kriging. While the table and graph provide a descriptive overview of the team's performance (in the table also league values are displayed), highlighting preferred shooting zones, the map excels in revealing subtle variations in shooting behavior, sometimes leading to more intricate insights. An illustrative example for EA7 Emporio Armani Milano, which is the winning team of the 2022/2023 LBA tournament, can be found in Figure 21, as well as Table 2.
- **1vs1:** Within this panel, users have the ability to choose two teams and specify the range of games they wish to analyse. It is also feasible to select the same team and compare their performance in two distinct periods of the season. As a result, the output includes two maps generated using Indicator Kriging, along with two tables like the one showcased in Table 2. By Comparing teams using both tables and maps, insightful conclusions about their shooting patterns can be drawn.
- **Splits:** This panel is dedicated to categorical variables. Users are required to choose a specific category and a team. Subsequently, a set of maps is generated corresponding to the different levels of the chosen categorical variable. By exploring this panel, users can employ the richness and uniqueness of our dataset to assess the influence of various categories on teams' shooting probabilities.
- **Players:** Within this panel, maps for evaluating players' performance can be viewed alongside their scatter plot and corresponding tables. The selection includes players who have attempted more than 200 shots throughout the season. However, even with this amount of data, it can be challenging for the model to generate meaningful scoring probability predictions for the half-court. Despite this limitation, we offer the opportunity to investigate players performances within our shiny application.
- **Models Comparison:** This section provides the opportunity to compare the statistical models used to address our challenge. Three maps are generated using Adaboost, Random Forest, and Indicator Kriging, respectively. The fourth map illustrates shooting statistics across predefined court sectors. Through this comparison, it becomes feasible to determine whether our shooting performance evaluation maps align with the actual team performance.

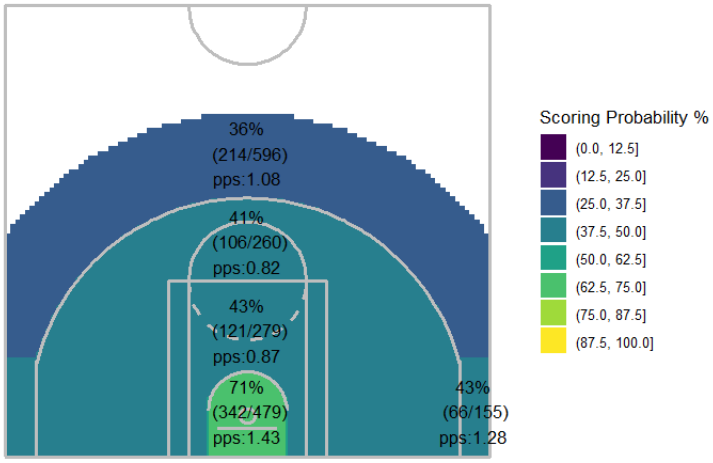
Users are also empowered to select whether to display the Field Goal Percentage (FG%) or the Effective Field Goal Percentage (EFG%) on the map. The FG% is the traditional measure of scoring probability, calculated as the ratio of successful shots to total attempts. In contrast EFG%, that is one of the Four Factors of Basketball Success introduced by Dean Oliver in his book [20], takes into account the value of the shots, that could be either 2 or 3. Let $2PM$ and $3PM$ the number of two-point and three-point shots made and FGA the total number of shots attempted. Then

$$FG\% = \frac{2PM + 3PM}{FGA},$$

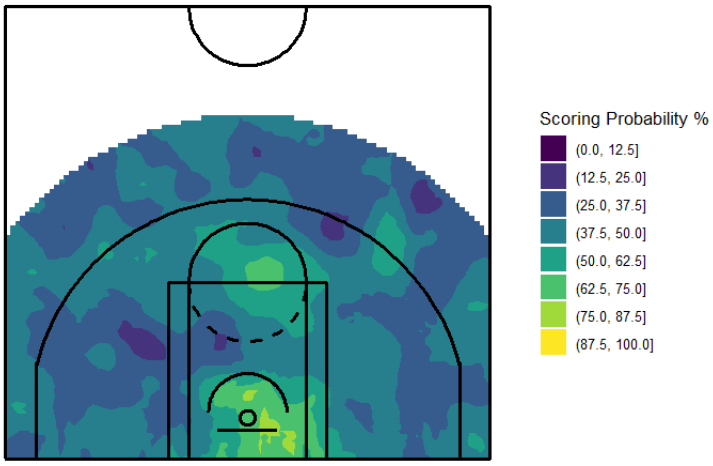
And

$$EFG\% = \frac{2PM + 1.5 \times 3PM}{FGA}.$$

In recent years, when analysing basketball matches, coaches and analytics tend to focus more on EFG% rather than the classic FG%, since it could capture better the shooting efficiency of the teams. We believed that incorporating this specific aspect would make our Shiny application more comprehensive in the analysis of shooting performance.



(a) Statistics by sector



(b) Indicator Kriging map

Figure 21: Example of complete shooting performance evaluation for shots taken by EA7 Emporio Armani Milano in 2022/2023 LBA tournament: At the upper portion (21a), there are statistics categorized by court sector, color-coded to indicate shooting percentages. Directly below (21b), there's a map generated through Indicator Kriging. This map shows the probability of scoring from each point on the court, given the available data. In the upper graph the following statistics are displayed from the top to the bottom: shooting percentage; (shots made/ shots missed); point per shot, i.e. (shots made/ shots missed) multiplied by the shot's points (2 or 3).

	RIM	3L	3C	PAINT	NON PAINT
FGM/FGA	342/479	214/596	66/155	121/279	106/260
FG%	71.4%	35.9%	42.6%	43.4%	40.8%
AVG FG%	67.9%	35.9%	39.3%	41.0%	37.9%
PPS	1.43	1.08	1.28	0.87	0.82
AVG PPS	1.36	1.08	1.18	0.82	0.76
FREQ	27.1%	33.7%	8.8%	15.8%	14.7%
AVG FREQ	26.7%	33.0%	7.6%	18.3%	14.4%

Table 2: Example of complete shooting performance evalutaion for shots taken by EA7 Emporio Armani Milano in 2022/2023 LBA tournament: table representing basic statistics divided by court-zones. FGM/FGA: shots made / shots missed; FG%: scoring probability, i.e. shots made / shots missed in percentage; AVG FG%: scoring probability for the whole league; PPS: point per shot, i.e. FG% times shot's points (2 or 3); AVG PPS: point per shot for the whole league; FREQ: relative shooting frequency from that zone; AVG FREQ: relative shooting frequency from that zone of the whole league.

Abstract in lingua italiana

Questo lavoro tratta il tema delle Basketball Analytics, concentrandosi sulla valutazione della performance al tiro mediante mappe innovative che, per ogni punto del campo, mostrano la probabilità di segnare un canestro. Queste mappe possono cogliere piccole aree a bassa o alta efficacia nella performance di tiro, fornendo utili analisi nel contesto della Pallacanestro. Esse sono create tramite un processo inferenziale: un modello viene addestrato sui dati di tiro, rappresentati dalle loro coordinate e dall'esito binario, e poi viene impiegato per produrre le previsioni per ogni punto sul campo. Proponiamo due approcci principali, il primo basato sul metodo Adaboost, che è una tecnica di boosting che utilizza alberi decisionali come modelli di base, e l'altro basato su Indicator Kriging, una tecnica di previsione nel contesto della geostatistica. Discutiamo anche la possibilità di aggiungere variabili categoriche al modello, portando alla creazione di mappe visivamente impattanti. Successivamente, affrontiamo la questione della densità non omogenea dei tiri sul campo da basket. Tutti i modelli impiegati vengono poi confrontati utilizzando un indice in grado di valutare la bontà grafica delle mappe. L'applicazione dei dati coinvolge un dataset estremamente ricco e preciso contenente tutti i tiri effettuati nel primo campionato Italiano di basket (LBA) nella stagione 2022/2023.

Parole chiave: Basketball Analytics, Analisi della Performance, Adaboost, Geostatistica, Indicator Kriging

Acknowledgements

Vorrei ringraziare il Professore Andrea Cappozzo, mio relatore in questa tesi, per essersi dimostrato estremamente disponibile ed entusiasta e per avermi garantito il massimo supporto. Indubbiamente, aver avuto al mio fianco una persona così, contribuirà a rendere più lieto il ricordo dell'esperienza di tesi.

Ringrazio le Professoresse Paola Zuccolotto e Marica Manisera per avermi dato la possibilità di lavorare con loro. Sono state prima di tutto fonte di ispirazione, tramite le loro pubblicazioni in materia di Basketball Analytics, e il fatto di continuare il loro studio riguardande l'analisi della performance di tiro è stato per me motivo di grande onore.

Un grande ringraziamento va alla Pallacanestro Varese, per avermi permesso di fare un'esperienza per me fondamentale, sia dal punto di vista della tesi sia per il mio percorso da allenatore. Ringrazio nello specifico Luis Scola, CEO, per avermi voluto e accolto con grande entusiasmo, e Luca Cappelletti, riferimento n.1 in Italia in ambito Basketball Analytics, per avermi insegnato tanto e per il grande supporto in fase di elaborazione della tesi. Avere avuto la possibilità di vivere dall'interno una realtà così innovativa ha sicuramente generato in me grande voglia di portare avanti ricerche in questo ambito.

Ringrazio la mia famiglia e i miei nonni, che mi hanno sempre supportato facendo in modo che coniugassi al meglio studio e passione. La dedizione al lavoro e la forza di volontà nel portare avanti sempre con successo scuola e sport la devo sicuramente a loro.

Tra gli amici ci tengo a ringraziare Luca F e Caterina, due persone per me fondamentali che hanno la grande capacità di farmi sentire sempre importante e apprezzato. Inoltre ringrazio Luca C e Francesca, compagni prima di liceo e poi di università, con cui ho condiviso tanto e che hanno contribuito a rendere questi anni un qualcosa che ricorderò sempre con felicità.

Un ringraziamento finale è dovuto al privilegio di aver potuto dedicarmi alla mia passione, la Pallacanestro, nella chiusura del mio percorso di studi.