

Two clusterings to capture basketball players' shooting tendencies using tracking data: clustering of shooting styles and the shots themselves

Kazuhiro Yamada^a and Keisuke Fujii^{a,b,c}

^aGraduate School of Informatics, Nagoya University, Nagoya, Aichi, Japan;

^bRIKEN Center for Advanced Intelligence Project, Suita, Osaka, Japan;

^cPRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama, Japan

Abstract

Studies to understand the shooting preferences of basketball players relied exclusively on data on shot location, which did not lead to concrete understandings because they contained no information on how they moved to that location. Therefore, this study tried to cluster the players' shooting tendencies using the tracking data of the players' movements during the game. To do this, we first created hand-crafted shot features that included information on the pre-shot movement. Using those features, the dissimilarity of shooting tendencies between players was computed by considering the shot set of each player as a probability distribution and calculating the Wasserstein distance between them. The clustering based on their dissimilarity resulted in more clusters than in previous studies and allowed for specific shooting styles to be defined. Clustering using Gower distance as a dissimilarity measure for shot features, including a categorical feature, extracted clusters of shots that are useful for understanding players' more detailed shooting tendencies. These results prove that it is not only the shot location but also how the player moved before the shot that is important to capture the player's shooting preferences.

KEYWORDS: BASKETBALL, CLUSTERING, GOWER DISTANCE, WASSERSTEIN DISTANCE.

Introduction

Sports analytics is gaining more and more attention every year, creating an effective method for solving issues in sports, the same is true in basketball. In basketball analysis, many studies have been conducted to accurately capture player characteristics as a means to some end or as an end in itself. Player characteristics here refer to the player's offensive and defensive abilities and playing styles, including shooting patterns and so on.

There is a great variety of methods for evaluating offensive and defensive ability. One of the most mainstream trends developed is the plus-minus-based evaluation method, which is based on points gained or lost during a player's appearance, as it allows for an all-inclusive evaluation of a player's contribution. Hvattum (2019) provides a comprehensive review of plus-minus ratings including other sports. Recently, Ishida et al. (2023) proposed a sophisticated player evaluation method with a Bayesian approach. In contrast, studies using tracking data, in which the coordinates of the players' and the ball's locations on the court during the game are continuously recorded, have conducted more detailed-level modeling and proposed new evaluation metrics. For example, Yanai et al. (2022) used the framework of reinforcement learning as a modeling tool for games in the National Basketball Association (NBA) and proposed a new impact metric for NBA players, and Barron et al. (2024) introduced density-functional fluctuation theory to enable the evaluation of NBA player positioning and gravity.

In order to correctly capture the playing styles of players, researchers have used diverse clustering methods to create new typologies. Alagappan (2012) redefined positions (playing styles) and used topological data analysis to group 13 new player types. Lutz (2012) examined the effects of individual types or combinations of types on wins and losses after grouping NBA players into types by Gaussian Mixture Model (GMM)-based clustering. Zhang et al. (2016) typify NBA guards by K-means clustering. Bianchi et al. (2017) used self-organizing maps and fuzzy clustering to type NBA players. Whitehead (2018) used hierarchical clustering to define the offensive roles of NBA players. Kalman & Bosch (2020) also used GMM-based clustering for soft cluster labeling of players and then examined what lineup compositions were effective. Muniz & Flamand (2022) proposed a method for clustering players based on six different data sets and then using weighted network clustering to comprehensively cluster the results. Wang et al. (2022) also conducted GMM-based clustering to compare their performance after typing native and foreign players in the Chinese Basketball Association (CBA). Chen et al. (2023) defined the offensive roles of CBA players through K-means clustering and analyzed the impact of each role on team performance. Hua & Su (2023) redefined the offensive positions of NBA players by K-means++ clustering and estimated the plus-minus of each position and examined its change between seasons. With regard to data, most of these studies have used statistics such as the simple results and percentages of types of play.

Contrary to this generality, studies that have attempted to grasp players' shooting styles, or shooting tendencies, have used shot chart data, in which the shot location where the player took the shot was recorded (Miller et al. 2014, Hu et al. 2020, Jiao et al. 2021, Fan et al. 2023). However, the shot chart lacked information on how the player was moving before the shot, making it impossible to distinguish between types of shots, especially at close range (e.g., a dunk from cutting and a layup from driving).

Therefore, the purpose of this study is to propose clustering methods that use tracking data to incorporate information not only on shot location but also on movement before the shot, which will be useful in understanding the shooting tendencies of players. The clustering methods proposed in this study includes two complementary approaches: one is a clustering of shooting styles, which clusters players based on their shooting tendencies, and the other is a clustering of the shots themselves, which groups the shots into types. The former clustering differs from the

clustering method using shot charts in that it does not allow for detailed locational visualization, but it does allow for the definition of typologies to understand players' shot tendencies in a way that better fits our intuition. The latter clustering is done to get a more specific understanding of shooting tendencies and facilitates comparisons between players by quantifying what types of shots each player is taking and at what rate. This quantification requires a large amount of data on each player's shooting but provides detailed information on shooting preferences that cannot be captured by clustering of shooting styles. The former clustering is particularly effective in capturing the shooting tendencies of players who have played less time because it does not require as much data on shots as the latter clustering.

Methodologically, various approaches have been considered with respect to summarizing tracking data. Previous studies on dimensionality reduction of basketball players' complex movements have applied techniques such as non-negative matrix factorization (Miller et al. 2014), topic modeling (Miller & Bornn 2017), and tensor decomposition (Papalexakis & Pelechrinis 2018). To capture more dynamic aspects of data, approaches like neural network-based image processing (Wang & Zemel 2016, Nistala 2018) and dynamic mode decomposition (Fujii et al. 2017, 2018, 2020) have been explored. However, in this study, we use interpretable hand-crafted features (e.g., McQueen et al. 2014, Hojo et al. 2018, McIntyre et al. 2016, Hojo et al. 2019, Zhang et al. 2022) to interpret the results of clustering in shooting styles. In clustering, where researchers aim to organize similar objects into groups in sports data, various similarity measures have been employed. These include dynamic time warping (Sha et al. 2016), Frechet distance (Kanda et al. 2020), Hausdorff distance (Bunker et al. 2023), Gaussian mixture model clustering (Per'se et al. 2009), and self-organizing maps (Kempe et al. 2014). However, in shooting style clustering, it is necessary to calculate the dissimilarity between distributions of the players' shots. To solve this issue, we use Wasserstein distance (Aaditya et al. 2017), which is a distance function between two distributions with fewer restrictions (for details, see Materials and Methods). We also used the Gower distance (Gower 1971), a dissimilarity measure applicable to mixed quantitative and qualitative data, to cluster shot data containing both quantitative and qualitative features.

The main contributions of this work are as follows: (i) Using the tracking data, we defined a typology of shooting styles that can be used to capture players' shooting tendencies in a way that fits our intuition, and a typology of the shots themselves to capture players' shooting tendencies quantitatively, by clustering them. (ii) We proposed a method of clustering players while considering the similarity between shots by using the Wasserstein distance as the dissimilarity in the clustering of shooting styles, and in the clustering of shots themselves, we calculated the dissimilarity while considering qualitative features by using the Gower distance, thus achieving a highly interpretable typology. (iii) We found that the extracted shooting styles are useful in comparing shot efficiency in the same style and that specific comparisons of shot tendencies among players can be made using shot typology. In summary, this study represents a new direction in the study of players' shooting preferences, which to date has been limited to the use of shot charts. In practical terms, the proposed method can help team decision-makers in scouting players and in thinking about roster and lineup organisation. It should also be emphasised that this method can be extended to other goal-oriented sports such as handball, where each player has a reasonable frequency of shots, although a little ingenuity is required.

Methods

This section is made up of three sub-sections: Dataset, Preprocessing and Clustering. Clustering has two further sub-sections: Clustering of shooting styles, Clustering. An overview of the proposed method is shown in Figure 1.

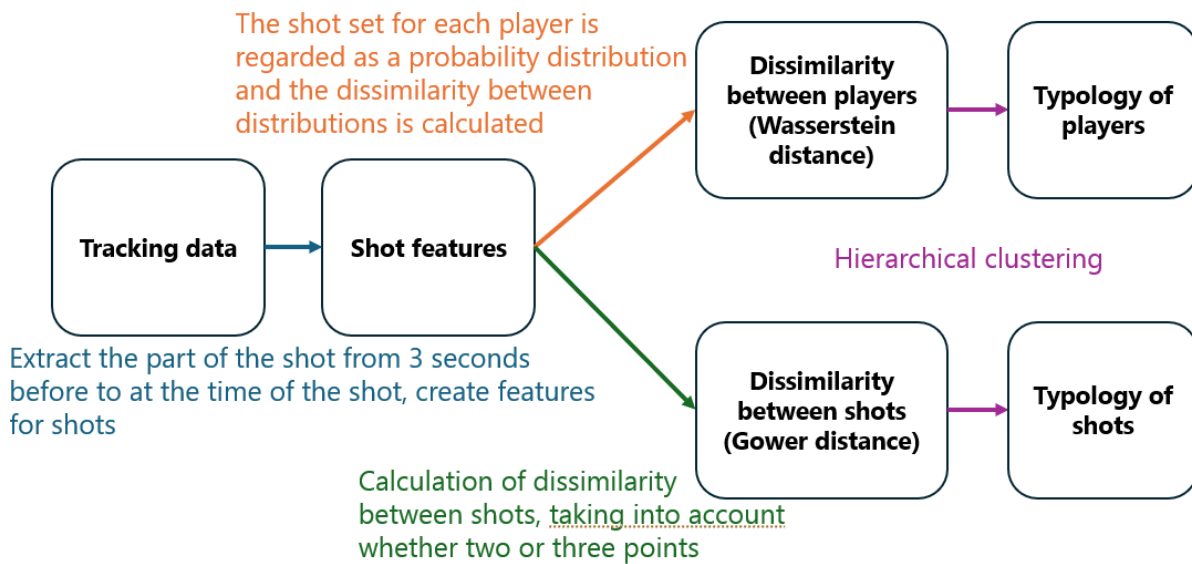


Figure 1: An overview of the proposed method

Dataset

We used tracking data from SportVU, a video-based tracking system from STATS LLC (currently Stats Perform). This dataset comprises attack segments, recorded from the moment the ball crosses the center line or is brought into the frontcourt for a throw-in until the end of the possession. The data consists of a time series of player and ball location coordinates. It is important to note that the sampling rate of the original data was 25 Hz but was downsampled to 10 Hz, and all offensive directions were unified. The dataset used in this study was offense segments from 630 games within the 2015-16 NBA season. A total of 41,160 shots (28,427 two-pointers and 12,733 three-pointers) were included in the analysis, limited to attack segments that lasted more than 3 seconds and ended with a shot. We analyzed the trajectory from three seconds before the shot to the moment of the shot.

Preprocessing

We used 17 features of the shot for the subsequent analysis: x,y coordinates of the shooter at the time of the shot and 1 second before the shot [m]; x,y coordinates at the time of receiving the ball [m]; distance to the rim at the time of the shot and 0.5 seconds, 1 second, 1.5 seconds, 2 seconds, 2.5 seconds, and 3 seconds before the shot [m]; distance to the rim when the ball was received [m]; distance traveled while holding the ball [m]; speed at the time of the shot [m/s]; and time of holding the ball [s]. The features of the position coordinates were standardized due to their larger scale compared to others, and their importance is lower than that of the distances. In addition, to reduce redundancies they were compressed to 9 dimensions using principal component analysis, which retained 99% of the cumulative contribution ratio. For visual confirmation, we further compressed them into two dimensions using UMAP (McInnes et al. 2018) and plotted them in Figure 2. UMAP is a non-linear dimensionality reduction method to visualize the structure of high-dimensional data and has the advantage of processing large amounts of data faster than t-SNE (van der Maaten & Hinton 2008), which is a well-known non-linear dimensionality reduction method. The plot revealed a bias of three-pointers towards the left, indicating their proximity in the feature space.

Table 1: Shot features and their units

Shot Feature	Unit
x, y coordinates on the court of the shooter at the time of the shot	meter
x, y coordinates on the court of the shooter 1 second before the shot	meter
x, y coordinates on the court of the shooter at the time of receiving the ball	meter
distance to the rim at the time of the shot	meter
distance to the rim 0.5 seconds, 1 second, 1.5 seconds, 2 seconds, 2.5 seconds, and 3 seconds before the shot	meter
distance to the rim when the ball was received	meter
distance traveled while holding the ball	meter
speed at the time of the shot	meter / second
time of holding the ball	second

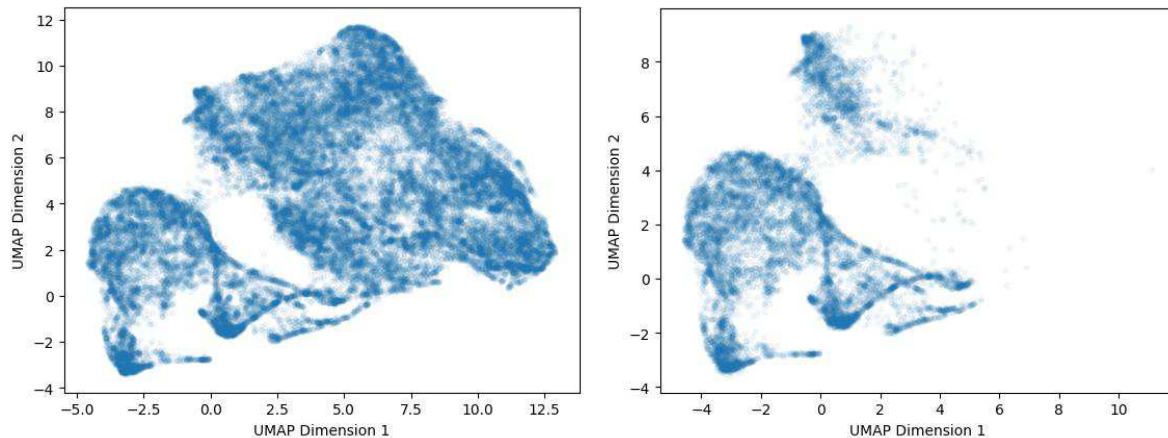


Figure 2: Scatterplots on the feature space of shots whose features are reduced to two dimensions by UMAP; left: all shots, right: 3-pointers.

Clustering

Clustering of shooting styles

First, it was necessary to calculate the dissimilarity between players for clustering. The dissimilarity between players is the dissimilarity in the propensity to shoot, i.e., the dissimilarity between sets of shots. However, inter-set similarity, such as the Jaccard coefficient, assumes that the same elements are present and could not be used in this case where no two shots are identical. Therefore, by considering the set of shots for each player as a discrete uniform distribution, we calculated the Wasserstein distance, which is the distance between distributions; the Wasserstein distance is a dissimilarity measure calculated based on the optimal transport problem, recently known as the loss function of the Wasserstein GAN (Arjovsky et al. 2017). Note that Kullback-Leibler divergence is a more well-known measure of dissimilarity of probability distributions, but it cannot be used when the supports of two discrete probability distributions do not completely overlap.

For example, for two random variables $X, Y \in \mathbb{R}^q$, given a probability mass function with m supports: $\mu(\mathbf{x})$ for $X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ and a probability mass function with n supports: $\nu(\mathbf{y})$ for

$Y = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$, and distance matrix $D \in \mathbb{R}^{m \times n}$ whose elements are the distances between $\mathbf{x}_i$ and $\mathbf{y}_j$ with parameter $p (\geq 1)$: $\|\mathbf{x}_i - \mathbf{y}_j\|_2^p$, the p-Wasserstein distance W_p between μ and ν is defined as follows:

$$W_p(\mu, \nu) := (\min \sum_{i=1}^m \sum_{j=1}^n D_{ij} P_{ij})^{\frac{1}{p}}, \quad (1)$$

where $P \in \mathbb{R}^{m \times n}$ is a matrix representing the transportation plan, and P_{ij} is the quantity (probability mass) to be transported from $\mathbf{x}_i$ to $\mathbf{y}_j$. In other words, equation (1) represents the total transportation cost in the most efficient case of carrying the probability mass of one distribution to match that of the other. In particular, when $p = 1$, it is also called Earth Mover's Distance (EMD). In this study, Earth Mover's Distance was calculated for simplicity by considering each shot set for each player as an empirical discrete uniform distribution. In other words, we calculated the dissimilarity of the distributions μ and ν of the shots of the two players when $q = 9$ (using principal components of the shot) and $p = 1$ (using distance matrix of Euclidean distance) in the above formula. Note that the data was limited to players for whom data for 30 or more shots existed, and 325 players were included in the analysis.

Next, the players were grouped using Ward's method (Murtagh & Legendre 2011), which is based on the calculated distance. Ward's method forms clusters so that the sum of squares of deviations within a cluster is minimized when merging two clusters. This method was used because it provides the most evenly distributed number of players in each cluster. Then, we calculated silhouette coefficients to determine the appropriate number of clusters. The silhouette coefficient s_i is defined for data point $\mathbf{x}_i$ using the following degree of cohesion a_i within the same cluster and the degree of deviation b_i from the nearest separate cluster:

$$a_i = \frac{1}{|C_{in}| - 1} \sum_{j \in C_{in}} d(\mathbf{x}_i, \mathbf{x}_j), \quad (2)$$

$$b_i = \frac{1}{|C_{near}|} \sum_{j \in C_{near}} d(\mathbf{x}_i, \mathbf{x}_j), \quad (3)$$

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad (4)$$

where C_{in} is the cluster to which $\mathbf{x}_i$ belongs, C_{near} is the nearest cluster from $\mathbf{x}_i$ other than C_{in} , and d is an arbitrary distance function. The silhouette coefficient takes values in $[-1, 1]$ from the definition above and is closer to 1 the more agglomerated the data points are in the same cluster and the further apart the data points are in different clusters. In other words, the closer to 1 can be interpreted as better clustering. In this method, the information on the Earth Mover's Distance between data points (i.e., players) was used to calculate the mean silhouette coefficients when the number of clusters was between 2 and 20. From the mean silhouette coefficients in Figure 3, a small number of clusters was considered relatively good. However, since a small number of clusters does not provide meaningful information, 13 was adopted as the number of clusters just before the silhouette coefficient decreased, while also taking interpretability into account.

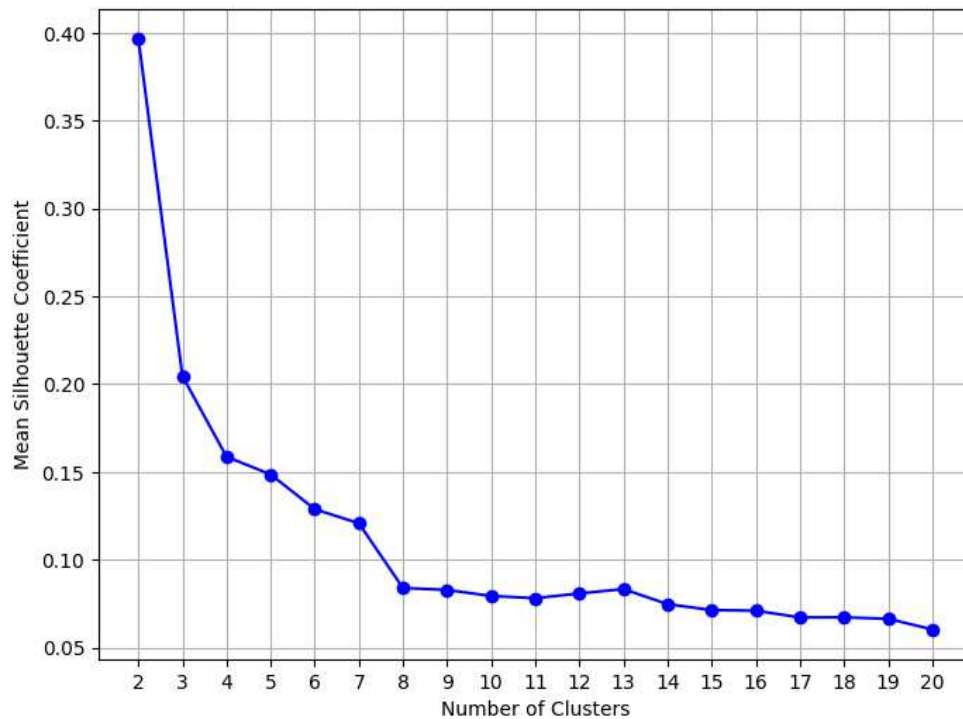


Figure 3: Mean Silhouette Coefficients in Clustering of Shooting Styles

The characteristics of each cluster were interpreted by looking at histograms of the features of the shots (before dimensionality reduction) contained in each cluster. This allowed us to confirm that the clustering results fit our intuition and to name the clusters.

As another verification, the shot efficiency of players between and within the clusters extracted was compared in True Shooting percentage (TS%) for the 2015-16 season. TS% is an advanced statistics that represents the efficiency of the shot, calculated by the following formula:

$$TS\% = \frac{PTS}{2 \times (FGA + 0.44 \times FTA)} \times 100, \quad (5)$$

where PTS is total points scored, FGA is Field Goal Attempts, and FTA is Free Throw Attempts. Note that the possible range of TS% is from 0 to 150. We examined the top five players in TS% within each cluster and see if they fit our intuition. This analysis is limited to players with 200 or more FGA during the 2015-16 season to ensure the reliability of the TS% value.

Clustering of the shots themselves

Clustering shooting styles, or typing players based on their shooting tendencies, provides an intuitive understanding of shooting tendencies. However, it is not very specific and does not quantitatively tell us what kind of shots they take more often. Therefore, by clustering the shots themselves, it is possible to calculate the percentage of how many shots a player is taking in each cluster, leading to a quantitative understanding.

We first attempted clustering using the 9 principal component dimensions of the shot features extracted in the Preprocessing Section. However, some clusters had a mixture of 2-point and 3-

point shots, which made interpretation difficult. To resolve this, categorical features with information on whether they were 2- or 3-point features were added. As a dissimilarity measure for clustering this quantity-quality mixed data, we used Gower distance, a dissimilarity measure that incorporates the Manhattan and Dice distances; Gower distance D_{gower} between data points $\mathbf{x}_i$ and $\mathbf{x}_j$ is expressed by the following equation:

$$D_{gower}(\mathbf{x}_i, \mathbf{x}_j) = 1 - \left\{ \frac{1}{K} \sum_{k=1}^K S_k(x_{ik}, x_{jk}) \right\}, \quad (6)$$

where K is the number of feature dimensions, x_{ik} and x_{jk} refer to the k -th feature dimension of the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively, and

$$S_k(x_{ik}, x_{jk}) = \begin{cases} 1 - \frac{|x_{ik} - x_{jk}|}{R_k} & \text{if } k\text{-th feature is a quantitative variable,} \\ 1 & \text{if } k\text{-th feature is a qualitative variable and } x_{ik} = x_{jk}, \\ 0 & \text{if } k\text{-th feature is a qualitative variable and } x_{ik} \neq x_{jk}, \end{cases} \quad (7)$$

where R_k is the range (difference between maximum and minimum) of the k -th feature.

Based on the calculated Gower distance, hierarchical clustering was performed using the Ward method. The Ward method was again employed here because, as before, it tended to produce clusters of uniform size and the clusters were easy to interpret. The number of clusters was determined from the change in the average silhouette coefficient as well as the clustering of shooting styles. From Figure 4, we decided to adopt the number of clusters 10 just before the mean of silhouette coefficient decreases.

Interpretation of the characteristics of each cluster was done by viewing images representation of the trajectory of the shots in that cluster.

As an example of how players' shooting tendencies can be captured, we will also present an actual case study comparing how the percentage of each shot cluster differs among players with the same shooting style.

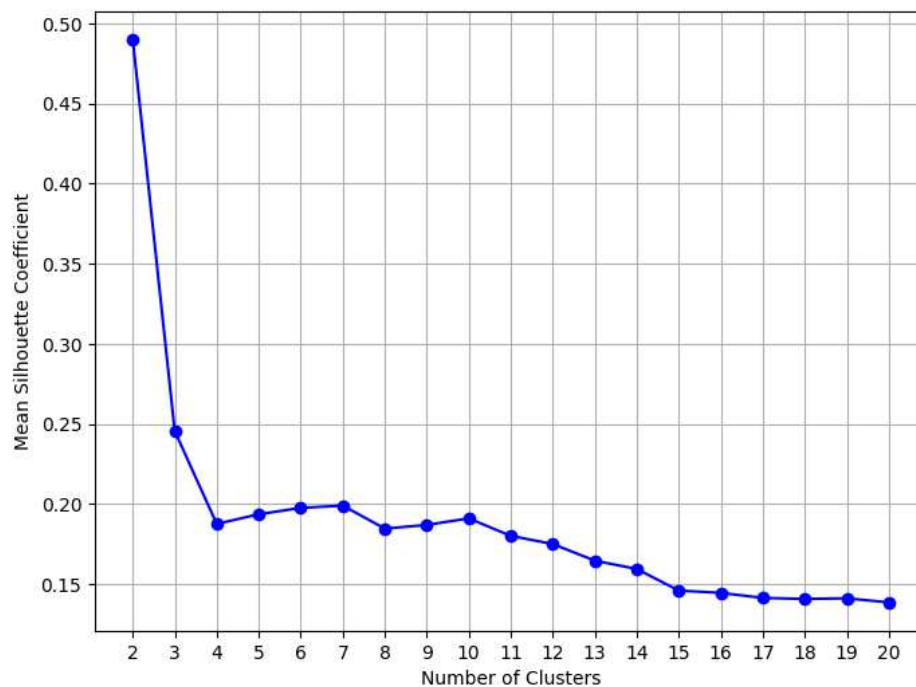


Figure 4: Mean Silhouette Coefficients in Clustering of the Shots Themselves

Results

Clustering of shooting styles

Cluster Interpretation

In the clustering of shooting styles, the individual clusters were interpreted using the features of the shots in each cluster. For example, regarding the distance between the shooter and the rim 3 seconds before the shot in cluster 1 and 2 (Figure 5), the players in these clusters were often about 4 meters from the rim in common. Compared to the other clusters, the players in these clusters are more likely to be near the rim before the shot, indicating that the players in these clusters is a Big Men (common name for large players). Regarding the location of the shot (Figure 6), both clusters prefer to shoot near the rim, but compared to cluster 1, players in cluster 2 tend to shoot mid-range shots as well. From the above, we named cluster 1 “Close-Range Big” and cluster 2 “Mid-Range Big”. For all clusters, we also confirmed that the clusters to which the players belonged were not counterintuitive.

Table 1 describes the name of each cluster, the players’ mean height, examples of players, and a brief description. The third cluster was named “Mid-Range All-Rounder” due to their tendency to attempt mid-range shots, especially near the high post. The fourth cluster was named “Mid-Range Slasher” owing to the tendency to hold the ball slightly longer than clusters 1 and 2, although it prefers shots near the rim. The fifth cluster was named “Driving All-rounder” because it tended to aim for shots from any location and was interpreted as being aggressive in driving while attempting triples. The sixth cluster was named “Pull-up Ball-Handler” due to the tendency to hold the ball longer and has more threes from outside the corners and shots just inside its arc. The seventh cluster was named “Driving Ball-Handler” because, like cluster 6, it had a longer ball-holding time and more shots were made near the rim. The eighth cluster was named the “Stretch Four” due to its higher frequency of 3-pointers from the corners and top and a slightly taller average height. The ninth cluster was named “Corner Shooter” because of its preference for shooting threes, especially from the corner. The tenth cluster preferred the 3-point, but unlike

the ninth cluster, it tended to aim from any position, hence the name “Pure Shooter”. The eleventh cluster was named “Slashing Finisher” because players tended to hold the ball longer and preferred to shoot near the rim but not so much for 3-pointers. The twelfth cluster was similar to the ninth cluster but was named “Driving Shooter” due to a slight preference for shots near the rim. The 13th cluster prefers close-range to mid-range shots, but can also attempt a reasonable number of 3-pointers, thus we named it “Stretch All-rounder”. Based on the distance matrix, the players were plotted in a 2-dimensional coordinate plane with cluster labels using t-SNE (Figure 7). In this figure, players with similar shot tendencies are placed closer together, while players with dissimilar shot tendencies are placed farther apart.

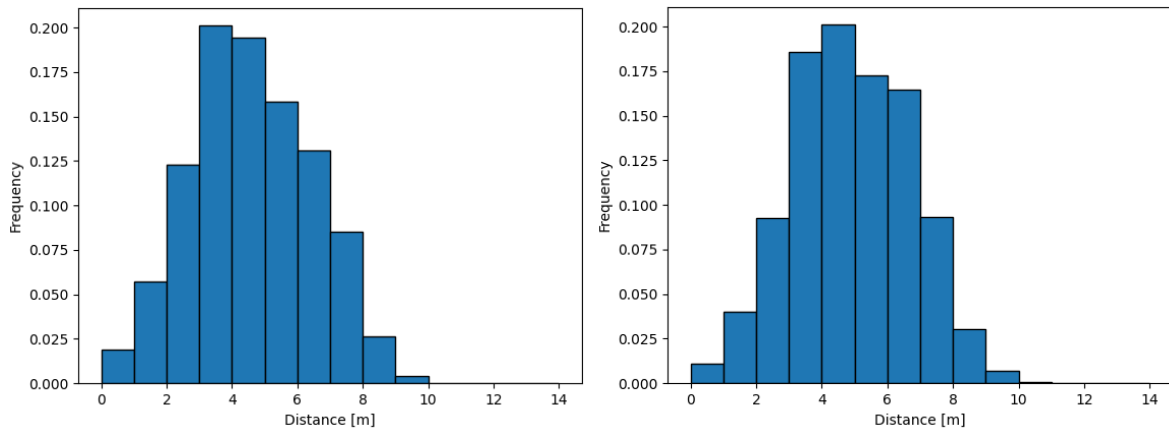


Figure 5: Distance between the shooter and the rim 3 seconds before shot in two clusters as examples. Normalized frequencies are displayed. On the left is Cluster 1, which tends to have the shortest distance from the rim 3 seconds before the shot. On the right is Cluster 2, which also tends to have a shorter distance from the rim 3 seconds before the shot.

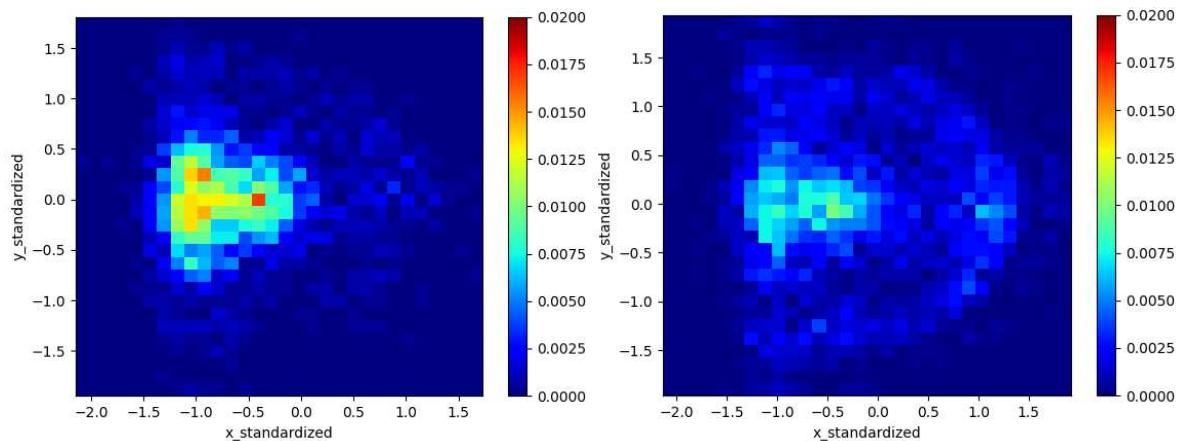


Figure 6: 2D histogram of shot location, in two clusters as examples. Normalized frequencies are displayed. On the left is Cluster 1, with shot locations concentrated near the rim. On the right is Cluster 2, which exhibits shots well distributed from the rim to mid-range.

Table 2: Shooting Style Cluster Description

Cluster Name	Description	Mean Height [cm]	Example Players
Close-Range Big	Big Men who attempt most shots from close-range.	210.6	Andre Drummond
			Dwight Howard
Mid-Range Big	Big Men who can shoot from close-range as well as mid-range.	209.9	Pau Gasol
			LaMarcus Aldridge
Mid-Range All-Rounder	Players who play offense from mid-range and often shoot from near the high post.	208.9	Kevin Garnet
			Dirk Nowitzki
Mid-Range Slasher	Players who attack the rim from mid-range through post-play or drive.	206.5	Shaun Livingston
			DeMarcus Cousins
Driving All-Rounder	Players who begin offense from beyond the arc and aim to shoot from anywhere with their versatile skills.	198.6	Stephen Curry
			Kevin Durant
Pull-up Ball-Handler	Players who often drive and aim for many pull-up jumpers.	189.9	Kyle Lowry
			Damian Lillard
Driving Ball-Handler	Players who often drive from the perimeter, but also attempt threes moderately.	190.7	Russell Westbrook
			James Harden
Stretch Four	Big shooter who often attempts corner threes or threes from the top position.	205.2	Nikola Mirotic
			Meyers Leonard
Corner Shooter	Shooter attempting mainly three-pointers from the corner.	198.6	Patrick Beverley
			Jason Terry
Pure Shooter	Shooter who attempts 3-pointers from any location.	194.4	Eric Gordon
			Kyle Korver
Slashing Finisher	Players who do not shoot many threes and prefer to drive from beyond the arc.	199.9	DeMar DeRozan
			Tony Parker
Driving Shooter	Similar to Corner Shooter, but shooter with a slight preference for drive.	201.5	Klay Thompson
			Vince Carter
Stretch All-Rounder	Players who shoot from close-range to mid-range, stretch and shoot 3-pointer as well.	205.7	Kristaps Porzingis
			Kevin Love

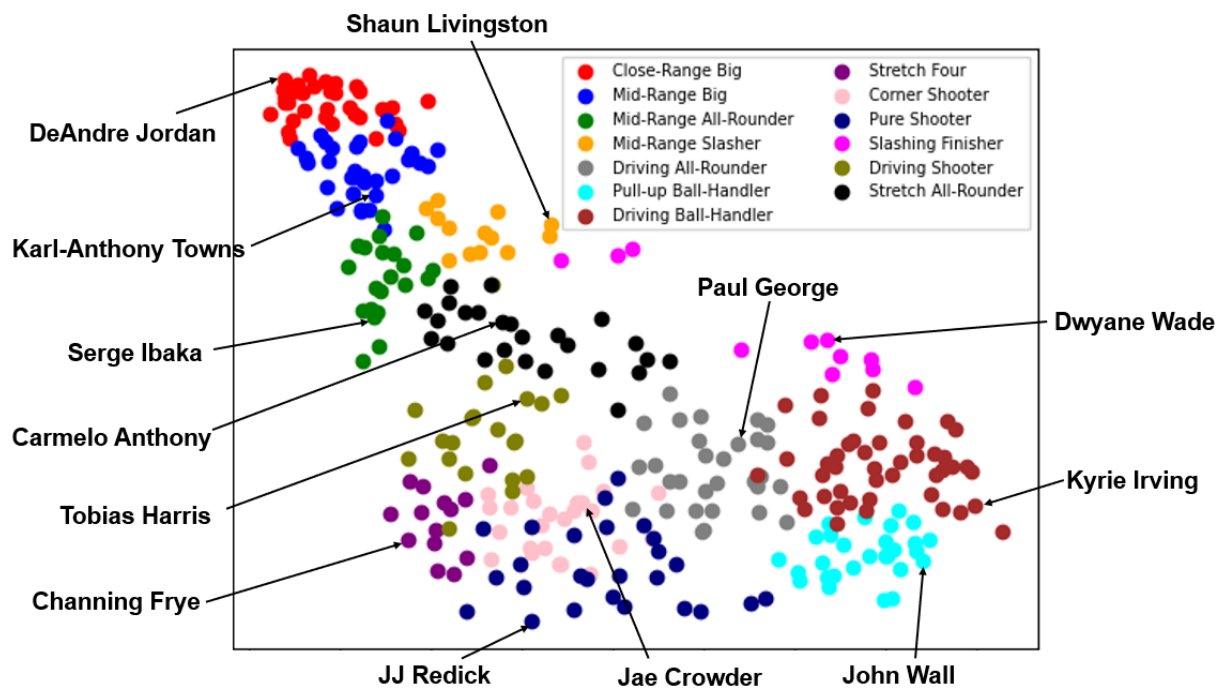


Figure 7: Visualization of shooting style clusters by t-SNE

Comparison of Shot Efficiency within Cluster

Tables 3 shows the top five TS% players and their respective field goal attempts for several clusters in the 2015-16 season (the higher the value, the more reliable the value of the TS%). The remaining clusters are listed in the Supplementary Materials. From the tables, we can see that many of the players are from teams that were strong at the time, such as Golden State Warriors (GSW), San Antonio Spurs (SAS), and Oklahoma City Thunder (OKC). This suggests that players from stronger teams tended to have better shooting efficiency (when viewed in conjunction with remaining clusters in the Supplementary Materials, many players from other strong teams of the time, the Los Angeles Clippers (LAC) and the Toronto Raptors (TOR), are also represented), which fits our intuition. The results indicate that this typology of shooting styles is highly valid, as well as its usefulness in comparing players' shooting efficiency.

Table 3: Top 5 players for Close-Range Big, Driving All-Rounder, Stretch All-Rounder, Slashing Finisher and Driving Shooter in TS%

Close-Range Big		
Player Name (Team)	FGA	TS%
Hassan Whiteside (MIA)	682	62.9
DeAndre Jordan (LAC)	508	62.8
Cole Aldrich (LAC)	225	62.6
Andrew Bogut (GSW)	279	62.3
Steven Adams (OKC)	426	62.1

Driving All-Rounder		
Player Name (Team)	FGA	TS%
Stephen Curry (GSW)	1598	66.9
Kevin Durant (OKC)	1381	63.4
Chandler Parsons (DAL)	651	58.9
Omri Casspi (SAC)	622	58.7
Evan Fournier (ORL)	929	58.7

Stretch All-Rounder		
Player Name (Team)	FGA	TS%
Kawhi Leonard (SAS)	1090	61.6
Draymond Green (GSW)	819	58.7
Mike Scott (ATL)	376	57.5
Chris Bosh (MIA)	767	57.1
Kelly Olynyk (BOS)	556	56.1

Slashing Finisher		
Player Name (Team)	FGA	TS%
Jonathon Simmons (SAS)	242	58.6
DeMar DeRozan (TOR)	1377	55
Tony Parker (SAS)	710	54.6
Donald Sloan (BKN)	350	53.6
James Johnson (TOR)	240	53.2

Driving Shooter		
Player Name (Team)	FGA	TS%
Klay Thompson (GSW)	1386	59.7
Thabo Sefolosha (ATL)	372	57.8
Alonzo Gee (NOP)	255	57.2
Andre Roberson (OKC)	777	57.1
Andre Iguodala (GSW)	609	56.4

Clustering of the shots themselves

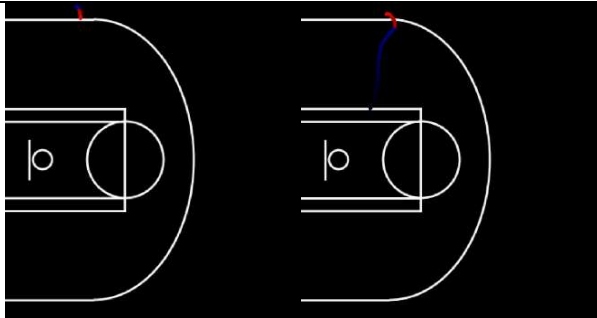
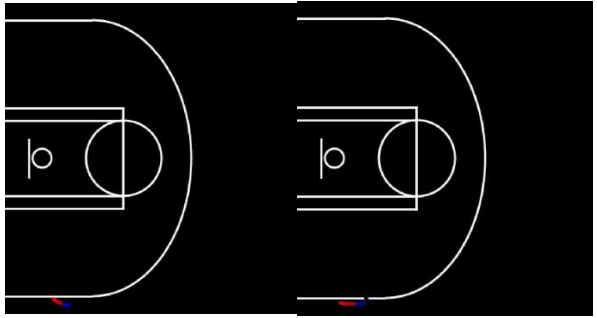
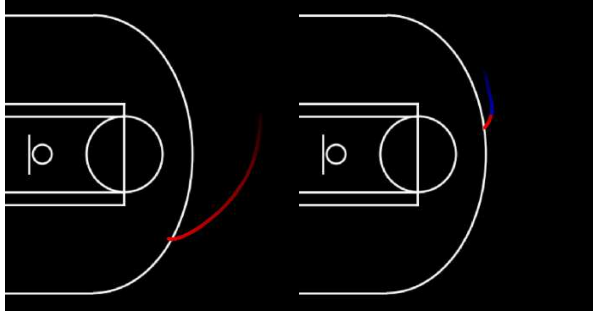
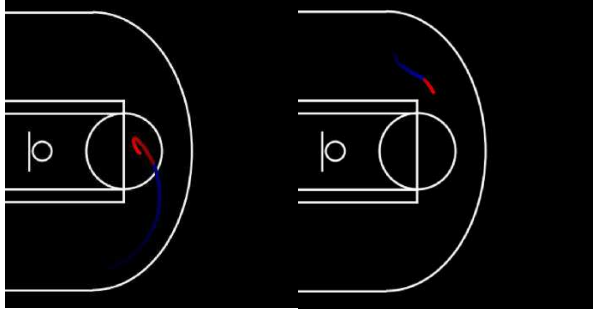
Cluster Interpretation

The results of the clustering of the shots themselves were determined primarily by looking at the visualized image of the trajectory of each cluster just prior to the shot, as described in Materials and Methods.

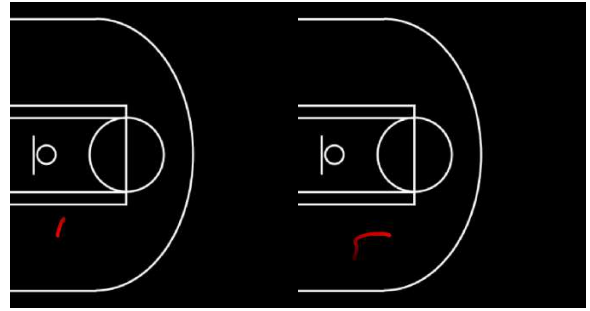
Cluster 1 was named ‘3-pointer from the right corner or right wing’ because the cluster consisted of 3-point shots from around the right corner or right wing. Cluster 2 is consisted of a three-point shot from the left corner, hence the name ‘3-pointer from the left corner’. Cluster 3 consisted of three-point shots near the top of the key or left wing, thus the name ‘3-pointer from the top of the key or left wing’. Cluster 4 consisted of shots near the high-post, hence the name ‘Mid-range shot near the high-post’. Cluster 5 was named ‘Shot from a mid-post move’ because it consisted of shots from post-play near the mid-post. Cluster 6 was named “Close-range shot from a post move” because it consisted of close-range shots from a post move. Cluster 7 was named “Cutting layup/dunk” because it consisted of shots after receiving the ball from the

cutting. Cluster 8 was named “Mid-range shot from the right side” because it consisted of mid-range shots, including pull-up jumper, on the right side. Cluster 9 was named “Mid-range shot from the left side” because it consists of mid-range shots from the left side, including pull-up jumpers. Cluster 10 was named “Driving layup/dunk” because it consisted of shots from the drive.

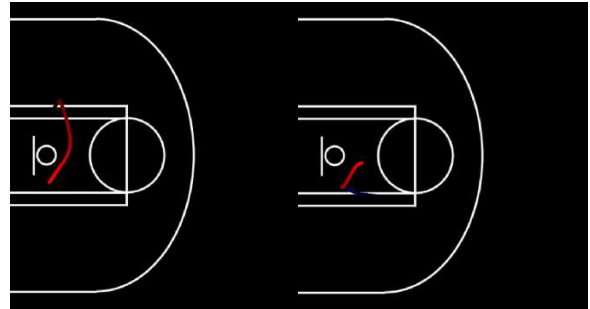
Table 4: Shot cluster names and examples of shots in each cluster. The red and blue lines in the image represent the shooter’s trajectory from 3 seconds before the shot to the time of the shot when holding the ball and when not holding the ball, respectively; the lighter the color, the later the time series.

Cluster Name	Example Images
3-pointer from the right corner or right wing	
3-pointer from the left corner	
3-pointer from the top of the key or left wing	
Mid-range shot near the high-post	

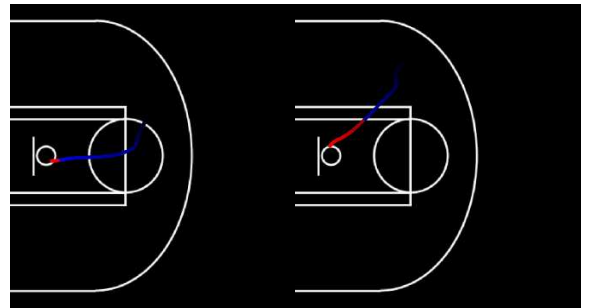
Shot from a mid-post move



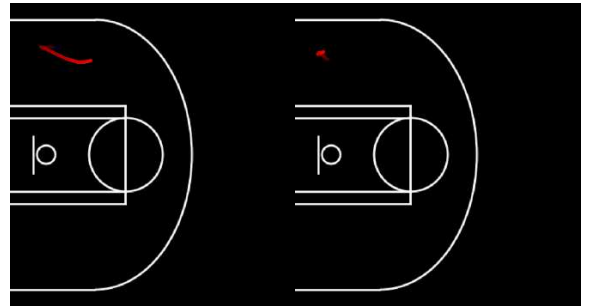
Close-range shot from a post move



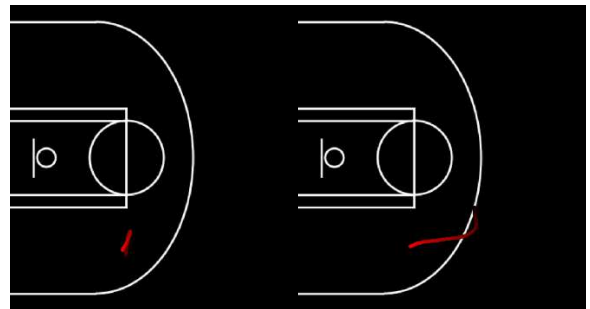
Cutting layup/dunk



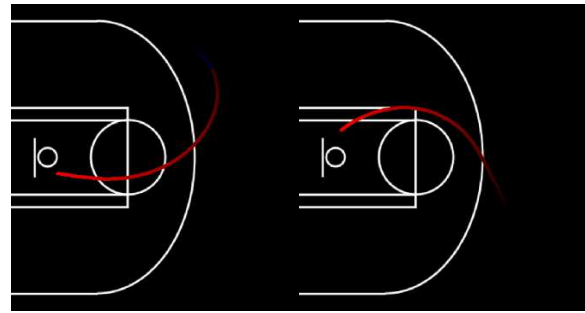
Mid-range shot from the right side



Mid-range shot from the left side



Driving layup/dunk



Detailed comparison of shot tendencies among players

Here, the detailed players' shooting tendencies are interpreted by comparing two players belonging to the same shooting style. As an example, the percentage of each shot cluster for Stephen Curry and Kevin Durant in the Driving All-Rounder and Anthony Davis and Dirk Nowitzki in the Mid-Range All-Rounder is shown in a side-by-side bar graph (Figure 8). Both Curry and Durant have driving layups/dunks that account for about 20% of their total shots, but there is a difference in their 3-point shot tendencies. Curry hits about 20% of his threes from the right corner and right wing and 10% from the left corner, while Durant hits about 10% and almost 0% of those respectively. There is also a difference in the middle shot near the high post, with Curry hitting this shot only about 5% of the time, while Durant seems to hit it nearly 20% of the time. As for the other pair, Davis and Novitsky, there are also differences in tendencies. First, Davis hits only about 5% of his 3-pointers near the top of the key or left wing, while Novitsky hits about 15%. There is also a difference in mid-range tendencies, with both players strongly favoring shots near the high post at around 25%, but Novitski's percentage is higher than Davis' for shots from the mid-post and mid-shots from the right side. Furthermore, Nowitzki's close-range shot from a post move, cutting layup/dunk, and driving layup/dunk all have smaller percentages, suggesting that Nowitzki did not take many close-range shots. In summary, these are considered to be consistent with the characteristics of each player's actual shot tendencies, and reflect each player's signature moves and preferences of shots.

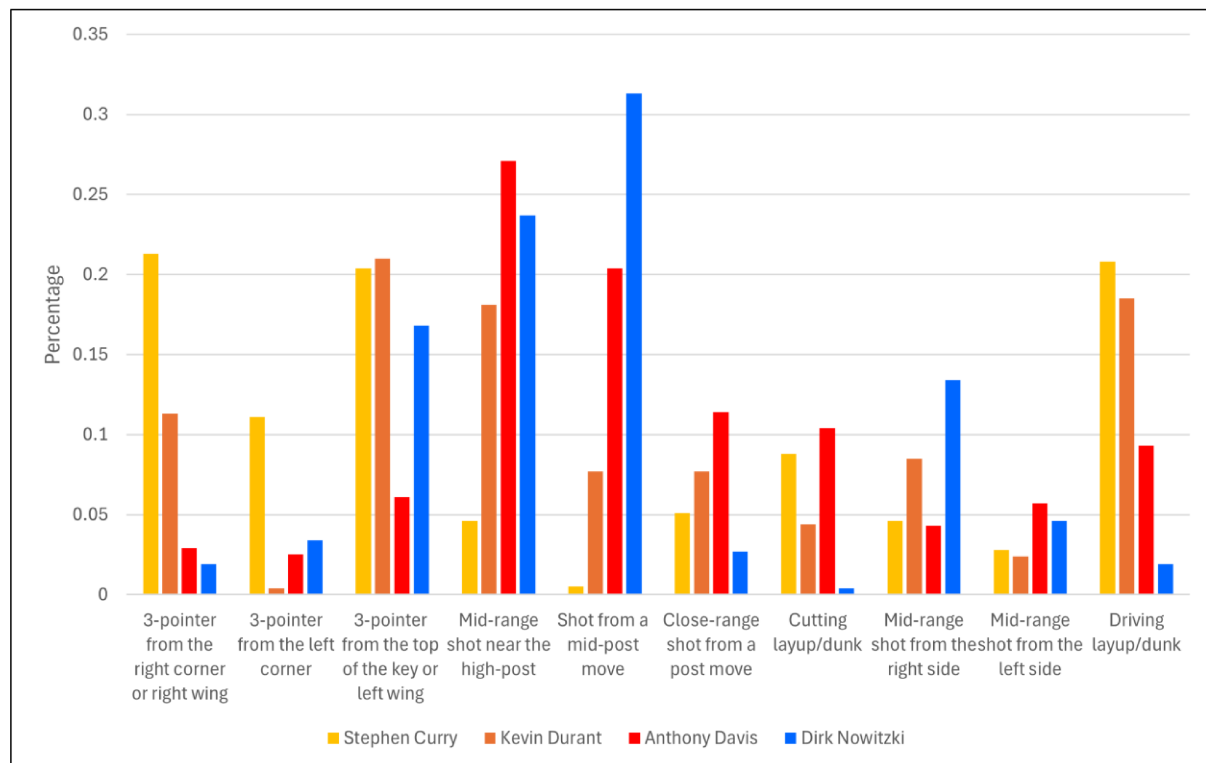


Figure 8: Percentage of each shot cluster for Stephen Curry, Kevin Durant, Anthony Davis, and Dirk Nowitzki. Note that the total number of shot data is 216, 248, 280, and 262, respectively.

Discussion

The clustering of shooting styles resulted in 13 interpretable clusters, which were considered to be specifically defined because they were larger in number than in previous studies (Hu et al. 2020, Fan et al. 2023) that used shot charts for clustering. Unlike these studies, our method does not allow us to visualize the distribution of each player's shot on the half-court. However, we were still able to gain a clearer understanding of a player's shooting tendencies by typifying the shots themselves, including the dynamic information immediately before the shot, and comparing how the percentage of each shot differs from player to player. We believe that quantifying the percentage of each shot type will make it easier to understand and capture players' shot tendencies.

The typology of shooting styles was found not only to aid in an intuitive understanding of players' shooting tendencies but also to provide a reasonable comparison of shot efficiency among players. We would argue that players should be evaluated relative to each other by making such comparisons between players with similar shooting tendencies, rather than simply comparing statistics such as TS% without taking into account players' shooting tendencies. Without doing so, it would not be a fair evaluation of the shot efficiency of players who have different roles in the offense.

With respect to the clustering method of the shooting styles, we were able to calculate the Wasserstein distance between players by taking into account the dissimilarity of their pre-shooting movements. The method used in this study, in which the set of shots for each player is considered to be a discrete uniform distribution, enables the calculation of dissimilarity between sets of vectors of the same number of dimensions and is therefore applicable to extracting some method of features and typifying the tendencies of the movements of each player for each

possession. Developing this method may allow for a data-driven typology of players' defensive roles that has not yet been done in previous studies.

One limitation of this study is that it only incorporates the planar movement of the players as dynamic information just prior to the shot. For example, although there are various types of close-range shots, such as dunks and layups, and various types of layups, such as overhand layups, underhand layups, and double clutches, we have not been able to distinguish between them. One way to solve this problem would be to include information on the player's posture estimated from the video image in the shot features and then perform clustering (example in soccer: Yeung et al. 2024).

In addition, the tracking data used in this study is from the 2015-16 season; different shooting styles and shot clusters would likely be extracted if data from later 3-point eras were used. In other words, the clustering results of this study are only one of the suggestions, and the main focus of this study is on the methodology itself, which is that tracking data should be used to capture players' shooting tendencies by taking movement information into account.

Conclusion

This study proposed a method for clustering the shooting styles of NBA players using tracking data and extracting shooting styles that conform to intuition. We also proposed a technique for successfully clustering the shots themselves to capture more specific shooting tendencies of players, and we were also able to extract clusters that are useful for comparison among players. These contributions may provide new directions for future research on capturing players' shooting preferences and encourage further development.

Data Availability Statements

The data and codes used in this study are available here: (the random seeds are fixed in the code to ensure reproducibility) <https://github.com/ky116/two-clusterings-to-capture-basketball-player-s-shooting-tendencies>.

Acknowledgements

This work was supported by Japan Society for the Promotion of Science KAKENHI (grant numbers 20H04075 and 23H03282) and Japan Science and Technology Agency PRESTO (grant number JPMJPR20CA).

References

- Aaditya, R., Nicol´as, T. & Marco, C. (2017), On wasserstein two-sample testing and related families of nonparametric tests, *Entropy*, 19(2), 47. doi: 10.3390/e19020047
- Alagappan, M. (2012), From 5 to 13: Redefining the positions in basketball, in ‘2012 MIT Sloan Sports Analytics Conference.’
- Arjovsky, M., Chintala, S. & Bottou, L. (2017), Wasserstein generative adversarial networks, in D. Precup & Y. W. Teh, eds, *Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, PMLR*, 70, 214–223. doi: 10.5555/3305381.3305404
- Barron, B., Sitaraman, S., N. & Arias, A., T. (2024), Analyzing NBA player positions and interactions with density-functional fluctuation theory, in ‘18th annual MIT Sloan Sports Analytics Conference’.
- Bianchi, F., Facchinetti, T. & Zuccolotto, P. (2017), Role revolution: towards a new meaning of positions in basketball, *Electronic Journal of Applied Statistical Analysis*, 10(3), 712–734.
- Bunker, R., Le Duy, V. N., Tabei, Y., Takeuchi, I. & Fujii, K. (2023), Multi-agent statistical discriminative sub-trajectory mining and an application to nba basketball, arXiv preprint arXiv:2311.16564 .
- Chen, R., Zhang, M. & Xu, X. (2023), Modeling the influence of basketball players’ offense roles on team performance, *Frontiers in Psychology*, 14. doi: 10.3389/fpsyg.2023.1256796
- Fan, Y., Guanyu, H. & Shen, W. (2023), Analysis of professional basketball field goal attempts via a bayesian matrix clustering approach, *Journal of Computational and Graphical Statistics*, 32(1), 49–60. doi: 10.1080/10618600.2022.2085727
- Fujii, K., Inaba, Y. & Kawahara, Y. (2017), Koopman spectral kernels for comparing complex dynamics: Application to multiagent sport plays, in ‘European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD’17)’, *Springer*, 127–139. doi: 10.1007/978-3-319-71273-4_11
- Fujii, K., Kawasaki, T., Inaba, Y. & Kawahara, Y. (2018), Prediction and classification in equation-free collective motion dynamics, *PLoS Computational Biology*, 14(11): e1006545.
- Fujii, K., Takeishi, N., Hojo, M., Inaba, Y., and Kawahara, Y. (2020). Physically-interpretable classification of network dynamics for complex collective motions. *Scientific Reports*, 10(3005).
- Gower, J. C. (1971), A general coefficient of similarity and some of its properties, *Biometrics*, 27(4), 857– 871. doi: 10.2307/2528823
- Hojo, M., Fujii, K., Inaba, Y., Motoyasu, Y. & Kawahara, Y. (2018), Automatically recognizing strategic cooperative behaviors in various situations of a team sport, *PLoS One*, 13(12): e0209247.
- Hojo, M., Fujii, K. & Kawahara, Y. (2019), Analysis of factors predicting who obtains a ball in basketball rebounding situations, *International Journal of Performance Analysis in Sport*, 1–14.
- Hu, G., Yang, H.-C. & Xue, Y. (2020), Bayesian group learning for shot selection of professional basketball players, *Stat*, 10(1), e324, doi: 10.1002/sta4.324.
- Hua, G. & Su, C. (2023), Estimating positional plus-minus in the nba, in ‘17th annual MIT Sloan Sports Analytics Conference.’
- Hvattum, Magnus, L. (2019), A comprehensive review of plus-minus ratings for evaluating individual players in team sports, *International Journal of Computer Science in Sport*, 18(1), 1–23.

- Ishida, A., Takayanagi, M., Hoshina, I. & Iwayama, K. (2023), Bayesian credible possession based player performance evaluation in basketball, *Keisankitoukeigaku [Journal of the Japanese Society of Computational Statistics]*, 36(2), 99–126.
- Jiao, J., Hu, G. & Yan, J. (2021), A bayesian marked spatial point processes model for basketball shot chart, *Journal of Quantitative Analysis in Sports*, 17(2), 77–90. doi: 10.1515/jqas-2019-0106
- Kalman, S. & Bosch, J. (2020), NBA lineup analysis on clustered player tendencies: A new approach to the positions of basketball & modeling lineup efficiency of soft lineup aggregates, in ‘14th annual MIT Sloan Sports Analytics Conference’.
- Kanda, S., Takeuchi, K., Fujii, K. & Tabei, Y. (2020), Succinct trit-array trie for scalable trajectory similarity search, *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 518–529.
- Kempe, M., Grunz, A. & Daniel, M. (2014), Detecting tactical patterns in basketball: Comparison of merge self-organising maps and dynamic controlled neural networks, *European Journal of Sport Science*, 15(4), 249–255.
- Lutz, D. (2012), A cluster analysis of nba players, in ‘2012 MIT Sloan Sports Analytics Conference.’
- McInnes, L., Healy, J., Saul, N. & Großberger, L. (2018), Umap: Uniform manifold approximation and projection, *Journal of Open Source Software*, 3(29), 861. doi: 10.21105/joss.00861
- McIntyre, A., Brooks, J., Guttag, J. & Wiens, J. (2016), Recognizing and analyzing ball screen defense in the NBA, *Proceedings of the MIT Sloan Sports Analytics Conference*, 11–12.
- McQueen, A., Wiens, J. & Guttag, J. (2014), Automatically recognizing on-ball screens, in *Proceedings of the MIT Sloan Sports Analytics Conference*.
- Miller, A., Bornn, L., Adams, R. & Goldsberry, K. (2014), Factorized point process intensities: A spatial analysis of professional basketball, in E. P. Xing & T. Jebara, eds, *Proceedings of the 31st International Conference on Machine Learning, Proceedings of Machine Learning Research, PMLR*, 32, 235–243. doi: 10.5555/3044805.3044833
- Miller, A. C. & Bornn, L. (2017), Possession sketches: Mapping NBA strategies, in *Proceedings of the MIT Sloan Sports Analytics Conference*.
- Muniz, M. & Flamand, T. (2022), A weighted network clustering approach in the nba, *Journal of Sports Analytics*, 8(4), 1–25.
- Murtagh, F. & Legendre, P. (2011), Ward’s hierarchical clustering method: Clustering criterion and agglomerative algorithm.
- Nistala, A. (2018), Using deep learning to understand patterns of player movement in basketball, PhD thesis, Massachusetts Institute of Technology.
- Papalexakis, E. & Pelechrinis, K. (2018), thoops: A multi-aspect analytical framework for spatiotemporal basketball data, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2223–2232.
- Peršič, M., Kristan, M., Kovačič, S., Vučković, G. & Peršič, J. (2009), A trajectory-based analysis of coordinated team activity in a basketball game, *Computer Vision and Image Understanding*, 113(5), 612–621.
- Sha, L., Lucey, P., Yue, Y., Carr, P., Rohlf, C. & Matthews, I. (2016), Chalkboarding: A new spatiotemporal query paradigm for sports play retrieval, in ‘International Conference on Intelligent User Interfaces’, 336–347.
- van der Maaten, L. & Hinton, G. (2008), Visualizing data using t-sne, *Journal of Machine Learning Research*, 9(86), 2579–2605.
- Wang, K.-C. & Zemel, R. (2016), Classifying nba offensive plays using neural networks, *Proceedings of the MIT Sloan Sports Analytics Conference*.

- Wang, X., Han, B., Zhang, S., Zhang, L., Lorenzo Calvo, A. & Gomez, M.- (2022), The differences in the performance profiles between native and foreign players in the chinese basketball association, *Frontiers in Psychology*, 12. doi: 10.3389/fpsyg.2021.788498
- Whitehead, T. (2018), Nylon calculus: Defining 23 offensive roles using the NBA's play-type data. Retrieved November 22, 2024, from <https://fansided.com/2018/10/15/nylon-calculus-defining-23-offensive-roles-nba-play-type-data>
- Yanai, C., Solomon, A., Katz, G., Shapira, B. & Rokach, L. (2022), Q-ball: Modeling basketball games using deep reinforcement learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 8806–8813.
- Yeung, C., Ide, K. & Fujii, K. (2024), Autosoccerpose: Automated 3d posture analysis of soccer shot movements, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 3214–3224.
- Zhang, L., Lu, F., Liu, A., Guo, P. & Liu, C. (2016), Application of k-means clustering algorithm for classification of nba guards, *International Journal of Science and Engineering Applications*, 5(1), 1–6.
- Zhang, Z., Takeda, K. & Fujii, K. (2022), Cooperative play classification in team sports via semisupervised learning, *International Journal of Computer Science in Sport*, 21(1), 111–121.