Computational Statistics (2025) 40:1731–1751
https://doi.org/10.1007/s00180-024-01564-4

ORIGINAL PAPER

Scoring probability maps in the basketball court with Indicator Kriging estimation

Mirko Luigi Carlesso[1] · **Andrea Cappozzo**[2] · **Marica Manisera**[1] · **Paola Zuccolotto**[1]

Received: 15 March 2024 / Accepted: 16 September 2024 / Published online: 7 November 2024

Abstract

Measuring players' and teams' shooting performance in the basketball court can give important information aimed to the definition of both game strategies and personalized training programs. From a methodological point of view, the estimation of the scoring probability can be faced by resorting to different tools in the field of statistical or algorithmic modelling. As a matter of fact, the most natural theoretical framework for this problem is that of spatial statistics, with the particularity that the analysis is based on the binary measurement variable informing about whether a shot is made or missed. In this paper we propose the use of spatial statistics tools suited to this specific context, namely lorelograms to investigate the spatial correlation and Indicator Kriging to draw scoring probability maps. A structured case study is presented, dealing with all the teams of the Italian Basketball First League, based on a non-public dataset containing substantive additional information, that allows interesting insights about assisted and uncontested shots.

Keywords Scoring probability · Shooting performance · Spatial statistics · Lorelogram · Indicator Kriging

✉ Andrea Cappozzo
andrea.cappozzo@unimi.it

Mirko Luigi Carlesso
mirko.carlesso@unibs.it

Marica Manisera
marica.manisera@unibs.it

Paola Zuccolotto
paola.zuccolotto@unibs.it

[1] Big&Open Data Innovation Laboratory (BODaI-Lab), University of Brescia, Brescia, Italy

[2] Department of Economics, Management, and Quantitative Methods, University of Milan, Milan, Italy

1 Introduction

Sports analytics is becoming increasingly important due to the expanded availability of sports data and a slow but steady change in mindset among sport professionals. A rising amount of data is available from various sources, both conventional and technologically advanced, and user-friendly softwares and applications able to provide analysis of this data to sports professionals are becoming more and more accessible. Connected with this, professionals across different sports domains progressively recognize the importance of extracting insights from data for several purposes. These include pinpointing factors influencing performance, devising effective game strategies, and customizing training sessions, among others. Consequently, a scientific literature has emerged within the realms of statistics, data science, and more generally in sports sciences, featuring books and articles dedicated to the various objectives of analytics in different sports. Focusing solely on the topic of basketball analytics, the literature is very extensive and includes, among many others, Albert et al. (2017), Oliver (2004), Kubatko et al. (2007), Passos et al. (2016), Zuccolotto et al. (2018), García et al. (2013), Bianchi et al. (2017), Passos et al. (2011), Lamas et al. (2011), Bornn et al. (2017), Metulini et al. (2018), Wu and Bornn (2018), Lopez and Matthews (2015), Ruiz and Perez-Cruz (2015), Sandri et al. (2020), Skinner and Goldman (2017), Santos-Fernandez et al. (2022), Macis et al. (2023); for a comprehensive review in basketball analytics, see Zuccolotto and Manisera (2020).

In this paper, we focus on spatial performance analysis in basketball. Performance analysis generally provides critical insights into players' strategies and game dynamics. In basketball, it can encompass various aspects, such as shooting performance, player movements, passing and ball movement, scoring trends, lineup analysis etc.

Studying shooting performance and, more generally, scoring trends in basketball involves analysing scoring probability, effectiveness of shooting, as well as patterns in how points are scored throughout a tournament or a match, in different game scenarios, for example in high-pressure situations, and within space. A spatial analysis of the shooting performance can yield intriguing results, shedding light on the specific spatial patterns and tendencies of players' shooting accuracy across the court.

Conventional basketball analytics predominantly rely on elementary statistical approaches to delineate spatial performance. These methods often involve employing fundamental metrics like shooting percentages, computed within predefined court sections, such as squares or slices (some functionalities for this kind of analysis are accessible in the R package BasketballAnalyzeR Sandri, 2020). Graphically, these basic statistics are commonly visualized through heatmaps of the basketball court, which use varying colors to identify regions exhibiting different shooting frequencies, offering a simplistic depiction of shooting intensity across the court.

Recently, spatial analyses of basketball performances have expanded to consider not only offensive statistics but also defensive metrics (among others, see Franks et al., 2015). Indeed, thanks to new technologies and systems used for player tracking, it is possible to obtain more accurate data on defensive performance compared to traditional defensive statistics, which are often affected by human biases and inaccuracies.

Sophisticated statistical and machine learning approaches have also been proposed to represent spatial structures. For example, Miller et al. (2014) focused on modeling the player shooting behaviour in space, by means of non-negative matrix factorization (NMF); using a time series of basketball players' coordinates; Metulini et al. (2018) investigated the relationship between patterns of space among players and team performance, from both an offensive and a defensive perspective; López Hernández et al. (2013) used a spatial cluster analysis technique in order to identify and visualize teams' or players' shooting patterns; Santos-Fernández et al. (2022) proposed to investigate the hidden dynamics of basketball interactions in space and time by means of a Bayesian mixture model.

Recent advancements in basketball analytics have introduced innovative methods for spatial perfomance analysis. Zuccolotto et al. (2021) proposed to use CART (Classification And Regression Trees, Breiman et al. (1984)) to segment the court into optimal rectangles with respect to a given player or team shooting performance. Building upon Zuccolotto et al. (2021), Zuccolotto et al. (2023) further developed this idea, by (*i*) replacing rectangles into slices (aligning more closely with the basketball court's geometry) obtained through the use of shot coordinates expressed in a polar coordinate system and (*ii*) mitigating the CART's inherent instability by resorting to ensemble learning algorithms (Random Forests, Breiman, 2001 and Extremely Randomized Trees, Geurts et al., 2006). These novel approaches estimate the conditional probability function for the binary outcome describing the made or missed shot (given the shot location), resulting in maps that illustrate the player's or team's scoring probability across the court.

In this paper, we introduce a novel approach for generating basketball scoring probability maps, which, similar to Zuccolotto et al. (2021, 2023), depicts the scoring probability of a player (or a team) across the court using a visually intuitive, color-coded basketball court, offering a comprehensive insight into the shooting performance across different court areas. The novelty relies in using a geostatistical technique, the Indicator Kriging (Solow 1986), which effectively addresses spatial dependence when estimating the scoring probability conditional on the shot location. Kriging encompasses a broad category of prediction methods based on the principle of best linear unbiased estimation (Chiles and Delfiner 2012). Specifically, Indicator Kriging is suited for spatial prediction of binary response variables (made/missed shot) and estimates the probability of an event occurring at given locations based on nearby observations, possibly adjusted for covariates effect.

The paper is organised as follows. Section 2 describes the underlying rationale that motivated our proposal, in Sect. 3 we recall the formalization of the spatial statistics methods employed to construct scoring probability maps, tailored specifically to the basketball context. Section 4 illustrates the application of the Indicator Kriging to real play-by-play data related to the 2022/2023 tournament of the LBA, the Italian Basketball First League (Serie A). Data are firstly described (Sect. 4.1), then used to obtain the scoring probability maps (Sect. 4.2), and finally results are compared to those obtained by using two learning ensemble methods (Sect. 4.3). Section 5 concludes the paper by outlining future research paths on some open questions.

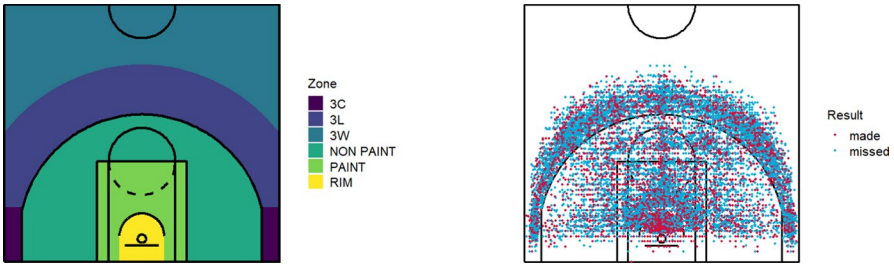

Fig. 1 Partition of the court induced by the the court lines (left) and shot chart of all the shots of the Italian Basketball First League (LBA) 2022/2023 tournament (right)

To make the methods discussed in the paper accessible to stakeholders with varying levels of statistical proficiency, we have developed a user-friendly Shiny web application. This application accompanies the manuscript and can be freely accessed.[1]

2 Motivation

Analysing the shooting spectrum of a game, which entails examining the distribution of shots across the court, holds immense significance for the definition of game strategies. Indeed, it allows the coaching staff to assess whether the observed shots distribution aligns with their strategic game plan and overall objectives. It can also be done on data of the other teams, which proves to be an important tool in pre-game scouting.

The analyses most commonly performed by practitioners for assessing shooting efficiency in high-level basketball consider the court divided into five main areas (Fig. 1, left), mainly identified by the court lines and commonly known among the basketball experts. This partition is widely adopted thanks to both its simplicity and the alignment with basketball principles. According to basketball experts, besides making sense in relation to the underlying game mechanism, this partition provides areas whose diversity of efficiency is confirmed in practice, as shown, for example, by results in Table 1, which presents some widespread shooting efficiency metrics from the LBA 2022/2023 tournament, categorized by the five main areas induced by the court lines.

Starting from this basic approach, we discuss in the following some evidences that motivate the necessity to adopt the formal analysis tool based on spatial statistics proposed in this paper.

The remarks we draw are based on some preliminary analyses we made using the data of all the shots taken during the LBA 2022/2023 tournament. The procedure used to collect and clean the data is explained in detail in Sect. 4.1.

[1] Link of Shiny web application: https://b9fagl-mirkoluigi99.shinyapps.io/census-app/.

Table 1 Table of shooting efficiency in the five conventional court zones, Italian Basketball First League (LBA) 2022/2023 tournament

	RIM	3C	3 L	PAINT	NON PAINT
FGM/FGA	5449/8028	902/2294	3554/9906	2252/5490	1639/4324
FG%	67.9%	39.3%	35.9%	41.0%	37.9%
PPS	1.36	1.18	1.08	0.82	0.76
FREQ	26.7%	7.7%	33.0%	18.3%	14.4%

FGM/FGA: field goals made / field goals attempted; FG%: shooting percentage; PPS: point-per-shot; FREQ: relative shooting frequency. The zones are ordered by PPS

We decided to exclude shots taken from more than 2.7 m beyond the 3-point line, because they are not typical of a 'standard' game play but usually attempted only in 'desperate' offensive situations. Those shots are notably fewer in number compared to all other shots (153 out of 30,197 total shots, in our dataset), so their deletion does not affect the overall results. By excluding these shots, the resulting maps will be unfilled in the corresponding zone of the court. The shot chart displayed in the right panel of Fig. 1 displays the shots of the LBA 2022/2023 tournament and clearly shows the non-homogeneous space distribution of shots, reflecting the tendency of modern basketball players to prefer shots from very close to the basket or very far away from it.

As introduced by Zuccolotto et al. (2021, 2023), a first approach to partition the basketball court according to shooting efficiency involves the use of CART (Breiman et al. 1984). Given a dataset consisting of the polar coordinates (distance and angle, with respect to the basket) and the result (made or missed) of each shot, this method produces a partition in slices maximizing the shooting percentage difference between zones. Figure 2 shows the results of this method fitted to all the shots taken in the LBA 2022/2023 tournament. Following Zuccolotto et al. (2023), we grew the CART using the pre-pruning strategy consisting in setting the minimum number of observations in any terminal leaf node equal to 10% of the total shots (about 3000). The resulting partition shows a large difference between the shots taken from very close to the basket, made with a success percentage of 67%, and the others. This result, consistent with the evidence in Table 1, confirms the validity of the coarse partition adopted by practitioners and technical experts. Nevertheless, at

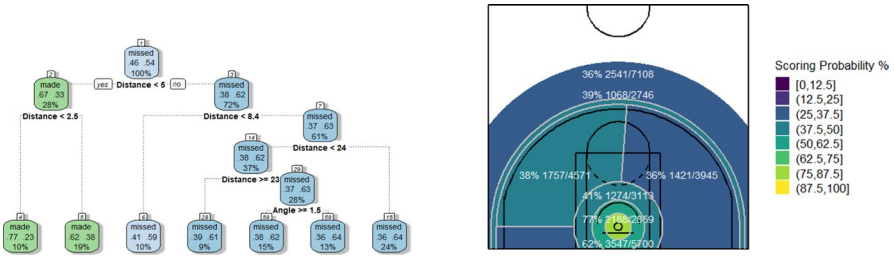

Fig. 2 CART grown on polar coordinates (left) and corresponding court partition induced by CART (right)—shots of the Italian Basketball First League (LBA) 2022/2023 tournament

the same time, the court partition suggested by the CART method also warns on the possible existence of patterns that would be very difficult to detect without a formal data analysis (*e.g*, the slight difference between shots taken from the left- and right-hand side of the basket).

Specifically, as expected, a remarkable spatial dependence among shots is evident. Thus, employing an appropriate analytical approach to capture and model this dependence could greatly enhance the resulting scoring probability maps. To this aim, we make use of geostatistics, a branch of statistics focused on the analysis of spatially dependent data. The subsequent section offers a brief overview of the general framework, followed by an in-depth discussion of the methods relevant to the application under study.

3 Methodology

Spatial statistics involves analysing data collected across various locations within a spatial domain (Chiles and Delfiner 2012; Cressie 2015; Pebesma and Bivand 2023). The goal of spatial statistics is to develop inferential methods that effectively address spatial dependence when working with georeferenced observations. Within this framework, the analysed phenomenon is typically represented using a random field:

$$\left\{Z_{\mathbf{s}}, \quad \mathbf{s} \in D\right\},$$

where $Z_{\mathbf{s}}$ is a random variable observed at location $\mathbf{s}$ within the spatial domain $D \subseteq \mathbb{R}^d$, with d generally equal to 2 or 3. Given a sample of M observed values at locations $\mathbf{s}_1, \ldots, \mathbf{s}_M$, standard geostatistical analyses typically involve identifying suitable models to characterize the spatial variability of the phenomenon and making predictions at unobserved locations within the system.

In the context of basketball analytics, the dependence between shots attempted from similar court zones and their subsequent outcomes—whether successful or not—is influenced by spatial factors such as shooting angles, proximity to the basket, and player positions. Therefore, it is sensible to leverage the geostatistics framework to construct scoring probability maps that can effectively accommodate the geometry of the basketball court. In detail, the spatial distribution of successful and unsuccessful shots delineates a set of geostatistical binary data, a type of spatial data assumed to stem from a binary random field (Heagerty and Zeger 1998; Oliveira 2000). In this context, the response variable $Z_{\mathbf{s}}$ assumes only two values for any $\mathbf{s} \in D$, conveniently coded as 0 (representing missed shots) and 1 (indicating made shots); with the primary objectives involving predicting the probability of successful shots at unavailable locations and estimating the influence of endogenous covariates on the map of estimated probabilities. To do so, we propose employing Indicator Kriging (Journel 1983; Solow 1986), a variant of universal kriging specifically designed for spatial analysis and prediction of binary data. This approach based on moment specifications has been selected for its computational simplicity, which is essential for the real-time construction of scoring probability maps in the web

application and for its proven effectiveness in various fields such as, among others, image segmentation (Wonho and Lindquist 1999) and soil quality evaluation (Smith et al. 1993). In addition, we also make use of the lorelogram tool (Heagerty and Zeger 1998; Iannarilli et al. 2019) to better describe and graphically visualize the spatial dependence structure of the missed/made shots.

In the following subsections, the considered geostatistical methods, namely Lorelogram and Indicator Kriging, are described in detail. For an up-to-date review of other statistical models for geostatistical binary data, the interested reader is referred to De Oliveira (2020); while further approaches such as Simplicial Indicator kriging and more general models for geostatistical count data not treated here can be found in Diggle et al. (1998), Tolosana-Delgado et al. (2008a, 2008b), Kazianka (2013), and references therein.

3.1 Lorelogram

The lorelogram is a statistical tool employed to assess and visualize dependency structures in binary data. Originally introduced by Heagerty and Zeger (1998) to describe dependence patterns in longitudinal categorical data, it can be used to measure spatial dependence in terms of the marginal pairwise log-odds ratio, thus providing an alternative to variograms and correlograms suited for dichotomous and/or categorical variables. More in detail, the lorelogram examines ratios of conditional odds to assess whether an event is more or less likely to occur based on the occurrence of another event nearby in space. In our context, we are interested in exploring how the odds of having successfully scored a basket at location $\mathbf{s}$ influences the odds of a shot being made at location $\mathbf{s}+h$, with h identifying a positive distance in any direction from $\mathbf{s}$:

$$\Psi\left(Z_{\mathbf{s}}, Z_{\mathbf{s}+h}\right) = \frac{\left[\frac{P\left[Z_{\mathbf{s}+h}=1|Z_{\mathbf{s}}=1\right]}{P\left[Z_{\mathbf{s}+h}=0|Z_{\mathbf{s}}=1\right]}\right]}{\left[\frac{P\left[Z_{\mathbf{s}+h}=1|Z_{\mathbf{s}}=0\right]}{P\left[Z_{\mathbf{s}+h}=0|Z_{\mathbf{s}}=0\right]}\right]} = \frac{P\left[Z_{\mathbf{s}+h}=1 \mid Z_{\mathbf{s}}=1\right]P\left[Z_{\mathbf{s}+h}=0 \mid Z_{\mathbf{s}}=0\right]}{P\left[Z_{\mathbf{s}+h}=0 \mid Z_{\mathbf{s}}=1\right]P\left[Z_{\mathbf{s}+h}=1 \mid Z_{\mathbf{s}}=0\right]}. \quad (1)$$

By direct application of the conditional probability rule, the expression in Equation (1) can be reformulated in terms of unconditional pairwise odds ratios. Under the isotropy assumption of the process (i.e., the dependence structure is homogeneous over all directions in $\mathbb{R}^2$), the quantity in Equation (1) can be directly estimated using absolute frequencies:

$$\hat{\Psi}\left(Z_{\mathbf{s}}, Z_{\mathbf{s}+h}\right) = \frac{\hat{P}\left[Z_{\mathbf{s}+h}=1, Z_{\mathbf{s}}=1\right]\hat{P}\left[Z_{\mathbf{s}+h}=0, Z_{\mathbf{s}}=0\right]}{\hat{P}\left[S_{\mathbf{s}+h}=0, Z_{\mathbf{s}}=1\right]\hat{P}\left[Z_{\mathbf{s}+h}=1, Z_{\mathbf{s}}=0\right]} = \frac{n_{11,h}n_{00,h}}{n_{01,h}n_{10,h}}, \quad (2)$$

where the term $\hat{P}\left[Z_{\mathbf{s}+h}=1, Z_{\mathbf{s}}=1\right]$ represents the estimated probability that two shots taken at a distance h both result in successful baskets. Specifically, $\hat{P}\left[Z_{\mathbf{s}+h}=1, Z_{\mathbf{s}}=1\right]$ is calculated as the ratio $n_{11,h}/n_h$, where n_h is the total number of observed shots taken at a distance h, and $n_{11,h}$ is the absolute frequency of shots at a distance h that both result in successful baskets. All other quantities in Equation (2)

are defined in a similar manner. The empirical lorelogram is finally constructed by considering the marginal pairwise log-odds ratio across different distances h:

$$\text{LOR}(h) = \log\left[\hat{\Psi}\left(Z_{\mathbf{s}}, Z_{\mathbf{s}+h}\right)\right]. \tag{3}$$

A zero pairwise log-odds ratio denotes independence between observations, while positive and negative values denote positive and negative dependency, respectively. By plotting LOR(h) as a function of h, we can visually observe how the dependency pattern changes with increasing distance between shots. The `CompRndFld R` package provides a convenient collection of functions for directly calculating and displaying such a metric (Padoan and Padoan 2015).

3.2 Indicator Kriging

Given a collection of binary observations distributed across a spatial domain $D \subseteq \mathbb{R}^2$, simple Indicator Kriging (IK) defines a statistical method for estimating the conditional probability that an unobserved location takes on one of the two values (Journel 1983; Solow 1986). In our context, the occurrences of missed or made shots on the basketball court represent the dichotomous observations, and the objective is to generate a map depicting estimated probabilities of scoring a basket. Formally, given the sample of M observed shots at locations $\mathbf{s}_1, \ldots, \mathbf{s}_M$, we wish, for a new location $\mathbf{s}_0$, to compute:

$$p = P(Z_{\mathbf{s}_0} = 1 | Z_{\mathbf{s}_1}, \ldots, Z_{\mathbf{s}_M}) = E(Z_{\mathbf{s}_0} | Z_{\mathbf{s}_1}, \ldots, Z_{\mathbf{s}_M}). \tag{4}$$

First off, assume that D can be partitioned into two subregions D_1 and D_2 such that:

$$\begin{aligned} Z_{\mathbf{s}} &= 1 \text{ if } \mathbf{s} \in D_1, \\ Z_{\mathbf{s}} &= 0 \text{ if } \mathbf{s} \in D_2. \end{aligned} \tag{5}$$

Secondly, it is required that the underlying random field enjoys the following stationarity and isotropy properties:

$$\begin{aligned} &P(\mathbf{s} \in D_1) = p_1 \text{ for all } \mathbf{s} \in D, \\ &P(\mathbf{s} \in D_i, \mathbf{s} + h \in D_j) = p_{ij}(h) \text{ for all } \mathbf{s}, \mathbf{s} + h \in D. \end{aligned} \tag{6}$$

To retrieve an estimate of the quantity p in Equation (4), Indicator Kriging considers the linear estimator

$$Z^*_{\mathbf{s}_0} = w_0 + \sum_{j=1}^{M} w_j Z_{\mathbf{s}_j}, \tag{7}$$

where the optimal set of $M + 1$ weights $w_j, j = 0, \ldots, M$, are chosen to minimize the expected squared error:

$$ESE(Z_{\mathbf{s}_0}) = E\left(\left(Z_{\mathbf{s}_0} - Z^*_{\mathbf{s}_0}\right)^2\right). \tag{8}$$

The simple Indicator Kriging estimator in Equation (7) defines the best linear unbiased estimator of p; the optimal set of weights are computed employing the spatial covariance model built on the empirical variogram. For further details the reader is referred to Solow (1986). By leveraging the Indicator Kriging approach, scoring probability maps for the LBA 2022/2023 tournament are constructed. The `gstat` `R` package (Pebesma 2004) provides routines for computing both total and local kriging. Classical kriging, employed in the former, uses all observations to compute each prediction, while the latter involves using only a neighborhood of points, resulting in a faster procedure. This approach is ideal for online web applications. Promising outcomes in terms of visual homogeneity and divergence from a uniform distribution, in accordance with Zuccolotto et al. (2023), are retrieved: results, employing total kriging, are discussed in the next section.

4 Scoring probability maps for the Italian Basketball First League (LBA) 2022/2023 tournament

4.1 Data

In recent years, basketball analytics rely more and more on play-by-play data, that is, a dataset that describes, line by line, every event in the game. The columns represent the basic information about the game (teams, day and time), the ten players on the court and various details about the occurred event. Obviously, the more detailed the dataset, the more useful information can be derived from it. In this respect, European basketball is still lagging behind the amount of information that is collected for the National Basketball Association (NBA) games, but some European organizations are helping to increase the awareness of how useful data analytics can be for the team success. In Italy, Openjobmetis Varese, provider of the data analysed in this paper, devotes a great deal of attention to analytics and commits considerable effort to data collection.

The data considered in this paper originate from the play-by-play dataset of the LBA 2022/2023 tournament,[2] webscraped from the official website. In this dataset, for each shot, information is available on its coordinates in the court, the player who attempts it, the assisting teammate (if present), its result (made or missed) and the game clock in the moment when the shot is taken. These data are freely accessible by webscraping, but unfortunately the accuracy of

[2] Teams: Banco di Sardegna Sassari, SAS; Bertram Yachts Derthona Tortona, TOR; Carpegna Prosciutto Pesaro, PES; Dolomiti Energia Trentino, TRN; EA7 Emporio Armani Milano, MIL; Germani Brescia, BRE; GeVi Napoli Basket, NAP; Givova Scafati, SCA; Happy Casa Brindisi, BRI; NutriBullet Treviso Basket, TVB; Openjobmetis Varese, VAR; Pallacanestro Trieste, TRI; Tezenis Verona, VER; Umana Reyer Venezia, VEN; UNAHOTELS Reggio Emilia, REM; Virtus Segafredo Bologna, BOL.

the downloaded data resulted compromised, especially with respect to shot coordinates.

To overcome this issue, the technical staff of Openjobmetis Varese created a dashboard that exploits video analysis to enable the manual adjustment of each shot's location, while allowing the recording of supplementary information. This correction and integration procedure led Openjobmetis Varese to have access to extremely accurate data for assessing the teams' shooting performance.

In detail, for each shot the following operations have been manually carried out: (*a*) shot location correction, (*b*) assist location correction, (*c*) 24-shotclock correction: seconds remaining at the time of the shot, (*d*) addition of two dichotomous variables indicating whether or not the shot was contested by the defense and assisted by a teammate (also for missed shots). We will make use of these accurately corrected and integrated data to construct scoring probability maps. Results are reported in the next subsection.

4.2 Scoring probability maps

Scoring probability maps are obtained for all LBA teams by means of Indicator Kriging. The substantial number of shots taken by each team, averaging around 1900, enables indeed a comprehensive distribution of shots across the court, as shown in the left panel of Fig. 3 for the team Banco di Sardegna Sassari. This enhances the reliability of the shooting probabilities predicted by the model.

The initial phase of geostatistical analysis involves an exploratory examination of the covariance structure of the process within the spatial domain. As mentioned before, different options are availaible for this purpose. In detail, since we are dealing with a binary measurement variable, we may decide to resort to the lorelogram or to more common procedures of variogram estimation adapted to the case of a binary variable. One common approach for variogram estimation involves computing the squared difference of binary measurements at pairs of points separated by a distance h:

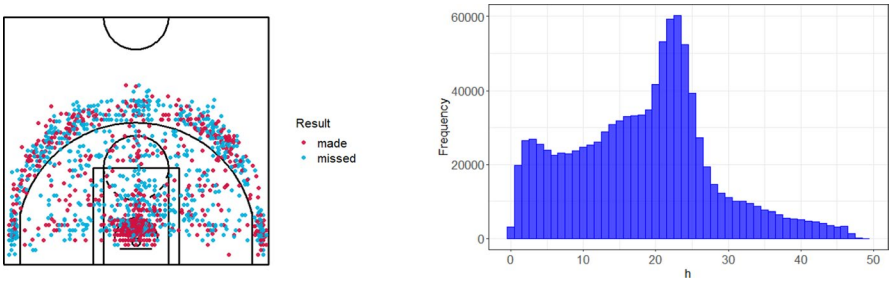

Fig. 3 Shot Chart of made and missed shots (left) and histogram of distances between shots in the court (right)—Banco di Sardegna Sassari, season 2022/2023

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} [Z_{\mathbf{s}_i} - Z_{\mathbf{s}_j}]^2, \tag{9}$$

where $N(h) = (i,j) : ||\mathbf{s}_i - \mathbf{s}_j|| = h$ and $|N(h)|$ is its cardinality. In real applications, the estimate of the variogram is typically obtained through the binned semivariogram, denoted as

$$\hat{\gamma}(\mathbf{h}) = (\hat{\gamma}(h_1), \ldots, \hat{\gamma}(h_K))^T, \tag{10}$$

which represents a discretized version of (9), computed as an average within K classes of distances of estimator (9).

Careful consideration was given to align the estimate with our specific problem by inspecting, for each team, the histogram of the distance between shots.

For example, the right panel of Fig. 3 shows the histogram of distances between all the pairs of shots taken by Banco di Sardegna Sassari during the 2022/2023 season; it reveals a pattern that is roughly alike for each team. Observing the histogram allows us to tune some relevant parameters in the lorelogram or variogram estimation, namely the *cutoff* (the spatial separation distance up to which point pairs are included in semivariance estimates) and the *boundaries* (a numerical vector with distance interval upper boundaries). We set a *cutoff* equal to 26 feet, to account for the peak of distances that each team has within the range of 22 to 27. This peak is attributed to the distances between shots close to the basket and 3-point shots, which represent the two zones characterized by the highest shot density. To calculate the empirical variogram, 9 bins were designated for distances less than 2 feet. The extensive set of distances ranging from 2 to 26 was evenly divided into 120 intervals, each with a width of 0.2. Using this approach, we have estimated both the lorelogram (3) and the binned variogram (10) for each team as shown in Fig. 4 for Banco di Sardegna Sassari. Despite the fact that the lorelogram highlights the covariance structure of the process more prominently, both approaches lead to the same conclusions: there is a correlation between the outcomes of shots taken from nearby points on the court, which rapidly decreases in the first 5 feet, then remains almost stable, albeit with some fluctuations. The presence of an alternating trend

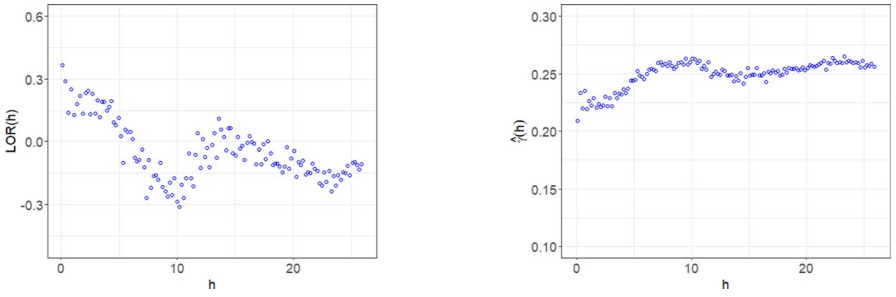

Fig. 4 Empirical lorelogram (left) and binned variogram (right)—Banco di Sardegna Sassari, season 2022/2023

is of particular interest, as it is probably due to the specific geometry of the court. For example, it makes sense to conjecture that 3-pointers taken from symmetrical regions in relation to the basket, on the right or left side of it, could share the same scoring probability. If this hypothesis were true, a correlation between the values of these two areas of the court would be observed, even though they are quite far apart. Indeed, all the graphs exhibit upward oscillations in correlation within the 10 to 20-feet range. Thanks to the substantial consistency between the evidences from the two presented approaches, in the following we will refer to the variogram for the scoring probability estimation using Indicator Kriging.

Proceeding with the analysis involves fitting a theoretical model to the empirical variogram. Before taking the next steps, it is important to point out that every team displays a finite sill, suggesting second-order stationarity in the process. This condition ensures that we can make use of the indicator kriging method. A combination of nugget and spherical theoretical models emerged as the most effective choice for fitting on the empirical variogram across all teams.

We employ Indicator Kriging for predicting scoring probabilities and generating the corresponding maps, as illustrated in Fig. 5. These maps offer a detailed court partition, revealing specific areas of high or low effectiveness. For example, in the case of Banco di Sardegna Sassari, an examination of the maps indicates a notably

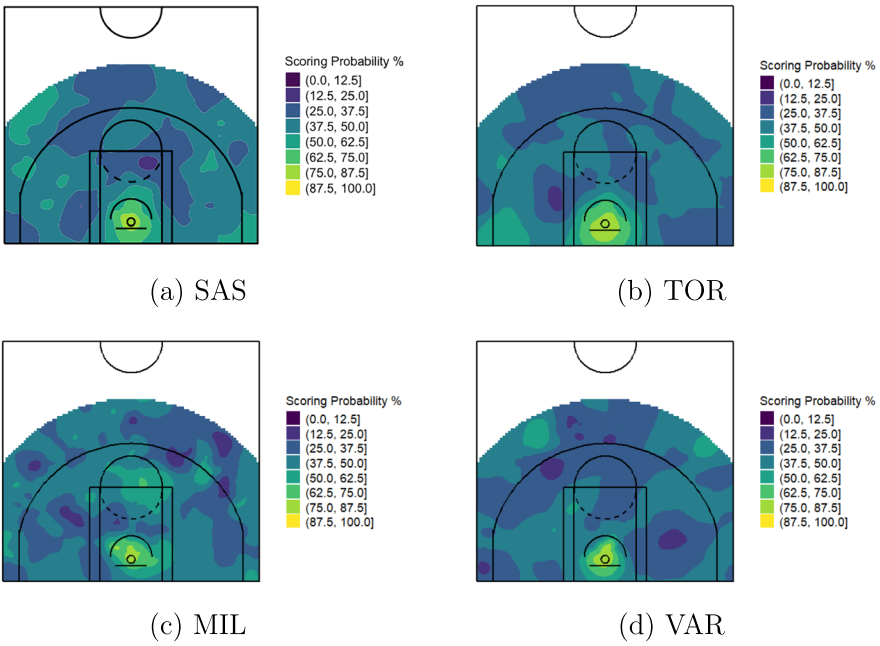

Fig. 5 Scoring probability maps produced via Indicator Kriging for the LBA teams—data are all the shots taken during the season 2022/2023

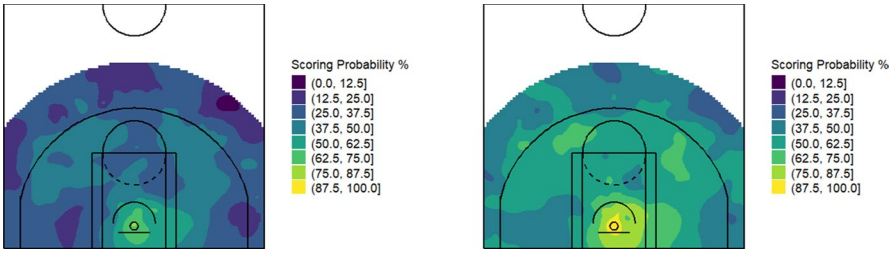

Fig. 6 Scoring probability maps produced via Indicator Kriging, for contested (left) and uncontested (right) shots—data of Virtus Segafredo Bologna, season 2022/2023

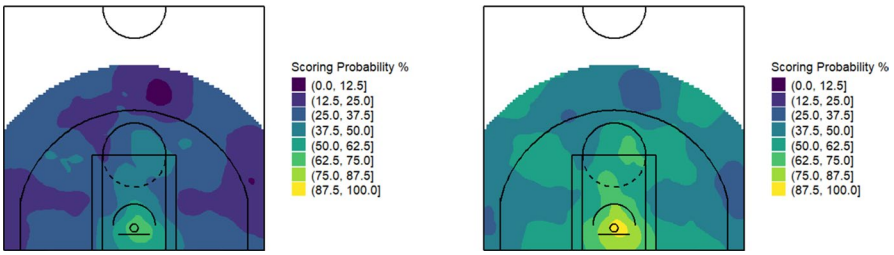

Fig. 7 Scoring probability maps produced via Indicator Kriging, for unassisted (left) and assisted (right) shots—data of Dolomiti Energia Trentino, season 2022/2023

effective shooting area in the right mid-range corner. A cross-reference with player profiles reveals that Jamal Jones[3] greatly contributes to the high effectiveness from that particular zone. An additional interesting point can be derived from the map of EA7 Emporio Armani Milano. In their playing style, there are numerous high-low post or short-roll situations, where players like Niccolò Melli[4] and Brandon Davies[5] take often spot-up shots near the free-throw line. The map, indeed, reveals a notable high-efficiency shooting area in this context. Upon closer inspection of the maps, a general trend emerges across almost all teams, indicating a slightly lower shooting efficiency from the middle 3-point area. This phenomenon could be attributed to the prevalence of isolation-pull-up threes from this zone, which tend to be less efficient compared to catch-and-shoot spot-up threes, often taken from the side. We are confident that maps like these, with their capacity to accentuate specific performance trends, offer meaningful value in assessing team shooting behavior while providing a visually engaging perspective.

Another analysis involves incorporating a categorical variable into the model, resulting in distinct maps for each category level. Figures 6 and 7 illustrate examples

[3] Jamal Jones, USA, forward, averages 11.5 points-per-game with Banco di Sardegna Sassari in the Italy first league 2022/2023.

[4] Nicolò Melli, ITA, forward/center, averages 8.8 points-per-game with EA7 Emporio Armani Milano in the Italy first league 2022/2023.

[5] Brandon Davies, USA/UGA, center, averages 13.5 points-per-game with EA7 Emporio Armani Milano in the Italy first league 2022/2023.

for two dichotomous variables that we have in our data, namely (*a*) contested/uncontested shots and (*b*) unassisted/assisted shots. The efficiency difference is substantial not only for the teams taken as an example, but also for the others. Therefore, it is worthwhile to develop game strategies that encourage players to deliver numerous assists, providing teammates with opportunities to shoot from positions without defensive coverage.

Readers are invited to explore both the basic maps and the categorical maps for all teams within the Shiny web application.

4.3 Comparison to other methods

In this section we compare the scoring probability maps obtained with Indicator Kriging to those generated by Random Forest (Breiman 2001) and AdaBoost (Schapire 2013), two prediction algorithms belonging to the class of tree-based learning ensembles. The use of algorithms based on decision trees for the definition of scoring probability maps has been proposed by Zuccolotto et al. (2023), as a possible solution to the drawbacks of the method based on CART (Zuccolotto et al. 2021) already presented in Fig. 2.

Two examples (Germani Brescia and Umana Reyer Venezia) are shown in Figs. 8 and 9, displaying sector graphs illustrating shooting percentages in pre-defined areas

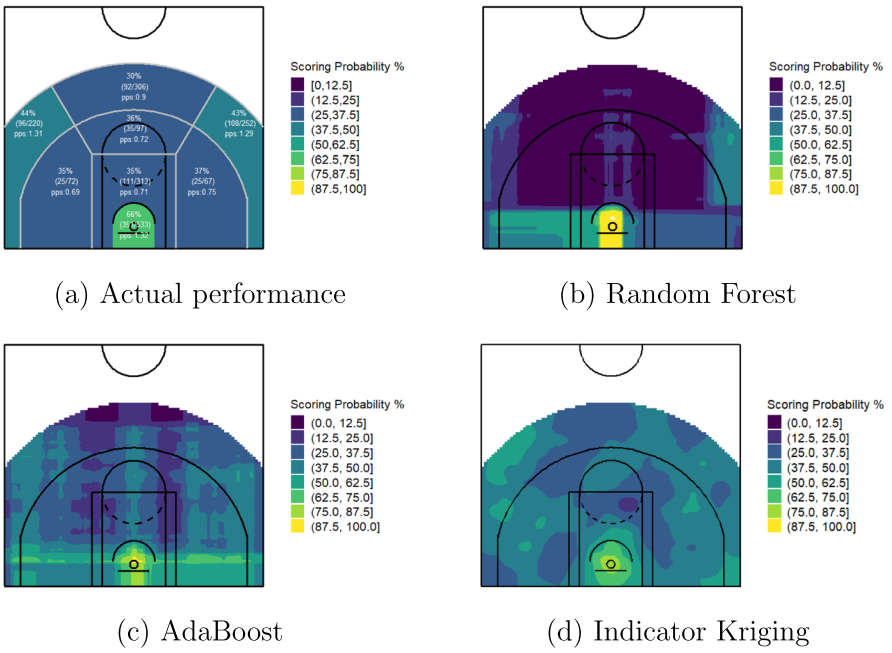

Fig. 8 Comparison between actual performance and scoring probability maps generated by Random Forest, AdaBoost and Indicator Kriging—data of Germani Brescia, season 2022/2023

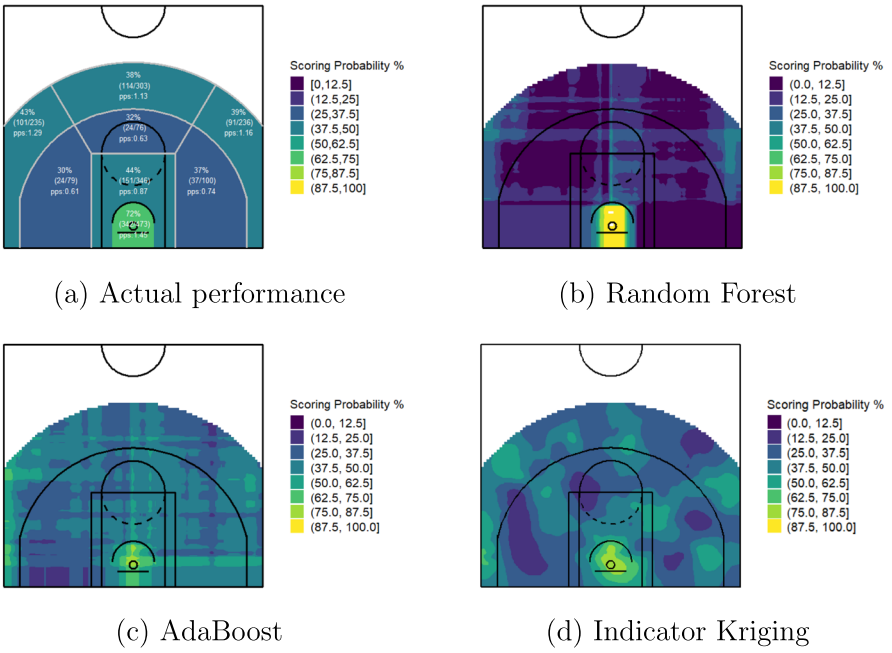

Fig. 9 Comparison between actual performance and scoring probability maps generated via Random Forest, AdaBoost and Indicator Kriging—data of Umana Reyer Venezia, season 2022/2023

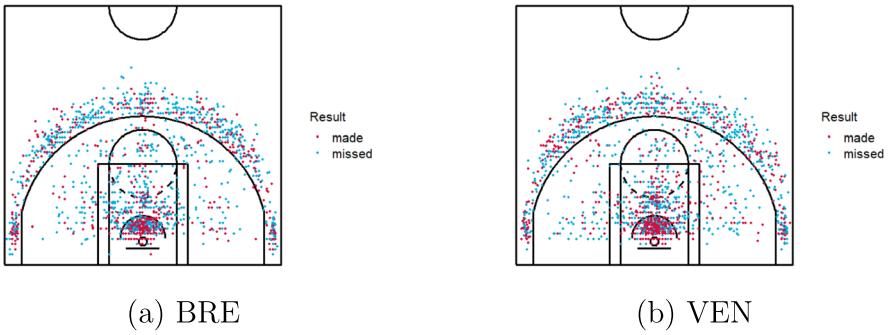

Fig. 10 Shot Chart of made and missed shots—data of Germani Brescia and Umana Reyer Venezia, season 2022/2023

of the court (panel (a)), and scoring probability maps generated using respectively Random Forest, AdaBoost, and Indicator Kriging (panels (b),(c), and (d)), all defined by using cartesian coordinates as predictors.

In both examples, Random Forest effectively identifies good and bad performance zones. However, it tends to generate scoring probability estimates that are too close to 0 or 1, diverging from the observed shooting percentages visualized in the sector graph. Adaboost, thanks to boosting, successfully addresses this issue by fitting the training

data more accurately, and offering predictions that align better with the observed shooting percentages. However, both algorithms based on decision trees encounter issues in predicting the shooting probability in subsets of the space where there are no observed data. In fact, court rectangles with few or no shots tend to be agglomerated to neighbor rectangles and be assigned the same prediction, which may even be a good choice when the empty rectangle is small, but it risks providing a misleading result when it is rather large. A clear illustration of this anomaly is the two rectangles with a high predicted shooting probability in the bottom of the court behind the basket for the case of Germani Brescia (Fig. 8, panels (b) and (c)), but similar instances can be observed also in other areas, corresponding to spots where few or no shots are taken (see the shot charts in Fig. 10).

With Indicator Kriging the problem of extrapolating predictions to zones where there are no observed data is faced by using the estimated covariance structure, which ensures a gradually decreasing effect of the evidences coming from neighboring areas. This seems to offer more credible results but, of course, in absence of shots, the goodness of the prediction cannot be checked and the best way to assign reliable prediction to areas where few or no shots are taken is still an open issue.

As the demand for accurate and actionable insights from these maps grows, it becomes essential to establish concrete criteria to evaluate the performance of different modeling techniques in capturing the complexities of shooting behavior. Our primary goal is not to create a model that performs optimally on new data or achieves high accuracy on training data. Instead, we aim to develop a procedure that produces high-quality maps accurately representing the team's shooting behavior. To assess this, we will employ the strategy introduced by Zuccolotto et al. (2023), as their methodology perfectly aligns with the objective of our analysis. In details, Zuccolotto et al. (2023) devised an index that rewards maps exhibiting two key characteristics: (1) low variance in the neighborhood of each grid point, indicating visual homogeneity, and (2) a cumulative probability distribution significantly divergent from a uniform distribution in the interval [0,1]. The justification for the latter characteristic is based on the empirical observation that, generally, a reasonable distribution of scoring probabilities should be strongly right-skewed (see Zuccolotto et al., 2023). The index, denoted as Φ, is calculated by taking the ratio of two quantities related to the aforementioned characteristics. Specifically, the formula for the index Φ reads as follows:

$$\Phi = \frac{\sigma_N}{H}. \tag{11}$$

In details, σ_N is defined as:

$$\sigma_N = \sqrt{\frac{1}{g}\sum_{i=1}^{g}\sigma_{N_i}^2}, \tag{12}$$

where 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 i-th grid point. The denominator H is defined as:

$$H = \sup_{y} |\hat{F}(y) - F_U(y)|, \tag{13}$$

where $\hat{F}(y)$ is the empirical distribution function of the estimated scoring probabilities of the map and $F_U(y)$ is the cumulative distribution function of a Uniform random variable. The numerator in (11) represents the first characteristic (a lower value results in low variance and visually homogeneous maps), while the denominator relates to the second attribute (higher values indicate a deviation from the uniform distribution). Therefore, a lower Φ index indicates a map that is preferable based on this metric. The selected method for comparing the three models involves calculating the index defined in Equation (11) across various teams. We calculated the Φ index for each competing model (i.e., Adaboost, Indicator Kriging and Random Forest) for every team. The resulting Φ values for each team are illustrated in Fig. 11. Following this rationale, Indicator Kriging stands out as the preferred model most of the time. Adaboost also performs well, achieving the lowest value of Φ in 4 out of the total 16 considered teams. On average, Indicator Kriging emerges as the top performer with a mean index value of 0.08 (SD= 0.027), followed by Adaboost with 0.11 (SD= 0.016) and Random Forest with 0.20 (SD= 0.052).

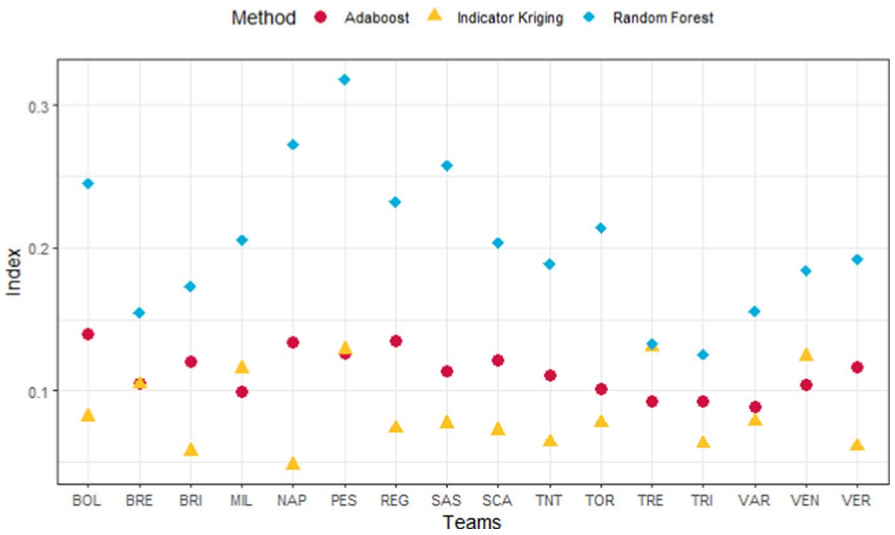

Fig. 11 Comparison between Random Forest, Adaboost and Indicator Kriging using the Φ index for each one of the 16 LBA teams

5 Concluding remarks

In this paper we have proposed the use of spatial statistics methods for binary data, with the specific aim of constructing scoring probability maps in basketball. After explaining the practical importance to draw such maps for a better knowledge of teams' shooting performance, we briefly recalled the adopted methodologies and we then showed the results achieved by applying them to the data from the 2022/2023 tournament of the Italian Basketball First League. The obtained maps clearly show some peculiarities of the analysed teams, as pointed out in the discussion. The employed play-by-play dataset, made available by Pallacanestro Varese after a complex operation of cleaning, error correction and information integration, also made possible to draw scoring probability maps for assisted/unassisted and uncontested/contested shots. This allowed to highlight further interesting evidences. Finally, a comparison to other methods, namely AdaBoost and Random Forest, enabled us to emphasize the strengths of the proposed approach based on geostatistical analysis of spatially dependent data. This comparison was carried out using a specifically designed index. All the results were presented only for some selected teams as an example, but those concerning all the teams of the championship can be obtained through the Shiny web application mentioned in the Introduction.

There are a number of open issues for future research in this context. The first idea is to extend the proposal of this paper to using polar instead of Cartesian coordinates, as done by Zuccolotto et al. (2023). In this context, the main advantage of polar coordinates is that they are consistent with the intrinsic geometry of a basketball court. From the point of view of the variogram estimation and the Indicator Kriging interpolation, the main effect of using polar coordinates is that, when turning back to Cartesian coordinates, points equidistant from a given location no longer lie on a circle, but on the perimeter of a figure obtained by a deformation of it. The consequences of using such distance measure should be carefully explored from a theoretical point of view, as it might result in a good way to address the waiving pattern exhibited by lorelograms/variograms and to better meet the isotropy requirement.

Other problems are concerned with the irregular spacing of the measurements in space. It is well known, and also evident from the shot charts, that shots tend to be taken mainly from very close to the basket or from the long distance. This makes more difficult to get robust estimates for the scoring probability of middle distance shots, as revealed also by the difficulties encountered by ensemble learning tools due to the presence of almost empty rectangles in the court. To tackle this problem, a number of additional shots taken from the middle distance could be sampled from data of other teams, with the aim of obtaining a preliminary reliable estimate that could be then refined by using data observed for the analysed team.

Lastly, interesting results could be obtained by considering the scored points as measurement variable. Such variable assumes values 0 (for all the missed shots), 2 or 3 for made shots, according to whether they are taken from inside or outside the 3-point line. This perspective would lead to extremely interesting results in terms of the shots effectiveness, but presents further challenges from a statistical point of view, as it exacerbates the issue of non-stationarity. In this case, the basketball court could be considered a complex domain, due to the presence of a well-defined boundary beyond

which the measurement variable has a completely different distribution than what is observed on the other side of the boundary. In nature, where spatial analysis is typically carried out, such complex domains may consist in regions with natural or artificial constraints, such as holes, mountain ranges or whatever barriers. As mentioned above, there are methods able to deal with non-stationary fields, for example by relying on local models describing spatial dependence within subregions of the spatial domain, where stationarity is considered a valid assumption (local stationarity). These methods could be leveraged in the basketball analytics context, treating the 3-point line as a spatial boundary.

All of the aforementioned research avenues are currently being explored and will be the focus of future work.

Acknowledgements The authors thank Pallacanestro Varese for granting the use of their manually corrected dataset.

Funding Open access funding provided by Università degli Studi di Milano within the CRUI-CARE Agreement. M.L. Carlesso, M. Manisera, P. Zuccolotto: Research Project PRIN 2022, granted by European Union - Next Generation EU, "Statistical Models and AlgoRiThms in sports (SMARTsports). Applications in professional and amateur contexts, with able-bodied and disabled athletes", project nr. 2022R74PLE, CUP: D53D23005950006. A. Cappozzo: Fund for Departments of Excellence academic funding, provided by the Ministero dell'Università e della Ricerca (MUR), established by Stability Law, namely 'Legge di Stabilità n.232/2016, 2017' - Project of the Department of Economics, Management, and Quantitative Methods, University of Milan.

References

Albert J, Glickman ME, Swartz TB, Koning RH (2017) Handbook of statistical methods and analyses in sports. CRC Press, Boca Raton

Bianchi F, Facchinetti T, Zuccolotto P (2017) Role revolution: towards a new meaning of positions in basketball. Electronic J Appl Stat Anal 10:712–734

Bornn L, Cervone D, Franks A, Miller A (2017) Studying basketball through the lens of player tracking data. In: Handbook of statistical methods and analyses in sports. Chapman and Hall/CRC, pp 245–269

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Friedman JH, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, Boca Raton

Chiles JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty, vol 713. John Wiley & Sons, New York

Cressie N (2015) Statistics for spatial data. John Wiley & Sons, New York

De Oliveira V (2020) Models for geostatistical binary data: properties and connections. Am Stat 74:72–79. https://doi.org/10.1080/00031305.2018.1444674

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-Based Geostatistics. J R Stat Soc Ser C Appl Stat 47:299–350. https://doi.org/10.1111/1467-9876.00113

Franks A, Miller A, Bornn L, Goldsberry K (2015) Characterizing the spatial structure of defensive skill in professional basketball. Ann Appl Stat 9:94–121

García J, Ibáñez SJ, De Santos RM, Leite N, Sampaio J (2013) Identifying basketball performance indicators in regular season and playoff games. J Human Kinetics 36:161–168
Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42
Heagerty PJ, Zeger SL (1998) Lorelogram: a regression approach to exploring dependence in longitudinal categorical responses. J Am Stat Assoc 93:150–162
Iannarilli F, Arnold TW, Erb J, Fieberg JR (2019) Using lorelograms to measure and model correlation in binary data: applications to ecological studies. Methods Ecol Evol 10:2153–2162. https://doi.org/10.1111/2041-210X.13308
Journel AG (1983) Nonparametric estimation of spatial distributions. J Int Assoc Math Geol 15:445–468. https://doi.org/10.1007/BF01031292
Kazianka H (2013) Approximate copula-based estimation and prediction of discrete spatial data. Stochast Environ Res Risk Assess 27:2015–2026. https://doi.org/10.1007/s00477-013-0737-7
Kubatko J, Oliver D, Pelton K, Rosenbaum DT (2007) A starting point for analyzing basketball statistics. J Quant Anal Sports 3:1–22
Lamas L, De Rose D Jr, Santana FL, Rostaiser E, Negretti L, Ugrinowitsch C (2011) Space creation dynamics in basketball offence: validation and evaluation of elite teams. Int J Perform Anal Sport 11:71–84
Lopez MJ, Matthews GJ (2015) Building an NCAA men's basketball predictive model and quantifying its success. J Quant Anal Sports 11:5–12
López Hernández F, Martinez J, Ruiz Marín M (2013) Spatial pattern analysis of shot attempts in basketball. Revista Internacional de Medicina y Ciencias de la Actividad Fisica y del Deporte, 13
Macis A, Manisera M, Zuccolotto P, Sandri M (2023) A survival analysis to discover which skills determine a higher scoring in basketball. Stat Appl Italian J Appl Stat, 35
Metulini R, Manisera M, Zuccolotto P (2018) Modelling the dynamic pattern of surface area in basketball and its effects on team performance. J Quant Anal Sports 14:117–130
Miller A, Bornn L, Adams R, Goldsberry K (2014) Factorized point process intensities: a spatial analysis of professional basketball. In: International conference on machine learning, PMLR, pp 235–243
Oliveira VD (2000) Bayesian prediction of clipped Gaussian random fields. Comput Stat Data Anal 34:299–314. https://doi.org/10.1016/S0167-9473(99)00103-6
Oliver D (2004) Basketball on paper: rules and tools for performance analysis. Potomac Books Inc, Sterling
Padoan S, Padoan MS (2015) Package 'comprandfld'
Passos P, Araújo D, Volossovitch A (2016) Performance analysis in team sports. Taylor & Francis, Milton Park
Passos P, Davids K, Araújo D, Paz N, Minguéns J, Mendes J (2011) Networks as a novel tool for studying team ball sports as complex social systems. J Sci Med Sport 14:170–176
Pebesma E, Bivand R (2023). Spatial Data Sci Appl R. https://doi.org/10.1201/9780429459016
Pebesma EJ (2004) Multivariable geostatistics in s: the gstat package. Comput Geosci 30:683–691
Ruiz FJ, Perez-Cruz F (2015) A generative model for predicting outcomes in college basketball. J Quant Anal Sports 11:39–52
Sandri M (2020) The R package `BasketballAnalyzeR`. In: Zuccolotto P, Manisera M (eds) Basketball data science—with applications in R. chapter 6. Chapman and Hall/CRC Press
Sandri M, Zuccolotto P, Manisera M (2020) Markov switching modelling of shooting performance variability and teammate interactions in basketball. J R Stat Soc Ser C Appl Stat 69:1337–1356
Santos-Fernandez E, Denti F, Mengersen K, Mira A (2022) The role of intrinsic dimension in high-resolution player tracking data-insights in basketball. Ann Appl Stat 16:326–348
Santos-Fernández E, Denti F, Mengersen K, Mira A (2022) The role of intrinsic dimension in high-resolution player tracking data. insights in basketball. Ann Appl Stat 16:326–348
Schapire RE (2013) Explaining adaboost. In: Empirical inference: festschrift in honor of Vladimir N. Vapnik. Springer, pp 37–52
Skinner B, Goldman M (2017) Optimal strategy in basketball. In: Handbook of statistical methods and analyses in sports. Chapman and Hall/CRC, pp 229–244
Smith JL, Halvorson JJ, Papendick RI (1993) Using multiple-variable indicator kriging for evaluating soil quality. Soil Sci Soc Am J 57:743–749. https://doi.org/10.2136/sssaj1993.03615995005700030020x
Solow AR (1986) Mapping by simple indicator kriging. Math Geol, 18
Tolosana-Delgado R, Pawlowsky-Glahn V, Egozcue JJ (2008) Indicator kriging without order relation violations. Math Geosci 40:327–347. https://doi.org/10.1007/s11004-008-9146-8
Tolosana-Delgado R, Pawlowsky-Glahn V, Egozcue JJ (2008) Simplicial indicator kriging. J China Univ Geosci 19:65–71. https://doi.org/10.1016/S1002-0705(08)60025-4

Wonho Oh, Lindquist B (1999) Image thresholding by indicator kriging. IEEE Trans Pattern Anal Mach Intell 21:590–602. https://doi.org/10.1109/34.777370
Wu S, Bornn L (2018) Modeling offensive player movement in professional basketball. Am Stat 72:72–79
Zuccolotto P, Manisera M (2020) Basketball data science—with applications in R. Chapman and Hall/CRC Press, Boca Raton
Zuccolotto P, Manisera M, Sandri M (2018) Big data analytics for modeling scoring probability in basketball: the effect of shooting under high-pressure conditions. Int J Sports Sci Coaching 13:569–589
Zuccolotto P, Sandri M, Manisera M (2021) Spatial performance indicators and graphs in basketball. Soc Indicators Res 156:725–738
Zuccolotto P, Sandri M, Manisera M (2023) Spatial performance analysis in basketball with cart, random forest and extremely randomized trees. Ann Oper Res 325:495–519