Computational Statistics (2025) 40:2031–2048
https://doi.org/10.1007/s00180-025-01599-1

ORIGINAL PAPER

A model-based approach to shot charts estimation in basketball

Luca Scrucca[1] · **Dimitris Karlis**[2]

Received: 28 April 2024 / Accepted: 6 January 2025 / Published online: 20 January 2025

Abstract

Shot charts in basketball analytics provide an indispensable tool for evaluating players' shooting performance by visually representing the distribution of field goal attempts across different court locations. However, conventional methods often overlook the bounded nature of the basketball court, leading to inaccurate representations, particularly along the boundaries and corners. In this paper, we propose a novel model-based approach to shot charts estimation and visualization that explicitly considers the physical boundaries of the basketball court. By employing Gaussian mixtures for bounded data, our methodology allows to obtain more accurate estimation of shot density distributions for both made and missed shots. Bayes' rule is then applied to derive estimates for the probability of successful shooting from any given locations, and to identify the regions with the highest expected scores. Additionally, calibration plots are introduced to compare the estimated scoring probabilities with the observed proportions of made shots across different offensive areas, complemented by the normalized calibration error to summarize the overall goodness-of-fit of the model-based estimates. To illustrate the efficacy of our proposal, we apply it to data from the 2022/2023 NBA regular season, showing its usefulness through detailed analyses of shot patterns and calibration performance for two prominent players.

Keywords Shot charts · Visualization of shooting patterns · Density estimation · Transformation-based Gaussian mixtures for bounded data · Probability of successful shooting · Expected points scored · Calibration plot · Normalized calibration error

✉ Luca Scrucca
luca.scrucca@unibo.it

[1] Department of Statistical Sciences, University of Bologna, Bologna, Italy

[2] Department of Statistics, Athens University of Economics, Athens, Greece

1 Introduction

Basketball is among the most popular sports game worldwide. It not only enjoys widespread popularity as a sport but has also generated substantial economic benefits through its associated industries. The National Basketball Association (NBA) is widely recognized as the world's leading league, attracting international interest with tremendous amounts spent in related marketing. In Europe, the Euroleague represents the pinnacle of professional men's club basketball competition and is regarded as the top-tier men's league on the continent. The increasing interest on basketball has led quite early to the development of advanced statistical methodologies for measuring performance (Kubatko et al. 2007), while several other proposal have been made after this. For a broad picture of academic and non-academic research on basketball analytics we recommend the book of Zuccolotto and Manisera (2020) and the broad review paper by Terner and Franks (2021).

One of the basic characteristics of basketball is that it is a fast-paced contact game in which the players are constantly moving in heated confrontations, thus leading to quick transitions from defense to offense or vice versa. In practice basketball is a game of space (see e.g. Goldsberry (2012), p.1). The teams that make better use of spatial aspects can have an advantage and hence several tactics related to better enhancement of spatio-temporal game aspects (Sandholtz et al. 2020).

Advancements in sports information systems and technology have allowed the collection of a number of detailed spatio-temporal data that capture various aspects of basketball (Papalexakis and Pelechrinis 2018; Shortridge et al. 2014). Such data can help considerably to understand the game and the effects of space on that while they also provide interesting information for all stakeholders of the game, including trainers, team managers, players, scouters of new players, spectators and journalists. Visualizations of basketball games can provide important information about the game (Perin et al. 2018). An increasing number of visualization research has been conducted that includes as visual analysis of player trajectories, visualization of field goals of a player, and visualization of basic statistics of different players in different games (Chen et al. 2016).

Shots are a key-ingredient of the sport. The final score of a team is defined by the number of successful shots and their quality, (2 or 3 points plus the 1 point for free throws). As such considerable interest has been made on understanding and predicting shot tactics and success. For example, Zuccolotto et al. (2018) utilized several techniques to model scoring probability under high-pressure conditions in basketball based play-by-play data from the Italian "Serie A2" Championship 2015/2016. Shortridge et al. (2014) discussed and proposed different measures about shot efficiency that take into account the spatial effect and they also proposed visualizations related to shot efficiency. Oughali et al. (2019) tried to predict shot success based on several machine learning approaches. Fichman and O'Brien (2019) discussed the optimal shot selection strategy for a basketball team. Jiao et al. (2021) proposed a marked spatial point process for modeling

basketball shots based on the observation that the success rate of a basketball shots may be higher at locations where a player makes more shots. Related to the spatial aspect are also the so-called corner 3's, which are those shots that while producing 3 points are taken closer to the basket, thus allowing for larger probability of success and distinguished tactic for that shots (Pelechrinis and Goldsberry 2021).

Visualizing shots can be a powerful tool for better understanding the different tactics. Quite early it has been noted that spatial visualizations like shot charts can be very valuable to reveal the tactical performance of the teams and hence be a valuable tool in the hands of trainers (Reich et al. 2006). For example, shot charts, that is, maps capturing locations of (made or missed) shots, and spatio-temporal trajectories for the players on the court can capture information about the offensive and defensive tendencies, as well as, schemes used by a team. Characterization of these processes is important for player and team comparisons, scouting, game preparation etc. Since then there has been extensive literature related to shots in basketball including effective visualizations that can produce insights. Since shots are the most important aspect as it leads to gaining points, it is quite common to produce statistics related to shot success but also to shot patterns, including spatio-temporal aspects of shots. The radical choice of most teams towards different shooting styles that include more 3-points is perhaps partially due to the improved visualizations available.

Shot charts in basketball analytics are a fundamental tool for visually examining the distribution of players' field goal attempts and their efficiency in different court locations. Typically such charts visualize the locations of all shots made, either by cutting the courts in cells (or hexagons or other areas) of the same size and representing their frequency by some color (Chu 2010). Ehrlich and Sanders (2024) proposed alternative ways to improve the information provided using some model based estimate of the shot efficiency. See also the work on Fu and Stasko (2024) about the importance of visualizing the shooting performance.

However, despite their utility, current shot charts representations face certain limitations. Predominantly, when constructed from observed data using hexagons or derived from standard density estimation procedures, they often fail to take into account the bounded nature of the basketball court. This limitation can result in misleading representations, especially at the boundaries and corners of the court. Consequently, the analysis may not fully account for the contextual constraints imposed by the court's physical boundaries, potentially skewing the assessment of shooting patterns and efficiency, particularly in areas where players are more inclined to attempt shots due to strategic advantages or positional play. These discrepancies underscore the importance of refining shot charts methodologies to accurately depict the nuanced spatial dynamics inherent in basketball shot data.

Figure 1a illustrates the approach commonly used to visualize the spatial distribution of a player's shot attempts. Typically, a two-dimensional kernel density estimate is used (Scott 2009). However, if the boundaries of a basketball court are not taken into account, some artifacts are noticeable, particularly in the corner 3-point areas, behind the backboard, and in front of the center 3-point line. In contrast, by adopting the methodology proposed in this paper we obtain a density

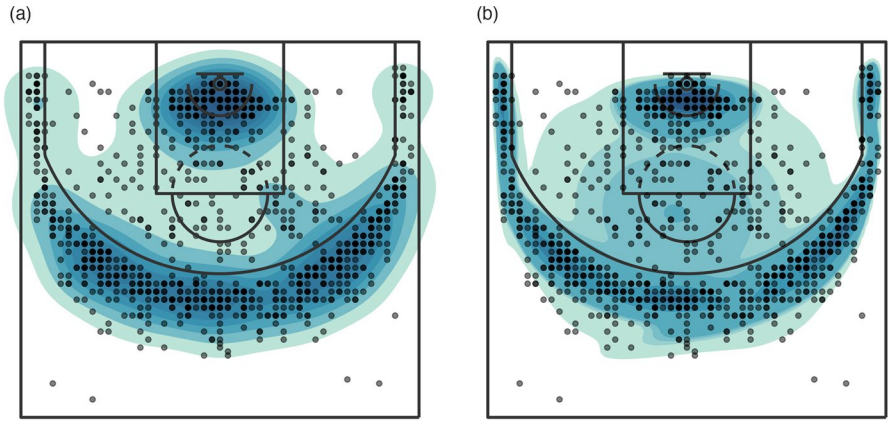


Fig. 1 Distribution of Stephen Curry's shot attempts during the 2022/2023 NBA regular season. Panel **a** shows the density estimate obtained using two-dimensional kernel density estimation, while panel **b** the estimate obtained by fitting Gaussian mixtures for bounded data, which allows the physical boundaries of the basketball court to be taken into account

estimate that remains confined within the physical boundaries of the basketball court and, by providing more accurate spatial estimates, effectively remove the above mentioned artifacts.

To summarize, in our proposal we embrace a model-based approach to shot charts estimation and visualization that: (1) employs Gaussian mixtures to estimate the density distribution of made and missed shots; (2) takes into account the physical boundaries of the basketball court; (3) applies Bayes' rule to derive estimates for the probability of successful shooting from any location; and (4) identifies regions with the highest expected scores.

The proposed approach is used to develop scoring probability maps. These maps offer a visualization of the court, displaying different levels of scoring probability for the analyzed player or team. Previous work on scoring probability maps includes Zuccolotto et al. (2021), which employs CART approaches, and Zuccolotto et al. (2023), which utilizes CART-based ensemble learning algorithms and polar coordinates. In these works, the court is split into areas determined by the algorithms rather than being predefined. Consequently, these partitions are optimal with respect to a given player (or team) shooting performance, instead of uniform grids or predefined regular slices. Additionally, Miller et al. (2014) proposed using point process models to estimate an intensity surface of NBA performance outcomes over the court.

To ensure an objective assessment of the performance of scoring probability maps, a graphical goodness-of-fit index should also be reported. For example, Zuccolotto et al. (2023) proposed an index calculated as the ratio of spatial variability in the neighborhoods of each data point to the Kolmogorov-Smirnov statistic measuring the discrepancy between the estimated probabilities and a uniform distribution. This index is useful both for selecting the tuning parameters of CART-based models and for comparing different decision-tree algorithms.

We propose a distinct evaluation strategy specifically tailored to our model-based approach. In this framework, shot probabilities estimated for a player or a team are compared to the corresponding observed proportions across a user-defined partition of the offensive basketball court. This comparison is carried out both graphically, using a calibration plot, and numerically, via a normalized calibration error index. Notably, our approach eliminates the need for hyper-parameter tuning, as model complexity is addressed through the adoption of model selection criteria such as the Bayesian Information Criterion (BIC).

The paper is organized as follows: Sect. 2 describes the model and the estimation procedure; Sect. 3 illustrates the proposed methodology using the data from the 2022/2023 NBA regular season for two players, namely Stephen Curry, perhaps the GOAT (*Greatest Of All Time*) 3-point shooter, and Joel Embiid, the MVP (*Most Valuable Player*) for that season; Section 4 discusses the evaluation of estimated shot probabilities across meaningful areas of the offensive basketball court, introducing the combined use of a calibration plot and the associated normalized calibration error. Section 5 contains some concluding remarks and potential future extensions to this paper.

2 Methods

Shot charts in basketball analytics provide a visual representation of a player or team's shooting performance by analyzing data on shots attempted from various spots on the court. However, basketball courts come in many different sizes. In the NBA, the court is 94 by 50 feet (28.7 by 15.2 m), while under the International Basketball Federation (FIBA) rules, the court is slightly smaller, measuring 28 by 15 ms (91.9 by 49.2 ft). The 3-point line is also different, being located at 23 feet 9 inches (7.24 m) from the center of the basket in the NBA (22 ft or 6.70 m at the corner), and 6.75 m (22 ft 1.75 in) for FIBA (6.60 m or 21 ft 8 in at the corner). As discussed in reference to the results shown in Fig. 1, these physical constraints on the basketball court must be given due consideration in density estimation from shots spatial information.

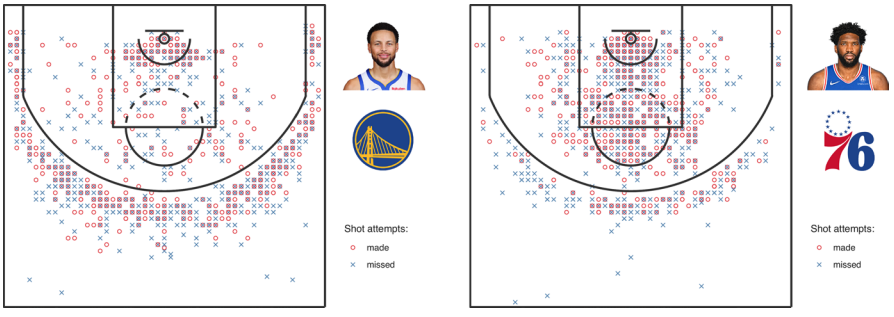


Fig. 2 Shots attempted by Stephen Curry and Joel Embiid during the 2022/2023 NBA regular season

Figure 2 shows the shots attempted by Stephen Curry (left panel) and Joel Embiid (right panel) during the 2022/2023 NBA regular season with each data point marked by shot outcome. The significant presence of shots from beyond the arc of the 3-point line is evident for Curry, while a greater number of attempts in the mid-range can be traced for Embiid. However, partly because of the presence of overlapping points, it is difficult to identify the spots from which the two players preferentially and most effectively shoot at the basket. Thereby, density estimation becomes crucial for gaining insights into shooting patterns and optimizing players performance, or to set up an efficient defense that limits shooting opportunities at preferred positions.

Gaussian mixtures (McLachlan and Peel 2000; Fraley and Raftery 2002) offer a semiparametric approach to density estimation. In this approach, the density of the data is expressed as a convex linear combination of one or more probability density functions. Gaussian mixtures are a popular choice obtained by using Gaussian densities as components of the mixture.

Gaussian Mixtures Models (GMMs) carry several advantages due to their intrinsic probabilistic generative nature. In particular, maximum likelihood estimation of parameters is available via the EM algorithm (see Sect. 2.2), with estimates that remain efficient even for multidimensional data. Moreover, GMMs require no hyperparameters tuning, with the problem of selecting the complexity of the mixture that can be recast as model selection problem (see Sect. 2.2).

Despite the fact that GMMs can approximate any continuous density with arbitrary accuracy, provided the mixture has an adequate number of components (see Ferguson (1983); Escobar and West (1995) among others), it is crucial to consider the inherent physical constraints of the basketball half-court when estimating densities in shot charts. This can be achieved by adopting the transformation-based approach to Gaussian mixture density estimation for bounded data proposed by Scrucca (2019). This approach is particularly suitable for this scenario because it explicitly considers the natural bounds of the basketball half-court. Next section briefly reviews the methodology of our proposal.

2.1 Model specification

The transformation-based approach for GMMs discussed in Scrucca (2019) extends density estimation using mixture modeling to the case of bounded variables. The basic idea is to carry out density estimation not on the original data but on appropriately transformed scale. Then, the density for the original data can be simply obtained by a change of variables.

Let (x_i, y_i) denote the coordinates of the position on the court where a player attempts a shot, for $i = 1, \ldots, n$, where n is the number of shots attempted, and $C_i = \{0, 1\}$ the corresponding binary outcome, where 1 indicates a made shot and 0 a missed shot. Consider the coordinate-wise range-logit transformation defined as

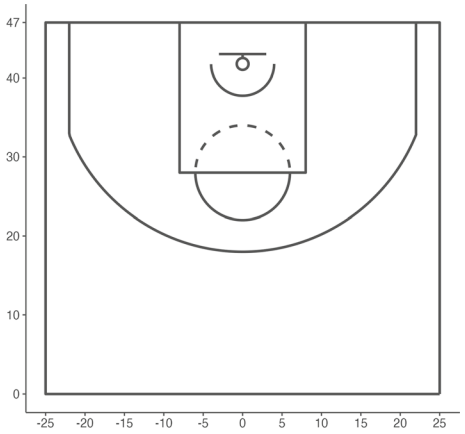


Fig. 3 NBA half-court dimensions and coordinates (in feet) used in the present paper

$$t(x,y) = \begin{bmatrix} t(x) \\ t(y) \end{bmatrix} = \begin{bmatrix} \log\left(\frac{x-\ell_x}{u_x-x}\right) \\ \log\left(\frac{y-\ell_y}{u_y-y}\right) \end{bmatrix},$$

where (ℓ_x, u_x) and (ℓ_y, u_y) are the lower and upper bounds along, respectively, the x-axis and the y-axis. Figure 3 shows the coordinates of the half-court we consider in our study for a 94 by 50 feet NBA basketball court. Thus, half-court court boundaries are set at $(\ell_x = -25, u_x = 25)$ and $(\ell_y = 0, u_y = 47)$.

In the logit-range transformed scale the density of a shot from location (x, y) can be expressed using the following Gaussian mixture

$$h(t(x,y)) = \sum_{g=1}^{G} \pi_g \, \mathcal{N}(t(x), t(y) \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{1}$$

where G is the number of mixture components, π_g the mixing probabilities (with $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$), $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, respectively, the mean vector and covariance matrix for Gaussian component g. Upon re-expressing it in the original coordinate scale, the density function can be formulated as follows:

$$f(x,y) = h(t(x,y)) \times |\boldsymbol{J}(t(x,y))|, \tag{2}$$

where $|\boldsymbol{J}(t(x,y))|$ is the Jacobian of the transformation. According to the coordinate-wise transformation approach adopted, the matrix of first derivatives is diagonal, so the Jacobian reduces to the product of first derivatives, i.e.

$$|\boldsymbol{J}(t(x,y))| = t'(x) \times t'(y) = \left(\frac{1}{x-\ell_x} + \frac{1}{u_x - x}\right) \times \left(\frac{1}{y-\ell_y} + \frac{1}{u_y - y}\right).$$

The density in the transformed coordinates from (1) can be estimated separately for made ($C = 1$) and missed shots ($C = 0$), and then back-transformed in the original scale using (2). Subsequently, the probability of scoring a basket from a specific

location can be calculated using Bayes' theorem. Specifically, the density at location (x, y) for shot outcome $C = k$, with $k = \{0, 1\}$, is given by

$$f(x, y \mid C = k) = \left(\sum_{g=1}^{G_k} \pi_{g|k} \, \mathcal{N}(t(x), t(y) \mid \boldsymbol{\mu}_{g|k}, \boldsymbol{\Sigma}_{g|k}) \right) \times |\boldsymbol{J}(t(x, y))|, \tag{3}$$

where G_k represents the number of mixture components for shot outcome $C = k$. The $\pi_{g|k}$ terms denote the mixing probabilities for outcome $C = k$ ($\pi_{g|k} > 0$ and $\sum_{g=1}^{G_k} \pi_{g|k} = 1$), and $\boldsymbol{\mu}_{g|k}$ along with $\boldsymbol{\Sigma}_{g|k}$ stand for the mean vectors and covariance matrices for component g of outcome $C = k$.

Once the density is estimated for both made shots, $f(x, y \mid C = 1)$, and missed shots, $f(x, y \mid C = 0)$, the probability of a successful shot can be obtained using Bayes' rule as:

$$\mathbb{P}(C = 1|x, y) = \frac{\tau_1 f(x, y \mid C = 1)}{\tau_0 f(x, y \mid C = 0) + \tau_1 f(x, y \mid C = 1)}, \tag{4}$$

where τ_1 and τ_0 are the outcome prior probabilities of, respectively, made and missed shots.

The estimated probabilities of making shots from various positions on the court in (4) can be multiplied by the point value of those shots (2 or 3 points) to derive the *expected points scored*:

$$\text{EPS}(x, y) = \begin{cases} 2 \times \mathbb{P}(C = 1|x, y) & \text{if } (x, y) \text{ is within the 3-point line} \\ 3 \times \mathbb{P}(C = 1|x, y) & \text{if } (x, y) \text{ is beyond the 3-point line} \end{cases}$$

This represents an important metric which provides valuable insights into offensive strategies and efficiency from different positions on the court.

2.2 Estimation and model selection

Estimation of unknown parameters, $\pi_{g|k}, \boldsymbol{\mu}_{g|k}, \boldsymbol{\Sigma}_{g|k}$, for $g = 1, \ldots, G_k$ and $k = \{0, 1\}$, in (3) can be pursued via the EM algorithm. For details see Scrucca (2019, Sect. 3.3). Moreover, outcome prior probabilities, τ_1 and τ_0, in (4) can be estimated from, respectively, the proportions of made and missed shots.

Without imposing any constraints on the covariance matrices of Gaussian components, empirical evidence suggests the inclusion of a Bayesian regularization prior to increase smoothness of the density estimate over the basketball court and avoid singularities and degeneracies in maximization of the likelihood. This can be achieved by adopting the approach of Fraley and Raftery (2007), who proposed weekly informative conjugate priors to regularize the estimation process. The EM algorithm can still be used for model fitting, but maximum likelihood estimates (MLEs) are replaced by maximum a posteriori (MAP) estimates. For details see Scrucca et al. (2023, Sect. 7.2).

A crucial aspect in mixture modeling is the choice of the number of mixture components, G_k, for each outcome. Typically, the Bayesian Information Criterion (BIC; Schwarz (1978)) is used as model selection criterion in finite mixture models. This choice is justified by Keribin (2000), who demonstrated that BIC is consistent for choosing the number of components in a mixture model, assuming a bounded likelihood (which is guaranteed by the introduction of the regularized prior mentioned earlier). However, when Bayesian regularization is introduced a slightly modified version of BIC should be used for model selection, with the maximized log-likelihood replaced by the log-likelihood evaluated at the MAP.

3 Applications

In this section we analyzed the player-by-player data of some selected NBA players for the 2022/2023 NBA regular season. The data are obtained from the `R` package `hoopR` (Gilani 2023), which provides easy access to data available on ESPN analytics at https://www.espn.com/nba/.

3.1 Stephen Curry

Figure 4 shows the estimated densities for made (a) and missed (b) shots, respectively $f(x, y|C = 1)$ and $f(x, y|C = 0)$ from (3). Regions are highlighted by highest density regions (HDRs) corresponding to specific percentages of the data. Note, however, that these cannot be directly compared, but they can be used for computing shot probabilities using (4). Required prior probabilities are estimated using proportions of made and missed shots during the regular season, giving $\hat{\tau}_1 = 0.4724$ and $\hat{\tau}_0 = 0.5276$.

Figure 5a presents the estimated shot chart highlighting regions of high and low probability for made shots by Stephen Curry. The chart reveals a remarkable consistency in Curry's shooting ability across various regions, with particularly high

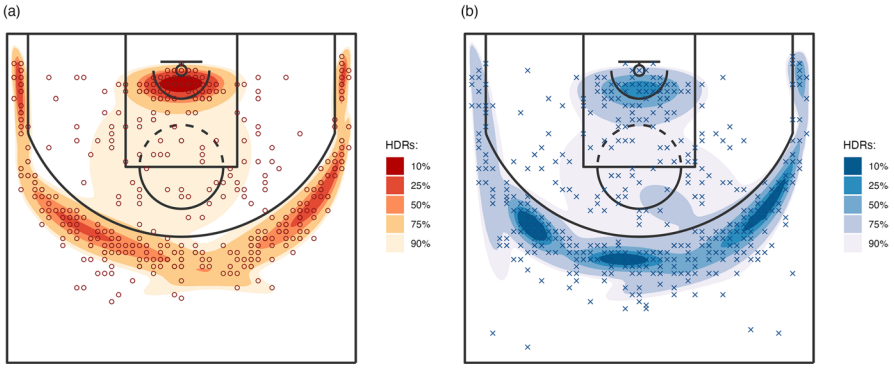
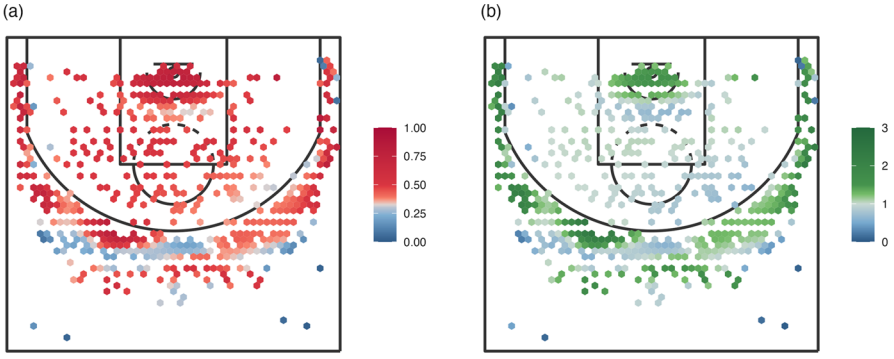


Fig. 4 Highest density regions (HDRs) from mixture-based estimated densities for made **a** and missed **b** shots for Stephen Curry during the 2022/2023 NBA regular season



ShotType	Attempts	ObsProp	AveProb	AvePts	ExpPts
2Pt	390	0.5410	0.5476	1.0923	1.0953
3Pt	626	0.4297	0.4211	1.2891	1.2632

Fig. 5 Shot charts depicting **a** estimated probabilities and **b** expected points scored per attempt for Stephen Curry during the 2022/2023 NBA regular season. The table below the charts reports a summary of empirical and estimated key statistics for both two-point and three-point shots

probabilities close to the basket and extending well beyond the three-point line. Notably, two key exceptions emerge: very far locations and positions approximately 2–3 feet from the three-point line at the top of the key. Additionally, a closer look suggests a reduced probability in the right mid-range area.

Building upon the estimated shot chart discussed above, Fig. 5b presents the corresponding graph of expected points scored. This visualization highlights regions of high scoring efficiency, primarily concentrated around close-range shots and extending to all areas beyond the three-point arc, with a notable preference for the left side. Interestingly, these high-efficiency regions align with areas of higher shot probability observed in Fig. 5a, while regions with lower expected points coincide with areas of lower shot probability.

Lastly, the table below Fig. 5 summarizes key statistics for both two-point and three-point attempts: number of attempts, observed made shot proportions, estimated average probabilities, observed average points per attempt, and estimated expected score. Notably, the empirical and estimated values exhibit close agreement, highlighting the accuracy of the model. These data showcase Stephen Curry's remarkable offensive efficiency beyond the three-point arc, reflected in an estimated expected score of approximately 1.26 points per attempt compared to 1.10 points for closer shots.

3.2 Joel Embiid

As a second player we analyze Joel Embiid of the Philadelphia 76ers. Compared to Stephen Curry's role as shooting guard, Embiid plays as a center, is much taller and stronger physically, but at the same time has an excellent aptitude for shooting from mid-range and beyond the arc. During the 2022/2023 regular season Embiid had the highest average points per game (30.6) and won the MVP award.

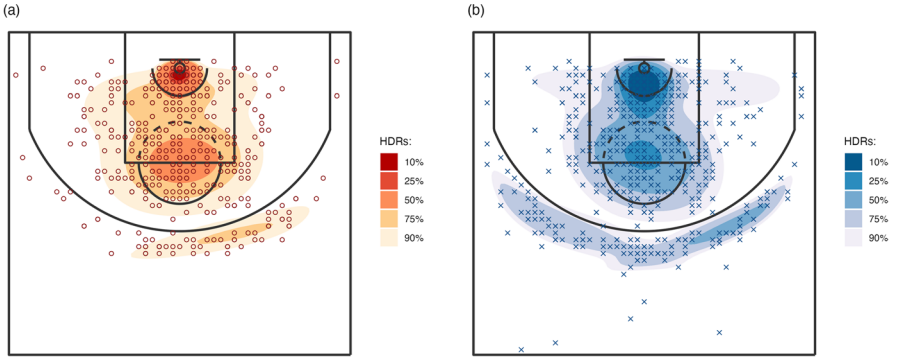
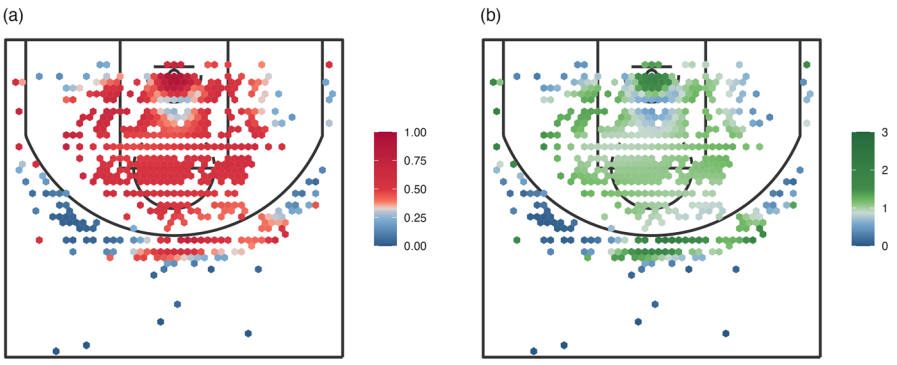


Fig. 6 Highest density regions (HDRs) from mixture-based estimated densities for made **a** and missed **b** shots for Joel Embiid during the 2022/2023 NBA regular season



ShotType	Attempts	ObsProp	AveProb	AvePts	ExpPts
2Pt	1083	0.5817	0.5840	1.1634	1.1680
3Pt	208	0.3365	0.3003	1.0096	0.9009

Fig. 7 Shot charts depicting **a** estimated probabilities and **b** expected points scored per attempt for Joel Embiid during the 2022/2023 NBA regular season. The table below the charts reports a summary of empirical and estimated key statistics for both two-point and three-point shots

Charts in Fig. 6 show the highest density regions (HDRs) obtained from mixture-based estimated densities for made (a) and missed (b) shots. The majority of shots are concentrated in the paint and near the free-throw line, while beyond the three-point arc Embiid's favorite position appears to be the central one.

Embiid's shooting efficiency is very high, as can be seen from the chart in Fig. 7a, with estimated success probabilities well above 50% in almost all mid-range and close-to-basket positions. For three-point shots, two preferred positions with very high success rates emerge: in front of the basket and slightly to the right. In other positions beyond the arc, the estimated probabilities appear significantly lower.

In terms of expected points scored from different positions, the most profitable ones are near the basket, thanks to the high shooting percentage, and those with the

highest efficiency beyond the arc, due to the fact that more points are obtained for each basket made.

Finally, it is interesting to compare the different shooting choices of Stephen Curry and Joel Embiid, and their relative effectiveness and efficiency (see tables at the bottom of Figs. 7 and 5). Curry favors long-distance shots, attempting approximately 60% more three-point shots (626 attempts compared to 390), while Embiid notably focuses on within the arc shots (1083 attempts compared to 208). Curry exhibits high estimated probabilities of scoring for both 2-point (54%) and, especially, 3-point (42%) shots, whereas Embiid demonstrates a higher percentage in the mid- and close-range shots (58%), but only a moderate 3-point percentage (30%), which is nonetheless excellent for his role. These translate into excellent expected points for both 2-point and 3-point attempts, with Curry astonishingly averaging about 1.26 points per 3-point attempt.

4 Evaluating estimated shot probabilities

The analysis of shooting performance in basketball often relies on partitioning the court into meaningful spatial regions. Such segmentation enables more detailed evaluation of player tendencies, shooting efficiency, and defensive strategies. Although several partitions could in principle be conceived, in our investigation we consider the areas shown in Fig. 8. Each area corresponds to distinct strategic zones in gameplay. For example, the restricted area and paint are high-percentage scoring areas, crucial for evaluating close-range efficiency, while three-point areas measure long-range shooting capabilities. Modern basketball has seen a rise in three-point attempts, with particular emphasis on the corners due to their shorter distance compared to other three-point shots. Moreover, shooting performance varies significantly by areas. For instance, shots in the restricted area tend to have the highest success rates, while midrange shots are less efficient compared to three-pointers due to reduced scoring value (Partnow 2021).

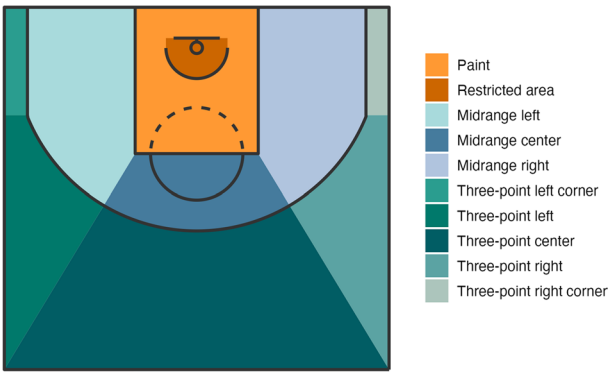


Fig. 8 Basketball court partition showing offensive shooting areas used in our analysis

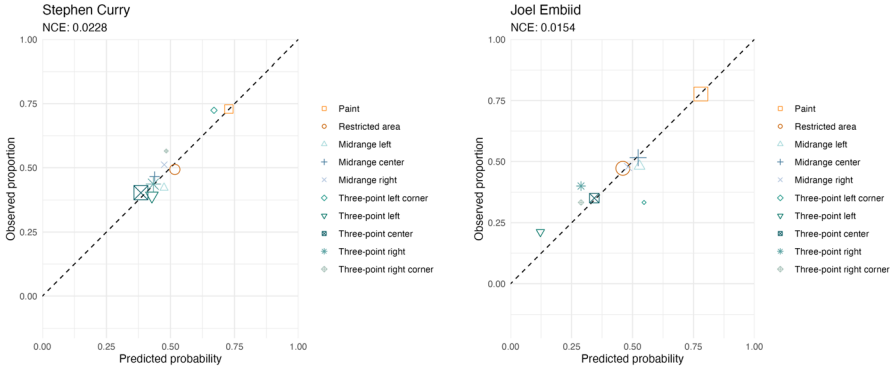


Fig. 9 Calibration plots showing observed proportions versus estimated probabilities for each player across the selected shooting areas, with point sizes proportional to the number of attempts from each area. The Normalized Calibration Error (NCE) is also reported for each player

Table 1 Summary of observed shooting statistics and model-based estimates for Stephen Curry

Area shooting	Number of attempts	Observed proportion	Average	Average points	Expected points
Restricted	96	0.7292	0.7277	1.4583	1.4553
Paint	148	0.4932	0.5169	0.9865	1.0339
Midrange left	45	0.4222	0.4747	0.8667	0.9495
Midrange center	58	0.4655	0.4381	0.9310	0.8762
Midrange right	43	0.5116	0.4755	1.0930	0.9511
Three point left corner	29	0.7241	0.6697	2.1724	2.0092
Three point left	120	0.3917	0.4273	1.1750	1.2820
Three point center	305	0.4033	0.3844	1.2098	1.1533
Three point right	149	0.4362	0.4330	1.3087	1.2989
Three point right corner	23	0.5652	0.4834	1.6957	1.4501

Based on the shooting areas reported in Fig. 8, we aim at evaluating the goodness-of-fit of model's estimated probabilities against observed proportions. A useful graphical check is provided by the *calibration plot* obtained by plotting the observed proportions of made shots against the average of estimated probabilities of making a shot in each area $\mathcal{A}_i$ of the basketball court ($i = 1, \ldots, A$, where $A = 10$ in our case). A well-calibrated model will have points closely clustered around the 45-degree line, indicating accurate predictions. Examples of calibration plots for the two players analyzed are shown in Fig. 9 presents examples of calibration plots for the two analyzed players, with point sizes proportional to the number of attempts from each area. These plots are based on the statistics reported in Tables 1 and 2, which suggest a good agreement between the model's estimates and the observed data.

Table 2 Summary of observed shooting statistics and model-based estimates for Joel Embiid

Shooting area	Number of attempts	Observed proportion	Average probability	Average points	Expected points
Restricted	354	0.7768	0.7802	1.5537	1.5604
Paint	360	0.4722	0.4602	0.9444	0.9205
Midrange left	79	0.4810	0.5287	0.9620	1.0573
Midrange center	223	0.5157	0.5228	1.0314	1.0456
Midrange right	67	0.4776	0.4814	0.9552	0.9628
Three point left corner	3	0.3333	0.5471	1.0000	1.6414
Three point left	33	0.2121	0.1220	0.6364	0.3660
Three point center	131	0.3511	0.3432	1.0534	1.0297
Three point right	35	0.4000	0.2886	1.2000	0.8659
Three point right corner	6	0.3333	0.2887	1.0000	0.8662

Among the several metrics that can be employed to summarize the calibration of probabilistic predictions, the *Normalized Calibration Error* (NCE) is defined as:

$$\text{NCE} = \frac{\sum_{i=1}^{A} |p_i - f_i| n_i}{\sum_{i=1}^{A} n_i},$$

where $p_i = \mathbb{P}(C = 1|(x, y) \in \mathcal{A}_i)$ is the estimated probability of making a shot within the area $\mathcal{A}_i$, $f_i = I(C = 1|(x, y) \in \mathcal{A}_i)/n_i$ is the observed proportions of successful attempts within area $\mathcal{A}_i$, $n_i = |\mathcal{A}_i|$ the number of attempts within area $\mathcal{A}_i$, and A is the total number of areas employed to divide the basketball court.

The metric NCE takes values on the range [0, 1], with minimum value in case of perfect fit, attained when the model perfectly estimates the observed proportions, i.e. $p_i = f_i$ for all basketball court areas $\mathcal{A}_i$. The maximum value occurs when the predicted probabilities p_i are at the maximum possible deviation from f_i, i.e. when $p_i = 0$ and $f_i = 1$, or when $p_i = 1$ and $f_i = 0$. This represents the worst fit, for which NCE = 1. The expected value under random fit is derived in the appendix, showing that $\mathbb{E}[\text{NCE}] = 1/3$.

The NEC values for Stephen Curry and Joel Embiid are 0.0228 and 0.0154, respectively, indicating a high overall goodness-of-fit of the proposed approach.

5 Conclusions

The availability of good quality spatial data in sports has increased a lot their usage, including spatial visualizations. For example, we are all familiar with heatmaps that represent the location density of players as an attempt to describe their playing behavior but also to identify tactics. Shot charts are pivotal tools in

basketball analytics, offering valuable insights into players' shooting tendencies and efficiencies across different areas of the court. Existing shot chart representations often fall short in accurately capturing shooting spatial distribution, primarily due to their inability to account for the bounded nature of the basketball court. In the present paper we proposed a new approach that employs Gaussian mixtures to estimate the density distribution using a transformation-based approach that takes into account the physical boundaries of the court. We demonstrate the effectiveness of our methodology through case studies involving real-world data from the 2022/2023 NBA regular season.

The easiness of applying and fitting Gaussian mixtures to estimate the spatial distribution creates additional opportunities. An explicit extension of the proposed work relates to all other sports where spatial location data are used. Recall also that this may extend to other non-sport related applications where boundaries need to be taken into account. As a proposal for further investigation, we also mention the use of mixture models as the basis for *conditional heatmaps*. So far, most of the sports visualization based on tracking data is based on the position of a player in the court. Sometimes it is interesting to visualize the conditional heatmap, i.e. the position of a player conditional on the position of some other player. For example, in basketball (but also in football and other team sports) this can reveal important tactical aspects and space creation strategies for the teams, which is an important ingredient of the game. Gaussian mixtures allow easily to work on that since one can easily obtain/estimate the joint distribution of the location of two players as the joint distribution in 4 dimensions, allowing also for dependence. From the joint distribution one can estimate the conditional density in a straightforward manner and thus produce a conditional heatmap.

Appendix A: Expected NEC under random probability assignment

Assume the probabilities are assigned at random to the different areas of the basketball court. In this random fit scenario, the predictions p_i do not systematically align with f_i. Then, the expected value can be computed as

$$\mathbb{E}\,[\mathrm{NCE}] = \frac{\sum_{i=1}^{A} \mathbb{E}\,[|p_i - f_i|]n_i}{\sum_{i=1}^{A} n_i}.$$

The expected value of $|p_i - f_i|$ when p_i and f_i are independent random variables uniformly distributed on [0, 1] is equal to 1/3. This result comes from integrating the absolute difference over the unit square:

$$\mathbb{E}\,[|p_i - f_i|] = \int_0^1 \int_0^1 |p_i - f_i| \; \mathrm{d}f_i \; \mathrm{d}p_i.$$

Consider splitting the unit square into two triangles, one for the case $p_i \geq f_i$, so $|p_i - f_i| = p_i - f_i$, and one for the case $p_i < f_i$, so $|p_i - f_i| = f_i - p_i$. By computing the integral over one triangle and, by symmetry, doubling its result:

$$\begin{aligned}\mathbb{E}\left[|p_i - f_i|\right] &= 2\int_0^1 \left(\int_0^{p_i} (p_i - f_i)\,\mathrm{d}f_i\right)\mathrm{d}p_i = 2\int_0^1 \left(\left[p_i f_i - \frac{f_i^2}{2}\right]_0^{p_i}\right)\mathrm{d}p_i \\ &= 2\int_0^1 \left(p_i^2 - \frac{p_i^2}{2}\right)\mathrm{d}p_i = 2\int_0^1 \left(\frac{p_i^2}{2}\right)dp_i = \int_0^1 p_i^2\,\mathrm{d}p_i = \left[\frac{1}{3}\,p_i^3\right]_0^1 = \frac{1}{3}.\end{aligned}$$

Substituting back, we get $\mathbb{E}\left[\mathrm{NCE}\right] = \dfrac{\sum_{i=1}^{A} 1/3\; n_i}{\sum_{i=1}^{A} n_i} = \dfrac{1}{3}$.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Data availability All the analyses have been conducted in R (R Core Team 2024) using the R package `GMMBasketShotCharts` (Scrucca 2024). Source code for the package, datasets, and a script to reproduce the analyses are available in a GitHub repository at https://github.com/luca-scr/GMMBasketShotCharts.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Ethical approval The research study did not involve any human participants or animals.

References

Chen W, Lao T, Xia J, Huang X, Zhu B, Hu W, Guan H (2016) Gameflow: narrative visualization of NBA basketball games. IEEE Trans Multimed 18(11):2247–2256

Chu S (2010) Information visualization in the NBA: The shot chart. Technical report, University of California, Berkeley

Ehrlich J, Sanders S (2024) Estimating NBA team shot selection efficiency from aggregations of true, continuous shot charts: A generalized additive model approach. Available at SSRN: https://ssrn.com/abstract=4697111

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90(430):577–588

Ferguson T (1983) Bayesian density estimation by mixtures of normal distributions. In: Rizvi MH, Rustagi JS, Siegmund D (eds) Recent Advances in Statistics. Academic Press, Cambridge, pp 287–302

Fichman M, O'Brien JR (2019) Optimal shot selection strategies for the NBA. J Quant Anal Sports 15(3):203–211
Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97(458):611–631
Fraley C, Raftery AE (2007) Bayesian regularization for normal mixture estimation and model-based clustering. J Classif 24(2):155–181
Fu Y, Stasko J (2024) Hoopinsight: analyzing and comparing basketball shooting performance through visualization. IEEE Trans Visual Comput Gr 30(1):858–868
Gilani S (2023) hoopR: Access Men's Basketball Play by Play Data. R package version 2.1.0
Goldsberry K (2012) CourtVision: new visual and spatial analytics for the NBA. In: 2012 MIT Sloan Sports Analytics Conference, volume 9, pages 12–15
Jiao J, Hu G, Yan J (2021) A Bayesian marked spatial point processes model for basketball shot chart. J Quant Anal Sports 17(2):77–90
Keribin C (2000) Consistent estimation of the order of mixture models. Sankhya Ser. A 62(1):49–66
Kubatko J, Oliver D, Pelton K, Rosenbaum DT (2007) A starting point for analyzing basketball statistics. J Quant Anal Sports. 3(3)
McLachlan GJ, Peel D (2000) Finite Mixture Models. Wiley, New York
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, pages 235–243. PMLR
Oughali MS, Bahloul M, El Rahman SA (2019) Analysis of NBA players and shot prediction using random forest and XGBoost models. In: 2019 International Conference on Computer and Information Sciences (ICCIS), pages 1–5. IEEE
Papalexakis E, Pelechrinis K (2018) thoops: A multi-aspect analytical framework for spatio-temporal basketball data. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pages 2223–2232
Partnow S (2021) The Midrange Theory. Triumph Books
Pelechrinis K, Goldsberry K (2021) The anatomy of corner 3s in the NBA: What makes them efficient, how are they generated and how can defenses respond? arXiv preprint arXiv:2105.12785
Perin C, Vuillemot R, Stolper CD, Stasko JT, Wood J, Carpendale S (2018) State of the art of sports data visualization. Comput Gr Forum 37(3):663–686
R Core Team (2024) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
Reich BJ, Hodges JS, Carlin BP, Reich AM (2006) A spatial analysis of basketball shot chart data. Am Stat 60(1):3–12
Sandholtz N, Mortensen J, Bornn L (2020) Measuring spatial allocative efficiency in basketball. J Quant Anal in Sports 16(4):271–289
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Scott DW (2009) Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, 2nd edition
Scrucca L (2019) A transformation-based approach to Gaussian mixture density estimation for bounded data. Biom J 61(4):873–888
Scrucca L (2024) GMMBasketShotCharts: Gaussian Mixture Model for Model-based Shot Charts Estimation in Basketball. R package version 0.1
Scrucca L, Fraley C, Murphy TB, Raftery AE (2023) Model-Based Clustering, Classification, and Density Estimation Using mclust in R. Chapman & Hall/CRC
Shortridge A, Goldsberry K, Adams M (2014) Creating space to shoot: quantifying spatial relative field goal efficiency in basketball. J Quant Anal Sports 10(3):303–313
Terner Z, Franks A (2021) Modeling player and team performance in basketball. Annu Rev Stat Appl 8:1–23
Zuccolotto P, Manisera M (2020) Basketball data science: with applications in R. CRC Press
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 Coach 13(4):569–589
Zuccolotto P, Sandri M, Manisera M (2021) Spatial performance indicators and graphs in basketball. Soc Indic 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(1):495–519